ML-Powered Triage and Queue Optimization for Resource-Constrained Free Clinics

Armaan V. Grewal AGREWAL@CPS.EDU
[Walter Payton College Prep, Chicago, IL]

Abstract

Free clinics serve about 1.7 million uninsured Americans annually in the US, yet operate under severe resource constraints that lead to missed urgent cases, inefficient patient flow, and long waits. Our machine learning (ML)-powered triage and queue optimization system is designed as a **decision support tool specifically for resource-constrained free clinics**. Our system combines a Random Forest (RF) classifier trained on MIMIC-IV-ED data with a multi-objective queue optimization algorithm presented via a staff-facing interface. Our triage model, when simulating free clinic deployment without vital sign equipment, achieves an 83.6% critical case detection rate and no dangerous misses on the test set (0.012% on hold-out); optimizing for patient safety over raw accuracy. Monte Carlo simulation across 1,000 clinic sessions demonstrates a 72% reduction in wait times ($p < 0.001$) for critical patients compared to first-come-first-served (FCFS) queue ordering. Unlike commercial triage systems that can be prohibitively expensive, our solution is built entirely on free and free-tier tools and designed for volunteer-staffed environments lacking trained intake nurses, vital sign monitors, and electronic health records (EHR). We developed the system as a tablet-optimized web application with real-time queue updates and physician queue overrides. Our work is entirely retrospective evaluation and simulation. Prospective studies in free clinic settings, in collaboration with clinicians and domain experts, are planned as a next step.

Data and Code Availability This paper uses data sampled from the MIMIC-IV-ED (Medical Information Mart for Intensive Care, Emergency Department module) dataset (Johnson et al., 2023), which is publicly available through PhysioNet to credentialed users. We use a stratified 80/20 train/test split. The dataset and model are available upon reasonable request.

Institutional Review Board (IRB) MIMIC-IV-ED is a publicly available, de-identified dataset. The original dataset was collected under IRB approval at Beth Israel Deaconess Medical Center (BIDMC) with informed consent waived for secondary research use of de-identified data. As our research involved secondary analysis of publicly available, de-identified data with no direct human subject involvement and no collection of new data, this study does not constitute human subject research, and IRB approval was not required.

1. Introduction

Free clinics represent a critical safety net in the American healthcare system, with approximately 1,400 clinics serving about 1.7 million uninsured and underinsured patients annually (National Association of Free & Charitable Clinics, 2023). These clinics operate under unique constraints: limited staffing with volunteer healthcare providers working ∼4-hour shifts, lack of EHR, limited or no diagnostic equipment, patient populations with high prevalence of chronic conditions, diverse cultural backgrounds that may amplify language barriers, and walk-in operations with unpredictable daily patient volumes. Addressing these challenges requires not just a triage model but a fully integrated system encompassing urgency classification, queue management, and a deployable user interface, the complete pipeline we present here.

Free clinics have two distinct volunteer types: *clinicians* (physicians and nurses) who provide patient care, and *intake staff* (untrained community volunteers) who handle patient registration and information collection at check-in. Critically, free clinic intake is often handled by untrained

community volunteers, not clinical staff; these volunteers can reliably record patient-reported information (symptom severity, duration, and red flags) but are not equipped to perform ESI-style triage, which requires trained nurses and vital sign collection and interpretation.

Current patient flow management in free clinics relies predominantly on First-Come-First-Served (FCFS) ordering, with untrained intake staff occasionally using subjective judgment to perform informal triage (Davidian and Bloom, 2024) that may help or hurt. This approach creates multiple failure modes. It is common for free clinics to have long wait times (Min et al., 2024), leading to patient deterioration or departure without care. FCFS systems also inadvertently favor patients with flexible schedules and greater access to transportation, creating inequitable access patterns.

Existing emergency triage systems such as the Emergency Severity Index (ESI) are designed for well-resourced hospital settings with trained triage nurses and vital sign monitoring equipment (Gilboy et al., 2020). Recent large-scale studies have revealed significant limitations: a study of over 5 million ED encounters found ESI achieves only 65.9% sensitivity for high-acuity cases with a 32.2% overall mis-triage rate (Brown et al., 2023). The study also identified racial disparities, with Black patients experiencing an 18.5% higher under-triage risk. Commercial healthcare queue management systems typically cost thousands of dollars annually, which is far beyond free clinic budgets.

1.1. Transfer Learning from ED to Free Clinic Settings

Our ML-powered triage and queue optimization system presented via a staff-facing interface is specifically designed for the free clinic context. This integrated scope reflects a deliberate design philosophy: a triage model alone is not clinically actionable without the queue management and interface layers that translate predictions into real free clinic workflows.

The intake interface is designed for use by untrained community volunteers who staff intake desks, while clinicians retain full authority over patient care decisions and can manually override queue ordering. Our contributions include:

- **Domain Adaptation:** Given the lack of publicly available free clinic data, we adapted an emergency department (ED) dataset to train a distribution-aware ML triage model for resource-constrained free clinic settings that may lack vital sign monitors, trained triage nurses, and electronic health records. We validate the model under both Full Vitals and Vitals-Imputed conditions to address distribution shift concerns.
- **Safety-Optimized Classification:** An RF model that prioritizes critical case detection (83.6%) and minimizing dangerous misclassifications over overall accuracy (77.0%).
- **Multi-Objective Queue Optimization:** An algorithm balancing medical urgency with fairness constraints, achieving a 72% reduction in critical patient wait times.
- **Practical Deployment:** An open-source implementation built entirely on free and free-tier tools, designed to be deployable on low-cost tablets in volunteer-staffed free clinics pending prospective clinical evaluation.
- **Domain Expertise:** Drawing on firsthand experience as free clinic volunteers and iterative consultation with clinic staff, nurses, and physicians, we designed a solution grounded in operational realities for both volunteer types and began establishing partnerships for future IRB-approved prospective evaluation, which is required to assess real-world impact.

We adopt a lifecycle approach to bias mitigation, acknowledging that harm can arise not just from data (representation bias) but also during model definition (aggregation bias) and deployment (deployment bias), particularly when tools trained in one context are applied to another (Suresh and Guttag, 2021).

2. Related Work

2.1. Machine Learning for ED Triage

ML approaches to ED triage have demonstrated consistent improvements over traditional scoring systems. Multiple systematic reviews have examined this: Martos-Cabrera et al. (2022) confirmed XGBoost and deep neural networks as top performers; and Miles et al. (2020) provided a comprehensive review of ML risk prediction models for triage. Raita et al. (2019) demonstrated that ML models achieved AUROC[1] of 0.86 for critical care outcome prediction compared to 0.74 for

1. AUROC: Area Under the Receiver Operating Characteristic Curve.

ESI-based models. Similarly, Hong et al. (2018) achieved AUROC of 0.91 for hospital admission prediction using XGBoost with triage information and patient history.

Deep learning has also demonstrated the ability to predict in-hospital mortality and readmission from raw EHR data with high accuracy (Rajkomar et al., 2018). However, applying these complex models in resource-constrained environments requires balancing performance with the interpretability and infrastructure limitations of free clinics, motivating our choice of RF.

Recent prospective studies have validated these findings. Yi et al. (2024) conducted a systematic review of prospective AI triage studies, finding prediction accuracy ranging from 80.5% to 99.1% across different implementations. Sitthiprawat et al. (2025) developed an XGBoost model achieving AUROC of 0.917 compared to 0.882 for the Canadian Triage and Acuity Scale (CTAS), demonstrating ML superiority in real-world validation.

2.2. MIMIC-IV-ED Benchmarks

The MIMIC-IV-ED dataset has emerged as the primary benchmark for ED prediction models. Xie et al. (2022) created a comprehensive benchmark from MIMIC-IV-ED containing over 400,000 ED visits with three clinical prediction tasks (72-hour ED revisit, hospitalization, and critical outcomes), finding that ML models consistently outperformed clinical scoring systems including MEWS, NEWS, and CART[2]. The original MIMIC-IV dataset paper (Johnson et al., 2023) established data quality and clinical validity standards that enable reproducible research.

2.3. Limitations of Traditional Triage Systems

The ESI, used in over 70% of US ED (Gilboy et al., 2020), exhibits significant limitations. Chmielewski and Moretz (2022) found that appropriate adherence to ESI may be as low as 60%, with wide variation across facilities. For pediatric populations, Sax et al. (2024) found only 34.1% of visits were correctly triaged using ESI version 4, with 58.5% over-triaged. Travers et al. (2009) established moderate interrater reliability (Cohen's Kappa, $\kappa = 0.57$) of the ESI for the double triage method by trained ED nurses.

