Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Optimization of Medical Resources within a Markov Framework: Integration of Queuing Theory and Deep Learning

PENGCHENG MA[1], HAN WANG[1], YUJING HUANG[2], JIACHENG LI[3], HONGYANG ZHANG[4] and BORAN CUI[5]

[1]School of Mathematics and Computing Science, Guilin University of Electronic Technology; Center for Applied Mathematics of Guangxi (GUET); Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation, Guilin, 541004, China
[2]Xiangsihu College of Guangxi Minzu University, Nanning, 530000, China
[3]School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, No.10 XiTu Cheng Road HaiDian District, Beijing, 100876, China
[4]College of Biological and Agricultural Engineering, Jilin University No.5988 Renmin Street Nanguan District,Changchun,130022,China
[5]School of Basic Medicine Qiqihar Medical University No. 197, Bukui South Street Jianhua District,Qiqihar, 161005,China

Corresponding author: Yujing Huang (e-mail: scipcma@163.com).

This work was supported by Innovation Project of Guangxi Graduate Education(JGY2023126), and the Project for Enhancing Young and Middle-Aged Teachers' Essential Research Capabilities in Colleges of Guangxi (2022KY0189), and Guangxi Higher Education Reform Project(2024JGA182), and the Project for Enhancing Young and Middle-Aged Teachers' Essential Research Capabilities in Colleges of Guangxi: 'Research on Key Technologies and Applications of Digital Animation in the Context of the Educational Metaverse' (2023KY1665), and partially supported by the Science and Technology Project of Guangxi (Guike AD23023002).

ABSTRACT Today, on a global scale, the distribution of health human resources is facing a considerable challenge, especially in low - and middle - income countries. In these countries, people face numerous problems in accessing medical services, which only exacerbates the unfairness in healthcare. Currently, the demand for healthcare services is constantly increasing, while the available resources are extremely limited. Such a situation leads to patients waiting longer for care and a decline in the quality of care. To address these problems, an innovative model integrating Markov Decision Processes, queuing theory, and Deep Neural Networks (DNN) is proposed to optimize the allocation of medical resources for medical purposes.This theoretical model queued for the dynamics of patient mobility and models for resource analysis, while neural networks (NNS) in depth to learn from past and predictable historical data for making decisions about freedom of opinion, the model according to fluctuations in demand, dynamically, Making adjustments in the allocation of resources would improve the efficiency of the health system and allow a more equitable distribution of medical resources, while improving the adaptability of the health system. This study has contributed many criticisms, including integrating markov processes, resources to capture dynamic characteristics, the queue to improve the flow of patient wait time as well as resources, the use of neural network depth for large-scale data analysis, for decision making through training, through this method multiple are not integrated, With this model allows to produce optimal resource allocation solutions, which significantly reduces patient wait times, improves resource efficiency and ensures equitable access to health care. This research provides a comprehensive framework for addressing the multifaceted challenges of medical resource allocation in diverse and dynamic healthcare environments, offering practical solutions for policymakers and healthcare adminis- trators.

INDEX TERMS Healthcare resource assignment, Markov processes, Queuing network, Deep Neural Network, Health system efficiency.

I. INTRODUCTION

QUEUING theory can be applied as a tool to analyze the phenomena of the treatment system and determine the features of the waiting time, queue length, and capacity of the system and results in improving the efficiency and decreasing the patient rejection. Global medical Configuration of human resource management, has become today hamper the proper functioning of the health system, should, by 2030, the world would be about 900 million of the shortage of health professionals, this includes doctors and nurses, etc., this shortage,

