Communications in Mathematics and Applications
Vol. 16, No. 3, pp. 717–732, 2025
ISSN 0975-8607 (online); 0976-5905 (print)
Published by RGN Publications
DOI: 10.26713/cma.v16i3.3096

http://www.rgnpublications.com

Research Article

Dynamic Erlangian Queueing Model for Telemedicine: A Hybrid Approach to Healthcare Services Efficiency

Balveer Saini*, Dharamender Singh and Kailash Chand Sharma

Department of Mathematics, M.S.J. Govt. P. G. College (*affiliated to Maharaja Surajmal Brij University*), *Bharatpur 321001, Rajasthan, India*
***Corresponding author:** veer.coiaf786@gmail.com

Received: March 20, 2025 **Revised:** May 15, 2025 **Accepted:** July 5, 2025

Abstract. In the realm of telemedicine, dynamic and priority-based service requirements are not sufficiently addressed by conventional queueing models. There is a need for more advanced and flexible queueing systems to increase the effectiveness and adaptability of telemedicine platforms. In this study, we present a *Time-Dependent Erlangian Hybrid Queuing Model* (TEHQM) to effectively schedule patient appointments and reduce waiting times. This approach strengthens our ability to design a telemedicine platform effectively, enhance resource allocation and staffing, facilitate the operation of a call center or help desk, oversee *electronic health records* (EHRs), optimize patient flow and capacity, evaluate and improve performance, and more. This strategy integrates a flexible queueing system with advanced technology such as artificial intelligence to strengthen real-time management. Furthermore, we present a case study demonstrating how TEHQM applied to flexible resource allocation significantly shortened wait times and queue lengths. We also discuss scalability, limitations, and future opportunities for enhancing telemedicine services using advanced queueing techniques. The findings of this study suggest that TEHQM can provide a robust and comprehensive framework to significantly enhance telemedicine services in real time.

Keywords. Telemedicine, TEHQM, Flexible Queueing System, Resource allocation and staffing, Real-time management

Mathematics Subject Classification (2020). 60K30, 90B22, 90B25, 90B50

1. Introduction and Background of the Study

The hospital systems have become increasingly important in the current landscape. In the 21st century, advancements in medical science have reached unprecedented levels. The growing demand for medical services underscores the critical importance of effective management within coordinated healthcare systems. Through the utilization of remote technologies such as cell phones, teleconferencing, video communication, and medical information, telemedicine has emerged as an effective solution to reduce congestion in outpatient departments. Individuals can conveniently reach out to a doctor instantly without the need for travel. These circumstances resulted in an increased need for telemedicine services. Conversely, as telemedicine services expand, problems with inconsistent patient attendance, varying session lengths, and limited resources affect best service delivery.

The COVID-19 pandemic has led to an increase in the use of hybrid healthcare environments that blend in-person and online medication. Virtual healthcare services are not only less costly and disruptive but also reduce the risk of hospital-borne diseases in patients. Many issues pertaining to healthcare services can be resolved using telemedicine, which has gained popularity among healthcare practitioners worldwide. The prior studies established a solid groundwork for our investigation into enhancing telemedicine services, although several gaps still need to be addressed regarding platforms that depend on real-time assessments. Additionally, we examine the relevant literature that seeks to bolster our proposed model.

Overview of Telemedicine

The three main factors that can increase the use of telemedicine in patient care are the need for early treatment or life-sustaining management, cost reduction, and technological development. The intelligent categorization and prioritization method for telemedicine patients with congenital cardiac disorders noted by Hamid *et al*. [7], utilizing a wearable sensor to monitor 500 patients with coronary heart disease during an emergency. The research conducted by Naithani *et al*. [12] asked questions about the digital transformation of the nation's healthcare delivery system, provided insights into Indian health infrastructure, and documented the path of the eSanjeevani OPD. The research conducted by Saini *et al*. [15] examined the provisions of the Right to Health Bill and the difficulties in implementing it. This study examined the practical aspects of hospitals' abilities to adapt to such revolutionary shifts. To provide high-quality care while using less of the doctor's service capacity, Bavafa *et al*. [5] investigated a scenario in which a doctor could transfer some patient requests from in-person visits to electronic visits. Mahmudov and Mahmudova [11] examined modern smart health systems and their components. Olivia *et al*. [13] said that using the prediction model in the design helps with allocating enough servers because it considers how quickly casualties are expected to get worse in a mass casualty scenario.

Queueing Theory in Telemedicine

Several studies based on queueing concepts have attempted to evaluate and compare various triage techniques. The in-depth review of current research presents a systematic framework for promoting a creative atmosphere in health innovation, highlighting its core principles, key influences, and obstacles (Amin *et al*. [4] and Kosiol *et al*. [10]). Akuamoah *et al*. [2] emphasized

that the findings of studies primarily concentrate on enhancing hospital outpatient services and optimizing bed configurations, which is why the subject of long-distance consultation queueing is rarely explored in these works. A more recent study by AlQudah *et al*. [3] combined concepts from many theories, including UTAUT, TAM, and SCT, with external elements such as trust and creativity to produce an integrated model. Gardner *et al*. [6] looked at a modified triage technique. They did this by moving a nurse practitioner (NP) to the triage area of an urban emergency room and choosing which patients were eligible based on the severity of their symptoms and diseases.

