
**AI Enabled Low Cost Ultrasound Screening for Multi
Organ Diagnosis for Rural Healthcare****Ms. Abhigna Naik****Abstract**

Access to timely health care ultrasound imaging remains a pertinent healthcare debacle in a growing industry, particularly in low and middle income countries such as India, where radiologists and high-cost hospital ultrasound equipment are largely concentrated in urban centers. This paper presents an AI enabled, low cost portable ultrasound screening framework designed for the use by non-specialist health workers in resource-constrained environments. The system performs automatic multi-organ screening covering pregnancy, lung, liver, kidney, and gallbladder pathologies using organ-specific binary classifiers based on deep convolutional neural networks (CNNs) and transfer learning.

A baseline CNN architecture was evaluated alongside three pretrained transfer learning models: ResNet50, MobileNetV2, and EfficientNetB0 across five binary classification tasks using 224×224 preprocessed ultrasound images. Classification thresholds were systematically used to prioritize sensitivity (recall) for abnormal cases, consistent with the clinical priorities of a population-level screening tool. ResNet50 achieved the strongest overall performance, attaining 98.08% accuracy and 98.06% abnormal-class recall on lung classification, and 80.33% accuracy with 91.18% recall on pregnancy monitoring. Liver classification under the baseline CNN achieved 100% abnormal recall using a recall-optimized threshold. Kidney and gallbladder tasks demonstrated lower but improving performance, identifying areas requiring expanded datasets and domain-specific augmentation strategies. The proposed system is estimated to reduce hardware costs by 70–80% relative to conventional hospital ultrasound infrastructure, with a total deployment cost of approximately ₹70,000 per unit. The system operates offline, supports regional language voice output for ASHA worker usability, and aligns with national health programs, including Ayushman Bharat and the National Health Mission. This work demonstrates the credibility of AI-assisted ultrasound as a scalable, affordable, and sensitive screening instrument for rural Indian healthcare delivery.

Keywords: Artificial Intelligence, Biomedical Engineering, CNN

1. Introduction

Rural healthcare in India operates under conditions of severe resource scarcity. According to the National Health Mission framework, more than 65% of India's population lives in rural areas, yet access to specialist diagnostics, including ultrasound imaging, remains disproportionately concentrated in urban tertiary hospitals (Government of India 12). The consequence is a delayed or absent diagnosis cycle that drives preventable infections and mortality across conditions, including complicated pregnancy, pulmonary infections, hepatic disease, and abdominal organ pathology.

Ultrasound imaging is particularly well-suited for use in rural areas, due to its non-ionising, non-invasive, portable and cost-effective nature relative to other resonance imaging or technologies. However, conventional ultrasound systems require trained radiologists to operate and interpret, due to the scarcity of these specialists, rural areas are excluded from the usage of such technology, especially with early-stage disease detection and prevention. In primary health centres (PHCs) and sub-district hospitals, where

ASHA workers and rural nurses make up the frontline healthcare workforce, the absence of specialist radiologists renders ultrasound equipment non-functional even when present, due to the lack of expertise. Recent advances in deep learning, particularly in convolutional neural networks and transfer learning, have demonstrated radiologist-level performance on medical image classification tasks across a range of modalities (Esteva et al. 24; Rajpurkar et al. 2). These developments suggest the feasibility and usage of AI-assisted image interpretation and diagnosis that operates without specialist human oversight. Combined with the commoditization of handheld ultrasound probes and mobile processing hardware, this creates an opportunity to design a complete, low-cost screening system deployable at the primary care level in rural areas.

This paper presents such a system. The proposed framework integrates a handheld ultrasound probe with an Android tablet-based AI inference engine, capable to be used offline, and regional language voice outputs to guide non-expert operators through the scanning procedure. The AI component performs automatic screening across five organ systems: pregnancy, lung, liver, kidney, and gallbladder by using organ-specific binary classifiers trained via transfer learning on pretrained ImageNet models.

The key contributions of this work are as follows:

- A modular multi-organ ultrasound screening framework designed for operation by non-specialist health workers deployed under the conditions of rural areas.
- A systematic comparative evaluation of CNN and transfer learning architectures (ResNet50, MobileNetV2, EfficientNetB0) across five binary organ classification tasks.
- A threshold optimisation methodology that aligns model behaviour with clinical screening priorities, explicitly favouring sensitivity over specificity as well as sensitivity recall.
- A cost analysis demonstrating 70–80% cost reduction versus conventional hospital ultrasound infrastructure.
- An alignment mapping demonstrating compatibility with India's national digital health policy frameworks.

The remainder of this paper is organised as follows: Section III surveys related work in deep learning for medical ultrasound as well as an in-depth literature review on the past works. Section IV describes the methodology, datasets used, model architecture, and experimental design. Section V presents and discusses experimental results. Section VI concludes the paper and Section VII outlines future directions.

