© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717976

ML-Based Smart Queue Management System for Public Services

DR BALAMURUGAN S[1], DHANUSH S[2]
[1]*Assistant Professor, Department of Computer Science & IT, Jain (Deemed-To-Be-University), Bangalore, India*
[2]*Scholar, Department of Computer Science & IT, Jain (Deemed-To-Be-University), Bangalore, India*

Abstract- ***Public-service queue systems in hospitals, banks, and citizen-service centers still face long waits, weak transparency, and inefficient use of counters. Recent studies show that machine learning and lightweight sensing can improve waiting-time estimation and queue visibility, but the most directly relevant evidence is concentrated in healthcare, banking, and service-center applications. This paper reviews ten highly relevant queue studies and uses them to propose an ML-based smart queue management framework for public-service delivery. The review identifies queue length, arrival rate, service duration, and real-time queue-state signals as the most useful predictive inputs, while the main unresolved gaps are cross-domain generalization, prediction-allocation integration, infrastructure-light deployment, and balanced evaluation of service outcomes. Based on these findings, the paper proposes a framework that combines event capture, preprocessing, feature engineering, waiting-time prediction, dynamic counter allocation, monitoring, and HOT-Fit evaluation. The framework is intended to improve transparency, reduce congestion, and support practical queue modernization in public-service settings.***

Keywords- ***Smart Queue Management System, Waiting-Time Prediction, Machine Learning, Dynamic Counter Allocation, Public Service Delivery, HOT-Fit Evaluation***

I. INTRODUCTION

1.1 Background of Study

Queue management is a critical operational issue in hospitals, banks, ticketing counters, and municipal service offices because delay directly affects satisfaction, fairness, throughput, and staff productivity. Earlier systems focused mainly on token generation and static FIFO handling, but recent work shows that queue behavior is dynamic and should be studied using queue length, arrival surges, service-time variability, and counter availability [1][2][3][4][5][6][8][9][10].

Digital transaction records, mobile interfaces, and camera-based tracking now make it feasible to build queue systems that do more than display token numbers. However, many current studies still separate prediction from operational control, or they describe architecture without strong predictive validation [5][6][7][8][9][10].

1.2 Problem Statement

Traditional queue systems in public-service settings still rely heavily on static first-come-first-served handling, fixed counters, and manual visibility of congestion. This creates long waiting times during peak periods, underutilization during off-peak periods, poor transparency for citizens or customers, and weak decision support for frontline staff. Although recent studies have shown that machine learning can improve waiting-time prediction and that digital queue systems can improve workflow visibility, a consistent journal-ready framework that combines queue-state prediction, dynamic counter allocation, low-cost sensing, and multi-dimensional evaluation is still missing. The problem is therefore not only prediction accuracy, but also how to translate queue intelligence into a realistic operational system that works across public-service conditions.

1.3 Motivation

The motivation for this research is the need for a deployable queue framework rather than another isolated prediction study. Hospitals require fairer service handling, banks need better overload control, citizen-service centers need clearer waiting-time communication, and low-resource environments need approaches that work without expensive new hardware [1][2][3][4][8][9][10].

1.4 Objectives of Study

The objective of this study is to develop a publication-ready ML-based smart queue management framework for public-service delivery using waiting-time prediction, dynamic counter allocation, and workflow-level evaluation grounded in the ten selected research papers.

Objective 1: To examine the methodological contribution of ten highly relevant queue-related studies across healthcare, banking, and public-service settings.

Objective 2: To identify the strongest predictive variables, modeling approaches, and deployment patterns reported in the selected studies.

Objective 3: To propose an integrated smart queue framework that combines queue-state data capture, feature engineering, waiting-time prediction, and dynamic Objective 4:counter-allocation logic.

Objective 4: To define an evaluation plan using technical metrics, service KPIs, and HOT-Fit dimensions for real public-service deployment.

1.5 Contributions of the Paper

This paper consolidates ten highly relevant queue-management studies into one journal-style literature base.

It organizes direct public-service queue papers across healthcare, banking, and service-center settings to strengthen methodological depth.

It proposes an ML-based smart queue framework that connects prediction with operational action instead of treating prediction as the endpoint.

It extends the discussion beyond technical metrics by incorporating deployment, service impact, social value, and academic value.

1.6 Organisation of the Paper

Section 2 reviews the selected ten studies, presents comparative analysis, and identifies research gaps. Section 3 proposes the smart queue methodology, architecture, workflow, and algorithmic design. Section 4 discusses expected outcomes, evaluation strategy, and practical implications. Section 5 outlines applications, social impact, policy relevance, and academic value. Section 6 concludes the paper and highlights the future direction.

