Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 47 No. 02
February 2026

IntelliQueue: An AI-Driven Adaptive Queue Orchestration Framework Using Machine Learning and Cloud–Edge Synergy

[1]Darshana K, [2]Mr. Vimit Varghese

[1]Department of M.Tech Computer science and engineering,

Sri Krishna College of Engineering and Technology, Kuniamuthur, Coimbatore, Tamil Nadu - 641008

dkdarshana03@gmail.com

[2]Assistant professor,

Department of Computer science and engineering,

Sri Krishna College of Engineering and Technology, Kuniamuthur, Coimbatore, Tamil Nadu - 641008

vimitvarghese@skcet.ac.in

Abstract - Public service oriented areas such as hospitals, banks, schools and government offices are often characterized by some inefficiency in operation since long waiting time for the people, inadequacy of space because of overcrowding use to cum lack of transparency during queue processing. Traditional queue management systems are based on static rules or manual interventions, which cannot adapt to variable traffic pattern and peak-hour crowdiness. In response to these challenges, in this paper, we develop IntelliQueue, an AI-powered adaptive queue orchestration framework which leverages ML-based prediction with unified queue management and run-time user interaction.

The system provides online token booking and offline walk-in registration allowing universal access by all users. A timestamp queue integration scheme combines tokens generated at different access points into a single common queue, so that fairness and first-come-first-served principle is achieved. Data of the queue, such as arrival pattern, state of the queue (number waiting), rate at which the customers are served and temporal aspect are used and analyzed. In this context, we apply supervised machine learning regression models of waiting time and queue length for real-time prediction, thereby achieving proactive congestion monitoring and service flow enhancement.

IntelliQueue utilizes a cloud-based architecture for multiuser scalability, energy-efficient decision-making support mechanisms, as well as automated notifications regarding queue status and estimated wait time to avoid unneeded physical waiting and overcrowding. Experimental results with synthetic data show that the proposed framework not only is a better predictor than traditional rule based queue management methods but also has significantly reduced waiting time and system utilization. The findings suggest that data-driven queue orchestrators have potential to improve transparency, operational effectiveness and user satisfaction in contemporary public service settings.

Keywords **-** Intelligent Queue Management, Waiting Time Prediction, Machine Learning, Public Service System, Predictive Analytics, Cloud Architecture.

I. INTRODUCTION

Locations where people traditionally visit for government or public services, such as hospitals, banks, schools and govern- ment buildings in typical include queues suffering from the effect of poor queue management. It can be seen that user satisfaction is easily inducible from long waiting time, overdraws and non-order in service while this may increase the operational burden of services providers. These problems intensify in rush hours, emergency situations or in times of high-demand, emphasizing the importance for efficient and adaptive queue management solutions. Most of the traditional queue management solutions are based

on static policies including first-come-first-serve and fixed appointment system, or manual token distribution. However, they work in a non-dynamic manner and are not updated based on current arrival rates or service capacity. Users have their waiting time unknown in advance and administrators miss useful predictive information on congestion control and resource allocation. These systems respond to problems once congestion has already taken place.

