Journal of Information Technology Management
ISSN-OnLine: 2980-7972
University of Tehran

A Two-Stage Progressive Transfer-Learning Framework for Binary Thyroid Nodule Classification in Ultrasound Imaging

Ali Retha Hasoon Khayeat*

*Corresponding author, Department of Computer Science, College of Computer Science & Information Technology, University of Kerbala, Karbala, Iraq. E-mail: ali.r@uokerbala.edu.iq

Ansam Abd Alrasool Ali AL-Obaidi

College of Pharmacy, University of Kerbala, Karbala, Iraq. E-mail: ansam.a@uokerbala.edu.iq

Muhanad Abdul Elah Alkhalisy

College of Business Informatics, University of Information Technology and Communications (UoITC), Baghdad, Iraq. E-mail: muhanad_alkhalisy@uoitc.edu.iq

Journal of Information Technology Management, 2026, Vol. 18, Special Issue, pp. 137-153.
Published by the University of Tehran, College of Management
doi: https://doi.org/10.22059/jitm.2026.109979
Article Type: Research Paper
© Authors
Received: August 14, 2026
Received in revised form: October 27, 2026
Accepted: November 08, 2026
Published online: December 01, 2026

Abstract

Thyroid nodules are highly common, mostly benign, and evaluated using ultrasound as the standard of care; however, their interpretation is often operator-dependent. This paper proposes an experimental approach based on deep transfer learning for binary thyroid nodule classification using a subset of 3,115 resized grayscale ultrasound images from Kaggle. We propose a two-stage progressive fine-tuning protocol with intermediate output suppression to prevent catastrophic forgetting, along with a preprocessing pipeline that combines bilateral filtering, contrast-limited adaptive histogram equalization (CLAHE), and single-image unsharp masking. Here, we perform a systematic architecture search, comparing six different architectures: our custom CNN trained from scratch; EfficientNetB0, DenseNet121, and ResNet152 (all with ImageNet-pretrained weights); and a CNN-handcrafted-feature-fusion model that is fine-tuned over all the compared backbone architectures and jointly evaluated against average classifier accuracies per species at 6-, 12-, and 20-fold imbalanced ratios. This yields a best-performing architecture (an ImageNet-pretrained ResNet50 with two-stage fine-tuning), achieving 77.61% test accuracy, with test AUC evaluated on a withheld test set of 469 images. Class 0 precision, recall, and F1-scores were 0.83, 0.80, and 0.81, respectively, compared with 0.70, 0.74, and 0.72, respectively, for class 1. Gradient-weighted class activation mapping (Grad-CAM) confirmed the qualitative clinical plausibility of the model, with network attention directed at hyperechoic nodule margins rather than background

artifacts. We further contextualize these findings in relation to the wider literature on thyroid ultrasound image classification using CNNs and discuss how our single-backbone scheme relates to ensemble approaches reported in concurrent literature. We also discuss lessons regarding generalization, gradient shock, and the extent to which preprocessing and feature engineering are necessary for small medical imaging datasets.

Keywords: Thyroid Nodule Classification, Convolutional Neural Network, Transfer Learning, Resnet50, Densenet121, Grad-Cam, Ultrasound Imaging, Computer-Aided Diagnosis

Introduction

Research indicates that thyroid nodules are common, appearing in one out of every two adults. Fortunately, they are rarely a cause for concern, as only 10% are cancerous (Durante et al., 2018). If nodules are common and mostly benign, why have thyroid cancer rates increased across all age, gender, and racial groups over the last 30 years, keeping ultrasound clinics very busy? Ultrasound remains the primary method for risk stratification of thyroid nodules and selection for fine-needle aspiration biopsy. However, ultrasound interpretation is known to be operator-dependent (Floridi et al., 2019), with interobserver variation among readers with different levels of experience.

This subjectivity has led to the development of computer-aided diagnosis (CAD) pipelines, most notably those based on convolutional neural networks (CNNs). Rather than presenting a new architecture, the current work provides an independent experimental case study. Instead of proposing a novel architecture, we conducted systematic experiments to compare commonly used CNN architectures on a single thyroid ultrasound dataset. We consider both the successful designs and the approaches that underperformed, including the specific reasons for their limitations, to be equally important to our analysis. Accordingly, we report these findings in equal detail as part of the ablation study.

The remainder of the paper is structured as follows. In Section 2, we present related work on CNN-based thyroid ultrasound classification. In Section 3, we describe the dataset, preprocessing pipeline, candidate architectures, and training protocols. The results of the architecture search and the specifications of the best-performing model are presented in Section 4. These findings and their comparison with previous work are discussed in Section 5. Section 6 concludes the paper.

