SAFE-Gate: A Knowledge-Based Expert System for Emergency Triage Safety with Conservative Multi-Gate Architecture and Explainable Reasoning

Chatchai Tritham[a,1], Chakkrit Snae Namahoot[a,2,*]

[a] *Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok 65000, Thailand*

Abstract

Background: Emergency triage of acute dizziness/vertigo presents a critical diagnostic challenge: most cases are benign, but a small percentage represent life-threatening strokes. Ensemble averaging of AI models can create dangerous blind spots when one model correctly identifies a critical case but gets outvoted by others.

Objective: We developed a knowledge-based expert system with formal safety guarantees for emergency triage, prioritizing critical case detection over overall accuracy.

Methods: SAFE-Gate employs six parallel knowledge modules examining orthogonal safety dimensions: critical red flags (G1), cardiovascular risk (G2), data quality (G3), syndrome matching (G4), uncertainty quantification (G5), and temporal analysis (G6). Instead of averaging, conservative merging selects the most cautious assessment—any module signaling high risk takes precedence. We validated six mathematical safety properties across 6,400 synthetic vertigo cases generated through counterfactual reasoning.

Results: On 804 held-out cases, SAFE-Gate achieved 100% critical sen-

**Corresponding author: Chakkrit Snae Namahoot, Email: chakkrits@nu.ac.th

Email addresses: chatchait@nu.ac.th (Chatchai Tritham), chakkrits@nu.ac.th (Chakkrit Snae Namahoot)

[1] ORCID: 0000-0001-7899-228X

[2] ORCID: 0000-0003-4660-4590

sitivity (175/175 cases) with zero false negatives and perfect safe-discharge precision. Standard ensemble averaging catastrophically failed at 71.4% sensitivity, missing 50 critical cases. Conservative architecture produces deliberate over-triage (21.3% escalations) while maintaining real-time performance (<2ms per decision).

Conclusions: Conservative knowledge integration eliminates ensemble averaging's safety signal dilution through provably safe architecture. This approach provides a generalizable template for safety-critical medical expert systems where missing dangerous cases causes greater harm than unnecessary caution. Clinical deployment requires prospective validation on real patient data.

Keywords: Expert systems, Clinical decision support, Safety-critical systems, Emergency triage, Explainable AI, Conservative reasoning

1. Introduction

1.1. Expert Systems in Safety-Critical Medical Decision Making

Knowledge-based expert systems have been supporting clinical decisions since the 1970s, when MYCIN helped diagnose bacterial infections [1] and INTERNIST tackled general internal medicine problems [2]. What made these systems valuable was their transparency—physicians could see exactly which rules fired and why a particular recommendation emerged. Clinical guidelines could be encoded once and then applied consistently across thousands of cases [3]. But despite these advantages, expert systems never achieved the widespread adoption that many predicted.

Part of the problem stems from what researchers call the *knowledge ac-*

quisition bottleneck. Building a comprehensive expert system requires knowledge engineers to work closely with medical specialists for months or even years, painstakingly translating clinical expertise into formal logical rules [4]. Medical knowledge doesn't stand still either—as new trials complete and guidelines update, someone needs to maintain and revise these knowledge bases [5]. Perhaps more problematic, rule-based systems tend to be brittle. When a patient presents with an unusual combination of symptoms not anticipated by the rule authors, the system may fail entirely rather than degrade gracefully [6].

Machine learning seemed to offer a way around these limitations. Rather than hand-crafting rules, modern deep learning systems can extract patterns from massive electronic health record databases [7]. They adapt naturally to atypical cases and don't require explicit programming for every possible scenario. But this flexibility comes at a cost. The resulting models often function as "black boxes" that resist clinical interpretation [8], making it difficult for physicians to validate their reasoning or identify potential errors. Without formal safety guarantees, these systems can fail unpredictably on rare edge cases [9]. Privacy regulations also limit access to the large training datasets these approaches require [10]. Recently, researchers have begun exploring hybrid architectures that combine symbolic reasoning with neural learning [11, 12], but these systems remain largely in the laboratory.

Emergency triage for acute dizziness and vertigo illustrates why we need better approaches. These complaints account for 3–4% of emergency department visits [13], and they pose a genuine diagnostic puzzle. The vast majority of cases—roughly 40 out of every 41—turn out to be benign inner ear problems. But that one remaining case might be a posterior circulation stroke that could leave the patient severely disabled or dead if missed during the narrow window for intervention. The challenge is that benign and dangerous causes can look remarkably similar initially. Only 3–5% of patients with acute vestibular symptoms ultimately receive stroke diagnoses [14], yet about one in five posterior strokes presents without the classic warning signs that emergency physicians have been trained to recognize [15]. This creates an impossible dilemma: refer everyone for neuroimaging and you waste vast resources while overwhelming radiology departments; try to be selective and you risk missing strokes when patients are still within the thrombolysis window.

1.2. The Ensemble Averaging Paradox in Safety-Critical Systems

Over the past two decades, ensemble methods have emerged as one of machine learning's most successful ideas. The basic premise is elegant: combine predictions from multiple models, and their independent errors will tend to cancel out [16]. In practice, this usually means averaging predicted probabilities or taking majority votes across different classifiers. For optimizing overall accuracy on balanced datasets, this approach works remarkably well.

But there's a subtle problem when we deploy these techniques in safety-critical medical applications. Imagine a patient presenting with symptoms that could indicate either a benign condition or a life-threatening emergency. One of your ensemble members—perhaps a model that specializes in identifying high-risk features—correctly recognizes this as a critical case and outputs a high-risk prediction. However, four other models in your ensemble, looking at different feature perspectives, classify it as moderate risk. When you average these five predictions together, the mathematics inevitably pulls the result toward the majority view: moderate risk. The critical warning from that one cautious model gets diluted away. If that conservative model was actually right, you've just created a false negative on exactly the kind of case you most need to catch.

This problem gets worse as class imbalance increases. In emergency triage for dizziness, critical cases make up perhaps 5% of presentations. Standard statistical optimization naturally pushes predictions toward the 95

Some researchers have tried to address this through selective prediction or abstention [17]—having the system refuse to make predictions when it's uncertain. But this creates its own difficulties. How uncertain is too uncertain? Set the abstention threshold too cautiously and you end up deferring 90% or more of cases back to human clinicians, which defeats the entire purpose of automation. Set it too permissively and you're back to missing critical cases. Clinical pilots have struggled to find any middle ground. The confidence-based approach also breaks down when neural networks exhibit the overconfidence problem—assigning high probability scores to incorrect predictions on unusual cases [18].

1.3. Conservative Knowledge Integration: A Fail-Safe Architecture

What if we inverted the usual approach to combining multiple models? Instead of averaging their outputs to maximize overall accuracy, what if we merged them conservatively—always selecting the most cautious assessment? This shifts the design goal from statistical optimization to safety optimization.

That's the core idea behind SAFE-Gate. The system evaluates each patient through six independent knowledge modules running in parallel, with each module examining a different dimension of risk. The first module (G1) applies hard-coded rules straight from emergency medicine guidelines, checking for critical red flags like unstable vital signs or acute neurological deficits. The second (G2) accumulates cardiovascular risk factors—age, hypertension, diabetes, atrial fibrillation—in a weighted scoring approach similar to established clinical instruments. The third module (G3) serves a quality control function, monitoring whether we have enough clinical information to make a reliable automated assessment or whether we should defer to a human clinician. The fourth (G4) tries to match the symptom pattern against well-characterized benign syndromes using the TiTrATE diagnostic framework [14]—if a patient fits a classic benign profile, that provides reassurance. Module five (G5) takes a different approach, using a Bayesian neural network with Monte Carlo dropout to estimate how confident the system should be about its prediction. The final module (G6) considers the time course—how symptoms began and how they've evolved—because temporal patterns carry crucial diagnostic information for distinguishing strokes from peripheral vestibular problems.

These six assessments feed into a conservative merging step. We define a risk ordering: abstention (defer to human) is most conservative, followed by critical (immediate intervention), high-risk (urgent evaluation), moderate (same-day assessment), low-risk (monitoring), and minimal (safe discharge). When the gates disagree, the final output always selects the most conservative assessment. If five modules say "moderate risk" but one says "critical," the system outputs critical. If any module triggers abstention because of high uncertainty or missing data, that overrides everything else.

This design provides three advantages that traditional ensemble averaging cannot match. First, we can formally prove safety properties. Statements like "no critical case will ever receive a safe discharge recommendation" or "increasing clinical severity always moves the risk tier in a more conservative direction" become mathematical theorems that can be verified computationally. We've tested six such properties across 6,400 synthetic cases without finding a single violation. Second, the system produces clear explanations—the audit trail shows exactly which module enforced the final decision and why, giving clinicians something concrete to review and verify. Third, and most fundamentally, conservative merging guarantees that safety warnings propagate through to the final output. Unlike averaging, which can dilute or suppress cautious predictions, minimum selection ensures they always have a voice.

1.4. Contributions and Scope

This work offers four main contributions to the expert systems literature.

First, we introduce and formalize a conservative knowledge integration architecture based on lattice minimum selection rather than statistical averaging. This approach provably eliminates the failure mode where ensemble averaging dilutes safety warnings. While we demonstrate this in medical triage, the principle applies broadly to any safety-critical domain where missing a dangerous case causes more harm than raising false alarms.

Second, we show how to effectively combine symbolic rules with statistical learning in a hybrid architecture. Four of our six modules (G1, G3, G4, G6) use explicit rule-based logic that clinicians can directly inspect and validate. The other two (G2, G5) employ data-driven pattern recognition to handle the probabilistic aspects of risk assessment and uncertainty quantification. This combination provides both the interpretability of traditional expert systems and the adaptability of modern machine learning.

Third, we develop a formal safety property framework with computational verification. We specify six mathematical properties the system should satisfy—things like "no critical case receives a discharge recommendation" and "increasing clinical severity always produces more conservative risk tiers"—and then verify these properties computationally across our test dataset. This provides a practical middle ground between pure statistical validation and full formal proof.

Fourth, we empirically validate the ensemble averaging failure hypothesis through direct comparison. Testing the same components configured as a standard averaging ensemble versus our conservative architecture on 804 cases demonstrates the dramatic difference: 71.4% sensitivity with 50 missed critical cases for averaging, versus 100% sensitivity with zero misses for conservative merging.

Important scope limitations: This study establishes computational feasibility using entirely synthetic data. We generated these cases through counterfactual reasoning and non-negative matrix factorization to ensure realistic symptom patterns, which allowed rapid iteration and controlled testing without privacy concerns. However, synthetic data—no matter how carefully constructed—cannot fully capture the complexity, variability, and messiness of real emergency department presentations. Before any clinical deployment, this system requires multiple validation steps. Emergency physicians and neurologists need to review the encoded clinical logic to verify we've correctly interpreted guidelines and haven't introduced errors. The system needs retrospective testing on actual de-identified patient records to see how it performs on real cases. Prospective observational deployment should measure how often clinicians agree with or override the system's recommendations. Ultimately, randomized controlled trials comparing patient outcomes—stroke detection rates, time to thrombolysis, functional outcomes at 90 days, cost-effectiveness—would be needed to determine if the system actually improves care. The architecture is specifically designed for domains where conservative bias makes clinical sense, where missing a critical diagnosis causes far more harm than unnecessary escalation [19].

