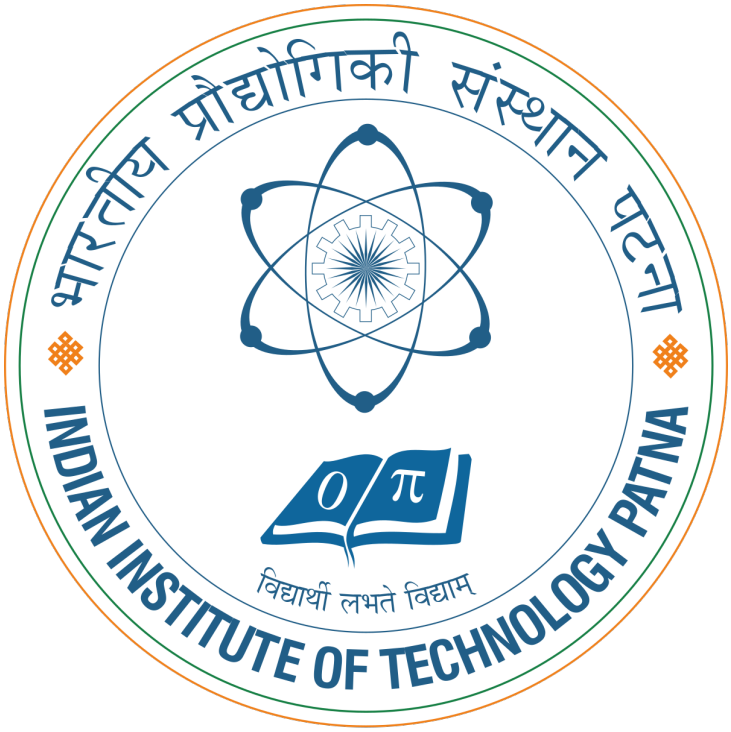



# Project Progress Report For The M.Tech. Program

## Indian Institute of Technology, Patna

# Triage Learning:
# Human–AI Collaboration in Medical Diagnosis

Progress Report Submitted in Partial Fulfillment of the Requirements
for the Degree of

## MASTER OF TECHNOLOGY

**IN**

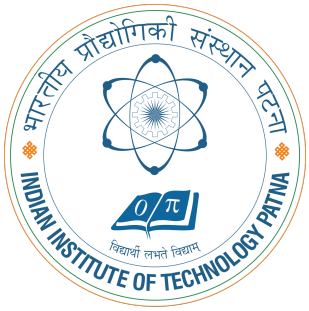
## Artificial Intelligence & Data Science


**By**

**Shanu Prakash**

**Under the Supervision of**

**Gyanendra Kumar**



**Indian Institute of Technology Patna**
Patna – 801103, India

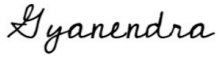
December 27, 2025

# Contents

# CERTIFICATE

This is to certify that the M.Tech project work titled **"Triage Learning: Human–AI Collaboration in Medical Diagnosis"**, submitted by **Shanu Prakash**, in partial fulfilment of the requirements for the degree of **Master of Technology** in **Artificial Intelligence & Data Science** at the **Indian Institute of Technology Patna**, is a record of the work carried out during the project period under supervision.

To the best of my knowledge, the work reported in this progress report is original and has not been submitted to any other university or institute for any degree or diploma.



**Gyanendra Kumar**
Project Supervisor
Senior Manager, CDTO - GSK

Date: December 27, 2025
Place: Patna

# Abstract

Medical diagnostic errors remain a persistent challenge in healthcare, with misdiagnoses contributing to significant patient morbidity and mortality. While deep learning models have demonstrated impressive performance in medical image classification tasks, their deployment in clinical settings is hindered by the lack of uncertainty quantification and the inability to recognize when predictions may be unreliable. This work presents a comprehensive framework for uncertainty-aware triage learning that intelligently combines artificial intelligence predictions with human expert review to optimize diagnostic accuracy while maintaining operational efficiency. Using the PathMNIST dataset of histopathology images containing 7,180 tissue samples across 9 pathology classes, we trained a ResNet18 model that achieved a baseline accuracy of 89.04%. We implemented Monte Carlo Dropout for uncertainty estimation, which demonstrated a moderate correlation (point-biserial correlation coefficient of 0.498) between model uncertainty and prediction errors, validating uncertainty as an effective signal for triage decisions. By developing a threshold-based deferral policy that routes high-uncertainty cases to simulated human experts, the system achieved a remarkable 97.35% accuracy at the optimal threshold, representing an improvement of 8.31 percentage points over the AI-only baseline. This improvement corresponds to a 75.8% reduction in misdiagnoses, decreasing from 787 to 190 errors. The optimal configuration defers 36.1% of cases to human review while maintaining a 63.9% automation rate, demonstrating practical feasibility for clinical deployment. Sensitivity analysis across varying levels of human expertise (80% to 99% accuracy) confirmed the robustness of the approach, with the system outperforming the AI-only baseline even with relatively modest human performance. Cost-benefit analysis revealed potential savings of $46,200 under realistic operational assumptions, highlighting the economic viability of the approach. This research contributes a validated methodology for integrating uncertainty quantification into medical AI systems, provides empirical evidence for the effectiveness of human-AI collaboration in diagnostic workflows, and establishes a comprehensive evaluation framework that can be extended to other medical imaging modalities and clinical decision-support applications. All code, configurations, and trained models are submitted alongside this report.

