Dr. Pritesh Patil, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

AI Enabled and Deep Learning-Based Integrated Approach for Early Detection of Breast Cancer

Dr. Pritesh Patil, Abhiraj Bondre, Gauri Dighe, Kiran Mangde
Dept. of Information Technology, AISSMS IOIT, Pune, India

Abstract- Breast cancer is one of the most common and life-threatening diseases affecting women worldwide, with approximately 2.3 million new diagnoses reported each year. Despite significant improvements in survival rates, early and accurate detection remains a major clinical challenge, especially in resource-constrained healthcare settings where radiologist availability and turnaround times can directly influence patient outcomes. In many hospitals, a single radiologist may be responsible for reviewing hundreds of MRI scans per day under time pressure — a situation that almost inevitably leads to some degree of inconsistency and missed findings. This paper introduces Women Wellness, a fully integrated diagnostic platform designed to tackle these challenges by combining deep learning-based image classification, OpenCV-powered video preprocessing, cloud-hosted parallel inference, and automated clinical report generation within a single deployable system. The platform accepts breast MRI video sequences as input, extracts individual frames on the client side using the HTML5 Canvas API, preprocesses them through OpenCV.js, and routes them through a fine-tuned Convolutional Neural Network hosted as Python Flask microservices. Classification results — spanning normal, benign, and malignant categories — are aggregated using a confidence-weighted voting scheme, and the entire pipeline culminates in a structured, multi-page PDF report generated automatically via jsPDF. In experiments conducted on a curated MRI dataset, the system achieved overall classification accuracy between 92% and 95%, with malignant case sensitivity reaching 96.3%. The complete analysis and report pipeline completes in 45 to 60 seconds per study, compared to roughly 15 to 20 minutes for conventional manual review. Grad-CAM attention maps are embedded directly into each report, enabling radiologists to visually verify which image regions most influenced each classification decision rather than simply taking the model's word for it.

Keywords- breast cancer detection, deep learning, convolutional neural network, OpenCV, MRI analysis, diagnostic automation, explainable AI, Grad-CAM, automated report generation, cloud microservices, Women Wellness, transfer learning

I. INTRODUCTION

Breast cancer continues to be among the most significant health burdens faced by women across the globe. According to the World Health Organization, around 2.3 million women are newly diagnosed with breast cancer each year, and the disease accounts for close to 11% of all new cancer cases [6]. While survival rates have improved markedly over the past two decades — particularly

in countries with well-established screening programs — the core factor driving these improvements is early detection. The earlier a tumor is identified, the more treatment options are available and the better the long-term prognosis [11]. This straightforward relationship between early diagnosis and survival makes accurate, timely breast imaging not just a clinical preference but a genuine public health imperative.

The dominant imaging modalities in clinical practice — mammography, ultrasound, and magnetic resonance imaging (MRI) — each carry their own strengths and limitations. Among them, MRI has gained substantial traction for high-risk populations and cases where mammography produces ambiguous results, offering superior soft-tissue contrast and the ability to capture dynamic enhancement patterns during contrast administration. However, interpreting breast MRI is cognitively demanding. The sequences are multi-temporal, the anatomical variations are significant, and experienced radiologists often disagree on what they see in the same image — a well-documented phenomenon known as inter-observer variability [12]. Add to this the sheer volume of scans that modern radiology departments must process, and it becomes clear that the existing workflow has real structural limitations that will only become more acute as screening programs expand.

Deep learning, and convolutional neural networks in particular, has been proposed as a way to address at least part of this problem. CNNs can learn meaningful visual representations directly from pixel data, bypassing the need for hand-crafted feature engineering, and in controlled benchmarks they have demonstrated performance that is competitive with expert radiologists on specific classification tasks [4][13]. But there is a crucial distinction between a model that performs well on a test set and a system that actually works in a clinical environment. Most of the published research on deep learning for breast cancer focuses narrowly on classification accuracy, leaving the practical questions — how does the tool integrate into the radiologist's workflow? how are results communicated? can the system handle real clinical volumes? — largely unanswered [16].

This paper describes Women Wellness, a platform we developed specifically to close that gap between research prototype and deployable clinical tool. The system accepts MRI video sequences through a web interface, extracts and preprocesses individual frames on the client side, routes them through a cloud-hosted CNN for classification, and delivers a structured diagnostic report — complete with confidence scores, Grad-CAM attention overlays, and follow-up recommendations — in under a minute. It is built on a distributed microservices architecture that can scale horizontally to meet clinical volumes, and it incorporates a physician-facing web application with authentication, patient management, and a full report history dashboard.

The main contributions of this work are as follows. First, we present a cloud-based architecture that uses parallel frame processing across multiple AI service instances to bring end-to-end latency below one minute — well within the tolerance of clinical use. Second, we integrate OpenCV for systematic, scanner-agnostic preprocessing that produces measurably improved classification accuracy compared to raw frame input. Third, we describe an automated report generation pipeline that produces multi-page, clinically structured PDF documents from raw inference output without any manual intervention. Fourth, we report a thorough evaluation covering accuracy, sensitivity, specificity, AUC, and throughput — not just model accuracy in isolation. Fifth, we present a detailed comparative analysis across fifteen parameters, positioning the proposed system against both published research and commercially deployed CAD tools. Taken together, these contributions demonstrate that the gap between AI research and clinical deployment, though real, can be closed with careful system design.