Literature Review

Since January 2020, the pool of existing CAD systems for thyroid ultrasound has become sufficiently broad to distinguish general types based on their technical foundations: traditional machine learning approaches, in which features are derived from a segmented region of interest and fed into a conventional supervised classifier; and deep learning methods, in which convolutional neural networks (CNNs) directly process input images and learn their own feature hierarchies during the training process (Hosny et al., 2018). To date, evidence appears to be consistently favorable toward both types of CAD systems when compared with unaided interpretation. Deep learning systems tend to perform similarly to, and potentially approach the diagnostic accuracy of, experienced sonographers and endocrinologists (Chambara & Ying, 2019; Wu et al., 2020).

These analyses lead to three standard ImageNet-pretrained networks: VGG16, ResNet50, and InceptionV3. No single model wins consistently because final performance depends on data size, preprocessing, and tuning choices. For example, Xu et al. (2025) performed a classification task using VGG16, ResNet50, and InceptionV3 to compare the three models quantitatively and found that InceptionV3 was the strongest, narrowly outperforming ResNet50 (AUC = 0.747) and VGG16 (AUC = 0.731). ImageNet-pretrained transfer learning generally results in a higher area under the ROC curve on thyroid ultrasound data than training the same architecture from random initialization (Lee et al., 2023; Shao et al., 2020), which served as inspiration for the transfer-learning procedure adopted in this study.

Another important methodological aspect is determining how many layers should be fine-tuned. In accordance with the gradient-shock failure mechanism discussed in Section 4, previous studies by Dar et al. (2020) and Gupta et al. (2020) show that, for smaller target datasets, freezing early convolutional layers while optimizing later ones appears to outperform both keeping the entire network frozen and fully unfreezing the entire network.

Another key trend in more recent literature is that CNNs and traditional radiomics, when used independently, may have limitations, leading researchers to combine them or recommend their integration. For instance, Xu et al. (2025) used a combination of LASSO-selected radiomics features and PCA-reduced deep features in an SVM classifier (AUC = 0.748), together with an InceptionV3 model (AUC = 0.763), using simple probability averaging to classify risk. The AUC of 0.783 for this hybrid model represents a statistically significant improvement compared with the standalone SVM model (DeLong p = .036). While some studies have fused features at the representation level, for instance, by combining local binary patterns (LBP), which are handcrafted texture descriptors, with deep features extracted by CNNs, we return in Section 5 to compare these results with those of our own hybrid CNN-LBP configuration. Surprisingly, this combined configuration does not outperform the best-performing standalone backbone model, contrasting with the fusion literature more broadly.

Furthermore, numerous large-scale studies have reported that CNN performance can exceed that of practicing radiologists in various scenarios. Xu et al. (2023) found that a thyroid ultrasound AI model trained using data from more than 200 hospitals improved radiologists' diagnostic accuracy when used as a decision aid. Tong et al. (2023) also found that a tailored balance between AI assistance and radiologist interpretation outperformed either approach alone. However, other validation studies offer a more cautious perspective. Using fully independent, multi-vendor data, Buda et al. (2019) found an AUC of 0.69 for a CNN, only slightly higher than that of four radiologists reading the same images (AUC range = 0.63–0.66).

This study differs substantially from much of this literature in scope rather than architecture. We do not propose a new network or attempt to develop a novel hybrid feature-learning approach beyond one hybrid ablation arm. Instead, we perform a controlled architecture search across different configurations for binary classification using shallow and deep transfer learning on a single dataset. The contribution of a classical preprocessing pipeline for this specific use case, together with a Grad-CAM-based qualitative audit of the best-performing model, is also examined. The most relevant literature in this regard concerns feature fusion, which we revisit in Section 5 as a useful point of comparison. Whereas the fusion literature tends to report gains from combining CNN and non-CNN features, our hybrid configuration produced the opposite result.

Materials and Methods

Dataset

The dataset uploaded in Kaggle consisted of a total of 3,115 grayscale thyroid US images which were labeled for binary classification class as well including benign (class 0) or malignant (class 1). Images were split roughly into a 70:15:15 scheme (training, validation, and test splits, respectively), resulting in 2180 training images and a held-out test set totaling 469 images. Throughout the duration, class 1 (malignant) was an underrepresented minority class, in accordance with the expected patterns of prevalence as reported in the literature (Section 2).

Image Preprocessing

But one obstacle in imaging is that raw ultrasound images are corrupted by acoustic noise, low soft tissue contrast, and inherent speckle noise. Accordingly, a standard preprocessing pipeline consisting of three stages was applied to each grayscaled image before the model input:

- Rather than reducing noise by directly adding isotropic Gaussian to the image, a bilateral filter (7, σ_color = 65, σ_space = 65) was applied that focuses more on suppressing high-frequency grains while lessening the decay of sharp anatomical edges.

- Local contrast enhancement: Contrast Limited Adaptive Histogram Equalization (CLAHE), clip limit 2.5 and an 8×8 tile grid, was selected to equalize shadowed regions without amplifying background noise.

- Texture sharpening: An unsharp-masking filter was used to enhance the submillimeter character of tissue boundaries and nodule margins.

The champion architecture is trained without this preprocessing pipeline in (Section 4.3) to separate its contribution.