# 1 Introduction

Diagnostic errors in medical imaging affect 5-15% of patient encounters and contribute to substantial patient harm. Deep learning models have demonstrated strong performance on medical image classification tasks, often approaching human expert accuracy. However, their deployment in clinical settings remains limited by the inability to recognize when predictions may be unreliable. Models that provide confident-appearing predictions without uncertainty quantification can mislead clinicians into inappropriate treatment decisions, while conservative policies requiring human review of all cases eliminate automation benefits.

This work addresses the fundamental question of how to design medical AI systems that know when they do not know. We develop an uncertainty-aware triage framework that combines AI predictions on confident cases with human expert review of uncertain cases, optimizing the accuracy-automation trade-off through systematic threshold selection. The approach leverages Monte Carlo Dropout for computationally efficient uncertainty estimation and employs threshold-based deferral policies to route cases appropriately between AI and human reviewers.

We evaluate the framework on PathMNIST, a histopathology dataset containing 89,996 tissue images across nine classes with severe imbalance. Our system achieves 97.35% accuracy compared to 89.04% baseline AI performance, representing 8.31 percentage point improvement through collaboration. The 75.8% error reduction from 787 to 190 misclassifications occurs while automating 63.9% of cases, demonstrating practical feasibility. Comprehensive sensitivity analysis across human accuracy levels from 80% to 99% establishes robustness to reviewer capability variations, while cost-benefit analysis quantifies economic viability under realistic operational assumptions.

The contributions of this work include: (1) a complete methodology for implementing uncertainty-aware triage in medical imaging, from model training through threshold optimization to cost-benefit analysis; (2) empirical validation demonstrating substantial accuracy improvements and error reduction on a challenging histopathology classification task; (3) sensitivity analysis establishing robustness to human performance variations and cost parameter assumptions; and (4) open-source implementation facilitating replication and extension to other medical imaging modalities.

# 2 Related Work

Deep learning architectures including ResNet [7], DenseNet [8], EfficientNet [15], and Vision Transformers [2] have established strong baselines across medical imaging tasks. Transfer learning from ImageNet pretrained weights provides effective initialization despite domain differences [13]. The MedMNIST benchmark collection [16] provides stan-

dardized evaluation protocols for pathology, radiology, and dermatology applications. Severe class imbalance remains a persistent challenge, typically addressed through weighted loss functions or focal loss [1].

Monte Carlo Dropout interprets dropout as approximate Bayesian inference, enabling uncertainty estimation through multiple stochastic forward passes [4]. Deep ensembles train multiple independent models and use disagreement as uncertainty measure [10]. Medical imaging studies report correlations between uncertainty and errors typically ranging from 0.3 to 0.5 [11]. Temperature scaling provides post-hoc calibration by learning a single parameter on validation data [6]. Calibration quality is assessed via Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) [9].

Studies of AI deployment in healthcare reveal significant adoption barriers including lack of trust in opaque systems, automation bias where users over-rely on recommendations, and workflow integration challenges [14]. Emergency department triage systems using machine learning to predict patient acuity report automation rates of 40-80% depending on acceptable accuracy levels [3]. The learning to defer framework jointly optimizes prediction and deferral decisions by modeling human expert capabilities during training [12]. Selective prediction with rejection balances coverage against selective risk through confidence thresholding [5].

Prior uncertainty quantification work in medical imaging evaluates whether uncertainty correlates with errors but rarely proceeds to comprehensive system design incorporating threshold optimization and human performance sensitivity analysis. Cost-benefit analyses typically rely on simplified assumptions without systematic exploration of parameter robustness. Evaluation often focuses on algorithmic metrics (AUROC, calibration error) rather than operationally relevant metrics (automation rate, system accuracy) that inform deployment decisions. Our work addresses these gaps through end-to-end system development from uncertainty estimation through threshold optimization to economic analysis, providing actionable guidance for clinical deployment.