2. MEWS: Modified Early Warning Score; NEWS: National Early Warning Score; CART: Classification and Regression Trees.

2.4. Fairness in Clinical Prediction Models

Algorithmic fairness in clinical prediction requires careful attention to how prediction problems are formulated. Benjamin (2019) argues that indifference to social reality in algorithm design, such as using historical data without accounting for structural inequities, can automate and deepen existing forms of racism in healthcare delivery. This concern is particularly acute for triage systems serving populations who have historically faced barriers to care.

A systematic review by Siddique et al. (2024) found that while some algorithms exacerbate disparities, those designed with intentionality can reduce them. Our safety-optimized approach, prioritizing critical detection over overall accuracy, represents such an intentional design choice. However, pre-emptive bias corrections based on training data from one population can introduce new distribution shift bias in different deployment settings; thus, we adopt a monitor-then-adjust framework that defers race-based interventions until prospective data from target free clinic populations is available.

2.5. Free Clinic Operations and Healthcare Access

Free clinics face unique operational challenges that existing triage systems do not address. Davidian and Bloom (2024) documented that understaffed and underfunded free clinics struggle with long waitlists and extended wait times while relying entirely on mostly untrained volunteer intake staff. Min et al. (2024)'s retrospective analysis found that medical complexity, volunteer training levels, and on-site laboratory availability significantly affected wait times, with triage time itself being a key bottleneck.

2.6. Social Determinants of Health

Free clinic patients face compounding social determinants of health that affect both disease burden and healthcare-seeking behavior. But traditional triage systems do not account for factors such as housing instability, food insecurity, limited transportation, and lack of health insurance. Our queue optimization algorithm's fairness weight helps ensure that patients facing longer travel times (hence, arriving late at the free clinic) or less schedule flexibility are not systematically disadvantaged.

2.7. Queue Optimization in Healthcare

Patient queueing in emergency settings has received substantial research attention. Elalouf and Wachtel (2022) reviewed 229 articles spanning seven decades on ED queueing problems, identifying key approaches including patient prioritization, fast-track systems, and dynamic resource allocation. Rashwan et al. (2022) combined discrete event simulation with neural networks and genetic algorithms to minimize patient waiting times. Lee and Lee (2020) applied deep reinforcement learning to patient scheduling, demonstrating that learned policies outperformed traditional dispatching rules.

2.8. Gap and Contribution

While ML triage systems have shown promise in well-resourced ED settings, to our knowledge, no prior work applies ML-triage and queue optimization to volunteer-staffed free clinics. Commercial solutions remain prohibitively expensive, and research prototypes assume infrastructure that free clinics lack. Our work bridges this gap by training on ED data while designing for contexts lacking infrastructure, optimizing for safety rather than overall accuracy, and building a practical software system built on free and free-tier infrastructure and deployable on cheap devices.

3. Methods

3.1. Dataset

We utilized the MIMIC-IV-ED, a publicly available dataset maintained by MIT and BIDMC containing de-identified records from over 425,000 ED encounters (Johnson et al., 2023; Xie et al., 2022).

The dataset's `acuity` field encodes ESI triage levels 1 (highest priority, immediate life threat) through 5 (lowest priority, non-urgent), as defined by Gilboy et al. (2020) and validated by Xie et al. (2022). ESI levels 1–2 correspond to cases requiring immediate intervention for life-threatening conditions. We train and evaluate our model using this 5-level schema directly; however, **in free clinic deployment, the model output functions as a decision-support urgency score, not a formal ESI assignment**. Free clinic intake volunteers do not perform ESI triage, and the model is not meant to replicate the full ESI protocol. Instead, it uses ESI-labeled ED data as a proxy for clinical urgency to drive queue prioritization in free clinic settings.

From 425,087 total records, we filtered to 395,464 records containing all essential vital signs. We then applied stratified random sampling by urgency level to extract 10,000 training records while maintaining the exact distribution from the full dataset (Table 1). Stratified sampling ensures proportional representation of all urgency levels, particularly the rare level 1 cases ($n = 370$, 3.7%). We also used stratified 80/20 split ($n = 8{,}000$ training, $n = 2{,}000$ test) with stratification by urgency level to ensure proportional representation of critical cases in both sets.

Table 1: Urgency level distribution in the stratified sample, matching the full MIMIC-IV-ED dataset. Levels correspond to ESI 1–5 (Gilboy et al., 2020; Xie et al., 2022).

Level	Description	Percentage	Count
1	Critical	3.7%	370
2	High	33.6%	3,360
3	Moderate	55.5%	5,550
4	Low	7.0%	700
5	Non-urgent	0.3%	20

This sample size balances statistical power with computational efficiency for deployment in resource-constrained settings: the resulting RF model requires only ∼50MB storage and achieves <100ms inference time on a standard CPU, enabling deployment via free-tier cloud infrastructure. Sample size planning guidelines suggest 5-25 samples per class for training and 75-100 samples for validation (Beleites et al., 2013). Recent empirical analysis of RF classifiers on clinical data found median sample sizes of 3,404 (range: 250-140,499) are needed to reach AUC stability, with more balanced classes reducing required sample sizes (Cho et al., 2024).

Our 10,000-record stratified sample balances statistical adequacy with deployment constraints: it provides 370 Level 1 and 3,360 Level 2 critical cases, exceeding minimum thresholds while enabling a computationally efficient model. Critically, the 10K sample size is a deliberate design constraint since our deployment target is free clinics in the equity gap, where free-tier infrastructure imposes hard limits on model storage and inference time. Experiments with larger training sets produced models exceeding these size limits without appreciable benefits. The 10K sample satisfies RF stability thresholds for clin-

ical data (Cho et al., 2024) while remaining deployable on free infrastructure (Sec. 3.8). To verify that subsampling did not materially inflate performance, we evaluated the trained model on the remaining 385,087 held-out MIMIC-IV-ED encounters; results are reported in Sec. 4.

3.2. Feature Engineering

We engineered 20 clinical features from raw patient data, guided by three design imperatives: (1) **Deployability**: each feature must be collectable by untrained intake staff without specialized equipment; (2) **Clinical grounding**: features align with ESI assessment criteria (Gilboy et al., 2020), ensuring clinical validity; and (3) **Parsimony**: limiting to 20 features reduces overfitting risk and keeps the intake form practical for a walk-in clinic setting. Features are organized into direct measurements and derived indicators, and detailed justification for each feature is provided in Table 6 (Appendix C):

Direct features (11): `age`, `gender_encoded`, `symptom_severity` (1–10 clinician-assessed scale), `symptom_duration_hours`, `symptom_onset_encoded` (sudden vs. gradual), `heart_rate`, `systolic_bp`, `diastolic_bp`, `temperature`, `oxygen_saturation`, `previous_visits`.

Derived features (9): `has_red_flag` (binary indicator for chest pain, difficulty breathing, altered mental status), `has_chronic_condition`, `high_risk_chronic` (diabetes, heart disease, or cancer), `hr_abnormal` (HR <60 or >100), `bp_abnormal` (SBP <90 or >180), `temp_abnormal` (temp <96 or >100.4°F), `spo2_abnormal` (SpO_2 <95%), `vital_abnormalities` (count of abnormal vitals, 0–4), `symptom_acuity` (composite severity × duration score).

To address the distribution shift between training data (MIMIC-IV-ED with complete vital signs) and deployment (free clinics potentially lacking vital sign equipment), we evaluate the model under two conditions:

- **Full Vitals Mode:** Evaluation using actual MIMIC-IV-ED vital sign measurements, representing deployment in clinics with vital sign equipment.
- **Vitals-Imputed Mode:** All vital sign features replaced with physiologically normal defaults (heart rate 75 bpm, blood pressure 120/80 mmHg, temperature 98.6°F, SpO_2 100%) and derived vital-abnormality features set to zero, simulating free clinic deployment without vital sign monitors.

This dual evaluation approach directly validates the deployment scenario and avoids the methodological flaw of training on vital-sign-rich data while deploying with imputed defaults without quantifying the impact.