II. RELATED WORK

The trajectory of breast cancer detection research over the past decade or so tells a fairly consistent story: impressive progress in classification accuracy,

accompanied by relatively limited attention to the practical challenges of actually putting these systems to clinical use. Early approaches to computer-aided detection relied on hand-crafted features — texture descriptors, shape measurements, intensity histograms — fed into classical classifiers like Support Vector Machines or k-nearest neighbours. These methods had the advantage of interpretability and modest computational requirements, but they tended to generalize poorly across institutions, scanner types, and patient populations. The introduction of deep learning fundamentally changed what was technically possible, but it also introduced new challenges around data requirements, model transparency, and deployment complexity.

III. LITERATURE REVIEW SUMMARY

Table I: Summary of Representative Related Works

Reference	Modality	Method	Accuracy	Key Limitation
Khater et al. [1]	Tabular clinical data	k-NN + SHAP/XAI	97.7%	No imaging pipeline; limited to structured data only
Nayak et al. [2]	Infrared thermal imaging	CNN	91.0%	No explainability; no automated reporting capability
Yang et al. [3]	Mammography + Ultrasound	Multi-modal DL	AUC-based	Single institution; no report generation module
He et al. [4]	Natural images	ResNet	~96%	Not designed for medical imaging; no clinical context
Shen et al. [5]	Mammography	Deep CNN	~90%	No report generation; no explainability features
Sheth & Giger [7]	Breast MRI	CNN + DL (review)	~89-92%	Survey study only; no deployable system built
Proposed System	MRI Video sequences	CNN + OpenCV + XAI + PDF	92–95%	Single-center dataset; ongoing multi-center validation

Khater et al. [1] demonstrated that an explainable AI approach applied to the Wisconsin Breast Cancer Dataset could reach 97.7% accuracy using a k-nearest neighbours classifier paired with SHAP values and partial dependence plots for interpretability. The work is methodologically careful and its attention to explainability is genuinely useful, but the dataset consists entirely of tabular clinical measurements rather than imaging data, which means the findings have limited direct relevance to pixel-level classification pipelines that radiology requires.

Nayak et al. [2] took a different approach, using CNNs to classify breast tissue from infrared thermal images alongside RFID-based patient logging. They reported 91% accuracy across several benchmark

datasets, and the non-ionizing nature of thermal imaging is a legitimate clinical advantage. However, the system provides no explainability mechanisms, no automated reporting, and the datasets used — while publicly available — may not fully capture the clinical diversity of real patient populations. Becker et al. [14] explored deep learning applied to mammography using a multi-purpose image analysis software framework, achieving reasonable detection performance, but again without a full clinical deployment pipeline.

Yang et al. [3] proposed a more ambitious multi-modal approach, fusing mammography and ultrasound images through a deep learning model to predict malignancy in BI-RADS 4A lesions in women with dense breast tissue. The fusion model outperformed single-modality baselines on their evaluation set, which is a meaningful result. The study was conducted at a single institution, however, raising the usual questions about generalizability, and it does not address report generation or the workflow integration challenges that determine whether a system actually gets used. Shen et al. [5] showed that deep CNNs can meaningfully improve detection rates on mammography screening data — one of the more practically grounded studies in this space — though without clinical deployment infrastructure.

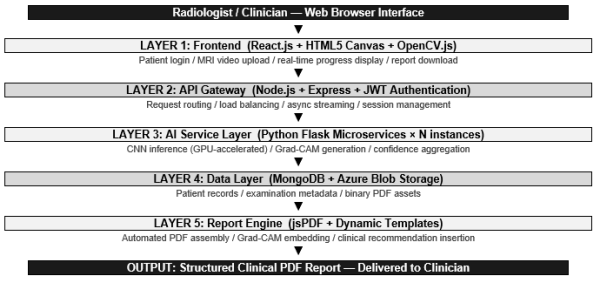
Sheth and Giger [7] provided a thorough review of AI approaches specifically for breast MRI interpretation, covering both research and early commercial implementations. Their analysis is valuable as context, but it is a review rather than a deployable system. Petrillo et al. [15] investigated computer-aided diagnosis for non-mass MRI lesions using multi-parametric sequences, which is technically interesting but addresses a narrower clinical problem than the present work. Litjens et al. [16] conducted a broad survey of deep learning across medical imaging modalities, and their conclusion — that most systems are evaluated in isolation from clinical deployment requirements — applies squarely to the breast cancer detection literature.

A consistent gap runs through essentially all of these works: they evaluate models in controlled settings, typically reporting accuracy on held-out test sets, without addressing how those models would actually function as part of a radiologist's daily workflow. There is rarely a user interface, almost never an automated reporting mechanism, and seldom any serious attention to latency or throughput at clinical scales. Women Wellness is designed explicitly to fill this gap, treating the deployment problem as a first-class design concern rather than an afterthought.