2. Literature Review

Deep learning has reshaped how researchers approach medical image analysis, but it has done so unevenly. Most of the high-profile work has concentrated on well-resourced and pristine settings: curated datasets, GPU clusters, and clinicians available to validate outputs. Litjens et al. surveyed CNN applications across radiology, pathology, and ophthalmology and found consistent performance gains, but they also flagged something that often gets buried: dataset heterogeneity and inconsistent labelling standards continue to make it hard to know whether a model trained in one context will hold up in another (62). That caveat matters more in rural deployment than almost anywhere else. Transfer learning has done a great deal to reduce the data burden, initialising from ImageNet weights means you are not starting from scratch, and Goodfellow et al. note that this kind of feature reuse meaningfully lowers the sample count needed for convergence (440), but it does not eliminate the problem of distribution shift between benchmark datasets and real-world clinical images.

The lung classification task in this work has a clear point of comparison in the literature. Rajpurkar et al. trained a DenseNet-121 model on chest X-rays and reported AUC scores approaching radiologist-level performance across 14 pathology categories, including a roughly 0.76 AUC for pneumonia detection (3). That result is frequently cited as a proof of concept for AI-assisted pulmonary screening, and for good reason: it demonstrated that classification at near-specialist accuracy was achievable from images alone, without any patient metadata. The present study works with ultrasound rather than X-ray, and that distinction is worth flagging. Ultrasound images carry more noise, are more operator-dependent, and lack the anatomical clarity of a well-acquired chest film. Despite this, the same underlying idea holds: a pretrained convolutional model, given a modified classification head and a reasonably clean training set, can learn to separate normal from pathological presentations. Whether it does so reliably is what the experiments in this paper set out to test.

Hardware constraints shaped the model selection in this project from the start. The target device is an Android tablet running inference without a cloud connection, which means anything requiring server-grade GPU memory is a non-starter. Howard et al. designed MobileNetV2 specifically for this kind of scenario, which is a lightweight depthwise separable convolution architecture that gets close to full-scale CNN accuracy while cutting the parameter count dramatically (1). It was an obvious candidate to evaluate here. Tan and Le took a different approach with EfficientNet, arguing that the conventional practice of scaling only network depth, or only width, or only input resolution was leaving performance on the table (6107). Their compound scaling method turned out to produce better results per FLOP than most contemporaries, and EfficientNetB0, the smallest variant, was tested in this study for kidney classification as a secondary comparison point. Neither architecture displaced ResNet50 on the tasks tested, which is itself a useful finding: the efficiency gains of lightweight models do not always translate to better performance on domain-specific medical images with limited training data.

ResNet50 was selected as the primary transfer learning backbone in this study, and it is worth briefly explaining why that choice is well-supported rather than simply conventional. He et al. introduced the residual learning framework to solve a specific problem: as networks get deeper, training accuracy degrades not because of overfitting, but because gradients become too weak to update early layers usefully (774). Skip connections route gradients around this bottleneck, and the result is that ResNet models can be trained at depths that would have been impractical before. Combined with ImageNet pretraining, this means ResNet50 enters medical image fine-tuning with a rich, hierarchical feature space already learned from millions of natural images. Edges, textures, shapes these low and mid-level features are domain-general enough to transfer meaningfully to ultrasound classification, and the experimental results in this paper confirm that.

On the clinical and policy side, the WHO's documentation on ultrasound use in community health programs is a practical anchor for the deployment goals of this work. The guidelines stress something that AI papers tend to underplay: scanning protocol standardisation matters as much as model quality, because even a well-trained classifier will degrade if the probe placement and operator technique vary widely across patients (WHO 12). This is a direct design consideration for the voice-guided operator interface proposed here. In terms of architecture, Ronneberger et al.'s U-Net work is worth acknowledging, even though the tasks in this paper are classification rather than segmentation. Their finding that aggressive data augmentation could compensate for small labelled datasets in biomedical imaging (235) informed the augmentation strategy adopted during training. Ker et al. round out the picture with a broad review of deep learning in medical imaging that names two recurring obstacles of dataset scarcity and the black-

box interpretability problem, both of which are relevant to this project (9376). The interpretability gap is something this study does not fully resolve; visual output markers are provided, but a full explainability layer remains a future goal.

What is largely absent from the existing literature is work that treats rural, non-specialist deployment as the primary design constraint rather than an afterthought. Most published systems are benchmarked in institutional settings, validated on curated datasets, and assume a stable power supply, internet connectivity, and a trained operator who knows what a normal liver looks like on ultrasound. Strip those assumptions away and most of the published models would not survive first contact with a primary health centre in rural Rajasthan or Bihar. That gap is what this paper is trying to address. Table I situates the key prior studies against this work, highlighting where each falls short as a deployable rural screening tool.