3.3. Triage Classification Model

We trained an RF classifier using scikit-learn 1.7.2 with the following hyperparameters (fixed a priori):

- `n_estimators`: 200 decision trees
- `max_depth`: 15 levels
- `class_weight`: 'balanced' to address class imbalance
- `random_state`: 42 for reproducibility
- `oob_score`: True for out-of-bag validation

The `class_weight='balanced'` parameter automatically adjusts weights inversely proportional to class frequencies, penalizing misclassification of rare critical cases (Levels 1–2) more heavily than common moderate cases. This design choice prioritizes patient safety over raw multi-class accuracy. Although RF natively supports probability outputs via `predict_proba()`, we use hard class predictions for two reasons. (1) `class_weight='balanced'` already encodes asymmetric misclassification costs by up-weighting rare critical cases during training, making this functionally equivalent to post-hoc threshold adjustment without requiring manual threshold tuning on a held-out calibration set. (2) the queue optimizer (Sec. 3.4) requires discrete urgency level inputs to compute priority scores. Full cost-sensitive thresholding using an explicit cost matrix is planned for the prospective evaluation, where real free clinic outcome data will enable empirically grounded cost estimation. The model achieved an Out-of-Bag (OOB) score of 76.6%, indicating good generalization.

3.4. Queue Optimization Algorithm

We developed a multi-objective weighted priority scoring system that computes dynamic priority scores for each patient in the queue:

$$\text{priority} = w_u \times f_u + w_w \times f_w + w_a \times f_a \quad (1)$$

where:

- $f_u = 6 - \text{urgency\_level}$ is the urgency factor (inverted so Level 1 = 5, Level 5 = 1)

- $f_w = \min(\text{wait\_minutes}/90, 1.0)$ is the normalized wait time factor
- $f_a = \max(0, (\text{age} - 60)/40)$ is the age-based vulnerability factor
- $w_u = 10.0$, $w_w = 0.15$, $w_a = 0.05$ are the respective weights

The urgency weight is designed for dominance ($w_u = 10.0$), ensuring critical patients are always prioritized. The fairness weight ($w_w = 0.15$) prevents extreme delays by gradually increasing priority for all waiting patients. The age weight ($w_a = 0.05$) provides a small additional factor for patients over 60 when other factors are equal.

3.5. Baseline Methods

To contextualize our results, we compared our balanced RF model against four baseline approaches on the same test set ($n = 2{,}000$):

- **Decision Tree:** Single tree classifier with default parameters
- **Logistic Regression:** Multinomial logistic regression with L2 regularization
- **Rule-Based Triage:** Deterministic rules based on vital sign thresholds and red flag symptoms
- **ESI-Inspired:** Simplified implementation of ESI logic using available features

We selected these baselines because: (1) Decision Tree represents simple ML, (2) Logistic Regression is a standard statistical approach, (3) Rule-Based represents current practice, and (4) ESI-Inspired represents the current state-of-the-art triage application.

3.6. Evaluation Metrics

We evaluated the triage model using metrics aligned with clinical priorities:

- **Overall Accuracy:** Correct predictions across all 5 urgency levels
- **Critical Detection Rate (Sensitivity):** TP/(TP+FN)[3] for binary critical (Levels 1–2) classification, our primary safety metric
- **Dangerous Misses:** Critical patients (Levels 1-2) misclassified as non-urgent (Levels 4–5). We consider this the worst possible triage error
- **Critical Accuracy:** Binary accuracy for critical/non-critical classification
- **Weighted F1:** F1 score weighted by class frequency
- **Out-of-Bag Score:** RF's internal validation

3. TP: True Positive; FN: False Negative.

We do not report AUROC because, one, our model outputs discrete urgency levels (1-5) rather than continuous risk scores, and two, AUROC treats all mis-rankings equally, whereas our use case has asymmetric costs: missing a critical patient (Levels 1-2) is catastrophically worse than mis-ranking moderate cases (Levels 3-4).

3.7. Queue Optimization Validation

We validated queue optimization benefits through Monte Carlo simulation with 1,000 synthetic clinic sessions with 2 physicians (4 hours each), and a variable consultation time with 18 min average:

- Generated patient arrivals using a Poisson process with $\lambda = 0.12$ patients/minute ($\sim$7 patients/hour), calibrated from NAFC data (National Association of Free & Charitable Clinics, 2023): 5.8M visits/year across 1,400 clinics yields $\sim$4,143 visits/clinic/year. Assuming 150 sessions/clinic/year (3 sessions/week $\times$ 50 weeks) implies $\sim$27.6 patients per 4-hour session ($\lambda = 27.6/240 \approx 0.12$ patients/minute), which is also consistent with the 25 patients/session from Min et al. (2024).
- Sampled urgency levels from observed MIMIC-IV distributions
- Simulated both FCFS and our multi-objective queue orderings for identical patient sequences
- Computed average, median, and critical patient wait times
- To stress-test robustness under realistic conditions, we additionally simulated a 2-state Markov-Modulated Poisson Process (MMPP) with batch arrivals across 13,000 sessions; results are reported in Sec. 4.

Statistical significance was assessed using paired t-tests across simulation iterations.

3.8. Implementation

The complete system comprises:

- **Clinic Staff Web App:** React 19 with Tailwind CSS 3.4, tablet-optimized with 44px touch targets
- **ML API:** Flask microservice deployed on Google Cloud Run (free tier)
- **Database:** Compliant Firebase Firestore for real-time queue synchronization
- **Demonstration Web App:** Streamlit application with model scoring visibility and Monte Carlo simulation visualization

- **Development Tools:** VS Code with Python virtual environment, Git/GitHub and npm (Node Package Manager)

All components use exclusively free and free-tier services, majorly reducing financial burden.

4. Results

4.1. Triage Classification Performance

We evaluated the final model under Full Vitals (actual MIMIC-IV-ED measurements) and Vitals-Imputed (simulating free clinic deployment without vital sign monitors) modes. Table 2 presents the comparison.

Table 2: Model performance: Full Vitals vs. Vitals-Imputed on test set ($n = 2{,}000$).

Metric	Full	Imputed	Δ
Overall Acc.	75.7%	77.0%	+1.3%
Critical Det.	76.0%	83.6%	+7.7%
Critical Acc.	80.0%	81.1%	+1.1%
Weighted F1	75.5%	76.6%	+1.0%
Dangerous Miss	5	0	-5
OOB Score	76.6%	76.6%	–

Counter-intuitively, the Vitals-Imputed evaluation outperforms Full Vitals across all metrics. Critical Detection improves from 76.0% to 83.6% (+7.7 percentage points), and Dangerous Misses decrease from 5 to 0 (95% CI: 0.00%–0.02%) on the test set. Confusion matrices for both modes are shown in Figure 2. This finding is analyzed in detail in Section 5.1.

We conducted two additional validations to assess the robustness of dangerous misses. We evaluated the Vitals-Imputed model on the full 385,087-record MIMIC-IV-ED dataset (excluding the 10,000 training records), comprising approximately 143,600 critical patients (Levels 1-2). The model produced a dangerous miss rate of 0.012% (17 misses; 95% CI: 0.007%–0.019%). This compares significantly favorably vs. all baseline methods. Additionally, to assess generalization beyond a single train/test split, we evaluated performance stability across 10 stratified splits (80/20, random seeds 0-9) of the full 10,000-record sample. Results were consistent: 0 dangerous misses across all 10 splits (0.0 ± 0.0), critical detection 83.5%±1.5% (CV[4] = 1.75%, Good), and overall accuracy 77.6%±0.6% (CV = 0.74%, Excellent). The low coefficients of variation confirm that results are not an artifact of a single favorable random split.

To understand model calibration, we constructed reliability diagrams for binary critical/non-critical classification. The uncalibrated model exhibits mild underconfidence (ECE = 0.122, Brier = 0.122); Figure 1. This conservative bias is acceptable given our focus on high critical detection rate and low dangerous misses. After Platt scaling on a held-out calibration set, calibration improved (ECE = 0.044, Brier = 0.103); Figure 3. The distribution of predicted probabilities before and after calibration (Figure 4) demonstrates that well-calibrated probability estimates can be obtained from our model for clinical decision support.

While prospective validation remains essential, these results support deployment as a decision-support tool under physician oversight.

4.2. Baseline Comparison

Table 3 compares our model against baseline methods. We report our model's Vitals-Imputed results since this represents the intended free clinic deployment scenario.

In clinical triage, a dangerous miss, i.e., classifying a critical patient as non-urgent, carries substantially higher morbidity/mortality risk than false positives (Gilboy et al., 2020). Our model offers the best outcome compared to all baselines (38–105 dangerous misses). The rule-based and ESI-inspired approaches perform poorly, confirming the limitations of traditional triage methods documented in prior literature (Brown et al., 2023; Sax et al., 2024). A visual comparison of dangerous misses across methods is provided in Figure 5.