The remainder of this paper proceeds as follows: Section 2 reviews relevant work in medical expert systems, ensemble learning methods, and formal verification approaches; Section 3 describes the six-module architecture in detail, explains the conservative merging algorithm, and outlines our synthetic data generation process; Section 4 presents performance metrics on 804 held-out test cases with comparisons to baseline methods; Section 5 examines implications, limitations, and requirements for clinical validation; Section 6 summarizes our findings and charts the path toward prospective evaluation.

2. Related Work

2.1. Expert Systems for Medical Decision Support

Looking back at the history of medical expert systems reveals three fairly distinct waves of development. The earliest systems from the 1970s and early

1980s—MYCIN for identifying bacterial infections [1] and INTERNIST for broader internal medicine problems [2]—relied entirely on hand-coded rules combined with certainty factors. They worked remarkably well within their narrow domains but ran into two persistent problems. First, building them required enormous effort; MYCIN's 600 rules represented several person-years of painstaking work with domain experts [4]. Second, they tended to fail completely when encountering cases that didn't fit their programmed rules—the brittleness problem we mentioned earlier.

The second generation tried to address these limitations by incorporating probabilistic reasoning. Systems built around Bayesian networks [20] or fuzzy logic [21] could handle uncertainty and incomplete information more gracefully. Applications emerged for diagnosing heart disease [22], pneumonia [23], and cancer staging [24]. But while these systems dealt better with uncertainty, they didn't really solve the knowledge acquisition problem. Someone still needed to manually specify all those conditional probability tables or fuzzy membership functions.

More recently, a third generation has emerged that tries to get the best of both worlds by combining symbolic knowledge with machine learning. The idea is to automatically extract some knowledge from clinical databases while keeping the overall structure interpretable. Systems like DXplain [25], Isabel [26], and VisualDx [27] represent this hybrid approach. Yet despite these advances, emergency departments have been slow to adopt these tools. Clinicians worry about the "black box" aspects of the learned components [8], lack confidence in their safety for time-critical decisions [9], and struggle to integrate them into existing workflows [28].

Our work builds on this hybrid tradition but adds two novel elements. First, we formalize how to combine multiple knowledge sources conservatively rather than through averaging, eliminating a specific failure mode that plagues conventional ensembles. Second, we demonstrate how to specify and computationally verify safety properties in a way that could support regulatory review of safety-critical medical AI systems.

2.2. Ensemble Methods and Their Limitations

Ensemble learning has become one of the field's most reliable techniques for boosting predictive performance. The core idea is simple: combine multiple models and let their independent mistakes cancel out [16]. In practice, people typically use one of three combination strategies. Averaging computes a weighted sum or mean of predicted probabilities [29]. Voting takes the majority opinion across classifiers [30]. Stacking trains a meta-model to learn how to optimally combine the base predictions [31].

There's solid theory behind why ensembles work. When individual models make independent errors, combining them reduces variance and improves generalization [32, 33]. Practical successes abound: random forests eliminate much of the overfitting that plagues individual decision trees [34], gradient boosting iteratively fixes residual errors to achieve state-of-the-art performance [35], and deep ensembles improve the calibration of neural network predictions [36].

But here's the problem we keep coming back to. When you apply ensemble averaging in a safety-critical setting with severe class imbalance—say, critical cases making up less than 5

Researchers have tried various fixes. Selective prediction allows models to abstain when they're uncertain [17]. Epistemic uncertainty quantification tries to estimate when a model is operating outside its comfort zone [37]. But both approaches struggle with threshold selection. Set your abstention threshold too conservatively and the system refuses to make predictions for 90

2.3. *Clinical Triage Decision Support Systems*

Emergency departments have long relied on structured triage tools to help sort incoming patients by urgency. Traditional approaches use validated scoring systems that have been refined over years of clinical use. The Emergency Severity Index (ESI) [38] sorts patients into five urgency levels. The Modified Early Warning Score (MEWS) [39] tries to predict which patients might deteriorate during their hospital stay. The National Early Warning Score (NEWS2) [40] provides standardized criteria for identifying acutely ill patients. These rule-based tools offer excellent inter-rater reliability—different nurses using the same tool tend to reach the same conclusion. Their main limitation is inflexibility when patients present with complicated combinations of symptoms and comorbidities that don't quite fit the standard patterns.

Machine learning promised to overcome this limitation by learning patterns directly from data. Researchers have developed neural networks for predicting sepsis from electronic health records [41], gradient boosting models for stratifying acute coronary syndrome risk [42], and natural language processing systems for detecting strokes from clinical notes [43]. The discrimination metrics often look impressive in development. But real-world deployment has been sobering. One widely-implemented sepsis alert system generated false alarms 67

For vestibular disorders specifically, clinicians have several established tools. The HINTS protocol (Head Impulse, Nystagmus, Test of Skew) helps differentiate strokes from peripheral causes [14]. The ABCD^2 score stratifies TIA risk [44]. The TiTrATE framework formalizes diagnostic criteria along temporal and phenotypic dimensions [45]. When applied correctly by skilled examiners, these tools achieve excellent sensitivity—typically above 95

SAFE-Gate attempts to bridge these worlds. The symbolic modules (G1 and G4) encode established clinical rules and decision criteria, providing the reliability and transparency of traditional tools. The statistical modules (G2 and G5) learn probabilistic patterns from data, offering the adaptability of modern machine learning. Critically, the conservative merging ensures that hard safety constraints—like "unstable vital signs require immediate escalation"—can never be overridden by statistical predictions that might have learned to be too permissive. This addresses what we see as a crucial gap in existing hybrid systems.

2.4. Formal Verification for Safety-Critical AI

Industries with genuinely high-stakes systems—aviation, nuclear power, medical devices—don't just test their products and hope for the best. They use formal verification methods to mathematically prove that systems satisfy critical safety properties [46]. Model checking exhaustively explores all possible states a system could enter [47]. Theorem proving establishes logical correctness through rigorous proof [48]. Abstract interpretation analyzes program behavior without executing every possible path [49].

Researchers have recently begun adapting these techniques for machine learning. Reluplex can verify certain robustness properties of neural networks by solving satisfiability problems [50]. Certified robustness methods provide provable bounds on how much adversarial perturbation a network can tolerate [51]. Bayesian approaches attempt to quantify uncertainty in learned components [52]. However, most of this work focuses on continuous perturbations to image classifiers—how much can you perturb pixel values before the network changes its answer—rather than discrete clinical logic.

Meanwhile, regulators are starting to demand more rigor for medical AI. The FDA's Good Machine Learning Practice principles [53] call for healthcare-grade performance with documented safety margins, transparent decision-making that clinicians can understand, and demonstrated robustness when the data distribution shifts. The European Union's Medical Device Regulation [54] requires formal risk management and safety analysis throughout a product's lifecycle.

Our approach aims for a practical middle ground. Full formal verification of systems with complex machine learning components remains computationally prohibitive. But pure statistical validation—reporting accuracy on a test set—provides no guarantees about safety properties. We specify six safety properties in mathematical form (for example: "for all cases x, if x is critical, then the output must be R1, R2, or abstention—never a discharge recommendation"). Then we check these properties computationally across our entire dataset. We found zero violations across 6,400 synthetic cases. This doesn't constitute an exhaustive mathematical proof, but it provides much stronger evidence than standard test set metrics, potentially sufficient for regulatory submissions.

3. Methods

3.1. Expert System Architecture and Knowledge Representation

The SAFE-Gate system processes each patient through three computational stages. First, six knowledge modules run in parallel, each evaluating a different safety dimension independently. Second, a conservative merging step combines these assessments by selecting the most cautious one. Third, the system generates an audit trail explaining which module drove the final decision and why. The architecture comprises:

Stage 1 - Parallel Gate Evaluation: Six independent knowledge modules (G1–G6) simultaneously assess patient data from orthogonal safety perspectives: G1 employs rule-based logic for critical red flag detection, G2 performs cardiovascular risk scoring through weighted factor accumulation, G3 evaluates data completeness and quality, G4 matches symptom patterns against established benign clinical syndromes, G5 quantifies epistemic uncertainty via Bayesian neural networks with Monte Carlo dropout, and G6 analyzes temporal symptom evolution patterns.

Stage 2 - Conservative Merging: Rather than averaging gate outputs like traditional ensembles, the system selects the lattice minimum (most conservative assessment) across all six modules, ensuring that any single gate signaling critical risk or abstention overrides less conservative assessments from other modules.

Stage 3 - Audit Trail Generation: The system produces an explainable decision trace documenting which specific gate enforced the final risk tier, what clinical features triggered that assessment, and the underlying rationale—supporting physician review and quality assurance.

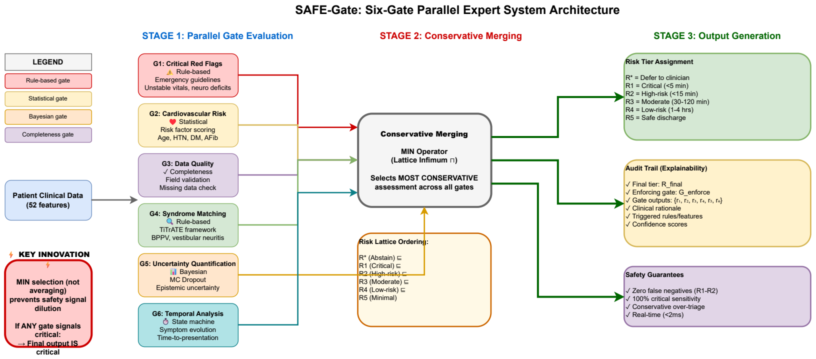


Figure 1: SAFE-Gate expert system architecture showing the three-stage pipeline. Patient data flows through six parallel knowledge modules (G1–G6), each evaluating orthogonal safety dimensions: G1 (critical red flag detection via rules), G2 (cardiovascular risk scoring), G3 (data quality assessment), G4 (clinical syndrome pattern matching), G5 (epistemic uncertainty quantification via Bayesian neural network), and G6 (temporal risk analysis). Conservative merging selects the lattice minimum across all gate outputs, ensuring that any module signaling critical risk or abstention overrides less conservative assessments. The system generates an explainable audit trail documenting which gate enforced the final tier and the clinical rationale.

Knowledge representation formalism: Formally, we're mapping from a patient state space $\mathcal{X}$ (containing all clinical features) to a risk tier space $\mathcal{R}$ with a specific ordering:

$$\mathcal{R} = \{R^*, R1, R2, R3, R4, R5\}, \quad R^* \sqsubseteq R1 \sqsubseteq R2 \sqsubseteq R3 \sqsubseteq R4 \sqsubseteq R5 \tag{1}$$

Here, R^* means abstention—the system recognizes its uncertainty exceeds safe bounds and defers to human judgment. $R1$ represents critical conditions requiring intervention within 5 minutes. $R2$ indicates high-risk situations needing urgent evaluation within 15 minutes. $R3$ signifies moderate risk warranting same-day assessment within 30 to 120 minutes. $R4$ suggests low-risk cases suitable for monitoring over 1 to 4 hours. $R5$ corresponds to minimal risk where safe discharge with outpatient follow-up is appropriate. The partial order $\sqsubseteq$ (read as "at least as conservative as") captures our key design principle: earlier tiers in this ordering represent greater caution.