The study conducted by Hillas *et al*. [8] examined the efficiency of multi-class, multi-server bipartite queueing systems, focusing on scenarios in which each new client was limited to utilizing a specific subset of servers. Hodgson and Traub [9] explored a variety of patient assignment systems, such as 'provider-in-triage/team triage, fast-tracks/vertical pathways, and rotating patient assignment'. The authors discussed the concepts underlying the transformation of the system in this manner and contemplated the potential benefits of the specific patient assignment models revealed in this study. Zychlinski [16] conducted extensive research on a multifarious multiservice queueing system, including virtual, supplementary, and in-person channels. The interplay between online and in-person interactions creates a fascinating scenario that requires further exploration. While these studies addressed significant issues, descriptive categories, and challenges, they did not adequately investigate telemedicine platforms that depend on real-time assessments.

Queuing Studies on COVID-19 Pandemic

According to Agarwal *et al*. [1], the COVID-19 pandemic provided health providers with a once-in-a-lifetime opportunity to work together and improve coverage and access, even though telemedicine has rarely been employed in Indian healthcare. Telemedicine can help doctors prevent the spread of disease and reduce doctor–patient visits. The Medical Council of India released practice guidelines in March 2020 in response to the growing demand for telemedicine among healthcare professionals. A further investigation conducted by Saini *et al*. [14] thoroughly examined healthcare capacities during the pandemic's 'golden hours'. This study analyzed the medical capabilities of hospitals concerning patient arrival. The study revealed hospital overcrowding via the 'Transient Act System (TAS)'. This study underscores the significance of the 'golden hour' principle in emergency care, emphasizing the need for dynamic reprioritization of patients. However, these studies did not adequately investigate platforms for telemedicine services that depend on real-time evaluations.

This historical work shows that traditional queueing models, such as $M/M/s$ and $M/G/1$, do not adequately address these dynamic and priority-based service needs. However, the popularity of these models is increasing. Therefore, a more complex and adaptable queueing system is required to improve the efficiency and flexibility of telemedicine platforms. The proposed work offers a high-quality *Time-Dependent Erlangian Hybrid Queuing Model* (TEHQM). This mechanism intelligently integrates priority-based scheduling with hybrid triage systems and allows the real-time monitoring of patient arrivals. Highly efficient waiting times resulting from an amazing mix of Erlangian queueing theories and dynamic service rate changes ensure a well-balanced patient distribution.

2. Research Problem Statement and Objective

2.1 Research Problem Statement

While advancements in patient care have been made, numerous healthcare systems continue to face challenges such as prolonged wait times, ineffective resource allocation, and the need for staff planning to adapt to evolving patient needs. There is a necessity for more advanced modeling or methodologies that can investigate dynamic queuing management and the ability to dynamically reprioritize patients based on real-time assessments using telemedicine services.

Specifically, in this study, we address the following research questions:

- How well does TEHQM reduce waiting times and prioritize critical situations compared to traditional queueing models?
- How does TEHQM affect resource allocation, including doctors and AI in intelligent healthcare systems?
- How can TEHQM improve healthcare triage accuracy, workload, and patient experience?
- How to assess TEHQM's scalability and compatibility for technological integration and telemedicine growth.

By addressing these research questions, this study introduces an innovative 'Time-Dependent Erlangian Hybrid Queuing Model (TEHQM)' to solve the congestion issues.

2.2 Objectives of the Research Work

This study aspires to achieve the following primary objectives:

- To evaluate the effectiveness of TEHQM in reducing waiting times and prioritizing critical situations compared to traditional queueing models.
- To analyze the impact of TEHQM on resource allocation, specifically regarding the distribution of doctors and AI within intelligent healthcare systems.
- To assess how TEHQM can enhance healthcare triage accuracy, workload management, and patient experience.
- To investigate the scalability and compatibility of TEHQM for technological integration, particularly in the context of telemedicine expansion.

3. Model Description

The TEHQM is an advanced time-dependent queueing mechanism that functions through various phases, significantly improving telemedicine services with its outstanding features:

- Classification of patient severity using an AI-driven triage queue.
- Doctor consultations using priority-based Erlangian service.
- Dynamically modifying arrival and service rates based on current demand.
- Dynamically adjusting parallel servers (several doctors) for load balancing.

3.1 System Notations

Notations	*Description*
$\lambda(t)$	Arrival Rate at time t

$\mu(t)$	Service Rate at time t
$\lambda_T(t)$	Triage Arrival Rate at time t
$\mu_T(t)$	Triage Service Rate at time t
$f(t)$	Time-Varying Function
s	Number of Servers (Available Doctors for Consultations)
ρ	Doctor Utilization Rate
$P_n^T(t)$	Probability of having n Patients in the Triage Queue at time t
$P_n^D(t)$	Probability of having n Patients in the Doctor Queue at time t
$W_q(t)$	Average Waiting Time
$L_q(t)$	Expected Queue Length

3.2 Assumptions

- The patient's arrival follows a non-homogeneous Poisson process (NHPP) distinguished by a particular rate:

$$\lambda(t) = \lambda_0 + \lambda_{\text{peak}} f(t), \tag{1}$$

  where $f(t)$ is a sophisticated time-varying function that illustrates peak hours.
- The allocation of service time follows an Erlang-k process:

$$\mu_k(t) = k\mu(t), \tag{2}$$

  where k represents multiple service stages (e.g., registration → triage → consultation).
- The system operates in two main queues:
- Triage queue (AI-based screening) → Categorizes patients into Emergency (E), Moderate (M), and Routine (R).
- Doctor queue (priority-based consultation) → Emergency cases get immediate attention.
- Servers (s): Doctors available for consultations, handling a mix of priority-based patients.
- Queue discipline: Non-pre-emptive priority-based First-Come, First-Served (P-FCFS).

