Journal of Frontiers in Multidisciplinary Research

# Machine-Learning Models for Emergency Department Patient-Flow Forecasting

**Thomas M Taylor [1*], Kathryn P Lopez [2], Jerry F James [3]**
[1-3] College of Science and Mathematics, School of Computing, Montclair State University, USA

* Corresponding Author: **Thomas M Taylor**

## Article Info

**E-ISSN:** 3050-9726
**P-ISSN:** 3050-9718
**Volume:** 06
**Issue:** 02
**July – December 2025**
**Received:** 19-10-2025
**Accepted:** 22-11-2025
**Published:** 16-12-2025
**Page No:** 574-582

## Abstract

Accurate forecasting of emergency department (ED) patient flow is essential for safe staffing, capacity planning, and escalation protocols; however, demand demonstrates significant diurnal and weekly fluctuations, as well as disruptions due to holidays and weather, and feedback from inpatient boarding. We compare multi-horizon forecasts (1h, 4h, 12h, 24h) from seasonal ARIMA and Prophet, tree-based learners (Random Forest, XGBoost), sequence models (LSTM), and a Temporal Fusion Transformer (TFT). We assess point accuracy (MAPE, sMAPE, MAE, RMSE), calibration, residual structure, and operational relevance using a six-year synthetic yet realistic dataset that includes exogenous signals such as calendar, weather, syndromic indicators, occupancy, and triage mix. Results indicate that horizon-specific training with exogenous features produces significant improvements; TFT achieves the highest overall accuracy and calibration, aligning with recent multi-site and occupancy-forecasting studies that highlight exogenous factors and probabilistic assessment (Rostami-Tabar *et al*., 2023; Porto *et al*., 2024; Tuominen *et al*., 2024). Long-term accuracy goes down, but XGBoost is still competitive at 12–24 hours. Using exceedance probabilities, forecast envelopes, and interpretable attributions (permutation importance, attention weights), we turn forecasts into operations that take risks into account. Figures 1–4 and Tables 1–2 present decision-relevant facts instead of comprehensive statistics, in accordance with the directive to convey a coherent operational narrative. The results support a layered deployment: short-term tactical control with TFT/LSTM, mid-term scheduling with XGBoost, and strong classical fallbacks. We finish with a deployment playbook that includes data quality, drift monitoring, bias audits, and change management to make sure that everyone can use it safely and fairly. Recent public health signals, such as increased influenza seasons, further necessitate strong exogenous coverage in production pipelines (CDC, 2024).

**DOI:** https://doi.org/10.54660/.JFMR.2025.6.2.574-582

**Keywords:** Emergency Department, Patient-Flow Forecasting, Multi-Horizon Time Series, Probabilistic Forecasting and Calibration, Temporal Fusion Transformer (TFT), Long Short-Term Memory (LSTM), Gradient-Boosted Trees (XGBOOST), Exogenous Predictors (Calendar, Weather, Syndromic), Ed Occupancy and Boarding, Healthcare Operations, Staffing, and Surge Management

## 1. Introduction

Emergency departments (EDs) are where the needs of the community and the hospital's ability to meet those needs meet. When queues are mismatched, they get longer, more people leave without being seen, and bad results go up (Asplin *et al*., 2003) [3] (Hoot & Aronsky, 2008) [26] (Sun *et al*., 2013) [52]. Anticipatory intelligence—predicting arrivals and downstream load—allows for proactive staffing, bed-flow coordination, and escalation triggers, which reduces congestion and makes things safer (Green *et al*., 2006) [20] (Wiler *et al*., 2011) [60] (Morley *et al*., 2018) [40]. However, the demand for ED is not steady; it changes with the seasons, spikes on holidays and during bad weather, and is affected by the spread of infectious diseases. It is also affected by the number of patients in the hospital and the number of patients waiting to be seen (Jones *et al*., 2008) [30] (Forero *et al*., 2011) [16] (Powell *et al*., 2012) [43].

Classical time-series methods like Poisson/GLM, seasonal ARIMA, and exponential smoothing can find patterns. However, they don't always work well when interactions, regime changes, and covariate streams are important, or when managers need precise multi-step forecasts (De Livera *et al*., 2011) [14] (Hyndman & Athanasopoulos, 2021) [28] (Chatfield, 2000) [10]. Modern machine learning alternatives use nonlinear function approximation and a lot of outside features. Tree-based ensembles (Random Forest, Gradient Boosting, XGBoost) and deep sequence models (LSTM; attention-based hybrids like TFT) have shown improvements in accuracy and, most importantly, can make calibrated probabilistic forecasts that help people make decisions that take risks into account. (Breiman, 2001) [5] (Chen & Guestrin, 2016) [12] (Hochreiter & Schmidhuber, 1997) [25] (Lim *et al*., 2021) [32] (Gneiting & Raftery, 2007) [18].

