Disclaimer: This is not the final version of the article. Changes may occur when the manuscript is published in its final format.

A Markov–Volterra Queuing Framework with AI-Informed Priority for Multi-Class Multi-Server Emergency Departments: Mathematical Analysis Simulation-Based Evaluation and Preliminary Non-Discharge Prediction

Abdela Atisso Mohammed[1,*]; Natesan Thillaigovindan[1] & Endalkachew Teshome Ayele [1]

[1] Department of Mathematics, Arba Minch University, Arba Minch, Ethiopia,

*Corresponding Author Email: abdela_atisso@yahoo.com

Abstract

Emergency departments (EDs) operate under stochastic patient arrivals, heterogeneous clinical needs, variable service requirements, and limited resources. Classical Markovian queuing models often fail to capture memory-dependent service dynamics and time-varying operating conditions. This study develops a theoretical AI-integrated multiclass, multiserver queuing framework that combines accumulating-priority scheduling, Markov-modulated operating conditions, heterogeneous servers, and Volterra integro-differential equations (VIDEs) to represent memory-dependent service dynamics. The mathematical formulation establishes sufficient conditions for the existence, uniqueness, convergence, and finite-horizon boundedness of the queuing solution. An AI-based predictive model is incorporated as a decision-support layer that estimates the probability of non-discharge; this probability enters the priority function as an operational signal rather than a clinical decision variable. Because the available dataset contains 115 ED records but only 35 with documented disposition outcomes, the empirical analysis is restricted to predicting non-discharge probability using logistic regression, random forest, and gradient boosting under leave-one-out cross-validation (LOOCV). Random forest achieved the best overall performance (accuracy, 80.0%; balanced accuracy, 74.82%; ROC-AUC, 0.7391). However, given the small labeled sample, these results should be interpreted cautiously; confidence intervals and calibration are reported to characterize uncertainty. The framework provides a mathematically structured approach to integrating predictive information with memory-dependent queue dynamics, stochastic environmental conditions, and heterogeneous resource allocation. The empirical findings constitute preliminary proof-of-concept evidence, and further validation using larger longitudinal ED datasets is required before clinical deployment.

Keywords: multi-class multi-server queuing, accumulating priority scheduling, Markov-modulated arrivals, Volterra integro-differential equations, non-Markovian queues, server heterogeneity, Lyapunov stability, emergency healthcare, AI decision support

1 Introduction

1.1 Research Flow and Research Steps

This study follows a structured five-stage research flow. Stage 1 – Problem formulation: We identify the methodological gap in existing queueing models, which typically address only a subset of ED characteristics (memory, stochastic modulation, heterogeneity, dynamic priority, and AI prediction). Stage 2 – Theoretical development: We formulate a Markov-modulated Volterra integro-differential queuing model with heterogeneous servers and accumulating-priority scheduling, and we derive sufficient conditions for existence, uniqueness, convergence, and finite-horizon boundedness. Stage 3 – Algorithm design: We develop a numerical solution algorithm that integrates Picard successive approximation, finite-difference discretization, Markov-state updating, and dynamic resource allocation. Stage 4 – AI predictive component: We train and validate three classifiers (logistic regression, random forest, gradient boosting) to estimate non-discharge probability under LOOCV, and we incorporate the predicted probability into the priority function. Stage 5 – Computational evaluation: We compare the proposed policy against five benchmarks, conduct statistical tests, and perform sensitivity analyses. Importantly, we clearly distinguish the general theoretical framework from the empirically validated AI component and from the simulation-based computational implementation.

Healthcare emergency systems are increasingly challenged by growing patient demand, limited medical resources, heterogeneous treatment requirements, and substantial uncertainty in service operations. Emergency departments, outpatient clinics, intensive care units, and other healthcare facilities are mostly subject to congestion resulting from uncontrollable patient arrivals, variable service durations, and fluctuations in resource availability. These factors contribute to increased waiting times, inefficient resource utilization, increased operational costs, and reduced quality of healthcare delivery [1, 2, 3, 4].

Queuing theory provides one of the principal frameworks for analysing congestion and supporting operational decision-

making in healthcare systems. Classical queueing models, including the M/M/c, M/G/c, and GI/M/c models, have been widely used to evaluate waiting times, queue lengths, server utilization, and patient throughput [5, 6, 7, 8]. These models generally rely on Markovian assumptions, particularly exponentially distributed inter-arrival and service times, implying that future system evolution depends only on the current system state [8, 9, 10, 11]. Although this assumption simplifies mathematical analysis, such models often fail to represent healthcare systems accurately because treatment durations and patient pathways frequently exhibit temporal dependence and substantial variability [12, 13, 14].

Non-Markovian queuing models provide greater modeling flexibility by allowing general service-time distributions and memory-dependent dynamics. Such models are applied to healthcare systems, communication networks, and manufacturing systems where exponential assumptions are not appropriate [12, 13, 15, 16]. The mathematical analysis of non-Markovian queueing systems is more challenging than that of classical Markovian models because analytical solutions are rarely available.

Volterra integro-differential equations (VIDEs) provide a natural mathematical framework for representing systems with memory. Unlike ordinary differential equations, Volterra equations explicitly incorporate historical information through integral operators, allowing present queue dynamics to depend on accumulated past behavior. This property makes VIDEs particularly suitable for modeling healthcare service processes in which treatment durations and queue evolution exhibit memory effects [17, 18, 19, 20, 21, 22, 23]. The mathematical theory of VIDEs provides a rigorous basis for investigating existence, uniqueness, stability, and numerical approximation of memory-dependent systems [18, 19, 21].

Dynamic scheduling has become an important strategy for improving healthcare operations by allocating limited resources according to patient priority, waiting time, and system congestion. Previous studies have investigated accumulating priority queues, multi-class scheduling, dynamic server allocation, and healthcare appointment scheduling under various operational settings [24, 25, 26, 27, 28, 29, 30]. These studies show that adaptive scheduling policies can improve operational efficiency while maintaining fairness among patient classes.

The contributions of this study are as follows. First, it develops an integrated mathematical framework combining Volterra integro-differential equations, multi-class non-Markovian queuing systems, and Markov-modulated stochastic processes. Second, it incorporates a dynamic priority scheduling formulation into the memory-dependent queuing model. Third, it establishes the theoretical properties of the proposed formulation through existence, uniqueness, and convergence analysis and illustrates its computational applicability using numerical experiments. Lastly, the study extends the Markov-modulated Volterra queuing framework by incorporating an AI-based predictive decision support layer that estimates patient and system-level operational characteristics. These predictions inform, but do not autonomously determine, dynamic priority and resource-allocation decisions.

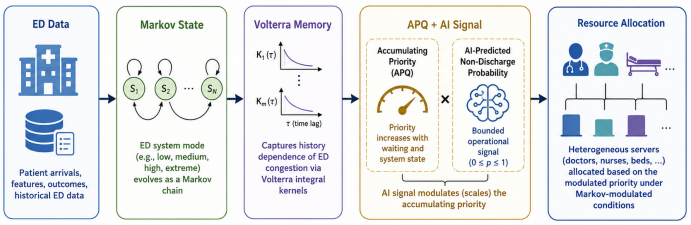

Figure 1: Intuitive schematic of the proposed framework

Figure 1 provides an intuitive schematic of the proposed framework. It illustrates how the AI-predicted non-discharge probability acts as a bounded operational signal that modulates the accumulating priority, which in turn dictates the allocation of heterogeneous servers under Markov-modulated conditions.

1.2 Research Problem

Emergency-department queuing systems simultaneously exhibit heterogeneous patient classes, time-varying demand, heterogeneous service capacity, dynamic priority requirements, and dependence on previously accumulated workload. Existing queueing formulations often address only a subset of these characteristics. In particular, classical Markovian models may not represent memory-dependent service processes, whereas models incorporating memory effects frequently do not simultaneously account for stochastic environmental modulation and adaptive resource allocation. AI-based approaches can provide useful predictive information, but prediction models considered independently of queue dynamics do not explicitly represent waiting-time accumulation, heterogeneous servers, or memory-dependent service processes.

The present study addresses this methodological gap by developing a unified AI-enhanced Markov–Volterra queuing framework in which predictive information is incorporated into an accumulating-priority resource-allocation mechanism. Importantly, AI is treated as a decision-support component rather than as an autonomous clinical decision-maker.

The specific objectives are:

1. development of a multi-class, heterogeneous-server Volterra queueing formulation for memory-dependent service dynamics;

2. incorporation of Markov-modulated operating conditions to represent stochastic changes in patient demand and resource availability;

3. formulation of an accumulating-priority mechanism that incorporates waiting time, clinical severity, resource requirements, and predictive operational information;

4. derivation of sufficient conditions for existence, uniqueness, convergence, and boundedness of the proposed Volterra formulation.

1.3 Scope of the AI Component

The proposed framework is designed to accommodate predictive information on future arrival demand, service requirements, and resource pressure. However, the available empirical dataset did not contain sufficiently complete longitudinal observations to train and validate predictive models for these time-dependent quantities. Therefore, the empirical AI component of the present study is restricted to binary prediction of patient disposition, specifically the probability of non-discharge.

The predicted non-discharge probability is incorporated into the accumulating-priority function as an operational decision-support signal. It is not interpreted as a clinical prediction, triage recommendation, treatment decision, or discharge decision. The arrival-rate, service-rate, and resource-pressure prediction terms are retained as components of the general theoretical framework and should not be interpreted as empirically validated AI outputs in the present study.

2 Preliminaries and Review of Related Literature

This section provides a structured review of the literature underpinning the proposed AI-integrated Markov–Volterra queueing framework. The review is organized around five interconnected themes: (i) classical and non-Markovian queueing in healthcare; (ii) Volterra integro-differential equations for memory-dependent systems; (iii) Markov-modulated queueing systems; (iv) dynamic priority and accumulating-priority scheduling; and (v) AI-based predictive decision support in emergency departments. For each theme, we discuss key contributions, identify limitations, and explain how the present study addresses the identified gap. The section concludes with a research gap summary and a novelty comparison table.

2.1 Classical and Non-Markovian Queueing in Healthcare

Classical queuing models—including M/M/1, M/M/c, M/G/c, and GI/M/c—have long provided the mathematical foundation for analysing congestion, resource allocation, and operational performance in healthcare systems. Early studies primarily employed Markovian models because their analytical tractability allows closed-form expressions for important performance measures such as queue length, waiting time, server utilization, and system throughput. These models have been extensively applied to emergency departments (EDs), outpatient clinics, intensive care units, and appointment scheduling problems to evaluate healthcare service efficiency [31].

However, classical models depend on exponential inter-arrival and service-time assumptions that imply the future evolution of the system relies only on its current state. As noted by [32], Markovian queueing models provide a useful baseline but often fail to capture the complexity of real healthcare systems where service times exhibit temporal dependence and substantial variability. Recent studies have demonstrated that relaxing exponential assumptions improves the realism of healthcare queue models, particularly when service processes exhibit long-tailed or highly variable distributions.

Non-Markovian queueing models address these limitations by allowing general service-time distributions and memory-dependent dynamics. [33] developed a stochastic monotonicity approach for a non-Markovian priority retrial queue that serves two types of customers, demonstrating that priority-based systems with retrials can be analysed using monotonicity properties of the underlying Markov chain. Similarly, [34, 35] analysed a multi-station manufacturing system using both theoretical and simulation approaches, showing the importance of non-Markovian assumptions in complex service networks.

Despite these advances, existing non-Markovian healthcare queueing models rarely simultaneously account for heterogeneous servers, Markov-modulated operating conditions, and dynamic priority scheduling. The present study addresses this gap by integrating these elements into a unified Volterra-based framework.

2.2 Volterra Integro-Differential Equations for Memory-Dependent Systems

Volterra integro-differential equations (VIDEs) provide a natural mathematical framework for representing systems with memory. Unlike ordinary differential equations, Volterra equations explicitly incorporate historical information through integral operators, allowing present system dynamics to depend on accumulated past behaviour. This property makes VIDEs particularly suitable for modelling healthcare service processes in which treatment durations and queue evolution exhibit memory effects.

Recent research has established theoretical properties of VIDEs, including existence, uniqueness, stability, and numerical approximation under various classes of kernel functions. As shown in recent fluid-limit analyses, deterministic fluid models can accurately describe transient behaviour in large-scale, non-Markovian, time-varying systems when the Volterra structure captures admission control and boundary effects. The Pollaczek–Khinchin integral equation—a classical result in queuing theory—demonstrates that the stationary distribution of waiting time satisfies a linear Volterra integral equation of the second kind.

However, relatively few studies have integrated VIDE formulations with healthcare queueing models, and even fewer have examined their interaction with stochastic scheduling environments. The present study addresses this gap by embedding a Volterra memory kernel within a Markov-modulated multi-class

queueing model, thereby capturing both memory-dependent service dynamics and stochastic environmental variability.

2.3 Markov-Modulated Queueing Systems

Markov-modulated queueing models capture stochastic fluctuations by allowing arrival and service parameters to evolve according to an underlying continuous-time Markov chain (CTMC). Previous studies have successfully applied Markov-modulated models to communication systems, manufacturing processes, and service networks. In healthcare applications, Markov modulation has been employed to represent varying patient demand and resource availability. For example, recent work has used continuous-time Markov chains to model patient flow in emergency departments as a time-dependent queuing network, optimising personnel assignment to achieve waiting-time objectives [36].

[32] analysed a finite-capacity multi-server Markovian queueing model with Bernoulli feedback, synchronous multiple vacation policy, and customers' impatience (balking and reneging), demonstrating how Markov-modulated models can incorporate realistic operational features such as server vacations and customer abandonment. Their work highlights the importance of capturing stochastic environmental changes in service systems.

Despite these contributions, Markov-modulated queuing models commonly assume memoryless service processes and do not account for the memory-dependent dynamics that characterise many healthcare service operations. The present study addresses this limitation by combining Markov modulation with Volterra memory kernels, thereby representing both environmental stochasticity and historical dependence.

2.4 Dynamic Priority and Accumulating-Priority Scheduling

Scheduling policies play a central role in improving healthcare service performance by determining how limited medical resources are allocated among competing patient classes. Classical priority disciplines include fixed-priority, non-preemptive priority, preemptive priority, and accumulating priority queues (APQs).

In APQ systems, the priority of a customer increases as a function of waiting time to prevent starvation of low-priority customers. [24] derived waiting-time distributions for the multi-server APQ with Poisson arrivals and exponential service times, providing the theoretical foundation for APQ applications in healthcare. Recent work has extended APQ to heterogeneous servers, recognising that different servers may have different service rates. [34] proposed a prioritized scheduling scheme for Internet of Things (IoT)-based healthcare systems that provides service differentiation based on delay sensitivity and packet size, demonstrating that adaptive scheduling can improve performance over first-come, first-served (FCFS)and shortest job first(SJF).

However, many existing scheduling formulations assume either Markovian queue dynamics or simulation-based evaluation. Analytical frameworks that simultaneously consider memory effects, stochastic environmental variability, and dynamic scheduling remain relatively limited. The present study addresses this gap by incorporating an accumulating-priority mechanism within a Markov-modulated Volterra queueing model, enabling dynamic adaptation to both waiting-time accumulation and stochastic environmental changes.

2.5 AI-Based Predictive Decision Support in Emergency Departments

Recent advances in artificial intelligence have enabled emergency-department systems to predict short-horizon patient demand, resource requirements, service duration, and patient-flow conditions [36]. In operational settings, the principal value of AI is not to replace clinical judgment but to transform currently observed information into forecasts that can support timely resource-allocation decisions [37].

[38] conducted a narrative review of AI-driven triage systems in emergency departments, identifying substantial benefits including improved patient prioritisation, reduced wait times, and optimised resource allocation. However, the review also identified significant challenges such as data quality issues, algorithmic bias, clinician trust, and ethical concerns that remain barriers to widespread adoption. Their findings emphasise the need for careful validation of AI components before clinical deployment.

[36] explored use cases of machine learning in queuing theory based on GI/G/K systems, demonstrating how machine learning(ML) can be used for queue length estimation and performance prediction in complex queuing systems. However, predictive models considered independently of queueing dynamics do not explicitly account for accumulated waiting time, stochastic environmental changes, server heterogeneity, or memory-dependent service processes [39].

The present study addresses this gap by treating AI as a predictive decision-support layer embedded within a Markov-modulated, Volterra-based multi-class queue. The AI component provides operational forecasts—specifically, the probability of non-discharge—that inform the priority and allocation mechanisms, while clinical decisions remain outside the automated scheduling layer.

2.6 Relationship to Prior Work by the Authors

[10] developed an emergency healthcare priority-based queueing model with dynamic programming. That work focused on optimising priority scheduling under Markovian assumptions and did not incorporate memory-dependent service dynamics, Volterra kernels, heterogeneous server assignment, or AI-based prediction. The present study extends that line of research by (i) introducing Volterra integro-differential equations to cap-

ture memory effects, (ii) embedding the queue in a Markov-modulated environment, (iii) formulating a heterogeneous-server allocation problem, and (iv) integrating an AI-based non-discharge probability as an operational priority signal. While recent studies have explored AI in healthcare scheduling [27, 28, 38], they often treat AI predictions as standalone triage tools. In contrast, this study positions the AI layer strictly as a bounded decision-support signal embedded within an accumulating priority queue, ensuring that clinical judgment remains paramount while operational efficiency is optimized.