Architecture Search

These matched data splits were used to train and evaluate six classification configurations.

- Casual Cononical Net (trained from scratch): a very slim convolutional network (with no pre-training) — included as a data-starvation baseline.

- EfficientNetB0: a compound-scaled network pretrained on ImageNet and then fine-tuned on the preprocessed images in which each grayscale channel has been replicated.

- DenseNet121 (fully unfrozen from epoch 1): all 121 layers were trainable in the first epoch, for a high-capacity feature-reuse architecture.

- ResNet152 (progressive fine-tuning): a 152-layer residual network, initially frozen to warm up, then fine-tuned with: page 1131 cosine-decay learning-rate schedule.

Deep learning model: Hybrid CNN + LBP fusion – A 32-dimensional hand-crafted Local Binary Pattern (LBP) texture vector concatenated with ResNet50 deep features before the classification.

- ResNet50 (progressive fine-tuning two-stage): the best configuration detailed fully below.

All architectures either had ImageNet-pretrained weights initialized where applicable, with input images resized to 224×224 pixels and the single grayscale channel tripled up as three channels at the input so that image depth would match what was expected from the pretrained backbones. Data augmentations (random rotation ±0.05 radians, horizontal flip,

random zoom ±5%, constant zero-fill at exposed borders) were performed automatically on the fly during training and switched off for validation/test evaluation.

Champion Model: Two-Stage Progressive ResNet50

Here, the top-performing configuration was a ResNet50 backbone (224×224×1 input, ImageNet-initialized) trained under a two-stage progressive protocol to prevent accidental catastrophic forgetting that can occur when deep pretrained networks are unfrozen too early (Section 4.2):

- Stage 1 (warmup): The complete ResNet50 backbone is frozen. A custom classification head (Dense (512, activation='relu'), BatchNorm, Dropout (rate = 0.4), Softmax (2 units), Adam ($lr=1\times10^{-4}$) was trained for 6 epochs.).
- Stage 2 (fine-tuning): Backbone fully unfrozen and trained together with classification head for a fraction of the original learning rate (1×10^{-5}) utilizing early stopping on validation loss (patience-based restoration of best weights) to avoid overfitting as training progresses.

This staged approach allows the randomly initialized classification head to first reach a stable error-minimizing configuration without allowing gradient signals to reach and disrupt the pretrained residual layers. This safeguard proved necessary, as observed in the DenseNet121 failure mode described in Section 4.2, where parameters drifted away from their pretrained values under our initial joint training setup.

Explainability

The Grad-CAM attention heatmaps for representative test images were created and overlaid on the original ultrasound scans (Section 4.4) to confirm that the predictions of the champion model depended on clinically plausible areas rather than background noise or artifacts.

Evaluation Metrics

Evaluation of performance was performed on the held-out test set via accuracy, per-class precision, recall, and F1-score, as well as area under the receiver operating characteristic curve (AUC). A confusion matrix was created for the champion model to describe error distribution, and based on this, the balance point between false positives (negative prediction by model vs. malignant actual class) and false negatives (positive prediction by model vs. benign actual class) in the clinically important minority (malignant) class was determined.

Results

Architecture Search Summary

Table 1 gives a summary of the test set accuracy achieved by each architecture on all preprocessing methods and data splits. The two-stage progressive ResNet50 setting was a golden architecture, which performed best among all the tested deeper or more sophisticated alternative architectures with 77.61% accuracy.

Table 1. Summary of test-set accuracy across the six evaluated architectures

Configuration	Preprocessing	Test Accuracy	Note
ResNet50 (two-stage fine-tuned)	Bilateral + CLAHE + USM	77.61%	Champion configuration
ResNet152 (progressively fine-tuned)	Bilateral + CLAHE + USM	74.41%	Overparameterized for available data
EfficientNetB0	Bilateral + CLAHE + USM	72.28%	Pretrained-domain mismatch
Hybrid CNN + LBP fusion	Dual-stream (image + handcrafted)	71.64%	Feature-distribution clash
Custom CNN (from scratch)	Grayscale, normalized.	70.58%	Limited capacity, low overfitting
DenseNet121 (fully unfrozen, epoch 1)	Bilateral + CLAHE + USM	56.93%	Gradient shock (Section 4.2)
ResNet50 (two-stage fine-tuned)	None (raw grayscale)	73.13%	Without Preprocessing

Table 1 demonstrates that model performance does not scale monotonically in terms of parameter count or architectural complexity. The 2 largest tested networks (ResNet152 and a fully unfrozen DenseNet121) performed worse than the comparatively modest ResNet50. DenseNet121 collapsed to near-baseline performance under naïve fine-tuning. In Section 4.2, we discuss the reason for this collapse.

