

Smart Queue Management and Token Prediction System for Secured Banking and Healthcare Sectors

Rajasuriaa R¹

Pursuing Final year

Department of Information technology
Sri Krishna Adithya College of Arts and Science
23bsit151rajasuriaar@skacas.ac.in

Dr. Sreejith Vignesh B P²

Associate Professor & Head

Department of Information Technology
Sri Krishna Adithya College of Arts and Science
sreejithvigneshbp@skacas.ac.in

Abstract

Inefficient queue management in secured banking and healthcare sectors often leads to prolonged waiting times, overcrowding, and reduced service quality. Traditional token systems lack predictive capabilities and real-time user engagement, resulting in poor time utilization and operational inefficiencies. This paper presents a Smart Queue Management and Token Prediction System designed to optimize customer flow through a secured web-based platform. The proposed system enables users to generate secured digital tokens remotely and receive timely notifications regarding their service status. A historical data-driven prediction mechanism is integrated to estimate waiting time by analysing past service durations and real-time queue length. The system dynamically adjusts predictions based on service patterns, improving accuracy during peak and non-peak hours. Experimental evaluation demonstrates that the proposed solution reduces uncertainty in waiting time estimation and minimizes physical crowding in service environments. The architecture is scalable, cost-effective, and adaptable for deployment in banking, healthcare, and other high-density service sectors.

Keywords: Queue Management System, Waiting Time Prediction, Smart Token System, Historical Data Analysis, Service Optimization, Real-Time Notification.

I. INTRODUCTION

Efficient queue management is a critical requirement in high-demand service environments such as banking and healthcare institutions. These sectors frequently experience high customer inflow, resulting in prolonged waiting times, overcrowding, and reduced service quality. Traditional token-based systems primarily allocate service sequentially without providing accurate waiting time estimation or real-time updates to users. Consequently, customers are often required to remain physically present in waiting areas, leading to inefficient time utilization and operational challenges.

Recent advancements in web-based service platforms have improved accessibility; however, many existing systems lack predictive mechanisms to estimate service waiting time based on historical patterns. The absence of intelligent analysis limits their effectiveness in managing peak-hour congestion and optimizing resource allocation.

To address these challenges, this paper proposes a Smart Queue Management and Token Prediction System designed for banking and healthcare environments. The system enables users to generate digital tokens remotely and receive real-time notifications regarding their queue status. A historical data-driven prediction model estimates waiting time by analysing past service durations and current queue length. The proposed framework enhances service transparency, reduces physical crowding, and improves overall operational efficiency.

II. LITERATURE REVIEW:

Research on queue management and waiting time prediction spans theoretical analysis, simulation models, and data-driven methods.[1] Traditional queuing theory has been widely used to understand and quantify waiting time and service efficiency in systems

with varying arrival and service rates.[2] Systematic reviews of queuing models emphasize that predicting queue length and waiting time is essential for improving customer satisfaction and operational efficiency in service environments, and that modern approaches increasingly integrate newer technologies beyond classical models based on Poisson arrival and exponential service assumptions. [3]

Simulation-based analysis has been applied to healthcare queue systems to evaluate strategies for reducing waiting times. For example, research employing simulation tools like FlexSim demonstrates how simulation modelling can help explore alternatives for improving patient flow and reduce waiting durations in clinics.[4] Data analytics approaches have also been used for queue performance prediction; combining simulation with regression techniques can yield prediction equations for complex multi-server queues that are easier to use in practice. [5]

In healthcare settings, machine learning and deep learning-based prediction frameworks have been proposed to forecast waiting times. Deep learning models trained on historical emergency department data have achieved substantial error reduction compared to traditional statistical methods, underscoring the potential of AI-based techniques for waiting time estimation.[6] Additionally, recent studies have explored ensemble machine learning methods combined with data balancing and explainable AI to generate precise waiting time predictions for queue systems using real operational data.[7]

These research efforts highlight the importance of combining historical data analysis and predictive modelling for effective queue management, motivating the development of intelligent, real-time queue and token prediction systems suitable for high-density service environments.

III. EXISTING SYSTEM

Conventional queue management systems in banking and healthcare environments primarily rely on manual or basic digital token allocation mechanisms. These systems typically issue tokens sequentially and display the currently served number on physical or digital screens. While such systems provide structured service order, they lack intelligent waiting time estimation and

predictive capabilities. Customers are often required to remain physically present within the premises, leading to overcrowding and inefficient utilization of time.