4.3. Queue Optimization Results

Monte Carlo simulation across 1,000 clinic sessions show significant improvements (Table 4).

The 72% reduction in critical patient wait times, from 24.4 (95% CI: 23.4-25.5) to 6.8 (95% CI: 6.5-7.2) minutes ($p < 0.001$), represents a potential clinical benefit. Under simulated assumptions, patients with chest pain, diabetic emergencies, or difficulty breathing receive substantially faster attention compared to FCFS. The median wait time reduction of 37% is also statistically significant ($p < 0.001$), while the minimal change in overall average wait time is not significant ($p = 0.356$).

4. CV: Coefficient of Variation.

Table 3: Comparison against baseline methods on MIMIC-IV-ED test set (n = 2,000). Dangerous Misses counts critical patients (Levels 1–2) misclassified as non-urgent (Levels 4–5).

Method	Overall Acc.	Crit. Det.	Crit. Acc.	F1	Dangerous Misses
Decision Tree	78.5%	87.5%	83.6%	80.0%	38
Logistic Regression	72.0%	86.0%	84.2%	78.1%	104
Ours (Vitals-Imputed)	77.0%	**83.6%**	81.1%	76.6%	**0**
Rule-Based	29.1%	56.5%	52.5%	33.0%	105
ESI-Inspired	38.9%	48.9%	53.9%	41.2%	105

Table 4: Monte Carlo simulation results (1,000 iterations) comparing FCFS vs. our optimizer. Times in mins.

Metric	FCFS	Ours	Improve.
Overall Avg	24.7	24.6	-0.1 (-0.2%)
Median Wait	23.7	14.9	-8.8 (-37%)
Critical Wait	24.4	6.8	-17.6 (-72%)

Sensitivity analysis with λ ranging from 0.08 to 0.15 patients/minute (5-9 patients/hour) showed consistent relative improvement in critical wait times (68-75% reduction). To stress-test robustness under realistic bursty conditions, we extended the simulation using a 2-state Markov-Modulated Poisson Process (MMPP) with batch arrivals across 13,000 sessions. Table 5 and Figure 6 summarize results.

Under bursty conditions, FCFS critical wait time rose 125% (24.4→54.7 min); our optimizer rose only 64% (6.7→11.0 min), yielding an 80% reduction vs. 72% under Poisson arrivals. Additionally, across 25 priority weight configurations ($w_u \in \{5, 7.5, 10, 15, 20\}$, $w_w \in \{0.05, 0.10, 0.15, 0.30, 0.50\}$), critical wait reduction under bursty arrivals ranged 78.4%–80.6% (Figure 6), confirming results are not sensitive to specific weight choices. Staffing level variation, triage time bottleneck modeling, higher critical prevalence scenarios, and partial vitals ablation are planned for the prospective evaluation as the next step.

Table 5: Critical wait under Smooth Poisson vs. Bursty MMPP arrivals (13,000 sessions). Times in mins.

Condition	FCFS	Ours	Improve.
Smooth	24.4	6.7	72.7%
Bursty	54.7	11.1	**79.7%**
25-config wt. sensitivity (bursty)			78.4–80.6%

4.4. Demographic Subgroup Analysis

We examined model performance across demographic subgroups:

- Critical accuracy for age <40: 74%
- Critical accuracy for age 40–65: 75%
- Critical accuracy for age >65: 79%
- Critical accuracy variation by sex: <2%

The model showed minimal demographic bias, with slight improvement for elderly patients (79% vs. 74% for younger patients), possibly reflecting clearer symptom presentation in older populations. Sex subgroup sample sizes: n_{male} = 1,087, n_{female} = 913. Critical accuracy: 77.8% (male) vs. 77.6% (female), difference not significant ($\chi^2 = 0.02$, $p = 0.89$).

While we analyzed age and sex independently, future audits must apply an intersectional approach. Buolamwini and Gebru (2018) demonstrated that error rates can be disproportionately high for specific subgroups (e.g., darker-skinned females) even when aggregate metrics appear robust, highlighting the necessity of intersectional auditing in clinical tools.

Race and Ethnicity While MIMIC-IV-ED contains race/ethnicity data, we did not include these variables as model features or perform race-stratified performance analysis for a few methodological reasons. First, our training data is drawn from a single urban academic medical center (BIDMC, Boston), and racial/ethnic distributions in free clinic populations may differ substantially. Pre-adjusting for racial disparities observed in one setting risks introducing distribution shift bias when deployed in another. Second, best practices in fair ML recommend prospective bias auditing in the deployment setting rather than pre-emptive correction based on potentially mismatched training data (Obermeyer et al., 2019; Siddique et al., 2024). Race-stratified performance monitoring and an audit

protocol on prospective free clinic populations are important next steps, discussed in Sec. 5.3.

5. Discussion

5.1. Why Vitals-Imputed Mode Outperforms Full Vitals

The counter-intuitive finding that Vitals-Imputed mode outperforms Full Vitals warrants detailed analysis. Feature importance analysis (Figure 7) reveals the mechanism. These are the key findings from our analysis:

- **Vital signs are noisy predictors:** While 85.1% of Level 1 (critical) patients have abnormal vitals, so do 60.6% of Level 3 (moderate) patients (Figure 8). This 24.5%-pt difference seems insufficient separation for reliable classification. A 60.6% abnormal vital rate among moderate patients means abnormal vitals are common in non-critical cases, reducing discriminative value. This problem is compounded in free clinic settings where vitals equipment may be absent or operated by untrained intake staff, further widening the noise.
- **Critical patients often have normal-looking vitals:** 33.6% of critical patients have zero abnormal vital signs. When the model sees 'normal' vitals in Full Vitals mode, it may incorrectly infer lower urgency, causing the 5 dangerous misses observed.
- **Imputation removes misleading signal:** By setting all vitals to 'normal' defaults, we remove the model's ability to be 'fooled' by normal-appearing vitals in actually-critical patients. The model then relies on symptom-based features (symptom_acuity, severity, red flags), which prove more predictive of true urgency.
- **Safer failure mode:** When predictions shift between modes, 10.9% shift toward MORE urgent (safer) while only 6.9% shift toward LESS urgent in the Vitals-Imputed mode. The net effect is a more conservative posture that significantly reduces dangerous misses.

This finding has important practical implications: for free clinic deployment, the absence of vital sign equipment is not a limitation. Symptom-based triage (severity, acuity, red flags) is what untrained volunteers can reliably collect, and our results validate this approach. Randomized device-noise simulation (Gaussian noise calibrated to low-cost device specifications) is a logical future experiment discussed in Sec. 5.4.

5.2. Comparison to Prior Work

Our critical detection rate (83.6%) is consistent with prior ML triage literature. Raita et al. (2019) reported AUROC of 0.86 for critical care prediction, and Xie et al. (2022) achieved similar or better performance with RF on MIMIC-IV-ED benchmarks. Notably, our model substantially outperforms traditional triage systems: Brown et al. (2023) found ESI achieves only 65.9% sensitivity for high-acuity patients across 5 million encounters. Our model's 83.6% critical detection represents an 18%-pt. improvement over this.

5.3. Practical Deployment Considerations

Our solution is designed for potential adoption in resource-constrained settings:

- **Cost:** All infrastructure uses free and free-tier software infrastructure and services
- **Hardware:** The tablet-optimized interface runs on any cheap device with a web browser
- **Training:** Volunteer intake staff need minimal training
- **Override:** The system design maintains human oversight through a manual queue override capability, ensuring physicians retain ultimate decision authority
- **Connectivity:** Patient intake requires brief internet access (∼1KB per triage prediction) to call the cloud-hosted ML API, ensuring model version control. Once urgency is classified, offline persistence enables all queue management without connectivity. This hybrid architecture requires minimal bandwidth, achievable even via a weak cellular hotspot

The hybrid human-AI design maintains human oversight while providing systematic, reproducible triage support. Figure 9 and Figure 10 illustrate how the deployed system integrates into free clinic operations without requiring clinical training from intake staff, how the priority algorithm escalates urgent cases in practice, and physician override capability.

To ensure transparency for volunteer staff, we propose accompanying any deployment with a **Model Card** (Mitchell et al., 2019). This document would explicitly detail the model's intended use, performance limitations across demographic subgroups, and training data composition to prevent misuse in out-of-scope contexts.