The DenseNet121 Failure Mode: Gradient Shock

We investigated in detail the fully unfrozen DenseNet121 configuration and why it fails. DenseNet121 has about 7.2 million trainable parameters, and when the entire architecture was unfrozen from epoch one, the classification head, whose weights were randomly initialized, produced large gradients in all loss functions initially. Since these gradients propagated all the way back through the network instead of being absorbed by a fixed and pre-warmed head, they disrupted standard ImageNet-feature representations before the network had time to stabilize. As a result of this, the validation loss jumped to 1.0491 at epoch #2, which was

early stopped to go back and use weights from epoch#1 instead. The weights can perform slightly better than chance but only achieved the 56.93% indicated in Table 1.

This negative effect of unfreezing the backbone too early is called a gradient shock, which agrees with prior transfer-learning literature [NLP, Section 2.1]. This was the observation which in turn led to a slightly different ResNet50 training process, where the classification head has been trained and then the backbone unfrozen [Section 3.4] to better avoid a loss of performance.

Champion Model Performance

The two-stage progressive ResNet50 model converged to its best validation loss after early stopping and was subsequently evaluated once on the held-out test set of 469 images. Table 2 presents the resulting classification metrics.

Table 2. The champion ResNet50 model details metrics

Metric	Class 0 (Benign)	Class 1 (Malignant)
Precision	0.83	0.70
Recall	0.80	0.74
F1-score	0.81	0.72

The test-set receiver operating characteristic (ROC) curve of the champion model (AUC = 0.810) is depicted in Fig. 1. As the precision/recall numbers in Table 2 indicate, the curve sharply rises to attain a true positive rate of 70%–80%, while keeping the false positive rate confined within its potential operating range (0.2–0.3).

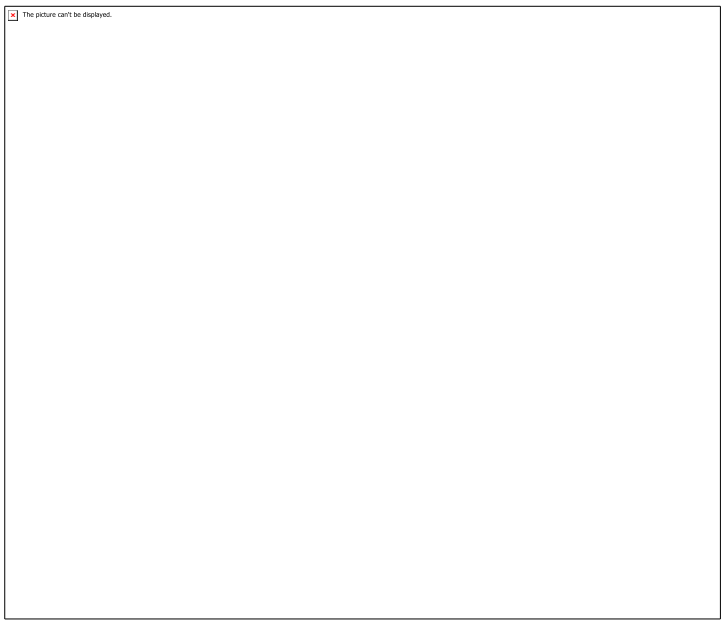


Figure 1. Receiver operating characteristic (ROC) curve for the ResNet50 model

The champion architecture confusion matrix is shown in Figure 2 for a random run. The irregularity found in this matrix is a standard characteristic of imbalanced binary tasks where benign cases (class 0) have higher absolute counts and lower relative errors than malignant ones (class 1). This is in line with the observer recall of minor classes, which was also inferior, as seen in Table 2, reflecting the dataset distribution. Usual false negatives in the malignant class are more costly in a clinical sense and, as mainly affects downstream clinical use, it is an important pattern.

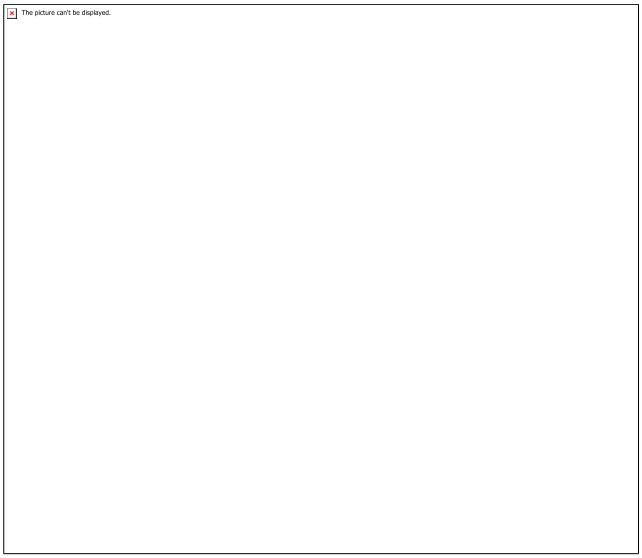


Figure 2. Confusion matrix for representative test-set evaluation of the ResNet50