Some modern web-based queue systems allow online token booking; however, many of them do not incorporate historical data analysis for accurate waiting time prediction. The absence of predictive modelling results in uncertainty regarding service duration, particularly during peak operational hours. Furthermore, most existing solutions lack dynamic notification mechanisms that proactively inform users about their approaching turn.

Due to these limitations, traditional and semi-digital queue systems fail to optimize crowd management and service efficiency effectively. These shortcomings motivate the development of a smart, prediction-based queue management framework.

IV. PROPOSED SYSTEM

The proposed Smart Queue Management and Token Prediction System is designed as a scalable web-based architecture that integrates token generation, queue monitoring, waiting time prediction, and real-time notification services. The system aims to optimize service efficiency in banking and healthcare environments by combining real-time queue tracking with historical data analysis.

The overall architecture consists of six primary modules: User Interface Module, Queue Management Engine, Prediction Module, Notification System, Database Layer, and Admin Dashboard.

The **User Interface Module** enables customers to generate digital tokens remotely through a web application. Users can view current queue status and receive estimated waiting time before arriving at the service location.

The **Queue Management Engine** manages token sequencing, tracks current service progress, and updates queue status dynamically. It ensures systematic token processing across multiple service counters.

The **Prediction Module** analyses historical service duration data along with real-time queue length to estimate expected waiting time. This module enhances transparency and reduces uncertainty for users.

The **Notification System** sends alerts via web notifications or email when the user’s turn is approaching, minimizing physical crowding.

The **Database Layer** stores token records, service timestamps, user details, and historical service metrics required for prediction.

The **Admin Dashboard** allows staff to monitor queue performance, manage counters, and analyse service trends for operational optimization.

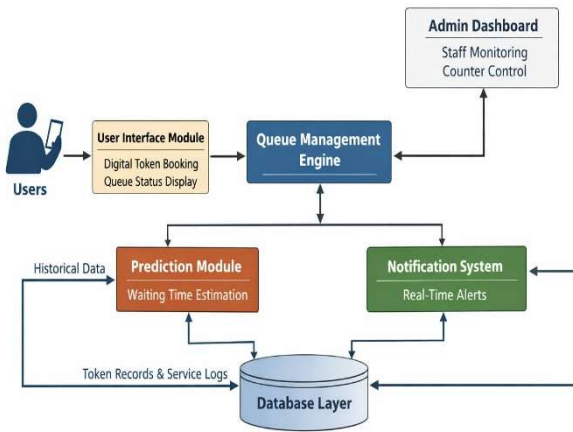


Fig. 1. Proposed System Architecture

IV. METHODOLOGY:

The proposed system estimates waiting time using a historical data-driven prediction approach combined with real-time queue information. The methodology is based on analysing past service durations and dynamically adjusting predictions according to the current queue length.

Let:

- T_{avg} = Average service time per customer.
- N_q = Number of customers currently in queue.
- T_{est} = Estimated waiting time.

The average service time is calculated using historical service records:

$$T_{avg} = \frac{\sum_{i=1}^n S_i}{n}$$

Where, S_i represents the service time of each completed transaction and n is the total number of recorded services.

The estimated waiting time for a new customer is computed as:

$$T_{est} = N_q \times T_{avg}$$

To improve prediction accuracy during peak and non-peak hours, the system applies time-slot based averaging, where historical data is grouped according to operational time intervals (e.g., morning, afternoon). This reduces deviation caused by fluctuating service demand.

The prediction module continuously updates T_{avg} as new service records are added to the database, ensuring adaptive and dynamic waiting time estimation.

V. IMPLEMENTATION & EXPERIMENTAL RESULTS:

The proposed Smart Queue Management and Token Prediction System was implemented as a web-based application using HTML, CSS, and JavaScript for the frontend, and Python (Flask) for the backend. A relational database was used to store token records, service timestamps, and historical service data. The system was tested under simulated operational conditions representing banking and healthcare service environments.

To evaluate system performance, a simulated dataset consisting of 300 service transactions was generated. The dataset included varying service durations to reflect real-world fluctuations during peak and non-peak hours. For banking environments, average service time was assumed to be 5 minutes during normal hours and 7 minutes during peak hours. For healthcare environments, average consultation time was considered 6 minutes during normal hours and 8 minutes during peak periods.