low and middle income countries are particularly affected. As in Malawi, 1000 people only 0.19 health care specialists, but in countries like Norway, more than 1000 people of 10 for administrators and this profound, to health care, exacerbating inequalities in rural areas and in underserved areas there are few ways to get medical care in time in these places. [1]. Recent reports point to a continuing shortage of health workers and, although there have been some improvements in areas such as coverage of HIV services, progress in ensuring financial protection and full coverage of health services has been more limited in low - and middle-income countries , the report also highlights the urgent need for increased investment in education and employment. [2], [3].To address the global shortage of health workers, while supporting the achievement of the sustainable development goals. [4], [5] These findings illustrate the worsening crisis of unequal distribution of medical resources and underline the need for systemic solutions to close disparities and ensure equitable access to health care worldwide. In addition to global health professionals distribution of issues on the agenda huge differences, even in the case of India, and in rural areas, urban areas per 1000 people the number of health professionals is three times higher than in rural areas, so access to medical services in rural areas is considerable. This unbalanced situation, which only limits health care coverage, also puts pressure on existing medical staff, causing problems of burnout and high turnover. Addressing these injustices is essential to improving the quality of the health care system and ensuring equitable access to resources, Although there is increasing interest in this issue, existing theoretical frameworks and optimization methods do not address the complex situation of the allocation of medical resources in a dynamic multiple medical environment. Many current models, a single institution either give priority to the optimization of resources but does not sufficiently consider dynamic operational challenges or fluctuations, operational healthcare the first study conducted recently by experts in the field of research by the static model of the dynamics of patient flows. For example, research by Professor Guanghua Wan and colleagues has demonstrated the need for more adaptive resource allocation strategies in healthcare systems. To address these gaps, this study proposes innovative approaches to optimize medical re- source allocation, ensuring fairness, efficiency, and adaptability in healthcare systems. How can we optimize the allocation of medical resources to ensure equitable and effective health- care delivery? This study addresses this central question by exploring three critical aspects. From the perspective of resource distribution, how can medical resources, particularly doctors and nurses, be efficiently allocated to meet varying patient demands across departments and institutions? From the perspective of demand management, how can fluctuating needs between emergency and non-emergency cases be balanced, especially during peak and off-peak periods? Finally, from the perspective of practical implementation, how can optimization models be effectively applied in real-world scenarios while accounting for patient flow complexities, resource limitations, and hospital operations? To answer this question, this study proposes a novel model integrating Markov processes, queueing theory, and neural networks to optimize medical resource allocation. The Markov Decision Process (MDP) component captures the dynamic nature of resource states, while queueing theory improves patient wait times and resource flows. The historical data are completely distracted from connecting neural networks, from the dynamic adjustment model in real time of resource allocation, the main contribution of this study isto integrate these three methods, in order to address the medical resources in several challenges to be met, and which scenarios or static single to focus only on existing models, This approach provides an integrated framework that ADAPTS to a diverse and dynamic health care environment. The use of neural networks for real-time decision making represents an important step forward that can bridge the gap between theoretical models and practical applications. This integrated approach has some potential to improve operational efficiency, reduce inequalities and support more effective health services. Bed allocation is an important aspect of hospital management as it helps to optimize the use of resources, minimize wait times for patients, and ensure that patients receive timely and appropriate care. The proposed model incorporates historical patient data and real-time information to estimate the probabilities of transitioning between different states. By analyzing this probabilistic information, hospitals can make informed decisions regarding bed allocation, ensuring that patients are admitted to the most appropriate beds based on their medical needs and the availability of resources. Through simulation experiments and data analysis, the effectiveness of the Markov chain model is evaluated in terms of bed occupancy rates, patient wait times, and overall resource utilization. The results demonstrate that the proposed model offers significant improvements over traditional bed allocation methods, resulting in reduced patient wait times, optimized bed occupancy, and better overall management of hospital resources. A study published in the Journal of Healthcare Engineering looked at the impact of bed allocation on patient flow in a hospital. The study found that an optimized bed allocation system could improve patient flow and reduce wait times, leading to better patient outcomes.

II. LITERATURE REVIEW

The literature on health care systems has been reviewed to answer research questions in order to focus on the challenges faced by health care services and the means to cope with them, such as Markov Decision Process (MDP) and queue theory, which are part of the approach taken to optimize the allocation of health resources.

A. LITERATURE ON SEVERE BARRIERS IN HEALTHCARE RESOURCE ALLOCATION

Global health care systems to ensure equitable and effective access to medical resources in this regard faced with various challenges, these challenges are primarily of geographical, financial origin and barriers to the exploitation of these barri-

ers would make the disparity in the distribution of resources, health care and barriers to the provision of health services must be integrated into the agreement. [6]–[8].

Economic disparities are a very important barrier and, in times of crisis, in the case of hospitals, those facing economic constraints generally reorganize the distribution of resources, which can lead to a decrease in the intensity of care in key areas. [6]. This paper concludes that geographic distance and cultural differences may limit patients' access to health services. Although telemedicine has undergone some development, physical distance and cultural barriers such as differences in language and values are still significant barriers to patients' access to health services. [7], [8].

The inefficiency of these obstacles has become even more serious, for example, this overcrowding of emergency rooms is a problem that has existed for a long time, in general, be hospitalized due to the relatively long situation and the efficiency of resource allocation, the policy of early bed distribution has been clearly confirmed the flow of emergency rooms, however, The success of this policy requires seamless coordination between emergency departments and hospitals. [9], [10]. At different levels of care medical resources, in a number of obstacles, for example, more resources for primary health care, the large hospital had an excessive burden, this situation of large hospitals also increased in relation to the provision of primary health care services played a role. [11]–[14].

B. LITERATURE ON METHODOLOGIES FOR OPTIMIZING HEALTHCARE SYSTEMS

To overcome these difficulties, researchers increasingly tend to advanced methods such as Markov Decision Process(MDP) and the tail theory of these methods to solve the complex situation of dynamic characteristics of health care and provides a reliable framework, Markov dynamic decision-making processes.

MDP has been widely used in the scenario, In optimizing patient processes and allocating resources. [9], [12], [15]–[22]. To overcome these difficulties, researchers increasingly tend to advanced methods such as Markov Decision Process (MDP) and the tail theory of these methods to solve the complex situation of dynamic characteristics of health care and provides a reliable framework, Markov Decision Process (MDP) has been widely used in the scenario, In optimizing patient processes and allocating resources. [12], [19].Markov Decision Process (MDP) based probabilistic predictions have also proven effective in emergency situations, allowing dynamic resource allocation strategies to reduce patient wait times and improve problems where the system is not efficient [9], [15].