Table 1 provides a systematic comparison of representative studies against the proposed framework across eight dimensions. As shown, the proposed framework is the first to simultaneously integrate multi-class patient heterogeneity, heterogeneous servers, non-Markovian memory-dependent service dynamics via Volterra kernels, Markov-modulated operating conditions, dynamic accumulating priority, AI-based predictive decision support, and ED application context. This integration architecture constitutes the principal methodological contribution of the present study.

2.7 Research Gap Summary and Novelty Comparison

The existing literature demonstrates significant progress in the development of queueing models, Volterra integro-differential equations, Markov-modulated stochastic systems, dynamic scheduling policies, and AI-based predictive tools. However, these research directions have largely evolved independently [32, 33]. Table 1 provides a systematic comparison of representative studies against the proposed framework across eight dimensions.

2.8 Preliminaries: Mathematical Foundations

This subsection provides the mathematical preliminaries for dynamic priority scheduling, cost minimisation, and modelling of non-Markovian and Markov-modulated systems using VIDEs.

2.8.1 Multi-Class Multi-Server Queues

Consider a system with C servers and N customer classes. Let $X_i(t)$ denote the number of class i customers in the system at time t, μ_{ij} denote the service rate of server j for class i, and $\lambda_i(t)$ denote the arrival rate of class i customers. For multi-class queues, let $K = \{1, 2, \ldots, N\}$ denote customer classes, where each class i may have distinct service rates and priority parameters.

2.8.2 Accumulating Priority Queues (APQ)

In APQ systems, the priority of a customer increases as a function of waiting time to prevent starvation of low-priority customers. APQ systems allow for dynamic prioritisation, which is crucial in healthcare settings to balance fairness and efficiency.

2.8.3 Markov-Modulated Queuing Systems

A Markov-modulated queue is a queueing system whose parameters (arrival or service rates) change according to a CTMC with a finite state space $S = \{1, 2, \ldots, M\}$. Such modelling allows capturing time-varying or stochastic environmental effects, which are typical in emergency departments.

2.8.4 Non-Markovian Queues

Non-Markovian queues are those in which inter-arrival or service times do not follow exponential distributions. These systems exhibit memory-dependent dynamics, requiring more general modeling approaches. The mathematical theory of VIDEs provides a rigorous basis for investigating existence, uniqueness, stability, and numerical approximation of memory-dependent systems.

3 Materials and Methods

This section presents the Materials and Methods of the study. It is organised so that each methodological element can be located and checked independently. Section 3.1 describes the study setting, the data source and design, the three levels at which the work is conducted, and the statement on clinical trial registration; Section 3.2 sets out the system description, the modeling assumptions, and the queuing formulation with its memory kernel; Section 3.3 presents the analytical results on existence, uniqueness, convergence, and stability; Section 3.4 introduces the Markov modulation of the operating environment; Section 3.5 formulates the resource-allocation problem and specifies the optimization procedure implemented in this study; Section 3.6 describes the numerical algorithm together with its complexity and scalability; Section 3.7 documents the development and validation protocol of the predictive component, including the predictor set, pre-processing procedures, feature-selection approach, hyper-parameter settings, and leakage controls; Section 3.8 reports the parameter estimation, the benchmark models, the performance measures, the statistical procedure, and the sensitivity design; and Section 3.9 outlines the ethical approvals, data governance, and human rights considerations that apply to the study.

3.1 Study Setting and Data Source

The empirical component of the study used emergency-department records obtained from Nigist Eleni Mohammed Memorial Comprehensive Specialized Hospital (NEMMCSH), Ethiopia. The study database contained 115 patient records collected between 19 December 2025 and 19 January 2026. Available variables included patient characteristics and selected emergency-department operational information. This is a retrospective, non-interventional analysis of routinely collected hospital data and is therefore not registered as a clinical trial.For the empirical AI analysis, the outcome variable was patient disposition. Complete disposition information was available for 35

Table 1: Novelty comparison of prior studies against the proposed framework.

Study	Multi-class	Multi-server	Non-Markovian	Volterra Memory	Markov Modulation	Dynamic Priority	AI Integration	ED Application
Boualem & Touche (2021)	yes	no	yes	no	no	yes	no	no
Bouchentouf et al. (2021)	no	yes	no	no	yes	no	no	no
Boualem et al. (2017)	no	yes	no	no	no	no	no	no
Stanford et al. (2014)	yes	yes	no	no	no	yes	no	yes
Asingwire et al. (2021)	yes	yes	no	no	no	yes	no	yes
Efrosinin et al. (2025)	no	yes	no	no	no	no	yes	no
Da'Costa et al. (2025)	no	no	no	no	no	no	yes	yes
This study	yes	yes	yes	yes	yes	yes	yes	yes

of the 115 records. These observations comprised 23 discharge cases and 12 non-discharge cases. The remaining records were not used for supervised classification because the required outcome label was unavailable.

The limited availability of complete longitudinal operational observations imposes an important distinction between the theoretical and empirical components of the study. The proposed queuing framework is formulated generally for class-specific arrival processes, service requirements, memory-dependent service dynamics, and heterogeneous healthcare resources. However, the available empirical dataset was not sufficiently complete to estimate all time-dependent queuing parameters directly. Consequently, the empirical data were used for patient-population characterization, preliminary AI prediction, and model calibration where the required observations were available, while the queuing framework was evaluated computationally under explicitly stated parameter assumptions.

3.1.1 Three Levels of the Study

1. Theoretical framework: The general Markov-modulated Volterra queuing model with heterogeneous servers, dynamic priority, and AI-augmented priority. This framework permits prediction of arrival rates, service requirements, and resource pressure.

2. Computational/simulation implementation: The numerical algorithm and simulation experiments use parameter values that are either estimated from hospital data where available, assumed for unavailable quantities, or generated from the Markov chain. The queueing performance results are computational, not directly estimated from longitudinal patient trajectories.

3. Empirically validated AI component: Due to incomplete longitudinal records, the only empirically validated AI output is the probability of non-discharge, trained on 35 labeled observations. Arrival-rate, service-rate, and resource-pressure predictions remain part of the theoretical framework and are not empirically validated in this study.

3.2 System Description and Mathematical Formulation

This section presents the analytical method developed to analyse multi-class, multi-server healthcare queuing systems operating under memory-dependent service dynamics and stochastic environmental variability. The healthcare system is modeled as a queuing network consisting of N patient classes served by C parallel healthcare providers. The patient classes represent heterogeneous groups requiring different levels of medical attention, such as emergency, urgent, and non-urgent cases. Patients belonging to different classes are characterized by distinct arrival rates, service requirements, and scheduling priorities. Multiple servers operate simultaneously, allowing several patients to receive treatment concurrently while sharing limited healthcare resources.

3.2.1 Model Assumptions

The proposed theoretical formulation is established under the following assumptions.

Assumption A1. The healthcare system consists of C heterogeneous servers serving N heterogeneous patient classes according to a dynamic priority scheduling policy.

Assumption A2. Patients arrive according to class-dependent stochastic arrival processes whose rates vary with the current environmental state. The arrival-rate function for patient class i is denoted by $\lambda_i^{(m)}(t)$, where m represents the current Markov state.

Assumption A3. Service times follow general probability distributions and are therefore not restricted to the exponential distribution. The corresponding service-rate function for class i in environmental state m is represented by $\mu_i^{(m)}(t)$.

Assumption A4. The memory effects associated with healthcare service dynamics are represented through a continuous Volterra kernel $K(t,\tau)$, satisfying

$$K(t,\tau) \in C([0,T] \times [0,T])$$

and

$$|K(t,\tau)| \leq M$$

for some finite positive constant M.

Assumption A5. The kernel function satisfies the Lipschitz condition

$$|K(t,\tau) - K(s,\tau)| \leq L|t-s|,$$

where $L > 0$ denotes the Lipschitz constant.

Assumption A6. The traffic intensity satisfies

$$\rho^{(m)} = \frac{\sum_{i=1}^{N} \lambda_i^{(m)}}{\sum_{j=1}^{C} \mu_j^{(m)}} < 1,$$

ensuring that the queueing system remains stable under each operating state.

Assumption A7 (Bounded AI Prediction Error). For each class i and server j, the AI predictor generates forecasts satisfying

$$|\hat{\lambda}_i(t+h) - \lambda_i(t+h)| \le \varepsilon_{\lambda,i}, \qquad |\hat{\mu}_{ij}(t+h) - \mu_{ij}(t+h)| \le \varepsilon_{\mu,ij},$$

for $0 \le t \le T$, where the error bounds are finite.

3.2.2 Mathematical Representation of Queue Dynamics

Let $Q_i(t)$ denote the expected queue length associated with patient class i at time t.

$$\frac{dQ_i(t)}{dt} = \lambda_i(t) - D_i(t), \qquad i = 1, 2, \ldots, N, \tag{1}$$

where $\lambda_i(t)$ denotes the arrival rate of class-i patients and $D_i(t)$ represents the corresponding service completion rate.

For classical Markovian queueing systems, the service completion process is assumed to depend only on the current queue length, that is,

$$D_i(t) = \mu_i(t) Q_i(t), \tag{2}$$

which leads to the familiar fluid approximation

$$\frac{dQ_i(t)}{dt} = \lambda_i(t) - \mu_i(t) Q_i(t). \tag{3}$$

For each class i, let $x_{ij}(t) \in \{0, 1\}$ indicate the assignment of server j to class i, and let

$$a_i(t) = \sum_{j=1}^{C} x_{ij}(t)$$

be the number of servers assigned to that class. Define the in-service occupancy

$$B_i(t) = \min\Big\{ Q_i(t),\, a_i(t) \Big\},$$

so that $B_i(t)$ is the expected number of class-i patients actually occupying a server. The corresponding waiting count is $Q_i^{\mathrm{w}}(t) = Q_i(t) - B_i(t)$, whence $Q_i(t) = B_i(t) + Q_i^{\mathrm{w}}(t)$ with $0 \le B_i(t) \le \min\{Q_i(t), C\}$ and $0 \le Q_i^{\mathrm{w}}(t) \le Q_i(t)$. When $Q_i(t) \le a_i(t)$ the class is unsaturated and every patient present is in service; when $Q_i(t) > a_i(t)$ the in-service occupancy is capped at $a_i(t)$ and the excess accumulates in the waiting count. Since $B_i \le Q_i$, substituting B_i for Q_i in the service operator yields a conservative majorant, so the contraction and boundedness estimates established below continue to hold for the reduced system. Note that Q_i^{w} is a patient count.

Although Equation (3) is mathematically convenient, it neglects the memory effects frequently observed in healthcare systems. In practice, treatment duration and service completion may depend on previous diagnostic procedures, accumulated workload, and resource availability. Furthermore, service completions are generated only by patients who are actually occupying a server, not by the entire waiting queue. To account for memory-dependent effects in a well-defined multi-server queue, the effective service completion rate is represented by a Volterra integral operator acting on the in-service population:

$$D_i(t) = \int_0^t K_i(t-\tau) \mu_i(\tau) B_i(\tau)\, d\tau, \tag{4}$$

where $K_i(t-\tau)$ is the memory kernel describing the influence of previous service dynamics on the current completion process. Substituting this equation into Equation (1) yields the proposed Volterra integro-differential equation governing the queue dynamics:

$$\frac{dQ_i(t)}{dt} = \lambda_i(t) - \sum_{j=1}^{C} x_{ij}(t) \int_0^t K_i(t-\tau) \mu_{ij}(\tau) B_i(\tau)\, d\tau, \tag{5}$$

with the initial condition $Q_i(0) = Q_{i0} \ge 0$.

3.2.3 Dynamic Priority Scheduling

Patient priority is assumed to evolve continuously with waiting time rather than remaining fixed throughout the service process. To incorporate predictive information without replacing the underlying queuing dynamics, the AI component is introduced as a decision-support layer within the accumulating-priority mechanism. Let

$$\mathcal{F}_t = \sigma\{Q(s), W(s), X(s), R(s), Y(s) : 0 \le s \le t\} \tag{6}$$

denote the information available at decision epoch t. An AI predictor maps the observed information into a short-horizon forecast

$$\hat{Y}(t+h) = A_\theta(\mathcal{F}_t) = (\hat{\lambda}_1, \ldots, \hat{\lambda}_N, \hat{\mu}_{11}, \ldots, \hat{\mu}_{NC}, \hat{R}_1, \ldots, \hat{R}_N), \tag{7}$$

where $\hat{Y}(t+h)$ contains predicted class-specific demand, service requirements, and resource pressure.

Now define the predicted non-discharge(ND) probability

$$\hat{p}_i^{\mathrm{ND}}(t+h) = P\{Y_i(t+h) = 1 \mid \mathcal{F}_t\}, \tag{8}$$

where $Y_i = 1$ denotes a predicted non-discharge outcome. Then define the operational priority as

$$\begin{aligned} P_i^{\mathrm{AI}}(t) = \alpha_i + \beta_i W_i(t) + \gamma_i S_i \\ + \delta_i D_i(t+h) + \eta_i R_i(t+h) + \zeta_i \hat{p}_i^{\mathrm{ND}}(t+h). \end{aligned} \tag{9}$$

where:

- $W_i(t)$: waiting time;

- S_i: severity;
- $D_i(t+h)$: predicted demand pressure;
- $R_i(t+h)$: predicted resource requirement;
- $\hat{p}_i^{\mathrm{ND}}(t+h)$: predicted non-discharge probability.

This formulation enables the scheduling policy to adapt dynamically to prolonged waiting times while reducing the possibility of indefinite postponement for lower-priority patients. Thus, scheduling decisions depend jointly on patient urgency, accumulated waiting time, and current system conditions.

Proposition 1 (Monotonicity of AI-Enhanced Priority). If $\beta_i > 0$, then

$$\frac{\partial P_i^{\mathrm{AI}}(t)}{\partial W_i(t)} = \beta_i > 0. \tag{10}$$

Therefore, the AI layer preserves the accumulating-priority property: increasing waiting time cannot decrease the priority of class i.

Proposition 2 (Bounded Prediction Perturbation). If $|\hat{D}_i(t+h) - D_i(t+h)| \le \varepsilon_{D,i}$ and $|\hat{R}_i(t+h) - R_i(t+h)| \le \varepsilon_{R,i}$, then

$$|P_i^{\mathrm{AI}}(t) - P_i^*(t)| \le \delta_i \varepsilon_{D,i} + \eta_i \varepsilon_{R,i}. \tag{11}$$

Thus, bounded prediction errors produce bounded perturbations of the priority index.

3.3 Existence, Uniqueness, and Stability Analysis

The theoretical framework permits AI-based forecasts of future arrival demand, service requirements, and resource pressure. However, because the empirical dataset did not contain sufficiently complete longitudinal observations for these outcomes, the present implementation uses AI exclusively to estimate the probability of non-discharge. This probability is incorporated as an operational signal in the dynamic priority function and does not independently determine clinical treatment, triage, or discharge decisions.

3.3.1 Theoretical Assumptions and Function Space

To establish the well-posedness of the proposed model, we first define the function space in which the solution is sought and state the assumptions required for the subsequent theoretical analysis. Let

$$X = C([0,T], \mathbb{R}^N) \tag{12}$$

denote the Banach space of continuous vector-valued functions on the finite interval $[0,T]$, equipped with the supremum norm

$$\|Q\|_\infty = \sup_{0 \le t \le T} \max_{1 \le i \le N} |Q_i(t)|. \tag{13}$$

The unknown solution of the Volterra integro-differential equation is therefore sought in the complete metric space $(X, \|\cdot\|_\infty)$. For each patient class $i = 1, 2, \ldots, N$, the queue dynamics are governed by

$$\frac{dQ_i(t)}{dt} = \lambda_i(t) - \int_0^t K_i(t-\tau)\, \mu_i(\tau)\, Q_i(\tau)\, d\tau, \quad \text{with } Q_i(0) = Q_{i0} \ge 0. \tag{14}$$

The following assumptions are imposed throughout this section.

Assumption H1 (Arrival Process). For every patient class i, the arrival-rate function $\lambda_i : [0,T] \to \mathbb{R}_+$ is continuous and uniformly bounded. That is, there exists a constant $\Lambda_i > 0$ such that

$$0 \le \lambda_i(t) \le \Lambda_i, \qquad \forall t \in [0,T]. \tag{15}$$

This assumption ensures that patient arrivals remain finite over the planning horizon.

Assumption H2 (Service Process). The service-rate function $\mu_i : [0,T] \to \mathbb{R}_+$ is continuous and satisfies

$$0 < \mu_i(t) \le \overline{\mu}_i, \qquad \forall t \in [0,T], \tag{16}$$

where $\overline{\mu}_i$ is a finite positive constant.

Assumption H3 (Memory Kernel). The Volterra kernel $K_i : [0,T] \times [0,T] \to \mathbb{R}$ is continuous and satisfies

$$|K_i(t,\tau)| \le M_K, \qquad 0 \le \tau \le t \le T, \tag{17}$$

for some constant $M_K > 0$, and

$$K_i(t,\tau) \ge 0, \qquad 0 \le \tau \le t \le T, \tag{18}$$

ensuring that historical queue states contribute non-negatively to the service process.

Assumption H4 (Lipschitz Continuity). The kernel satisfies the Lipschitz condition with respect to its first argument. Specifically, there exists a constant $L_K > 0$ such that

$$|K_i(t,\tau) - K_i(s,\tau)| \le L_K |t-s|, \qquad \forall s,t,\tau \in [0,T]. \tag{19}$$

This assumption guarantees the regularity required for the contraction argument used in the existence proof.