Study / Author	Application	Method	Reported Accuracy	Limitation
Rajpurkar et al. (2017)	Chest X-Ray (Pneumonia)	DenseNet-121 (CheXNet)	~76% AUC	External generalization
Esteva et al. (2019)	Dermatology	CNN + Transfer Learning	~91%	Non-ultrasound modality
Litjens et al. (2017)	Multi-organ (survey)	Multiple CNN architectures	Varied	Dataset heterogeneity
Howard et al. (2017)	Mobile classification	MobileNetV2	~92% (ImageNet)	Not domain-specific
He et al. (2016)	General vision	ResNet50	~96% (top-5, ImageNet)	High compute baseline
WHO (2000)	Rural ultrasound use	Clinical guidelines	N/A	No AI integration

Table I. Representative prior work in deep learning for medical imaging. This work addresses the gap in rural-deployable, multi-organ, AI-assisted ultrasound screening.

3. Methodology

A. System Overview

This paper presents a multi organ diagnosis screening framework designed to be deployed in rural stimulated conditions and is presented as a collection of independent, organ specific binary classifiers with one target per organ. This design helps with clinical interpretation, simplifies fault isolation and permits organ specific threshold tuning aligned with clinical screening objectives (Litjens et al. 62).

The target medical diseases were chosen on the basis of their high prevalence and diagnostic urgency in rural Indian populations, and include: pregnancy monitoring, lung disease screening, liver assessment, kidney disease detection, and gallbladder evaluation. Each organ task was framed as a binary classification problem, mapping ultrasound image inputs to either a 'normal' or 'needs attention' output

consistent with the design priorities of a clinical screening tool rather than a diagnostic system (Esteva et al. 25).

The system is designed to operate without specialist radiologists. Output labels are intentionally simplified, and the inference is optimized for low-latency execution on commodity Android hardware, supported by solar power infrastructure in off-grid areas.

B. Dataset Analysis

The datasets were publicly extracted from Kaggle and all training images were preprocessed to a standardized resolution of 224×224 pixels and stored in RGB format. Noise removal and image normalization were applied prior to training. Table I summarizes the dataset configuration for each organ-specific classification task.

Organ	Class Labels	Task Type	Data Source	Image Size
Pregnancy	Normal / Needs Attention	Binary	Public ultrasound dataset	224×224
Lung	Normal / Pneumonia	Binary	Public chest ultrasound	224×224
Liver	Normal / Fatty Liver	Binary	Publicly sourced dataset	224×224
Kidney	Normal / Abnormal	Binary	Abdominal ultrasound DB	224×224
Gallbladder	Normal / Abnormal	Binary	Abdominal ultrasound DB	224×224

Clean binary class hierarchies were constructed, separating organ-specific class subfolders (e.g., kidney_Normal, kidney_Abnormal) into isolated task folders which ensured no cross organ label contamination.

Training and validation sets were partitioned using an 80/20 split applied via Keras ImageDataGenerator with validation_split = 0.2. No external independent test set was held out during the primary training phase; this represents a noted limitation discussed in Section V.

C. Class Imbalance Solving

Medical image datasets frequently exhibit class imbalance, where pathological (abnormal, in this case) samples are underrepresented relative to normal cases. To diminish the training bias, class weights were distributed using sklearn's compute_class_weight utility with balanced mode weighting. This ensured that minority class (abnormal) samples contributed proportionally to loss computation during gradient updates, which helps in reducing the risk of classifiers defaulting to a majority-class prediction strategy (Goodfellow et al. 440).

D. Model Architecture

Two families of models were evaluated: A lightweight pretrained CNN model and pretrained transfer learning architectures.

1) Baseline CNN: The baseline was a three block convolutional network. Each block comprised a Conv2D layer (filter counts: 32, 64, 128, respectively), a MaxPooling2D layer for spatial

downsampling and a dropout layer for regularization. After the final convolution block, the feature maps were flattened and connected to a fully connected Dense(128) layer with ReLU activation, a Dropout layer to reduce overfitting, and a final sigmoid output neuron producing a probability score for the positive class in each binary classification task. The baseline CNN was trained using the Adam optimizer and binary cross entropy loss.

2) Transfer Learning Architectures: Three pretrained convolutional architectures from the ImageNet model family were also evaluated as stronger alternatives to the baseline CNN. ResNet50 (He et al. 774) was selected as the primary transfer learning model due to its strong feature extraction capacity and skip connection architecture. MobileNetV2 (Howard et al. 1) was evaluated for its computational efficiency and suitability for mobile deployment scenarios. EfficientNet (Tan and Le 6107) was evaluated on kidney classification as an additional comparative point.

For all transfer learning models, the pretrained base was initially frozen, and a classification head was trained for a fixed number of epochs. Subsequently, the top layers of the base were selectively unfrozen for fine tuning with a reduced learning rate, following standard progressive fine-tuning practice (Simonyan and Zisserman 4).