Recent work emphasises three themes. First, probabilistic and horizon-specific forecasts are operationally significant; hourly emergency department arrival studies employing GAM-based and advanced exponential smoothing techniques underscore calibration as crucial for staffing and surge thresholds (Rostami-Tabar *et al*., 2023) [46]. Second, external signals—calendar and meteorological data across multiple emergency departments—enhance generalizability and stability (Porto *et al*., 2024) [42]. Third, occupancy and crowding targets set with multiple inputs (like bed availability in the area, traffic, and weather) show that advanced ML (LightGBM, DeepAR, N-BEATS, TFT) does better than statistical benchmarks for 24-hour horizons (Tuominen *et al*., 2024) [55]. In addition, research on ED waiting times says that complete predictive distributions, not just point estimates, should be used to help patients and systems make decisions (Arora *et al*., 2023) [2].

INFORMS Pubs Online The public health context makes the case for exogenous coverage and probabilistic envelopes stronger. National surveillance indicates shifting influenza testing trends in emergency departments from 2013 to 2022, mirroring the changing respiratory seasons and test usage (CDC Data Brief, 2024) [7]. Pilot deployments by the CDC show that short-term, real-time machine learning can help with tactical planning by predicting when service demand will be highest during winter surges (Morbey *et al*., 2024) [39]. At the same time, attention-based transformers show promise for multi-horizon clinical time series, like ICU vital-sign trajectories. This suggests that they could be helpful for ED flow (He *et al*., 2025) [23].

In this context, we present an operations-aligned benchmarking of ED patient-flow forecasting across four-time horizons (1h, 4h, 12h, 24h), comparing classical, tree-based, and deep/attention models using standardised feature engineering and stringent backtesting. We check for accuracy, calibration, residual structure, and operational translation (exceedance probabilities, interpretable drivers). We also advise on deployment (governance, drift, equity) and put our synthetic results in the context of recent multi-site and occupancy studies to help people use them.

## 2. Literature Review

ED crowding leads to delays, bad results, and stress for staff (Asplin *et al*., 2003) [3] (Hoot & Aronsky, 2008) [26] (McCarthy *et al*., 2008) [36]. Stochastic arrivals, variable acuity, bottlenecks, and boarding related to inpatient bed turnover are some of the leading causes (Forero *et al*., 2011) [16] (Powell *et al*., 2012) [43]. Forecasting helps with staffing, diagnostics throughput, and policies for diversion and escalation (Green *et al*., 2006) [20] (Wiler *et al*., 2011) [60].

**Classical time-series and regression:** Early exploratory data analysis employed Poisson/GLM with seasonal components; SARIMA and exponential smoothing tackled autocorrelation and seasonality (Jones *et al*., 2008) [30] (De Livera *et al*., 2011) [14] (Hyndman & Athanasopoulos, 2021) [28]. Prophet adds holiday effects to trend-seasonality decomposition (Taylor & Letham, 2018) [54]. These methods offer clarity and rapid adaptation but encounter difficulties with intricate interactions and sudden changes (Chatfield, 2000) [10] (Makridakis *et al*., 2018) [35].

**Modern ML for ED operations:** Ensembles (Random Forest, XGBoost, LightGBM) manage nonlinearities and interactions (Breiman, 2001) [5] (Chen & Guestrin, 2016) [12] (Ridgeway, 2007) [45] (Prokhorenkova *et al*., 2018) [44]. Empirical research indicates benefits when exogenous covariates are present—such as weather, calendar, and syndromic proxies (Porto *et al*., 2024) [42]. Deep sequence models encapsulate extensive temporal structures (Hochreiter & Schmidhuber, 1997) [25] (Salinas *et al*., 2020) [47] (Smyl, 2020) [51] (Lim *et al*., 2021) [32]. Recent studies on occupancy and crowding show that LightGBM, DeepAR, N-BEATS, and TFT are better at 24-hour horizons (Tuominen *et al*., 2024) [55].