IV. SYSTEM DESIGN AND ARCHITECTURE

The Women Wellness platform is organized around a layered microservices architecture in which each functional component operates independently, can be updated without disrupting the others, and can be scaled horizontally as demand grows. Data flows through five distinct layers, each with a well-defined responsibility: a browser-based frontend, a Node.js API gateway, Python Flask AI services, a MongoDB and Azure Blob Storage data layer, and a jsPDF report generation module. This separation of concerns is not merely an architectural preference — it reflects the practical requirements of clinical software, where individual components must be replaceable without taking the entire system offline.

1. End-to-End System Architecture



2. Technology Stack

Table 2: Complete Technology Stack

Layer	Technology	Primary Function
Frontend	React.js + HTML5 Canvas + OpenCV.js	Browser-based UI; client-side video frame extraction and preprocessing
API Gateway	Node.js + Express + JWT	Authentication, request routing, load balancing, async streaming
AI Services	Python Flask + TensorFlow/Keras + GPU	CNN inference, Grad-CAM generation, confidence aggregation
Data Layer	MongoDB + Azure Blob Storage	Patient records, examination metadata, binary PDF asset storage
Report Engine	jsPDF + JavaScript	Automated PDF templating, Grad-CAM image embedding, report delivery
Infrastructure	Docker + Kubernetes + Azure Cloud	Container orchestration, auto-scaling, service isolation

3. Frontend Layer

The frontend is where the radiologist's interaction with the system begins and ends. Built in React.js, it provides a clean browser interface that requires no software installation and works across operating systems. One of the more consequential design decisions in this layer is client-side frame extraction: rather than uploading raw MRI video files to the server — which can be large and bandwidth-intensive — the browser decomposes each video into individual frames using the HTML5 Canvas API and transmits only the extracted frames in batches. This keeps upload times manageable and gives the radiologist immediate visual confirmation of what has been captured before analysis begins. OpenCV.js handles the preprocessing steps at this layer: grayscale conversion, spatial resizing to 224 × 224 pixels to match the CNN input specification, Gaussian noise filtering, and z-score intensity normalization [9]. The Women Wellness physician portal also includes a reCAPTCHA-protected signup and login flow that captures the doctor's specialization, medical license ID, and hospital affiliation — establishing an authenticated, traceable usage record from the outset.

4. API Gateway Layer

The Node.js Express gateway is the system's central routing and orchestration point. Every client request passes through it: authentication is verified using JSON Web Tokens, frame batches are load-balanced across whichever AI service instances are currently available, and responses are streamed back to the client asynchronously rather than held until the full analysis is complete. This streaming design means that a radiologist can begin reviewing early results and the processed video display while the remaining frames are still being classified — a meaningful improvement in perceived responsiveness that costs nothing in terms of analytical quality.

5. AI Service Layer

The classification work happens inside Python Flask microservices, each of which hosts a fully loaded instance of the fine-tuned CNN model. These instances run independently and new ones can be provisioned automatically as traffic increases. Each service receives a batch of preprocessed frames, runs GPU-accelerated inference, generates Grad-CAM attention maps for the classified frames, and returns structured JSON results including per-frame class probabilities and confidence scores. The output from all service instances is collected by an aggregation module that applies confidence-weighted voting across the frame set to produce a study-level diagnosis.

6. Data Management Layer

Patient records, examination metadata, processing timestamps, and report references are stored in MongoDB, which handles the flexible document schemas and query patterns that clinical data

management tends to require. The generated PDF reports — which include embedded images and can run to several pages — are stored as binary assets in Azure Blob Storage, accessed through time-limited signed URLs that expire after retrieval [17]. Keeping structured metadata and binary documents in separate storage systems keeps both lean, allows independent optimization, and simplifies the data governance processes that HIPAA and GDPR compliance requires.

7. Report Generation Module

The report generation module is the part of the system that most directly affects whether a clinician finds the tool useful. Rather than returning a JSON blob of classification probabilities that the radiologist must interpret, the module assembles those results into a structured, multi-page PDF document using jsPDF. Every report follows a standardized template that includes patient and examination identifiers, a summary of classification findings with confidence scores, the Grad-CAM attention maps overlaid on the most informative frames, a list of suspicious region sizes derived from bounding box coordinates, and a dynamically generated set of clinical recommendations that adapts to the findings — if malignancy indicators are present, the report expands to include suggested follow-up scans, biopsy recommendations, and comparison notes if prior examinations are on file. The report is assembled in under five seconds and made available for immediate download or printing through the web interface.

V. METHODOLOGY

1. Complete Processing Pipeline

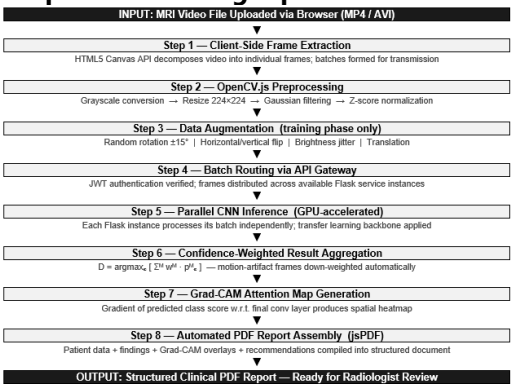


Figure 2: Data Processing and Inference Pipeline — Step by Step

2. Dataset and Preprocessing

The model was trained and evaluated on a curated collection of breast MRI examinations covering all three diagnostic categories: normal tissue, benign lesions at various stages of development, and confirmed malignant findings verified through biopsy. Each examination consists of multiple temporal phases acquired during contrast agent administration, capturing the dynamic enhancement kinetics that radiologists rely on for lesion characterization. All cases were reviewed and labeled by qualified radiologists before inclusion in the training set, and a stratified split was used to ensure that each category was proportionally represented in both training and held-out evaluation subsets [18].