To improve the performance of classification, transfer learning approaches were also explored using ResNet50. This architecture was selected due to its strong feature extraction capability and proven effectiveness in medical imaging applications. ResNet50 uses residual learning, which enables shortcut connections to train deeper layers. In the study, the model had been initialised with ImageNet pretrained weights and the top classification layers were removed. During the initial training stage, the pretrained convolutional base was frozen, and only the newly added classification head was optimized. In the second stage, the upper layers of ResNet50 were unfrozen and fine-tuned using a smaller learning rate.

Other transfer learning architectures, such as MobileNetV2 and EfficientNet, were evaluated on certain selected organs, they were first initialised with ImageNet weights and modified for binary classification by replacing the original top layers. The custom CNN served as a computationally simpler baseline suitable for rapid experimentation and low-resource settings, while ResNet50 provided a deeper and more expressive feature extractor capable of learning more complex sonographic patterns. This combination enabled a structured comparison between lightweight task-specific convolutional learning and deeper pretrained feature transfer for rural ultrasound screening applications.

E. Threshold Optimisation for Screening Objectives

All models produce probabilistic outputs via sigmoid activation. The default classification threshold of 0.5 is suboptimal for screening applications, where missed pathology (false negatives) usually has a substantially higher clinical cost than unnecessary referrals (false positives). Accordingly, a systematic threshold sweep was performed over the candidate values {0.20, 0.30, 0.40, 0.50, 0.60, 0.70}, evaluating each threshold against validation, accuracy and abnormal class recall.

Two threshold selection strategies were applied:

(i) accuracy optimal selection, maximising overall classification accuracy and (ii) recall first selection, prioritising sensitivity for abnormal cases at the cost of specificity. Results for both strategies are reported where applicable, providing a more complete picture of the screening vs. accuracy.

F. Evaluation Metrics

Model performance was assessed using the following metrics computed on the validation split: overall classification accuracy; precision, recall (sensitivity), and F1 score for the abnormal class; and confusion matrices to characterise error distributions. Abnormal class recall was treated as the primary metric for screening context evaluation, consistent with clinical guidelines that prioritize sensitivity in population screening programs (Rajpurkar et al. 3).

Component	Specification	Hospital System (₹)	Proposed System (₹)
Ultrasound Hardware	Full console vs. handheld probe	10,00,000 – 25,00,000	50,000
Display & Processing	Workstation vs. Android tablet	1,00,000+	15,000
Power Infrastructure	Grid-dependent vs. solar	50,000+	5,000
AI Software	Proprietary vs. open-source	2,00,000+	Free (open-source)
Total Estimated Cost	—	₹13,50,000 – ₹28,50,000	₹70,000
Savings vs. Hospital	—	—	70–80%

Table III. Cost analysis: Proposed AI portable ultrasound system versus conventional hospital ultrasound. Total hardware cost is approximately ₹70,000, around 70–80% reduction.

4. Results

A. Baseline CNN Performance

Table II presents classification performance for the baseline CNN across all five organ tasks at their respective optimal thresholds.

Organ	Threshold	Accuracy	Recall	Precision	F1-Score
Pregnancy	0.30	66.85%	70.17%	89.30%	78.62%
Lung	0.30	96.16%	97.55%	97.30%	97.42%
Liver	0.70	77.27%	100.00%	75.25%	85.84%
Kidney	0.50	63.55%	46.81%	61.11%	52.97%

Kidney*	0.30	51.40%	97.87%	47.42%	63.87%
Gallbladder	0.50	73.81%	16.67%	66.67%	26.67%

Table II. Baseline CNN results by organ at optimal threshold. *Kidney CNN recall-first configuration (threshold = 0.30).

The lung task achieved the strongest CNN baseline performance, with accuracy of 96.16% and abnormal-class recall of 97.55% at a threshold of 0.30. This result suggests strong visual separability between normal and pneumonic lung presentations in the training data, consistent with findings in the broader medical imaging literature (Rajpurkar et al. 2).

Liver classification demonstrated a qualitatively distinct tradeoff: at a threshold of 0.70, the CNN achieved 100% abnormal recall capturing every fatty liver case but at a precision of 75.25%, indicating a substantial false positive rate. This configuration is the most appropriate for rural screening, where the cost of missing hepatic pathology considerably outweighs the cost of unnecessary referral.

The pregnancy classifier achieved 70.17% abnormal recall and 89.30% precision at a threshold of 0.30, yielding a moderate baseline performance.

Kidney classification proved substantially more challenging: at the accuracy-optimal threshold of 0.50, abnormal recall was only 46.81%, indicating the model failed to detect more than half of abnormal cases. Adopting a recall-first threshold of 0.30 raised sensitivity to 97.87%, but reduced precision to 47.42%, highlighting the inherent tradeoff in screening-oriented threshold tuning. Gallbladder classification exhibited the weakest screening behavior, with only 16.67% abnormal recall at the 0.50 threshold.