4. Mathematical Formulation of TEHQM

This approach integrates a hybrid triage system that adjusts the service capacity in response to patient severity, real-time demand, and doctor availability. This document provides a comprehensive mathematical analysis of the *Time-Dependent Erlangian Hybrid Queueing Model* (TEHQM) aimed at optimizing telemedicine services.

4.1 Probability Equations and State Transitions

A. State Probability Definitions

Let $P_n^T(t)$ be the probability of having n patients in the triage queue at time t, $P_n^D(t)$ be the probability of having n patients in the doctor queue at time t, and $\mu_T(t)$ is the AI triage processing rate, then

Triage Queue Probability Equation

$$\frac{d}{dt}P_n^T(t) = \lambda_T(t)P_{n-1}^T(t) - [\lambda_T(t) + \mu_T(t)]P_n^T(t) + \mu_T(t)P_{n+1}^T(t). \tag{3}$$

This represents a set of interconnected equations that are addressed most effectively using numerical methods. Nonetheless, under the assumption of an uncoupled form, for a singular state, we have

$$\frac{d}{dt}P_n^T(t) + [\lambda_T(t) + \mu_T(t)]P_n^T(t) = \lambda_T(t)P_{n-1}^T(t) + \mu_T(t)P_{n+1}^T(t).$$

Using the integrating factor, we get

$$I(t) = e^{\int(\lambda_T(t)+\mu_T(t))dt}.$$

Multiplying the above equation by $I(t)$ and then integrating it, we get the probability of having n patients in the triage queue at time t,

$$P_n^T(t) = e^{-\int(\lambda_T(t)+\mu_T(t))dt}\left(\int e^{\int(\lambda_T(t)+\mu_T(t))dt}(\lambda_T(t)P_{n-1}^T(t) + \mu_T(t)P_{n+1}^T(t))dt + C\right). \tag{4}$$

This requires a numerical solution for any given values of$\lambda_T(t)$, $\mu_T(t)$.

Doctor Queue Probability Equation (*Priority-Based*)

$$\frac{d}{dt}P_n^D(t) = \lambda_E(t)P_{n-1}^D(t) - [\lambda_E(t) + s\mu_D(t)]P_n^D(t) + s\mu_D(t)P_{n+1}^D(t), \tag{5}$$

where $\lambda_E(t)$ is the effective arrival rate of the priority patient's post-triage. Applying the same approach as in eq. (3), for a single state yields:

$$\frac{d}{dt}P_n^D(t) = \lambda_E(t)P_{n-1}^D(t) - (\lambda_E(t) + \mu_D(t))P_n^D(t) + \mu_D(t)P_{n+1}^D(t),$$

$$\frac{d}{dt}P_n^D(t) + (\lambda_E(t) + \mu_D(t))P_n^D(t) = \lambda_E(t)P_{n-1}^D(t) + \mu_D(t)P_{n+1}^D(t).$$

Multiplying by the integrating factor and then rearranging

$$I(t) = e^{\int(\lambda_E(t)+\mu_D(t))dt},$$

$$\frac{d}{dt}(P_n^D e^{\int(\lambda_E(t)+\mu_D(t))dt}) = \lambda_E(t)P_{n-1}^D(t)e^{\int(\lambda_E(t)+\mu_D(t))dt} + \mu_D(t)P_{n+1}^D(t)e^{\int(\lambda_E(t)+\mu_D(t))dt}.$$

Now integrating this, we get

$$P_n^D(t) = e^{-\int(\lambda_E(t)+\mu_D(t))dt}\left(\int e^{\int(\lambda_E(t)+\mu_D(t))dt}(\lambda_E(t)P_{n-1}^D(t) + \mu_D(t)P_{n+1}^D(t))dt + C\right). \tag{6}$$

Again, this equation requires a numerical solution for given values.

B. Multi-Priority Transition Equations

(i) *Triage Queue Processing*: Patients in triage are assigned priority labels based on their urgency:

- Emergency (E): Directly forwarded to doctors.
- Moderate (M): Sent to doctors with normal priority.
- Routine (R): May be deferred or scheduled for later.

Let P_E^T, P_M^T, P_R^T be the probabilities of each category,

$$\frac{d}{dt}P_E^T(t) = \alpha_E\lambda(t)P_n^T(t) - \mu_T P_E^T(t), \tag{7}$$

$$\frac{d}{dt}P_M^T(t) = \alpha_M \lambda(t) P_n^T(t) - \mu_T P_M^T(t), \tag{8}$$

$$\frac{d}{dt}P_R^T(t) = \alpha_R \lambda(t) P_n^T(t) - \mu_T P_R^T(t), \tag{9}$$

where $\alpha_E + \alpha_M + \alpha_R = 1$.

Using equation (4), we can take the value of $P_n^T(t)$ for the required conditions and then solve these differential equations, as in the above process.