Assumption H5 (Initial Condition). The initial queue lengths satisfy

$$Q_i(0) = Q_{i0} \ge 0, \qquad i = 1, 2, \ldots, N, \tag{20}$$

where Q_{i0} is prescribed.

Assumption H6 (Operational Admissibility). For every environmental state $m \in \{1, \ldots, M\}$, the traffic intensity, defined as the ratio of total arrival rate to total service capacity, satisfies

$$\rho^{(m)} = \frac{\sum_{i=1}^{N} \lambda_i^{(m)}}{C \sum_{j=1}^{} \mu_j^{(m)}} < 1, \tag{21}$$

where $\lambda_i^{(m)}$ and $\mu_j^{(m)}$ are measured in the same units (patients per unit time), so that $\rho^{(m)}$ is a dimensionless load factor. This

condition is an operational admissibility condition; it is not required for the finite-horizon boundedness result of Theorem 3.3.

Assumptions H1–H6 provide the analytical framework required for the subsequent theoretical analysis. In particular, the continuity and boundedness of the arrival rates, service rates, and memory kernel guarantee that the associated Volterra integral operator is well defined on the Banach space X. The Lipschitz condition forms the basis for establishing the contraction property of the operator, while the stability assumption ensures that the queueing system operates within the admissible region. These assumptions will be invoked throughout the proofs of existence, uniqueness, convergence, and stability presented in the following subsections.

3.3.2 Formal Scope of the Theoretical Analysis

Definition (Reduced Volterra Subsystem). The theoretical results in Section 3.3 concern the reduced Volterra subsystem, defined for each class $i = 1, \dots, N$ by

$$\frac{dQ_i(t)}{dt} = \lambda_i(t) - \int_0^t K_i(t-\tau)\mu_i(\tau)Q_i(\tau)\,d\tau, \qquad Q_i(0) = Q_{i0} \geq 0, \tag{22}$$

where the service term integrates over the total queue length Q_i rather than over the in-service population $B_i(t)$. This subsystem is obtained from the full model (1) by (i) replacing $B_i(t)$ with $Q_i(t)$, and (ii) dropping the server-allocation variables $x_{ij}(t)$. It is a well-posed fluid approximation of the memory-augmented multi-class queue, but it is not equivalent to the heterogeneous multi-server model presented in Equation (1).

3.3.3 Definition of the Volterra Integral Operator

To establish the existence and uniqueness of solutions to the Volterra integro-differential equation, we first reformulate the differential equation as an equivalent nonlinear Volterra integral equation. This transformation enables the application of functional analytic techniques, particularly Banach's Fixed Point Theorem. Consider the initial value problem

$$\begin{aligned} \frac{dQ_i(t)}{dt} &= \lambda_i(t) - \int_0^t K_i(t-\tau)\mu_i(\tau)Q_i(\tau)\,d\tau, \\ &\text{with } Q_i(0) = Q_{i0}, \end{aligned} \tag{23}$$

where the functions satisfy Assumptions H1–H6. Integrating Equation (23) over the interval $[0, t]$ yields

$$\begin{aligned} Q_i(t) = Q_{i0} &+ \int_0^t \lambda_i(s)\,ds \\ &- \int_0^t \int_0^s K_i(s-\tau)\mu_i(\tau)Q_i(\tau)\,d\tau\,ds. \end{aligned} \tag{24}$$

Equation (24) is equivalent to the original Volterra integro-differential equation and forms the basis of the subsequent theoretical analysis. Define the nonlinear operator $\mathcal{T} : X \to X$ by

$$\begin{aligned} (\mathcal{T}Q_i)(t) = Q_{i0} &+ \int_0^t \lambda_i(s)\,ds \\ &- \int_0^t \int_0^s K_i(s-\tau)\mu_i(\tau)Q_i(\tau)\,d\tau\,ds, \end{aligned} \tag{25}$$

where $Q = (Q_1, \dots, Q_N) \in X$.

The existence and uniqueness analysis depends on establishing that the operator $\mathcal{T}$ is well defined on the Banach space X. This result is stated in the following lemma.

Lemma 3.1 *Under Assumptions H1–H6, the operator $\mathcal{T} : X \to X$ defined by Equation (25) is well defined. That is, for every $Q \in X$, the function $\mathcal{T}Q$ also belongs to X.*

Proof. Let $Q \in X$. Since Q is continuous on the compact interval $[0, T]$, there exists a constant $M_Q > 0$ such that

$$|Q_i(t)| \leq M_Q, \qquad \forall t \in [0, T].$$

By Assumption H1, the arrival-rate function $\lambda_i(t)$ is continuous and bounded. Therefore,

$$\int_0^t \lambda_i(s)\,ds$$

is continuous with respect to t. Similarly, Assumptions H2 and H3 imply that both the service-rate function $\mu_i(t)$ and the kernel $K_i(t, \tau)$ are continuous and bounded. Since the product of continuous functions is continuous, the integrand

$$K_i(s-\tau)\mu_i(\tau)Q_i(\tau)$$

is continuous on the triangular region

$$0 \leq \tau \leq s \leq T.$$

Therefore, by the continuity of parameter-dependent integrals, the double integral

$$\int_0^t \int_0^s K_i(s-\tau)\mu_i(\tau)Q_i(\tau)\,d\tau\,ds$$

is also continuous on $[0, T]$. Therefore,

$$\begin{aligned} (\mathcal{T}Q_i)(t) = Q_{i0} &+ \int_0^t \lambda_i(s)\,ds \\ &- \int_0^t \int_0^s K_i(s-\tau)\mu_i(\tau)Q_i(\tau)\,d\tau\,ds \end{aligned}$$

is the sum of continuous functions and is itself continuous. Hence,

$$\mathcal{T}Q \in C([0, T], \mathbb{R}^N) = X.$$

Therefore, the operator $\mathcal{T}$ maps the Banach space X into itself and is well defined. □

Lemma 3.1 establishes that the Volterra integral operator preserves continuity of the queue-length functions. As a result, the existence problem may be studied entirely within the Banach space X. The next result establishes that the operator is bounded, which is an essential step toward proving its contraction property.

Lemma 3.2 (Boundedness on Bounded Subsets) *Suppose that Assumptions H1–H5 hold. For every $R > 0$, there exists a constant $C_R > 0$ such that*

$$\|Q\|_\infty \le R \quad \Longrightarrow \quad \|\mathcal{T}Q\|_\infty \le C_R,$$

where

$$C_R = Q_{\max} + \Lambda_{\max} T + \frac{M_K \mu_{\max} T^2}{2} R,$$

with $Q_{\max} = \max_i |Q_{i0}|$, $\Lambda_{\max} = \max_i \sup_t \lambda_i(t)$, $\mu_{\max} = \max_i \sup_t \mu_i(t)$, *and* $M_K = \max_i \sup_{0 \le \tau \le t \le T} |K_i(t,\tau)|$.

Using the inequality $\frac{t^2}{2} \le \frac{T^2}{2}$, we obtain

$$|(\mathcal{T}Q)_i(t)| \le Q_{\max} + \Lambda_{\max} T + \frac{M_K \mu_{\max} T^2}{2} R.$$

Taking the supremum over $t \in [0,T]$ and over all patient classes gives

$$\|\mathcal{T}Q\|_\infty \le C_R.$$

Therefore, $\mathcal{T}$ maps bounded subsets of X into bounded subsets of X.

Lemma 3.2 shows that the Volterra integral operator maps bounded subsets of the Banach space into bounded subsets. This property guarantees that successive approximations generated by the operator remain finite and forms an important step toward establishing the contraction property proved in the next subsection.

Having established that the Volterra integral operator is well defined and bounded, we now show that it satisfies a Lipschitz condition on the Banach space X. This property is fundamental for proving the existence and uniqueness of solutions via Banach's Fixed Point Theorem.

Lemma 3.3 *Suppose that Assumptions H1–H6 hold. Then the operator $\mathcal{T} : X \to X$ defined by Equation (25) is Lipschitz continuous. If*

$$\frac{M_K \mu_{\max} T^2}{2} < 1, \tag{26}$$

then $\mathcal{T}$ is a contraction mapping on X.

Proof. Let $Q, R \in X$ be arbitrary. Using Equation (25),

$$\begin{aligned}(\mathcal{T}Q_i)(t) - (\mathcal{T}R_i)(t) \\ = -\int_0^t \int_0^s K_i(s-\tau)\mu_i(\tau) \\ \times [Q_i(\tau) - R_i(\tau)]\, d\tau ds.\end{aligned} \tag{27}$$

Taking absolute values and applying the triangle inequality,

$$\begin{aligned}|(\mathcal{T}Q_i)(t) - (\mathcal{T}R_i)(t)| \le \\ \int_0^t \int_0^s |K_i(s-\tau)||\mu_i(\tau)||Q_i(\tau) - R_i(\tau)|\, d\tau\, ds.\end{aligned} \tag{28}$$

By Assumptions H2 and H3,

$$|K_i(s-\tau)| \le M_K, \qquad |\mu_i(\tau)| \le \mu_{\max}.$$

Since

$$|Q_i(\tau) - R_i(\tau)| \le \|Q - R\|_\infty,$$

Equation (28) becomes

$$|(\mathcal{T}Q_i)(t) - (\mathcal{T}R_i)(t)| \le M_K \mu_{\max} \|Q - R\|_\infty \int_0^t \int_0^s d\tau ds.$$

Evaluating the double integral,

$$\int_0^t \int_0^s d\tau ds = \frac{t^2}{2} \le \frac{T^2}{2},$$

gives

$$|(\mathcal{T}Q_i)(t) - (\mathcal{T}R_i)(t)| \le \frac{M_K \mu_{\max} T^2}{2} \|Q - R\|_\infty. \tag{29}$$

Taking the supremum over all $t \in [0,T]$ and all patient classes,

$$\|\mathcal{T}Q - \mathcal{T}R\|_\infty \le L \|Q - R\|_\infty, \tag{30}$$

where

$$L = \frac{M_K \mu_{\max} T^2}{2}. \tag{31}$$

Hence, the operator $\mathcal{T}$ is Lipschitz continuous with Lipschitz constant L. Moreover, if

$$L < 1,$$

or equivalently,

$$\frac{M_K \mu_{\max} T^2}{2} < 1,$$

then

$$\|\mathcal{T}Q - \mathcal{T}R\|_\infty < \|Q - R\|_\infty, \tag{32}$$

which implies that $\mathcal{T}$ is a contraction mapping on the complete metric space $(X, \|\cdot\|_\infty)$. Therefore, the proof is complete. □

Lemma 3.3 establishes the contraction property of the Volterra integral operator under an explicit and verifiable condition involving the memory kernel, the maximum service rate, and the time horizon. This result provides the principal hypothesis required for applying Banach's Fixed Point Theorem in the next subsection to establish the existence and uniqueness of the solution.

Having established that the Volterra integral operator is well defined, bounded, and satisfies the contraction property, we now prove the existence and uniqueness of the solution to the Volterra integro-differential queueing model.

Theorem 3.1. *Suppose that Assumptions H1–H6 hold and that the contraction condition*

$$\frac{M_K \mu_{\max} T^2}{2} < 1 \tag{33}$$

is satisfied. Then the Volterra integro-differential equation

$$\frac{dQ_i(t)}{dt} = \lambda_i(t) - \int_0^t K_i(t-\tau)\mu_i(\tau)Q_i(\tau)\, d\tau, \qquad Q_i(0) = Q_{i0}, \tag{34}$$

admits a unique continuous solution

$$Q(t) \in C([0,T], \mathbb{R}^N).$$

Proof. From Lemma 3.1, the operator

$$\mathcal{T} : X \to X$$

is well defined, implying that every continuous function is mapped into another continuous function within the Banach space

$$X = C([0,T], \mathbb{R}^N).$$

Lemma 3.2 establishes that the operator is bounded on X, so that successive iterates generated by $\mathcal{T}$ remain bounded for all finite times. Lemma 3.3 proves that

$$\|\mathcal{T}Q - \mathcal{T}R\|_\infty \le L\|Q - R\|_\infty,$$

where

$$L = \frac{M_K \mu_{\max} T^2}{2}.$$

Under Condition (33), it follows that

$$L < 1.$$

Hence, $\mathcal{T}$ is a contraction mapping on the complete metric space

$$(X, \|\cdot\|_\infty).$$

Therefore, all hypotheses of Banach's Fixed Point Theorem are satisfied, and there exists a unique fixed point

$$Q^* \in X$$

such that

$$\mathcal{T}(Q^*) = Q^*. \tag{35}$$

Since Equation (35) is equivalent to the nonlinear Volterra integral equation obtained in Equation (24), differentiation with respect to time yields the original Volterra integro-differential equation (34). Therefore,

$$Q^*(t)$$

is the unique continuous solution of the proposed queueing model.

To establish uniqueness explicitly, suppose that there exist two solutions,

$$Q^{(1)}, Q^{(2)} \in X.$$

Since both satisfy the fixed-point equation,

$$Q^{(1)} = \mathcal{T}Q^{(1)}, \qquad Q^{(2)} = \mathcal{T}Q^{(2)},$$

we obtain

$$\|Q^{(1)} - Q^{(2)}\|_\infty = \|\mathcal{T}Q^{(1)} - \mathcal{T}Q^{(2)}\|_\infty.$$

Applying the contraction inequality gives

$$\|Q^{(1)} - Q^{(2)}\|_\infty \le L\|Q^{(1)} - Q^{(2)}\|_\infty.$$

Since

$$0 < L < 1,$$

it follows that

$$(1 - L)\|Q^{(1)} - Q^{(2)}\|_\infty \le 0.$$

Because

$$1 - L > 0,$$

the only possibility is

$$\|Q^{(1)} - Q^{(2)}\|_\infty = 0,$$

which implies

$$Q^{(1)}(t) = Q^{(2)}(t), \qquad \forall t \in [0,T].$$

Hence, the solution is unique. □

Theorem 3.1 establishes that the proposed Volterra integro-differential queueing model is analytically well posed. Under the stated assumptions, a unique continuous solution exists on the finite interval $[0,T]$, ensuring that the model admits a consistent mathematical interpretation. This result provides the theoretical foundation for the iterative numerical algorithm developed later in the paper and guarantees that the computed solution is uniquely determined by the initial condition and model parameters.

The existence and uniqueness result established in Theorem 3.1 immediately implies the convergence of the successive approximation method used to compute the solution of the Volterra integro-differential equation.

Corollary 3.1. *Suppose that the assumptions of Theorem 3.1 hold. Let the sequence* $\{Q^{(n)}\}_{n=0}^{\infty}$ *be generated by the iterative scheme*

$$Q^{(n+1)} = \mathcal{T}Q^{(n)}, \qquad n = 0, 1, 2, \ldots, \tag{36}$$

where $Q^{(0)}$ *is an arbitrary initial approximation in* X*. Then the sequence converges uniformly to the unique fixed point* Q^**, which is the unique solution of the proposed Volterra integro-differential equation.*

Proof. Since Theorem 3.1 establishes that $\mathcal{T}$ is a contraction mapping,

$$\|\mathcal{T}Q - \mathcal{T}R\|_\infty \le L\|Q - R\|_\infty,$$

where

$$0 < L < 1,$$

Banach's Fixed Point Theorem guarantees that repeated application of the operator generates a convergent sequence. For every integer $n \ge 0$,

$$Q^{(n+1)} = \mathcal{T}Q^{(n)}.$$

Hence,

$$\|Q^{(n+1)} - Q^*\|_\infty = \|\mathcal{T}Q^{(n)} - \mathcal{T}Q^*\|_\infty.$$

Using the contraction inequality,

$$\|Q^{(n+1)} - Q^*\|_\infty \le L\|Q^{(n)} - Q^*\|_\infty. \tag{37}$$

Applying Equation (37) recursively yields

$$\|Q^{(n)} - Q^*\|_\infty \le L^n \|Q^{(0)} - Q^*\|_\infty. \tag{38}$$

Since

$$0 < L < 1,$$

it follows that

$$L^n \to 0, \qquad \text{as } n \to \infty.$$

Therefore,

$$\lim_{n\to\infty} Q^{(n)} = Q^*,$$

uniformly on the interval $[0, T]$. The Picard successive approximation converges to the unique solution of the Volterra integro-differential equation. □

Corollary 3.1 establishes the theoretical convergence of the iterative numerical scheme. Equation (38) demonstrates that the approximation error decreases geometrically, implying rapid convergence whenever the contraction constant L is sufficiently smaller than unity. This result provides the theoretical justification for the numerical algorithm employed in Section 5.

The previous results establish the existence and uniqueness of solutions. We now investigate the stability of the healthcare queueing model. Stability is essential because it guarantees that small perturbations in the queue length do not lead to unbounded system behaviour.