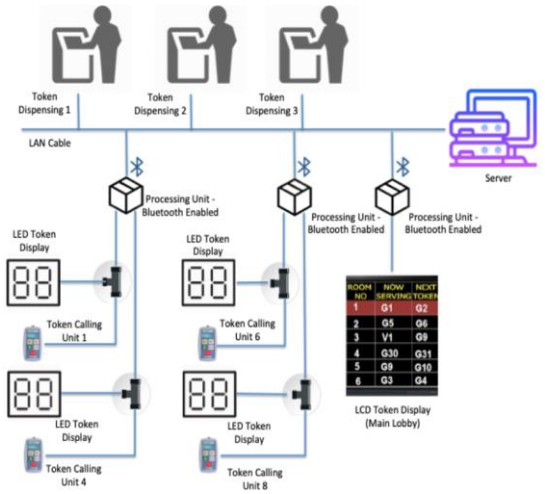

Fig. 1. Conceptual Overview of the IntelliQueue Framework

Recent progress in machine learning and cloud computing has made it possible to build intelligent systems that can analyze both historical and real time data to foretell system behavior. In queue management, after accounting for non-linear relationships between the queue size and arrival process, service rate and time-of-day factors can improve the prediction of waiting time using machine learning models. Nevertheless, some current arrangements concentrate their service on either the booking process online or display functions. The existing systems are based on the distinction between "online" and "offline" user statuses, which causes fairness problems and disjointed service flow.To address these issues, in this paper we present IntelliQueue, an AI-powered adaptive queue orchestration system that uses integrated online-offline token management, ML based waiting time prediction and real-time user notifications. An integrated queueing mechanism based on Timestamp provides support for fair Ordering of Service (FAOS), whereas predictive analysis allows the users to have an in advance congestion control. By reframing queue management from a static to an intelligent and predictive process IntelliQueue is helping its customers address efficiency, transparency and user experience challenges in the delivery of public services.

II. RELATED WORK

Intelligent queue management has been studied increasingly in recent years to eliminate waste in public service. Machine learning based data-driven approaches have achieved much better performances than rule-based queue systems. If you want to predict waiting times or behavior of queues in large scale public services, ensemble and supervised learning approaches have been effectively used [1-3].

There are a number of studies for improving the performance of queues in specialized domains, such as healthcare and public services. Models based on machine learning have achieved less waiting time and better congestion detecting performance through considering arrival rates, service rates, and time varying features [4-6]. This allows for proactive queues control, especially in periods of peak demand.

Developments in the smart city infrastructures have paved the way for cloud-based and cloud-edge QM systems. Recent works emphasize distributed intelligence and IoT-inclusion to realize scalable and real-time queue orchestration [7, 10,12]. Deep learning systems, such as LSTM-based models, have also been explored to exploit temporal dependencies on queue dynamics but they have high computational complexity and a lack of interpretability issue in terms of practical implementation [8,9].

Recent survey works point out that there is a gap to develop practical, predictive and safety-aware queue management mechanisms involving machine learning methods, transmission of information in real time, as well as adaptive service control techniques [11,13-15]. Inspired by these insights, in the current paper we aim to alleviate these knowledge gaps through a novel solution named IntelliQueue system, which integrates both online-offline queue management and predictive

analytics along with safety-aware automation, into an end-to-end framework.

III. PROBLEM STATEMENT

Public service places like hospitals, banks, educational institutions and public agencies often suffer from inefficient queue management that leads to long waiting times and crowding as well as user dissatisfaction. The queue management systems used by most companies are based on static practices such as FIFO (First in first out) or LIFO (Last in last out) rule of managing queues, fixed time appointments and manual issuance of token. These methods cannot adjust for variations in arrival rate and service capacity, as well as peak-hour load.

In addition, many digital queue systems are not fully transparent as they only provide minimal displays for queues or token booking capabilities with inaccurate waiting time estimates. Disjoint queue systems support the two forms of users and results in the imbalances and uneven service sorting. Software control wise, forward looking congestion control and resource allocation is not possible through the absence of predictive information. These limits support the need to adopt a smart, information-enhanced queue handling system that forecasts queue behavior and enables real time decision making.

IV. SYSTEM ARCHITECTURE

The IntelliQueue system that was designed is based on the layered and modular architecture of the intelligent queue management, predicative analytics, and real-time interaction with the user in public service conditions. The designs are aimed at scaling and fairness and good decision support by incorporating machine learning orientated models through a single unit of queue management, as well as cloud services.

Although the system we study is largely based on the cloud-based system, which creates its own issues with the scale of data analysis (particularly with specific issues regarding real-time processing of various forms of data), edge measurements (via future work) or local queue sensing can be naturally utilized in the architecture to offer more specific-purpose ML as a cloud-edge synergy.

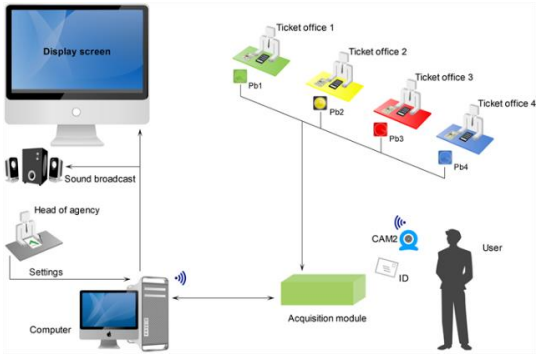

Fig. 2. Overall Architecture of the IntelliQueue framework

A. User Interaction Layer

This middle layer allows for engagement with various actors. Online users reserve service tokens via a web interface, and offline walk-in users are checked in at the service counter with help of the staff. Admins view centralized dashboard to keep track of queue status, estimated waiting time and congestion level indicators.