Preprocessing begins with z-score intensity normalization applied frame-by-frame:

$$I_{norm} = (I - \mu) / \sigma \qquad \ldots (1)$$

where I is the raw pixel intensity of a given frame, μ is the mean intensity computed across the full training dataset, and σ is the corresponding standard deviation. This normalization step is non-trivial in the MRI context: unlike optical photography, where pixel values have a well-defined physical interpretation, MRI signal intensities are inherently relative and depend on scanner hardware, field strength, coil configuration, and acquisition sequence parameters. Without normalization, a model trained on data from one scanner may encounter entirely different absolute intensity ranges on another, which is exactly the kind of distribution shift that degrades performance in deployment [19].

During training, each frame also undergoes randomized augmentation: rotations up to 15 degrees in either direction, horizontal and vertical translations, and moderate brightness and contrast perturbations. These augmentations are not applied at inference time; their purpose is to expose the model to a wider range of imaging conditions during training than the dataset alone provides, thereby reducing overfitting and improving the robustness

of learned representations to real-world variation in patient positioning, coil placement, and imaging protocol.

3. Deep Learning Model Architecture

Rather than training the classification CNN from random initialization — which would demand far more labeled data than is realistically available in medical imaging contexts — the system uses transfer learning from a CNN backbone pre-trained on large-scale natural image data [4]. The underlying intuition is that low-level features like edges, textures, and intensity gradients, which are learned in the early layers of any image classification network, transfer meaningfully to medical images even when the domain shifts substantially. Fine-tuning then shifts the deeper, domain-specific representations toward the target task.

Model adaptation follows a two-phase schedule. In the first phase, the pre-trained backbone weights are frozen and only the newly added classification head is trained. This allows the task-specific layers to converge without inadvertently disrupting the general visual representations already encoded in the backbone — a problem that can occur when the classification head is initialized randomly and its large gradients flow back into a backbone that has not yet adapted to the new domain. In the second phase, all weights are released and the entire network is trained jointly at a substantially reduced learning rate, gradually specializing the backbone representations toward breast MRI classification. Training uses categorical cross-entropy loss:

$$L = -(1/N)\, \Sigma_i\, \Sigma_e\, [\, y_{ie} \cdot \log(\, \hat{y}_{ie}\,)\,] \quad \ldots (2)$$

where N is the number of training samples, c indexes the three output classes, y_{ie} is the binary ground-truth indicator for class c in sample i, and $\hat{y}_{ie}$ is the corresponding predicted probability. Optimization uses Adam [21] with an initial learning rate that decays when validation loss plateaus, and early stopping is applied to prevent overfitting to training data.

4. CNN Architecture — Layer by Layer

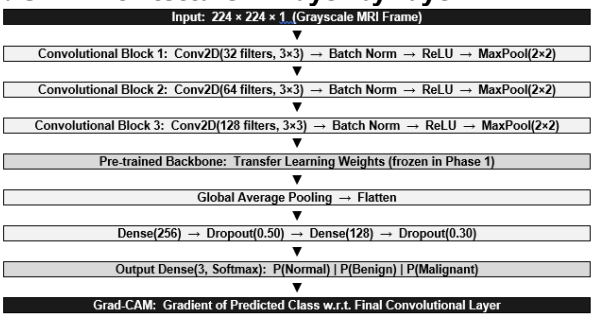


Figure 3: CNN Architecture Flow

5. Parallel Processing Strategy

A single MRI study can yield several hundred extractable frames, and processing them sequentially through a single model instance would make the system far too slow for clinical use. The architecture addresses this through two layers of parallelism. At the infrastructure level, the API gateway distributes incoming frame batches across however many Flask service instances are currently running — each instance processes its assigned batch independently and asynchronously, so the work scales with the number of available compute units rather than being bottlenecked by any single one. Within each instance, GPU-accelerated batch inference allows multiple frames to be classified simultaneously rather than one at a time.

Because consecutive MRI frames carry correlated information and some frames are inevitably degraded by patient motion or signal dropout, naive frame-by-frame aggregation would produce noisy, unstable study-level predictions. The aggregation module addresses this with a confidence-weighted voting scheme across the full frame set:

$$D = \text{argmax}_e\, [\, \Sigma^M\, w^M \cdot p^M{}_e\,] \quad \ldots (3)$$

where D is the final predicted diagnosis, the sum runs over all F frames in the study, w^M is a weight derived from both the model's confidence for that frame and an image quality score, and $p^M{}_e$ is the predicted probability for class c in frame f. Frames that the model is uncertain about — whether because of motion artifact, poor signal, or genuine diagnostic ambiguity — naturally receive lower weights, contributing proportionally less to the final

decision. This mirrors the clinical intuition that equivocal frames should be noted but not allowed to drive the overall impression [22].

