Translating Classifier Scores into Clinical Impact: Calibrated Risk and Queueing Simulation for AI-Assisted Radiology Worklist Triage

Tirthajit Baruah[1], **Punit Rathore**[1]

[1]Indian Institute of Science, Bengaluru
tirthajitoff@gmail.com, prathore@iisc.ac.in

Abstract

Radiology worklists are typically processed on a first-in, first-out (FIFO) basis, even when studies differ greatly in clinical urgency. We propose a pragmatic alternative: using calibrated probabilities of intracranial hemorrhage (ICH) to prioritize head CT exams for earlier reading. Using the public RSNA-ICH dataset, we train slice-level detectors, aggregate them to the exam level, apply post-hoc calibration, and feed these scores into a transparent discrete-event simulator of the reading queue. The simulator quantifies how triage benefits reduction in median time-to-read (TTR) for ICH, which scales with classifier AUC, workload (arrival rate), staffing, prevalence, and calibration. Across realistic loads, score-based prioritization yields substantial TTR reductions for ICH with minimal delay to non-ICH studies. We release a configuration-driven, reproducible pipeline that translates AI risk scores into operational metrics (minutes saved), enabling safe and data-driven evaluation before PACS/RIS[1] deployment.

Introduction

Radiology worklists are the operational front door to diagnostic care. When the list grows and all studies are treated alike, urgent cases can wait behind routine exams. In neuroimaging, such delays are undesirable. Time to interpretation can influence downstream clinical decisions, escalation pathways, and resource use. Although recent clinical reports suggest that AI-based flags can shorten turnaround for critical findings, most pre-deployment evaluations still end at diagnostic curves[2] (AUROC/AUPRC). These summarize discrimination but remain silent about certain critical operational questions.

This paper takes a deployment-centric view. We argue that decision makers need a mapping from model quality (discrimination and calibration) and site conditions (arrival rate, service times, staffing) to concrete workflow outcomes before altering the live worklist. Such a mapping should be transparent, configuration-driven, and reproducible across sites. It should quantify both gains (faster time-to-read for urgent studies) and trade-offs (impact on the overall cohort), support guardrails (tiered priority), and make explicit where benefits are largest (moderate–high utilization) and where they taper (overstaffed regimes). Finally, because priority acts on probabilities (model scores), calibration must be surfaced as a first-class concern alongside AUROC/AUPRC.

Problem. Given exam-level ICH probabilities from a detector, should a site replace FIFO with score-based priority? If so, how many minutes are saved for positives (ICH) under the site's workload? What happens to the rest of the queue (overall cohort)? How do these trade-offs depend on discrimination (AUC)? Will benefits persist at different staffing levels or if prevalence shifts?

Approach. We build an end-to-end, reproducible pipeline that: (i) trains slice-level detectors on RSNA-ICH, (ii) aggregates and calibrates exam probabilities, and (iii) evaluates *priority vs. FIFO* in a discrete-event, non-preemptive queue simulator (single and multi-server). Primary outcomes are median time-to-read (TTR) for ICH and service level attainment (fraction of ICH read within τ minutes).

Scope. The simulator translates model scores into operational metrics (minutes saved), supports stress tests across workload and staffing, and enables guardrail design (two-tier STAT policies[3], alert budgets, drift monitoring) prior to go live, aligning with responsible deployment goals around safety, transparency, and reliability.

Contributions.

- A compact, reproducible RSNA-ICH baseline (slice CNNs → exam aggregation) with post-hoc calibration and reliability analysis.
- A transparent worklist simulator for non-preemptive AI-priority vs. FIFO (single and multi-server), parameterized by arrival rate, service distribution, and prevalence.
- An empirical map from discrimination, calibration, workload, and staffing to operational benefit (minutes saved for ICH; impact on the overall cohort). A practical guidance to deploy AI based models in the clinical setup.