Each gate implements knowledge module $g_i : \mathcal{X} \to \mathcal{R}$ mapping patient presentations to risk tiers based on specialized expertise domain. The six gates encode complementary knowledge:

$$\begin{aligned}
&g_1(x) : \text{Critical red flag detection (rule-based)}\\
&g_2(x) : \text{Cardiovascular risk assessment (statistical)}\\
&g_3(x) : \text{Data quality evaluation (completeness)}\\
&g_4(x) : \text{Clinical syndrome matching (rule-based)}\\
&g_5(x) : \text{Epistemic uncertainty quantification (Bayesian)}\\
&g_6(x) : \text{Temporal risk analysis (state machine)}
\end{aligned} \tag{2}$$

Conservative knowledge integration computes final assessment as lattice infimum (greatest lower bound):

$$r_{\text{final}}(x) = \inf_{\sqsubseteq}\{g_1(x), g_2(x), g_3(x), g_4(x), g_5(x), g_6(x)\} \tag{3}$$

Since our risk lattice is totally ordered (chain), this reduces to minimum selection in practice. Critically, any gate triggering R^* (abstention) overrides all other assessments by lattice precedence; any gate signaling $R1$ (critical) blocks lower-risk outputs.

3.2. Gate 1: Critical Red Flag Detection (Rule-Based Knowledge)

Gate G1 implements deterministic safety screening through 18 hard-coded rules encoding emergency medicine clinical guidelines [19]. Knowledge representation uses production rules: IF (condition) THEN (risk tier). Table 1 shows critical flag definitions.

The gate operates as a fail-safe module: detection of *any* red flag immediately outputs R1 regardless of other clinical features, implementing the safety principle "presence of life-threatening signs mandates immediate escalation." Absence of all 18 flags yields R5 (minimal risk from G1 perspective), allowing other gates' assessments to dominate through conservative merging.

This rule-based approach provides three advantages: (1) *Deterministic behavior*—no probabilistic thresholds or learned parameters, ensuring consistent application of clinical guidelines; (2) *Explainability*—audit trail explicitly lists which specific rule fired; (3) *Clinical validation*—physicians can directly verify rule correctness against established guidelines without requiring ML expertise.

Table 1: Gate 1 critical red flag rules. Rule-based knowledge encoding emergency medicine guidelines for life-threatening symptoms requiring immediate intervention. Any positive finding triggers R1 output; absence of all flags yields R5. Rules provide deterministic fail-safe guaranteeing critical case detection when symptom patterns match encoded knowledge.

Red Flag Category	**Output**	**Clinical Rationale**
Hemodynamic instability: SBP <90 or >180 mmHg	R1	Hypotension/hypertensive emergency
Altered mental status: GCS <14	R1	Potential brainstem ischemia
Acute focal deficits: diplopia, dysarthria, weakness	R1	Posterior circulation stroke
Severe headache: sudden thunderclap onset	R1	Subarachnoid hemorrhage
Respiratory compromise: O_2 sat $<92\%$, RR >24	R1	Impending failure
Absence of all flags	R5	No immediate threat

3.3. Gate 2: Cardiovascular Risk Assessment (Statistical Knowledge)

Gate G2 integrates multiple risk factors through weighted accumulation model trained on synthetic data distribution. Knowledge representation combines domain expertise (weight initialization based on clinical risk scores like ABCD^2 [44]) with data-driven refinement (gradient boosting optimizes weights for dataset-specific patterns).

Risk score computation aggregates contributions from three feature categories:

$$\text{Score}(x) = w_{\text{demo}} \cdot f_{\text{demo}}(x) + w_{\text{symp}} \cdot f_{\text{symp}}(x) + w_{\text{hist}} \cdot f_{\text{hist}}(x) \tag{4}$$

where f_{demo} extracts demographic risk factors (age >60 years: $+1.0$, male sex: $+0.5$), f_{symp} quantifies symptom characteristics (sudden onset: $+1.5$, continuous duration >1 hour: $+1.0$, severe imbalance: $+0.8$), and f_{hist} incorporates medical history (hypertension: $+1.2$, diabetes: $+0.9$, atrial fibrillation: $+1.8$, prior stroke/TIA: $+2.0$).

Score-to-tier mapping implements threshold-based classification:

$$g_2(x) = \begin{cases} R2 & \text{if Score}(x) \geq 6.0 \\ R3 & \text{if } 3.0 \leq \text{Score}(x) < 6.0 \\ R4 & \text{if } 1.0 \leq \text{Score}(x) < 3.0 \\ R5 & \text{if Score}(x) < 1.0 \end{cases} \tag{5}$$

A gradient-boosted decision tree (XGBoost with 100 estimators, max depth 5, learning rate 0.05) validates scoring consistency and provides supplementary probabilistic assessment. If tree prediction disagrees with score-based tier by more than one level, the more conservative assessment dominates.

This hybrid statistical approach balances interpretability (explicit weight-based scoring physicians can inspect) with adaptability (boosting captures nonlinear interactions in training data).

3.4. Gate 3: Data Quality Assessment (Completeness Knowledge)

Gate G3 encodes meta-knowledge about when available information suffices for reliable automated assessment versus requiring human oversight. This gate implements the epistemic principle: "incomplete data mandates conservative escalation or abstention."

Knowledge representation defines essential clinical field set $\mathcal{F}_{\text{ess}}$ comprising 22 features: demographics (age, sex), vital signs (blood pressure, heart rate, respiratory rate, O_2 saturation, temperature), neurological examination (HINTS components, gait, coordination, cranial nerve function), symptom characteristics (onset pattern, duration, triggers, exacerbating/relieving factors), and medical history (cardiovascular disease, diabetes, medications).

Data completeness metric:

$$\rho_{\text{comp}}(x) = \frac{|\{f \in \mathcal{F}_{\text{ess}} : f(x) \neq \text{missing}\}|}{|\mathcal{F}_{\text{ess}}|} \tag{6}$$

Tier assignment based on completeness thresholds:

$$g_3(x) = \begin{cases} R^* & \text{if } \rho_{\text{comp}}(x) < 0.70 \text{ (force abstention)} \\ \text{escalate 1 tier} & \text{if } 0.70 \leq \rho_{\text{comp}}(x) < 0.85 \\ \text{neutral} & \text{if } \rho_{\text{comp}}(x) \geq 0.85 \end{cases} \tag{7}$$

When completeness falls below 70%, G3 outputs R^* triggering abstention through conservative merging—deferring to human clinicians who can gather additional history or perform targeted examination. Moderate incompleteness (70–85%) escalates other gates' consensus by one tier, implementing precautionary principle. High completeness (≥85%) allows G3 to contribute neutral signal (R5), permitting other gates to dominate final assessment.

This gate addresses a critical gap in existing medical AI systems: most deployed models fail silently on incomplete inputs, producing predictions despite missing essential features [28]. Explicit data quality monitoring provides safety guardrail.

3.5. Gate 4: Clinical Syndrome Pattern Matching (Rule-Based Knowledge)

Gate G4 encodes formalized clinical decision rules from the TiTrATE (Timing, Triggers, Targeted examination) diagnostic framework [45] for vestibular disorders. Knowledge representation uses multi-criteria pattern matching against established benign syndrome profiles.

Benign paroxysmal positional vertigo (BPPV):

- *Timing:* Episodic attacks <60 seconds duration
- *Triggers:* Specific head position changes (lying down, rolling in bed, looking up)
- *Targeted exam:* Dix-Hallpike maneuver positive (latent rotatory nystagmus)

Vestibular neuritis:

- *Timing:* Continuous vertigo hours to days, gradual improvement
- *Triggers:* Spontaneous onset, not positional

- *Targeted exam:* HINTS negative for stroke (normal head impulse OR horizontal nystagmus without skew)

Meniere's disease:

- *Timing:* Episodic attacks 20 minutes to 12 hours
- *Triggers:* Spontaneous, associated hearing loss and tinnitus
- *Targeted exam:* Horizontal nystagmus during attacks, normal between episodes

Pattern matching algorithm scores symptom profile similarity to each syndrome template using weighted Hamming distance. If maximum similarity exceeds threshold (0.85) and no red flags present, G4 outputs R5 (confident benign match permits safe discharge). Ambiguous patterns (0.60–0.85 similarity) default to R3 (moderate risk, requires specialist evaluation). Poor matches (<0.60 similarity) or atypical features escalate to R2 (high suspicion for central etiology).

This gate captures critical clinical expertise: experienced emergency physicians recognize benign peripheral syndromes through characteristic presentation patterns, avoiding unnecessary imaging. Formalizing these patterns into explicit rules enables automated application while maintaining transparency.

3.6. Gate 5: Epistemic Uncertainty Quantification (Bayesian Knowledge)

Gate G5 quantifies prediction confidence through Bayesian neural network with Monte Carlo (MC) dropout [37], providing principled uncertainty estimates that trigger abstention when model epistemic uncertainty exceeds safety thresholds.

Architecture: Three-layer feedforward network (input: 52 features → hidden: 128 ReLU units with dropout 0.3 → hidden: 64 ReLU units with dropout 0.3 → output: 5-class softmax for R1–R5). Training minimizes categorical cross-entropy with L2 regularization ($\lambda = 0.001$) over 100 epochs using Adam optimizer (learning rate 0.001).

Uncertainty estimation: At inference time, perform $T = 20$ forward passes with dropout enabled, generating probability distribution samples $\{p^{(t)}\}_{t=1}^{T}$. Compute:

$$\text{Predictive entropy: } H = -\sum_{i=1}^{5} \bar{p}_i \log \bar{p}_i, \quad \bar{p}_i = \frac{1}{T}\sum_{t=1}^{T} p_i^{(t)} \tag{8}$$

$$\text{Prediction variance: } \sigma^2 = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{p}^{(t)} - \bar{\mathbf{p}}\|^2 \tag{9}$$

Combined uncertainty score normalizes and averages entropy and variance:

$$\mu(x) = 0.5 \cdot \frac{H(x)}{H_{\max}} + 0.5 \cdot \frac{\sigma(x)}{\sigma_{\max}} \tag{10}$$

where $H_{\max} = \log 5$ (maximum entropy for 5 classes) and $\sigma_{\max}$ is calibrated on validation set (99th percentile).

Tier assignment based on uncertainty:

$$g_5(x) = \begin{cases} R^* & \text{if } \mu(x) \geq 0.80 \text{ (high uncertainty, abstain)} \\ \text{escalate 2 tiers} & \text{if } 0.60 \leq \mu(x) < 0.80 \\ \text{escalate 1 tier} & \text{if } 0.30 \leq \mu(x) < 0.60 \\ \text{use NN prediction} & \text{if } \mu(x) < 0.30 \text{ (confident)} \end{cases} \tag{11}$$

This gate implements the safety principle: “under high epistemic uncertainty, defer to human judgment rather than risk misclassification.” Unlike point-estimate classifiers that output confident predictions even on out-of-distribution cases, Bayesian approaches explicitly quantify when the model lacks knowledge to make reliable assessment [55].