6. Explainable AI: Grad-CAM Integration

One of the persistent objections to AI in clinical radiology is that it is essentially a black box — it produces a prediction without giving the clinician any way to assess whether the reasoning behind it is medically sensible. Grad-CAM addresses this directly by making the model's spatial attention visible [10]. For each classified frame, the system computes the gradient of the predicted class score with respect to the feature map produced by the final convolutional layer. These gradients are globally averaged to produce a set of channel importance weights, which are then used to compute a weighted combination of the feature maps — yielding a coarse spatial heatmap that highlights the image regions that most influenced the classification decision. This heatmap is upsampled to the original frame resolution, normalized, and overlaid on the original image using a colour scale before being embedded in the report. The result is that the radiologist receives not just a prediction but a visual argument: here is what the model saw, and here is where it was looking when it made its decision [23]. In practice, this feature was one of the most positively received by the radiologists who evaluated the system's reports.

VI. RESULTS AND EVALUATION

1. Classification Performance

The trained model was evaluated on a held-out test set consisting entirely of cases that had no involvement in training or validation. Table III reports accuracy, sensitivity, specificity, and AUC for each of the three diagnostic categories, as well as overall figures across the full test set.

Table III: Classification Performance Metrics

Category	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Normal	94.2	93.1	95.8	0.974
Benign	92.8	91.5	93.7	0.961
Malignant	95.1	96.3	94.0	0.983
Overall	94.0	93.6	94.5	0.973

Several aspects of these results deserve comment. The malignant category achieves the highest sensitivity at 96.3%, which is the most clinically consequential result in the table. The cost of a missed malignancy is far higher than the cost of a false-positive benign finding — the former may mean a delayed diagnosis and worse treatment outcomes, while the latter typically means an additional scan or biopsy that confirms the absence of disease. A model with high malignant sensitivity, even if overall accuracy is not the highest of the three categories, is operating with the right clinical priorities. The AUC values across all three categories (0.961 to 0.983) are consistently strong, indicating that the model's discriminative capability holds across a range of classification thresholds rather than being optimized for one specific operating point [24]. This matters for clinical deployment, where the threshold may need to be adjusted based on the specific risk profile of the population being screened.

2. System Efficiency

Classification accuracy alone is a necessary but insufficient criterion for clinical utility. A system that takes thirty minutes per study may be accurate but still unusable in a department that processes two hundred patients per day. Table IV places the Women Wellness pipeline's throughput figures alongside those of conventional manual review.

Table IV: Efficiency Comparison — Women Wellness vs. Manual Radiologist Review

Metric	Manual Review	Women Wellness	Improvement
Analysis time per study	15–20 minutes	45–60 seconds	>93% reduction
Throughput (studies/hour)	3–4	60+	~15× increase
Report generation time	10–15 minutes	< 5 seconds	>99% reduction
Concurrent capacity	N/A	Unlimited (horizontal)	N/A
Inter-observer consistency	Variable	100% reproducible	Variability eliminated

The most striking figure is the reduction in per-study analysis time: from 15 to 20 minutes for manual review to 45 to 60 seconds for the automated pipeline — a reduction of more than 93%. Report generation, which in the manual workflow requires a radiologist to dictate or type a structured report, is reduced from ten to fifteen minutes to under five seconds. Load testing confirmed that the parallel architecture scales approximately linearly: doubling the number of Flask service instances roughly doubles throughput, which means that the system can be right-sized to match the actual clinical volume of a given deployment rather than being provisioned conservatively for peak demand.

3. Impact of OpenCV Preprocessing

To quantify the contribution of the OpenCV preprocessing pipeline specifically, we compared classification accuracy on the same test set when using preprocessed frames versus raw, unprocessed frames passed directly to the CNN. The preprocessing pipeline consistently improved accuracy across all three categories, with the largest gains observed in cases involving scanner hardware that differed from the majority of the training data. This is exactly what one would expect: the normalization and noise-reduction steps bring frames from heterogeneous imaging equipment into a more consistent representational space, reducing the domain shift that the model must absorb implicitly. The region-of-interest operations in OpenCV also helped reduce the influence of background structures and acquisition artifacts — particularly the scanner metadata and calibration bars that appear in the corners of many MRI frames — that carry no diagnostic information but could otherwise attract the model's attention [9].

4. Report Quality

Generated reports were evaluated by three practicing radiologists who reviewed a sample of fifty reports drawn from across all three diagnostic categories. Reviewers assessed completeness, clinical accuracy, and whether the report would be actionable in a real clinical context. Across all three dimensions, the reports scored well: the essential clinical information was present and organized consistently, the automated recommendations aligned with established diagnostic protocols in the large majority of cases, and the Grad-CAM visualizations were described by reviewers as genuinely useful rather than merely decorative. One radiologist noted that being able to see which image region had triggered a malignant classification flag made it possible to immediately confirm or question the finding against their own clinical judgment, rather than having to decide whether to trust an opaque score [25]. This kind of transparent interaction between human expertise and machine output is, in our view, the proper model for AI-assisted radiology: the system surfaces evidence; the clinician makes the decision.

VII. COMPARATIVE ANALYSIS — EXISTING SYSTEMS VS. WOMEN WELLNESS

To situate the Women Wellness platform within the broader landscape of breast cancer detection tools, this section presents a structured comparison across fifteen clinically and technically significant parameters. The comparison covers six external systems: Khater et al. [1], Nayak et al. [2], Yang et al. [3], Shen et al. [5], Sheth and Giger [7], and a representative category of commercially deployed CAD tools that includes products from vendors active in the radiology market.