II. LITERATURE REVIEW

Research on waiting-time prediction and smart queue management has accelerated because many service systems now produce queue traces that can be analyzed using machine learning and real-time sensing. The selected studies cover healthcare, banking, general service centers, and live queue-tracking settings.

2.1 Review of Related Studies

Study	Method and Key Limitation
Parthasarathi Pattnayak et al. (2023)	Multiple ML/DL models evaluated for patient waiting-time prediction.; limitation: Detailed model reproducibility specifics are limited in extracted text.
Ajinkya Mishra et al. (2023)	Random Forest, Monte Carlo simulation, queue-theory methods, and statistical estimation.; limitation: Paper is more system-proposal oriented; large-scale quantitative validation is limited.
Mayank Dutta et al. (2023)	Priority equation with Tribonacci-based score adjustment for late patients and dynamic reordering.; limitation: Evidence is case-study based; broader external validation remains needed.
Lin Lin Guo et al. (2025)	Linear regression plus eight ML models including tree-based methods.; limitation: Performance varied by task type, showing limited one-model-fits-all behavior.
Tapodhir Karmakar Taton et al. (2024)	Multiple regression models, LSTM, and Voting Regressor ensemble.; limitation: Exact dataset provenance and full reproducibility details are limited in the article text.

Table 1a: Comparative Analysis of Existing Queue Waiting-Time Prediction and Smart Queue Management Studies (Part 1)

Study	Method and Key Limitation
Carolina Loureiro et al. (2023)	DT, RF, GBT, ANN, and AutoML.; limitation: Domain transferability may still vary with organization-specific queue policies.
Athanasios I. Kyritsis et al. (2019)	Fully connected neural network; queue-specific model training; simulator-backed validation.; limitation: Performance may depend on data quality and on retraining frequency per queue type.
Dipta Gomes et al. (2020)	SVR (selected main model), with comparisons against RNN and K-means based approaches.; limitation: Limited explicit reporting of full metric tables in extracted sections reduces direct reproducibility.
Roberto Moraso et al. (2017)	Queueing-theory formula baseline, Deep Learning, Gradient Boost Machine (GBM), and Random Forest (RF).; limitation: Results may depend on local branch operations and regulatory thresholds.
Begukan Gezler et al. (2022)	YOLOv4 object detection and MOSSE tracking.; limitation: Performance can vary with camera angle, crowd density, and occlusion conditions.

Table 1b: Comparative Analysis of Existing Queue Waiting-Time Prediction and Smart Queue Management Studies (Part 2)

Healthcare queue prediction and patient prioritization

The first theme covers healthcare-oriented queue studies, where the objective is not merely to predict delay but to improve fairness, triage sensitivity, transparency, and patient experience.

i. Parthasarathi Pattnayak et al. (2023) - Deep Learning based Patient Queue Time Forecasting in the Emergency Room
The study aimed to predict emergency-room patient queue times using ml/dl to improve prioritization and response in the context of 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS). It addressed the gap that Conventional queueing-theory assumptions are often inadequate for real-world ER variability, and it relied on Real ER dataset with 29,909 patients from 2019 in India [1].

Methodologically, it used multiple ml/dl models evaluated for patient waiting-time prediction with preprocessing centered on Clinical queue records prepared for model training and comparison across methods. The evaluation used Prediction-error comparison across models, and the study found that Data-driven models can better capture ER queue dynamics for practical planning. Its main limitation is Detailed model reproducibility specifics are limited in extracted text, but it remains useful because it strengthens evidence for ml/dl in healthcare queue forecasting [1].

From a framework-design perspective, this paper is especially valuable because its strongest reported strength is Uses real clinical data and directly targets patient-centric operations. This makes it a practical reference for our proposed public-service queue architecture, particularly where transparent waiting-time communication and operational action need to be connected in one system [1].

ii. Ajinkya Mishra et al. (2025) - Hospital Queue Wait Time Prediction
The study aimed to design an automated hospital queue waiting-time prediction system using iot data capture and ml/simulation approaches in the context of International Journal of Creative Research Thoughts (IJCRT). It addressed the gap that Hospital queues often rely on manual tracking and provide poor visibility of expected waiting time, and it relied on RFID-based time-stamped hospital process data collected via proposed IoT setup [2].