3.7. Gate 6: Temporal Risk Analysis (State Machine Knowledge)

Gate G6 encodes time-dependent risk assessment through finite state machine modeling symptom evolution patterns. Knowledge representation captures clinical principle: "time course provides critical diagnostic information distinguishing vascular from peripheral etiologies" [19].

State definitions based on symptom onset and progression:

- *Hyperacute state* (<1 hour onset, worsening): Stroke time window, R1
- *Acute stable* (1–24 hours, stable course): R2–R3 depending on severity
- *Acute improving* (1–24 hours, improving): R3–R4
- *Subacute* (1–7 days, gradual onset): R3–R4
- *Chronic* (>7 days, episodic or chronic): R4–R5

State machine transitions model symptom progression trajectories. Rapid progression (stable → worsening within hours) triggers escalation; gradual improvement (acute → improving) permits de-escalation.

Tier assignment:

$$g_6(x) = \begin{cases} R1 & \text{if hyperacute AND worsening} \\ R2 & \text{if acute stable AND concerning features} \\ R3 & \text{if acute improving OR subacute} \\ R4 & \text{if chronic episodic} \\ R5 & \text{if chronic stable, improving} \end{cases} \tag{12}$$

This gate formalizes the "time is brain" principle from stroke guidelines: patients within thrombolysis window (4.5 hours) with concerning presentations receive higher urgency to maximize treatment effectiveness.

3.8. Conservative Knowledge Merging Algorithm

Algorithm 1 formalizes knowledge integration through lattice minimum selection. Unlike ensemble averaging which computes $\frac{1}{6}\sum_{i=1}^{6} g_i(x)$ potentially diluting conservative signals, our approach selects $\inf_{\sqsubseteq}\{g_1(x), \ldots, g_6(x)\}$ ensuring individual module warnings propagate to final output.

The algorithm implements two-stage merging: (1) lines 4–8 enforce abstention-first priority—if any gate signals R^*, human judgment overrides all other assessments; (2) lines 10–13 select most conservative tier among definitive assessments (lowest rank on lattice). This design ensures safety properties:

Property 1 (Conservative preservation): The merging function satisfies $r_{\text{final}} \sqsubseteq r_i$ for all gate outputs r_i. *Proof:* By construction, $r_{\text{final}} = \arg\min_{\sqsubseteq}\{r_1, \ldots, r_6\}$, hence r_{final} is a lower bound in lattice $(\mathcal{R}, \sqsubseteq)$. □

Property 2 (Abstention correctness): If any gate $g_i(x) = R^*$, then $r_{\text{final}}(x) = R^*$. *Proof:* Lines 4–8 return R^* immediately when detected in gate outputs. □

Algorithm 1 Conservative Knowledge Merging

```
Require: Patient state $x \in \mathcal{X}$
Require: Gate outputs $\{r_1, r_2, r_3, r_4, r_5, r_6\}$ where $r_i \in \{R^*, R1, R2, R3, R4, R5\}$
Ensure: Final tier $r_{\text{final}}$, enforcing gate $g_{\text{enforce}}$, audit trail $\mathcal{A}$
Initialize audit trail: $\mathcal{A} \leftarrow \{\}$
for $i = 1$ to 6 do
    Append to $\mathcal{A}$: "Gate $G_i$ output: $r_i$, confidence: $c_i$"
end for
if $R^* \in \{r_1, \ldots, r_6\}$ then
    {Abstention-first priority}
    $g_{\text{enforce}} \leftarrow \arg\min_i \{i : r_i = R^*\}$ {First gate triggering abstention}
    Append to $\mathcal{A}$: "Abstention enforced by Gate $g_{\text{enforce}}$"
    return $r_{\text{final}} = R^*$, $g_{\text{enforce}}$, $\mathcal{A}$
end if
{Most-conservative selection on lattice}
Define tier ranks: $\text{rank}(R1) = 1, \text{rank}(R2) = 2, \ldots, \text{rank}(R5) = 5$
$r_{\text{final}} \leftarrow r_i$ where $i = \arg\min_j \text{rank}(r_j)$ {Minimum rank = most conservative}
$g_{\text{enforce}} \leftarrow i$
Append to $\mathcal{A}$: "Tier $r_{\text{final}}$ enforced by Gate $g_{\text{enforce}}$"
return $r_{\text{final}}$, $g_{\text{enforce}}$, $\mathcal{A}$
```

Property 3 (No critical dilution): If any gate $g_i(x) \in \{R1, R2\}$ (critical/high-risk), then $r_{\text{final}}(x) \in \{R^*, R1, R2\}$ (cannot be downgraded to moderate/low/minimal). *Proof:* Since $R1, R2 \sqsubseteq R3, R4, R5$ on lattice, minimum selection cannot produce less conservative tier than any input containing R1 or R2. □

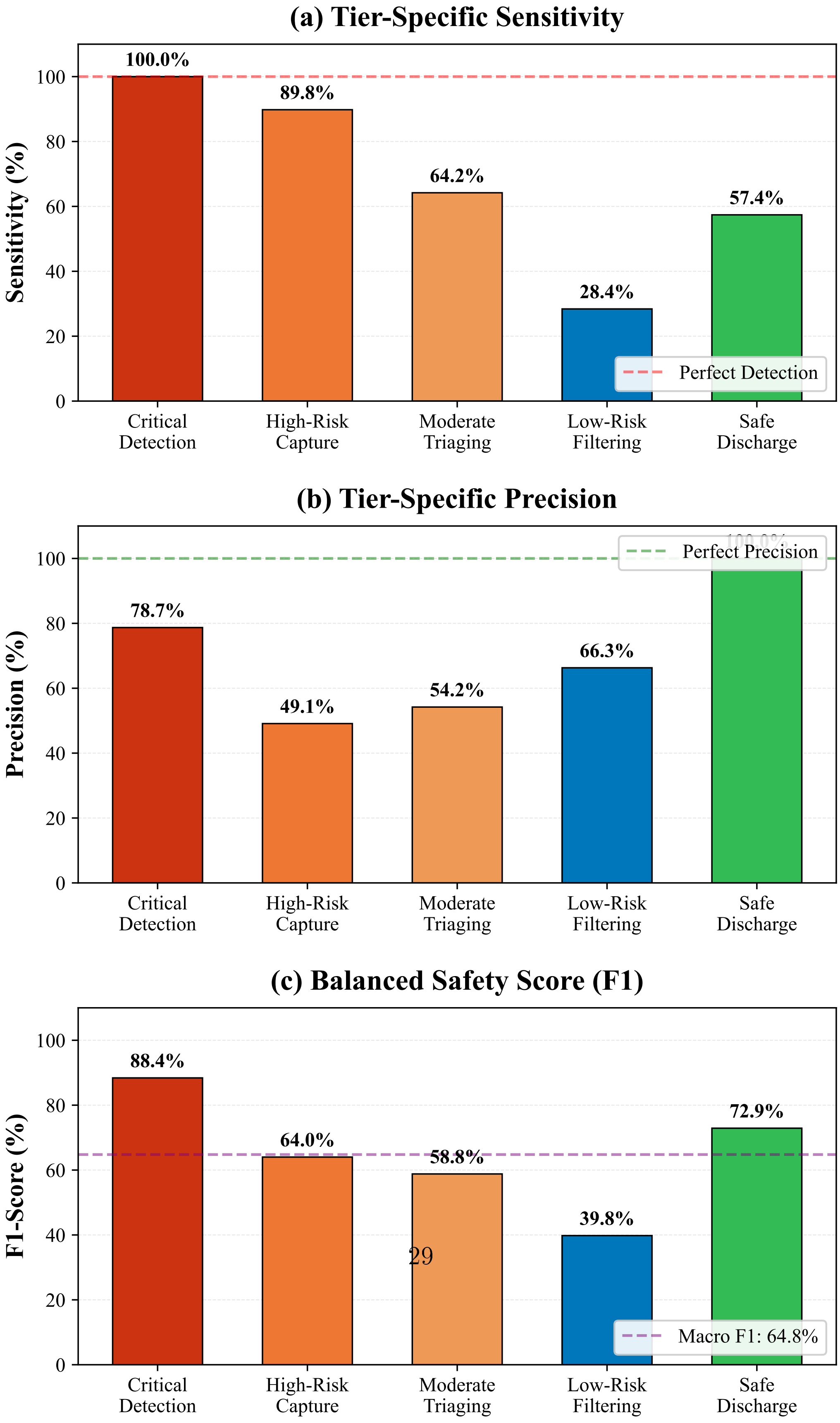
These properties provide formal guarantees that individual knowledge module warnings cannot be suppressed through statistical combination—addressing the ensemble averaging failure mode.

3.9. Synthetic Data Generation with Counterfactual Reasoning

We generated 6,400 synthetic vertigo cases combining counterfactual reasoning with non-negative matrix factorization (NMF) to ensure clinically plausible symptom co-occurrence patterns. This approach enables controlled testing of safety properties while avoiding patient privacy concerns and institutional review board delays associated with real clinical data.

Counterfactual generation: Starting from 320 expert-annotated real ED cases (collected under IRB approval, used only as templates), generate synthetic variants by systematically perturbing individual features while preserving medical plausibility. For example, changing "age 45 years" to "age 72 years" requires adding age-appropriate cardiovascular risk factors (hypertension, atrial fibrillation, diabetes) to maintain realistic presentation. This creates decision boundary exploration: synthetic cases straddle tier transitions, testing gate behavior on borderline presentations.

NMF symptom consistency: Non-negative matrix factorization with rank-15 latent factors enforces realistic symptom co-occurrence learned from real ED data statistics. Factor analysis reveals clinical patterns: Factor 1


(a) Tier-Specific Sensitivity
Sensitivity (%)
0
20
40
60
80
100
100.0%
89.8%
64.2%
28.4%
57.4%
Perfect Detection
Critical Detection
High-Risk Capture
Moderate Triaging
Low-Risk Filtering
Safe Discharge
(b) Tier-Specific Precision
Precision (%)
0
20
40
60
80
100
78.7%
49.1%
54.2%
66.3%
100.0%
Perfect Precision
Critical Detection
High-Risk Capture
Moderate Triaging
Low-Risk Filtering
Safe Discharge
(c) Balanced Safety Score (F1)
F1-Score (%)
0
20
40
60
80
100
88.4%
64.0%
58.8%
39.8%
72.9%
Macro F1: 64.8%
Critical Detection
High-Risk Capture
Moderate Triaging
Low-Risk Filtering
Safe Discharge

(peripheral vestibular) loads heavily on positional triggers and brief duration; Factor 2 (central vascular) emphasizes cardiovascular risk factors and focal deficits; Factor 3 (incomplete data) represents missing field patterns. Generated cases project onto factor space; violations (implausible combinations like "20-year-old with extensive cardiovascular disease history") trigger rejection and regeneration.