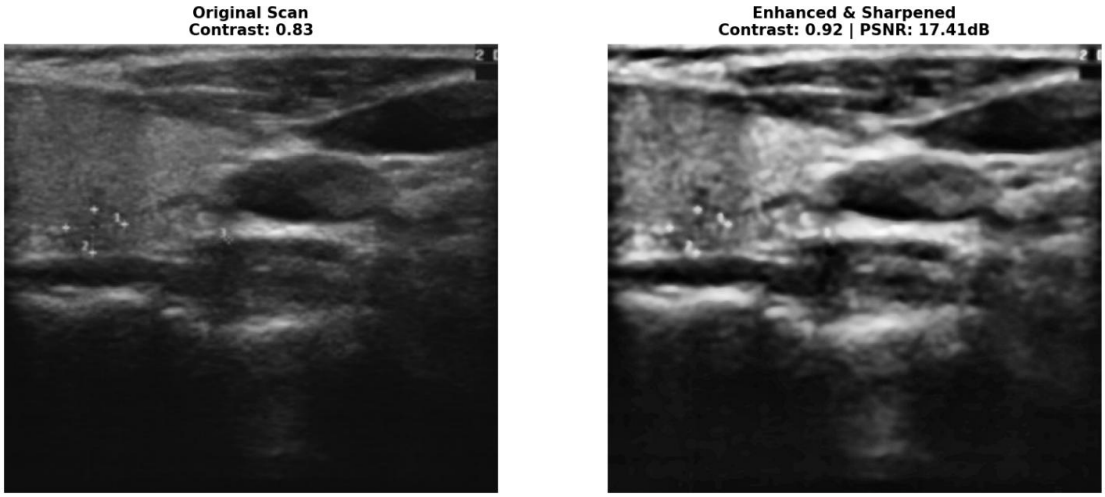
The training and validation accuracy and loss curves over 15 epochs during fine-tuning in stage 2 (progressive run) are demonstrated in Figure 3. Validation accuracy follows training accuracy reasonably closely, with a peak of 0.706 at epoch 14 but without the sharp divergence feature typically showing severe overfitting. The validation loss curve shows a local peak around epoch 6 (≈0.80), and another peak at epoch 9, but goes downwards overall to approach ≈0.63 by the end of training. This overall trend reaffirms that the early-stopping criterion managed to find a stable convergence before diverging toward end epochs.



Figure 3. Training and validation accuracy (left) and loss (right) over the course of fine-tuning

Preprocessing Ablation

Removing the preprocessing steps (bilateral filter, CLAHE, and unsharp masking) from the ResNet50 pipeline reduced the test accuracy from 77.61% to 73.13%, see Table 1. This is consistent with the preprocessing pipeline implicitly assisting the network's task and producing a difference of 4.5 percentage points.



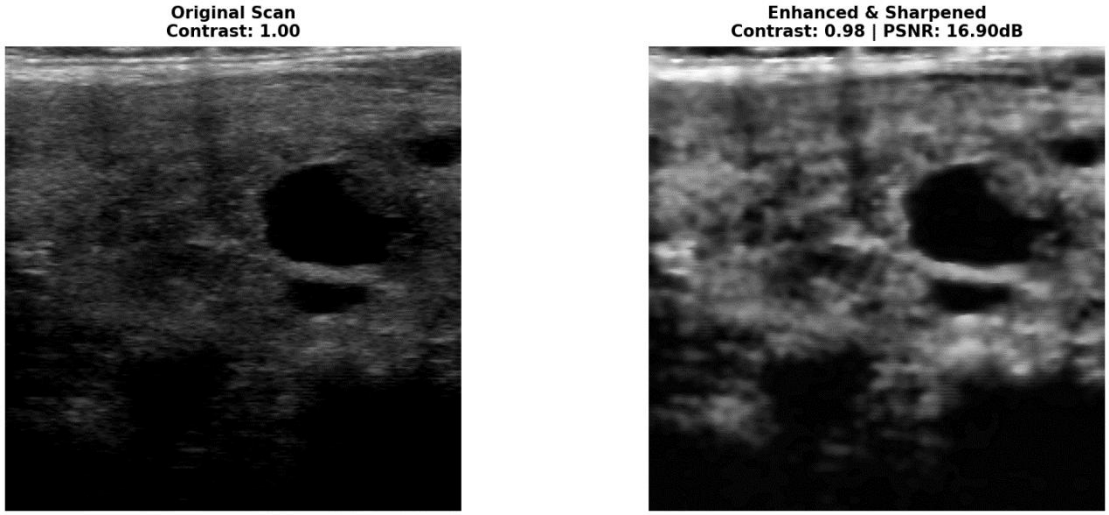


Figure 4. Examples of preprocessing results

Qualitative Explainability: Grad-CAM

Figure 4: Grad-CAM attention overlays for two test images. In both cases, the model's peak attention (red/orange regions) was in hyperechoic interfaces and focal masses consistent with nodule margins. Meanwhile, background speckles and shadow artifacts have been unmarked in the frame. On the left image, the model classified Class 0 as being with an 84.0% confidence and most of its attention was assigned to a hyperechoic band above the principal mass tissue. On the right image, the model assigned Class 0 with a confidence of 96.1% and indicated that two attention foci corresponded to adjacent hyperechoic structures.