Methodologically, it used random forest, monte carlo simulation, queue-theory methods, and statistical estimation with preprocessing centered on Automated data collection and central storage pipeline proposed for real-time monitoring. The evaluation used Comparative predictive accuracy and real-time queue visibility metrics, and the study found that Automated sensing plus prediction can improve transparency in hospital queues. Its main limitation is Paper is more system-proposal oriented; large-scale quantitative validation is limited, but it remains useful because it adds an iot-enabled architecture for practical hospital queue prediction [2].

From a framework-design perspective, this paper is especially valuable because its strongest reported strength is Strong end-to-end system framing from sensing to dashboard. This makes it a practical reference for our proposed public-service queue architecture, particularly where transparent waiting-time communication and operational action need to be connected in one system [2].

iii. Mayank Dutta et al. (2023) - Smart Queueing Management System for Digital Healthcare
The study aimed to design a smart digital-healthcare queue system that prioritizes patients by age, appointment time, severity, and waiting fairness in the context of 2023 International Conference on Recent Advances in Electrical, Electronics & Digital Healthcare Technologies (REEDCON). It addressed the gap that Conventional appointment queues often fail to balance urgency, fairness, and late-arrival handling, and it relied on Case-study deployment data from healthcare setting used to evaluate queue performance outcomes [3].

Methodologically, it used priority equation with tribonacci-based score adjustment for late patients and dynamic reordering with preprocessing centered on Patient attributes normalized into a dynamic scoring mechanism with timeout and fairness controls. The evaluation used Waiting-time reduction, response-time improvement, and patient satisfaction changes, and the study found that Dynamic score-based prioritization can improve both fairness and urgency handling. Its main limitation is Evidence is case-study based; broader external validation remains needed, but it remains useful because it contributes a configurable fairness-aware priority framework for healthcare queues [3].

From a framework-design perspective, this paper is especially valuable because its strongest reported strength is Practical algorithmic design tailored for healthcare triage-like conditions. This makes it a practical reference for our proposed public-service queue architecture, particularly where transparent waiting-time communication and operational action need to be connected in one system [3].

iv. Lin Lin Guo et al. (2025) - Predicting Waiting Times for Medical Tasks in a Pediatric Hospital Using Machine Learning: Comprehensive, Retrospective, Real-World Study
The study aimed to develop task-specific ml models to predict pediatric-hospital medical-task waiting times using queue-theory guided features in the context of JMIR Medical Informatics. It addressed the gap that Overcrowding and pediatric resource constraints require accurate per-task waiting-time predictions; generic models may be insufficient, and it relied on 230,864 time-stamped records from Nov 1, 2024 to Mar 13, 2025 from pediatric hospital systems [4].

Methodologically, it used linear regression plus eight ml models including tree-based methods with preprocessing centered on Data preprocessing of laboratory and radiology task records before model training and validation. The evaluation used MAE, MSE, RMSE, and R2; feature importance via SHAP, and the study found that Queue-related predictors, especially queued-patient count, were most influential for waiting-time prediction. Its main limitation is Performance varied by task type, showing limited one-model-fits-all behavior, but it remains useful because it provides robust evidence for task-specific queue-time modeling in pediatric care [4].

From a framework-design perspective, this paper is especially valuable because its strongest reported strength is Large real-world dataset and rigorous evaluation design with explainability. This makes it a practical reference for our proposed public-service queue architecture, particularly where transparent waiting-time communication and operational action need to be connected in one system [4].

Cross-domain and service-center waiting-time analytics
The second theme focuses on general service queues and banking-style service centers, where waiting-time prediction is tied to customer communication, counter efficiency, and overflow prevention.

v. Tapodhir Karmakar Taton et al. (2024) - Waiting Time Prediction in Queue Management: Leveraging Machine Learning Approach
The study aimed to develop a scalable queue management approach that predicts client waiting time in real time using ml and lstm-based models in the context of 2024 International Conference on Recent Progresses in Science, Engineering and Technology (ICRPSET). It addressed the gap that Traditional physical queue systems require physical presence and provide limited real-time prediction and scalability, and it relied on Simulation-based queue scenarios and service flow records (exact raw source details not fully specified in paper) [5].

Methodologically, it used multiple regression models, lstm, and voting regressor ensemble with preprocessing centered on General preprocessing for regression/deep learning pipeline (cleaning and split; detailed steps not fully specified). The evaluation used R2 score and MAE (minutes), and the study found that Ensemble regression performed best for waiting-time prediction and supports dynamic counter allocation decisions. Its main limitation is Exact dataset provenance and full reproducibility details are limited in the article text, but it remains useful because it shows how ml-based waiting-time estimation can be operationalized in practical queue systems [5].