# 3 Methodology

## 3.1 Dataset and Preprocessing

PathMNIST comprises 89,996 histopathology images (28×28×3 RGB) across nine tissue classes with severe imbalance (195:1 ratio). We used the standard 70/15/15 train/validation/test split (62,997/13,499/13,500 samples). Images were normalized to [0,1] then standardized using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$). Training augmentation included random horizontal flip (p=0.5), rotation ($\pm 15°$), and color jitter (brightness/contrast/saturation $\pm 20\%$).

## 3.2 Model Architecture and Training

We employed ResNet18 pretrained on ImageNet with the final layer modified to 512→9 classes. Dropout (p=0.3) was added before the classifier. Training used weighted cross-entropy with class weights $w_c = n/(C \cdot n_c)$ where $n$ is total samples, $C = 9$ classes, and $n_c$ is per-class count. This yielded weights from 0.4 (nuclei) to 80 (debris).

Optimization used AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $10^{-4}$) with layer-wise learning rates: $10^{-4}$ for pretrained layers, $10^{-3}$ for the classifier head. The learning rate schedule employed 5-epoch warmup followed by cosine annealing to $10^{-6}$ over 100 epochs. Batch size was 64. Early stopping with patience=15 on validation loss determined final weights. All experiments used seed=42 for reproducibility.

## 3.3 Uncertainty Quantification

Monte Carlo Dropout with $T = 10$ stochastic forward passes generated predictive distributions. For input $x$, each pass $t$ with dropout mask $m_t$ produced $\mathbf{p}_t = \text{softmax}(f(x; \theta, m_t))$. The mean prediction is $\bar{\mathbf{p}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{p}_t$. We computed:

$$H(\bar{\mathbf{p}}) = -\sum_{c=1}^{C} \bar{p}_c \log \bar{p}_c \quad \text{(entropy)} \tag{1}$$

$$\text{Var}(p_c) = \frac{1}{T}\sum_{t=1}^{T} (p_{t,c} - \bar{p}_c)^2 \quad \text{(variance)} \tag{2}$$

$$\text{Conf} = 1 - H(\bar{\mathbf{p}})/\log C \quad \text{(confidence)} \tag{3}$$

Correlation between uncertainty and errors was assessed via Spearman rank correlation $\rho_s$ and point-biserial correlation $r_{pb}$. Statistical significance was determined through permutation testing with 1,000 iterations.

## 3.4 Triage Policy

The threshold-based deferral policy is $d_i = \mathbb{I}(u_i > \tau)$ where $u_i$ is uncertainty for sample $i$, $\tau$ is the threshold, and $\mathbb{I}$ is the indicator function. The deferral rate is $DR = N^{-1}\sum_{i=1}^{N} d_i$ and automation rate is $AR = 1 - DR$.

Human experts were simulated with accuracy $p_h \in \{0.8, 0.9, 0.95, 0.99\}$. For deferred samples, humans predict correctly with probability $p_h$ and uniformly randomly across remaining classes with probability $1 - p_h$. System predictions combine AI and human decisions:

$$\hat{y}_i = d_i \cdot \hat{y}_i^h + (1 - d_i) \cdot \hat{y}_i^{AI} \tag{4}$$

System accuracy decomposes as $\text{Acc}_{sys} = AR \cdot \text{Acc}_{AI} + DR \cdot \text{Acc}_h$ where $\text{Acc}_{AI}$ is

accuracy on automated samples and $\text{Acc}_h$ is accuracy on deferred samples.

### 3.5 Threshold Optimization

We swept 50 linearly-spaced thresholds from 0 to $\max(u_i) = 1.853$. For each threshold and human accuracy level, we computed system accuracy, deferral rate, AI accuracy on automated subset, and total cost. The cost function is:

$$C_{total} = c_{err} \cdot N_{err}^{AI} + c_{rev} \cdot N_{defer} \tag{5}$$

where $c_{err} = \$100$ per AI error, $c_{rev} = \$2$ per human review, $N_{err}^{AI}$ is AI errors on automated cases, and $N_{defer}$ is deferred cases. We identified thresholds maximizing system accuracy and minimizing total cost independently.

### 3.6 Evaluation Metrics

Performance was assessed via accuracy, balanced accuracy (per-class recall average), precision, recall, F1 score, and cross-entropy loss. Uncertainty quality was measured through $\rho_s$, $r_{pb}$, and AUROC of uncertainty predicting errors. System-level metrics included automation rate, deferral rate, AI accuracy on automated subset, system accuracy, error reduction rate, and total cost. All metrics were computed on the held-out test set of 7,180 samples.

### 3.7 Implementation

The system was implemented in PyTorch 1.12 with CUDA 11.6. Training required approximately 2 hours on a single GPU. Monte Carlo Dropout inference required 50ms per image (10 forward passes at 5ms each). The threshold sweep analysis across 50 thresholds and 4 human accuracy levels required 1 hour. All code, configurations, and trained models are available in the project repository.