To sum it up, our contribution differs from prior work by providing a fully open, calibration-aware simulation pipeline that translates ICH detector probabilities into queue-level

[1] Picture-Archiving-and-Communication-Systems/Radiology-Information-System

[2] Reciever operating characteristics and Precision recall curves

[3] Clinical prioritization tiers, e.g., urgent (STAT) vs. routine cases.

metrics, enabling prospective, site-specific evaluation of AI-priority policies before PACS integration.

Related Work

ICH detection on head CT. Deep learning has reached near expert-level performance for detecting acute intracranial hemorrhage (ICH) on non-contrast head CT (Kuo et al. 2019). The RSNA Intracranial Hemorrhage (RSNA-ICH) challenge (Flanders et al. 2020) enabled standardized, slice-level benchmarks, facilitating reproducible pipelines for ICH detection. Such models form a practical foundation for triage systems in which AI outputs inform worklist order and urgency.

Calibration for decision support. Neural networks are often miscalibrated, producing unreliable probability estimates even when classification accuracy is high. Post-hoc calibrators such as temperature scaling (Guo et al. 2017) and isotonic regression (Zadrozny and Elkan 2002) improve probability reliability with minimal complexity. In clinical AI, well-calibrated probabilities support thresholding, abstention, and, essewntially, risk-based prioritization. We quantify reliability using Expected Calibration Error (ECE) (Naeini, Cooper, and Hauskrecht 2015) and the Brier score (Glenn et al. 1950). Related early work includes Platt-style scaling for SVMs and logistic models (Niculescu-Mizil and Caruana 2005).

Queueing and priority scheduling in healthcare. Queueing theory provides tools to analyze system performance under load. Classical M/G/1 priority queues characterize how waiting times shift with workload and service variability (Harchol-Balter 2013; Chou 1977). These ideas have been applied to clinical operations including patient scheduling and emergency flow (Green 2006; Jun, Jacobson, and Swisher 1999). Our simulator instantiates a minimal, transparent comparison of non-preemptive priority versus FIFO for radiology worklists, linking model discrimination (AUROC/AUPRC) and calibration to operational turnaround time (minutes saved).

Workflow triage with AI. Public evidence on AI-assisted triage in imaging workflows remains limited, particularly on open datasets that enable reproducible, end-to-end evaluation. A notable example is AI-driven chest radiograph triage reducing mean reporting delays for critical findings (Annarumma et al. 2019), though based on in-house data.

Data and Preprocessing

Dataset. We use the public RSNA Intracranial Hemorrhage (RSNA-ICH) dataset (Flanders et al. 2020), which provides non-contrast head CT DICOMs with slice-level labels for five subtypes (epidural, intraparenchymal, intraventricular, subarachnoid, subdural). We derive an *any-ICH* indicator as the slice-wise max across subtypes and later aggregate to the study level (exam).

Exam mapping and split. Each DICOM slice is associated to an exam via `StudyInstanceUID`. We ensure no patient (exam) leakage by splitting by exam into 70% train, 10% validation, and 20% test (fixed seed). All reported classifier and simulation metrics are computed at the exam level.

HU conversion and clinical windows. Raw pixel intensities are converted to Hounsfield units (HU) using the DICOM rescale parameters:

$$\mathrm{HU} = \texttt{RescaleSlope} \times \texttt{PixelValue}$$

$$+\texttt{RescaleIntercept}.$$

Then we apply three standard clinical windows and stack them as pseudo-RGB channels: brain (WL=40, WW=80), subdural (WL=80, WW=200), and bone (WL=600, WW=2800). For a window with level L (WL) and width W (WW), intensities are clipped to $[L-\frac{W}{2},\, L+\frac{W}{2}]$ and linearly scaled to $[0,1]$:

$$I_{\mathrm{win}}(x) = \mathrm{clip}\left(\frac{x-(L-W/2)}{W},\, 0,\, 1\right).$$

The three I_{win} channels are concatenated and resized to 224×224. These preprocessing steps are standard practice in head CT pipelines for both clinical visualization and deep learning (Chilamkurthy et al. 2018; Arbabshirani et al. 2018)

Artifacts and exclusions. Slices with unreadable headers, missing rescale fields, or corrupted pixels are skipped; exams with no usable slices are dropped. This affects only a small fraction of studies and does not change the class balance materially.