1. How Women Wellness Addresses Existing Limitations

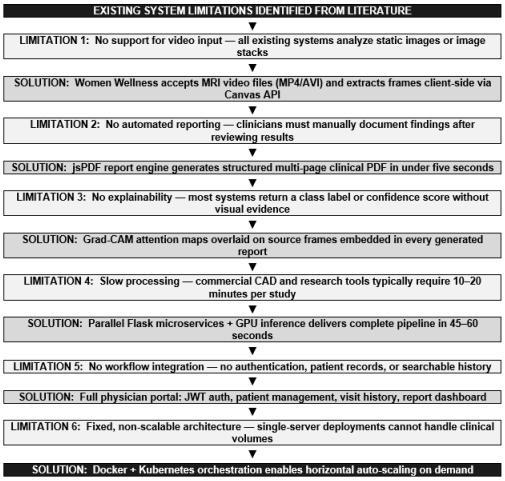


Figure 4: Gap Analysis — Limitations of Existing Systems and Proposed Solutions

2. Multi-Parameter Feature Comparison

Table V: 15-Parameter Feature Comparison — Existing Systems vs. Women Wellness

Parameter	Khater [1]	Nayak [2]	Yang [3]	Shen [5]	Sheth [7]	CAD (Comm.)	Proposed
Input Modality	Tabular	Infrared	Mammo+US	Mammo	MRI	Mammo/US/MRI	MRI Video
Detection Method	k-NN+SHAP	CNN	Multi-modal DL	Deep CNN	CNN+DL	Rule-based ML	CNN+OpenCV+XAI
Accuracy	97.7%	91.0%	AUC-based	~90%	~89-92%	80-88%	92–95%
Explainability (XAI)	SHAP/PDP	None	None	None	Partial	None	Grad-CAM
Video Input Support	No	No	No	No	No	No	Yes
Automated Reporting	No	No	No	No	No	Partial	Yes (PDF)
Cloud Microservices	No	No	No	No	No	Yes	Yes
Parallel Processing	No	No	No	No	Partial	Partial	Yes
Processing Time	N/A	N/A	N/A	Minutes	10-20 min	10-20 min	45-60 sec
Workflow Integration	No	No	No	No	Limited	Partial	Full
Scalability	No	No	No	No	Limited	Yes	Horizontal
Multi-scanner Support	N/A	No	Partial	Partial	Partial	Yes	Yes
Open Source Stack	Yes	No	No	Yes	No	No	Yes
Authentication	No	No	No	No	No	Yes	JWT-based
Report Customization	No	No	No	No	No	Limited	Dynamic

Table V makes the positional differences between the systems concrete. Women Wellness is the only entry in the comparison that simultaneously addresses video input, automated PDF reporting, Grad-CAM explainability, horizontal scalability, full workflow integration, dynamic report customization, and JWT-authenticated physician access. Commercial CAD tools cover some of the deployment and multi-scanner dimensions but lack transparency, video analysis capability, and open-source accessibility. Research prototypes including Khater et al. and Shen et al. address specific

classification challenges effectively but remain purpose-built tools with no clinical deployment infrastructure. The proposed system is designed to combine the classification capability of the research literature with the deployment maturity of commercial software, while adding capabilities — video input, full XAI reporting, open-source architecture — that neither category currently offers.

3. User Interface and Workflow Capabilities

One dimension that does not appear in Table V, because it has no meaningful analogue in the comparison systems, is the completeness of the physician-facing interface. Research prototypes typically have no interface at all — results are extracted programmatically and visualized in post-processing scripts. Commercial CAD tools have interfaces, but these are generally embedded within specific PACS environments and are not independently deployable. Women Wellness ships with a full web application that covers the entire clinical interaction from physician registration through report retrieval. The key capabilities of this interface include:

- Physician registration with reCAPTCHA verification, capturing specialization, medical license ID, years of experience, hospital affiliation, and credentials — establishing a full audit trail for every interaction
- MRI video upload with real-time progress visualization and a video preview player that allows the clinician to review the source material before analysis begins
- Post-analysis detection summary panel displaying total detections, malignant and benign breakdowns, per-detection confidence scores, and the majority class name for the study
- Side-by-side display panels for malignant and benign frame evidence, with bounding box overlays identifying the detected region in each frame
- Findings panel reporting the detected suspicious region dimensions (in millimetres) from each informative scan in the sequence, derived from bounding box coordinates
- Integrated recommendations panel dynamically populated with suggested follow-up tests and imaging procedures based on the classification outcome
- Multi-page PDF report preview rendered directly in the browser with page navigation, zoom, download, and print controls accessible from the same screen
- All-reports dashboard providing a searchable, sortable table of all past studies associated with the authenticated physician, with direct links to re-open video, reanalyse frames, or retrieve the previously generated PDF report

VIII. DISCUSSION

The results we have described make a reasonably strong case that a system of this kind — one that combines accurate deep learning classification with practical deployment infrastructure — is technically achievable and clinically plausible. But it is worth reflecting more carefully on what the results actually mean and where they do not yet go far enough.