Triage racial disparities are well-established, including an 18.5% higher under-triage risk for Black patients documented by Brown et al. (2023), making prospective bias auditing a clinical and ethical imperative. Deployment must also comply with the 2024 **HHS**[5] **Section 1557 Final Rule** (HHS Office for Civil Rights, 2024), which mandates that covered entities identify and mitigate discrimination in patient care decision support tools, including ensuring the tool does not rely on input variables that act as proxies for protected characteristics.

We deferred race/ethnicity analysis in this work due to a population mismatch. Our training data derives from a single urban medical center (BIDMC, Boston), and applying race-based corrections derived from that population risks false assurance and distribution shift bias when deployed in heterogeneous free clinic settings (Obermeyer et al., 2019). A structured prospective audit protocol must comprise: (1) a pre-deployment proxy discrimination audit of all input features for protected characteristic correlation, and document them in the Model Card (Mitchell et al., 2019); (2) prospective collection of de-identified race, ethnicity, and preferred language with stratified critical detection and dangerous miss rates; (3) automated retraining trigger if any inter-group performance gap exceeds 5 percentage points; (4) mandatory physician review of all Level 3 cases during prospective evaluation; and (5) intersectional audit across race, sex, age, and language subgroups (Buolamwini and Gebru, 2018).

Preliminary qualitative feedback from free clinic staff and physicians has been enthusiastic and informative. Clinic staff expressed enthusiasm for a pilot and noted the potential of reliable, systematic triage compared to current ad hoc practice. One physician indicated willingness to advocate for the system among their peer network, supporting organic peer dissemination as a viable deployment pathway. Another physician described how free clinic disorganization makes it difficult to attend to patients effectively and noted the system could be a significant benefit for volunteer physicians as well as patients. A prospective evaluation in partnership with domain experts is an important next step to evaluate effectiveness in real free-clinic workflows.

5. HHS: Department of Health and Human Services.

5.4. Limitations

Several limitations warrant consideration:

- **Training-Deployment Distribution Shift:** The distribution shift between ED training data and free clinic deployment populations represents the primary unresolved risk of this work, as symptom presentations and patient demographics may differ between populations. Our Vitals-Imputed evaluation directly simulates the free clinic scenario, providing more realistic performance estimates than Full Vitals evaluation alone. Still, prospective validation is essential to quantify its impact.
- **Single Institution:** MIMIC-IV-ED is drawn from BIDMC (urban, academic medical center). Free clinic populations may have different chronic disease profiles, medication access, and symptom presentation patterns. Future work should be validated on heterogeneous free clinic populations.
- **Retrospective Validation:** We use retrospective data and simulation. While large-scale holdout evaluation and 10-fold cross-validation stability strengthen confidence in the results, prospective clinical trials are needed to confirm our analysis.
- **Vital Signs:** Future work exploring a more discriminative vital sign feature construction could improve specificity by unlocking the signal carried by raw vitals. Additionally, a partial-vitals scenario is a logical future experiment using randomized device-noise simulation (Gaussian noise).
- **Multilingual Interface:** Language is particularly salient for free clinics serving non-English-speaking populations since symptom reporting differences across languages may affect how chief complaint features are elicited and recorded, introducing a potential source of representation bias. This makes multilingual interface support and language-stratified performance auditing important next steps.

5.5. Ethical Considerations

AI-assisted triage raises important ethical considerations. This is why our system is designed as a decision-support tool to improve recognition and queuing of critical cases. It's not a replacement for a physician's clinical judgment. With our approach, the clinic physician can manually adjust queue ordering when they observe factors the model cannot capture, hence maintaining hu-

man accountability. To mitigate **automation bias**, where clinicians may over-rely on algorithmic output, our interface requires active physician confirmation for queue reordering. Also, Hanna et al. (2025) emphasize that maintaining human oversight is critical to preserving patient autonomy and preventing 'interaction bias' in high-stakes clinical settings.

While our demographic analysis showed minimal bias, deployment in different populations requires ongoing monitoring for fairness. We acknowledge the risk of **label choice bias**, as demonstrated by Obermeyer et al. (2019), who found that algorithms trained on healthcare costs or utilization as a proxy for health needs can systematically under-triage Black patients, who historically face barriers to access despite having higher disease burdens.

We must also guard against **proxy discrimination**, where a model discriminates using variables correlated with protected traits even if demographic data is excluded (La Cava et al., 2025). Ongoing monitoring of outcome disparities is essential to detect this form of bias. Our deployment protocol addresses this through the prospective audit framework detailed in Sec. 5.3. We intentionally defer bias mitigation interventions until deployment data is available, as pre-correction based on MIMIC-IV-ED's Boston population could introduce distribution shift bias in the diverse free clinic settings we aim to serve.

Finally, transparency is maintained as staff can see model output and the classification rationale (urgency level, contributing factors, etc.), and if needed, physicians retain manual override capability. Any deployment should include regular audits comparing model and queue recommendations to clinical outcomes.

6. Conclusion

We presented our ML-powered triage and queue optimization **decision support system** designed for resource-constrained free clinics and evaluated it on retrospective MIMIC-IV-ED data with additional simulation-based analyses. Our RF classifier achieves 83.6% critical detection in Vitals-Imputed mode with zero dangerous misses observed in test set evaluation ($n = 2{,}000$), the only evaluated method to achieve this outcome, prioritizing patient safety over raw accuracy. Monte Carlo simulation demonstrates a potential 72% reduction in critical patient wait times (24.4 to 6.8 minutes) compared to FCFS ordering, with all improvements statistically significant ($p < 0.001$). Our work indicates that this approach may offer meaningful improvements if confirmed in prospective free-clinic studies.

A key finding is that the Vitals-Imputed mode, which simulates free clinic deployment without vital sign equipment, actually outperforms the Full Vitals mode. Analysis reveals that vital signs are noisy predictors in this dataset, and symptom-based features (severity, acuity, red flags) provide more reliable urgency signals. This validates the practical deployment scenario where volunteer-staffed clinics lack medical equipment.

The system addresses a genuine healthcare access gap: free clinics lack resources for commercial triage systems, yet their patient populations would particularly benefit from systematic, unbiased urgency assessment. By developing a deployable system on free and free-tier infrastructure, designed for untrained volunteer-staffed environments and built with physician overrides, our work illustrates that thoughtful application of established ML methods can help address underserved populations once clinical validation is completed.

Future work includes IRB-approved prospective clinical validation with proper audit protocols, multi-site validation, and expansion to multilingual interfaces.

Acknowledgments

This work is dedicated to all volunteers serving underserved populations around the globe. The author extends sincere thanks to the physicians, staff, and volunteers that provided their inputs and is grateful to the research teams at MIT and BIDMC for developing and maintaining the MIMIC-IV-ED dataset.

The author expresses gratitude to Kailen Lee, Tiffany Batiste-Gilmore, Chuka Emezue, Ph.D., Manraj Gill, Ph.D., and David Thoele, MD for their mentorship and guidance. The author also wishes to thank his family for their constant support, enabling him to dedicate his time to his passion, and for believing in this work. Claude (Anthropic) was used as an assistant; the author independently verified all content, claims, experimental results, figures, and references, and bears full responsibility for accuracy and originality.

References

Claudia Beleites, Ute Neugebauer, Thomas Bocklitz, Christoph Krafft, and Jürgen Popp. Sample size planning for classification models. *Analytica Chimica Acta*, 760:25–33, 2013.

Ruha Benjamin. Assessing risk, automating racism. *Science*, 366(6464):421–422, 2019.

Sarah R Brown, Diana Martinez Garcia, David J Nauman, Sriram Ramgopal, James M Chamberlain, Elizabeth R Alpern, and Paul L Aronson. Evaluation of version 4 of the emergency severity index in US emergency departments. *JAMA Network Open*, 6(2):e2255089, 2023.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81:1–15, 2018.

Nicole Chmielewski and Jill Moretz. ESI triage distribution in U.S. emergency departments. *Advanced Emergency Nursing Journal*, 44(1): 58–67, 2022.

Hyewon Cho, Jianhui She, Dario De Marchi, Hanin El-Zaatari, Edward L Barnes, Anna R Kahkoska, Michael R Kosorok, and Anjali V Virkud. Sample size requirements for popular classification algorithms in tabular clinical data: Empirical study. *JMIR AI*, 3:e62354, 2024.