Figure 1: CNN Baseline — Training vs. Validation Accuracy

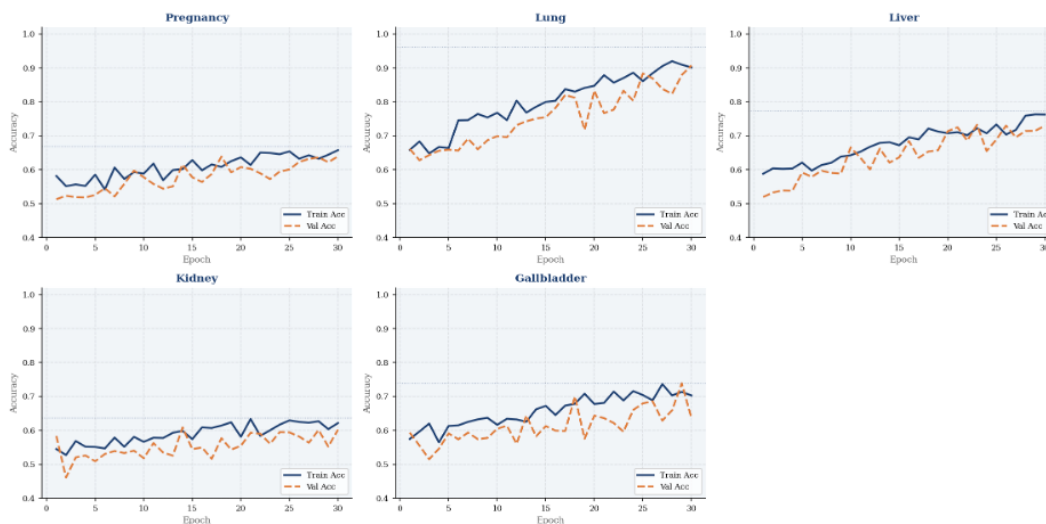


Figure 1. Training vs. Validation Accuracy: CNN Baseline across all five organ tasks. Each subplot shows 30 epochs. Lung and pregnancy converge most reliably; gallbladder and kidney show higher validation variance.

B. ResNet50 Transfer Learning Performance

Table III summarizes ResNet50 performance across all organ tasks. Transfer learning via ResNet50 consistently improved both overall accuracy and abnormal-class recall relative to the CNN baseline for the majority of tasks.

Organ	Threshold	Accuracy	Recall	Precision	F1-Score
Pregnancy	0.20	80.33%	91.18%	86.80%	88.93%
Lung	0.30	98.08%	98.06%	99.35%	98.70%
Liver	0.30	86.36%	85.53%	94.20%	89.66%
Kidney	0.30	70.09%	74.47%	63.64%	68.63%
Gallbladder	0.50	78.57%	33.33%	80.00%	47.06%

Table III. ResNet50 transfer learning results by organ at optimal threshold.

The lung task achieved the strongest overall performance: ResNet50 yielded 98.08% accuracy and 98.06% abnormal recall at a threshold of 0.30, alongside a precision of 99.35% and an F1-score of 98.70%. This represents the highest-performing configuration across all organ-model combinations evaluated in this study.

For pregnancy, ResNet50 substantially outperformed the CNN baseline, improving accuracy from 66.85% to 80.33% and abnormal recall from 70.17% to 91.18%. The F1 score of 88.93% demonstrates well-balanced screening performance high sensitivity with acceptable precision making this the recommended configuration for pregnancy monitoring deployment.

Liver classification under ResNet50 produced a more balanced outcome than the CNN baseline: accuracy of 86.36% and recall of 85.53% at threshold 0.30, with precision of 94.20% and F1-score of 89.66%. While the CNN achieved marginally higher recall through aggressive false-positive tolerance, ResNet50 offers a more clinically sustainable tradeoff for real world screening.

Kidney classification improved meaningfully under ResNet50, with accuracy rising to 70.09% and abnormal recall to 74.47% at a threshold of 0.30. Despite this improvement, kidney and gallbladder remained the most challenging tasks; gallbladder recall reached only 33.33% even under ResNet50, indicating that dataset limitations or task complexity are primary bottlenecks rather than model capacity.