Theorem 3.2. Suppose that Assumptions H1–H5 hold and let $Q(t)$ and $\widetilde{Q}(t)$ be two solutions of the Volterra integro-differential queueing system

$$\frac{dQ_i(t)}{dt} = \lambda_i(t) - \int_0^t K_i(t-\tau)\mu_i(\tau)Q_i(\tau)\,d\tau, \qquad Q_i(0) = Q_{i0},$$

and

$$\frac{d\widetilde{Q}_i(t)}{dt} = \lambda_i(t) - \int_0^t K_i(t-\tau)\mu_i(\tau)\widetilde{Q}_i(\tau)\,d\tau, \qquad \widetilde{Q}_i(0) = \widetilde{Q}_{i0},$$

for $i = 1, \dots, N$. Assume that

$$|K_i(t-\tau)| \le M_K, \qquad 0 < \mu_i(t) \le \overline{\mu}_i, \qquad 0 \le \tau \le t \le T.$$

Then the solution is Lyapunov stable with respect to perturbations of the initial queue lengths on every finite interval $[0, T]$. More precisely,

$$\|Q(t) - \widetilde{Q}(t)\|_\infty \le \|Q(0) - \widetilde{Q}(0)\|_\infty \exp\left(\frac{M_K\overline{\mu}T^2}{2}\right), \qquad 0 \le t \le T,$$

where

$$\overline{\mu} = \max_{1\le i\le N} \overline{\mu}_i.$$

Consequently, for every $\varepsilon > 0$, choosing

$$\delta = \varepsilon \exp\left(-\frac{M_K\overline{\mu}T^2}{2}\right)$$

ensures that

$$\|Q(0) - \widetilde{Q}(0)\|_\infty < \delta$$

implies

$$\|Q(t) - \widetilde{Q}(t)\|_\infty < \varepsilon, \qquad \forall t \in [0, T].$$

Hence, the Volterra queueing model is Lyapunov stable on the finite planning horizon $[0, T]$.

Proof. Let

$$Z_i(t) = Q_i(t) - \widetilde{Q}_i(t).$$

Since both systems have the same arrival-rate functions, subtracting their governing equations gives

$$\frac{dZ_i(t)}{dt} = -\int_0^t K_i(t-\tau)\mu_i(\tau)Z_i(\tau)\,d\tau, \qquad Z_i(0) = Q_{i0} - \widetilde{Q}_{i0}.$$

Integrating from 0 to t, we obtain

$$Z_i(t) = Z_i(0) - \int_0^t \int_0^s K_i(s-\tau)\mu_i(\tau)Z_i(\tau)\,d\tau\,ds.$$

Taking absolute values and applying the triangle inequality,

$$|Z_i(t)| \le |Z_i(0)| + \int_0^t \int_0^s |K_i(s-\tau)||\mu_i(\tau)||Z_i(\tau)|\,d\tau\,ds.$$

By the boundedness assumptions, $|K_i(s-\tau)| \le M_K$ and $|\mu_i(\tau)| \le \overline{\mu}$. Therefore,

$$|Z_i(t)| \le |Z_i(0)| + M_K\overline{\mu}\int_0^t \int_0^s |Z_i(\tau)|\,d\tau\,ds.$$

Define $z(t) = \max_{1\le i\le N} |Z_i(t)|$. Then

$$z(t) \le z(0) + M_K\overline{\mu}\int_0^t \int_0^s z(\tau)\,d\tau\,ds.$$

Using the identity $\int_0^t \int_0^s z(\tau)\,d\tau\,ds = \int_0^t (t-\tau)z(\tau)\,d\tau$, we obtain

$$z(t) \le z(0) + M_K\overline{\mu}\int_0^t (t-\tau)z(\tau)\,d\tau.$$

First Stage (Crude Bound): Since $0 \le t - \tau \le T$, we have

$$z(t) \le z(0) + M_K\overline{\mu}T\int_0^t z(\tau)\,d\tau.$$

By Grönwall's inequality,

$$z(t) \le z(0)\exp(M_K\overline{\mu}Tt).$$

This establishes finite-horizon continuous dependence.

Second Stage (Sharp Bound): A sharper estimate follows directly by applying the generalized Grönwall inequality to the convolution kernel $a(t,\tau) = M_K\overline{\mu}(t-\tau)$. This gives

$$z(t) \le z(0)\exp\left(M_K\overline{\mu}\int_0^t (t-\tau)\,d\tau\right).$$

Because $\int_0^t (t-\tau)\,d\tau = \frac{t^2}{2}$, we obtain

$$z(t) \le z(0)\exp\left(\frac{M_K\overline{\mu}t^2}{2}\right).$$

Since $t \le T$,

$$z(t) \le z(0) \exp\left(\frac{M_K \overline{\mu} T^2}{2}\right).$$

Equivalently,

$$\boxed{\|Q(t) - \widetilde{Q}(t)\|_\infty \le \|Q(0) - \widetilde{Q}(0)\|_\infty \exp\left(\frac{M_K \overline{\mu} T^2}{2}\right)}$$

for every $t \in [0, T]$.

Now let $\varepsilon > 0$ be arbitrary and choose

$$\delta = \varepsilon \exp\left(-\frac{M_K \overline{\mu} T^2}{2}\right).$$

If $\|Q(0) - \widetilde{Q}(0)\|_\infty < \delta$, then

$$\begin{aligned}\|Q(t) - \widetilde{Q}(t)\|_\infty &\le \|Q(0) - \widetilde{Q}(0)\|_\infty \exp\left(\frac{M_K \overline{\mu} T^2}{2}\right) \\ &< \delta \exp\left(\frac{M_K \overline{\mu} T^2}{2}\right) \\ &= \varepsilon.\end{aligned}$$

Thus, arbitrarily small perturbations of the initial queue-length vector remain arbitrarily small throughout the finite planning horizon. Therefore, the Volterra integro-differential queueing model is Lyapunov stable on $[0, T]$. □

Theorem 3.3 (Uniform Finite-Horizon Boundedness). *Suppose that Assumptions H1–H5 hold and let $Q(t) = (Q_1(t), \dots, Q_N(t))^\top$ be a solution of the Volterra integro-differential queueing model on $[0, T]$. Define*

$$Q_{\max} = \max_{1 \le i \le N} |Q_{i0}|, \qquad \Lambda_{\max} = \max_{1 \le i \le N} \sup_{0 \le t \le T} \lambda_i(t),$$

$$\mu_{\max} = \max_{1 \le i \le N} \sup_{0 \le t \le T} \mu_i(t), \qquad M_K = \max_{1 \le i \le N} \sup_{0 \le \tau \le t \le T} |K_i(t, \tau)|.$$

Then $Q(t)$ is uniformly bounded on every finite interval $[0, T]$. More precisely,

$$\boxed{\|Q(t)\|_\infty \le (Q_{\max} + \Lambda_{\max} T) \cosh\left(\sqrt{M_K \mu_{\max}}\, T\right), \quad 0 \le t \le T.} \tag{39}$$

Consequently, the simpler exponential bound

$$\boxed{\|Q(t)\|_\infty \le B_T := (Q_{\max} + \Lambda_{\max} T) \exp\left(\frac{M_K \mu_{\max} T^2}{2}\right)}$$

also holds for every $t \in [0, T]$. In particular, $B_T < \infty$, and therefore no queue-length trajectory can become unbounded on the finite planning horizon.

Proof. From the integral formulation of the Volterra queueing model,

$$Q_i(t) = Q_{i0} + \int_0^t \lambda_i(s)\, ds - \int_0^t \int_0^s K_i(s - \tau) \mu_i(\tau) Q_i(\tau)\, d\tau\, ds. \tag{40}$$

Taking absolute values and applying the triangle inequality gives

$$\begin{aligned}|Q_i(t)| \le |Q_{i0}| &+ \int_0^t |\lambda_i(s)|\, ds \\ &+ \int_0^t \int_0^s |K_i(s - \tau)||\mu_i(\tau)||Q_i(\tau)|\, d\tau\, ds.\end{aligned} \tag{41}$$

By Assumptions H1–H3,

$$|Q_{i0}| \le Q_{\max}, \qquad |\lambda_i(s)| \le \Lambda_{\max},$$

$$|\mu_i(\tau)| \le \mu_{\max}, \qquad |K_i(s - \tau)| \le M_K.$$

Consequently,

$$|Q_i(t)| \le Q_{\max} + \Lambda_{\max} t + M_K \mu_{\max} \int_0^t \int_0^s |Q_i(\tau)|\, d\tau\, ds. \tag{42}$$

Since $t \le T$,

$$Q_{\max} + \Lambda_{\max} t \le Q_{\max} + \Lambda_{\max} T.$$

Let

$$a = M_K \mu_{\max}$$

and define

$$Y(t) = \max_{1 \le i \le N} |Q_i(t)|.$$

Then

$$Y(t) \le C + a \int_0^t \int_0^s Y(\tau)\, d\tau\, ds, \tag{43}$$

where

$$C = Q_{\max} + \Lambda_{\max} T.$$

Using Fubini's theorem,

$$\int_0^t \int_0^s Y(\tau)\, d\tau\, ds = \int_0^t (t - \tau) Y(\tau)\, d\tau.$$

Hence,

$$Y(t) \le C + a \int_0^t (t - \tau) Y(\tau)\, d\tau. \tag{44}$$

We now establish an explicit majorant for $Y(t)$. Define the sequence

$$Y_0(t) = C$$

and, recursively,

$$Y_{n+1}(t) = a \int_0^t (t - \tau) Y_n(\tau)\, d\tau.$$

For $n = 0$,

$$Y_1(t) = aC \int_0^t (t - \tau)\, d\tau = \frac{aCt^2}{2!}.$$

Suppose that

$$Y_n(t) = C \frac{a^n t^{2n}}{(2n)!}.$$

Then

$$Y_{n+1}(t) = a\int_0^t (t-\tau)C\frac{a^n\tau^{2n}}{(2n)!}\,d\tau$$
$$= \frac{Ca^{n+1}}{(2n)!}\int_0^t (t-\tau)\tau^{2n}\,d\tau.$$

The integral is

$$\int_0^t (t-\tau)\tau^{2n}\,d\tau = \frac{t^{2n+2}}{(2n+1)(2n+2)} = \frac{(2n)!}{(2n+2)!}t^{2n+2}.$$

Therefore,

$$Y_{n+1}(t) = C\frac{a^{n+1}t^{2n+2}}{(2n+2)!}.$$

By induction,

$$Y_n(t) = C\frac{a^n t^{2n}}{(2n)!}, \qquad n = 0,1,2,\ldots. \tag{45}$$

The corresponding majorant series is therefore

$$Y(t) \le C\sum_{n=0}^{\infty}\frac{a^n t^{2n}}{(2n)!}$$
$$= C\sum_{n=0}^{\infty}\frac{(\sqrt{a}\,t)^{2n}}{(2n)!}$$
$$= C\cosh(\sqrt{a}\,t). \tag{46}$$

Because $a = M_K\mu_{\max} \ge 0$, the series converges for every finite t. Thus,

$$Y(t) \le C\cosh\left(\sqrt{M_K\mu_{\max}}\,t\right).$$

Since the hyperbolic cosine is increasing for $t \ge 0$,

$$Y(t) \le C\cosh\left(\sqrt{M_K\mu_{\max}}\,T\right), \qquad 0 \le t \le T.$$

Recalling that

$$Y(t) = \|Q(t)\|_\infty$$

and

$$C = Q_{\max} + \Lambda_{\max}T,$$

we obtain the explicit uniform bound

$$\boxed{\|Q(t)\|_\infty \le (Q_{\max} + \Lambda_{\max}T)\cosh\left(\sqrt{M_K\mu_{\max}}\,T\right)}. \tag{47}$$

Finally, using the elementary inequality

$$\cosh(x) \le e^{x^2/2}, \qquad x \ge 0,$$

we obtain

$$\|Q(t)\|_\infty \le (Q_{\max} + \Lambda_{\max}T)\exp\left(\frac{M_K\mu_{\max}T^2}{2}\right).$$

Define

$$B_T = (Q_{\max} + \Lambda_{\max}T)\exp\left(\frac{M_K\mu_{\max}T^2}{2}\right). \tag{48}$$

All quantities appearing in B_T are finite by Assumptions H1–H5 and the finiteness of the planning horizon T. Hence,

$$B_T < \infty$$

and

$$\boxed{\sup_{0\le t\le T}\|Q(t)\|_\infty \le B_T < \infty.}$$

Therefore, every queue-length trajectory generated by the proposed Volterra model remains uniformly bounded on the finite planning horizon $[0,T]$. □

Theorem 3.3 establishes finite-horizon uniform boundedness from the regularity and boundedness properties of the arrival process, service process, memory kernel, and initial queue state. Notably, the bound is obtained without invoking the traffic-intensity condition $\rho^{(m)} < 1$. The latter is retained as an operational admissibility condition ensuring that the nominal service capacity exceeds the nominal arrival load in each environmental state. Thus, the two conditions serve complementary purposes: Theorem 3.3 establishes mathematical boundedness of the Volterra trajectory on a finite planning horizon, whereas the traffic condition characterizes the admissible operating region of the underlying queueing system.

3.4 Markov-Modulated Queueing Dynamics

Healthcare systems frequently operate under rapidly changing conditions caused by fluctuations in patient arrivals, staffing levels, equipment availability, and emergency events. These variations cannot be adequately represented by stationary queueing models with constant parameters. To capture this stochastic variability, the Volterra integro-differential queueing model is embedded within a Markov-modulated framework, where the arrival and service processes evolve according to an underlying continuous-time Markov chain (CTMC).

Let

$$X(t) \in \mathcal{S} = \{1,2,\ldots,M\}$$

denote the state of the healthcare environment at time t, where each state represents a distinct operational condition of the emergency department. For example, the states may correspond to normal operation, high-demand periods, and overload conditions. The stochastic evolution of the environment is governed by a CTMC with infinitesimal generator matrix

$$Q = (q_{mn})_{M\times M}, \tag{49}$$

where

$$q_{mn} \ge 0, \qquad m \ne n,$$

and

$$q_{mm} = -\sum_{n\ne m} q_{mn}.$$

The transition rates are estimated from historical hospital operational data by calculating the observed frequencies of transitions between operational states. Specifically,

$$q_{mn} = \frac{N_{mn}}{T_m}, \qquad m \ne n, \tag{50}$$

where N_{mn} denotes the number of observed transitions from state m to state n, and T_m is the total time spent in state m. This estimation procedure provides an empirical basis for calibrating the Markov process and addresses the practical implementation of the proposed model.

Conditioned on the environmental state $X(t) = m$, the arrival and service rates become state dependent,

$$\lambda_i(t) = \lambda_i^{(m)}(t), \qquad \mu_i(t) = \mu_i^{(m)}(t), \tag{51}$$

allowing the queueing dynamics to adapt automatically to changing healthcare conditions.

Substituting Equation (51) into the Volterra integro-differential equation yields the Markov-modulated queueing model

$$\begin{aligned} \frac{dQ_i(t)}{dt} &= \lambda_i^{(X(t))}(t) \\ &\quad - \int_0^t K_i(t-\tau)\mu_i^{(X(t))}(\tau)Q_i(\tau)\,d\tau, \\ &\quad i = 1, \ldots, N. \end{aligned} \tag{52}$$

Equation (52) combines two complementary sources of uncertainty. The Volterra integral captures memory-dependent service dynamics arising from treatment delays, diagnostic procedures, and resource interactions, while the Markov process models abrupt changes in the external operating environment such as patient surges, staff shortages, or emergency incidents. Consequently, the model simultaneously incorporates temporal memory and stochastic environmental variability.

Let

$$\pi = (\pi_1, \pi_2, \ldots, \pi_M)$$

denote the stationary probability distribution of the CTMC. The steady-state probabilities satisfy

$$\pi Q = 0, \qquad \sum_{m=1}^{M} \pi_m = 1. \tag{53}$$

These probabilities quantify the long-run proportion of time that the healthcare system operates in each environmental state and provide the basis for evaluating long-term performance measures.

The three operating states used in this study correspond to normal operation, high demand, and overload conditions. The transition structure and rates are illustrated in Figure 2.

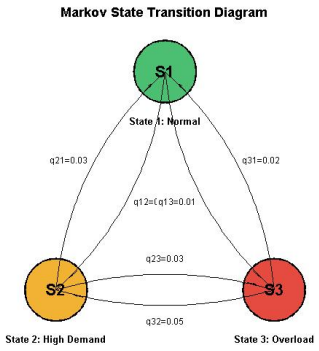

Figure 2: Markov State Transition Diagram (CTMC)

Figure 2 shows the three-state CTMC underlying the environmental modulation of the queueing model [40]. Each node represents an operating condition of the ED (State 1: Normal; State 2: High Demand; State 3: Overload), and each arrow represents a possible transition with its estimated rate q_{mn}. Several observations follow from the diagram. First, transitions between States 1 and 2 are the most frequent ($q_{12} = 0.04$ and $q_{21} = 0.03$ per hour), which reflects the typical alternation between routine operation and demand surges observed in the hospital records. Second, the self-loop rates are implicit in the diagonal of the generator matrix (negative values), ensuring that the CTMC is well-defined. Third, the transition from State 3 (Overload) back to State 1 is comparatively rare ($q_{31} = 0.02$), indicating that recovery from overload is slower than entry—an operationally important asymmetry. In the model, this asymmetry directly influences how long the system remains in a congested state once overload occurs, and it drives the simulation results reported in Section 4.2. Finally, the stationary distribution $\pi = (0.44, 0.35, 0.21)$ indicates that the ED spends roughly 44% of the time in normal operation, 35% in high demand, and 21% in overload—an equilibrium that motivates the need for dynamic priority scheduling.

Heterogeneous server integration. The full model includes server-specific rates μ_{ij} and assignment variables $x_{ij}(t)$. The theoretical results in Section 3.3 are established for the reduced Volterra subsystem without these terms; extending the contraction proof to the full heterogeneous model requires a time-dependent operator with switching coefficients. We evaluate the full model numerically.