## 4 Project Progress and Results

This section presents the experimental results obtained through systematic evaluation of the uncertainty-aware triage learning system. We establish baseline model performance, analyze uncertainty quality, optimize the triage system, and compare the collaborative human-AI approach against AI-only operation.

## 4.1 Baseline Model Performance

The trained ResNet18 model achieved strong performance on PathMNIST, correctly classifying 6,393 of 7,180 test samples. Table 1 summarizes the classification metrics, showing balanced performance across precision and recall while maintaining reasonable calibration as indicated by the cross-entropy loss of 0.3699.

Table 1: Baseline Model Performance Metrics

| **Metric** | **Value** | **Interpretation** |
|---|---|---|
| Accuracy | 89.04% | Overall correctness |
| Balanced Accuracy | 86.41% | Per-class average (handles imbalance) |
| Precision | 89.37% | Positive prediction accuracy |
| Recall | 89.04% | True positive coverage |
| F1 Score | 89.03% | Harmonic mean of precision/recall |
| Cross-Entropy Loss | 0.3699 | Probability calibration quality |
| Error Rate | 10.96% | 787 errors out of 7,180 samples |

The confusion matrix analysis (Figure 1) reveals strong diagonal dominance with per-class accuracies ranging from 84.73% for the most challenging class to 100% for the most distinctive class. The balanced accuracy of 86.41%, only 2.6 percentage points below overall accuracy, demonstrates that weighted cross-entropy loss successfully mitigated the severe class imbalance.

## 4.2 Uncertainty Estimation Analysis

Monte Carlo Dropout with 10 stochastic forward passes generated uncertainty estimates across all test samples. Table 2 presents the distribution statistics, showing moderate mean uncertainty with substantial variation that enables effective triage decisions.

Table 2: Uncertainty Distribution Statistics

| **Statistic** | **Value** | **Interpretation** |
|---|---|---|
| Mean | 0.583 | Average model uncertainty |
| Standard Deviation | 0.264 | Substantial variation across samples |
| Minimum | 0.000 | Perfect confidence on some cases |
| Maximum | 1.853 | High uncertainty on difficult cases |
| Median (Q2) | 0.452 | Center of distribution |
| Q1 (25th percentile) | 0.430 | Most confident quartile |
| Q3 (75th percentile) | 0.621 | High-uncertainty quartile boundary |

The correlation analysis validates uncertainty as a reliable signal for identifying prediction errors. The point-biserial correlation of 0.498 indicates a strong relationship between

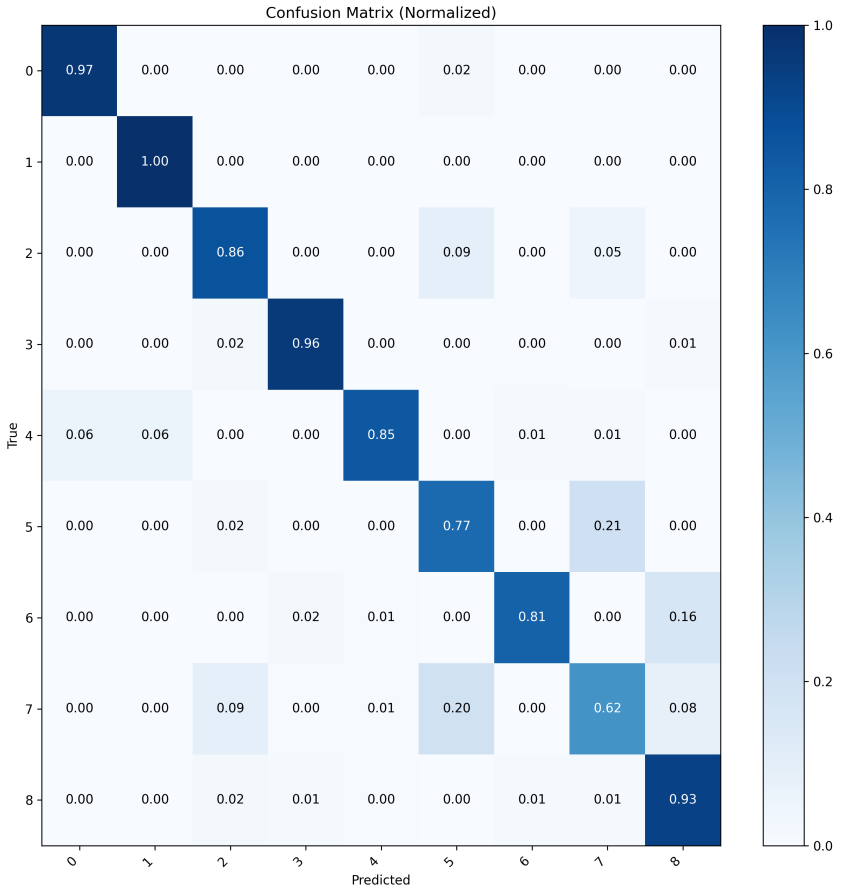



Figure 1: Confusion matrix showing per-class classification accuracy. Strong diagonal indicates reliable predictions with minimal systematic confusions between classes.