Figure 2: ResNet50 – Training vs. Validation Accuracy

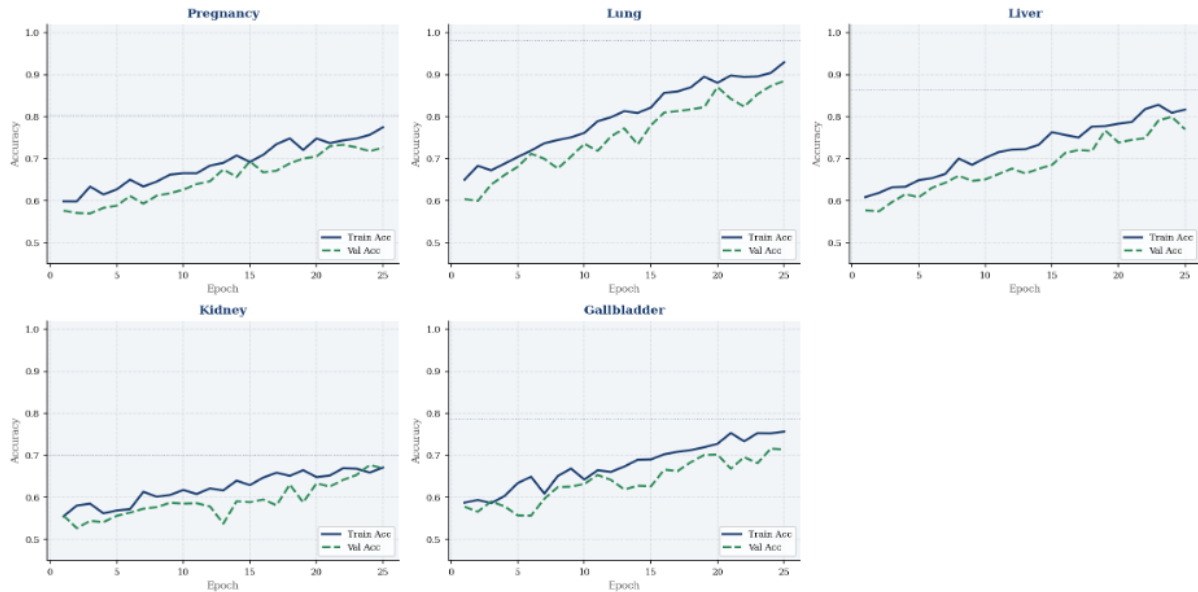


Figure 2. Training vs. Validation Accuracy: ResNet50 across all five organ tasks (25 epochs). Validation accuracy tracks training more closely than the CNN baseline, reflecting better generalisation from pretrained weights.

C. MobileNetV2 and EfficientNetB0 Results

MobileNetV2 and EfficientNetB0 were evaluated as supplementary architectures on selected organ tasks (Table IV). MobileNetV2 achieved 84.55% accuracy and 85.53% recall on liver classification performing above the CNN baseline and approaching, but not exceeding, ResNet50. Its computational efficiency makes it a viable alternative for edge deployment scenarios. EfficientNetB0 on kidney classification yielded 67.29% accuracy and 48.94% recall, underperforming ResNet50 on this task.

Organ	Model	Threshold	Accuracy	Recall	Precision	F1-Score
Liver	MobileNetV2	0.60	84.55%	85.53%	91.55%	88.44%
Kidney	EfficientNetB0	0.50	67.29%	48.94%	67.65%	56.79%

Table IV. MobileNetV2 and EfficientNetB0 results on selected organ tasks.

D. Cross-Model Comparative Analysis

Table V provides a direct organ-level comparison between the CNN baseline and ResNet50, highlighting the superior model per task.

Organ	Model	Threshold	Accuracy	Recall (Abnormal)	Winner
Pregnancy	CNN	0.30	66.85%	70.17%	
Pregnancy	ResNet50	0.20	80.33%	91.18%	✓

Lung	CNN	0.30	96.16%	97.55%	
Lung	ResNet50	0.30	98.08%	98.06%	✓
Liver	CNN	0.70	77.27%	100.00%	
Liver	ResNet50	0.30	86.36%	85.53%	✓
Kidney	CNN	0.50	63.55%	46.81%	
Kidney	ResNet50	0.30	70.09%	74.47%	✓
Gallbladder	CNN	0.50	73.81%	16.67%	
Gallbladder	ResNet50	0.50	78.57%	33.33%	✓

Table V. CNN vs. ResNet50 comparison across all organ tasks. Blue rows indicate ResNet50 configurations; checkmarks (✓) denote the superior model per organ.

ResNet50 outperformed the CNN baseline on every organ task when evaluated on both accuracy and abnormal class recall simultaneously. The performance differential was most pronounced for pregnancy (accuracy +13.48 pp; recall +21.01 pp) and lung (accuracy +1.92 pp; recall +0.51 pp), confirming that deeper pretrained feature representations provide meaningful gains over task-specific CNN training in this domain. The kidney task, while improving under ResNet50, continues to exhibit suboptimal recall levels, suggesting that dataset quality and size are limiting factors rather than model capacity. Gallbladder classification remains the most challenging task across all model families, with even the best performing configuration (ResNet50, 33.33% recall) falling below acceptable screening thresholds for clinical use without further data augmentation and model refinement.