Storage and normalization. Windowed stacks are saved as 8-bit PNGs (per-channel $[0,255]$ from the $[0,1]$ scaling above). At training time we normalize to $[0,1]$ (no per-image standardization), then apply model-specific preprocessing from `torchvision` for pretrained backbones.

Labels and aggregation. Slice-level labels are taken from the RSNA CSV (`ID_`*`image_subtype`*) and joined by `image_id`. For evaluation and simulation for each patient study (exam-level), we max-pool the probabilities over the CT slices. Then, the probabilities correspond to any slice positive label (any-ICH).

Methods

Slice-Level Classifier and Exam Aggregation

We train lightweight 2D convolutional and transformer backbones: ResNet-18 (He et al. 2016), EfficientNet-B0 (Tan and Le 2019), and ViT-B/16 (Dosovitskiy et al. 2020). Each equipped with a shared multi-label head predicting five ICH subtypes and the derived *any-ICH* label. Let $x_i \in \mathbb{R}^{224\times224\times3}$ denote an input slice, and $f_\theta(x_i) \in [0,1]^K$ the model output probabilities over K=6 classes. Exam-level probability for *any-ICH* is obtained via slice aggregation:

$$p_{\text{exam}} = \begin{cases} \max_i f_\theta(x_i), & \text{(max pooling)} \\ \frac{1}{n}\sum_i f_\theta(x_i), & \text{(mean pooling, ablation)} \end{cases}$$

where n is the number of slices in the study. All classifier metrics (AUROC, AUPRC) are reported at the exam level, reflecting the granularity of clinical prioritization.

Calibration

The triage policy acts on *absolute risk estimates* rather than ranks. We therefore apply post-hoc calibration on the validation split. For temperature scaling (Guo et al. 2017), the logits z are rescaled as

$$p_T = \sigma(z/T), \quad T > 0,$$

where T minimizes the negative log-likelihood on validation data. For isotonic regression (Zadrozny and Elkan 2002), a non-decreasing function $g(\cdot)$ is fitted such that $p_{\text{cal}} = g(p_{\text{raw}})$. Calibration quality is measured by:

$$\begin{aligned} \text{ECE} &= \sum_{b=1}^{B} \frac{|B_b|}{N} \left| \text{acc}(B_b) - \text{conf}(B_b) \right|, \\ \text{Brier} &= \frac{1}{N} \sum_i (p_i - y_i)^2. \end{aligned} \tag{1}$$

where B_b denotes the b-th confidence bin, and y_i is the binary target. These metrics quantify reliability which is crucial when probabilities drive operational priority.

Worklist Simulation

We design a discrete-event simulator for a radiology reading queue modeled as an $M/G/c$ system with *non-preemptive service*. Exam arrivals follow a Poisson process with rate λ (studies/hour). Each service time S is sampled from a lognormal distribution:

$$S \sim \text{LogNormal}(\mu, \sigma), \qquad \mathbb{E}[S] = e^{\mu + \sigma^2/2}.$$

The number of concurrent readers (radiologists) is c, defining system utilization

$$\rho = \frac{\lambda\, \mathbb{E}[S]}{c}.$$

Two scheduling policies are compared:

- **FIFO:** Exams are served in chronological order of arrival.
- **Priority:** Exams are sorted by descending calibrated ICH probability s, with FIFO tie-breaking.

At any event, if a reader is free and the queue is non-empty, the next exam is dispatched, and start/completion times are recorded.