Queue theory has been widely used to improve operational efficiency by optimizing resource utilization and patient mobility. [15], [17]–[19]. For example, queue models specifically tailored to healthcare scenarios can solve complex problems related to different patient needs, improve the balance of resources, and reduce bottlenecks. [15], [19]. In emergency departments, the policy of gradation of early bed allocation has already shown strong potential to improve system responsiveness during peak periods of demand. [15], [18].

Deep neural networks (DNN) also offer great potential to optimize health systems. Deep neural networks have considerable capacity for data analysis and pattern recognition, which can be used to predict patient flows and disease progression, researchers like li jie. [10] Results show that deep neural networks have the ability to extract valuable information from a large number of medical history data, which can serve as a basis for the prior allocation of medical resources, which can improve resource efficiency, (cheni et al.). [9] Results show that deep neural networks have the ability to extract valuable information from a large number of medical history data, which can serve as a basis for the prior allocation of medical resources, which can improve resource efficiency, (cheni et al.).

Markov Decision Process (MDP) and tail theory of maximizing potential in the field of integrated health care has shown a development this hybrid model Markov to decision making and tail to combine the two theoretical approaches, dynamically, of resource allocation to solve the problem of uncertainty that exists in health care systems and the constraints of various operations. [16], [17], [19], [22]. For example, this field-proven model reduces hospital staff shortages by 15%,increases patient throughput by 25% reduces operating costs by 10% [19], [22].The results of this research clearly show that Markov Decision Process (MDP), queue theory and deep neural networks (NNS) have unique and complementary benefits in addressing the multiple aspects of the challenges facing health care systems. [12], [16], [19]–[22].

Whether it's MDP, queue theory and DNN alone or in combination, these approaches can be powerful tools to optimize health care operations and solve the complex problems of resource allocation in a dynamic environment.

III. PROBLEM DESCRIPTION

Patients in a hospital, seeking to get health care are usually structured according to tasks, this system seems however, when they are for the most part a serious problem of inefficiency and bottlenecks, care provided by general care or registration of the reception, in this place, provide their personal data of Patients, you will then receive instructions on the next steps. On the problem of sorting links. , medical staff for each case of the emergency, then patients to the appropriate area, after, patients wait outside the physician designated by the advisory office, they may seek care and may perform diagnostic tests, subject to recommendations or his hospital physician. All stages of this process, some difficulties arise for most causes delays, inefficiency, dissatisfied patients, registration, a long queue, it is often at peak times, it is because the number of patients was too, associated with the administrative capacity of staff is limited. These delays, which occur during the first phase, often set the tone for a situation of ineffectiveness for the patient throughout the treatment, Problem to a set of additional challenges at

a triage stage. Patients are difficult to predict, which could lead to a sharp increase in demand exceeds the triage treatment capacity problem. Who patients needing emergency care to assess delays could follow, so it is possible to make their medical situation. In addition, inaccurate or inconsistent triage decisions can lead to patients being misguided, which can hinder the hospital workflow and increase patient wait times in the downstream phase, When the patient arrives at the consultation stage, there are generally additional delays due to the limited time available to the physician to provide services and the time available to the consultation. The need for physicians to allocate time according to the complexity of each case creates uncertainty throughout the system. This inaccuracy in time often leads to queues outside the consultation room, which can be frustrating for waiting patients and can also create an inefficient use of resources, The diagnosis and treatment phases further complicate the problem. The limited number of diagnostic equipment and overcrowded testing facilities often create bottlenecks that delay the completion of tests and treatments. These delays can prolong patients' stay in hospital, as patient movements become irregular, and create congestion in other areas of the hospital.

These challenges actually lies ina fundamental problem: of the principle of subsidiarity, the needs of hospital patients limited resources, of the mismatch between the patient for hospitals, essentially by time, disease, and many factors, such as external emergency. Hospital resources, including staff, equipment and space, are either fixed or difficult to develop in a short time. This inconsistency creates queues at every stage of the workflow, resulting in wasted time, reduced operational efficiency and a deterioration in the quality of patient care.

IV. DNN AND QUEUING INTEGRATED METHOD

To address the complex dilemmas that exist in the distribution of medical resources, an approach that integrates queue networks and deep neural networks. This approach is used in combination with the queuing theory approach and the neural network approach, which improves the optimization capacity of the health system in dynamic terms, it also ensures a fair distribution of resources, greater efficiency and some adaptability. A kind of open network, queued for patient flows and modelling of the situation of resources while the depth of neural networks based on historical data including the presence of learning, real-time forecasting and optimizing the operation of the application of the theory of queues in this scenario analysis, clearly define, Decision making data collection and the circumstances associated with pretreatment are described in detail. The DNN model design and training process is also discussed in detail. In addition, the performance of the model is checked. This integrated system enables dynamic optimization of medical resource allocation, improving hospital operating conditions and patient experience.