From a framework-design perspective, this paper is especially valuable because its strongest reported strength is Strong predictive accuracy and practical relevance for real-time queue operations. This makes it a practical reference for our proposed public-service queue architecture, particularly where transparent waiting-time communication and operational action need to be connected in one system [5].

vi. Carolina Loureiro et al. (2023) - Predicting Multiple Domain Queue Waiting Time via Machine Learning
The study aimed to predict queue waiting time across multiple service domains using crisp-dm and machine learning models in the context of Computational Science and Its Applications - ICCSA 2023 Workshops (Lecture Notes in Computer Science). It addressed the gap that Existing rigid business formulas were insufficient for accurate multi-domain queue-time estimation, and it relied on Millions of ticket records from 58 stores across five domains; one-year modeling and two-year deployment simulation periods [6].

Methodologically, it used dt, rf, gbt, ann, and automl with preprocessing centered on Data cleaning, scaling, and feature engineering under CRISP-DM workflow. The evaluation used MAE, NMAE, RMSE, and AREC; plus computational effort (training and prediction time), and the study found that Feature engineering and AutoML improved queue-time prediction quality across domains. Its main limitation is Domain transferability may still vary with organization-specific queue policies, but it remains useful because it demonstrates scalable enterprise-ready queue prediction methodology [6].

From a framework-design perspective, this paper is especially valuable because its strongest reported strength is Large real-world dataset and rigorous validation design. This makes it a practical reference for our proposed public-service queue architecture, particularly where transparent waiting-time communication and operational action need to be connected in one system [6].

vii. Athanasios I. Kyritsis et al. (2019) - A Machine Learning Approach to Waiting Time Prediction in Queueing Scenarios
The study aimed to assess whether machine learning can predict queue waiting times across industries and support automated queue management in the context of 2019 Second International Conference on Artificial Intelligence for Industries (AI4I). It addressed the gap that Classical queueing-theory approaches are often rigid and less adaptable to heterogeneous real queue scenarios, and it relied on Public bank-queue dataset plus simulator-generated multi-scenario queue data for system validation [7].

Methodologically, it used fully connected neural network; queue-specific model training; simulator-backed validation with preprocessing centered on Data preparation and per-queue model training workflow in the proposed web system. The evaluation used MAE in minutes for waiting-time prediction, and the study found that ML is a viable alternative to analytical queue formulas for practical waiting-time estimation. Its main limitation is Performance may depend on data quality and on retraining frequency per queue type, but it remains useful because it bridges ml prediction and deployable queue-management software in one framework [7].

From a framework-design perspective, this paper is especially valuable because its strongest reported strength is Generalizable architecture with adaptive per-queue training for different service contexts. This makes it a practical reference for our proposed public-service queue architecture, particularly where transparent waiting-time communication and operational action need to be connected in one system [7].

viii. Dipta Gomes et al. (2020) - Banking Queue Waiting Time Prediction based on Predicted Service Time using Support Vector Regression
The study aimed to predict banking queue waiting time from service-time related features using machine learning methods in the context of 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM). It addressed the gap that Customers face uncertain waits; practical queue prediction in developing-country banking contexts remains limited, and it relied on Real-life dataset of queues from multiple banks [8].

Methodologically, it used svr (selected main model), with comparisons against knn and k-means based approaches with preprocessing centered on Authors emphasize use of a well-preprocessed structured dataset before modeling. The evaluation used Comparative model evaluation (exact metric list not clearly detailed in extracted sections), and the study

found that SVR showed practical feasibility for real-world bank queue waiting-time prediction. Its main limitation is Limited explicit reporting of full metric tables in extracted sections reduces direct reproducibility, but it remains useful because it adds evidence for svr-based queue prediction in practical banking scenarios [8].

From a framework-design perspective, this paper is especially valuable because its strongest reported strength is Uses real multi-bank data and compares regression/classification/clustering perspectives. This makes it a practical reference for our proposed public-service queue architecture, particularly where transparent waiting-time communication and operational action need to be connected in one system [8].

ix. Roberto Mourao et al. (2017) - Predicting Waiting Time Overflow on Bank Teller Queues
The study aimed to predict time-overflow risk in bank teller queues early enough to support proactive intervention in the context of 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). It addressed the gap that Static warning thresholds (for example 5-minute alerts) are often insufficient to prevent waiting-time violations, and it relied on Operational bank queue records labeled by overflow behavior (paper context) [9].