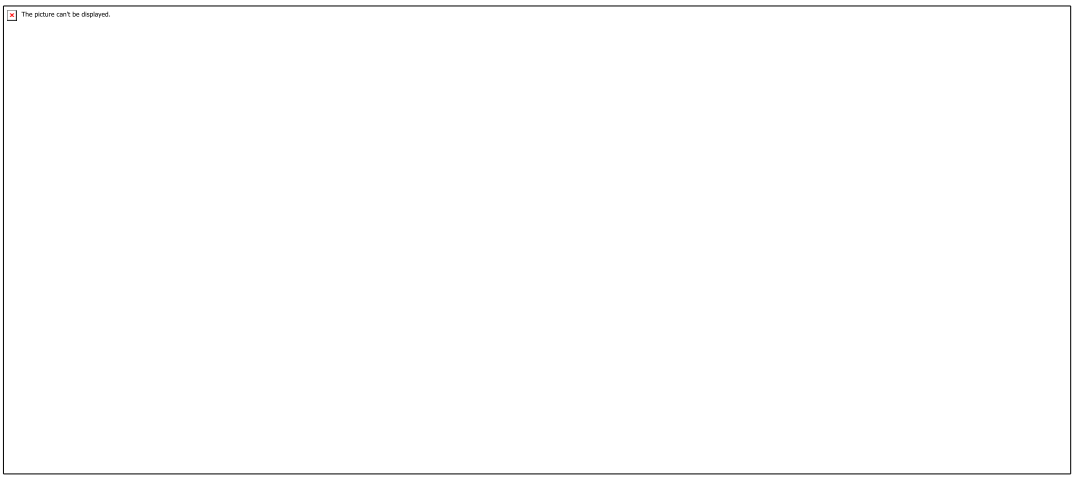


Figure 5a. Grad-CAM attention overlay, test image 0379.jpg (class 0, 84.0% confidence)



Figure 5b. Grad-CAM attention overlay, test image 0646.jpg (class 0, 96.1% confidence)

Grad-CAM is a qualitative approach, and the two examples given here are for demonstration purposes only. Nonetheless, the attention of the model always lands on anatomical structures sensible enough. Well away from image borders or background noise. At least this non-obvious checksum gives you some assurance that the champion model is not simply learning incidental artifacts, which is a typical failure mode for medical imaging CNNs and worth checking for as a standard.

Discussion

Why ResNet50 Outperformed Deeper and More Elaborate Alternatives

Main finding of the experiment: more parameters and architectural complexity do not improve performance in this dataset. This was more than three percentage points lower than ResNet152, which possesses around 3 times the residual capacity, and under similar aggressive unfrozen settings, DenseNet121 completely broke. This shape represents data starvation: 2,180 images are not enough non-noisy training data to keep the biggest models from overfitting noise instead of generalizing on the real signal during training. A second potential reason that ResNet50's residual addition blocks (x + F(x)) could hold a structural advantage on ultrasound data specifically: while DenseNet's dense concatenation pathways tend to retain and propagate low-level texture and grain information, the high-pass filtering action of residual addition can attenuate background speckle while simultaneously preserving large-scale structural boundaries, an action which seems plausibly well aligned with our bilateral-filtered, CLAHE-enhanced inputs. This explanation is not definitively confirmed by the current results but will be an interesting hypothesis to test further under controlled conditions comparing residual and simple dense architectures at matched parameter budgets.

Why the Hybrid Fusion Model Did Not Win

It is important to understand why the hybrid CNN + LBP configuration does not give better results. This threw up quite a few other potential issues (more tractable than others): A 1D mismatch: stacking a small static (32-length) LBP vector on top of the higher-order ResNet50 feature representation may leave the model's classification head with reconciling two very different scale and character feature distributions — effectively just learning noise instead of complementary signal. Second, and perhaps more importantly, the classic pre-processing pipeline applied throughout (bilateral filtering -> CLAHE -> unsharp masking) might have removed a lot of local-texture information that a crafted LBP descriptor would pass on anyway, leaving very little for fusion to add. Interesting general point — if the second explains most of the marginal value for explicit feature fusion, then that may be mostly conditioned on how much texture-relevant preprocessing was performed pre-fusion upstream in a way not always noted in this literature (i.e., fusing features).

Comparison with Published Fusion-Based Approaches

Next, we compare the current results with Xu et al. (2025), which is a good read that combines an SVM with InceptionV3 with post-fusion for AUC 0.783 on a similar-sized dataset (1,134 images, 907/227 train/test split). With a more complex classical preprocessing pipeline and tested on a much larger dataset (3,115 images), the present study's single-backbone ResNet50 model achieved an overall performance of AUC 0.810, which is broadly similar to our result. In a strict statistical sense, the two results are not directly comparable (the datasets, label definitions, and preprocessing are somewhat different), but their rough parities remain suggestive: if a single-backbone transfer-learning model is carefully pretrained/fine-tuned vs a well-delimited domain of other inputs (with some careful pre-processing), then it can compete with explicit fusion approaches — at least on these specific datasets. This should not be read as evidence against fusion modelling in general; rather, it indicates that the two design choices analysed in this work, training protocol and preprocessing, can have at least equal (or better) impact than whether any given fusion step is added or not at all.