A. INTEGRATING QUEUEING NETWORK AND DEEP NEURAL NETWORK

1) Combining Queueing Network and DNN for Dynamic Resource Assignment

When it comes to healthcare systems, and the depth of neural networks of the network queued (DNN) integration of dynamic optimization resource allocation, in fact a particularly effective method, the queue of the network offers a quantifiable, this approach can go patient mobility to hospital, also evaluate the efficiency of resource utilization. While the depth of neural networks, it is very talented at learning historical data in the existing model and gives a result of forecasting or providing assistance, if using combining these two methods queued us to overcome the limitations of the network network queued a set of computation that however it produce data very well. We can also deeply compute powerful neural networks, of course, the depth of neural networks to provide real-time forecasts and recommendations that need to have data as support, our method using models queued for the arrival time of patients and the contingencies of service duration, and then calculate the wait time, and the use of resources. While the neural networks in depth data resulting from the use of models queued for optimization of resource allocation, the opening, queue of the cohort problem model triage. Nurses and general practitioner, quantification of cohort flows the depth of neural networks, processing these data quantified to optimize resource allocation.

The combination of the two methods makes it possible to capture all the complexity of the real words and find an optimal solution.

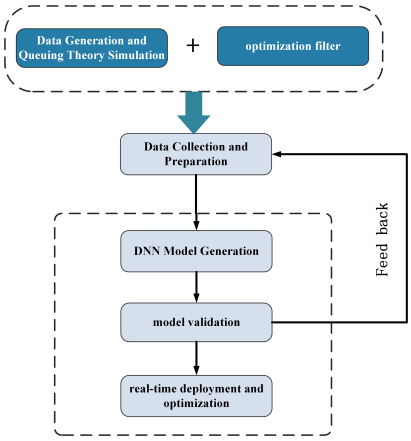


FIGURE 1. Overall flowchart

2) Model Architecture and Workflow

The flows from models queued, this model the queue presents the flows of patients in the hospital and the use of resources, the use of resources and an evaluation of people suffering from this model can be captured the arrival time and random of the length of service, one can also the waiting time and use of these resources. Following the entry of these deep neural network indicators into historical data for training and research, its optimal distribution of emission projections, the

depth of the capture of neural networks adopted the three main ones: patient arrival rates ((λ), N_n) within the limits of the quantities of nurses and doctors (N_d). Based on these data the depth of neural networks to predict system performance indicators such as wait times and salary costs, the method applied is the length of service to calculate the time wait time, as well as the use of resources and then this model dynamically number of nurses and doctors to make adjustments for the quality of services and efficiency of resource use.

This integrated approach offers several advantages:

- **Simplified Inputs**: The model requires only three key inputs (λ, N_n, and N_d), reducing the complexity of data collection and processing.
- **Improved Interpretability**: It is possible to isolate and analyze the effects of changes in staffing levels in order to obtain a clear view of how exactly the staffing changes have impacted system performance.

B. QUEUEING THEORY IN HEALTHCARE SYSTEMS

1) Basics of Queueing Theory in Healthcare

Queue the theory of plastic joints to resource sharing, random health provides a method of quantifying the distribution of patient flows in a hospital of resource utilization and resource utilization have a critical role to play, it relies on key performance indicators, wait times and quantify the rate of resource utilization, to help them optimize resource allocation.

The health care system is very uncertain:

- Patient arrivals following a Poisson process (λ_n for nurses, λ_d for doctors)
- Service times obeying exponential distributions (μ_n for nurses, μ_d for doctors)
- Need to balance service quality with constrained resources

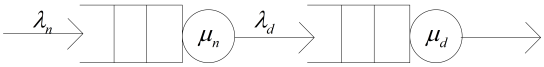


FIGURE 2. Queueing Process Flow Diagram

C. NOTATION

Notations	Description
λ_n, λ_d	The arrival rate at nurse station, and doctor's department
μ_n, μ_d	The service rate, for nurse and doctor's department
s_n, s_d	The number of nurses, and doctors
ρ_n, ρ_d	The utilization rate of nurse station, and doctor's department
$P_{0,n}, P_{0,d}$	The probability of an empty nurse station , and doctor's department q
W_n, W_d	The average waiting time at nurse station , and doctor's department
T_n, T_d, T_{total}	The total time at nurse station , doctor's department and system

FIGURE 3. Symbol callouts

1) Queueing Models in Healthcare

Queueing theory is used to describe patient flows and resource utilization within the hospital by quantifying key performance metrics. The service time is used to calculate waiting times, throughput times, and resource utilization. Two primary M/M/s queueing models are employed:

1. **Nurse Triage Queue**:
   - Arrival process: Poisson(λ_n)
   - Service process: Exponential(μ_n) with s_n servers
2. **Doctor Consultation Queue**:
   - Arrival process: Poisson(λ_d)
   - Service process: Exponential(μ_d) with s_d servers

These queues are connected in a tandem network to avoid oversimplification. System stability requires:

$$\lambda_n < s_n \mu_n \quad \text{(Nurse Station)}$$

$$\lambda_d < s_d \mu_d \quad \text{(Doctor Consultation)}$$

2) Key Performance Metrics

Nurse Station Analysis

$$W_{q,n} = \frac{P_{0,n} \left(\frac{\lambda_n}{\mu_n}\right)^{s_n} \mu_n}{s_n! \left(\mu_n s_n - \lambda_n\right) (1-\rho_n)^2} \tag{1}$$

where the steady-state probability is:

$$P_{0,n} = \left[\sum_{k=0}^{s_n-1} \frac{\left(\frac{\lambda_n}{\mu_n}\right)^k}{k!} + \frac{\left(\frac{\lambda_n}{\mu_n}\right)^{s_n}}{s_n!(1-\rho_n)} \right]^{-1} \tag{2}$$