Alec H Davidian and Garrett M Bloom. Improving patient care: Expansion of access to free clinics. *Patient Experience Journal*, 11(2):115–120, 2024.

Amir Elalouf and Gad Wachtel. Queueing problems in emergency departments: A review of practical approaches and research methodologies. *SN Operations Research Forum*, 3(1):2, 2022.

Nicki Gilboy, Paula Tanabe, Debbie Travers, A Michael Rosenau, and David R Eitel. Emergency severity index (ESI): A triage tool for emergency department care, version 4. Technical report, Emergency Nurses Association, 2020.

M G Hanna et al. Ethical and bias considerations in artificial intelligence/machine learning. *Modern Pathology*, 38(3):100686, 2025.

HHS Office for Civil Rights. Nondiscrimination in health programs and activities (section 1557 final rule), 2024.

Weng Seong Hong, Adrian D Haimovich, and R Andrew Taylor. Predicting hospital admission at emergency department triage using machine learning. *PLoS ONE*, 13(7):e0201016, 2018.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

William G La Cava, I Glenn Cohen, and Jaya Aysola. The future of algorithmic nondiscrimination compliance in the affordable care act. *npj Digital Medicine*, 9(1):49, December 2025.

Seunghyun Lee and Young Hoon Lee. Improving emergency department efficiency by patient scheduling using deep reinforcement learning. *Healthcare*, 8(2):77, 2020.

María Belén Martos-Cabrera et al. Machine learning methods applied to triage in emergency services: A systematic review. *International Emergency Nursing*, 60:101109, 2022.

Jennifer Miles et al. Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: A systematic review. *Diagnostic and Prognostic Research*, 4:16, 2020.

Emily Min et al. Factors that affect patient wait times at a free clinic. *Journal of Health Care for the Poor and Underserved*, 35(1):285–298, 2024.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, pages 220–229, 2019.

National Association of Free & Charitable Clinics. Free and charitable clinic data report. https://nafcclinics.org/, 2023.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Yoshihiko Raita, Tadahiro Goto, Mohammad Kamal Faridi, David FM Brown, Carlos A Camargo Jr, and Kohei Hasegawa. Emergency department triage prediction of clinical outcomes using machine learning models. *Critical Care*, 23:64, 2019.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1:18, 2018.

Waleed Rashwan et al. Optimization of service process in emergency department using discrete event simulation and machine learning algorithm. *BMC Health Services Research*, 22: 706, 2022.

Daniel R Sax et al. Emergency severity index version 4 and triage of pediatric emergency department patients. *JAMA Pediatrics*, 178(10): 969–978, 2024.

Shria M Siddique et al. The impact of health care algorithms on racial and ethnic disparities: A systematic review. *Annals of Internal Medicine*, 177(4):484–496, 2024.

Phichayut Sitthiprawat et al. Development and internal validation of an AI-based emergency triage model for predicting critical outcomes in emergency department. *Scientific Reports*, 15:17180, 2025.

Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of EAAMO '21*, pages 1–9, 2021.

Debbie A Travers et al. Reliability and validity of the emergency severity index for pediatric triage. *Academic Emergency Medicine*, 16(9): 843–849, 2009.

Feng Xie et al. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9:658, 2022.

Soowon Yi et al. The effects of applying artificial intelligence to triage in the emergency department: A systematic review of prospective studies. *Journal of Nursing Scholarship*, 56(3): 380–392, 2024.

Appendix A. Figures

This appendix contains all figures supporting the main text analysis.

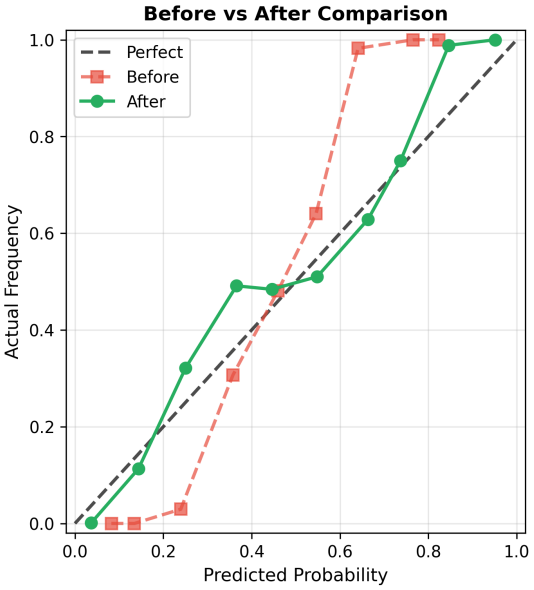

Figure 1: Reliability diagram comparing calibration before/after Platt scaling. Dashed line = perfect calibration. Uncalibrated (red) shows underconfidence; calibrated (green) aligns better with diagonal.

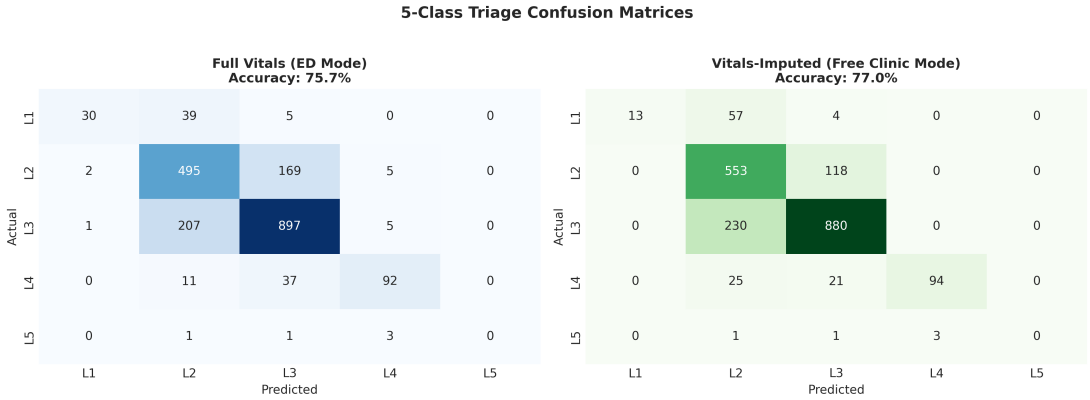

Figure 2: Confusion matrices comparing Full Vitals mode (left) vs. Vitals-Imputed mode (right) on test set (n = 2,000). Vitals-Imputed shows improved critical case detection with fewer dangerous misses.

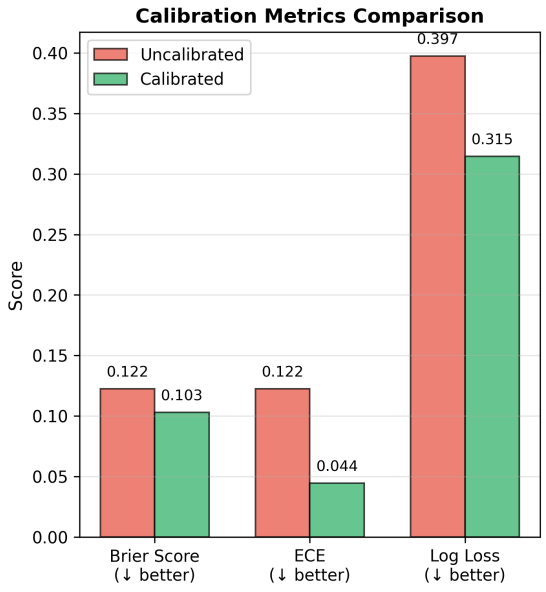

Figure 3: Calibration metrics before and after Platt scaling. All three metrics improved, with ECE showing the largest reduction (64%).

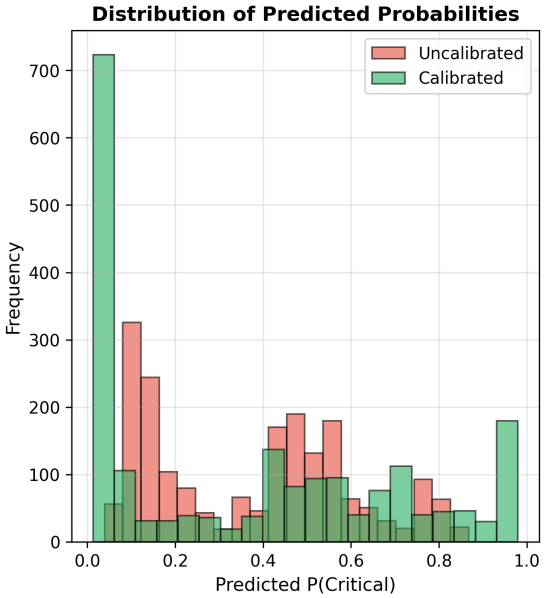

Figure 4: Distribution of predicted critical probabilities before/after Platt scaling. Calibration produces more decisive predictions at the extremes.