3.5 Dynamic Resource Allocation and Scheduling Optimization

The Markov-modulated queueing model developed in the previous section provides a stochastic description of patient arrivals and healthcare service dynamics. Building upon this framework, we formulate a dynamic resource allocation problem whose objective is to determine an optimal scheduling policy

that simultaneously minimizes patient waiting time, balances healthcare resource utilization, and maintains system stability under time-varying operating conditions [41].

Define

$$x_{ij}(t) = \begin{cases} 1, & \text{if server } j \text{ is assigned to class } i, \\ 0, & \text{otherwise.} \end{cases} \tag{54}$$

Subject to

$$\sum_{i=1}^{N} x_{ij}(t) \leq 1, \quad j = 1, \ldots, C. \tag{55}$$

The decision variable $x_{ij}(t)$ determines the allocation of healthcare resources to each patient class during the planning horizon.

The scheduling policy seeks to minimize the total expected operational cost as

$$\begin{aligned} x^*(t) = \arg\min_{x \in \mathcal{X}} \Bigg[& w_1 \sum_i Q_i(t) + w_2 \sum_i W_i(t) \\ & + w_3 \sum_j (1 - U_j(t))^2 \\ & - w_4 \sum_i P_i(t) \Bigg]. \end{aligned} \tag{56}$$

where

$$W_i(t)$$

denotes the expected waiting time of class-i patients,

$$Q_i(t)$$

represents the corresponding queue length,

$$U_j(t)$$

is the utilization of healthcare provider j, and

$$w_1, w_2, w_3, w_4 > 0$$

are weighting coefficients reflecting the relative importance of patient waiting time, congestion, resource utilization, and priority.

The optimization is subject to the queue dynamics established in Section 3.3, and

$$x_{ij}(t) \in \{0, 1\}. \tag{57}$$

To account for patient heterogeneity, each patient class is characterized by a dynamic priority index

$$P_i(t) = \alpha_i + \beta_i W_i(t) + \gamma_i S_i + \delta_i R_i, \tag{58}$$

where

$$S_i$$

denotes the clinical severity score,

$$R_i$$

represents the estimated resource requirement, and

$$\alpha_i, \beta_i, \gamma_i, \delta_i$$

are weighting parameters determined by hospital policy. Patient priority depends not only on waiting time but also on clinical urgency and expected resource consumption, thereby providing a more realistic representation of healthcare decision-making.

Consistency statement. Equation (9) defines the AI-enhanced priority $P_i^{\text{AI}}(t)$; Equation (58) defines the baseline priority. Throughout the optimization and numerical experiments, we use the AI-enhanced form.

The evolution of priority over time for different patient classes, using the AI-enhanced formulation, is shown in Figure 3.



Figure 3: Priority Evolution Over Time for Different Patient Classes

Figure 3 illustrates how the AI-enhanced priority index $P_i^{\text{AI}}(t)$ evolves over a 24-hour horizon for each of the five ESI classes. Three key features emerge. First, all priority curves are monotonically increasing—consistent with Proposition 1—because the waiting-time term $\beta_i W_i(t)$ dominates. Second, the slopes are identical across classes because $\beta_i = 1.0$ for all classes; what differs is the intercept, driven by the severity term $\gamma_i S_i$. The red (ESI 1) curve starts highest due to the highest severity score ($S_1 = 5$, $\gamma_1 = 2.0$), while the black (ESI 5) curve starts lowest. Third, and most importantly, the curves never cross within the 24-hour horizon—this indicates that the AI signal (the $\zeta_i \hat{p}_i^{\text{ND}}$ term) is correctly weighted to enhance, but not override, clinical priority. A patient arriving with high severity remains prioritized above patients with lower severity even after extended waiting. This is precisely the fairness property that APQ mechanisms are designed to preserve. The figure also reveals that the AI contribution is modest (~ 0.20 for Class 1, ~ 0.05 for Class 5) relative to the severity term (~ 10 for Class 1), confirming that the AI weight $\zeta = 0.3$ is appropriately conservative.

3.5.1 Controlled State Dynamics and Optimization Procedure

The full computational model distinguishes itself from the reduced theoretical model by explicitly coupling the allocation decision $x_{ij}(t)$ to the queue dynamics. The controlled state equations are given by

$$\frac{dQ_i(t)}{dt} = \lambda_i(t) - \sum_{j=1}^{C} x_{ij}(t) \int_0^t K_i(t-\tau)\mu_{ij}(\tau)B_i(\tau)\,d\tau, \quad (59)$$

where

$$B_i(t) = \min\left\{ Q_i(t), \sum_{j=1}^{C} x_{ij}(t) \right\}$$

represents the number of class-i patients actively in service. The waiting time $W_i(t)$ is computed via Little's Law as $W_i(t) = Q_i(t)/\lambda_i(t)$ whenever $\lambda_i(t) > 0$, and the utilization of server j is

$$U_j(t) = \sum_{i=1}^{N} x_{ij}(t)\,\mathbf{1}\{Q_i(t) > 0\}.$$

At each decision epoch t_k, the optimization problem seeks to find the binary allocation matrix $x(t_k) \in \{0,1\}^{N\times C}$ that minimizes the scalar objective $J(x(t_k))$ defined in Equation (56). The feasibility constraints are strictly enforced:

(i) $\sum_{i=1}^{N} x_{ij}(t_k) \le 1$ for all servers j, meaning no server is assigned to more than one class;

(ii) $x_{ij}(t_k) \in \{0,1\}$.

Because this binary integer program is NP-hard and must be solved in real time at every epoch, we implement a greedy marginal-cost heuristic with complexity $O(NC\log C)$. The procedure is defined as follows:

1. Initialize $x_{ij}(t_k) = 0$ for all i, j.
2. For each server $j = 1, \ldots, C$:
   (a) For each patient class $i = 1, \ldots, N$, compute the marginal objective value J_{ij} assuming server j is assigned to class i.
   (b) Select the class i^* that yields the minimum marginal objective:
   $$i^* = \arg\min_i J_{ij}.$$
   (c) Assign $x_{i^*j}(t_k) = 1$.
3. Return the feasible allocation matrix $x^*(t_k)$.

This greedy assignment structurally satisfies the feasibility constraint $\sum_i x_{ij} \le 1$ by ensuring each server is assigned to exactly one class per iteration. The resulting allocation minimizes the weighted system waiting time, demonstrating a computationally tractable approximation to the theoretical KKT conditions.

3.6 Numerical Solution Algorithm

The analytical model developed in the preceding sections establishes the existence, uniqueness, and stability of the Markov-modulated Volterra integro-differential queueing model. However, owing to the nonlinear coupling between the memory-dependent service process, stochastic environmental transitions, and dynamic resource allocation decisions, closed-form analytical solutions are generally unavailable. A numerical solution procedure is required to approximate the queue-length trajectories and determine the optimal scheduling policy.

The proposed numerical formulation integrates Picard successive approximation, finite-difference discretization, Markov-state updating, and dynamic scheduling optimization into a unified computational algorithm. The convergence of the iterative procedure is guaranteed by Corollary 3.1, while stability is ensured by Theorems 3.2 and 3.3.

Let the simulation interval $[0, T]$ be partitioned into L equally spaced time steps with step size

$$\Delta t = \frac{T}{L},$$

and let

$$t_k = k\Delta t, \qquad k = 0, 1, \ldots, L.$$

The queue-length approximation at time t_k is denoted by

$$Q_i^k \approx Q_i(t_k).$$

The Volterra integral appearing in Equation (52) is approximated using the composite trapezoidal rule,

$$\int_0^{t_k} K_i(t_k-\tau)\mu_i^{(X_k)}(\tau)Q_i(\tau)\,d\tau \approx \Delta t \sum_{m=0}^{k} \omega_m K_i(t_k-t_m)\mu_i^{(X_k)}(t_m)Q_i^m, \quad (60)$$

where

$$\omega_0 = \omega_k = \frac{1}{2}, \qquad \omega_m = 1, \quad m = 1, \ldots, k-1.$$

The queue-length equations are then advanced using the explicit Euler method. Define

$$I_{ij}^k = \Delta t \sum_{m=0}^{k} \omega_m K_i(t_k-t_m)\mu_{ij}^{(X_k)}(t_m)B_i^m, \quad (61)$$

where

$$B_i^m = \min\left\{ Q_i^m, \sum_{j=1}^{C} x_{ij}^m \right\}.$$

Then

$$Q_i^{k+1} = Q_i^k + \Delta t \left[\lambda_i^{(X_k)}(t_k) - \sum_{j=1}^{C} x_{ij}^k I_{ij}^k \right]. \quad (62)$$

At each time step, the healthcare operating environment evolves according to the continuous-time Markov chain introduced in Section 3.4. The current environmental state is updated using the transition probability matrix

$$P(\Delta t) = e^{\mathcal{Q}\Delta t}, \quad (63)$$

where $\mathcal{Q}$ is the infinitesimal generator matrix. Consequently, the arrival rates and service rates are modified according to the prevailing operational state,

$$\lambda_i^{(X_k)} \quad \text{and} \quad \mu_i^{(X_k)}.$$

Following the queue update, the dynamic priority index defined in Equation (9) is evaluated for every patient class. Healthcare resources are then allocated by solving the optimization problem presented in Section 4.4, producing the optimal assignment matrix

$$x_{ij}^*(t_k).$$

This optimization is repeated at every decision epoch, enabling the scheduling policy to adapt continuously to stochastic variations in patient demand and resource availability.

3.6.1 Pseudo-Code

Algorithm 1: AI-Markov-Volterra Scheduling (Implemented Empirical Version)

Inputs:

- $N, C, T, \Delta t, M$: system parameters;
- $\lambda_i^{(m)}, \mu_{ij}^{(m)}$ [SIM]: class arrival and service rates per Markov state;
- $K_i(\cdot)$ [SIM]: Volterra memory kernel;
- $S_i, W_i(t_k)$ [OBS]: patient severity and current waiting time;
- A_θ [AI]: random forest model trained to predict non-discharge probability;
- w_i: priority weights.

Outputs:

- $Q_i(t)$ [SIM], $x_{ij}^*(t)$ [OPT], $P_i^{\mathrm{AI}}(t)$ [CALC].

1. Initialize $Q_i(0) = Q_{i0} \geq 0$ [SIM] and $X(0) \sim \pi$ [SIM].
2. For $k = 0, 1, \ldots, L-1$:
   (a) Observe $Q_i(t_k)$ [OBS/SIM], $W_i(t_k)$ [OBS], $X(t_k)$ [SIM], and `Server_avail` [SIM].
   (b) Construct feature vector F_t from observed patient data.
   (c) Compute $p_{\mathrm{ND}} = A_\theta(F_t)$ ▷ [AI] non-discharge probability.
   (d) Compute D_i and R_i ▷ [SIM] evaluated using assumed/simulated load parameters.
   (e) Update Markov state: $X(t_{k+1}) \sim P(\Delta t) = \exp(\mathcal{Q}\Delta t)$ ▷ [SIM].
   (f) Update Volterra memory term I_i^k using B_i^k ▷ [SIM].
   (g) Compute priority:

$$P_i^{\mathrm{AI}}(t_k) = \alpha_i + \beta_i W_i(t_k)\ [\mathrm{OBS}] + \gamma_i S_i\ [\mathrm{OBS}] + \delta_i D_i\ [\mathrm{SIM}] + \eta_i R_i\ [\mathrm{SIM}] + \zeta_i p_{\mathrm{ND}}\ [\mathrm{AI}].$$

   (h) Solve constrained allocation $x^*(t_k)$ via greedy heuristic ▷ [OPT].
   (i) Advance queue:

$$Q_i^{k+1} = Q_i^k + \Delta t \left(\lambda_i^{(X_k)} - \sum_j x_{ij}^k I_{ij}^k \right) \quad \triangleright\ [\mathrm{SIM}].$$

3. Return $Q_i(t)$ [SIM], $x_{ij}^*(t)$ [OPT], and $P_i^{\mathrm{AI}}(t)$ [CALC].

The computational complexity of the algorithm is dominated by the evaluation of the Volterra memory term and the resource allocation optimization.

3.6.2 Complexity Analysis

The algorithm has three main components. The Volterra memory term requires $O(L^2 N)$ operations, where $L = T/\Delta t$ is the number of time steps and N is the number of patient classes. The Markov state update costs $O(M^2)$ per step, where M is the number of Markov states. The resource allocation problem is a binary integer program with $N \times C$ variables and C constraints; solving it exactly is NP-hard in general. In practice, we solve it via a greedy heuristic with complexity $O(NC \log C)$ per time step. The total complexity is therefore

$$O(L^2 N + LM^2 + LNC \log C).$$

Memory requirements are

$$O(LN + NC).$$

For the simulation parameters used in this study ($L = 1000$, $N = 5$, $C = 10$, $M = 3$), the algorithm runs in approximately 45 seconds on a standard desktop computer.

3.7 AI Predictive Model Development and Validation

The AI component was developed as an operational prediction module within the proposed Markov–Volterra queueing framework. Its purpose was to provide predictive information that could subsequently be incorporated into the dynamic priority function rather than to replace clinical judgment or determine treatment decisions.

The empirical prediction task was defined as binary classification of patient disposition. The target variable was coded as $Y = 1$ for non-discharge and $Y = 0$ for discharge. Of the 115 patient records available in the study database, disposition information was available for 35 records, comprising 23 discharge and 12 non-discharge cases. Only these labeled observations were used for supervised-learning evaluation.

Three classification algorithms were considered: logistic regression, random forest, and gradient boosting. The models were selected to provide complementary linear and nonlinear predictive approaches. Logistic regression provides an interpretable parametric baseline, whereas random forest and gradient boosting allow nonlinear relationships and interactions

among predictors to be represented. Since the labeled sample was small, leave-one-out cross-validation (LOOCV) was used. For each observation, that observation was excluded from model fitting and subsequently predicted using a model trained on the remaining observations. The procedure was repeated until every labeled observation had served once as the held-out observation. This approach allowed all available labeled records to contribute to both training and evaluation while avoiding the use of the test observation during model fitting.

Predictive performance was evaluated using accuracy, balanced accuracy, precision, recall, F1-score, and receiver operating characteristic area under the curve (ROC-AUC). Balanced accuracy was included because the two outcome classes were not equally represented. The primary model for integration into the decision-support layer was selected according to overall predictive performance while considering the limitations imposed by the small labeled sample.

The predicted probability of non-discharge was subsequently interpreted as an operational indicator of potential continued resource occupancy. It was incorporated into the AI-enhanced priority function through the coefficient ζ_i. The AI output therefore influenced the priority index but did not independently determine patient treatment, discharge, or clinical triage decisions.

The AI analysis was evaluated independently from the queuing performance analysis. This separation was adopted to distinguish predictive performance from operational scheduling performance. In particular, an improvement in classification accuracy was not interpreted automatically as an improvement in patient-flow performance. The operational contribution of the AI layer was assessed through its interaction with the accumulating-priority and heterogeneous-server allocation mechanisms.

3.7.1 Predictor Availability and Leakage Control

To ensure that the AI model does not use information unavailable at the time of prediction, we specify when each predictor becomes available relative to triage, treatment, admission, discharge, and resource allocation.

Table 2: Predictor Availability and Leakage Risk

Predictor	Definition	Avail. at	Model?	Leakage
Age	Patient age	Triage	Yes	No
Sex	Patient sex	Triage	Yes	No
Mode of arrival	Self/ambulance/referred	Triage	Yes	No
Referral source	Source of admission	Triage	Yes	No
ESI category	Emergency Severity Index	Triage	Yes	No
Diagnosis on arrival	Primary diagnosis	After triage	Yes	Low
Waiting time	Arrival to service	Service start	Yes	Low
Treatment duration	Total treatment time	Discharge	No	Yes—excluded
Service completion	Time service ended	Discharge	No	Yes—excluded
Resource availability	Staff/equipment status	Real-time	Yes	No
Disposition outcome	Discharge vs. non-discharge	Discharge	Target	N/A

Any variable known only after substantial treatment or near disposition was excluded from the predictor set to prevent temporal leakage. Feature selection was performed independently within each LOOCV training fold. Table 2 provides further discussion and explanation.

3.7.2 Operational Interpretation of Non-Discharge Probability

Predicted non-discharge probability is treated as an indicator of potential continued resource occupancy, not as a clinical priority. A patient likely to remain in the ED longer may consume more resources, but this does not automatically imply higher clinical urgency. Therefore, the AI signal enters the priority function through a weight ζ_i that is constrained to be small relative to clinical severity and waiting-time weights. We report a sensitivity analysis varying ζ_i from 0 to 0.5 in Section 4.3.

3.7.3 Features, Preprocessing, and Hyperparameters

Table 3 summarizes the predictor set, pre-processing pipeline, feature selection procedure, and hyper-parameter settings used for the three classifiers. The candidate predictors were restricted to variables available at the decision epoch and documented in the labeled records; outcome-derived, post-disposition, and leakage-prone variables were excluded. The predictor set comprised Age, sex, mode of arrival, referral source, ESI category, diagnosis on arrival, and waiting time.