Doctor Consultation Analysis

$$W_{q,d} = \frac{P_{0,d} \left(\frac{\lambda_d}{\mu_d}\right)^{s_d} \mu_d}{s_d! \left(\mu_d s_d - \lambda_d\right) (1-\rho_d)^2} \tag{3}$$

with analogous derivation for $P_{0,d}$

System Throughput

$$T_{\text{total}} = W_{q,n} + \frac{1}{\mu_n} + W_{q,d} + \frac{1}{\mu_d} \tag{4}$$

3) Steady-State Conditions

For reliable system operation:

1. Nurse station utilization: $\rho_n = \frac{\lambda_n}{s_n \mu_n} < 1$
2. Doctor consultation utilization: $\rho_d = \frac{\lambda_d}{s_d \mu_d} < 1$

Under integrated system analysis:

$$\lambda_n = \lambda_d \quad \Rightarrow \quad \rho_{\text{total}} = \frac{\lambda_d}{s_d \mu_d} < 1$$

4) Managerial Implications

This modeling approach enables:

1. Predictive analysis of service efficiency under varying resource configurations through the use of queuing models to quantify key metrics.
2. Dynamic staffing optimization by adjusting staffing levels based on the model's calculations to balance service quality and resource efficiency.
3. Reduction of unnecessary waiting times by calculating waiting and throughput times to ensure timely patient care without compromising care quality.

4. Operational decision support for hospital administrators through the use of model outputs to inform strategic staffing and operational decisions.

The proposed framework not only improves resource utilization but also enhances patient experience through evidence-based operational management, forming the theoretical foundation of our healthcare optimization research.

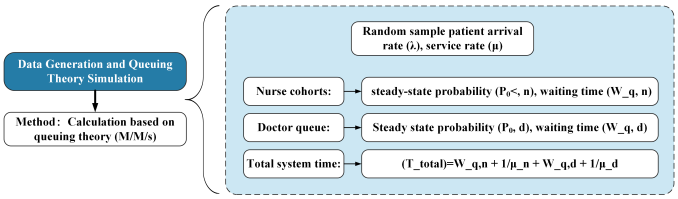


FIGURE 4. Queueing Theory

D. OPTIONS AND DECISION CRITERIA

Resource allocation must achieve synergistic optimization between nurse and doctor queues. This study proposes the following decision criteria, validated by real-world cases:

1) Core Optimization Objectives

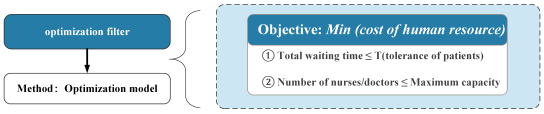


FIGURE 5. Options and Decision Criteria

2) Real-World Case: Johns Hopkins Hospital Outpatient Optimization

1. Problem Context (Annals of Emergency Medicine, 2019):

- Nurse queue peak wait time: 25 minutes, with a target of less than or equal to 15 minutes.
- Doctor queue average wait time: 40 minutes, with a target of less than or equal to 30 minutes.
- Fixed staff ratios, with $N_n = 4$ and $N_d = 6$, caused inefficiency: nurse utilization was 60%, while doctor utilization exceeded 95%.

2. Optimization Strategy:

- Input Features: Patient arrival rate (λ), N_n, and N_d.
- DNN Prediction: The model was trained on historical data, mapping λ, N_n, and N_d to wait times and utilization.
- Dynamic Adjustment: The optimal N_n and N_d were determined based on real-time λ .

3. Results:

- Nurse queue wait time decreased from 25 to 12 minutes.
- Doctor queue wait time decreased from 40 to 28 minutes.
- Utilization improved: nurses to 78%, doctors to 88%.
- Labor costs decreased by 9%.

E. DATA COLLECTION AND PREPARATION

Applications the depth of neural networks (DNN), the first step make data collection and data collection for accurate models, representative data on the formation of our model of data used comes from the queues, this containing the coverage rates of arrival rates in Germany, and the number of workers, etc. As these data are not time series, but are independent of each other, we opted for a fully connected feedforward neural network. In a context such as the allocation of medical resources, factors such as the flow of patients to hospitals, the use of resources and the assessment of resource use are complex and interdependent. With the collection of these data, sufficient information can be provided to the models to accurately predict the number of medical personnel required to prepare additional data for the effective operation of our method is essential, as only quality data must support the work of studying training models and optimizing depth. During the data processing, we applied several techniques to enhance data quality and model performance. Using queueing formulas and optimization models, we calculated the performance data (including utilization, waiting time, and cost), and applied filtering techniques on the data, resulting in the filtered dataset. In this process, we also employed the One-Hot Encoding technique, transforming categorical data into numerical values so that machine learning algorithms could process it more effectively. To eliminate the impact of different scales, we performed data normalization, scaling the data to a range of 0 to 1, ensuring the stability and accuracy of the model during training.