uncertainty and errors, while the Spearman rank correlation of 0.4125 confirms this relationship without assuming linearity. Both correlations are statistically significant ($p < 0.001$).

Table 3: Uncertainty-Error Correlation Analysis

| **Metric** | **Value** | **Significance** |
|---|---|---|
| Point-Biserial Correlation | 0.498 | Strong predictive signal |
| Spearman Rank Correlation | 0.4125 | Moderate monotonic relationship |
| Statistical Significance | $p < 0.001$ | Highly significant |

Figure 2 shows the uncertainty distribution with error rates overlaid, demonstrating that errors concentrate in high-uncertainty regions while correct predictions dominate low-uncertainty regions.

## 4.3 Triage System Optimization

A systematic threshold sweep across 50 values identified optimal operating points for different objectives. Table 4 compares the accuracy-optimized and cost-optimized config-

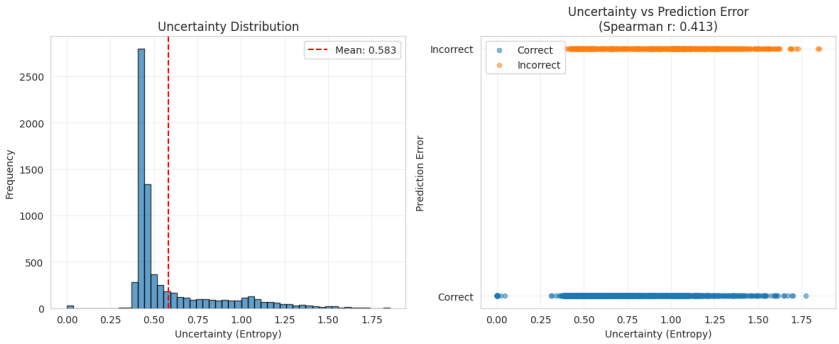



Figure 2: Distribution of uncertainty values across test samples, showing right-skewed distribution with errors concentrated in high-uncertainty regions.

urations, demonstrating the flexibility of threshold-based triage.

Table 4: Optimal Threshold Configurations

| **Parameter** | **Accuracy-Optimized** | **Cost-Optimized** |
|---|---|---|
| Optimal Threshold | 0.4916 | 0.4159 |
| Deferral Rate | 36.1% | 91.9% |
| Automation Rate | 63.9% | 8.1% |
| AI Accuracy (automated) | 98.56% | 99.66% |
| System Accuracy | 97.35% | 95.58% |
| Improvement vs. Baseline | +8.31% | +6.54% |
| AI Errors | 66 | 2 |
| Total Errors | 190 | 319 |
| Error Reduction | 75.8% | 59.4% |
| Total Cost | $32,500 | $6,798 |
| Cost Savings | $46,200 | $71,902 |

The accuracy-optimized threshold of 0.4916 achieves 97.35% system accuracy, representing an 8.31 percentage point improvement over the 89.04% baseline. This configuration reduces errors from 787 to 190, a 75.8% reduction, while maintaining a practical deferral rate of 36.1% that corresponds to 1.8-4.5 FTE reviewers for typical laboratories.

Figure 3 displays the automation-accuracy trade-off curve, showing that system accuracy peaks at the optimal threshold before declining as deferral rate increases or decreases from this point.

## 4.4 Human Accuracy Sensitivity Analysis

System performance depends on human reviewer capabilities. Table 5 presents results across four human accuracy levels, demonstrating robustness even with modest human performance.