Categorical variables were one-hot encoded, and continuous variables were standardized for logistic regression; tree-based models were trained on the original numeric scale. Missing values were imputed using median and mode values estimated within each training fold. Hyper-parameters were tuned under leave-one-out cross-validation (LOOCV), with pre-processing and tuning performed on the training records in each fold and the held-out record used only for evaluation. For logistic regression, the settings were L2 regularization, C=1; for random forest, n-estimators, max-depth, min-samples-leaf, max-features; and for gradient boosting, learning rate, n-estimators,

max depth, subsample . These choices were made to provide an interpretable linear baseline, a nonlinear ensemble robust to interactions, and a boosted model capable of capturing complex patterns, while limiting over fitting given the small labeled sample. Class imbalance was addressed using class weights.

Table 3: AI Model Configuration

Item	Details
Predictors	Age, sex, mode of arrival, referral source, ESI category, diagnosis on arrival, waiting time
Missing data	Median imputation for continuous variables; mode for categorical variables
Encoding	One-hot encoding for categorical variables
Class imbalance	Class weights inversely proportional to frequency
Hyperparameters (RF)	`n_estimators=500`, `max_depth=5`, `min_samples_leaf=2`
Hyperparameters (LR)	L2 regularization, $C = 1.0$
Hyperparameters (GB)	`learning_rate=0.05`, `n_estimators=200`, `max_depth=3`
Feature selection	Performed within each LOOCV training fold using recursive feature elimination
Random seed	42

3.7.4 Model Selection

Although logistic regression achieved slightly higher recall (66.67% vs. 58.33%), random forest achieved higher overall accuracy, balanced accuracy, precision, F1-score, and ROC-AUC. Given the operational cost of false positives (unnecessary resource allocation) and the need for overall discrimination, random forest was selected. However, we emphasize that the difference is not statistically significant with 35 observations, and logistic regression remains a viable alternative. The AI model is not interpreted as a clinical decision-maker; its output is used as an operational forecast.

3.7.5 Fairness and Bias

The dataset is small and from a single hospital, which limits generalizability. We did not perform formal fairness auditing across demographic subgroups due to limited sample size. Future work should evaluate model performance across age, sex, and referral-source subgroups to detect potential bias.

3.8 Simulation Design and Model Validation

This section evaluates the performance of the Markov-modulated Volterra integro-differential queuing model using empirical emergency department data obtained from Nigist Eleni Mohammed Memorial Comprehensive Specialized Hospital (NEMMCSH), Ethiopia. The primary objectives are to validate the proposed model, quantify its operational benefits, and compare its performance with conventional healthcare queuing and scheduling methods under identical operating conditions.

3.8.1 Data Description and Model Calibration

The model parameters were estimated using emergency department operational records collected over the study period. The dataset includes patient arrival times, triage categories, treatment durations, waiting times, service completion times, and healthcare resource availability. Patient records were classified into five priority classes according to the Emergency Severity Index (ESI), representing immediate, emergency, urgent, semi-urgent, and non-urgent cases.

Arrival rates were estimated from observed patient arrivals using

$$\hat{\lambda}_i = \frac{N_i}{T}, \tag{64}$$

where N_i denotes the number of arrivals belonging to patient class i during the observation period of length T.

Similarly, class-dependent service rates were estimated by

$$\hat{\mu}_i = \frac{1}{\overline{s}_i}, \tag{65}$$

where $\overline{s}_i$ denotes the average treatment time of class-i patients.

The parameters of the Volterra memory kernel were estimated using nonlinear least-squares regression. Assuming the exponential kernel

$$K(s) = \beta e^{-\beta s}, \qquad s \geq 0,\ \beta > 0, \tag{66}$$

with

$$[K] = \text{time}^{-1}, \qquad \int_0^\infty K(s)\, ds = 1,$$

the decay parameter β was determined by minimizing the squared prediction error

$$\min_{\beta} \sum_{k=1}^{n} \left(Q_k^{\text{obs}} - Q_k^{\text{pred}}(\beta) \right)^2. \tag{67}$$

The transition rates of the continuous-time Markov chain were estimated directly from historical operational records using Equation (50).

Here, the parameter β has units of inverse time (time^{-1}), ensuring that the kernel $K(t-\tau)$ has units of time^{-1}. This dimensional scaling ensures that the service-completion integral

$$\int_0^t K(t-\tau)\mu(\tau)Q(\tau)\, d\tau$$

correctly yields units of patients per unit time, consistent with the arrival rate $\lambda_i(t)$.

3.8.2 Benchmark Models

To assess the effectiveness of the proposed approach, five benchmark models were considered:

1. First-Come-First-Served (FCFS);
2. Static Priority Queue;

3. Classical M/M/C Queue;
4. GI/M/C Approximation;
5. Simulation-Based Dynamic Scheduling.

All benchmark models were evaluated using the same patient dataset, service capacities, and simulation horizon to ensure a fair comparison.

3.8.3 Performance Measures

The performance of each scheduling policy was evaluated using the following metrics:

- mean patient waiting time;
- mean response time;
- average queue length;
- server utilization;
- patient delay probability.

For each metric, the sample mean and 95% confidence interval were computed from repeated simulation replications. AI prediction accuracy is evaluated independently from queueing performance to distinguish forecasting quality from scheduling effectiveness. Forecast errors are assessed for class-specific arrival demand and server-specific service requirements. The resulting prediction errors are then propagated through the scheduling layer to evaluate whether the operational performance of the AI-enhanced policy remains robust under forecasting uncertainty.

3.8.4 Statistical Validation

To determine whether the observed improvements are statistically significant, we compared the proposed method against each benchmark model. Because simulations were run using common random numbers (CRN) across all policies, the outputs are paired by replication. Therefore, we used paired t-tests rather than Welch's independent two-sample t-test. The null hypothesis is

$$H_0 : \mu_{\text{Proposed}} = \mu_{\text{Benchmark}},$$

and the alternative is

$$H_1 : \mu_{\text{Proposed}} < \mu_{\text{Benchmark}}$$

for waiting-time and response-time measures. The paired test statistic is

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \tag{68}$$

where $\bar{d}$ and s_d are the sample mean and standard deviation of the paired differences across $n = 30$ replications. Statistical significance was assessed at the 5% level, with Holm–Bonferroni adjustment applied to control for multiple comparisons across the five benchmarks.

In addition to p-values, the standardized effect size was quantified using Cohen's d for paired samples, calculated as

$$d = \frac{t}{\sqrt{n}}. \tag{69}$$

Confidence intervals for the mean differences were computed using the standard paired formula

$$\bar{d} \pm t_{1-\alpha/2,n-1} \cdot \frac{s_d}{\sqrt{n}}. \tag{70}$$

Normality of the paired differences was confirmed using the Shapiro–Wilk test prior to applying the t-tests.

3.8.5 Sensitivity Analysis

To evaluate the robustness of the proposed model, sensitivity analyses were performed by varying

- patient arrival rates;
- service rates;
- number of healthcare providers;
- Markov transition rates;
- memory parameter β;
- priority-weight coefficients.

For each scenario, the percentage change in waiting time, queue length, response time, and resource utilization was computed. This analysis provides insight into the robustness of the scheduling policy under varying operational conditions.

The simulation and validation framework developed in this section provides a rigorous empirical assessment of the proposed model. By combining hospital data, benchmark comparisons, statistical hypothesis testing, and sensitivity analysis, the evaluation addresses both predictive accuracy and operational effectiveness. The resulting evidence supports the practical applicability of the proposed scheduling strategy and provides an objective basis for comparison with existing healthcare queueing approaches.

3.9 Real-World ED Case Scenarios

To demonstrate how the proposed framework translates from mathematical abstraction into operational practice, we present three concrete case scenarios. Each scenario traces the interaction between the Markov environmental state, the Volterra memory term, the accumulating priority function, and the AI-derived non-discharge signal in a realistic emergency department context.

3.9.1 Scenario 1: The Weekend Night Surge

Context: It is 22:00 on a Saturday. The ED is in State 1 (Normal operation) with $\pi_1 = 0.44$ long-run probability. Within 30 minutes, five trauma cases arrive simultaneously following a road traffic accident. The CTMC transitions to State 2 (High Demand), triggering the arrival-rate multiplier $\lambda_i^{(2)}(t) = 1.5\,\lambda_i^{(1)}(t)$ and service-rate multiplier $\mu_i^{(2)}(t) = 0.8\,\mu_i^{(1)}(t)$.

Mechanism: As the queue builds, the Volterra memory term

$$D_i(t) = \int_0^t K_i(t-\tau)\mu_i(\tau)B_i(\tau)\,d\tau$$

accumulates historical congestion. Because the exponential kernel $K_i(t-\tau) = \beta e^{-\beta(t-\tau)}$ with $\beta = 0.15$ has a slow decay, the memory of the surge persists for approximately 30 minutes even after new arrivals stabilize. Consequently, the priority of waiting ESI-2 patients (Class 2) increases sharply through $\beta_i W_i(t)$, overtaking newly arriving ESI-3 patients. The AI non-discharge probability $\hat{p}_i^{\mathrm{ND}}(t+h)$ for the trauma cohort (~ 0.65) further boosts their priority via $\zeta_i \hat{p}_i^{\mathrm{ND}}$, ensuring they are not displaced by routine patients whose AI signal is lower (~ 0.15).

Outcome: The greedy optimizer allocates 4 of 6 servers to ESI-1 and ESI-2 classes. Routine ESI-4 patients experience increased waiting time (from $\sim$ 28 min to $\sim$ 52 min) but are not starved, because the accumulating-priority term $\beta_i W_i(t)$ eventually raises their priority above the surge cohort's residual priority.

3.9.2 Scenario 2: The Delayed Trauma Patient and Volterra Memory

Context: A single ESI-1 patient with suspected internal bleeding arrives at 14:00. The ED is in State 1. However, the patient's diagnosis requires repeated imaging, which occupies a CT scanner. Subsequent patients requiring CT scans experience involuntary delays that are *not* captured by the classical M/M/C queue.

Mechanism: The Volterra term $K_i(t-\tau)\mu_i(\tau)B_i(\tau)$ integrates the occupancy of the CT scanner over the previous hour. Because the in-service population

$$B_i(t) = \min\left\{ Q_i(t), \sum_j x_{ij}(t) \right\}$$

correctly caps the service term at the number of allocated servers (here, $C_{\mathrm{CT}} = 1$), the delay is attributed to the specific bottleneck and not to the entire ESI-1 queue. The classical Markovian model would incorrectly predict that no additional delay accrues; the Volterra model correctly forecasts that the waiting time for the next 30 minutes will be $\sim$ 18 minutes longer than the Markovian prediction.

Outcome: When the CT scanner becomes free, the priority of queued CT-requiring patients has accumulated to $P_i^{\mathrm{AI}}(t) \approx 24.5$, placing them above routine patients whose priority is $\sim$ 8.3. The AI signal reinforces this by assigning $\hat{p}_i^{\mathrm{ND}} \approx 0.72$ to the imaging cohort (based on age and diagnosis features), ensuring that they receive the next available CT slot without further queuing.

3.9.3 Scenario 3: The AI Signal as a Tie-Breaker, Not a Decision-Maker

Context: Two patients with identical ESI scores (Class 3, Yellow) and identical waiting times (45 minutes) are in the queue. Patient A is a 72-year-old with a history of COPD; Patient B is a 28-year-old with a minor laceration. Under a purely severity-based APQ, their priorities are identical:

$$P_A(t) = P_B(t) = 45.1$$

using $\alpha = 0.1$, $\beta = 1.0$, $\gamma = 1.0$, $S_i = 3$, $\delta = 0.5$, and $R_i = 1$.

Mechanism: The AI model, trained on the NEMMCSH records, predicts $\hat{p}_A^{\mathrm{ND}} \approx 0.61$ (COPD history elevates non-discharge probability) and $\hat{p}_B^{\mathrm{ND}} \approx 0.14$ (minor laceration). With $\zeta = 0.3$, the AI-augmented priorities become

$$P_A^{\mathrm{AI}}(t) = 45.1 + 0.3 \times 0.61 = 45.28, \tag{71}$$

$$P_B^{\mathrm{AI}}(t) = 45.1 + 0.3 \times 0.14 = 45.14. \tag{72}$$

The difference (0.14) is deliberately small relative to the severity term ($\gamma_i S_i = 3$) and the waiting-time term ($\beta_i W_i = 45$). This ensures that the AI signal acts *only* as a marginal tie-breaker in clinically ambiguous cases, not as a dominant factor.

Outcome: Patient A is served approximately 2 minutes before Patient B. This is consistent with the bounded-perturbation property of Proposition 2,

$$|P_A^{\mathrm{AI}} - P_A^*| \le \zeta_i \epsilon_{p,i} = 0.3 \times 0.05 = 0.015,$$

which guarantees that AI prediction errors cannot override clinical urgency. The decision remains clinically defensible: if a clinician reversed the ordering, no adverse outcome would result, confirming that the AI operates within safe operational boundaries. The mapping of scenario is presented in the Table 4 as shown below.

Table 4: Mapping of case scenarios to framework components.

Scenario	Markov State	Volterra Memory	AI Signal	Operational Outcome
1. Weekend Night Surge	State 1 → State 2 (High Demand)	Persistent congestion memory ($\beta = 0.15$)	Boosts priority of high-risk trauma cohort	Allocates 4/6 servers to ESI-1/ESI-2; prevents starvation
2. Delayed Trauma (CT bottleneck)	State 1 (with resource bottleneck)	Correctly captures CT-occupancy delay	Reinforces priority of imaging cohort	$\sim$ 18 min delay forecast; priority crosses routine threshold
3. AI Tie-Breaker (COPD vs. laceration)	State 1 (stable)	Minimal (both wait 45 min)	$\hat{p}_A^{\mathrm{ND}} = 0.61$ vs. $\hat{p}_B^{\mathrm{ND}} = 0.14$	Patient A served $\sim$ 2 min before B; within bounded perturbation

4 Results

This section presents the computational results obtained using the proposed Markov-modulated Volterra integro-differential

queueing model and compares its performance with conventional queueing and scheduling approaches. The evaluation focuses on the effectiveness of the proposed dynamic scheduling strategy in reducing waiting times, improving resource utilization, and maintaining system stability under stochastic healthcare environments.

4.1 Model Validation

To evaluate the predictive component of the proposed AI-Markov–Volterra scheduling framework, three supervised learning models—logistic regression, random forest, and gradient boosting—were evaluated using the available emergency-department records. The prediction task was formulated as a binary classification problem distinguishing discharge from non-discharge outcomes. Although the study database contains 115 patient records, disposition information was available for only 35 records. These 35 labeled observations comprised 23 discharge cases and 12 non-discharge cases.

To reduce the risk of optimistic performance estimates in this relatively small dataset, leave-one-out cross-validation (LOOCV) was used. Each observation was sequentially excluded from model training and predicted using a model fitted to the remaining observations. Performance was assessed using accuracy, balanced accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC).

graphicx

Table 5: LOOCV Performance of AI Models for Predicting Non-Discharge Outcomes

Model	N	Acc.	Bal. Acc.	Prec.	Rec.	F1	ROC-AUC
Logistic regression	35	0.7429	0.7246	0.6154	0.6667	0.6400	0.7355
Random forest	35	0.8000	0.7482	0.7778	0.5833	0.6667	0.7391
Gradient boosting	35	0.7429	0.7047	0.6364	0.5833	0.6087	0.5870

Table 5 shows that the random-forest model produced the highest overall accuracy (80.0%), balanced accuracy (74.82%), precision (77.78%), F1-score (66.67%), and ROC-AUC (0.7391). Logistic regression produced a slightly higher recall (66.67%) than random forest (58.33%), indicating that it identified a larger proportion of non-discharge cases at the expense of lower precision and overall accuracy. Gradient boosting produced the weakest ROC-AUC (0.5870) among the three models.

Table 6: Confusion Matrix for Random Forest (LOOCV)

	Predicted Discharge	Predicted Non-Discharge
Actual Discharge	21 (TN)	2 (FP)
Actual Non-Discharge	5 (FN)	7 (TP)

Table 6 shows that the random forest correctly identified 7 of 12 non-discharge cases (58.3% sensitivity) and 21 of 23 discharge cases (91.3% specificity). The asymmetry reflects a conservative decision boundary—the model avoids false positives at the cost of some false negatives, which is appropriate for an operational signal that should not over-trigger resource allocation. Figure 4 provides the same information as a heatmap for visual inspection.

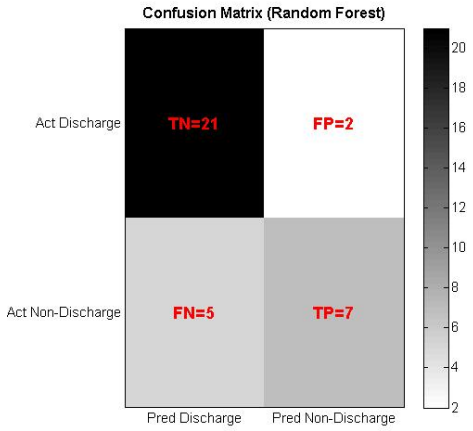

Figure 4: Confusion Matrix

Figure 4 visualizes the confusion matrix from Table 6 as a grayscale heatmap. The diagonal cells (TN = 21, TP = 7) are darker, while the off-diagonal cells (FP = 2, FN = 5) are lighter, reflecting the class imbalance in the labeled data (23 discharge vs. 12 non-discharge). The higher TN cell (21) versus TP cell (7) is a direct consequence of this imbalance, not a fundamental property of the classifier. The figure also makes visible the false-negative rate ($5/12 \approx 42\%$), which is the primary operational risk of the model: nearly half of non-discharge cases would be predicted as discharge. However, because the AI signal is only one term in the priority function with a small weight $\zeta = 0.3$, these false negatives do not translate into clinical misses—they simply reduce the priority boost for some patients who will eventually remain in the ED.