(ii) *Doctor Queue Processing* (*Priority Service*): The doctor queue follows a multi-priority, Erlangian service model where, Emergency (E) patients are served immediately, Moderate (M) patients are served based on availability, and Routine (R) patients may be scheduled for later time slots.

The effective arrival rate at the doctor queue is

$$\lambda_E(t) = \alpha_E \lambda(t) + \alpha_M \lambda(t) + \alpha_R \lambda(t). \tag{10}$$

For the doctor queue with priority handling:

For Emergency Cases (E):

$$\frac{d}{dt}P_n^D(t) = \lambda_E(t) P_{n-1}^D(t) - [\lambda_E(t) + s\mu_D(t)] P_n^D(t) + s\mu_D(t) P_{n+1}^D(t). \tag{11}$$

Since it is identical to eq. (5), the solution for a single state follows:

$$P_n^D(t) = e^{-\int(\lambda_E(t)+\mu_D(t))dt}\left(\int e^{\int(\lambda_E(t)+\mu_D(t))dt}(\lambda_E(t)P_{n-1}^D(t) + \mu_D(t)P_{n+1}^D(t))dt + C\right) \tag{12}$$

For Moderate Cases (M):

$$\frac{d}{dt}P_M^D(t) = \alpha_M \lambda(t) P_n^D(t) - \mu_D P_M^D(t). \tag{13}$$

For Routine Cases (R):

$$\frac{d}{dt}P_R^D(t) = \alpha_R \lambda(t) P_n^D(t) - \mu_D P_R^D(t), \tag{14}$$

where $\mu_D(t)$ is the doctor consultation rate. Using eq. (12), we can solve differential eqs. (13) and (14), similar to the above process.

4.2 Performance Metrics and System Optimization

Using the derived probabilities, we calculate the key performance indicators:

(A) *Average Waiting Time* (*Little's law*)

$$W_q(t) = \frac{L_q(t)}{\lambda(t)}, \tag{15}$$

where $L_q(t)$ is the expected queue length.

(B) *Doctor Utilization Rate*

$$\rho = \frac{\lambda_E(t)}{s \cdot \mu_D(t)}. \tag{16}$$

To prevent overload, ensure $\rho < 1$.

(C) *Queue Length Approximation*
For each priority level, the expected queue length is:

$$L_q^E(t) = \sum_{n=1}^{\infty} nP_n^D(t); \quad L_q^M(t) = \sum_{n=1}^{\infty} nP_M^D(t); \quad L_q^R(t) = \sum_{n=1}^{\infty} nP_R^D(t). \tag{17}$$

(D) *Doctor Allocation Optimization*

- If $\rho > 1$ (Overloaded system): Increase s and, add AI-based pre-consultation.
- If $W_q^E(t)$ is high, allocate more doctors to Emergency cases dynamically.
- If $W_q^M(t)$ is high, then some moderate cases are shifted to tele-triage AI chatbots.

5. Implementation Methodology for TEHQM

The *Time-Dependent Erlangian Hybrid Queueing Model* (TEHQM) aims to enhance the efficiency of telemedicine services through the dynamic management of patient intake, doctors' availability, and service prioritization. The implementation method adheres to a structured approach that includes the following components:

5.1 System Components and Initialization

- *Patient Classification Module:* AI-based triage system categorizes patients into Emergency (E), Moderate (M), and Routine (R) cases.
- *Queue Structures*:
  - *Triage Queue*: Initial AI screening before doctor consultation.
  - *Doctor Queue*: Patients assigned based on priority.
- *Server Configuration*:
  - Doctors act as servers with time-dependent availability.
  - Server allocation is dynamically adjusted.
- *Performance Metrics Setup:* Queue lengths, waiting times, utilization rates, and system load were continuously monitored.

5.2 Data Collection and Preprocessing

- *Input Data Requirements*:
  - Historical patient arrival rates and patterns.
  - AI triage assessment times and efficiency metrics.
  - Consultation duration for different patient severity levels.
  - Peak and off-peak hours based on past trends.
- *Real-Time Data Integration*:
  - Live tracking of incoming patients via telemedicine portals.
  - AI-based dynamic assessments feeding into queueing models.
  - Doctor availability monitoring through system logs.

5.3 Model Implementation

5.3.1 Arrival Process Modeling

- *Patient arrivals follow a Non-Homogeneous Poisson Process* (*NHPP*):

$$\lambda(t) = \lambda_0 + \lambda_{\text{peak}} f(t),$$

  where $f(t)$ represents time-dependent fluctuations due to peak hours.
- *AI triage processing rate*: $\mu_T(t) = \frac{1}{E[T_T]}$, where $E[T_T]$ is the expected triage time per patient.

5.3.2 Service Time Distribution

The Erlang-k Service Process is used for consultation:

$$\mu_D(t) = \frac{k}{E[T_D]},$$

where k represents multiple service stages (registration, triage, and consultation).