Methodologically, it used queueing-theory formula baseline, deep learning, gradient boost machine (gbm), and random forest (rf) with preprocessing centered on Data preparation for predictive modeling with validation over historical queue records. The evaluation used Accuracy and F1-measure, and the study found that GBM outperformed compared approaches for early overflow-risk detection in teller queues. Its main limitation is Results may depend on local branch operations and regulatory thresholds, but it remains useful because it provides actionable early-warning modeling for bank queue governance [9].
From a framework-design perspective, this paper is especially valuable because its strongest reported strength is Clear operational framing and strong predictive performance for compliance-oriented use cases. This makes it a practical reference for our proposed public-service queue architecture, particularly where transparent waiting-time communication and operational action need to be connected in one system [9].

Real-time sensing and lightweight queue monitoring
The third theme examines real-time sensing methods that estimate queue conditions from live video inputs without depending entirely on formal token systems.

x. Dogukan Gozler et al. (2022) - A Real-time Queue Tracking Method for Waiting Time Estimation
The study aimed to estimate average queue waiting time in real time by detecting and tracking people in queue videos in the context of 2022 30th Signal Processing and Communications Applications Conference (SIU). It addressed the gap that Many queue locations lack token systems; manual waiting-time estimation is unreliable and not real time, and it relied on Experimental queue video data at 640x480 resolution [10].

Methodologically, it used yolov4 object detection and mosse tracking with preprocessing centered on Object detection plus tracking pipeline optimized for real-time execution. The evaluation used Accuracy and processing speed (FPS), and the study found that Tracking-assisted pipeline enables practical real-time wait estimation from video streams. Its main limitation is Performance can vary with camera angle, crowd density, and occlusion conditions, but it remains useful because it contributes an implementable cv method for real-time queue wait estimation [10].

From a framework-design perspective, this paper is especially valuable because its strongest reported strength is Strong real-time performance for a queue-monitoring vision application. This makes it a practical reference for our proposed public-service queue architecture, particularly where transparent waiting-time communication and operational action need to be connected in one system [10].

2.2 Comparative Analysis of Existing Methods
Across the ten studies, several method families recur. Direct public-service queue papers mainly rely on supervised regression, ensemble learning, SVR, or deep learning to predict waiting time from queue-state and service-duration variables. Cross-domain

papers emphasize feature engineering and model adaptation, while banking studies place more emphasis on service-time estimation and overflow control [1][2][3][4][6][7][8][9].

The sensing study adds live queue visibility through video-based monitoring, showing that practical queue estimation can also work where formal token systems are incomplete [10]. Overall, the literature shows a shift from static rules toward adaptive queue prediction systems.

2.3 Critical Review

The literature offers strong evidence that waiting-time prediction is feasible, yet three limitations remain visible. First, most papers are domain-bounded. Second, even when prediction is strong, the link to resource reallocation is often weak. Third, evaluation criteria vary widely, which makes direct comparison difficult [2][3][4][5][6][8][9][10].

At the same time, the papers collectively show a useful design pattern: successful queue studies tend to combine clean event records, context-aware features, and action-oriented outputs rather than relying on prediction alone. This insight directly shapes the proposed framework in the next section.

2.4 Identified Research Gaps

The review points to four practical gaps: no generalized framework across public-service domains, limited integration of real-time signals with queue decisions, weak low-cost deployment guidance, and uneven evaluation of technical and service outcomes.

III. PROPOSED METHODOLOGY

The proposed methodology develops a smart queue management framework for public-service delivery by combining queue-event capture, feature engineering, waiting-time prediction, dynamic counter allocation, and service evaluation. The framework is influenced most strongly by the healthcare prediction papers, the multi-domain queue analytics papers, the banking studies, and the sensing study [1][2][3][5][6][8][10].

3.1 System Overview

The proposed smart queue system consists of the following modules:

- Queue Event Module
- Structured Data Processing Module
- Prediction Module
- Counter Allocation Module
- Dashboard and Alerting Module
- Real-Time Monitoring Module
- HOT-Fit Evaluation Module

The system processes transaction records, queue-state observations, and resource-status signals to calculate waiting-time predictions and congestion indicators that can be acted on immediately by the service operator.

3.2 Framework Architecture

This section presents the major components required for the proposed public-service smart queue framework.

3.2.1 Data Collection

The framework collects queue-event logs, arrival timestamps, service-start and service-end records, counter availability, service category, token progression, and optional sensing inputs where feasible. This design is informed by healthcare, banking, and queue-tracking studies [1][3][8][10].