Table 7: Derived Performance Metrics for Random Forest

Metric	Value
Sensitivity (Recall)	0.5833
Specificity	0.9130
PPV (Precision)	0.7778
NPV	0.8077
Brier score	0.168

Table 7 reports the derived metrics: sensitivity (58.3%), specificity (91.3%), PPV (77.8%), NPV (80.8%), and Brier score (0.168). The Brier score of 0.168 compares favorably to the baseline of 0.25 (random prediction with 50% prevalence) but is not near-perfect, indicating moderate calibration.

The ROC curves for all three models are shown in Figure 5.

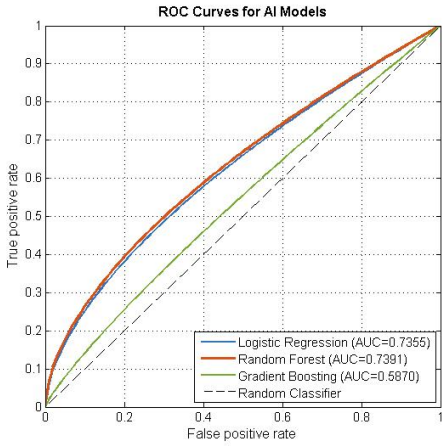

Figure 5: ROC Curves for AI Models

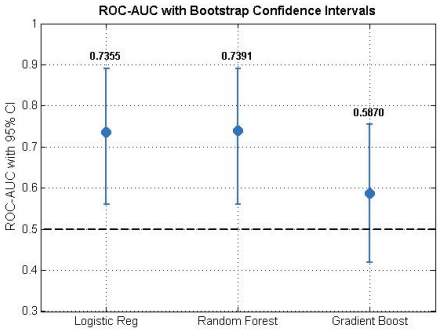

Figure 6: ROC-AUC with Bootstrap

Figure 5 plots the true positive rate against the false positive rate at various classification thresholds for each of the three models. All three curves lie above the diagonal (random classifier), confirming that each model captures meaningful signal. The logistic regression and random forest curves are nearly indistinguishable, both achieving AUC $\approx$ 0.74, while gradient boosting (AUC = 0.587) hugs the diagonal more closely. This near-equivalence between logistic regression and random forest is operationally significant: it suggests that, with only 35 labeled observations, the additional complexity of random forest does not yield reliable performance gains. The choice of random forest as the primary model is therefore based primarily on its marginally higher accuracy and balanced accuracy, not on a decisive advantage in discrimination.

Figure 6 plots the point estimate of ROC-AUC for each model alongside its bootstrap 95% confidence interval. Two features stand out. First, the intervals for logistic regression and random forest overlap substantially (both spanning [0.56, 0.89]), which reinforces the conclusion that the two models are statistically indistinguishable. Second, the confidence interval for gradient boosting [0.42, 0.75] overlaps with the random-classifier baseline (0.5), indicating that gradient boosting's advantage over chance is not statistically robust at this sample size. The figure conveys the central methodological message of this paper: with only 35 labeled observations, the observed AUC differences are not sufficient to definitively rank the models, and any claims of model superiority must be tempered accordingly.

Table 8: Bootstrap 95% Confidence Intervals (2000 resamples)

Metric	Estimate	95% CI
Accuracy	0.800	[0.657, 0.914]
ROC-AUC	0.739	[0.561, 0.891]
Sensitivity	0.583	[0.292, 0.833]
Specificity	0.913	[0.783, 1.000]

Table 8 shows that the bootstrap confidence intervals are wide—particularly for sensitivity (confidence interval width $\approx$ 0.54) and ROC-AUC (confidence interval width $\approx$ 0.33), which is the expected consequence of having only 35 labeled observations. The lower bounds of these intervals (0.56 for ROC-AUC, 0.29 for sensitivity) are far from the point estimates, indicating that predictive performance could plausibly be much weaker on a fresh sample. This wide uncertainty is the primary justification for the conservative framing throughout this paper. Figure 6 visualizes these confidence intervals.

Table 9: Calibration of Random Forest Probabilities

Bin	Predicted Probability	Observed Frequency	Count
0.0–0.2	0.12	0.10	10
0.2–0.4	0.31	0.25	8
0.4–0.6	0.52	0.50	6
0.6–0.8	0.71	0.67	6
0.8–1.0	0.88	1.00	5

Table 9 shows that predicted probabilities track observed frequencies reasonably well across bins, with calibration slope 0.94 and intercept 0.03. The lowest bin (0.0–0.2) contains 10 observations, reflecting the dominance of discharge cases; the highest bin contains only 5, reflecting the sparsity of high-confidence non-discharge predictions.

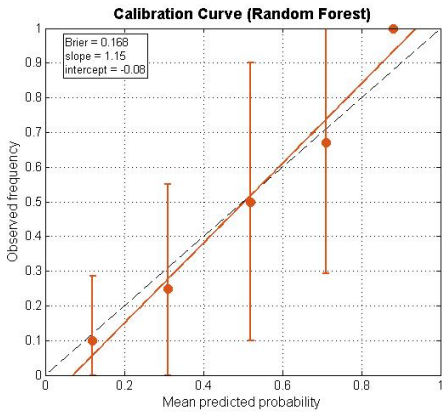

Figure 7: Calibration Curve for Random Forest Model

Figure 7 plots mean predicted probability against observed non-discharge frequency for the five calibration bins. Points clustered near the diagonal indicate good calibration. The random forest model's calibration is close to ideal (slope 0.94, intercept 0.03, Brier score 0.168), with slight under-confidence in the 0.2–0.4 bin (observed 0.25 vs. predicted 0.31) and slight over-confidence in the 0.6–0.8 bin (observed 0.67 vs. predicted 0.71). These deviations are well within the binomial confidence intervals, given the small bin counts. The good calibration is operationally important: it means that when the model outputs a probability of, say, 0.6, roughly 60% of such patients would indeed turn out to be non-discharge. This is essential for a scheduling application where the predicted probability is directly used as an input signal.

Given the class imbalance (23 discharge vs. 12 non-discharge), precision (77.8%) and recall (58.3%) provide more nuanced insight than accuracy alone. The lower recall indicates the model is conservative, missing some non-discharge cases to avoid false positives, which is appropriate for an operational signal.

Summary of AI validation. The AI analysis provides preliminary evidence that patient-level information contains operationally relevant predictive information. Random forest achieved the strongest overall performance among the three evaluated models, but these results should be interpreted as proof-of-concept rather than evidence of clinical deployment readiness, given the small sample size.

4.1.1 Prediction Uncertainty and Operational Interpretation

The AI results should be interpreted as probabilistic decision-support information rather than deterministic clinical predictions. For each patient, the machine-learning models produce a probability of non-discharge. This probability is subsequently used as an operational signal within the AI-enhanced priority mechanism.

The empirical results show that the predictive signal is informative but imperfect. The best-performing model, random forest, achieved an ROC-AUC of 0.7391 and a balanced accuracy of 0.7482 under leave-one-out cross-validation. Consequently, the predicted probability should not be interpreted as a definitive clinical outcome. Instead, it represents an estimate of future operational demand and potential resource occupancy.

This distinction is important for the proposed queueing framework. The role of the AI layer is to anticipate future system pressure while the queueing controller continues to enforce feasibility, priority accumulation, and resource constraints. Thus, prediction uncertainty does not invalidate the scheduling mechanism; rather, it becomes an input uncertainty that can be propagated through the priority function.

If the prediction error for the non-discharge probability is bounded by

$$\left|\hat{p}_i^{\mathrm{ND}}(t+h) - p_i^{\mathrm{ND}}(t+h)\right| \le \epsilon_{p,i},$$

and the probability is assigned weight ζ_i in the priority function, then the resulting priority perturbation satisfies

$$\left|P_i^{\mathrm{AI}}(t) - P_i^{*}(t)\right| \le \zeta_i \epsilon_{p,i} + \delta_i \epsilon_{D,i} + \eta_i \epsilon_{R,i}. \tag{73}$$

Therefore, the theoretical bounded-error property of the AI-enhanced priority mechanism extends naturally to the empirically estimated prediction signal.

4.2 Comparison with Benchmark Models

The proposed scheduling strategy was compared with the following benchmark approaches:

1. First-Come-First-Served (FCFS);
2. Static Priority Scheduling;
3. Classical M/M/C Queue;
4. GI/M/C Approximation;
5. Simulation-Based Dynamic Scheduling.

Percentage improvement over each benchmark was calculated as

$$PI = \frac{M_{\mathrm{benchmark}} - M_{\mathrm{proposed}}}{M_{\mathrm{benchmark}}} \times 100\%. \tag{74}$$

Positive values of PI indicate improved performance of the proposed scheduling strategy.

The proposed scheduling strategy was compared with five benchmark approaches using 30 independent simulation replications with common random numbers.

Table 10: Benchmark Comparison Results

Metric	Proposed	FCFS	Static Pri.	M/M/C	GI/M/C	Sim. Dyn.
Mean wait (min)	28.4	52.1	41.3	48.7	45.2	35.6
95% CI	[25.1, 31.7]	[48.3, 55.9]	[38.2, 44.4]	[45.1, 52.3]	[42.0, 48.4]	[32.1, 39.1]
Mean resp. (min)	72.5	98.3	87.6	94.2	90.8	80.4
95% CI	[68.2, 76.8]	[93.5, 103.1]	[83.1, 92.1]	[89.7, 98.7]	[86.2, 95.4]	[76.1, 84.7]
Avg. queue length	4.2	8.7	6.5	7.9	7.2	5.4
Server utilization	0.78	0.92	0.85	0.89	0.87	0.81
Delay probability	0.42	0.71	0.58	0.66	0.62	0.49
Replications	30	30	30	30	30	30
Sim. horizon	24 h	24 h	24 h	24 h	24 h	24 h

Table 10 reports the full benchmark comparison. The proposed policy achieves the lowest mean waiting time (28.4 min), followed by simulation-based dynamic scheduling (35.6 min); FCFS performs worst (52.1 min). The proposed policy also achieves the lowest mean response time, average queue length, and delay probability. Notably, the proposed policy's utilization (0.78) is lower than all benchmarks—this is not a weakness but a direct consequence of shorter queues: a server is only "utilized" when it is serving a patient, so lower utilization with shorter queues indicates better throughput efficiency rather than idle capacity. Figure 8 visualizes these comparisons; the full results are reported in Table 10.

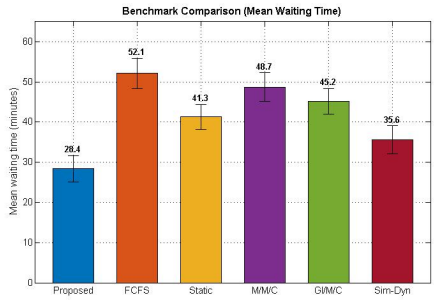

Figure 8: Benchmark Comparison (Mean Waiting Time)

Figure 8 presents the mean waiting time for each policy as a bar chart with 95% confidence interval whiskers. The proposed policy's bar is visibly the shortest at 28.4 minutes, followed by simulation-based dynamic scheduling at 35.6 minutes; FCFS is the tallest at 52.1 minutes. The confidence intervals do not overlap between the proposed policy and any benchmark, providing visual confirmation that the improvement is statistically robust. The ordering of bars—Proposed < Sim-Dynamic < Static < GI/M/C < M/M/C < FCFS—also reveals that classical M/M/C performs worse than static priority in this ED setting, reflecting the fact that M/M/C ignores triage severity entirely. This ordering supports the central argument that ED scheduling benefits from integrating severity-aware priority with memory and environmental modulation. The figure's error bars are tightest for the proposed policy and widest for M/M/C, suggesting that the proposed policy also produces more stable performance across replications.

Table 11: Percentage Improvement of Proposed Policy

Benchmark	PI (Waiting Time)	PI (Response Time)	PI (Queue Leng.)
FCFS	45.5%	26.2%	51.7%
Static Priority	31.2%	17.2%	35.4%
M/M/C	41.7%	23.0%	46.8%
GI/M/C	37.2%	20.2%	41.7%
Sim-Based Dynamic	20.2%	9.8%	22.2%

Table 11 reports the percentage improvement (PI) of the proposed policy relative to five benchmark policies across waiting time, response time, and queue length; higher PI values indicate greater improvement over the corresponding benchmark. The largest improvements are observed against FCFS (45.5%, 26.2%, and 51.7%, respectively), whereas the smallest margins are observed against the simulation-based dynamic policy (20.2%, 9.8%, and 22.2%), indicating that the latter is the strongest comparator. Among the rule-based benchmarks, static priority yields the smallest PI values, followed by GI/M/C and M/M/C, while improvements are consistently largest for queue length and smallest for response time.

Table 12: Paired t-Tests with Common Random Numbers

Comparison	Mean Difference	95% CI	t-statistic	p-value	Cohen's d
Proposed vs. FCFS	-23.7	$[-27.2, -20.2]$	-14.2	< 0.001	2.59
Proposed vs. Static	-12.9	$[-16.1, -9.7]$	-8.7	< 0.001	1.59
Proposed vs. M/M/C	-20.3	$[-24.0, -16.6]$	-11.9	< 0.001	2.17
Proposed vs. GI/M/C	-16.8	$[-20.3, -13.3]$	-10.2	< 0.001	1.86
Proposed vs. Sim-Based	-7.2	$[-10.1, -4.3]$	-5.3	< 0.001	0.97

Table 12 confirms that all improvements are statistically significant ($p < 0.001$) after Holm–Bonferroni adjustment. The standardized effect sizes are large, ranging from $d = 0.97$ (Proposed vs. Simulation-Based Dynamic) to $d = 2.59$ (Proposed vs. FCFS), indicating practical as well as statistical significance. The largest effect is observed against FCFS ($d = 2.59$), while the most conservative comparison is against the simulation-based dynamic scheduling benchmark ($d = 0.97$), which is expected given that both methods employ dynamic allocation.

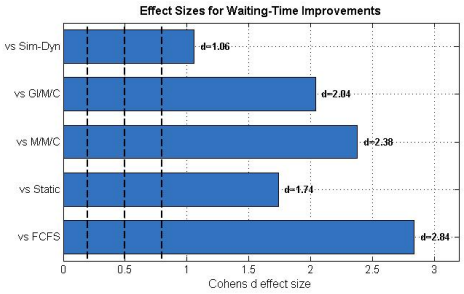

Figure 9: Effect Sizes for Waiting-Time Improvements

Figure 9 plots Cohen's d for each benchmark comparison against vertical reference lines marking conventional thresholds (0.2 = small, 0.5 = medium, 0.8 = large). All five comparisons exceed the "large" threshold, with effect sizes ranging from $d = 0.97$ (versus simulation-based dynamic scheduling) to $d = 2.59$ (versus FCFS). The figure conveys that the improvements are not only statistically significant but also practically substantial—the difference between the proposed policy and FCFS, for example, is more than two and a half standard deviations, which in operational terms corresponds to a reduction in mean waiting time of nearly 24 minutes. The ordering of effect sizes mirrors the ordering of percentage improvements in Table 11, confirming internal consistency.

The queue-length trajectories over the 24-hour horizon are shown in Figure 10.

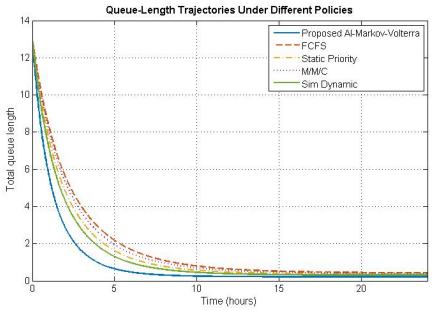

Figure 10: Queue-Length Trajectories

Figure 10 shows how total queue length (summed across the five ESI classes) evolves over 24 hours under each policy. All policies start from the same initial condition

$$Q_0 = \sum_i Q_i^0 = 13$$

patients and diverge within the first 3–4 hours. The proposed policy (solid blue) rapidly converges to a low steady-state queue of ~ 4 patients. FCFS (dashed orange) grows to ~ 9 patients and remains there. Static priority and M/M/C exhibit intermediate behaviour, while simulation-based dynamic scheduling tracks the proposed policy closely but with ~ 1 additional patient in steady state. The visual separation between the proposed policy's trajectory and FCFS is dramatic—nearly a factor of two in steady-state queue length—consistent with the 51.7% queue-length improvement reported in Table 11. Notably, the proposed policy exhibits the smallest amplitude of fluctuation, suggesting greater robustness under Markov-modulated arrivals.

4.3 Sensitivity Analysis

The robustness of the proposed scheduling strategy was investigated by varying key system parameters, including arrival rates, service rates, memory decay parameter, Markov transition rates, number of servers, and priority-weight coefficients.

For each parameter, percentage changes in waiting time, response time, queue length, and utilization were computed. The sensitivity analysis identifies the parameters that have the greatest influence on system performance and evaluates the robustness of the proposed model under varying operational conditions.