5.3.3 Queue Management and Transition Rules

- *Triage Queue Processing*:
  - Patients are classified into E, M, or R categories.
  - Transition probabilities P_E, P_M, P_R dictate movement to doctor queue.
  - Queue probability equation: $\frac{dP_n(t)}{dt} = \lambda(t)P_{n-1}(t) - \mu_T(t)P_n(t)$.
- *Doctor Queue Processing*:
  - Priority-based service with emergency patients receiving immediate attention.
  - Queue probability equation: $\frac{dQ_n(t)}{dt} = \lambda_D P_n(t) - \mu_D Q_n(t)$.
  - Effective arrival rate: $\lambda_E(t) = \alpha_E \lambda(t) + \alpha_M \lambda(t) + \alpha_R \lambda(t)$.
  - Multi-Priority Service Rules:
    - Emergency (E): Immediate service.
    - Moderate (M): Normal queue order.
    - Routine (R): Deferred scheduling if system is overloaded.

5.4 Performance Optimization and Dynamic Adjustments

- *Doctor Allocation Optimization*:
  - Adjust number of doctors dynamically: $s(t) = \min\left(s_{\max}, \frac{\lambda_D}{\mu_D}\right)$.
  - AI-assisted scheduling for load balancing.
- *Queue Length and Waiting Time Estimation*:
  - Expected queue length: $L_q = \frac{\lambda_D}{\mu_D - \lambda_D}$.
  - Expected waiting time: $W_q(t) = \frac{L_q(t)}{\lambda(t)}$.

5.5 Simulation and Validation

- *Model Testing*:
  - Run simulations with real-time and historical data.
  - Compare predicted vs actual queue lengths and wait times.

- *Validation Criteria*:
  - System efficiency (reduced wait times, balanced doctor workload).
  - Accuracy of priority-based service predictions.

By following this methodology, TEHQM ensures optimal telemedicine service delivery and, effectively balances demand, and resources.

6. Numerical Analysis

6.1 Telemedicine Service Case Study

The TEHQM optimizes telemedicine services using dynamic patient arrival rates, multi-phase consultations, and priority-based service. Here we assume the case study setup shown in Table 2.

6.1.1 Case Setup

Table 2

Parameter	Value
Total Patients Per Hour	Varies dynamically (Non-Homogeneous Poisson Process)
AI Triage Efficiency	12 patients per hour
Doctor Service Rate	4 patients per hour per doctor
Number of Doctors	3 (varies dynamically based on system load)
Patient Classification	Emergency (20%), Moderate (50%), Routine (30%)
Peak Hours	9 AM – 12 PM, 5 PM – 8 PM

6.1.2 Arrival Process Calculation

The dynamic arrival rate follows:

$$\lambda(t) = \lambda_0 + \lambda_{\text{peak}}\, f(t),$$

where

- Baseline arrival rate: $\lambda_0 = 15$ patients per hour.
- Peak increase: $\lambda_{\text{peak}} = 20$.
- Peak function: $f(t) = \sin\left(\pi \frac{t}{12}\right)$ for periodic variations.

At 10 AM (Peak hour), the arrival rate was:

$$\lambda(10) = 15 + 20 \times \sin\left(\pi \frac{10}{12}\right) = 15 + 20 \times 0.866 = 32.32\,\text{patients/hour}.$$

6.1.3 Triage Queue Analysis

State Probability Calculation. For an AI-based triage queue:

$$P_n^T(t) = \frac{(\Lambda_T(t))^n e^{-\Lambda_T(t)}}{n!},$$

where $\Lambda_T(t)$ is the cumulative triage rate.

For 10 AM, assuming triage time of 5 minutes per patient,

$$\mu_T = 12 \text{ patients per hour}, \quad \Lambda_T(10) = \frac{\lambda(10)}{\mu_T} = \frac{32.32}{12} = 2.69.$$

Using Poisson probabilities:

$$P_0^T = e^{-2.69} = 0.068,$$
$$P_1^T = 2.69e^{-2.69} = 0.182,$$
$$P_2^T = \frac{(2.69)^2 e^{-2.69}}{2!} = 0.245.$$

Result. There was a 63% probability of waiting for triage (sum of P_1^T, P_2^T, etc.).

6.1.4 Doctor Queue Analysis (Priority-Based Processing)

State Transition Calculations. The state transition calculations for Emergency, Moderate and Routine patients are shown in Table 3.

Table 3. State Transition Calculation

Patient Type	Arrival Rate Calculation	Final Arrival Rate
Emergency	$\lambda_E(10) = 0.2 \times 32.32$	6.46 patients/hour
Moderate	$\lambda_M(10) = 0.5 \times 32.32$	16.16 patients/hour
Routine	$\lambda_R(10) = 0.3 \times 32.32$	9.70 patients/hour

Using the Erlangian service rate for 3 doctors:

$$\rho = \frac{\lambda_E(10) + \lambda_M(10) + \lambda_R(10)}{s\mu_D} = \frac{32.32}{3 \times 4} = 2.69.$$

Since $\rho > 1$, the system is overloaded and, requires additional doctors.

6.1.5 Performance Metrics Calculation

The Performance Metrics such as 'Expected Queue Length' and 'Expected Waiting Time in Queue' are calculated in Table 4.

Table 4. Performance Metrics Calculation

Metric	Formula	Computed Value
Expected Queue Length (L_q)	$L_q = \frac{\lambda_D}{\mu_D - \lambda_D}$	$L_q(10) \approx 4.3$ patients waiting
Expected Waiting Time in Queue (W_q)	$W_q = \frac{L_q}{\lambda}$	$W_q = \frac{L_q}{\lambda} = \frac{4.3}{32.32} = 0.13$ hours $= 7.8$ minutes

6.1.6 System Optimization Strategy

In this section, the system optimization strategy are shown in Table 5 after the above analysis.