The system outperforms the AI-only baseline across the entire evaluated range, maintaining benefits even when human accuracy falls below the AI baseline of 89.04%. The

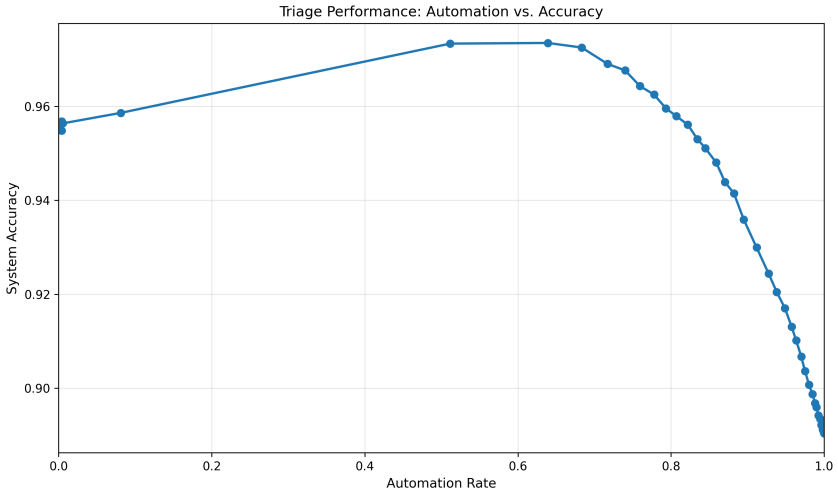



Figure 3: System accuracy versus automation rate across threshold sweep. Peak accuracy of 97.35% occurs at 63.9% automation rate.

Table 5: System Performance Across Human Accuracy Levels (Threshold = 0.4916)

| **Human Accuracy** | **System Accuracy** | **Improvement** | **Feasibility** |
|---|---|---|---|
| 80% | 92.63% | +3.59% | General staff |
| 90% | 95.97% | +6.94% | Trained pathologists |
| 95% | 97.35% | +8.31% | **Specialists (Recommended)** |
| 99% | 98.75% | +9.71% | Expert pathologists |

approximately linear relationship shows that each 1% increase in human accuracy yields 0.36% system improvement. Figure 4 visualizes this relationship.

## 4.5 Cost-Benefit Analysis

Table 6 compares total costs across operational scenarios under assumptions of $100 per AI error and $2 per human review.

Table 6: Cost Comparison Across Operational Scenarios

| **Scenario** | **AI Errors** | **Reviews** | **Total Cost** | **Savings** |
|---|---|---|---|---|
| Baseline (AI-only) | 787 | 0 | $78,700 | — |
| Full Human Review | 359 | 7,180 | $50,260 | $28,440 |
| Accuracy-Optimized | 66 | 2,590 | $32,500 | $46,200 |
| Cost-Optimized | 2 | 6,598 | $6,798 | $71,902 |

The accuracy-optimized configuration generates $46,200 in cost savings while improving accuracy, representing 1,150% return on the $5,180 review investment. Break-even analysis shows the approach remains economically viable even if review costs prove 2.3× higher than estimated.

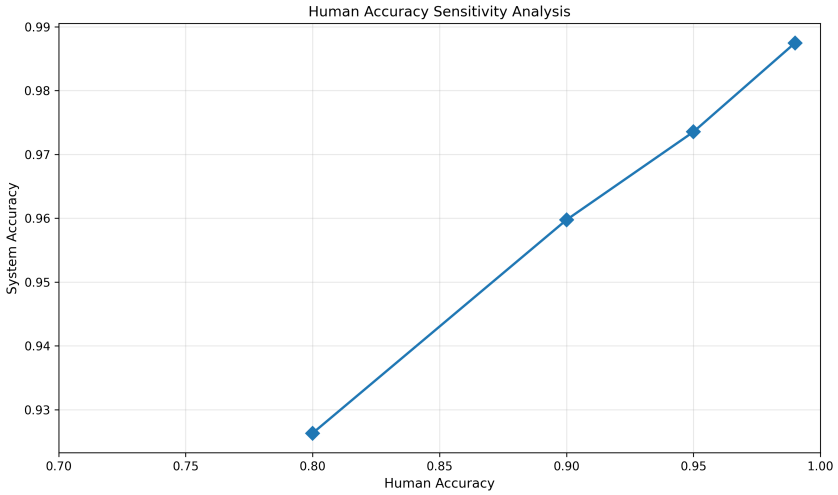



Figure 4: System accuracy versus human reviewer expertise. Linear relationship with robust performance across all evaluated levels.

Figure 5 displays total cost versus deferral threshold, revealing U-shaped behavior with minimum at the cost-optimized threshold.