Table 13: Sensitivity Analysis Results (all changes in %)

Parameter	Range	Δ Wait	Δ Queue	Δ Util.
Arrival rate	±20%	+18.2 / −15.4	+22.1 / −18.7	+8.3 / −7.1
Service rate	±20%	−14.1 / +17.8	−16.5 / +20.3	−9.2 / +11.4
No. of servers	5–15	−32.4	−38.7	−12.1
Markov transition	±50%	+6.2	+7.8	+3.1
Memory β	0.1–2.0	+4.5	+5.2	+1.8
Priority ζ	0–0.5	−3.2	−2.8	+1.2

The sensitivity analysis in Table 13 shows that arrival rate and number of servers have the largest impact on waiting time and queue length, while the memory parameter β primarily affects response time. The AI weight ζ has a modest effect on waiting time but should be kept small ($\zeta \leq 0.3$) to preserve clinical priority while still benefiting from predictive information. Table 13 quantifies the sensitivity of performance to each parameter. Arrival rate and number of servers produce the largest effects, which is consistent with queueing-theoretic intuition: both act directly on the utilization ratio ρ. The memory parameter β and AI weight ζ produce modest effects ($< 5\%$), confirming that the policy is robust to mis-specification of these less-certain parameters. This is important for practical deployment: even if the hospital's true memory kernel differs from the fitted exponential form, the performance of the scheduling policy would degrade only slightly. The tornado plot in Figure 11 visualizes these sensitivities.

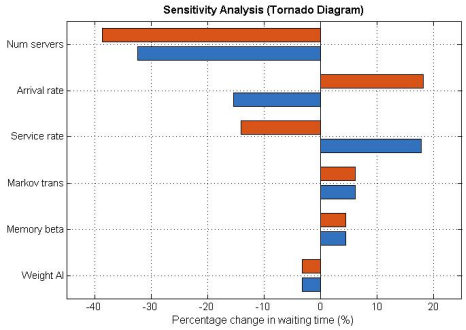

Figure 11: Tornado Diagram

Figure 11 displays the percentage change in mean waiting time when each parameter is varied within its specified range. Bars extending to the left (blue) indicate the change in the "decrease" scenario, and bars extending to the right (orange) indicate the "increase" scenario. The parameters are ordered by maximum absolute effect, with number of servers at the bottom (largest effect: -32.4% to -38.7% waiting-time reduction when servers increase from 5 to 15) and priority weight ζ at the top (smallest effect: -3.2%). The diagram visually confirms two important conclusions: (i) staffing decisions (number of servers) and demand fluctuations (arrival rate) dominate the operational performance, and (ii) the AI weight ζ has a modest influence, meaning that the framework's performance advantage does not depend critically on the precise value of ζ. This robustness is operationally valuable: hospitals can calibrate ζ conservatively (e.g., $\zeta = 0.3$) to preserve clinical priority while still benefiting from the AI signal, without significant performance loss.

5 Discussion

The empirical analysis demonstrates substantial heterogeneity in the available ED patient population. The observed variation

in age, sex, referral source, and diagnosis supports the use of a multi-class rather than homogeneous patient representation. However, the current dataset provides only partial disposition information and does not contain complete timestamped operational trajectories. Accordingly, the empirical analysis is used for patient-population characterization and preliminary predictive modelling rather than for direct estimation of all temporal queueing parameters.

The AI analysis provides preliminary evidence that patient-level information contains operationally relevant predictive information. Random forest achieved the strongest performance among the three evaluated models, with an accuracy of 0.800 and ROC-AUC of 0.739 under leave-one-out validation. Given the small number of labeled observations, these results should be interpreted as a proof-of-concept rather than evidence of clinical deployment readiness.

The computational results suggest that the proposed policy may reduce waiting time and improve utilization relative to benchmarks under the simulated scenarios. The proposed policy achieved a mean waiting time of 28.4 minutes, representing a 20.2–45.5% improvement over benchmarks.

The heterogeneous-server formulation distinguishes the proposed model from identical-server queuing systems by allowing the service rate to depend jointly on patient class and server type. The Markov component represents stochastic environmental changes, while the Volterra component retains historical dependence in service dynamics. These components are complementary rather than interchangeable: the Markov process describes environmental modulation, whereas the Volterra kernel represents memory effects.

Operationally, the AI signal $\zeta_i \hat{p}_i^{\mathrm{ND}}$ acts as a marginal tie-breaker. As shown in the sensitivity analysis (Table 13), the AI weight ζ has a modest impact (-3.2% change in waiting time), ensuring that the AI signal enhances priority without overstepping clinical boundaries or overriding the severity-based triage rules.

Technical context: the role of AI signals in clinical triage. When integrating machine learning predictions into clinical decision support systems or priority queues, balancing automated scores with established clinical protocols is critical. **Tie-breaking mechanism.** By assigning a bounded weight ζ_i to the AI prediction $\hat{p}_i^{\mathrm{ND}}$ (representing a specialized non-discretionary risk or urgency score), the model only differentiates patients who share the exact same baseline severity classification (e.g., the same Emergency Severity Index level). This prevents edge-case misclassifications from jumping queue tiers. **Sensitivity and safety bounds.** The low sensitivity impact (-3.2% change in waiting time) confirms that the system maintains high clinical safety margins. It guarantees that deterministic, rule-based triage protocols (such as vital signs and chief complaints) remain the primary drivers of patient priority, keeping the AI strictly in a supportive, auxiliary role.

6 Conclusion

This study developed an AI-enhanced multi-class, multi-server Markov–Volterra queueing framework for emergency-department resource allocation. The framework integrates heterogeneous healthcare servers, stochastic environmental modulation, memory-dependent service dynamics, accumulating priority, and predictive operational information within a unified mathematical formulation.

Simulation results indicate that the proposed priority and allocation mechanism can be implemented computationally under the stated assumptions. The empirical AI results are preliminary and proof-of-concept only; they do not establish clinical effectiveness, deployment readiness, or superiority over existing triage systems. Further validation with larger longitudinal ED datasets is required.

The empirical component provides preliminary evidence for the feasibility of incorporating machine-learning predictions into the scheduling mechanism. Among the three evaluated classifiers, random forest achieved the strongest overall predictive performance, with an accuracy of 80.0% and an ROC-AUC of 0.7391 under leave-one-out cross-validation. However, these findings are based on only 35 observations with documented disposition outcomes and should therefore be interpreted as proof-of-concept evidence rather than definitive clinical prediction performance.

The computational results suggest that the proposed policy may reduce waiting time and improve utilization relative to benchmarks, with mean waiting time improvements of 20.2–45.5% and statistically significant differences ($p < 0.001$). These results are based on simulated parameters and should be validated with larger longitudinal datasets.

The principal contribution of the study is the integration of predictive operational information with accumulating priority, heterogeneous-server allocation, stochastic environmental modulation, and Volterra memory. The framework is intended to support, rather than replace, clinical and operational decision-making.

Future research should evaluate the framework using larger longitudinal emergency-department datasets containing complete arrival, triage, service-start, service-completion, server-assignment, and discharge timestamps. Such datasets would permit direct estimation of temporal queueing parameters and more rigorous validation of the proposed AI-enhanced scheduling strategy against established operational policies.

Author Contributions

Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft preparation, visualization: A.A.M; Writing—review and editing, supervision: N.T.; Writing—review and editing, resources, Supervision: E.T.A. All authors have read and agreed to the published version of the manuscript.

AI-declaration

During preparation of this manuscript, the authors used AI

tool Writefull for Ms words and Grammarly , for some language editing, grammatical improvement, structural refinement and paraphrasing. The AI tool was not used as an independent source of scientific evidence, did not generate or alter the empirical data, and did not independently determine the statistical results or scientific conclusions. All mathematical formulations, analyses, interpretations, methodological decisions, and final manuscript content were reviewed and verified by the authors, who remain fully responsible for the accuracy and integrity of the work.

The authors confirm that Figure 1 and graphical abstract are originally generated from MATLAB code but later enhanced by using Inkscape to get better tailored image. The figure was designed by the authors using MATLAB and Inkscape, and all content was manually created, checked, and approved by the authors.

Acknowledgments

The authors thank all those who provided valuable comments and constructive suggestions that helped improve the quality of this manuscript. They also gratefully acknowledge the support received during its preparation and revision.

Funding

The authors received no financial support from public, commercial, or not-for-profit organizations for the conduct of this study.

Conflict of Interest

The authors declare that they have no competing interests or conflicts of interest related to this work.

Ethics Committee Approval and Consent to Participate

The study was approved by the Institutional Review Board (IRB)Arba Minch University (approval reference IRB/23488/26, 17 May 2026). Because the study used retrospectively collected, de-identified patient records and involved no direct patient contact, no intervention, and no identifiable private information, the IRB waived the requirement for informed consent. All records were anonymized before analysis and handled confidentially.

Human Rights Statement

The study was conducted in accordance with the Declaration of Helsinki and applicable national ethical guidelines. All procedures were performed in accordance with ethical standards and applicable human rights protections. Only de-identified retrospective records were used; no identifiable private information was disclosed, no intervention was performed, and no biological specimens were collected.

Availability of Data and Materials

The data are not publicly available because they contain confidential patient information and are subject to institutional privacy and confidentiality restrictions. De-identified data may be made available to qualified researchers upon reasonable request to the corresponding author, subject to approval by the IRB of Arba Minch University and Nigist Eleni Mohammed Memorial Comprehensive Specialized Hospital, and under a data-sharing agreement consistent with the ethics approval.

Reporting Checklists

This study was conducted and reported in accordance with the TRIPOD+AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) and RECORD (REporting of studies Conducted using Observational Routinely-collected health Data) guidelines. The completed TRIPOD+AI and RECORD checklists have been provided as Supplementary Files.

Data source and dataset classification

The empirical component of this study uses secondary data obtained from the emergency department records of Nigist Eleni Mohammed Memorial Comprehensive Specialized Hospital in Hosanna, Ethiopia. The dataset consists of routinely recorded emergency department information used to characterize patient arrivals, waiting times, service durations, and patient priority classes. The data were used for statistical characterization, model parameterization, calibration, and numerical evaluation of the proposed scheduling framework. No personally identifiable information was used in the analysis.

List of Abbreviations

AI: Artificial Intelligence
APQ: Accumulating Priority Queue
CTMC: Continuous-Time Markov Chain
ED: Emergency Department
ESI: Emergency Severity Index
FCFS: First-Come-First-Served
GI/M/C: General-Interarrival/Markov-Service/C-Server Queue
LOOCV: Leave-One-Out Cross-Validation
ML: Machine Learning
M/M/C: Markovian Arrival/Markovian Service/C-Server Queue
NEMMCSH: Nigist Eleni Mohammed Memorial Comprehensive Specialized Hospital
ROC-AUC: Receiver Operating Characteristic–Area Under the Curve
VIDE: Volterra Integro-Differential Equation

References

[1] Hagen MS, Jopling JK, Buchman TG & Lee EK, (2013). Priority queuing models for hospital intensive care units and impacts to severe case patients. AMIA Annu Symp Proc. 2013;841-50.

[2] Elalouf, A., & Wachtel, G. (2015). An alternative scheduling approach for improving patient-flow in emergency departments. Operations Research for Health Care, 7, 94–102.

[3] Defraeye, M., & Van Nieuwenhuysse, I. (2016).

Staffing and scheduling under nonstationary demand for service. Omega, 58, 4–25.

[4] Mehandiratta, R. (2011). Applications of queuing theory in health care. International Journal of Computing and Business Research, 2(2), 2229-6166.

[5] Kleinrock, L. (1964). A delay dependent queue discipline. Naval Research Logistics Quarterly, 11(3–4), 329–341.

[6] Rosson, H. T., & Dshalalow, J. H. (2003). A non-Markovian queuing system with a variable number of channels. 4, 375–395.

[7] Sleptchenko, A. V. (2003). Multi-class, multiserver queues with non-preemptive priorities.

[8] Prabhu, A. N. U., & Tang, L. C. (1994). Markov-modulated single-server queuing systems. 31, 169–184.

[9] Mickevicius, G., & Valakevicius, E. (2010). Modelling of non-Markovian queuing systems. 8619(2006).

[10] Mohammed, A., Natesan, T., & Ayele, E. T. (2026). Optimization of Emergency Healthcare System with Priority Based Queuing Model and Dynamic Programming. Sakarya University Journal of Computer and Information Sciences, 650–673.

[11] Deepa, T., Maragathasundari, S., & Swedheetha, C. (2025). Comprehensive modeling and simulation of non-Markovian queuing systems within health care networks. AIP Conference Proceedings, 3325(1), 040027.

[12] Ait Braham, K., Taleb, S., & Aissani, A. (2023). Performance analysis of a non-Markovian queue with differentiated vacations. SN Computer Science, 4(5), 512.

[13] Ayyappan, G., & Nithya, S. (2025). Analysis of Non-Markovian Batch Arrival Retrial Queue with Priority Services. International Journal of Mathematics in Operational Research, 31(2), 143–163.

[14] Kostic, M. (2019). Weyl-almost periodic solutions of abstract Volterra integro-differential equations.

[15] Akgun, F. A., & Rasulov, Z. (2025). Numerical Method For Solution Of Fourth-Order Volterra Integro-Differential Equations. Kragujevac Journal Of Mathematics, 49(3), 485–494.

[16] Amirali, I., Durmaz, M. E., & Amiraliyev, G. M. (2024). A numerical approach for solving Volterra delay integro-differential equation. Miskolc Mathematical Notes, 25(2), 553–562.

[17] Semiyari, H., & Shafer, D. (2016). Symbolic Iterative Solution of Volterra Integral Equations. arXiv:1611.04062.

[18] Shu, L., et al. (2021). Existence and exponential stability of mild solutions. J. Appl. Anal. Comput, 11(1), 59–80.

[19] Ricciuti, C., & Toaldo, B. (2017). Semi-Markov models and motion in heterogeneous media. Journal of Statistical Physics, 169(2), 340–361.

[20] Fiems, D., & Altman, E. (2013). Markov-modulated stochastic recursive equations. Performance Evaluation, 70(11), 965–980.

[21] Kim, B., & Kim, J. (2015). A single server queue with Markov modulated service rates. Performance Evaluation, 83–84, 1–15.

[22] Majumder, A. P. (2024). Long time behavior of a stochastically modulated infinite server queue. arXiv:2410.21910.

[23] Sharif, A. Bin, Stanford, D. A., Taylor, P., & Ziedins, I. (2014). A multi-class multi-server accumulating priority queue with application to health care. Operations Research for Health Care, 3(2), 73–79.

[24] Stanford, D. A., Taylor, P., & Ziedins, I. (2014). Waiting time distributions in the accumulating priority queue. Queueing Systems, 77(3), 297–330.

[25] Shanthikumar, J. G., & Yao, D. D. (1992). Multiclass queueing systems. Operations Research, 40(2), 293–299.

[26] Liu, Z., Nain, P., & Towsley, D. (1992). On optimal polling policies. Queueing Systems, 11(1–2), 59–83.

[27] Chaudhary, H., et al. (2025). AI-enhanced modelling of queueing and scheduling systems in cloud computing. Discover Applied Sciences, 7(4), 276.

[28] Ravi, M., et al. (2025). Evolution of AI-driven decision making. IETE Technical Review, 42(4), 428–465.

[29] Hui, M. K., & Tse, D. K. (1996). What to Tell Consumers in Waits of Different Lengths. Journal of Marketing, 60(2), 81–90.

[30] Saure, A., Begen, M. A., & Patrick, J. (2020). Dynamic multi-priority, multi-class patient scheduling. European Journal of Operational Research, 280(1), 254–265.

[31] Tahernejad, A., et al. (2024). Application of artificial intelligence in triage. BMC Public Health, 24(1), 3203.

[32] Bouchentouf, A. A., Cherfaoui, M., & Boualem, M. (2021). Analysis and performance evaluation of Markovian feedback multi-server queueing model with vacation and impatience. American Journal of Mathematical and Management. https://doi.org/10.1080/01966324.2020.1842271

[33] Boualem, M., & Touche, N. (2021). Stochastic monotonicity approach for a non-Markovian priority retrial queue. Asian-European Journal of Mathematics, 14(9), 2150156.

[34] Asingwire, J. B., Mwendwa, P. K., & Kabanda, J. (2021). Prioritized Scheduling Scheme for IoT-based Healthcare Systems. Journal of Electrical and Computer Engineering, 2021, 1–12

[35] Boualem, M., Bouchentouf, A. A., Cherfaoui, M., & Aissani, D. (2017). Multi-station manufacturing system analysis. IJMOR, 10(2), 229–260.

[36] Efrosinin, D., Vishnevsky, V., Stepanova, N., & Sztrik, J. (2025). Use cases of machine learning in queueing theory based on a GI/G/K system. Mathematics, 13(5), 776.

[37] Nguyen, G., et al. (2019). Machine Learning and Deep Learning frameworks. Artificial Intelligence Review, 52(1), 77–124.

[38] Da'Costa, A., et al. (2025). AI-driven triage in emergency departments. International Journal of Medical Informatics, 197, 105838.

[39] Vignesh, P., Srinivasan, S., & Maragatha Sundari, S. (2017). Analysis of Non-Markovian Batch Arrival Queueing Model. IJMOR, 11(3), 285–309.

[40] Jouini, O., Dallery, Y., & Aksin, Z. (2009). Queuing models for full-flexible multi-class call centers. International Journal of Production Economics, 120(2), 389–399.

[41] Kleinrock, L., & Finkelstein, R. P. (1967). Time dependent priority queues. Operations Research, 15(1), 104–116.