B. Token Management and Queue Integration Layer

Online tokens and offline tokens are timestamped upon entry. A combined queue is formed by arranging tokens in the order of their time stamps so that regardless of access mode, all users are treated with equally fair first-come-first-serve behavior.

C. Data Processing Layer

This layer extracts queue relevant information such as arrival time, latest queue length, service completing time and time. Preprocessing tasks such as cleaning, normalization and feature extraction are carried out to produce structured data, which can serve as input for models based on machine learning.

D. Intelligence Layer (Machine Learning Engine)

The intelligence layer consists of supervised machine learning regression models for real-time prediction of the waiting time and queue length. Ensemble learning methods aggregate predictions of several models to enhance accuracy and stability against the highly fluctuating service demands.

E. Decision Support and Notification Layer

Projected queue states are examined to identify overload for decisions in action. Users receive automatic notifications on the progress of the queue or the estimated wait time, and admins get predictive intelligence via visual analytics to help manage the queue.

V. METHODOLOGY

This suggested system is named IntelliQueue, which is an adaptive and intelligent system to manage queues in the public service organizations. It incorporates cohesive queue management with real time monitoring and queue service efficiency improving through the acquisition of waiting time prediction. This offers a subtlety where mathematical notation is used sparingly to describe the operation and prediction of queues in form, but without any extra computational this makes sense in light of the complexity of the processing.

A. Unified Queue Formation

Upon a user logging in either remotely or locally, the entry time has a separate token of the entry time produced. The time at which the user i enters, denoted as: t_i

All stones are tossed into one of the pipes and sorted by its entry time. The line is governed by a simple law:

$t_1 \leq t_2 \leq \ldots \leq t_n$

Therefore, services are attended to in sequential order of arrival, first-come first-serve maintaining equity (not only among the users of online but also off-line) users).

Algorithm 1: Timestamp-Based Queue Ordering

Description:

The tokens are placed in queue based on their arrival time. When the service has been served, the token which has been served is dequeued and the successor of the token is processed.

Logic:

Earlier entry time = More priority to services.

B. Queue Data Collection and Feature Representation

The system is continuously keeping records of queue related data including:

- Number of users waiting
- Service completion rate
- Time of the day

These values can be classified into a simple feature set: F=[Q, S, T].

Where,

- Q=current queue length
- S=average service rate
- T= time-related information

This feature set is put into use as input in prediction.

Algorithm 2: *Feature Preparation*

Description:
The queue is purged off the logs and represented in numeric forms as (queue length, speed of service, patterns in time). Such pre-processing ensures that there exist uniform inputs to prediction models.

C. Waiting Time Prediction

Estimation of waiting time is considered a regression. The forecasted waiting time is denoted as: $\hat{W} = f(F)$.

Where,

$\hat{W}$ = predicted waiting time, f = running machine learning regression model, F = queue feature set

In an attempt to increase accuracy, two-model predictions are averaged:

$\hat{W} = (\hat{W}1 + \hat{W}2) / 2$.

Such averaging in a simple form helps to minimize the error in prediction and enhance stability.

Algorithm 3: *Waiting Time Estimation*

Description:
ML models predict waiting time based on current state of the queue. Final waiting time is amalgamation of all model predictions.

D. Congestion Detection

Congestion is identified using simple comparison logic. If the predicted waiting time exceeds a predefined limit, congestion is assumed: $\hat{W} > W_{limit}$

This allows the system to detect peak-hour conditions early.

Algorithm 4: *Congestion Identification*

Description:
When waiting time crosses the acceptable limit, the system flags congestion and alerts administrators for proactive control.

VI. ADVANCED FEATURES OF INTELLIQUEUE

A. Intelligent Missed Token Rescheduling

In traditional queue management systems, users are disenrolled from the queue if they do not show up at their allocated time for service which results in inefficiencies in the service or loss of customers. To overcome this shortcoming, IntelliQueue proposes an Intelligent Missed Token Rescheduling method for flexibly reusing service positions of the missed tokens in a rescheduling way rather than discharging them.