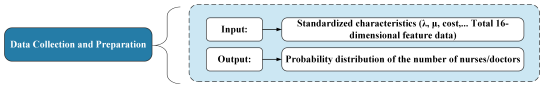


FIGURE 6. Data Collection and Preparation

1) Normalization Formula

Data normalization is applied to ensure the stability of the model during processing, preventing certain features from having too much influence on the training process. The formula for normalization is:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Where: - $X_{\min}$ and $X_{\max}$ are the minimum and maximum values of the feature data. - X_{norm} is the normalized data, ensuring that the data is scaled to the range [0, 1].

F. DNN MODEL GENERATION

In the Model Generation Process In the model generation process, we used a Fully Connected Feedforward Neural Network. The DNN architecture consists of three hidden layers, with 128 neurons in the first layer, 64 neurons in the second layer, and 32 neurons in the third layer. The input layer receives sixteen key parameters, including patient

TABLE 1. Core Optimization Objectives

Objective	Nurse Queue	Doctor Queue
Wait Time	Triage wait time less than or equal to 60 minutes	Consultation wait time less than or equal to 120 minutes
Utilization	Nurse staff count less than or equal to N_n	Doctor staff count less than or equal to N_d
Cost Limits	Nurse labor cost less than or equal to 200 per hour	Doctor labor cost less than or equal to 800 per hour

arrival rate , number of nurses , number of doctors , and other relevant metrics related to patient flow and resource utilization. The output layer predicts system performance metrics, such as wait times and resource utilization rates. This model mentioned in the text, all layers hidden activation function main goal, while the output layer linear function has been activated to do to ensure the accuracy of the predictions can be relatively high, chosen the scanning network because the data we have, and these data are independent between this situation gives the network a unidirectional character. This means that the data circulates in a certain order in the network, without any temporal dependence, which is not the same as for circulating neural networks (RNNs).Additionally, since we cannot preinterpret the initial data, we chose not to use Convolutional Layers. Instead, we opted for fully connected layers, which allow the network to better present the complex relationships between different inputs. For the network architecture, we set up multiple hidden layers and an appropriate number of neurons to ensure that our approach can learn sufficient features and make effective predictions.

During training, we used Cross-Entropy Loss as the loss function to evaluate the performance of our approach. The Cross-Entropy Loss measures the difference between the predicted classification and the true labels, and is defined as:

$$L_{\text{train}} = -\sum_{i=1}^{C} y_{\text{true},i} \log(p_{\text{pred},i})$$

Where $y_{\text{true},i}$ is the true label for sample i, and $p_{\text{pred},i}$ is the predicted class probability for sample i.

To optimize the training process, we selected the Adaptive Moment Estimation (Adam) optimization algorithm, which updates the model's parameters by combining first- and second-order moment estimates, and improves training stability and speed.

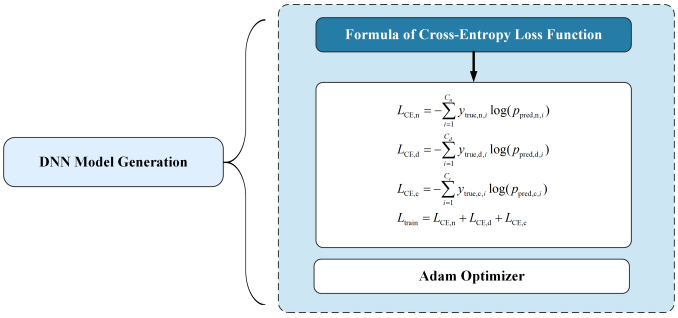


FIGURE 7. DNN Model Generation

1) Cross-Entropy Loss Formula

During training, we use Cross-Entropy Loss to handle the classification tasks. The formula for Cross-Entropy Loss is defined as follows:

$$L_{\text{CE,n}} = -\sum_{i=1}^{C_n} y_{\text{true,n},i} \log(p_{\text{pred,n},i})$$

$$L_{\text{CE,d}} = -\sum_{i=1}^{C_d} y_{\text{true,d},i} \log(p_{\text{pred,d},i})$$

$$L_{\text{CE,c}} = -\sum_{i=1}^{C_c} y_{\text{true,c},i} \log(p_{\text{pred,c},i})$$

Where: - C_n, C_d, and C_c represent the number of classes for nurses, doctors, and cost, respectively. - $y_{\text{true,n},i}$, $y_{\text{true,d},i}$, and $y_{\text{true,c},i}$ are the true labels for sample i. - $p_{\text{pred,n},i}$, $p_{\text{pred,d},i}$, and $p_{\text{pred,c},i}$ are the predicted probabilities for the nurse, doctor, and cost tasks, respectively.

The overall training loss can be represented as:

$$L_{\text{train}} = L_{\text{CE,n}} + L_{\text{CE,d}} + L_{\text{CE,c}}$$

2) Adam Optimization Algorithm

The Adam optimizer is used to efficiently update the model's parameters. The update rules for Adam are as follows:

$$\hat{m}_t = \beta_1 \hat{m}_{t-1} + (1-\beta_1) g_t$$

$$\hat{v}_t = \beta_2 \hat{v}_{t-1} + (1-\beta_2) g_t^2$$

$$\hat{m}'_t = \frac{\hat{m}_t}{1-\beta_1^t}$$

$$\hat{v}'_t = \frac{\hat{v}_t}{1-\beta_2^t}$$

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{\hat{v}'_t} + \epsilon} \hat{m}'_t$$

Where:

- g_t is the current gradient,
- $\hat{m}_t$ and $\hat{v}_t$ are the first and second moment estimates,
- β_1 and β_2 are the decay rates,
- η is the learning rate, and ϵ is a small constant to prevent division by zero.