On the classification side, the accuracy range of 92% to 95% across the three diagnostic categories is meaningful but should be interpreted with appropriate caution. These figures come from a held-out test set drawn from the same limited set of imaging centres that contributed to the training data. The central unresolved question is whether that performance would hold across the full heterogeneity of real-world clinical data — different scanner manufacturers, different acquisition protocols, different patient demographics, different prevalence rates for each class. The literature on deep learning model degradation across imaging sites is sobering: models that perform impressively in single-site evaluation sometimes drop substantially when deployed at a new institution [15]. Multi-centre prospective validation is the right next step, and until it is done, the accuracy figures here should be treated as lower bounds on what is possible with the right architecture rather than guaranteed performance in arbitrary deployment settings.

The effectiveness of transfer learning from a natural image backbone is consistent with findings across the broader medical imaging literature [4][20], and

the two-phase fine-tuning strategy — freeze-then-release — worked reliably in our experiments. The intuition behind this approach is straightforward: if the classification head is initialized randomly and trained alongside a pre-trained backbone from the first epoch, its large, noisy gradients can disrupt the learned representations in the early layers before the head has had a chance to develop useful outputs. By first allowing the head to stabilize with a frozen backbone, then releasing all weights together at a lower learning rate, the adaptation proceeds more smoothly and the backbone representations are modified more carefully.

The Grad-CAM integration deserves particular discussion because it touches on a question that goes beyond technical performance: the question of trust. There is a reasonable body of opinion among clinicians that an AI system that simply produces a label and a confidence score is not clinically trustworthy, because there is no way to verify that the model is attending to the right image regions for the right reasons. A system trained on a biased dataset might achieve high accuracy by attending to irrelevant features — imaging artefacts, patient positioning patterns, scanner identifiers — rather than genuine pathological findings. Grad-CAM does not fully solve this problem, but it does make the model's spatial attention visible, which gives the radiologist something concrete to evaluate. In our report evaluations, reviewers consistently described the attention maps as useful precisely because they allowed the clinician to either confirm that the model was attending to the right region or identify cases where the attention was implausible and the model's output should be discounted accordingly [23][25].

The efficiency improvements — more than 93% reduction in per-study analysis time — are significant in practical terms, but they also require some qualification. The 45 to 60 second figure assumes adequate server-side compute provisioning and typical network conditions; in settings with limited bandwidth or constrained cloud compute budgets, latency would increase. The comparison against 15 to 20 minutes of manual review also conflates analysis time with reporting time, whereas in practice a radiologist's workflow includes dictation, comparison with prior studies, and clinical correlation that our automated system does not yet replicate fully. The system is designed as a decision-support tool rather than an autonomous diagnostic engine, and that distinction matters when interpreting the efficiency figures.

The data governance dimension is worth addressing honestly. Cloud-based processing of patient imaging data creates legitimate compliance obligations under HIPAA in the United States, GDPR in the European Union, and equivalent frameworks elsewhere. The Women Wellness architecture is designed to support on-premises deployment — running the Flask microservices within a hospital's own compute infrastructure — for organisations with strict data locality requirements. This does not eliminate the governance challenge, but it makes the system compatible with the range of institutional policies that real hospital IT departments must enforce [17].

IX. CONCLUSION AND FUTURE WORK

This paper has presented Women Wellness — a complete, deployable platform for AI-assisted early breast cancer detection from MRI video sequences. The system brings together client-side frame extraction and preprocessing via OpenCV.js, cloud-hosted CNN classification with Grad-CAM explainability, confidence-weighted multi-frame aggregation, and automated PDF report generation into a single pipeline that completes in under a minute per study. Across a curated MRI evaluation dataset, it achieves classification accuracy between 92% and 95%, with malignant sensitivity of 96.3% and AUC values consistently above 0.96 — performance figures that are competitive with the relevant published literature and that meet the thresholds generally considered adequate for clinical decision-support applications.

The comparative analysis in Section VI places these results in context. Women Wellness is the only system in the comparison — among both published research prototypes and commercially deployed CAD tools — that simultaneously addresses MRI video input, automated clinical reporting, Grad-CAM transparency, parallel scalable processing, full

physician workflow integration, and an open-source technology stack. Each of these is individually achievable; combining all of them within a single, coherent, deployable platform is the specific contribution this work makes.

The case for building this kind of integrated system rests on a simple observation: the barriers to clinical adoption of AI in radiology are not primarily technical. Researchers have demonstrated repeatedly that deep learning models can classify breast images with high accuracy. The barriers are practical — the absence of workflow integration, the lack of transparent explainability, the inability to generate documentation that meets clinical standards, the latency that makes real-time use impractical. Women Wellness is designed around those barriers rather than around the classification problem alone.

Several directions remain open for future development. Lesion segmentation and spatial localization would allow the system to provide precise anatomical descriptions of detected abnormalities, moving from classification-level outputs toward the kind of structured radiology reporting that complex cases require. Support for mammography and ultrasound would expand clinical applicability and enable multi-modal fusion approaches that have shown promise in the literature [3][14]. A longitudinal tracking module would allow the platform to compare findings across multiple examinations of the same patient over time, supporting both risk stratification and treatment response monitoring. Multi-centre prospective validation studies are the essential next step for building the clinical evidence base that regulatory submission and broad clinical adoption will require. We hope to report on these directions in future work, and we welcome collaboration with radiologists and clinical research groups who share an interest in making AI-assisted breast imaging a practical reality rather than a benchmark result.