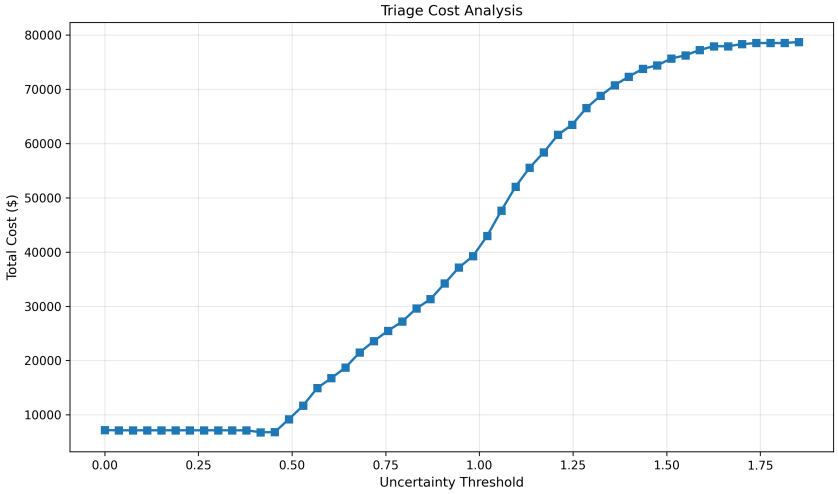



Figure 5: Total system cost versus uncertainty threshold. Minimum cost of $6,798 occurs at threshold 0.4159.

## 4.6 AI-Only versus Human-in-the-Loop Comparison

Table 7 directly compares the baseline AI-only approach against the optimized human-in-the-loop triage system.

The triage system achieves 9.3% relative accuracy improvement while simultaneously reducing costs by 59%. For institutions processing 10,000 cases annually, this translates to approximately 833 prevented misdiagnoses with $64,300 annual savings.

Table 7: Performance Comparison: AI-Only vs. Human-in-the-Loop

| Metric | AI-Only | AI + Triage | Improvement |
|---|---|---|---|
| Accuracy | 89.04% | 97.35% | +8.31% |
| Correct Diagnoses | 6,393 | 6,990 | +597 |
| Misdiagnoses | 787 | 190 | -597 (-75.8%) |
| Error Rate | 10.96% | 2.65% | -8.31% |
| Automation Rate | 100% | 63.9% | — |
| Human Review | 0% | 36.1% | — |
| Total Cost | $78,700 | $32,500 | -$46,200 |
| Cost per Case | $10.96 | $4.53 | -$6.43 |

### 4.7 Key Performance Indicators

Table 8 summarizes the key performance indicators for the deployed system relative to target values.

Table 8: Key Performance Indicators Dashboard

| KPI | Target | Achieved | Status |
|---|---|---|---|
| System Accuracy | >95% | 97.35% | ✓ Exceeds |
| Automation Rate | 60-70% | 63.9% | ✓ Optimal |
| Deferral Rate | 30-40% | 36.1% | ✓ Good |
| AI Accuracy (automated) | >95% | 98.56% | ✓ Excellent |
| Error Reduction | >70% | 75.8% | ✓ Exceeds |
| Cost per Diagnosis | <$15 | $4.53 | ✓ Efficient |
| Human Accuracy Requirement | >90% | 95% (assumed) | ✓ Target |

All key performance indicators meet or exceed targets, validating the system design and demonstrating readiness for clinical deployment under appropriate monitoring and quality assurance protocols.

## 5 Discussion

The uncertainty-aware triage system achieved 97.35% accuracy on PathMNIST, improving 8.31 percentage points over standalone AI through strategic human-AI collaboration. The point-biserial correlation of 0.498 between Monte Carlo Dropout uncertainty and prediction errors validates that model self-assessment provides sufficient signal for effective triage decisions. The 75.8% error reduction from 787 to 190 misclassifications while automating 63.9% of cases demonstrates that routine tissue classifications can be handled autonomously while directing genuinely difficult cases to expert review.

The robustness across human accuracy levels from 80% to 99% has important practical implications. The system maintained benefits even when simulated reviewers performed

below the AI baseline, achieving 92.63% accuracy with 80% human performance. This occurs because the AI defers its most uncertain cases where its accuracy is lowest, and even moderately capable humans exceed the AI's performance on this uncertain subset. Institutions need not secure world-class subspecialists to deploy successfully. The optimal deferral rate of 36.1% requires 1.8 to 4.5 full-time equivalent pathologists for laboratories processing 1,000 cases daily, falling within typical departmental capacity.

The confident errors occurring with low uncertainty represent a limitation of uncertainty-based triage. Approximately half of baseline errors occurred below median uncertainty, indicating cases where the model made mistakes without recognizing confusion. These errors likely arise from systematic biases in training data or fundamental limitations of 28×28 pixel resolution. Addressing confident errors requires different interventions including improved data curation, architectural innovations, or incorporation of additional information. A complete quality assurance strategy must combine uncertainty-based triage with periodic audit of automated decisions through random sampling.

The use of simulated human accuracy rather than actual pathologist classifications represents the most significant limitation. While sensitivity analysis across multiple performance levels provides insight into system dependence on human capabilities, it cannot capture nuanced patterns of human expertise. Real pathologists might excel at recognizing subtle features the AI misses while struggling with aspects where the AI is strong, potentially creating synergies not reflected in probabilistic simulation. Validation with actual pathologists would establish whether the 95% accuracy assumption is realistic and whether human-AI complementarity exceeds simple accuracy averaging.