G. DNN MODEL VALIDATION

Our test dataset is derived from a larger dataset by removing the training dataset. Specifically, the training dataset includes various features that are used to train our neural network to identify and predict the optimal allocation of healthcare resources. The features in the training dataset include patient arrival rates, the number of doctors and nurses, waiting times, and salary costs. These characteristics will be used to form the models so that they can more effectively predict results from unknown data.

According to the first formula, the relationship between the two test data sets and the training data sets is as follows:

$$\alpha_V = \alpha_S \backslash \alpha_T$$

Of these, α_V represents the test set, α_S he complete dataset and α_T the training set. This configuration ensures that the test data is what remains after the training data is removed.

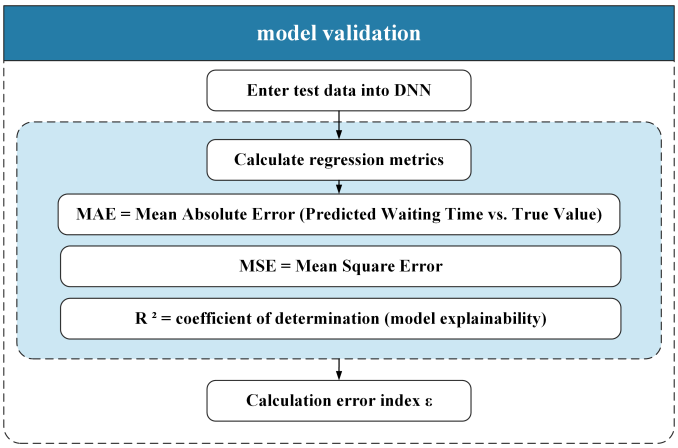


FIGURE 8. DNN Model Validation

In order to validate the proposed model, the results actually produced by the model and the expected results are compared. This comparison exercise is carried out using the error calculation method, the specific formulas for which are presented below:

$$\varepsilon = \frac{1}{|\alpha_V|} \sum_{V \in \alpha_V} \left| \frac{T_{\max} - T_V}{T_{\max}} \right|$$

Where:

- $T_{\max} = \max_{V \in \alpha_V} T_V$: The maximum true value in the set α_V.
- T_V: The true value of the current sample.

This formula clearly indicates the need to calculate the difference between the actual value and the expected value and normalize it by dividing this difference by the maximum actual value, which helps us measure the model's prediction error in the dataset performance test. Although we calculated the errors using the formulas mentioned above, we found that the robustness of the model was not as strong and there was some bias. We rely on increasing the number of data to improve the model's generalization capacity and forecast accuracy. This will improve the robustness of the model and reduce errors due to insufficient data. To enable the evaluation of our model, we used mean square error (MMR) as a performance indicator. Mean square error (MMR) is a key measure of the predictive ability of a regression model. It first calculates the square differences between the expected and actual values and then averages those square differences. His formula is as follows:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_{\text{pred},i} - y_{\text{true},i})^2$$

MSE effectively penalizes larger errors, especially in predicting more complex or harder- to-predict samples. This paper is based on MSE minimization, which allows the model to better adjust training data and improve its ability to predict new samples. Once the model is optimized and errors calculated, the output data is used as an optimal allocation of medical resources. The data from these outputs are derived from the results obtained through continuous optimization and training of deep neural network models. Based on more training data and validated data, this article wants to give a precise pattern for resource allocation in the health field.

V. NUMERICAL EXPERIMENT

As part of this study, we wanted to explore the impact of different optimization algorithms on the performance of our models. A series of numerical experiments were conducted to test the effectiveness of the Adam algorithm and compare it with other optimization methods such as AdamW, RMSprop and SGD.

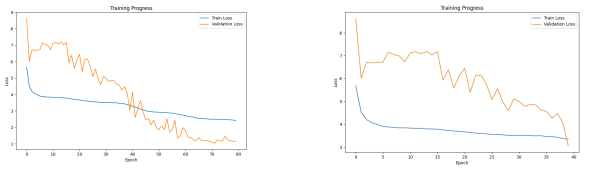
The experiment evaluated the performance of the model under different training conditions as well as different performance indicators. The diagram below gives a detailed comparison of the performance of the different optimizers. The two previous graphs show the progression of the model drive and the variation of losses with the optimizers Adam, AdamW, RMSprop and SGD. By comparing these results, we gain a better understand- ing of how each optimizer influences the convergence behavior and performance metrics.

In addition, the subsequent figures focus on the optimization effects in terms of perfor- mance improvement. The results after optimization are contrasted with the performance before optimization, showing the substantial improvements in the model's efficiency and accuracy.

Furthermore, we visualize the performance in terms of the total combined absolute error for nurse and doctor combinations after applying different optimizers. The final figure highlights the comparison of optimization effects on salary and total wait time, providing a comprehensive overview of the optimization's impact.