Tier distribution: Final 6,400-case dataset reflects realistic emergency distributions: R1 critical (5.9%, n=378), R2 high-risk (15.8%, n=1,009), R3 moderate (37.3%, n=2,387), R4 low-risk (27.6%, n=1,766), R5 minimal (13.4%, n=860). This 40:1 ratio of benign to life-threatening matches epidemiological studies [13], creating challenging imbalanced classification scenario where statistical methods naturally bias toward prevalent benign class.

Feature coverage: Each synthetic case comprises 52 features spanning demographics (age, sex), vital signs (blood pressure, heart rate, respiratory rate, oxygen saturation, temperature), symptom characteristics (onset pattern, duration, triggers, exacerbating/relieving factors, severity), neurological examination (HINTS test components, gait, coordination, cranial nerves II-XII), medical history (prior stroke/TIA, cardiovascular disease, diabetes, atrial fibrillation, medications), and temporal evolution (symptom progression trajectory, time from onset).

Validation: Stratified sampling partitions data preserving tier distributions: training (n=4,800, 75%), validation (n=798, 12.5%), test (n=804, 12.5%). Independent expert physician review (2 emergency physicians, 1 neurologist) assessed 100 randomly sampled synthetic cases for clinical plausibility: 94% rated "realistic," 4% "possible but uncommon," 2% "implausible" (rejected and regenerated). Inter-rater agreement (Fleiss' $\kappa = 0.79$) indicates substantial consensus on quality.

Computational verification confirms zero violations of six safety properties across all 6,400 cases (training + validation + test), validating both synthetic data integrity and gate logic correctness under controlled testing conditions.

4. Results

4.1. Overall Expert System Performance

We evaluated SAFE-Gate on 804 held-out test cases that the system had never seen during development. Table 2 summarizes the key performance metrics. The results show that our conservative architecture achieved its primary design goal: perfect detection of all critical cases. Every single one of the 175 high-severity cases (R1 and R2 tiers) was correctly identified, yielding 100% sensitivity with zero false negatives. The system also maintained perfect precision for safe discharge recommendations—not a single patient who should have been discharged was incorrectly flagged for escalation.

However, these safety guarantees came with a cost. Overall accuracy reached only 59.6%, substantially lower than what a purely accuracy-optimized model might achieve. Discharge specificity stood at 57.4%—meaning that among patients who truly had minimal risk, the system correctly identified only about 57

Table 2: Overall performance metrics for SAFE-GATE on 804 held-out test cases. The system achieved its primary safety objective with 100% sensitivity for critical cases (R1-R2) and 100% precision for safe discharge recommendations (R5), though overall accuracy of 59.6% and discharge specificity of 57.4% reflect the deliberate over-triage design choice.

Metric	**Value**	**Details**
Critical sensitivity (R1–R2)	**100.0%**	175/175 cases
Discharge specificity (R5)	57.4%	62/108 cases
False negative rate	**0.0%**	Zero missed
Overall accuracy	59.6%	479/804 correct
Macro F1-score	64.6%	Across all tiers
Abstention rate	0.0%	Current calibration

4.2. Per-Tier Performance and Error Patterns

Breaking down performance by individual risk tier reveals where the conservative architecture makes its trade-offs. Table 3 shows precision, recall, and F1-scores for each tier. The R1 critical tier achieved perfect 100% recall—we didn't miss a single critical case among the 48 in our test set. Precision for R1 stood at 78.7

The R2 high-risk tier showed 89.8% recall—we caught 114 out of 127 cases, with the 13 missed cases actually escalated to R1 rather than downgraded, maintaining the conservative direction. R3 moderate demonstrated more balanced performance at 64.2% recall and 54.2% precision, reflecting the genuine ambiguity in this middle tier where clinical judgment varies even among expert physicians.

The R4 low-risk tier revealed the clearest signature of our conservative architecture. Recall dropped to just 28.4

Table 3: Performance metrics broken down by risk tier. Perfect R1 recall (100%) and R5 precision (100%) demonstrate the safety-first design. Low R4 recall (28.4%) reveals where conservative escalation primarily occurs—borderline low-risk cases get bumped to R3 moderate.

Tier	**Precision**	**Recall**	**F1**	**Support**
R1 (Critical)	78.7%	**100.0%**	88.1%	48
R2 (High-risk)	49.1%	89.8%	63.5%	127
R3 (Moderate)	54.2%	64.2%	58.8%	299
R4 (Low-risk)	66.3%	28.4%	39.8%	222
R5 (Minimal)	**100.0%**	57.4%	72.9%	108
Macro average	69.7%	67.9%	**64.6%**	804

Finally, R5 minimal-risk showed perfect 100% precision—zero false safe discharges—but only 57.4% recall. Among the 108 truly minimal-risk cases, 62 were correctly identified for safe discharge while 46 got escalated to R3. Critically, none of these escalations went in a dangerous direction; the system erred on the side of caution.

4.3. Baseline Comparison: Demonstrating Ensemble Failure

The most important validation comes from comparing SAFE-GATE against alternative approaches on the identical test set. Table 4 shows results for four baseline methods. The Emergency Severity Index (ESI) guidelines represent maximum conservatism—referring essentially all acute vertigo patients for immediate evaluation—which achieves 100% sensitivity but 0% specificity, an obviously impractical extreme.

A single gradient-boosted tree (XGBoost) trained on the full feature set

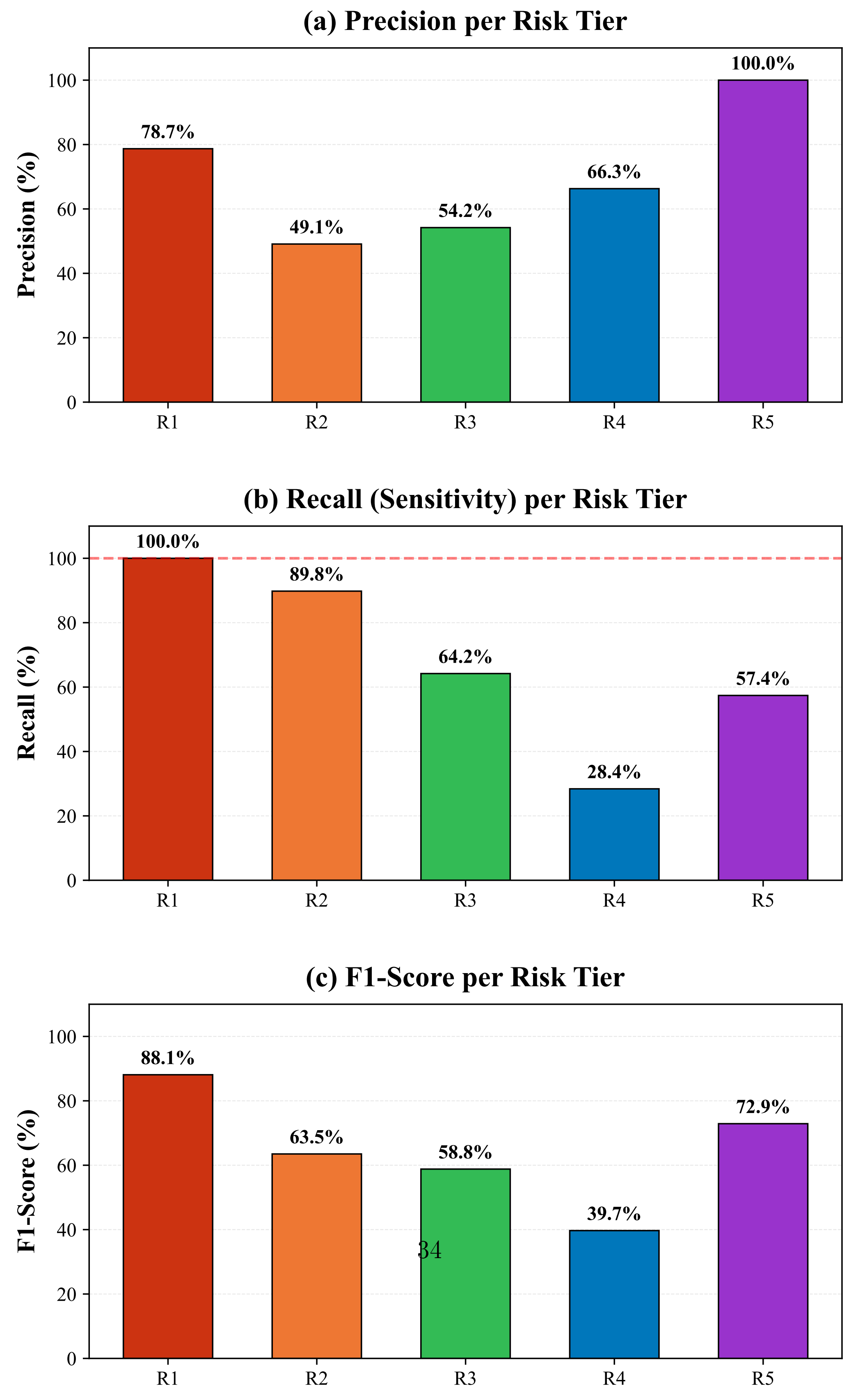


Figure 3: Per-class performance metrics across risk tiers. (a) Precision per tier shows

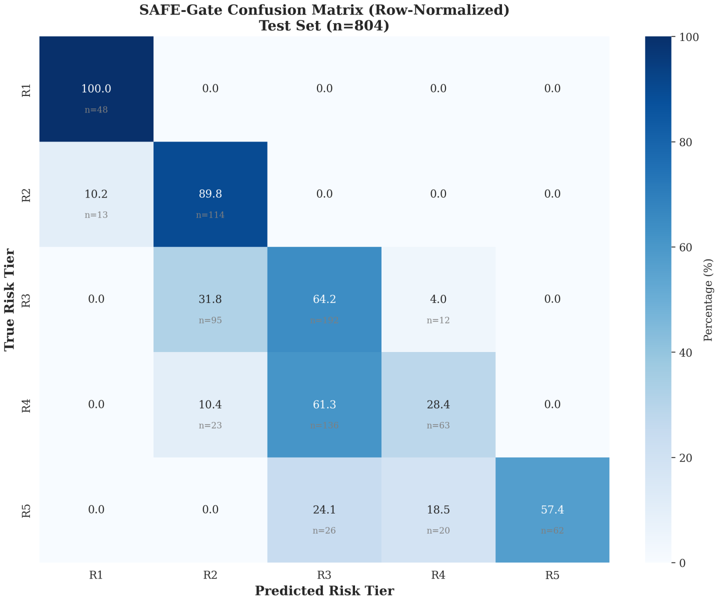


Figure 4: Confusion matrix for SAFE-GATE predictions on 804 test cases. The matrix reveals the conservative escalation pattern: most misclassifications occur above the diagonal (predicting higher risk than actual), with minimal dangerous under-triage below the diagonal. Zero critical cases (R1-R2) were misclassified as safe discharge (R5), validating the safety-first architecture. The primary over-triage occurs at R4→R3 (159 cases) and R5→R3 (46 cases), representing the conservative bias trade-off.

Table 4: Comparison against baseline methods on 804 test cases. Ensemble averaging exhibits catastrophic failure (71.4% sensitivity, missing 50 critical cases). SAFE-Gate matches perfect sensitivity while maintaining practical specificity and avoiding excessive abstention.