We simulate 8-hour sessions (plus a short warm-up) with $N_{\text{sess}} = 100$ independent replications to estimate the following metrics:

- **Median time-to-read for ICH** (TTR^+)**:** median waiting time plus service time for positive exams.
- **Service level attainment** (SLA-10 / SLA-20): fraction of ICH exams read within $\tau \in \{10, 20\}$ minutes:

  $$\text{SLA}_\tau = \frac{1}{N_+} \sum_{i=1}^{N_+} \nVdash [\, \text{TTR}_i^+ \le \tau \,].$$

  The SLA metric is then multiplied by 100, to get the percentage of ICH-positive exams interpreted within 10 or 20 minutes of arrival
- **Relative gain over FIFO:**

  $$\Delta\text{TTR(ICH)} = \text{TTR}^+_{\text{FIFO}} - \text{TTR}^+_{\text{Priority}},$$

  $$\Delta\text{TTR(all)} = \text{TTR}^{\text{all}}_{\text{FIFO}} - \text{TTR}^{\text{all}}_{\text{Priority}}.$$

  Positive Δ indicates faster turnaround under AI-driven prioritization.

This setup provides a transparent link from classifier calibration to operational time savings under realistic workload and staffing.

Priority Queueing Algorithm

Algorithm 1 Non-Preemptive Exam Triage by Calibrated Risk

```
Require: Arrival stream of completed CT exams; calibrated ICH
    scores $s \in [0, 1]$; $c$ readers; service time parameters $(\mu, \sigma)$.
for each arriving exam do
    if policy = FIFO then
        Enqueue at the tail (chronological order).
    else
        Enqueue with key $-s$ {higher $s \Rightarrow$ higher priority}
    end if
end for
while a reader is available and queue non-empty do
    Pop next exam (FIFO or highest-$s$); assign to reader.
    Sample service duration $S \sim \text{LogNormal}(\mu, \sigma)$.
    Record exam start and completion times.
end while
```

Experiments

We evaluate both *diagnostic* quality and *operational* impact of AI-assisted worklist triage through five focused studies. Unless stated otherwise, exam-level metrics are reported; simulations average 100 independent 8-hour sessions with 95% CIs.

E1: Baseline classification. Slice classifiers (ResNet-18, EfficientNet-B0, ViT-B/16) are aggregated to the exam level (max pooling). We report AUROC, AUPRC, Expected Calibration Error (ECE), and Brier after standard post-hoc calibration. Backbone comparison appears in Table 5.

E2: Calibration ablation. We compare *uncalibrated* probabilities to *isotonic* calibration fitted on validation data and evaluated on test set. Table 2 shows reliability gains (ECE, Brier) and downstream change in SLA attainment.

E3: Workload–AUC grid. To probe robustness, we inject controlled Gaussian noise in logit space to match target AUROC levels $\{0.70, 0.80, 0.85, 0.90\}$ and sweep arrival rates $\lambda \in \{4, 5, 6\}$ studies/hour. For each (λ, AUC) we compute $\Delta\text{TTR(ICH)}$ and $\Delta\text{TTR(all)}$ (FIFO–Priority). The isobenefit plot (Figure 2) visualizes how gains scale jointly with workload and discrimination.

E4: Staffing sensitivity. We vary concurrent readers $c \in \{1, 2, 3, 4\}$ under a fixed (λ, AUC) to quantify how lower utilization reduces queueing and narrows the marginal benefit of AI-priority as staffing increases. Figure 3 summarizes minutes saved for ICH reading with CIs.

E5: Prevalence sensitivity. Holding workload fixed, we resample the test cohort to emulate ICH prevalence $\{5\%, 10\%\}$. Table 4 reports SLA-10/20, highlighting that prevalence mainly shifts the fraction of positives meeting time targets, while absolute $\Delta\mathrm{TTR}$ is driven by load and AUC.

Reference setting for point estimates. Where a single setting is needed for comparison (e.g., Table 1 and Figure 1), we use λ=5 studies/hour and target AUC=0.85.