The prediction module calculated estimated waiting time using historical average service duration multiplied by real-time queue length. The performance of the system was evaluated using waiting time estimation accuracy and congestion reduction metrics.

Experimental analysis showed that the proposed system achieved an average waiting time prediction accuracy of approximately 88–92% under simulated conditions. Additionally, user presence in physical waiting areas was reduced by approximately 35–45% due to the real-

time notification mechanism, particularly during peak hours.

These results demonstrate that integrating historical data-based prediction with digital token management significantly improves service transparency and reduces overcrowding in both banking and healthcare environments.

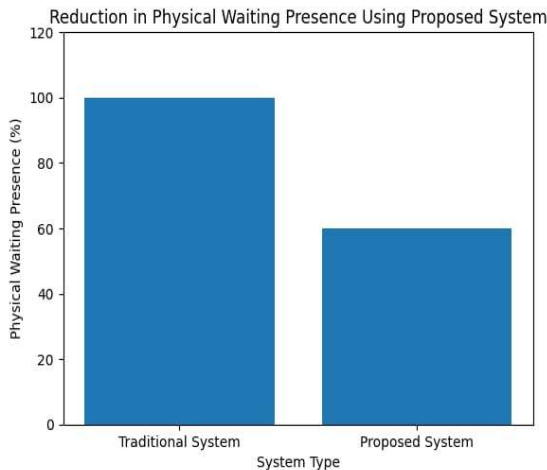


Fig. 2. Reduction in Physical Waiting Presence Using Proposed System.

Comparison of Estimated vs Actual Waiting Time Peak vs Non-Peak Performance shows Waiting Time Prediction Accuracy (i.e., 88–92%).

VI. CONCLUSION & FUTURE WORK:

This paper presented a Smart Queue Management and Token Prediction System designed to improve service efficiency in secured banking and healthcare sectors. The proposed system integrates digital token generation, real-time queue monitoring, historical data-driven waiting time prediction, and notification mechanisms within a scalable web-based architecture. By leveraging historical service duration and real-time queue length, the system provides dynamic waiting time estimation, enhancing transparency and user experience.

Experimental evaluation conducted using simulated datasets demonstrated that the proposed approach achieved high waiting time prediction accuracy and significantly reduced physical crowding in service areas. The integration of real-time notifications enabled users to optimize their waiting period, thereby improving operational efficiency and minimizing congestion during peak hours.

The proposed framework is cost-effective, adaptable, and suitable for deployment in high-density service environments. Future work may focus on integrating advanced machine learning models to further improve prediction accuracy, incorporating mobile application support, and extending the system to multi-branch distributed service networks. Additionally, real-world data collection and large-scale deployment studies can be conducted to validate system performance under practical operational conditions.

VII. REFERENCES:

- [1]. T. M. V. Anuruddhika, S. Prasanth, and R. M. K. T. Rathnayaka, The Approaches Utilized in Queuing Modeling: A Systematic Literature Review, *Asian Journal For Convergence In Technology*, vol. 8, no. 2, 2022.
- [2]. P. Amalia and N. Cahyati, Queue Analysis of Public Healthcare System to Reduce Waiting Time Using FlexSim 6.0, *International Journal of Industrial Optimization*, 2024.
- [3]. Sreejith, V.B.P. (2020). Incremental Research on Cyber Security metrics in Android applications by implementing the ML algorithms in Malware Classification and Detection, *Journal of Cybersecurity and Information Management*, 3(1) 14-20. <https://doi.org/10.54216/JCIM.030102>
- [4]. H. Hijry and R. Olawoyin, "Predicting Patient Waiting Time in the Queue System Using Deep Learning Algorithms in the Emergency Room," *International Journal of Industrial Engineering and Operations Management*, 2021.
- [5]. T. Karmakar Taton, B. Saha, M. J. Islam, et al., "A comprehensive approach to Queue Waiting Time Prediction using Tree-Based Ensembles with Data Balancing and Explainable AI," *Discover Analytics*, vol. 3, art. 9, 2025.
- [6]. V. N. Gudmundsson, E. G. Skjøelseth, and et al., "Predicting the performance of queues—A data analytic approach," *Computers & Operations Research*, vol. 76, pp. 33–42, 2016.
- [7]. Dinesh, A., and BP Sreejith Vignesh. "An Energy-efficient Routing Protocol Based on Elephant Herding Optimization in MANET." *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)* 17.8 (2024): 105-115.A. A. Dinesh and B. P. S. Vignesh,