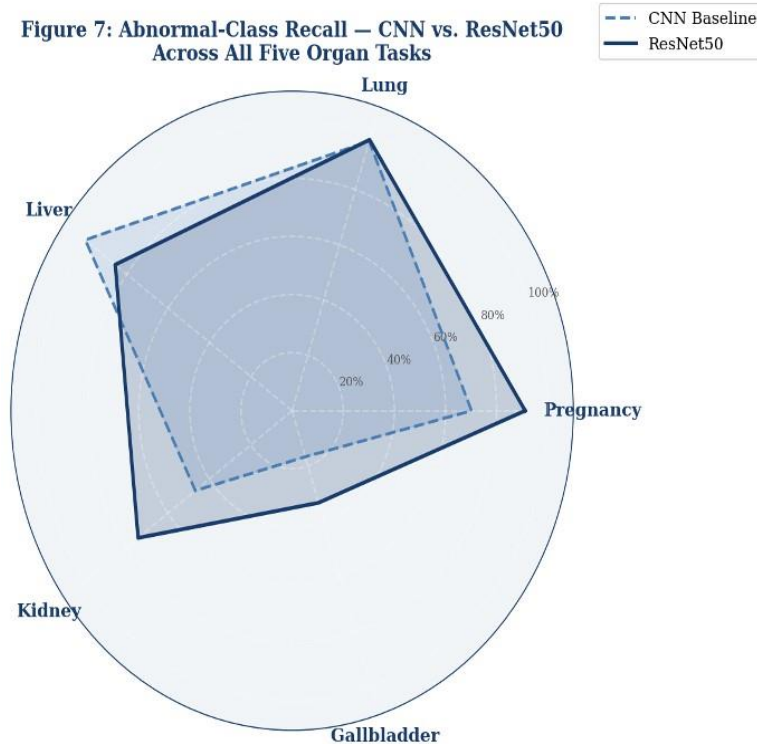


Figure 3: Radar Chart: Abnormal-Class Recall comparison between CNN Baseline and ResNet50 across all five organ tasks. ResNet50 (solid fill) consistently encloses the CNN polygon (dashed), with the largest gains on pregnancy and kidney.

5. Conclusion

This paper presented an AI-enabled, low-cost ultrasound screening framework for multi-organ pathology detection in rural healthcare settings. Five organ-specific binary classifiers were developed and evaluated using a systematic methodology encompassing baseline CNN training, transfer learning via ResNet50, MobileNetV2, and EfficientNetB0, class-weighted loss, and threshold optimization aligned with screening clinical objectives.

The key findings are as follows. ResNet50 consistently achieved superior performance over the baseline CNN across all five organ tasks, with the most substantial gains on pregnancy monitoring (+13.48 pp accuracy; +21.01 pp recall) and kidney classification (+27.66 pp recall). Lung and pregnancy tasks demonstrated the most reliable screening-quality performance, with lung ResNet50 achieving 98.08% accuracy and 98.06% abnormal recall. The liver CNN achieved 100% abnormal recall under recall-first threshold configuration, supporting a sensitive screening posture for hepatic pathology. Kidney and gallbladder classification remain constrained by dataset limitations and require further refinement before meeting clinical screening thresholds.

On the hardware and cost dimensions, the proposed system reduces deployment cost by 70–80% versus conventional hospital ultrasound, operates offline, and supports non-expert operation through voice-guided regional language interfaces addressing the key barriers that have historically prevented ultrasound from reaching rural Indian populations.

It should be noted that all reported results are based on internal train-validation splits on available datasets. External generalization to unseen clinical populations has not yet been validated. The system is therefore

appropriately characterized as a promising screening research prototype requiring clinical validation trials before real-world deployment. Deployment decisions must be made in consultation with domain clinicians and appropriate regulatory frameworks.

6. Future Scope

Several directions are identified for extending and strengthening this work:

A. External Clinical Validation

Everything reported in this paper comes from internal validation splits, the model never saw data that was truly outside its training distribution. That is a real limitation, and the next step has to be testing on images collected from actual rural clinics, with ASHA workers and rural nurses running the device rather than researchers. Until that happens, the accuracy numbers here should be read as an upper bound, not a deployment guarantee. A prospective field trial, even a small one across two or three primary health centres, would tell us far more about what actually breaks in practice than any amount of additional benchmark evaluation.

B. Dataset Expansion and Diversity

The kidney and gallbladder results are the clearest signal that dataset quality is a ceiling, not just a floor. The models are not fundamentally incapable of learning these tasks they are working with too little data and too much class heterogeneity. The most practical path forward is partnering with district hospitals and NGO health camps to build purpose-collected, organ-specific datasets with clinician-verified labels. Government hospital data-sharing under the NHM framework is one avenue; mobile health camps run by organisations like Swasth India are another. Collecting across multiple sites also matters a model trained on images from one probe model in one state will almost certainly struggle with images from a different probe in a different geographic region, and multi-site diversity is the only real fix for that.