Results

AI-priority accelerates ICH interpretation with minimal global delay. Table 1 provides the baseline comparison (FIFO vs AI based priority) which underlines this analysis. Under a representative mid-load condition (λ=5 studies/hour, AUC=0.85), *EfficientNet-B0* reduces the median ICH time-to-read by ≈102 min relative to FIFO, while the overall cohort remains slightly faster ($\Delta\mathrm{TTR}(\mathrm{all})$=7.8 min). *ResNet-18* and *ViT-B/16* show comparable directional gains. Figure 1 summarizes per-session trade-offs: most sessions fall in the upper-right quadrant, where both ICH and overall turnaround improve; or the upper-left quadrant, where ICH reads are faster with small overall delay. Sessions in the lower half (no triage benefit) are rare, indicating a consistent positive impact of AI-based triaging across replications.

Table 1: FIFO baseline vs. AI-priority at λ=5, AUC= 0.85.

Model	$\mathrm{TTR}_{\mathrm{ICH}}$↓	SLA10↑	SLA20↑	$\Delta\mathrm{TTR}_{\mathrm{ICH}}$ / $\Delta\mathrm{TTR}_{\mathrm{all}}$ ↑
FIFO	198.6	0.2	1.3	- -
ResNet	102.4	0.7	6.5	96.2 / **8.1**
EfficientNet	**96.6**	**0.9**	6.5	**102.0** / 7.8
ViT	99.4	0.8	**7.5**	99.1 / 7.9

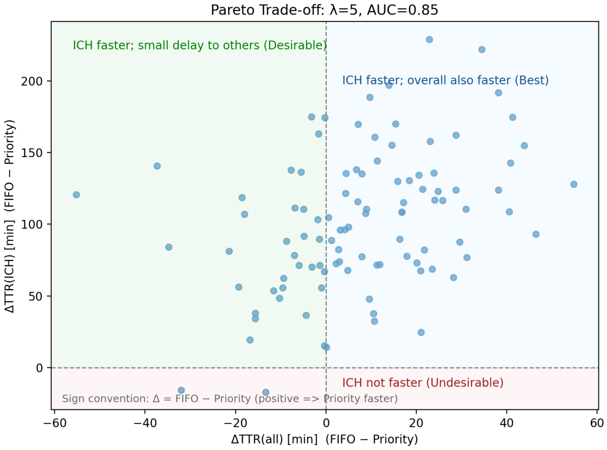

Figure 1: **Pareto trade-off between ICH and overall turnaround (EfficientNet-B0).** Each point represents one 8-hour simulated session at λ=5 (studies/hour) and AUC= 0.85. Most sessions lie where AI-priority yields faster ICH reading (upper half); the upper-right quadrant corresponds to simultaneous improvement for ICH and the overall cohort.

Workload–AUC grid: larger benefit with higher load and better discrimination. Across the (λ, AUC) grid, the mean benefit $\Delta\mathrm{TTR}(\mathrm{ICH})$ grows smoothly with both arrival rate and detector quality (Figure 2). At low loads (λ=4) and moderate AUC (≤ 0.8), queues are short and triage gains are modest (< 60 min). As workload intensifies (λ=6) or model quality improves (AUC> 0.85), median minutes saved exceed 120. The trend confirms queueing-theoretic expectations: stronger discrimination and heavier traffic amplify the advantage of prioritization. This mapping helps practitioners anticipate operational ROI for different model qualities and staffing levels before deployment.

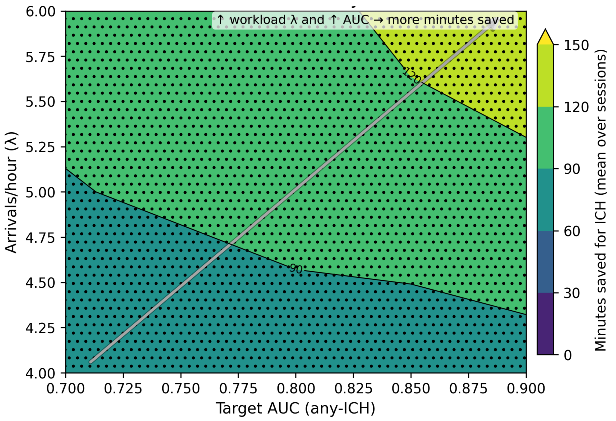

Figure 2: **Isobenefit map**: mean $\Delta\mathrm{TTR}(\mathrm{ICH})$. Benefit increases with both arrival rate λ (heavier workload) and AUC (better discrimination). Hatched cells mark 95% CIs excluding zero benefit.