If a customer does not report in the predetermined tolerance period, it compares with the real-time queue statistics, predicted waiting time and current service rate of the system or system load. Under this analysis, a new service PILOT position is dynamically assigned without sacrificing the fairness of the integrated queue. The rescheduled token is requeued at a suitable position which balances service continuity with total level of queue efficiency. This way of working can reduce idle service capacity, manual operation, and improve the user experience by granting control over how much to recover from missed sessions.

B. Safety-Aware Multi-Entry Control with Automated Staff Reallocation

To address crowd congestion and emergency situations in public services, IntelliQueue adopts a Safety-Aware Multi-Entry Control scheme. When the crowd is heavy or demand peaks, the system can flexibly open extra service entrances and re-distribute agents who work at different counters according to estimated queue load and service request.

The population of queue density and congestion indicators is dynamically maintained by the system. Then, automatic decision making is given to reshare load among services, when established safety constraints become over performed (to avoid one counter being too crowded). During emergencies like fire or a stampeding crowd, IntelliQueue triggers emergency exit procedures as well as site safety alerting through visual and audible alarming. This enables a rapid evacuation guidance and the best possible line control in order to safely guide the public.

IntelliQueue – And How It Can Help Service Optimization IntelliQueue's combination of safety-aware automation and predictive queue analytics extends service planning to the domain of crowd management and emergency preparedness for a public sector large scale high traffic environment.

VII. RESULTS & DISCUSSION

The simulation of the new proposed IntelliQueue was tested in the simulated situations of queues which were relevant in the real conditions of the standard of the public service, such as hospitals or administration. The evaluation criteria is the accuracy of such wait time predictions, efficiency in queue management and capability to cope with congestion. The results of IntelliQueue were contrasted with a typical first-come-first-served (FCFS) queue-based approach.

Fig. 3. Comparison of Average Waiting Time between Traditional FCFS and IntelliQueue

A. Experimental Setup

Simulated arrival and service distributions were used to generate queue data that describe natural variation. The characteristics we gathered included arrival rate, queue length, service completion time and a timestamp of when the capture was made. Machine learning regression models were developed based on historical queue information and predictions tested on previously not observed data. The performance was measured with metrics in the categories of average and MAE, and the time readied as an output variable.

B. Waiting Time Prediction Accuracy

A comparison on waiting time prediction accuracy is given in Table 1 between traditional FCFS-based estimation and the proposed IntelliQueue system.

It is seen that IntelliQueue performs and significantly outperforms, regarding prediction error, any of the traditional estimating methods. Using historical and real-time queue data, machine learning models give much more accurate waiting time predictions (especially in peak hours).

Table 1. Waiting Time Prediction Performance

Method	**MAE (minutes)**	**Average Error (%)**
FCFS-based estimation	7.8	21.4
IntelliQueue (ML-based)	3.1	9.2

C. Queue Efficiency Analysis

The average waiting time and the queue throughput were used to assess the impact of IntelliQueue on overall queue efficiency.

Table 2. Queue Efficiency Comparison

Metric	**Traditional Queue**	**IntelliQueue**
Average Waiting Time (min)	**28.5**	**17.2**
Queue Throughput (users/hr)	**42**	**55**

The results show the average waiting time reduced by around 39.6% using IntelliQueue, thus providing high service rate. The consistent-signature clock-based queue system provides fair ordering and offers the ability for a network administrator to perform proactive congestion management based on predictive knowledge.

D. Congestion Detection and Management

The performance of the congestion detection was evaluated by finding peak-hour contexts where the predicted waiting time was greater than a predefined threshold value. IntelliQueue was able to detect congestion prior to legacy switches, which allowed administrative action in a timely manner. In this way, service counters may tailor operations and avoid extended queue buildup.

This contrasts with reactive systems, which can take action only after congestion has already happened. Early warning resulted in even queue and less overcrowding at the service counter.

E. User Experience Improvement

User performance was evaluated by the system behavior qualitatively. Livestream updates and announcements facilitated users to have a sense of queue situation and estimated time. This minimized uncertainty, as well as physical waiting, in crowded situations. Users would be able to time their arrival better with less frustration and anxiety.