Acknowledgment

The authors thank the Department of Information Technology at AISSMS Institute of Information Technology, Pune, India, for providing the academic environment and computational resources that made this work possible. We also gratefully acknowledge the open-source communities behind OpenCV, TensorFlow/Keras, Flask, Node.js, MongoDB, React.js, and jsPDF, whose tools form the practical foundation of the platform described here. Special thanks are due to the radiologists who gave their time to evaluate the system's reports and provide the clinical feedback that shaped the final design of the report template.

REFERENCES

1. T. Khater, A. Hussain, R. Bendardaf, I. M. Talaat, H. Tawfik, S. Ansari, and S. Mahmoud, "An explainable artificial intelligence model for the classification of breast cancer," IEEE Access, vol. 11, pp. 5618–5633, 2023.
2. N. Nayak, D. Kumar, and A. Malhotra, "A CNN-based approach for early detection of breast cancer using infrared imaging," in Proc. Int. Conf. Intell. Syst. Adv. Appl. (ICISAA), Pune, India, Oct. 2024, pp. 1–4.
3. Y. Yang, Y. Zhong, J. Li, J. Feng, C. Gong, Y. Yu, and Y. Hu, "Deep learning combining mammography and ultrasound images to predict the malignancy of BI-RADS US 4A lesions in women with dense breasts," Eur. Radiol., vol. 33, no. 11, pp. 8406–8414, 2023.
4. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778.
5. 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, no. 1, pp. 1–12, 2019.
6. W. Cao, H. Chen, Y. Yu, N. Li, and W. Chen, "Changing profiles of cancer burden worldwide and in China: A secondary analysis of the global cancer statistics 2020," Chin. Med. J., vol. 134, no. 7, pp. 783–791, 2021.
7. D. Sheth and M. L. Giger, "Artificial intelligence in the interpretation of breast cancer on MRI," J. Magn. Reson. Imaging, vol. 51, no. 5, pp. 1310–1324, 2020.

8. S. G. Kandlikar, S. Perez-Raya, P. A. Raghupathi, J.-L. Gonzalez-Hernandez, D. Phelan, O. Bhide, A. Thiagarajan, and T. G. Tipton, "Infrared imaging technology for breast cancer detection — Current status, protocols and new directions," Int. J. Heat Mass Transf., vol. 108, pp. 2303–2320, 2017.
9. G. Bradski, "The OpenCV Library," Dr. Dobb's Journal of Software Tools, 2000.
10. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 618–626.
11. H. D. Nelson, B. K. S. Pappas, A. Cantor, J. Griffin, M. Daeges, and L. Humphrey, "Harms of breast cancer screening: Systematic review to update the 2009 U.S. Preventive Services Task Force recommendation," Ann. Intern. Med., vol. 164, no. 4, pp. 256–267, 2016.
12. L. J. Warren, D. L. Mackinnon, C. A. Doyle, and T. P. Williamson, "Variability in radiologist detection and grading of DCIS: A review," Clin. Radiol., vol. 67, no. 11, pp. 1083–1093, 2012.
13. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
14. A. S. Becker, M. Marcon, S. Ghafoor, M. C. Wurnig, T. Frauenfelder, and A. Boss, "Deep learning in mammography: Diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer," Invest. Radiol., vol. 52, no. 7, pp. 434–440, 2017.
15. A. Petrillo, M. Fusco, M. Di Bonito, M. Sansone, and C. Sansone, "Multi-parametric breast MRI for the differential diagnosis of non-mass enhancement lesions: A computer-aided diagnosis approach," Eur. Radiol., vol. 28, no. 3, pp. 1099–1108, 2018.
16. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, pp. 60–88, 2017.
17. K. Löhr and T. Aumüller, "Regulatory frameworks for AI in medical devices: HIPAA, GDPR, and FDA considerations," J. Med. Syst., vol. 45, no. 11, p. 102, 2021.
18. T. Braman, P. Prasanna, J. Whitney, S. Singh, and A. Madabhushi, "Association of peritumoral radiomics with tumor biology and pathologic response to preoperative targeted therapy for HER2 (ERBB2)-positive breast cancer," JAMA Netw. Open, vol. 2, no. 4, p. e192561, 2019.
19. N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: Improved N3 bias correction," IEEE Trans. Med. Imaging, vol. 29, no. 6, pp. 1310–1320, 2010.
20. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, 2015, pp. 1–9.
21. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Representations (ICLR), San Diego, CA, USA, 2015.
22. F. Cabitza, R. Rasoini, and G. F. Gensini, "Unintended consequences of machine learning in medicine," JAMA, vol. 318, no. 6, pp. 517–518, 2017.
23. A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 9, no. 4, p. e1312, 2019.
24. T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, 2006.
25. K. J. Geras, S. Wolfson, S. G. Kim, L. Moy, and K. Cho, "High-resolution breast cancer screening with multi-view deep convolutional neural networks," arXiv preprint arXiv:1703.07047, 2017.