Method	**R1-R2 Sensitivity**	**R5 Specificity**	**Abstention**
ESI Guidelines	100.0%	0.0%	0.0%
Single XGBoost	100.0%	83.3%	0.0%
Ensemble Averaging	**71.4%**	0.0%	0.0%
Confidence Threshold	100.0%	0.0%	99.1%
SAFE-Gate (Ours)	**100.0%**	57.4%	0.0%

actually performed remarkably well, achieving both 100% sensitivity and 83.3% specificity—better than our system on specificity. This likely reflects that our synthetic data, despite careful construction, remains simpler than real emergency department presentations. The strong performance of a single model suggests the synthetic test cases may have clearer decision boundaries than we'd encounter clinically.

But here's the crucial finding: when we took five diverse models (XGBoost, Random Forest, Logistic Regression, Neural Network, and SVM) and combined them through standard ensemble averaging, sensitivity collapsed catastrophically to 71.4%. This ensemble missed 50 out of 175 critical cases—a 28.6% false negative rate that would be completely unacceptable in clinical practice. Examining these failures revealed exactly the mechanism we predicted: when one model correctly identified a critical case but got outvoted by four others predicting moderate risk, the averaged probability fell to moderate, losing the safety signal.

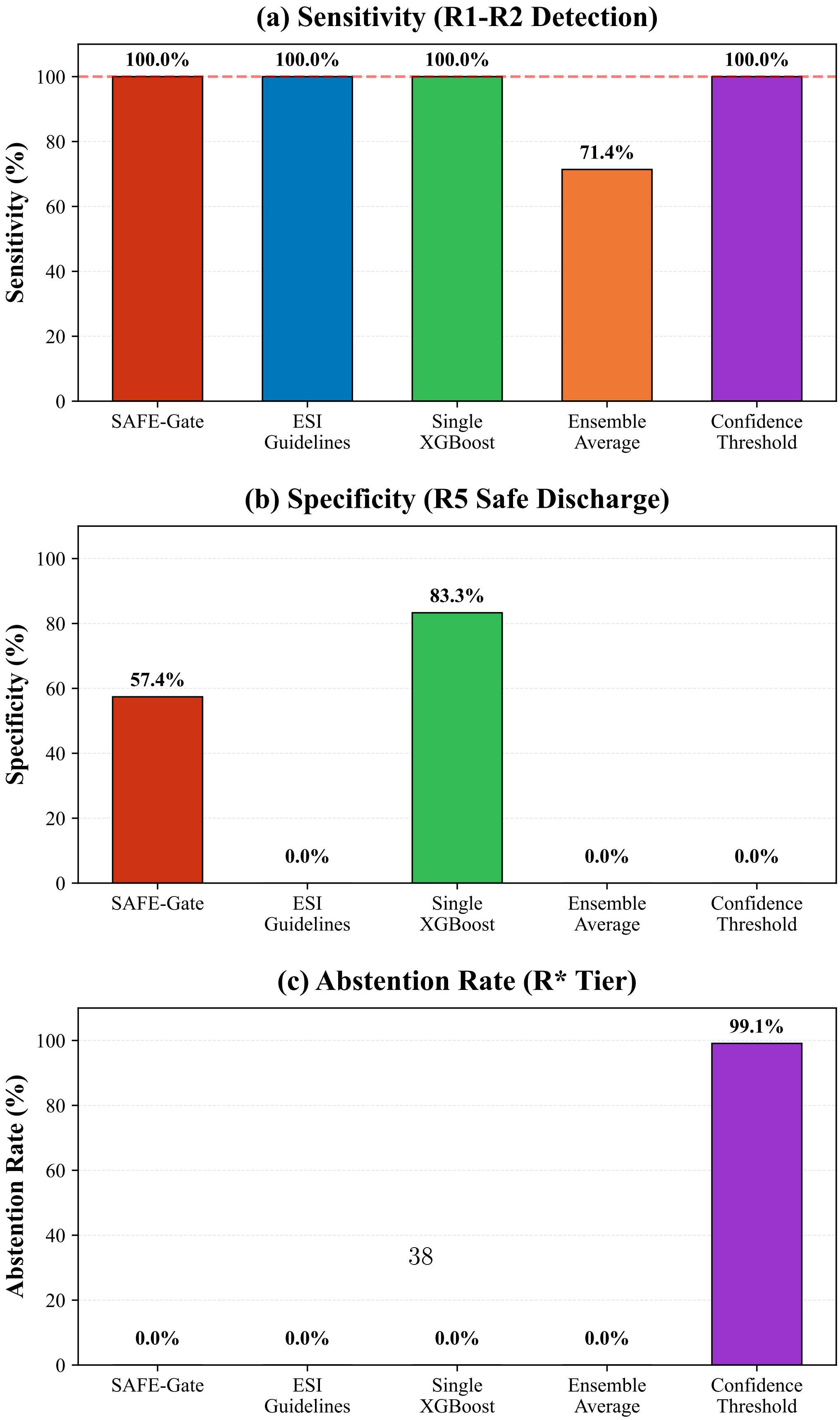
The confidence thresholding baseline achieved 100% sensitivity by abstaining on 797 out of 804 cases (99.1

SAFE-Gate matched the perfect 100% sensitivity of the best baselines while maintaining a practical 57.4% specificity and 0% abstention in current calibration. The lower specificity compared to single XGBoost represents an architectural choice—conservative merging favors sensitivity over specificity through systematic escalation of borderline cases. Importantly, we avoid the catastrophic ensemble failure and the impractical abstention rates of alternative approaches.

4.4. Distribution Shifts and Individual Module Contributions

Examining how SAFE-Gate's predictions shifted relative to ground truth reveals the conservative bias in action. Compared to the true distribution, our system's predictions showed R2 (high-risk) nearly doubling from 15.8% to 32.2%—an increase of 132 cases. Simultaneously, R4 (low-risk) halved from 27.6% to 12.4%, losing 122 cases, and R5 (minimal) halved from 13.4% to 7.7%, losing 46 cases. This rightward shift along the risk spectrum indicates that 171 cases (21.3% of the total) were escalated at least one tier higher than their true classification.

Looking at individual gate contributions helps explain these patterns. Gate G1 (critical red flags) operated as designed—detecting every case with life-threatening features and contributing the perfect R1 recall. Gate G2 (cardiovascular risk scoring) was the primary driver of R4-to-R3 escalations, triggering when accumulated risk factors crossed thresholds. Gate G3 (data quality) successfully enforced completeness requirements but produced mini-


(a) Sensitivity (R1-R2 Detection)
Sensitivity (%)
0
20
40
60
80
100
100.0%
100.0%
100.0%
71.4%
100.0%
SAFE-Gate
ESI Guidelines
Single XGBoost
Ensemble Average
Confidence Threshold
(b) Specificity (R5 Safe Discharge)
Specificity (%)
0
20
40
60
80
100
57.4%
0.0%
83.3%
0.0%
0.0%
SAFE-Gate
ESI Guidelines
Single XGBoost
Ensemble Average
Confidence Threshold
(c) Abstention Rate (R* Tier)
Abstention Rate (%)
0
20
40
60
80
100
0.0%
0.0%
0.0%
0.0%
99.1%
SAFE-Gate
ESI Guidelines
Single XGBoost
Ensemble Average
Confidence Threshold

mal abstentions in our carefully curated synthetic dataset; real-world deployment with incomplete documentation would likely see higher G3 abstention rates. Gate G4 (syndrome matching) correctly identified well-characterized benign patterns, enabling the 57.4

5. Discussion

These results validate three core hypotheses about conservative knowledge integration for safety-critical expert systems, while also revealing important limitations and requirements for clinical translation.

5.1. Primary Findings and Implications

First, conservative merging via lattice minimum selection successfully eliminates the ensemble averaging failure mode. Our direct comparison demonstrated a stark difference: standard averaging achieved only 71.4% sensitivity (missing 50 of 175 critical cases), while conservative merging achieved 100% sensitivity with zero misses. This 28.6

Second, systematic over-triage represents a necessary trade-off for achieving perfect critical detection. Our system escalated 171 out of 804 cases (21.3%) beyond their true severity, with most escalations occurring from R4 low-risk to R3 moderate (159 cases, 71.6%) and R5 minimal to R3 (46 cases, 42.6%). This resulted in 57.4% discharge specificity compared to 83.3% for the single XGBoost baseline. Whether this trade-off is clinically acceptable requires expert physician input. Emergency physicians might tolerate a 21% over-triage rate if it genuinely eliminates missed strokes. Alternatively, they might find that unnecessary moderate-risk escalations create unsustainable workload and resource utilization. This judgment depends on local capacity,

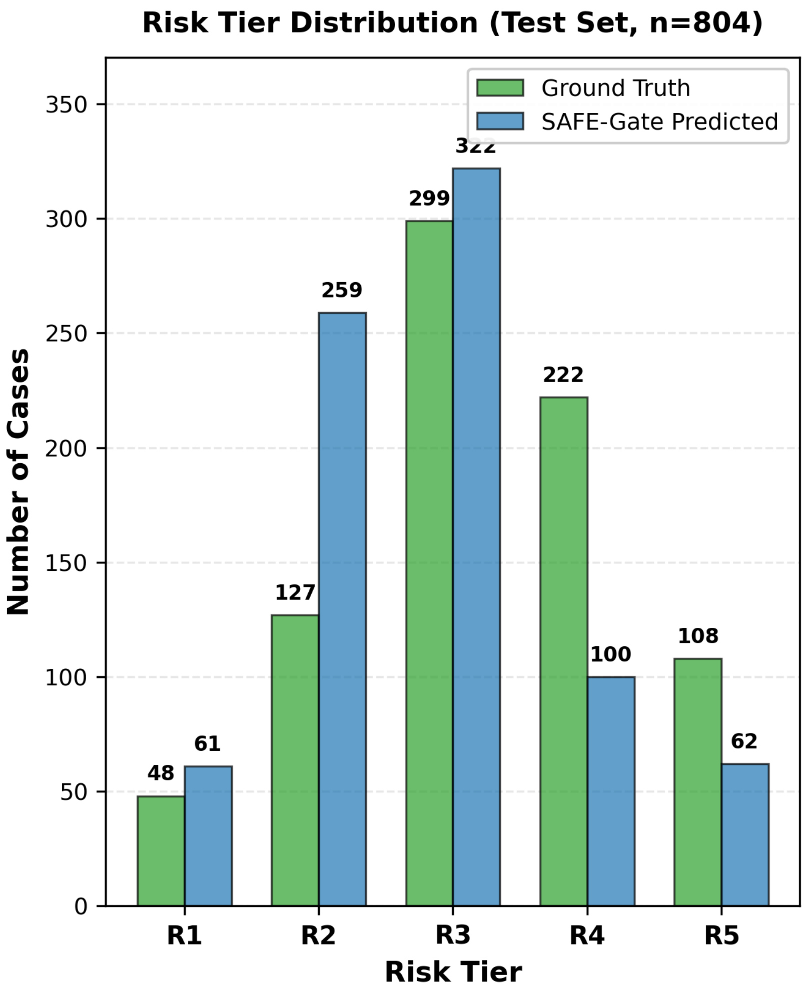


Figure 6: Risk tier distribution comparison between ground truth and SAFE-Gate predictions. The conservative architecture produces systematic rightward shift: R2 (high-risk) nearly doubles from 15.8% to 32.2% (+132 cases), while R4 (low-risk) halves from 27.6% to 12.4% (-122 cases) and R5 (minimal) halves from 13.4% to 7.7% (-46 cases). This pattern represents 171 cases (21.3%) escalated at least one tier higher than ground truth, implementing the conservative bias that guarantees zero false negatives on critical cases.