F. Discussion

The experimental results show that combining ML based prediction with UQM enhances queueing performance to a great extent. We show that IntelliQueue achieves better prediction accuracy, and results in lower waiting time reduction and congestion treatment by comparison with classical rule-based systems. The results also underscore the critical role of predictive analytics in turning queue management from a reactive to proactive operation.

While the present analysis are conducted on simulated data, there were significantly better performance and the results are very promising for actual use. The modularity of IntelliQueue makes it very easy to be extended into a variety of public service domains with little effort.

VIII. CONCLUSION

This work introduced IntelliQueue, an intelligent queue management system that combines unified timestamp-based queue processing with machine learning-based waiting time prediction. The system creates more efficiency and transparency in the service by pooling online and offline users into one fair queue and taking real-time waiting time into account. We have experimentally observed a reduction in waiting time and an increase of the flow across the queue as compared to rule based systems under simulated conditions. This new methodology demonstrates the capability of predictive analytics to be used in modern day public

service queue management, and offers a scalable paradigm on how the application could happen in a real world.

REFERENCES

[1] Ahmed, S., Mahmood, K., and Rahman, M., "Data-driven waiting time prediction for a large-scale public service queue using an ensemble learning," Expert Systems with Applications, pp. 245, Article 122456, 2024.

[2] Kumar, R., Sharma, A., and Verma, P., "Intelligent Queue Orchestration for Smart Public Service Market by machine learning", IEEE Access2019. 12, pp. 21456-21470, 2024.

[3] L. Chen, Y. Li, and H. Zhao, “Predictive queue management in smart cities with supervised learning techniques,” Knowledge-Based Systems, vol. 279, Article 110345, 2024.

[4] Singh, N. B., and Malathi, T., "Adaptive service queue optimization using machine learning in healthcare environment," As referred in Journal of Ambient Intelligence and Humanized Computing [5]. 15, no. 2, pp. 2141-2156, 2024."

[5] Verma, P., Gupta, S., and Jain, R. Machine learning-based queue congestion detection for public governance systems Expert Systems with Applications. 238, Article 121789, 2024.

[6] Zhang, Y., Wang, L., and Liu, X., “Dynamic waiting time estimation using gradient boosting models for public service queues,” Applied Soft Computing, 10.1016/ j.asoc.2018.09.publ. 145, Article 110567, 2024.

[7] Hassan, M., Al-Fuqaha, A., and Guizani, M. 116 Smart queue management using cloud-edge intelligence for urban services," IEEE Internet of Things Journal, vol. 11, no. 6, pp. 10422-10435, 2024.)

[8] Li, Q., Zhou, S., and Chen, X. Deep learning based temporal modeling for queue length and waiting time prediction Neural Networks vol. 172, pp. 352-364, 2024.

[9] Park, J., and Kim, H., "LSTM-based waiting time prediction of a real-time service system," IEEE Access, vol. 12, pp. 33110-33122, 2024.

[10] Ghosh, A.; Banerjee, S., "Cloud-enabled intelligent queue management framework for smart governance," Future Generation Computer Systems, vol. 150, pp. 302-315, 2024.

[11] Nguyen, T. H., Tran, D., and Jung, J., ``Predictive analytics for service operations: Recent advances and applications,'' Decision Support Systems, vol. 181, Article 113956, 2024.

[12] Al-Turjman, F., Zahmatkesh, H., and Mostarda, L., "AI-enabled crowd-aware queue management systems for smart public infrastructures," Sustainable Cities and Society. 104, Article 105287, 2024.

[13] Li, J., Xie, K., and Sun, Y., Intelligent service systems using machine learning: A comprehensive review, IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 54, no. 3, pp. 1781–1796, 2024.

[14] Kumar, S., Gupta, D., and Malhotra, S., "Machine learning for public service optimization: Trends, challenges and future directions," Sustainable Cities and Society, 2020. 110, Article 105982, 2025.

[15] Rahman, M., Islam, S. and Hossain, M., "Next generation intelligent queue management using cloud-edge synergy," IEEE Access, vol. 13, pp. 55601-55615, 2025.