Limitations

- Single-dataset, single-institution study: like much of the literature covered in Section 2, all reported results here are from one dataset and have not been externally validated on a different institution's scanners or patient population – the setting in which prior work (Section 2.3) suggests that performance is typically degraded.
- Ground-truth labeling: existing binary labels (Class 0/Class 1) were used as is; we did not perform an independent audit of the proportion of breeds originating from surgical

pathology versus FNA cytology, a potential source of label noise flagged in the wider literature on this dataset.

- Single-run ablations: Both the preprocessing ablation (Section [4.4]) and architecture comparisons in Table [1] reflect a single run for each configuration rather than replicated trials with variance estimates; P-values should be interpreted as suggestive, not definitive.
- Static-image classification: In line with many of the studies found in Section 2, we classify only a single representative static image per case and not multi-view or video-based information, which could lead to an underestimation of performance if ultrasound examination is represented more comprehensively.
- Qualitative, not Quantitative Explainability: The Grad-CAM analysis in Section 4.5 was indicative and based on two example cases; a systematic, quantitative audit across the entire test set (e.g., computing overlap in attention with radiographer-defined nodule boundaries) lay beyond the scope of this work and therefore represents natural future directions.

Practical Lessons

Beyond the specific accuracy figures, this architecture search yields a few practical lessons that may be of more general use to others working with modestly sized medical ultrasound datasets. First, the order of operations in transfer learning matters considerably more than is sometimes appreciated: a brief frozen-backbone warm-up stage before any fine-tuning appears to be the difference between a stable 77.61% result and a collapsed 56.93% one, using the same underlying pretrained weights. Second, bigger is not automatically better: both of the largest networks tested here underperformed a comparatively modest ResNet50, which is a useful caution against defaulting to the deepest available architecture. Third, classical preprocessing is not merely cosmetic: removing it cost the champion model over four points of accuracy, suggesting that careful denoising and contrast normalization upstream of the network remains a worthwhile investment even in the transfer-learning era.

IT Managerial Implications

The experimental results of this research provide the healthcare IT executives, clinical Chief Information Officers (CIOs), and medical software engineers who are responsible for implementing a computer-aided diagnosis (CAD) system with actionable strategic and operational insights:

1. Optimization of Infrastructure and Computational Costs

A key finding of this research is that architectural complexity and higher parameter counts do not monotonically translate to higher clinical accuracy. The comparatively modest ResNet50 backbone outperformed significantly larger networks like ResNet152 and DenseNet121. For

IT management, this implies that organizations can avoid over-provisioning expensive cloud computing resources or high-end local GPU workstations. Deploying optimized, single-backbone transfer-learning frameworks lowers data storage and computational overhead, reducing the total cost of ownership (TCO) and facilitating edge-computing deployments directly on local clinic ultrasound hardware.

2. Technical Risk Management and Algorithmic Guardrails

A very serious technical risk is indicated by the failure of DenseNet121 running with a naive, completely unfrozen setup: gradient shock. Software managers need to impose strong architectural guardrails when engineering deep learning pipelines for clinical environments. IT leadership must really enforce development guidelines mandating frozen-backbone warm-up first as a leading step to executing full fine-tuning. Filling in classification heads after stabilizing random choices before filling the pretrained part protects algorithmic assets against catastrophic forgetting, thus leading to stable learning.

3. Investment Priorities in Data Pipeline Engineering

Often, IT managers prefer to buy newer deep learning models instead of optimizing the data pipeline available. Yet the study shows that eliminating classical preprocessing (bilateral filtering, CLAHE, and unsharp masking) reduces accuracy by 4.5 percentage points here. Funding for and engineering of reliable upstream data pipelines should be prioritized in healthcare IT infrastructure. Denoising and contrast normalization should be established as scaffolding middleware components to best capitalize on the transfer-learning-based models applied against clinically heterogeneous, noisy ultrasound data.

4. Governance, Liability, and Clinical Trust

Due to their inherent operator dependence, the deployment of "black-box" neural networks in radiology carries high compliance, liability, and adoption risks. Incorporating qualitative explainability tools within IT governance frameworks is a prerequisite to building clinical trust. The use of Grad-CAM overlays with various architectures accomplishes an important visual checksum for clinical sonography practice, indicating that the network is attending to clinically plausible nodule margins rather than background artifacts or incidental noise.