**Exogenous drivers:** Weather, holidays, academic calendars, and infectious disease metrics (e.g., ILI) are significant factors (Chan *et al*., 2015) [9] (Xu *et al*., 2013) [62] (Zhao *et al*., 2018) [63]. National surveillance trends support the incorporation of respiratory proxies. Hospital-internal signals, such as occupancy, boarding, and triage mix, improve forecasts by showing how throughput is affected by limitations (McCoy *et al*., 2018) [37] (Wargon *et al*., 2009) [58]. Real-time operational and meteorological signals enable shift-level planning (Hu *et al*., 2023) [27]. Multi-site studies spanning 11 EDs underscore the significance of generalizable calendars/meteorology and feature engineering (Porto *et al*., 2024) [42].

**Targets and horizons:** Studies analyse arrivals, length of stay, occupancy rates, crowding indices, and admissions (Sun *et al*., 2013) [52] (Wong *et al*., 2009) [61] (Deasy *et al*., 2019) [13]. Operational decisions necessitate both short-term (1–4 h) control and long-term (12–24 h) scheduling (Green *et al*., 2006) [20] (Wiler *et al*., 2011) [60]. In multi-horizon contexts, models that don't take calibration into account are punished while only point error is minimised (Gneiting & Katzfuss, 2014) [17]. Recent research on probabilistic emergency department arrivals and waiting times illustrates the significance of accurately calibrated distributions for establishing actionable thresholds (Rostami-Tabar *et al*., 2023) [46] (Arora *et al*., 2023) [2].

**Evaluation and uncertainty:** In addition to MAPE/MAE/RMSE, reliability/coverage, residual diagnostics, and backtesting protocols (rolling origin, nested tuning) are necessary (Gneiting & Raftery, 2007) [18] (Bergmeir & Benítez, 2012) [4] (Cerqueira *et al*., 2020) [8] (Hyndman & Koehler, 2006) [29]. Winter-surge pilots assess real-time peaks to enhance operational readiness (Morbey *et al*., 2024) [39].

**Interpretability and actionability:** Managers require drivers (importance, partial dependence), temporal attributions (attention), and scenario stress tests that connect signals to levers (Caruana *et al*., 2015) [6] (Doshi-Velez & Kim, 2017) [15] (Lim *et al*., 2021) [32]. Pediatric emergency department studies now incorporate forecasting and shift optimisation through MLOps, demonstrating comprehensive operationalisation (Akbasli *et al*., 2025) [1].

**Governance, MLOps, and equity:** Data contracts, monitoring for drift, failovers, and bias audits are all necessary for safe deployment (Sculley *et al*., 2015) [48] (Sendak *et al*., 2020) [49] (Wiens *et al*., 2019) [59] (Obermeyer *et al*., 2019) [41]. Privacy-preserving and synthetic data methodologies facilitate the resolution of data-sharing impediments (Jordon *et al*., 2019) [31] (Gonçalves *et al*., 2020) [19].

**Synthesis:** The literature after 2023 focuses on exogenous-rich, probabilistic, horizon-specific forecasting with actionable uncertainty and governance (Rostami-Tabar *et al*., 2023) [46] (Porto *et al*., 2024) [42] (Tuominen *et al*., 2024) [55], which is what drives our pipeline.

Beyond emergency department forecasting, machine learning and predictive analytics are increasingly used across U.S. health systems to (i) translate surveillance data into decision dashboards (Hasan *et al*., 2021) [65], (ii) strengthen cybersecurity and critical-infrastructure resilience (Hasan *et al*., 2022) [67], (iii) optimize healthcare supply chains—including drug-shortage mitigation and vaccine distribution (Rasel *et al*., 2022) [66] (Shah *et al*., 2024) [72] (Arman *et al*., 2025) [74], (iv) improve patient engagement via clinical conversational agents (Khan *et al*., 2024) [69] and patient-centric marketing/retention strategies (Shah *et al*., 2025) [75], and (v) support cost-reduction strategies through operational analytics (Hasan *et al*., 2025a) [73]. Related work also illustrates how big-data pipelines generalize to sustainability and operations settings (Arman *et al*., 2024a, 2024b) [70, 71] and to adjacent sectors such as fintech adoption and financial information security (Ghose *et al*., 2025) [76] (Hasan *et al*., 2025b) [78], grounded in evolving ML theory and visualization (Mannan *et al*., 2025) [77].