- Historical queue-event records from hospitals, banks, and service counters
- Service timestamps, priority flags, and counter-status information
- Optional sensing signals from cameras or service-area sensors
- Operational indicators such as no-shows, backlog, and peak-hour demand

3.2.2 Data Preprocessing

The preprocessing stage performs time alignment, duplicate removal, missing-value handling, event harmonization, and service-stage normalization. The selected papers repeatedly show that waiting-time prediction is highly sensitive to timestamp quality and queue-state consistency [2][6][7][8].

- Timestamp synchronization and record ordering

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717976

- Queue-stage normalization and duplicate handling
- Outlier filtering for abnormal service times
- Aggregation into operational windows for near-term prediction

3.2.3 Feature Engineering

Feature engineering converts raw queue traces into predictive operational variables. The most useful variables reported across the studies include queue length, arrival rate, service rate, service-duration history, priority class, counter status, and sensing-derived occupancy signals [4][5][6][10].

Waiting-Time Prediction Function

$$W_{pred} = f(Lq, \lambda, \mu, c, s, p)$$

Figure 1: Waiting-time prediction function for queue-state and service-state features

3.2.4 Waiting-Time Prediction

After feature engineering, the system predicts near-term waiting time using candidate models such as Random Forest, Gradient Boosting, SVR, feed-forward neural models, and LSTM-based predictors. The goal is to use a repeatable protocol that identifies the most reliable model for each service environment [2][5][6][7][8].

Prediction outputs are stored as waiting minutes and congestion indicators so that the framework can trigger action before conditions deteriorate [8][9].

3.2.5 Dynamic Counter Allocation

The main operational extension beyond waiting-time prediction is dynamic counter allocation. When predicted waiting time, congestion risk, or overflow probability crosses configured thresholds, the system recommends actions such as opening counters, reassigning staff, or rerouting service categories [3][5][8][9].

Counter Allocation Priority Score

$$S = \alpha R + \beta Q + \gamma D - \delta I$$

Figure 2: Counter-allocation priority score for congestion-sensitive queue control

3.2.6 Real-Time Monitoring and Feedback

The framework also includes a monitoring layer that updates dashboard metrics, citizen-facing waiting-time displays, and queue alerts. Depending on the deployment context, the monitoring layer can rely on token events alone or combine those events with video signals. The literature shows that lightweight monitoring improves both operational visibility and user trust because the queue no longer behaves like a black box [10].

Research Gap	Proposed Solution
Prediction without action	Prediction linked to counter reallocation
Domain-specific models	Shared feature framework with local tuning
Weak real-time visibility	Dashboards supported by token, video, or mobile signals
Narrow evaluation	Technical, operational, and HOT-Fit evaluation
Hardware-heavy systems	Infrastructure-light deployment options

Table 2: Research gaps, methodology used, and proposed solutions for the smart queue framework

3.2.7 HOT-Fit Evaluation

The Human dimension evaluates user satisfaction and staff acceptance. The Organizational dimension evaluates throughput, congestion reduction, and counter utilization. The Technological dimension evaluates prediction accuracy, sensing feasibility, and scalability.

Using HOT-Fit in this context is important because queue systems succeed only when predictions are understood and acted upon by both staff and end users. A technically accurate model that is ignored operationally would have limited public-service value.

3.2.8 Deployment Considerations

For practical deployment, the framework should support both low-infrastructure and medium-infrastructure service environments. In smaller offices, the system may operate mainly from token logs and staff inputs, while larger facilities can add sensor or camera-assisted monitoring.

This staged deployment approach helps institutions modernize queue operations without requiring a full technology overhaul at the first step. It also makes the framework more realistic for public-service rollout.

3.3 Workflow Diagram

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717976

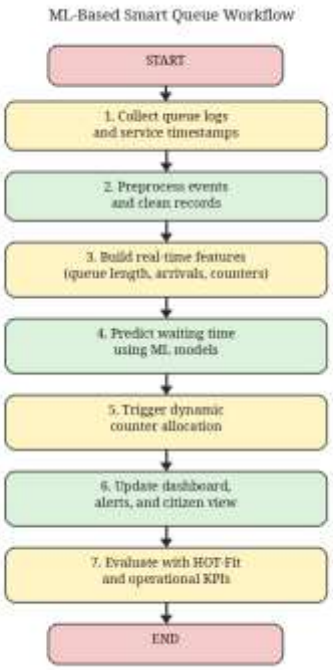

Figure 3: Workflow diagram of the ML-based smart queue management system

3.4 Algorithm Used

3.4.1 Ensemble and Tree-Based Models

Tree-based and ensemble models are useful because they handle heterogeneous queue variables without strong parametric assumptions. Random Forest and Gradient Boosting repeatedly appear in healthcare and banking queue studies [2][4][8][9].