Conclusion

In this study, we first presented the results of a systematic architecture search on a dataset of 3,115 ultrasound images for binary thyroid nodule classification (yes/no). Custom CNN, EfficientNetB0, DenseNet121, and ResNet152 were run with the same preprocessing pipeline with this output in mind (H5 models noted), along with the hybrid CNN-LBP fusion model [25] and ResNet50. Notably, a two-stage progressive fine-tuning protocol applied to ResNet50—frozen-backbone warmup followed by low-learning-rate full fine-tuning—yielded the strongest result, with 77.61% test accuracy and an AUC of 0.810, ahead of every deeper

or more elaborate alternative tested, such as our fully unfrozen DenseNet121 confounder, which collapsed under gradient shock. A preprocessing ablation found that the classical bilateral-filter, CLAHE, and unsharp-masking pipeline were meaningful contributors to this result, and Grad-CAM analysis provided qualitative evidence that model predictions in fact localized to clinically plausible image regions rather than background artifacts. A systematic, quantitative explainability audit and external validation across institutions are still important directions for future work before any such system could be clinically deployable. However, taken together with the broader literature reviewed in Section 2, these results suggest that, through careful training protocol design and classical preprocessing, it is possible to make well-tuned single-backbone transfer learning competitive with more complex fusion architectures.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

The author's received no financial support for the research, authorship, and/or publication of this article.

References

Buda, M., Wildman-Tobriner, B., Hoang, J. K., Thayer, D., Tessler, F. N., Middleton, W. D., & Mazurowski, M. A. (2019). Management of thyroid nodules seen on US images: Deep learning may match performance of radiologists. *Radiology*, *292*(3), 695–701.

Chambara, N., & Ying, M. (2019). The diagnostic efficiency of ultrasound computer-aided diagnosis in differentiating thyroid nodules: A systematic review and narrative synthesis. *Cancers*, *11*(11), 1759.

Dar, S. U. H., Ozbey, M., Catli, A. B., & Cukur, T. (2020). A transfer-learning approach for accelerated MRI using deep neural networks. *Magnetic Resonance in Medicine*, *84*(2), 663–685.

Durante, C., Grani, G., Lamartina, L., Filetti, S., Mandel, S. J., & Cooper, D. S. (2018). The diagnosis and management of thyroid nodules: A review. *JAMA*, *319*(9), 914–924.

Floridi, C., Cellina, M., Buccimazza, G., et al. (2019). Ultrasound imaging classifications of thyroid nodules for malignancy risk stratification and clinical management: State of the art. *Gland Surgery*, *8*(Suppl. 3), S233.

Gupta, P., Malhotra, P., Narwariya, J., Vig, L., & Shroff, G. (2020). Transfer learning for clinical time series analysis using deep neural networks. *Journal of Healthcare Informatics Research*, *4*(2), 112–137.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer, 18*(8), 500–510.

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4700–4708).

Jia, D., Wei, D., Socher, R., Li, L.-J., Kai, L., & Li, F.-F. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255).

Lee, J. H., Kim, Y. G., Ahn, Y., et al. (2023). Investigation of optimal convolutional neural network conditions for thyroid ultrasound image analysis. *Scientific Reports, 13*, 1360.

Lim, H., Devesa, S. S., Sosa, J. A., Check, D., & Kitahara, C. M. (2017). Trends in thyroid cancer incidence and mortality in the United States, 1974–2013. *JAMA, 317*(13), 1338–1348.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626).

Shao, J., Zheng, J., & Zhang, B. (2020). Deep convolutional neural networks for thyroid tumor grading using ultrasound B-mode images. *The Journal of the Acoustical Society of America, 148*(3), 1529.

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 6105–6114).

Tong, W.-J., Wu, S.-H., Cheng, M.-Q., et al. (2023). Integration of artificial intelligence decision aids to reduce workload and enhance efficiency in thyroid nodule management. *JAMA Network Open, 6*(5), e2313674.

Wu, M. H., Chen, K. Y., Shih, S. R., et al. (2020). Multi-reader multi-case study for performance evaluation of high-risk thyroid ultrasound with computer-aided detection. *Cancers, 12*(2), 373.

Xu, W., Jia, X., Mei, Z., et al. (2023). Generalizability and diagnostic performance of AI models for thyroid US. *Radiology, 307*(5), e221157.

Xu, Y., Xu, M., Geng, Z., Liu, J., & Meng, B. (2025). Thyroid nodule classification in ultrasound imaging using deep transfer learning. *BMC Cancer, 25*, 544.

Bibliographic information of this paper for citing:

Khayeat, Ali Retha Hasoon; AL-Obaidi, Ansam Abd Alrasool Ali & Alkhalisy, Muhanad Abdul Elah (2026). A Two-Stage Progressive Transfer-Learning Framework for Binary Thyroid Nodule Classification in Ultrasound Imaging. *Journal of Information Technology Management,* 18 (Special Issue), 137-153. https://doi.org/10.22059/jitm.2026.109979

Copyright © 2026, Ali Retha Hasoon Khayeat, Ansam Abd Alrasool Ali AL-Obaidi and Muhanad Abdul Elah Alkhalisy