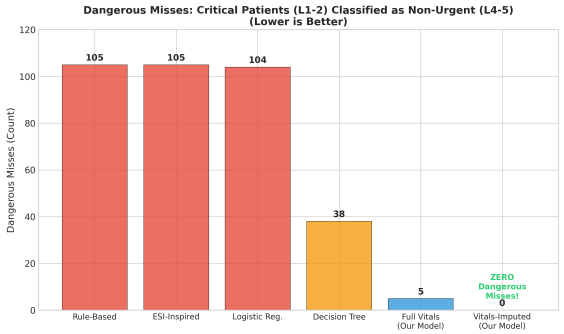

Figure 5: Dangerous Misses comparison across baseline methods. Our Vitals-Imputed model achieves zero dangerous misses, the only method to do so.

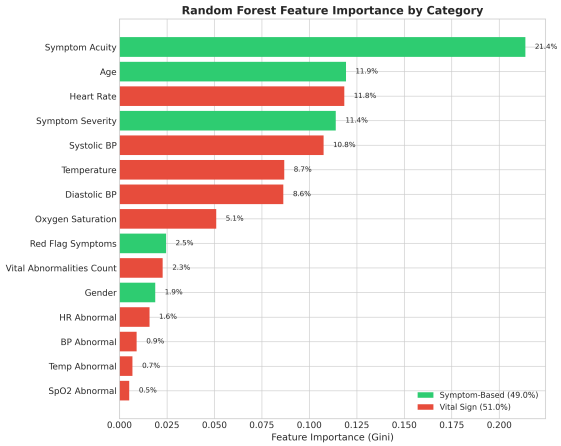

Figure 7: Feature importance from RF model. Symptom-based features dominate; vital signs show lower importance, explaining Vitals-Imputed success.

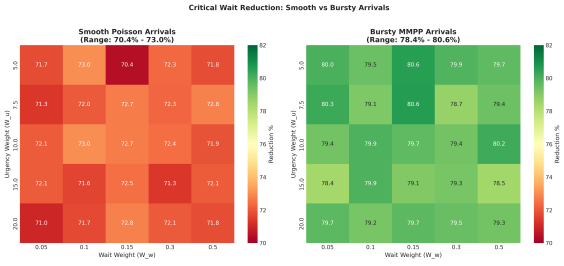

Figure 6: Critical wait reduction (%) across 25 weight configurations under Smooth Poisson (left) and Bursty MMPP (right) arrivals. Bursty yields higher reductions (78–81%).

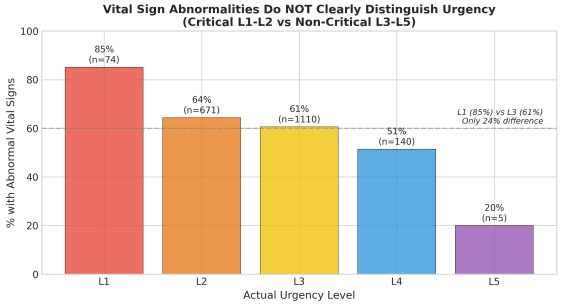

Figure 8: Vital sign abnormalities across urgency levels. 85.1% of Level 1 patients have abnormal vitals, but so do 60.6% of Level 3, showing insufficient separation.

Figure 9: Patient-flow diagram showing the end-to-end workflow for a single clinic session using ClinicTriage. Volunteer roles (intake staff vs. physician) are indicated below each step. Vital signs are optional (Vitals-Imputed mode used if absent).

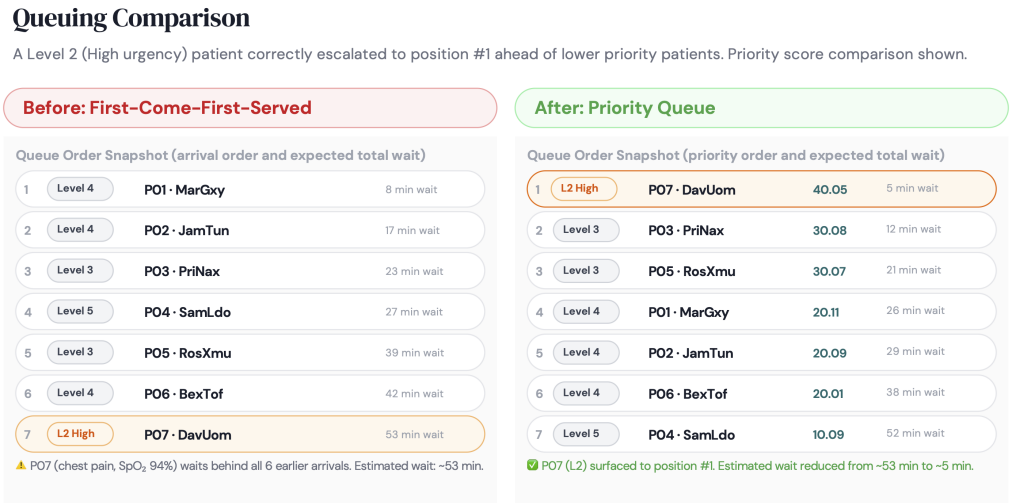

Figure 10: Queuing comparison showing Patient P07 (Level 2, chest pain, SpO_2 94%) correctly escalated to queue position #1 ahead of 6 lower priority patients.

Appendix B. Additional Visualizations

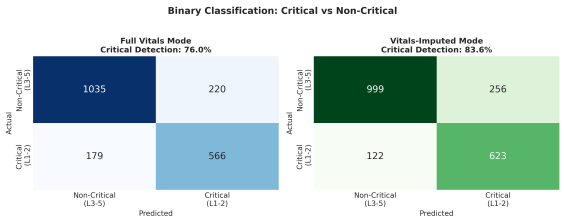

Figure 11: Binary classification: Critical vs. Non-Critical confusion matrices for both modes.

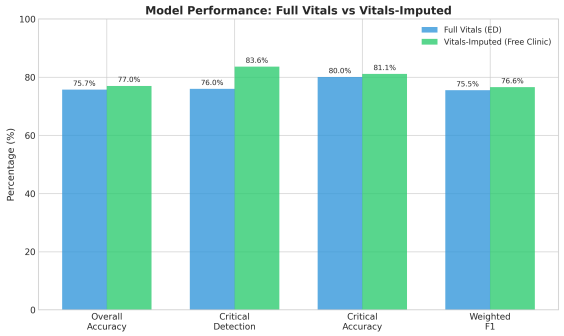

Figure 12: Model performance comparison: Full Vitals vs. Vitals-Imputed across all metrics.

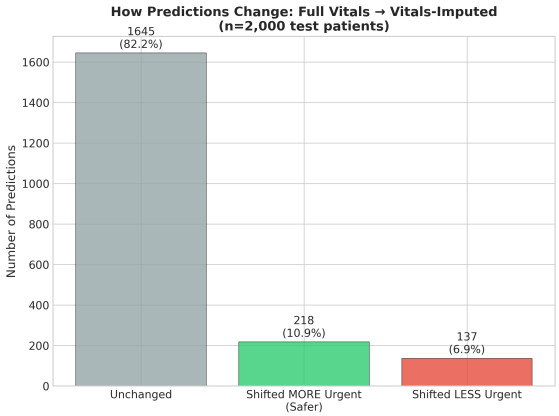

Figure 13: Prediction shift analysis: how classifications change between Full Vitals and Vitals-Imputed modes.

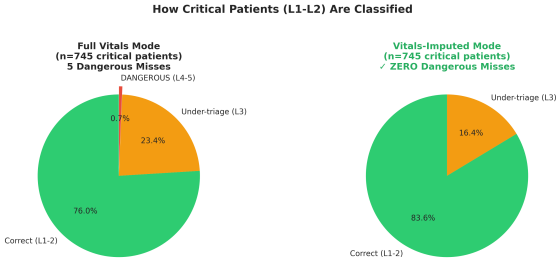

Figure 14: Critical patient classification breakdown showing correct, under-triage, and dangerous miss rates.