3.4.2 Neural and Temporal Prediction Models

Neural models are most valuable when queue behavior is highly time dependent. The literature includes feed-forward predictors, ER forecasting, and LSTM-based waiting-time prediction where sequence behavior matters [1][5][7].

3.4.3 Real-Time Tracking and Sensing

Real-time tracking methods such as YOLO-based queue detection add a useful deployment pathway where formal ticket systems are incomplete [10].

3.4.4 Rule-Based Operational Decision Layer

The final layer translates model output into actions such as opening counters, staff reassignment, overflow alerts, and citizen-facing updates.

IV. EXPECTED RESULTS AND DISCUSSION

The proposed smart queue management framework is expected to improve waiting-time transparency, operational efficiency, and resource utilization in service environments that currently rely on static queue handling. The reviewed literature suggests that accurate waiting-time prediction can reduce uncertainty for users, while dynamic queue control can reduce preventable congestion [1][2][6][8][9].

A second expected outcome is better service coordination. When the system predicts congestion early and links it to counter-allocation rules, managers can intervene before queue conditions deteriorate.

4.1 Expected Outcomes

i. Performance Improvements

The framework is expected to improve waiting-time prediction accuracy and enable earlier congestion intervention than static first-come queue handling.

ii. Robustness

The system should remain useful across multiple queue settings because the feature pipeline is designed around common operational variables rather than one institution-specific heuristic.

iii. Scalability

The architecture can be extended from one service site to multi-counter and multi-department environments using the same event-driven queue core.

Metric	Expected Value
MAE	< 10 minutes
Average Waiting Time	Lower than static FCFS
Throughput	Higher counter utilization
Service Transparency	Visible real-time updates
HOT-Fit Outcome	Positive user, organizational, and technical gains

Table 3: Expected performance metrics of the proposed smart queue management system

4.2 Comparative Evaluation Plan

The proposed system will be evaluated using model-error metrics, service KPIs, and implementation measures.

- Mean Absolute Error (MAE)
- RMSE and R2 where applicable

- Average waiting time and queue-length reduction
- Counter utilization and throughput improvement
- User satisfaction, staff acceptance, and service transparency

The results will be compared with static queue handling, non-predictive digital queue workflows, and where possible, simpler baseline models such as rule-only or naive estimates.

A phased evaluation strategy is also recommended. The first phase should use retrospective queue records to compare models, the second should test decision rules in controlled operational settings, and the third should evaluate user-facing transparency and staff adoption in live deployment.

4.3 Discussion

Unlike conventional queue systems that mainly display queue position or token number, the proposed framework treats prediction as a decision-support service. This makes the architecture more valuable for public-service management because it not only informs the citizen but also informs the service operator. The integration of queue analytics with HOT-Fit evaluation also gives the framework a stronger implementation focus than many purely technical studies.

The broader implication is that queue modernization should not be defined only by faster software or digitized tokens. A truly modern queue system should understand service demand, anticipate delay, support timely intervention, and remain feasible in institutions that have uneven digital maturity. This is the niche the proposed research framework is intended to address.

The ten-paper evidence base also suggests that implementation success depends on matching model complexity to operational maturity. In some settings, simpler explainable models may be more practical than deeper models if they support consistent intervention and easier staff adoption.

V. APPLICATIONS AND USE CASES

The proposed smart queue management system can be applied in hospitals, banks, citizen-service centers, appointment-driven service environments, and infrastructure-light facilities where queue visibility is limited.

5.1 Industry Use

Hospitals can use the framework to estimate task-specific waits and manage bottlenecks across registration, laboratory, billing, and consultation counters. Banks can use it to monitor teller congestion and predict overflow risk. Citizen-facing service offices can use it to reduce crowding at grievance, licensing, and utility counters.

Because the framework is event-driven, the same core design can also be adapted to appointment-backed counters, campus service centers, and other multi-desk public environments where demand changes throughout the day.

5.2 Social Impact

The proposed system can improve fairness and transparency in service delivery. By making waiting-time estimates visible and supporting more consistent counter allocation, the system can reduce uncertainty and improve the service experience for citizens, patients, and customers.

This is especially relevant in high-pressure service environments where long and poorly explained waits often create frustration. Better queue visibility can support trust even before total waiting time is fully optimized.

5.3 Policy Relevance

Queue modernization is increasingly relevant to digital-governance and service-improvement policies. A predictive queue framework supports transparent service standards, performance monitoring, and operational accountability. Because the framework can be deployed with varying levels of infrastructure, it is useful for phased modernization strategies.