Staffing sensitivity. Increasing the number of concurrent readers c reduces utilization $\rho = \lambda\mathbb{E}[S]/c$ and shortens queues, diminishing marginal benefit (Figure 3). Even so, at realistic workloads (λ=5–6$/hr$, $\rho \approx$ 0.4–0.8), median ICH savings remain substantial (40–100 min). The effect size scales inversely with staffing, confirming that triage adds most value when resources are constrained.

Calibration improves reliability and SLA attainment. Isotonic calibration improves both probabilistic reliability (ECE and Brier) and service-level attainment (SLA-20; Table 2). For instance, *EfficientNet-B0* shows ECE drop from $0.094 \rightarrow 0.014$ and Brier from $0.071 \rightarrow 0.050$, while SLA-20 rises from $6.7 \rightarrow 7.17\%$. The consistent trend across backbones emphasises the value of simple post-hoc calibration before deploying score-based triage.

Table 2: Calibration ablation (uncalibrated vs. isotonic).

Model	ECE↓	Brier↓	SLA-20↑
ResNet	0.066 / **0.010**	0.060 / **0.051**	7.1 / 7.0
EfficientNet	0.094 / **0.014**	0.071 / **0.050**	6.7 / **7.1**
ViT	0.114 / **0.023**	0.113 / **0.094**	6.7 / **7.3**

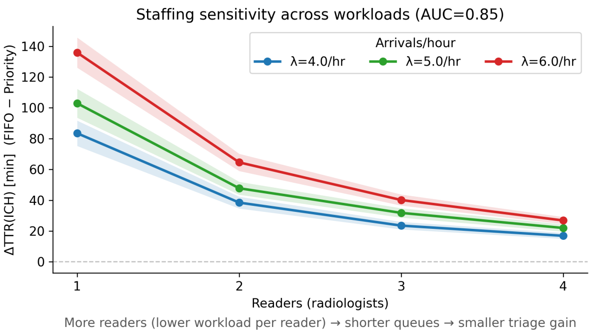

Figure 3: **Staffing sensitivity across workloads (AUC$=$ 0.85).** As readers increase, utilization ρ decreases and queues shorten, reducing triage benefit. Yet AI-priority consistently saves $40 - 100$ min for ICH across realistic staffing levels. Shaded bands denote 95% CIs over 100 simulated sessions.

Aggregation ablation. At the exam level, taking the maximum probability across slices outperforms mean pooling for SLA metrics (Table 3). This aligns with the sparsity of the nature of the hemorrhagic dataset: a single highly positive slice is sufficient for detection, and max-pooling preserves this signal.

Table 3: Aggregation (CT slices) ablation at exam level.

Model	SLA-10 (max/mean)	SLA-20 (max/mean)
ResNet	0.9 / 0.9	7.0 / 7.2
EfficientNet	**0.8** / 0.7	**7.1** / 6.4
ViT	0.9 / 0.9	**7.3** / 7.1

Prevalence sensitivity. Higher ICH prevalence mechanically increases queue congestion and reduces SLA attainment (Table 4). The absolute $\Delta\mathrm{TTR}$ (minutes saved) depends more on workload and AUC than on prevalence. Prevalence mainly shifts the fraction of urgent cases meeting time targets.

Table 4: Prevalence sensitivity at SLA-10/20 attainment.