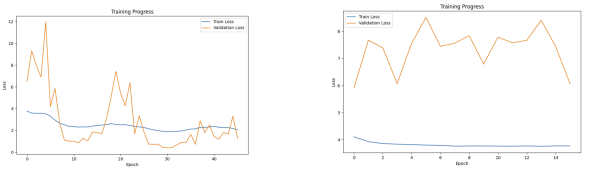
VI. CONCLUSION

This study presents a novel approach to optimizing medical resource allocation by integrating Markov processes, queueing theory, and deep neural networks (DNN). The proposed



(a) Loss Variation with Training Epochs (Adam)

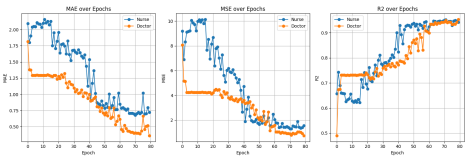
(b) Performance on Training Dataset (AdamW)



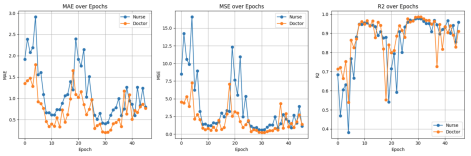
(c) Performance on Test Dataset (RMSprop)

(d) Absolute Error of Nurse and Doctor Combinations (SGD)

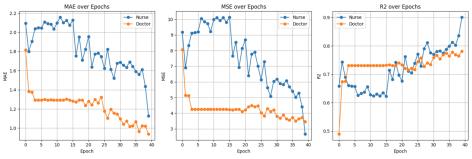
FIGURE 9. Comparison of Training and Test Performance with Different Optimizers



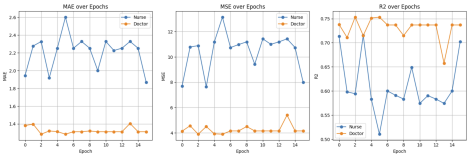
(a) Loss Variation with Training Epochs (Adam)



(b) Performance on Training Dataset (AdamW)



(c) Performance on Test Dataset (RMSprop)



(d) Absolute Error of Nurse and Doctor Combinations (SGD)

FIGURE 10. Training Progress and Error Analysis with Different Optimizers

framework addresses the multifaceted challenges of healthcare systems by capturing the dynamic nature of resource states and enabling data-driven adjustments based on historical patterns. Ineligibility raised, the theory of the mathematical model tail provides a solid framework for analyzing patient flows and resource allocation can be optimized integrating neural networks in depth, forecasting to obtain

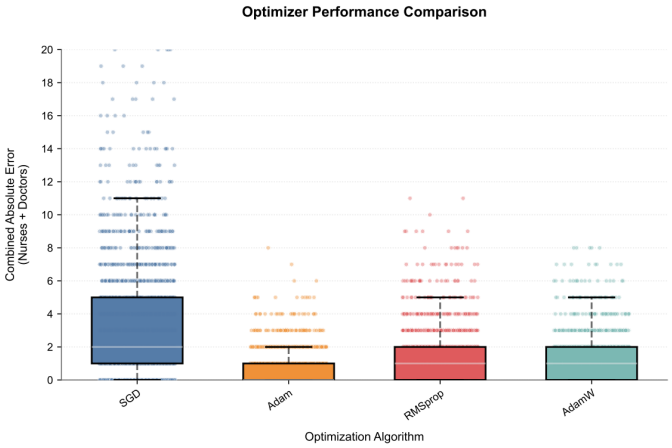


FIGURE 11. Comparison of Optimizer Performance on Combined Absolute Error

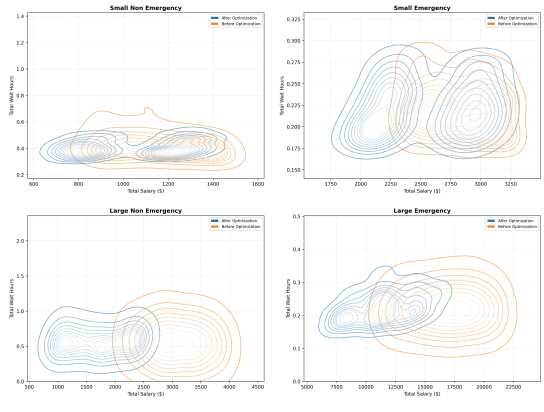


FIGURE 12. Comparison of Optimization Effects on Salary and Wait Time

freedom of opinion, making decisions highly adaptable, this system allows to respond to fluctuations in demand. This combination of different approaches represents a major advance in the field of health care operational research. It has proven to be effective in reducing patient wait times, improving resource efficiency and reducing labor costs. Results show that this model clearly improves the efficiency and equity of health services, providing practical solutions for decision makers and health care managers. However, this model exists some limitations, for now essentially focus on the implementation of the tertiary health care model, this situation risks its application to other health care institutions, for example community health care centers, Reliance on historical data from this model this reliance may not be able to fully explain unexpected events or rapidly changing health care needs. Taking into consideration the geographical factors of space-time, such as the effects of climate change on the spread of disease, this isan area of research to be explored, saying that, although this learning model wants to do to increase the protection of personal data related to the abuse of maize and the sharing of data remains the greatest technological challenge among future work, The focus will be on the applicability of the