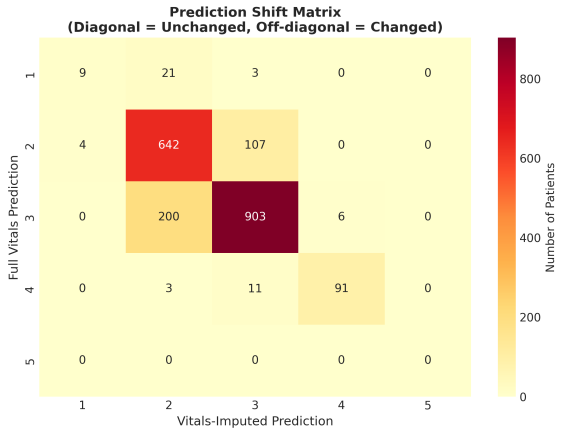

Figure 15: Prediction shift matrix: diagonal = unchanged, off-diagonal = changed between modes.

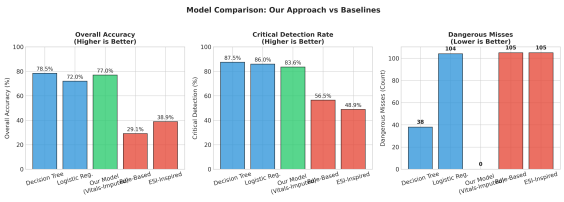

Figure 16: Model comparison vs. baselines: accuracy, critical detection, and dangerous misses.

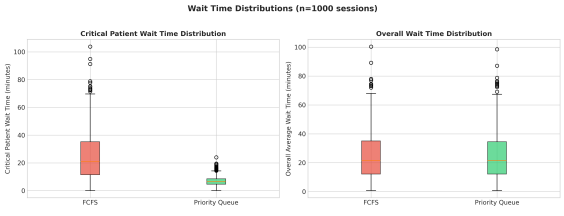

Figure 17: Wait time distributions from Monte Carlo simulation (1,000 sessions): FCFS vs. Priority Queue.

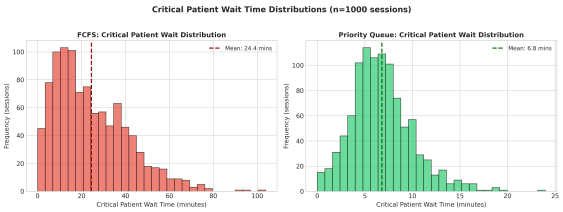

Figure 18: Critical patient wait time distributions: FCFS (left) vs. Priority Queue (right).

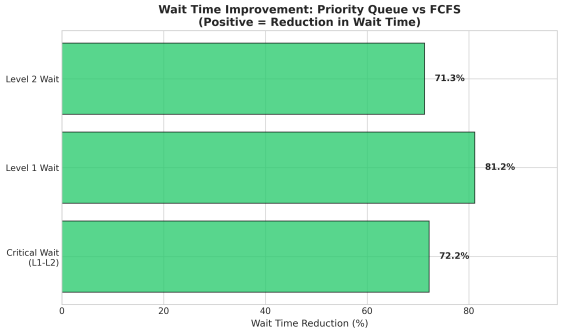

Figure 19: Wait time improvement by urgency level: Priority Queue vs. FCFS (positive = reduction).

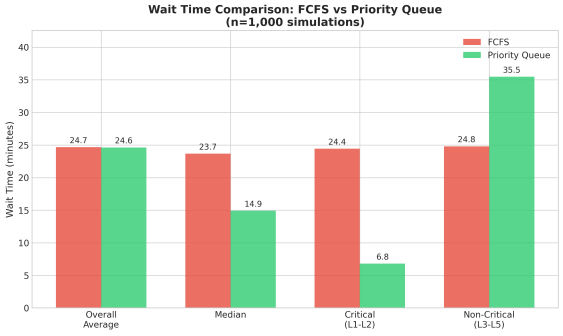

Figure 20: Wait time comparison: FCFS vs. Priority Queue across overall, median, critical, and non-critical metrics.

Appendix C. Feature Engineering

Table 6: Justification for the 20 clinical features used in the triage model. Vitals-Imputed defaults noted where applicable.

Feature	Type	Clinical / Operational Justification	Data Source
Direct Features (11)			
`age`	Direct	Age stratifies physiological reserve and risk of deterioration; elderly patients are expected to have higher acuity for identical presentations.	Patient self-report
`gender_encoded`	Direct	Gender-linked physiological differences affect symptom presentation (e.g., atypical MI in females).	Patient self-report
`symptom_severity`	Direct	Patient-reported severity (1–10 scale) is the most direct proxy for subjective distress and urgency perception; correlates with acuity across triage systems.	Patient self-report
`symptom_duration_hours`	Direct	Duration distinguishes acute emergencies from chronic complaints.	Patient self-report
`symptom_onset_encoded`	Direct	Sudden onset is a hallmark of life-threatening conditions (e.g., aortic dissection, subarachnoid hemorrhage); gradual onset may suggest chronic exacerbation.	Patient self-report
`heart_rate`	Direct	HR outside normal range can be a primary indicator of hemodynamic instability.	Pulse oximeter or manual palpation; Vitals-Imputed: 75 bpm
`systolic_bp`	Direct	SBP outside normal range may indicate hypotensive shock or markedly elevated blood pressure requiring clinical attention.	BP cuff; Vitals-Imputed: 120 mmHg
`diastolic_bp`	Direct	Standard vital sign providing discriminative signal alongside SBP for hypotensive and hypertensive risk assessment.	BP cuff; Vitals-Imputed: 80 mmHg
`temperature`	Direct	Fever may indicate infection/sepsis; hypothermia indicates severe systemic illness; both elevate acuity.	Thermometer; Vitals-Imputed: 98.6 °F
`oxygen_saturation`	Direct	SpO_2 below threshold may indicate hypoxemia requiring intervention.	Pulse oximeter; Vitals-Imputed: 100%
`previous_visits`	Direct	Provides context for disease trajectory and chronic care gaps; frequent visitors may represent undertreated chronic conditions or social determinants of health needs.	Patient self-report
Derived Features (9)			
`has_red_flag`	Derived	Chest pain, difficulty breathing, and altered mental status are canonical high-acuity presenting complaints per ESI (Gilboy et al., 2020); any one in isolation warrants physician attention; binary.	Structured intake checklist; no equipment needed
`has_chronic_condition`	Derived	Chronic conditions increase baseline risk and the likelihood that acute presentations represent decompensation; binary.	Patient self-report
`high_risk_chronic`	Derived	Diabetes, heart disease, and cancer represent the highest-risk comorbidity triad for acute decompensation and are potentially life-threatening; binary.	Patient self-report
`hr_abnormal`	Derived	Encodes the clinical decision boundary for HR: <60 bpm (bradycardia, potential heart block) or >100 bpm (tachycardia, potential sepsis/hemorrhage/PE); binary.	Derived from `heart_rate`
`bp_abnormal`	Derived	Encodes critical BP boundaries: SBP <90 mmHg (hypotension/shock) or SBP >180 mmHg (markedly elevated BP threshold); binary.	Derived from `systolic_bp`
`temp_abnormal`	Derived	Encodes clinical fever (>100.4 °F, the standard febrile threshold) and a conservative hypothermia screening threshold (<96 °F); binary.	Derived from `temperature`
`spo2_abnormal`	Derived	SpO_2 <95% is a conservative triage screening threshold; binary.	Derived from `oxygen_saturation`
`vital_abnormalities`	Derived	Composite count of abnormal vitals (0–4) captures cumulative physiological burden; multiple simultaneous abnormalities are a stronger acuity signal, consistent with SIRS/qSOFA[a]multi-criteria logic; integer.	Composite from vital features
`symptom_acuity`	Derived	Novel engineered composite: product of `symptom_severity` and `symptom_duration_hours`, prioritizing high-severity and prolonged acute symptoms as a combined urgency signal (not externally validated).	`symptom_severity` × `symptom_duration_hours`

[a] **SIRS** (Systemic Inflammatory Response Syndrome): a clinical screening criterion requiring ≥2 of: fever (>38 °C) or hypothermia (<36 °C), tachycardia (>90 bpm), tachypnea (>20 breaths/min), and abnormal WBC count. **qSOFA** (quick Sequential Organ Failure Assessment): a bedside sepsis risk score awarding one point each for altered mentation, respiratory rate ≥22 breaths/min, and SBP ≤100 mmHg; a score ≥2 identifies high-risk patients. Both criteria use multi-vital composite logic.