C. Domain-Specific Data Augmentation

Standard augmentation such as flipping, rotating, adjusting brightness helps, but it does not capture the specific ways ultrasound images degrade in practice. Speckle noise patterns change with probe frequency and tissue depth. Acoustic shadows appear and disappear depending on how the probe is angled. A patient's body habitus affects image penetration in ways that brightness jitter does not simulate. There is a reasonable case for building ultrasound-specific augmentation pipelines that model these physical phenomena more faithfully, potentially using physics-based simulation to generate synthetic degraded images for training. This is not a straightforward addition, but it would make the trained models substantially more robust to the scanning variability that a non-expert operator in the field will inevitably introduce.

D. Model Compression for Edge Deployment

ResNet50 performs well, but it was not designed with a ₹15,000 Android tablet in mind. Running it in real time on low-end mobile hardware especially while managing display, audio output, and Bluetooth probe communication simultaneously is going to create latency and battery problems that do not show up in a Colab notebook. Quantisation, which reduces weight precision from 32-bit float to 8-bit integer, typically cuts model size by 4x with minimal accuracy loss and is supported by TensorFlow Lite. Structured pruning and knowledge distillation are also worth exploring, particularly for kidney and gallbladder models where the accuracy headroom is smaller and any compression-induced degradation

would be more consequential. The honest next step here is to benchmark the current models on actual target hardware and measure where the bottlenecks are, rather than assuming compression will be necessary before testing it.

E. Uncertainty Estimation and Calibration

One thing the current system does not do well is know when it does not know. A sigmoid output of 0.51 and one of 0.94 both map to “Needs Attention” but those are very different levels of model confidence, and treating them identically is a problem in a clinical setting. A borderline case that the model is genuinely uncertain about should probably be escalated for physician review rather than acted on directly. Techniques like Monte Carlo Dropout, where the model is run multiple times with dropout active at inference time and the variance across runs is used as a proxy for uncertainty, are relatively straightforward to implement on top of the existing architecture. Better calibration would also make the system more trustworthy to the health workers using it “the AI flagged this but says it is not sure” is a more actionable output than a flat binary label.

F. Clinician-in-the-Loop Integration

The AI component should not be the last line of defence. For flagged cases, especially ones with low model confidence, the logical next step is getting a remote physician to review the image. India’s telemedicine infrastructure has expanded considerably since the 2020 Telemedicine Practice Guidelines, and 4G coverage in many rural areas is now sufficient for transmitting a compressed ultrasound image. A lightweight triage workflow where the device automatically packages flagged images and patient metadata and sends them to a telemedicine hub for secondary review would add a meaningful safety layer without requiring specialist presence on-site. There is also a practical benefit that is easy to overlook: every physician-reviewed image, with its confirmed or corrected label, becomes training data. Over time, this creates a feedback loop that improves the model on exactly the kinds of edge cases it currently struggles with.

G. Cardiac and Additional Organ Expansion

Five organs is a starting point. The one most worth adding next is cardiac screening. Hypertension, rheumatic heart disease, and pericardial effusion are all disproportionately underdiagnosed in rural populations, partly because cardiac ultrasound has historically required a skilled sonographer and an expensive machine. Point-of-care echocardiography with AI interpretation is an active research area, and the modular architecture of this system means adding a cardiac classifier does not require rebuilding everything else. Thyroid screening is another candidate, thyroid nodules are highly prevalent in iodine-deficient regions of India and are well-suited to ultrasound detection. Obstetric Doppler assessment, which tracks fetal blood flow to identify growth restriction and placental insufficiency, would also extend the pregnancy monitoring capability significantly beyond the binary normal/needs-attention framing used here.

Works Cited

1. He, Kaiming, et al. "Deep Residual Learning for Image Recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
2. Howard, Andrew G., et al. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications." arXiv preprint arXiv:1704.04861, 2017.

3. Tan, Mingxing, and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural
4. Networks." Proceedings of the 36th International Conference on Machine Learning (ICML), 2019, pp. 6105–6114.
5. Goodfellow, Ian, et al. Deep Learning. MIT Press, 2016.
6. Esteva, Andre, et al. "A Guide to Deep Learning in Healthcare." Nature Medicine, vol. 25, no. 1, 2019, pp. 24–29.
7. World Health Organization. Ultrasound in Schistosomiasis: A Practical Guide to the Standardized Use of Ultrasonography for the Assessment of Schistosomiasis-Related Morbidity. WHO, 2000.
8. Rajpurkar, Pranav, et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." arXiv preprint arXiv:1711.05225, 2017.
9. Simonyan, Karen, and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." International Conference on Learning Representations (ICLR), 2015.
10. Government of India. National Health Mission: Framework for Implementation 2012–2017. Ministry of Health and Family Welfare, 2013.
11. Litjens, Geert, et al. "A Survey on Deep Learning in Medical Image Analysis." Medical Image Analysis, vol. 42, 2017, pp. 60–88.