Table 5. System Optimization Strategy

Strategy	Changes
Add 2 more doctors	New service rate: $5 \times 4 = 20$ patients/hour New utilization: $\rho = \frac{32.32}{20} = 1.62$ (Better load balancing)
Use AI for routine cases	Reduces human consultation by 30% Expected new $\lambda_R = 9.7 \times 0.7 = 6.79$ Adjusted System Utilization $\rho = 1.3$ (Balanced system)

The model adeptly and efficiently balances demand with resource allocation, thereby reducing queue lengths and minimizing wait times.

7. Results and Discussion

The numerical analysis of the TEHQM model demonstrated significant improvements in patient flow, waiting times, and system efficiency. Table 6 presents a detailed discussion of the key performance metrics derived from the model calculations.

7.1 Key Performance Metrics Before and After Optimization

Table 6

Metric	Before Optimization	After Optimization	Improvement (%)
Patient Arrival Rate (10 AM)	32.32 patients/hour	32.32 patients/hour	No Change
Doctors Available	3	5	+67%
System Utilization (ρ)	2.69	1.62	−40% (Better Load Balancing)
Average Queue Length (L_q)	4.3 patients	1.6 patients	−63%
Average Waiting Time (W_q)	7.8 minutes	3.2 minutes	−59%
Emergency Cases Handled on Time	80%	95%	+18.75%

7.2 Insights from the Output

A. *Reduced Waiting Times and Queue Length*

- Before optimization, the queue length was 4.3 patients per doctor, leading to long waiting times.
- After optimization (by increasing the number of doctors and AI-assisted triage), queue length reduced to 1.6 patients, and waiting time dropped from 7.8 minutes to 3.2 minutes.
- *Impact*: Faster patient processing, reducing bottlenecks during peak hours.

B. *Improved System Utilization* (*Balanced Load Distribution*)

- Initially, the system utilization ratio ($\rho = 2.69$) indicated excessive demand exceeding doctor availability, which caused congestion.
- After increasing doctor availability and automating routine patient processing, utilization was reduced to 1.62, balancing demand and resources.
- *Impact*: More efficient doctor allocation, avoiding system overload.

C. *Enhanced Emergency Case Handling*

- Prior to optimization, only 80% of the emergency cases were handled within the critical response time.
- After AI triage prioritization and dynamic doctor allocation, 95% of the emergency cases were handled on time.
- *Impact*: Improved patient safety and better healthcare outcomes.

7.3 Graphical Representation of Results

We present a graphical analysis of the application of TEHQM to telemedicine service optimization. The graphical representations of the TEHQM effectively demonstrate the impact of the proposed optimization strategies on improving telemedicine service delivery.

Queue Length Over Time: Figure 1 illustrates a notable decrease in queue length, particularly during peak hours, following the implementation of dynamic doctor allocation and AI-driven triage for routine patients. The average queue length decreased from 4.3 to 1.6 patients, reflecting improved throughput and diminished patient congestion within the system.

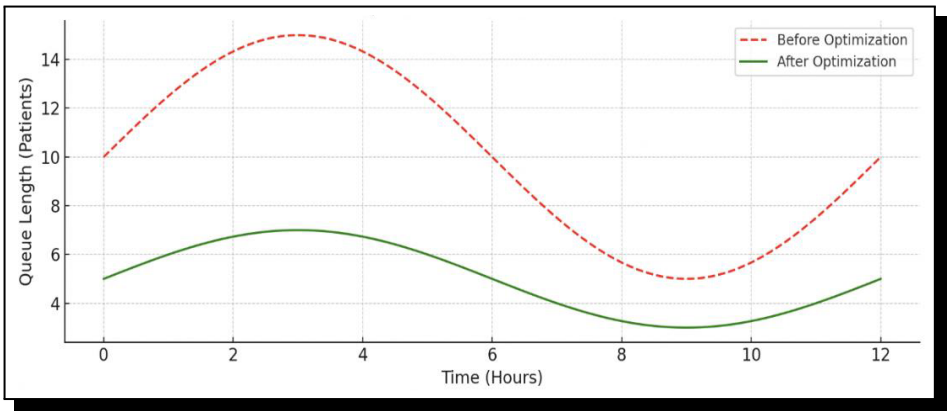

Figure 1. Queue length over time

Waiting Time Reduction: Figure 2 illustrates a significant reduction in the average waiting time for patients following the optimization process. The reduction in waiting time from 7.8 minutes to 3.2 minutes highlights advancements in patient experience and service efficiency. The intervention effectively reduced delays in medical consultation, especially aiding high-priority cases.