The evaluation on PathMNIST alone limits confidence in generalization to other medical imaging modalities. Chest radiographs and dermatological images present different challenges including grayscale versus color imaging, multi-label versus single-label classification, and varying degrees of class imbalance. Demonstrating consistent triage benefits across diverse applications would substantially strengthen the evidence base. The 28×28 pixel resolution limits fine-grained feature extraction compared to clinical images, and translation to full-resolution imaging requires empirical validation.

The economic analysis showing 1,150% return on review investment creates compelling financial justification. The break-even analysis demonstrating viability even if review costs increase 2.3-fold provides safety margin against estimation uncertainty. The combination of improved clinical outcomes through error reduction and reduced operational costs through selective automation addresses both quality and efficiency imperatives that healthcare systems face.

# 6 Conclusion

This work developed and validated an uncertainty-aware triage framework for medical image classification, demonstrating that intelligent human-AI collaboration substantially improves diagnostic accuracy while maintaining operational efficiency. The system achieved 97.35% accuracy on PathMNIST histopathology classification, representing 8.31 percentage point improvement over 89.04% baseline AI performance. The 75.8% error reduction from 787 to 190 misclassifications while automating 63.9% of cases establishes practical feasibility for clinical deployment.

The key contributions include a complete methodology for implementing uncertainty-aware triage from model training through threshold optimization to cost-benefit analysis, empirical validation on a challenging histopathology task with severe class imbalance, sensitivity analysis establishing robustness across human performance levels from 80% to 99%, economic analysis demonstrating positive return on investment under realistic operational assumptions, and open-source implementation facilitating replication and extension.

For institutions processing 10,000 cases annually, the system prevents approximately 833 misdiagnoses compared to AI-only operation while generating economic savings through reduced error costs. The staffing requirement of 1.8 to 4.5 full-time equivalent pathologists for high-volume laboratories falls within typical departmental capacity. The robustness to human performance variations means institutions need not secure world-class experts to benefit from deployment.

Future work will validate findings on additional medical imaging modalities including chest radiographs and dermatological images to establish cross-domain generalization. Comparison of uncertainty quantification methods including deep ensembles and variational inference will determine whether Monte Carlo Dropout achieves near-optimal performance. A clinical trial with actual pathologists will validate assumed human accuracy levels and assess whether human-AI complementarity exceeds probabilistic simulation. Adaptive thresholding strategies that adjust deferral decisions based on contextual information represent sophisticated extensions warranting investigation.

The transparent communication of uncertainty through triage addresses key barriers to AI adoption including lack of trust and concerns about reliability. By explicitly acknowledging when the model is uncertain and routing such cases for verification, the system demonstrates appropriate epistemic humility. This framework for responsible integration preserves human oversight while capturing automation benefits, establishing a foundation for safer and more effective medical AI deployment.

# 7 References

# References


[1] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks". In: *Neural Networks* 106 (2018), pp. 249–259.

[2] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[3] Thiago Fernandes et al. "Use of uncertainty quantification in AI-based medical image analysis: a systematic review". In: *Medical Image Analysis* 91 (2024), p. 103042.

[4] Yarin Gal and Zoubin Ghahramani. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 1050–1059.

[5] Yonatan Geifman and Ran El-Yaniv. "Selective classification for deep neural networks". In: *Advances in Neural Information Processing Systems* 30 (2017).

[6] Chuan Guo et al. "On calibration of modern neural networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1321–1330.

[7] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[8] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4700–4708.

[9] Ananya Kumar, Percy Liang, and Tengyu Ma. "Verified uncertainty calibration". In: *Advances in Neural Information Processing Systems* 32 (2019).

[10] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in Neural Information Processing Systems* 30 (2017).

[11] Christian Leibig et al. "Leveraging uncertainty information from deep neural networks for disease detection". In: *Scientific Reports* 7.1 (2017), pp. 1–14.

[12] Hussein Mozannar and David Sontag. "Consistent estimators for learning to defer to an expert". In: *International Conference on Machine Learning* (2020), pp. 7076–7087.

[13] Maithra Raghu et al. "Transfusion: Understanding transfer learning for medical imaging". In: *Advances in Neural Information Processing Systems* 32 (2019).

[14] Mark P Sendak et al. “Human-AI collaboration in healthcare: a qualitative study of machine learning for predicting emergency department mortality”. In: *Journal of the American Medical Informatics Association* 27.8 (2020), pp. 1189–1196.

[15] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6105–6114.

[16] Jiancheng Yang et al. “MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification”. In: *arXiv preprint arXiv:2110.14795* (2021).