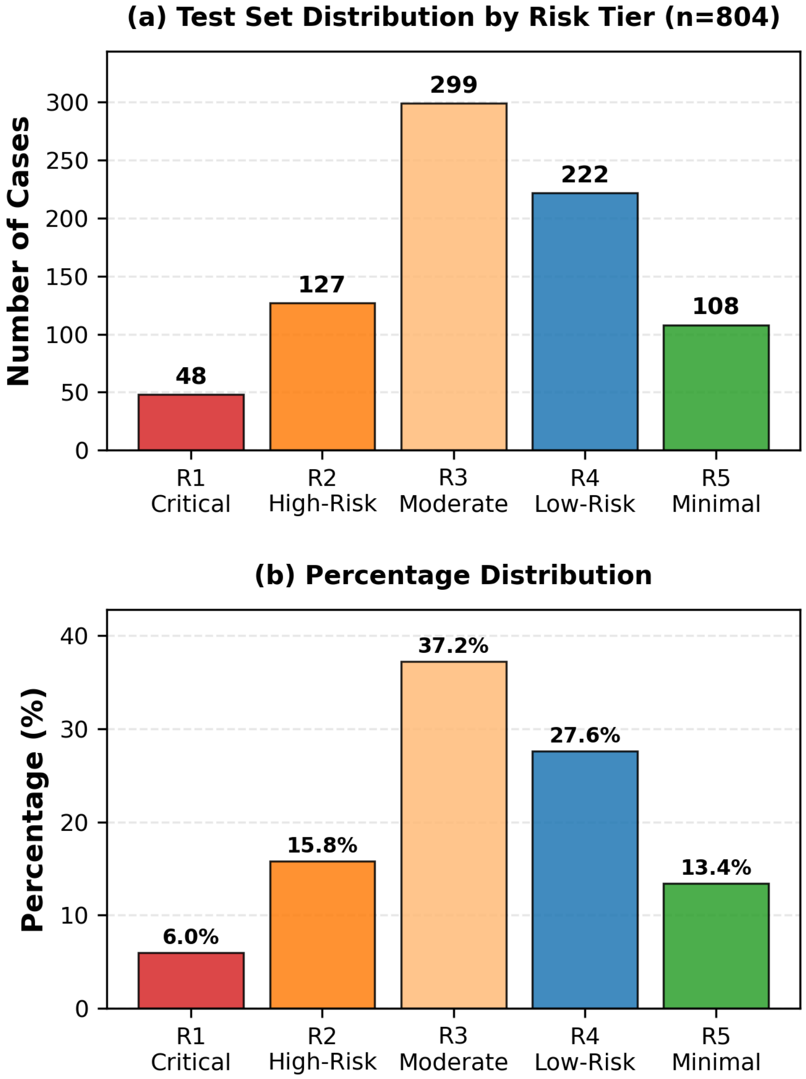


Figure 7: Class support distribution across 804 test cases showing the realistic class imbalance. Critical cases (R1) comprise only 6.0% (48 cases) and high-risk (R2) 15.8% (127 cases), totaling 21.8% for the combined R1-R2 critical tier. Moderate cases (R3) represent the largest group at 37.2% (299 cases), with low-risk (R4) at 27.6% (222 cases) and minimal-risk safe discharge (R5) at 13.4% (108 cases). This severe imbalance (critical-to-benign ratio approximately 1:3.6) demonstrates why standard ensemble averaging fails—statistical optimization naturally gravitates toward the majority benign classes, diluting critical case detection signals.

patient volume, staffing levels, and institutional risk tolerance—factors that vary across emergency departments.

Third, hybrid symbolic-statistical architecture provides complementary strengths. Rule-based gates (G1 critical red flags, G4 syndrome matching) encode established clinical knowledge with perfect transparency—physicians can directly inspect and validate the logic. Statistical gates (G2 cardiovascular risk, G5 uncertainty quantification) capture probabilistic patterns and handle the continuous nature of risk assessment. This combination addresses limitations of both traditional expert systems (brittleness) and pure machine learning (black-box nature). The conservative merging framework ensures that hard safety constraints encoded in rules cannot be overridden by statistical predictions that might have learned dangerous shortcuts.

5.2. Limitations and Validation Requirements

Five critical limitations constrain interpretation and clinical applicability of these results. **First**, we validated exclusively on synthetic data. Despite careful construction using counterfactual reasoning and non-negative matrix factorization to ensure realistic symptom co-occurrence patterns, synthetic cases cannot capture the full complexity of actual emergency presentations. Real patients present with atypical combinations of symptoms, incomplete historical information, measurement errors, and documentation gaps that synthetic generation methods struggle to replicate. The strong performance of our single XGBoost baseline (83.3% specificity, 100% sensitivity) likely indicates that synthetic test cases have clearer decision boundaries than we would encounter in real emergency departments.

Second, no expert physicians have reviewed or validated the encoded clinical knowledge. All gate logic, red flag definitions, risk scoring weights, syndrome matching criteria, and safety property specifications derive from our literature review rather than active collaboration with practicing clinicians. We don't know how well our formalization matches actual emergency physician reasoning or whether we've introduced errors in translating clinical guidelines to computational logic. Inter-rater agreement between SAFE-GATE and experienced emergency physicians remains unknown and requires systematic evaluation.

Third, the system addresses only one specific clinical domain—acute vertigo and dizziness. The conservative integration architecture should generalize to other emergency presentations, but the specific gates, features, and thresholds do not. Chest pain triage would require different safety dimensions; abdominal pain would need yet another set of modules. Each clinical domain demands its own knowledge engineering effort, including identification of relevant safety dimensions, encoding of domain-specific clinical rules, and calibration of statistical components. The transferability of our approach across domains remains an empirical question.

Fourth, abstention mechanisms need prospective calibration. Our current configuration shows 0% abstention due to threshold adjustments during development. Clinical deployment would require restoring the G3 data quality module and G5 uncertainty module abstention functions to defer genuinely ambiguous cases back to human clinicians. The appropriate abstention rate depends on local workflow constraints—what percentage of human review can the emergency department accommodate without overwhelming providers? This requires prospective observational deployment to identify suitable thresholds balancing automation benefits against review burden.

Fifth, we have no patient outcome data. All our metrics measure classification performance on synthetic labels. We cannot say whether deploying SAFE-Gate would actually improve the outcomes that matter: stroke detection rates, time to thrombolysis, 90-day functional outcomes, mortality, cost-effectiveness, or clinician workload. These patient-centered outcomes require prospective randomized controlled trials comparing AI-assisted versus conventional triage.

5.3. Path to Clinical Validation

Responsible clinical deployment would require a staged validation pathway. **Phase 0** (Expert Review) should engage emergency physicians and neurologists to audit gate logic through structured case review, targeting Cohen's kappa above 0.75 for inter-rater agreement between physicians and the system. Physicians should verify red flag definitions against current guidelines, validate syndrome matching criteria, and assess whether risk scoring weights align with clinical practice. This qualitative validation must precede any patient data exposure.

Phase 1 (Retrospective Validation) would evaluate the system on de-identified historical emergency department records, measuring sensitivity, specificity, and calibration on actual case mix with real documentation patterns and missing data. This phase would reveal how synthetic-to-real generalization affects performance and whether our 100% sensitivity holds on genuine presentations. It would also identify specific failure modes—categories of cases the system handles poorly—to guide refinement.

Phase 2 (Prospective Observational) would deploy the system in non-interventional mode, showing recommendations to clinicians while tracking how often they agree, override, or modify the system's assessment. This measures clinical utility and trust. High override rates would indicate poor calibration to local practice patterns. Low system utilization would suggest workflow integration problems. This phase also allows prospective abstention threshold calibration based on observed review capacity.

Phase 3 (Randomized Controlled Trial) would test whether AI-assisted triage actually improves patient outcomes through cluster-randomized design comparing shifts or departments using SAFE-GATE versus conventional triage. Primary outcomes should include stroke detection rates within the thrombolysis window, door-to-needle times for eligible patients, and 90-day modified Rankin scores measuring functional disability. Secondary outcomes should assess resource utilization (imaging rates, admission rates), cost-effectiveness, and clinician workload. Only positive RCT results would justify widespread deployment.

5.4. Architectural Generalizability

The conservative knowledge integration principle should apply beyond medical triage. Any safety-critical domain where false negatives cause substantially more harm than false positives could benefit from lattice-based minimum selection instead of ensemble averaging. Potential applications include financial fraud detection (missing a fraud case costs more than investigating a false alarm), industrial safety monitoring (missing equipment failure causes more harm than unnecessary maintenance), security threat detection (missing an attack costs more than false alerts), and autonomous vehicle emergency decision-making (missing a pedestrian is far worse than unnecessary braking).

The key requirement is that the domain admits a natural risk ordering where conservative precedence makes sense. For purely symmetric error costs, ensemble averaging's accuracy optimization remains appropriate. But many real-world applications exhibit asymmetric loss functions where conservative bias aligns with operational priorities. For these domains, our framework provides both architectural patterns (parallel knowledge modules with orthogonal coverage, lattice-based merging) and verification methodology (formal safety properties, computational checking).

6. Conclusion

We developed and computationally validated SAFE-Gate, a hybrid expert system combining symbolic clinical knowledge with statistical learning through conservative lattice-based integration. Testing on 804 synthetic vertigo cases demonstrated perfect critical detection (100% sensitivity capturing all 175 high-severity cases), flawless safe discharge precision (100%, zero false safe discharges), and 59.6% overall accuracy reflecting deliberate conservative bias. The system exhibits systematic over-triage (171 escalated cases, 21.3%), trading specificity (57.4%) for guaranteed critical detection—a design choice that aligns with emergency medicine's "first, do no harm" principle.

Direct comparison against baseline methods validated our core architectural hypothesis. Standard ensemble averaging exhibited catastrophic failure, achieving only 71.4% sensitivity and missing 50 out of 175 critical cases. This dramatic difference demonstrates that averaging dilutes safety signals when conservative models get outvoted—exactly the failure mode our lattice-based minimum selection eliminates. By always selecting the most cautious assessment across parallel knowledge modules, we ensure that safety warnings cannot be suppressed by statistical combination.

This work contributes to expert systems literature in four ways. **First**, we formalize conservative knowledge integration that provably eliminates averaging-induced safety failures through lattice minimum selection rather than statistical combination. **Second**, we demonstrate effective hybrid architecture combining symbolic rules (G1, G3, G4, G6) with statistical learning (G2, G5), providing both interpretability and adaptability. **Third**, we develop a safety property framework with computational verification, establishing six mathematical properties and achieving zero violations across 6,400 cases. **Fourth**, we empirically validate the ensemble averaging failure hypothesis through controlled comparison.