5.4 Academic Value

Academically, the proposed research contributes to machine learning, service operations, public-service systems, and applied queue analytics. It brings together evidence from healthcare, service counters,

computer vision, banking operations, and queue-state prediction into one queue-management framework.

The study also shows how literature-backed design can be translated into a publishable research framework.

A further academic contribution is the explicit linkage between prediction models, counter-allocation logic, and service-evaluation criteria. This linkage can support future empirical studies that want to move beyond algorithm comparison toward operational research impact.

5.5 Research and Deployment Limitations

Even with a strong ten-paper evidence base, some limitations remain. Many selected studies are still restricted to one institution, one queue type, or one operational dataset, so broader public-service validation is still necessary before general deployment claims can be made.

There are also implementation constraints such as data quality, staff readiness, counter reallocation policies, and uneven digital infrastructure. Recognizing these limits early improves the realism of the proposed framework and helps position future pilot studies more credibly.

VI. CONCLUSION

This research proposes an ML-based smart queue management framework for public-service delivery using waiting-time prediction and dynamic counter allocation. The review of ten highly relevant papers shows that queue prediction is feasible across hospitals, banks, general service counters, and tracking-based queue environments [1][2][6][7][8][10].

The framework extends this literature by emphasizing that prediction alone is insufficient. Public-service value emerges when queue intelligence is connected to service action, visibility, and evaluation. Future work may include live multi-site deployment, adaptive online learning, and stronger validation across public-service sectors.

Overall, the paper positions smart queue management as both a machine-learning problem and an operational decision problem. This dual emphasis is what makes the framework relevant for publication-oriented research as well as future real-world implementation.

REFERENCES

[1] Parthasarathi Pattnayak, Tulip Das, Arpeeta Mohanty, and Sanghamitra Patnaik, "Deep Learning based Patient Queue Time Forecasting in the Emergency Room," 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), pp. 541-545, 2023.

[2] Ajinkya Mishra, Aryan Kapse, Kunal Nikam, Rajat Ingale, and Suresh Jajoo, "Hospital Queue Wait Time Prediction," International Journal of Creative Research Thoughts (IJCRT), pp. e570-e574, 2025.

[3] Mayank Dutta, Fakhra Najm, Dhruv Tomar, and Shabana Mehfuz, "Smart Queueing Management System for Digital Healthcare," 2023 International Conference on Recent Advances in Electrical, Electronics & Digital Healthcare Technologies (REEDCON), pp. 541-546, 2023.

[4] Lin Lin Guo, Rui Tang, Jia Yang Wang, Si Zheng, Yin Zeng, Jun Hou, Mo Chen Dong, Jiao Li, and Ying Cui, "Predicting Waiting Times for Medical Tasks in a Pediatric Hospital Using Machine Learning: Comprehensive, Retrospective, Real-World Study," JMIR Medical Informatics, pp. e77297-e77297, 2025.

[5] Tapodhir Karmakar Taton, Bipin Saha, Amena Akter, Md. Johirul Islam, and Shaikh Khaled Mostaque, "Waiting Time Prediction in Queue Management: Leveraging Machine Learning Approach," 2024 International Conference on Recent Progresses in Science, Engineering and Technology (ICRPSET), pp. 1-5, 2024.

[6] Carolina Loureiro, Pedro José Pereira, Paulo Cortez, Pedro Guimarães, Carlos Moreira, and André Pinho, "Predicting Multiple Domain Queue Waiting Time via Machine Learning," Computational Science and Its Applications -

ICCSA 2023 Workshops (Lecture Notes in Computer Science), pp. 404-421, 2023.

[7] Athanasios I. Kyritsis and Michel Deriaz, "A Machine Learning Approach to Waiting Time Prediction in Queueing Scenarios," 2019 Second International Conference on Artificial Intelligence for Industries (AI4I), pp. 17-21, 2019.

[8] Dipta Gomes, Rashidul Hasan Nabil, and Kamruddin Nur, "Banking Queue Waiting Time Prediction based on Predicted Service Time using Support Vector Regression," 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM), pp. 145-149, 2020.

[9] Roberto Mourao, Ricardo Carvalho, Rommel Carvalho, and Guilherme Novaes Ramos, "Predicting Waiting Time Overflow on Bank Teller Queues," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 842-847, 2017.

[10] Dogukan Gozler, Beyazit Isik, and Cihan Topal, "A Real-time Queue Tracking Method for Waiting Time Estimation," 2022 30th Signal Processing and Communications Applications Conference (SIU), pp. 1-4, 2022.