Model	SLA-10 (5%/10%)	SLA-20 (5%/10%)
ResNet	**4.5** / 2.1	**26.7** / 18.2
EfficientNet	**3.1** /1.6	**34.5** / 22.0
ViT	3.9 /4.5	**23.9** / 20.6

Backbone comparison. Finally we perform ablation over the backbone models used in this study (Table 5). *EfficientNet-B0* achieves the highest exam-level discrimination (AUROC/AUPRC≈0.98) and best overall reliability (ECE≈0.08 post-calibration), translating into the largest operational gain. *ViT-B/16* underperforms, likely due to limited training data and 2D supervision (ViTs typically require stronger augmentation and longer schedules). Our primary goal is to evaluate triage dynamics rather than architecture design. The backbone ablation is included for completeness. The results across all the experiments validate that AI-based priority consistently accelerates urgent reads regardless of backbone.

Table 5: Backbone ablation (exam-level, post-calibration).

Model	AUROC	AUPRC	ECE	Brier
ResNet	0.976	0.974	**0.045**	**0.059**
EfficientNet	**0.977**	**0.975**	0.077	0.073
ViT	0.935	0.928	0.098	0.114

Discussion

Triaging helps. In non-preemptive priority queues, any score correlated with true urgency reorders service in favor of positives and shortens their expected delay (Chou 1977; Harchol-Balter 2013). Our workload–AUC grid reproduces this classical behavior that benefit grows as utilization rises (queues form) and as model discrimination improves. The Pareto frontier makes the trade space explicit. Large ICH gains can be achieved with only modest impact on the overall cohort.

Calibration matters. Priority acts on *absolute* risk. Poorly calibrated models compress or distort score scales near decision-critical thresholds. This degrades both rank consistency and fairness of prioritization (Guo et al. 2017; Naeini, Cooper, and Hauskrecht 2015). In our experiments, simple isotonic scaling reduced ECE and Brier scores and modestly improved SLA attainment, demonstrating that even lightweight post-hoc calibration can translate into measurable operational benefit.

Aggregation and sparsity. For focal pathologies such as ICH, exam-level max pooling across slices consistently outperforms mean pooling on SLA metrics. This reflects the sparse lesion burden: a single highly suspicious slice should dominate the decision. Such aggregation remains a pragmatic proxy for full 3D or temporal models when only 2D supervision is available.

Operational guardrails. Deployment requires design beyond algorithmic accuracy. Sites should bound alert volume, use multi-tier policies (like STAT/High/Normal) with clinician override and audit trails, and pilot in "silent" mode with drift monitoring and periodic recalibration. Our simulator enables prospective stress testing of these human-in-the-loop policies before PACS/RIS integration, turning theoretical gains into validated operational design.

Data and research gaps. Public, workflow-level datasets for AI triage remain scarce, especially in radiology. MIMIC-IV-Ext offers partial coverage (vitals, labs, and clinical notes) but lacks end-to-end imaging workflow timestamps and routing signals needed to evaluate queueing interventions (Johnson et al. 2023). This gap leaves substantial room for multimodal, workflow-aware benchmarks that link perception models to system-level outcomes, including real-time dispatch and escalation.

Limitations

Our simulation is deliberately stylized for clarity and reproducibility. (i) *Synthetic arrivals and service.* We assume Poisson arrivals and lognormal service times for a single specialty. Real clinical worklists include batching, interruptions, and multitasking. (ii) *Dataset scope.* RSNA-ICH provides slice-level labels but no operational timestamps or outcomes. Hence, we report *time-to-read* improvements as an operational proxy rather than clinical outcome. (iii) *Modeling choices.* We use 3-window 2D slices and short training schedules. Stronger volumetric or hybrid architectures with longer ViT training could improve discrimination and magnify the simulated benefits. (iv) *Generality.* We study ICH as an exemplar of acute, sparse findings. Other pathologies with different prevalence, calibration, or aggregation dynamics may yield distinct behaviors.