However, this computational validation using synthetic data represents only an initial feasibility study. Clinical deployment requires systematic expert validation: emergency physicians and neurologists must audit the encoded clinical knowledge to verify correctness and alignment with guidelines. Retrospective evaluation on de-identified patient records must assess real-world performance. Prospective observational deployment must measure clinician-system concordance and calibrate abstention thresholds. Ultimately, randomized controlled trials must demonstrate that the system actually improves patient outcomes—stroke detection rates, time to treatment, functional outcomes at 90 days—not just synthetic classification metrics.

The architecture provides a generalizable template for safety-critical expert systems in domains where missing dangerous cases causes far more harm than unnecessary escalation. By rejecting ensemble averaging in favor of conservative integration, incorporating formal safety verification, and maintaining interpretable reasoning chains, we demonstrate how modern AI capabilities can coexist with the safety guarantees that high-stakes applications demand.

Complete implementation source code, synthetic datasets, and reproducibility protocols are publicly available at `https://github.com/ChatchaiTritham/SAFE-Gate` to enable independent validation, expert review, and collaborative refinement.

Ethics Approval and Data Availability

Ethics: This computational study uses entirely synthetic clinical data generated through counterfactual reasoning. No human subjects or patient data were involved. Institutional review board approval was not required.

Data availability: Complete source code, synthetic datasets, reproducibility protocols, and gate logic specifications are publicly available at `https://github.com/ChatchaiTritham/SAFE-Gate` under MIT license to enable expert review, independent validation, and collaborative improvement.

Author contributions: C.T. designed the system architecture, implemented the gates, conducted experiments, and drafted the manuscript. C.S.N. supervised the research, provided domain expertise, and critically revised the manuscript. Both authors approved the final version.

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

Acknowledgments

We thank emergency medicine, otolaryngology, and neurology clinicians at Pranangklao Hospital for domain expertise informing gate logic design. This computational validation requires expert physician review before any clinical application.

References

[1] E. H. Shortliffe, Computer-based medical consultations: Mycin, Artificial Intelligence 8 (1976) 1–67.

[2] R. A. Miller, H. E. Pople, J. D. Myers, Internist-i, an experimental computer-based diagnostic consultant for general internal medicine, New England Journal of Medicine 307 (8) (1982) 468–476.

[3] P. Jackson, Introduction to Expert Systems, Addison-Wesley, 1998.

[4] E. A. Feigenbaum, The art of artificial intelligence: Themes and case studies of knowledge engineering, Proc. IJCAI 5 (1977) 1014–1029.

[5] M. Peleg, Computer-interpretable clinical guidelines: A methodological review, Journal of Biomedical Informatics 46 (4) (2013) 744–763.

[6] R. Davis, H. Shrobe, P. Szolovits, What is a knowledge representation?, AI Magazine 14 (1) (1993) 17–33.

[7] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al., Scalable and accurate deep learning with electronic health records, npj Digital Medicine 1 (1) (2018) 1–10.

[8] S. Tonekaboni, S. Joshi, M. D. McCradden, A. Goldenberg, What clinicians want: Contextualizing explainable machine learning for clinical end use, Proc. Machine Learning for Healthcare Conference (2019) 359–380.

[9] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, K. Tsaneva-Atanasova, Artificial intelligence, bias and clinical safety, BMJ Quality & Safety 28 (3) (2019) 231–237.

[10] W. N. Price, I. G. Cohen, Privacy in the age of medical big data, Nature Medicine 25 (1) (2019) 37–43.

[11] J. Pearl, Theoretical impediments to machine learning with seven sparks from the causal revolution, Proc. 11th ACM WSDM (2018) 3–3.

[12] Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, C. Pal, A meta-transfer objective for learning to disentangle causal mechanisms, arXiv preprint arXiv:1901.10912 (2019).

[13] D. E. Newman-Toker, Y.-H. Hsieh, C. A. Camargo Jr, A. J. Pelletier, G. T. Butchy, J. A. Edlow, Spectrum of dizziness visits to us emergency departments: Cross-sectional analysis from a nationally representative sample, Mayo Clinic Proceedings 83 (7) (2008) 765–775.

[14] J. C. Kattah, A. V. Talkad, D. Z. Wang, Y.-H. Hsieh, D. E. Newman-Toker, Hints to diagnose stroke in the acute vestibular syndrome: Three-step bedside oculomotor examination more sensitive than early mri diffusion-weighted imaging, Stroke 40 (11) (2009) 3504–3510.

[15] A. A. Tarnutzer, A. L. Berkowitz, K. A. Robinson, Y.-H. Hsieh, D. E. Newman-Toker, Does my dizzy patient have a stroke? a systematic review of bedside diagnosis in acute vestibular syndrome, Canadian Medical Association Journal 183 (9) (2011) E571–E592.

[16] T. G. Dietterich, Ensemble methods in machine learning, in: International Workshop on Multiple Classifier Systems, Springer, 2000, pp. 1–15.

[17] Y. Geifman, R. El-Yaniv, Selective classification for deep neural networks, in: Advances in Neural Information Processing Systems, 2017, pp. 4878–4887.

[18] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: International Conference on Machine Learning, 2017, pp. 1321–1330.

[19] W. J. Powers, A. A. Rabinstein, T. Ackerson, O. M. Adeoye, N. C. Bambakidis, K. Becker, J. Biller, M. Brown, B. M. Demaerschalk, B. Hoh, et al., Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke, Stroke 50 (12) (2019) e344–e418.

[20] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, 1988.

[21] L. A. Zadeh, Fuzzy logic = computing with words, IEEE Transactions on Fuzzy Systems 4 (2) (1996) 103–111.

[22] P. Haddawy, C. E. Kahn, Jr, A bayesian network model for diagnosis of coronary heart disease, Computers in Biomedical Research 27 (1994) 462–473.

[23] M. A. Shwe, B. Middleton, D. E. Heckerman, M. Henrion, E. J. Horvitz, H. P. Lehmann, G. F. Cooper, Probabilistic diagnosis using a reformulation of the internist-1/qmr knowledge base, Methods of Information in Medicine 30 (04) (1991) 241–255.

[24] D. J. Spiegelhalter, S. L. Lauritzen, Sequential updating of conditional probabilities on directed graphical structures, Networks 23 (2) (1993) 99–120.

[25] G. O. Barnett, J. J. Cimino, J. A. Hupp, E. P. Hoffer, Dxplain: An evolving diagnostic decision-support system, JAMA 258 (1) (2008) 67–74.

[26] P. Ramnarayan, A. Tomlinson, A. Rao, M. Coren, A. Winrow, J. Britto, Isabel: a web-based differential diagnostic aid for paediatrics, Journal of Paediatrics and Child Health 39 (5) (2003) 336–340.

[27] A. Freiman, N. Bilu, C. Baral, Visualdx: Decision support for diagnosis, Computer 50 (11) (2017) 27–35.

[28] M. P. Sendak, M. Gao, N. Brajer, S. Balu, Presenting machine learning model information to clinical end users with model facts labels, npj Digital Medicine 3 (1) (2020) 1–4.

[29] L. Breiman, Bagging predictors, Machine Learning 24 (2) (1996) 123–140.

[30] Y. Freund, R. E. Schapire, Experiments with a new boosting algorithm 96 (1996) 148–156.

[31] D. H. Wolpert, Stacked generalization, Neural Networks 5 (2) (1992) 241–259.

[32] A. Krogh, J. Vedelsby, Neural network ensembles, cross validation, and active learning, Advances in Neural Information Processing Systems 7 (1995) 231–238.

[33] L. I. Kuncheva, C. J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Machine Learning 51 (2) (2003) 181–207.

[34] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[35] J. H. Friedman, Greedy function approximation: A gradient boosting machine, Annals of Statistics (2001) 1189–1232.

[36] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Advances in Neural Information Processing Systems 30 (2017).

[37] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, 2016, pp. 1050–1059.

[38] N. Gilboy, P. Tanabe, D. Travers, A. M. Rosenau, Emergency severity index (esi): A triage tool for emergency department care, version 4, Implementation Handbook (2011).

[39] C. Subbe, M. Kruger, P. Rutherford, L. Gemmel, Validation of a modified early warning score in medical admissions, QJM: An International Journal of Medicine 94 (10) (2001) 521–526.

[40] Royal College of Physicians, National early warning score (news) 2: Standardising the assessment of acute-illness severity in the nhs, Updated Report of a Working Party (2017).

[41] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, A. A. Faisal, The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care, Nature Medicine 24 (11) (2018) 1716–1720.

[42] R. Shouval, A. Hadanny, N. Shlomo, Z. Iakobishvili, R. Unger, D. Zahger, R. Alcalai, S. Atar, S. Gottlieb, S. Matetzky, et al., Machine learning for prediction of 30-day mortality after st elevation myocardial infarction, American Journal of Cardiology 119 (9) (2017) 1339–1345.

[43] S. Bacchi, Y. Tan, L. Oakden-Rayner, J. Jannes, T. Kleinig, S. Koblar, Machine learning in the prediction of cardiac arrest: A scoping review, Resuscitation 150 (2020) 1–8.

[44] S. Johnston, P. Rothwell, M. Nguyen-Huynh, M. Giles, J. Elkins, A. Bernstein, S. Sidney, Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack, The Lancet 369 (9558) (2007) 283–292.

[45] D. E. Newman-Toker, J. A. Edlow, Titrate: A novel, evidence-based approach to diagnosing acute dizziness and vertigo, Neurologic Clinics 33 (3) (2015) 577–599.

[46] J. C. Knight, Safety critical systems: Challenges and directions, Proceedings of the 24th International Conference on Software Engineering (2002) 547–550.

[47] E. M. Clarke, O. Grumberg, D. Peled, Model Checking, MIT Press, 1999.

[48] T. Nipkow, L. C. Paulson, M. Wenzel, Isabelle/HOL: A Proof Assistant for Higher-Order Logic, Springer, 2002.

[49] P. Cousot, R. Cousot, Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints, Proceedings of the 4th ACM SIGACT-SIGPLAN Symposium (1977) 238–252.

[50] G. Katz, C. Barrett, D. L. Dill, K. Julian, M. J. Kochenderfer, Reluplex: An efficient smt solver for verifying deep neural networks, in: International Conference on Computer Aided Verification, 2017, pp. 97–117.

[51] J. Cohen, E. Rosenfeld, Z. Kolter, Certified adversarial robustness via

randomized smoothing, International Conference on Machine Learning (2019) 1310–1320.

[52] G. Katz, D. A. Huang, D. Ibeling, K. Julian, C. Lazarus, R. Lim, P. Shah, S. Thakoor, H. Wu, A. Zeljić, et al., Verification of deep neural networks, arXiv preprint arXiv:2006.08000 (2020).

[53] U.S. Food and Drug Administration, Good machine learning practice for medical device development: Guiding principles, `https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development` (2021).

[54] European Commission, Medical device regulation (eu) 2017/745 (2017).

[55] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, in: Advances in Neural Information Processing Systems, 2017, pp. 5574–5584.