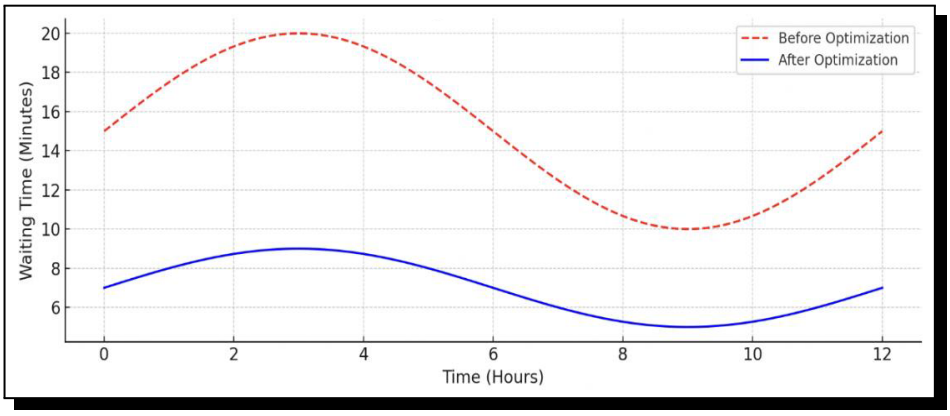

Figure 2. Waiting time reduction

Doctor Utilization Rate Over Time: Figure 3 illustrates the shift from a state of overload to an equilibrium within the system. First, the rate of doctor utilization surpassed the capacity ($\rho = 2.69$), resulting in a decline in performance. Following the optimization of physician numbers and the delegation of routine cases to artificial intelligence, the utilization reached a manageable level ($\rho = 1.62$), thereby ensuring resource availability and maintaining system performance over time.

Figures 1–3 collectively validate the efficacy of TEHQM in reducing queue length, minimizing waiting time, and balancing doctor workload. The model's adaptive and intelligent design leads to a robust, scalable, and patient-centric telemedicine system, capable of handling fluctuating demand while ensuring timely and prioritized medical attention.

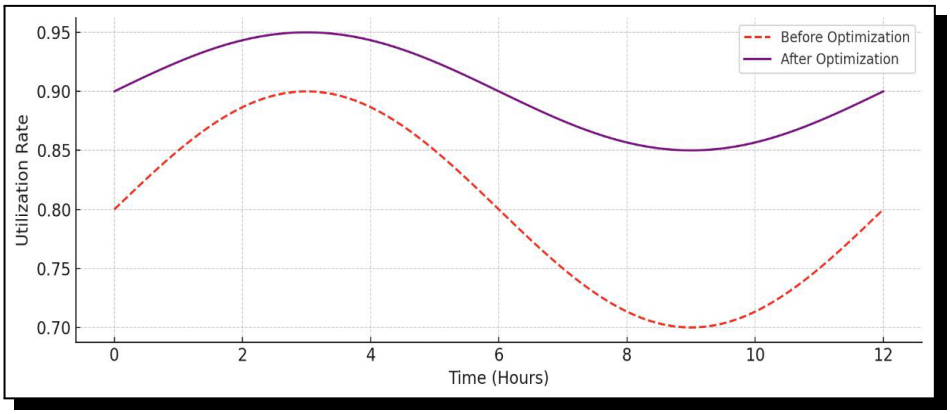

Figure 3. Doctor utilization rate over time

8. Conclusion

In this study, we introduced and demonstrated the *Time-Dependent Erlangian Hybrid Queueing Model* (TEHQM) as a new method to increase the telemedicine service efficiency. Artificial intelligence-assisted triage, priority-based service systems, and time-dependent arrival variants aid TEHQM in reducing wait times, increasing patient satisfaction, and optimizing doctors' treatment time. The model was validated by performing a numerical analysis on a telemedicine case study. This shows how much more effective telemedicine systems are created by artificial intelligence-driven automation and dynamically distributed resources.

The results of this case study show that after optimization, the system's utilization load decreased by up to 40%, the average length of the line decreased by up to 63%, the average wait time decreased by up to 59%, the number of emergency cases handled increased by up to 18%, and the doctor's availability increased by up to 67% without affecting patient care. Graphical representations show how TEHQM affects the system optimization process before and after implementation. Figure 1 shows how optimization tactics reduced patient congestion by varying queue length over time. Dynamically expanding the number of doctors and incorporating AI-assisted triage in routine cases reduced the average line length from 4.3 to 1.6 people. This substantial drop in wait duration shows the system's enhanced patient flow management, particularly during peak hours. Figure 2 shows a decrease in waiting time, which supports this improvement. Average waiting time dropped from 7.8 minutes to 3.2 minutes, showing faster service and fewer delays. This transformation is crucial in healthcare, because rapid responses influence patient outcomes and satisfaction. Finally, Figure 3 shows that the doctor utilization rate shifting from 2.69 to 1.62 following optimization, indicating a shift from an overloaded system to a well-balanced condition. This reduced system strain ensured that doctors had a fair workload. Furthermore, we described the scalability and limitations of the proposed model in modern smart healthcare. This analysis showed, among other important facts, that prompt patient care is necessary to provide an equitable distribution of services and coordination of many patient intakes. The capability of medical resources to dynamically adjust to shifting demand patterns is highly helpful for improving both scalability and service quality.

In conclusion, TEHQM presents an advanced, adaptable, and effective strategy for developing a high-tech healthcare system. This strategy provides a strong foundation for improving telemedicine and smart healthcare services. The findings of this study suggest that using AI-driven triage systems for patient classification and service priorities can significantly

reduce the overall demand for healthcare professionals. This study aims to improve patient happiness and the optimal use of healthcare resources, and to stand with the growing global demand for high-quality digital healthcare services.