Despite these simplifications, the qualitative patterns of triage benefit increase with load and discrimination, and the clear advantage of calibration is theoretically grounded and robust across configurations.

Broader opportunity. Integrating structured EHR, imaging, and clinical notes with large language models and agentic AI planners could enable richer triage signals (symptom-aware risk, escalation budgets, or time-critical task routing). Our configuration-driven simulator provides a controllable testbed for such extensions once appropriately consented multimodal workflow logs become available.

Conclusion

Calibrated, score-based prioritization offers a practical means of translating model discrimination into operational benefit. On RSNA-ICH, *AI-priority* consistently accelerates interpretation of hemorrhagic cases with minimal impact on other studies, and gains scale with both discrimination and workload. By releasing a transparent pipeline, *from calibration to discrete-event simulation*, we enable sites to quantify AI-driven benefit in *minutes saved* before deployment and to stress-test operational guardrails such as tiered alerts, staffing, and service thresholds. This work establishes a reproducible foundation for linking perception-level AI models to workflow-level outcomes and opens the path towards multimodal, agentic, and data-driven triage systems for responsible clinical deployment.

Acknowledgments

We thank collaborators and domain experts their helpful feedback. Large language models (LLMs) was used only for minor text refinement. All experiments and analyses were concieved and verified by the authors.

Ethics Statement

This study used only publicly available, de-identified data (RSNA-ICH). No human subjects or patient interaction were involved, and no institutional approval was required.

Code and Data Availability

The RSNA Intracranial Hemorrhage (RSNA-ICH) dataset is publicly available. All preprocessing, training, calibration, and simulation scripts will be released in the GitHub repository of the primary author upon publication to ensure full reproducibility.

References

Annarumma, M.; Withey, S. J.; Bakewell, R. J.; Pesce, E.; Goh, V.; and Montana, G. 2019. Automated triaging of adult chest radiographs with deep artificial neural networks. *Radiology*, 291(1): 196–202.

Arbabshirani, M. R.; Fornwalt, B. K.; Mongelluzzo, G. J.; Suever, J. D.; Geise, B. D.; Patel, A. A.; and Moore, G. J. 2018. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *NPJ digital medicine*, 1(1): 9.

Chilamkurthy, S.; Ghosh, R.; Tanamala, S.; Biviji, M.; Campeau, N. G.; Venugopal, V. K.; Mahajan, V.; Rao, P.; and Warier, P. 2018. Development and validation of deep learning algorithms for detection of critical findings in head CT scans. *arXiv preprint arXiv:1803.05854*.

Chou, W. 1977. Queueing Systems, Volume II: Computer Applications-Leonard Kleinrock. *IEEE Transactions on Communications*, 25(1): 180–180.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*.

Flanders, A. E.; Prevedello, L. M.; Shih, G.; Halabi, S. S.; Kalpathy-Cramer, J.; Ball, R.; Mongan, J. T.; Stein, A.; Kitamura, F. C.; Lungren, M. P.; et al. 2020. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3): e190211.

Glenn, W. B.; et al. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1): 1–3.

Green, L. 2006. Queueing analysis in healthcare. In *Patient flow: reducing delay in healthcare delivery*, 281–307. Springer.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Harchol-Balter, M. 2013. *Performance modeling and design of computer systems: queueing theory in action*. Cambridge University Press.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Johnson, A. E.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T. J.; Hao, S.; Moody, B.; Gow, B.; et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1.

Jun, J. B.; Jacobson, S. H.; and Swisher, J. R. 1999. Application of discrete-event simulation in health care clinics: A survey. *Journal of the operational research society*, 50(2): 109–123.

Kuo, W.; Hne, C.; Mukherjee, P.; Malik, J.; and Yuh, E. L. 2019. Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. *Proceedings of the National Academy of Sciences*, 116(45): 22737–22745.

Naeini, M. P.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Niculescu-Mizil, A.; and Caruana, R. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, 625–632.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Zadrozny, B.; and Elkan, C. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 694–699.