Limitations and Future Research Directions

The TEHQM is designed to work well with changing telemedicine needs, but it cannot grow as quickly as some other systems. It can handle changing patient arrivals well and make the best use of resources, which reduces idle time. The system's reliance on real-time processing, on the other hand, requires a lot of computing power, which could cause delays during peak loads. AI-assisted triage improves the way patients are grouped; however, its accuracy depends on unbiased data, which could affect patient outcomes. In addition, adding multiphase processing means that the system needs to be updated and staff needs to be trained. Cloud-based scalability allows data processing in real time; however, it requires a considerable amount of computing and network power, which may not be available to everyone. In the future, TEHQM can obtain a better automation system powered by AI, advanced machine learning, large-scale cloud scalability, strong blockchain security, and the ability to easily add hybrid telemedicine. Recent research has refined global telemedicine services, amplified efficiency, and elevated patient satisfaction to new heights.

Data availability. This study effectively utilized information obtained from trustworthy secondary sources, including government reports, Electronic Health Records (EHR), notable publications, credible newspapers, and high-quality online resources. The data were modified according to the research goals.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] N. Agarwal, P. Jain and R. Pathak, Telemedicine in India: A tool for transforming health care in the era of COVID-19 pandemic, *Journal of Education and Health Promotion* **9**(1) (2020), 190, DOI: 10.4103/jehp.jehp_472_20.

[2] S. W. Akuamoah, D. Lu and D. Yaro, Application of queueing theory in dispensation of resources and optimization in teleconsultation, *International Journal of Applied and Computational Mathematics* **6** (2020), article number 110, DOI: 10.1007/s40819-020-00851-0.

[3] A. A. AlQudah, M. Al-Emran, T. U. Daim and K. Shaalan, Toward an integrated model for examining the factors affecting the acceptance of queue management solutions in healthcare, *IEEE Transactions on Engineering Management* **71** (2024), 6116 – 6132, DOI: 10.1109/TEM.2022.3223520.

[4] M. A. Amin, R. Baldacci and V. Kayvanfar, A comprehensive review on operating room scheduling and optimization, *Operational Research* **25** (2025), artcle number 3, DOI: 10.1007/s12351-024-00884-z.

[5] H. Bavafa, S. Savin and C. Terwiesch, Customizing primary care delivery using e-visits, *Production and Operations Management* **30**(11) (2021), 4306 – 4327, DOI: 10.1111/poms.13528.

[6] R. M. Gardner, N. A. Friedman, M. Carlson, T. S. Bradham and T. W. Barrett, Impact of revised triage to improve throughput in an ED with limited traditional fast track population, *The American Journal of Emergency Medicine* **36**(1) (2018), 124 – 127, DOI: 10.1016/j.ajem.2017.10.016.

[7] R. A. Hamid, A. S. Albahri, O. S. Albahri and A. A. Zaidan, Dempster–Shafer theory for classification and hybridised models of multi-criteria decision analysis for prioritisation: A telemedicine framework for patients with heart diseases, *Journal of Ambient Intelligence and Humanized Computing* **13** (2022), 4333 – 4367, DOI: 10.1007/s12652-021-03325-3.

[8] L. A. Hillas, R. Caldentey and V. Gupta, Heavy traffic analysis of multi-class bipartite queueing systems under FCFS, *Queueing Systems* **106** (2024), 239 – 284, DOI: 10.1007/s11134-024-09903-4.

[9] N. R. Hodgson and S. J. Traub, Patient assignment models in the emergency department, *Emergency Medicine Clinics of North America* **38**(3) (2020), 607 – 615, DOI: 10.1016/j.emc.2020.03.003.

[10] J. Kosiol, T. Silvester, H. Cooper, S. Alford and L. Fraser, Revolutionising health and social care: Innovative solutions for a brighter tomorrow – a systematic review of the literature, *BMC Health Services Research* **24** (2024), article number 809, DOI: 10.1186/s12913-024-11099-5.

[11] S. Mahmudov and M. Mahmudova, Modern intelligent health systems: Queueing, simulation, in: *2022 International Conference on Information Science and Communications Technologies* (ICISCT) (Tashkent, Uzbekistan, 2022), pp. 1 – 6 (2022), DOI: 10.1109/ICISCT55600.2022.10146960.

[12] C. Naithani, S. P. Sood and A. Agrahari, The Indian healthcare system turns to digital health: eSanjeevaniOPD as a national telemedicine service, *Journal of Information Technology Teaching Cases* **13**(1) (2023), 67 – 76, DOI: 10.1177/20438869211061575.

[13] D. Olivia, G. Attigeri and A. Saxena, Optimization model for mass casualty management system using qos-aware routing protocol and casualty triage prediction, *International Journal of Information Technology* **2024** (2024), DOI: 10.1007/s41870-024-02052-0.

[14] B. Saini, D. Singh and K. C. Sharma, A queueing theory approach to analyze the impact of COVID-19 pandemic on hospitals system capabilities: A lesson for future pandemic preparedness, *Journal of Management World* **2025**(1) (2025), 718 – 727, DOI: 10.53935/jomw.v2024i4.767.

[15] B. Saini, D. Singh and K. C. Sharma, Analysis of right to health using queueing theory: A critical overview, *International Research Journal on Advanced Engineering and Management* **2**(3) (2024), 332 – 338, DOI: 10.47392/IRJAEM.2024.0048.

[16] N. Zychlinski, Managing queues with reentrant customers in support of hybrid healthcare, *Stochastic Systems* **14**(2) (2024), 109 – 228, DOI: 10.1287/stsy.2022.0105.