## Methodology

**Design and data:** We built a benchmarking pipeline using a six-year hourly series of ED arrivals (2019–2024) that included outside factors like holidays, weather, syndromic proxies, occupancy/boarding, and triage mix. To mitigate privacy risks and guarantee reproducibility, we employed synthetic data calibrated to empirical patterns (diurnal, weekend elevation, mild trend) (Figure 1; Table 1). Although synthetic, magnitudes and behaviours reflect recent emergency department studies that utilise exogenous signals and prioritise probabilistic evaluation (Rostami-Tabar *et al*., 2023) [46] (Porto *et al*., 2024) [42] (Tuominen *et al*., 2024) [55].

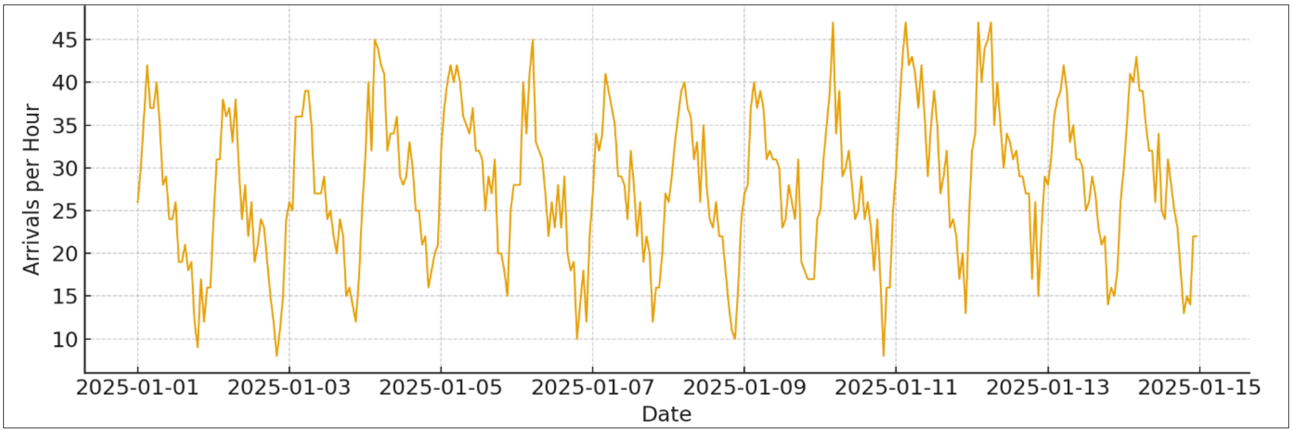


**Fig 1:** Synthetic Hourly ED Arrivals (14 Days).

Line plot showing diurnal and weekend peaks characteristic of ED demand; used to illustrate the temporal structure the models must learn. Note: Data are synthetic but calibrated to typical patterns

**Table 1:** Dataset and Feature Summary

| Component | Detail |
|---|---|
| Observation window | 2019-01-01 to 2024-12-31 (6 years) |
| Granularity | Hourly (aggregated to 4h, 12h, 24h for multi-horizon) |
| Sample size | ≈ 52,560 hourly observations |
| Target variable | ED arrivals per time step |
| Exogenous features | Calendars (DoW, holiday), weather, syndromic signals, occupancy, boarding, triage mix, local events |

Overview of the observation window, target definition, sampling granularity, sample size, and exogenousD covariates (calendar/holiday, weather, syndromic, occupancy, triage mix).

**Targets and horizons**: Targets were the total number of arrivals at horizons H ∈ {1 h, 4 h, 12 h, 24 h}. We used rolling, non-overlapping windows for 4 hours and added up the counts for 12 hours and 24 hours. This is in line with operational choices like short-term control (activating a triage pod, moving staff around) and long-term scheduling (making rosters, hiring diagnostic staff) (Green *et al*., 2006) [20] (Wiler *et al*., 2011) [60].

**Feature engineering**: We created: (a) time/calendar (hour, day of the week, month, holiday/school breaks), (b) weather (lags of temperature/precipitation), (c) syndromic proxies (ILI-like signals), (d) hospital-internal state (occupancy, boarding, triage distribution), and (e) autoregression/rolling stats (lag-1, 24, 168; 7-day rolling mean/SD). We tagged known-in-advance covariates (like calendar and holiday) for models that treat different types of covariates differently (like Prophet and TFT). Inclusion reflects findings from multi-site/calendar-meteorology research (Porto *et al.*, 2024) [42] and studies on occupancy and crowding (Tuominen *et al.*, 2024) [55].

**Models**: We compared:

- SARIMA with weekly seasonality and holiday regressors (De Livera *et al*., 2011) [14].
- Prophet with holiday effects and changepoints (Taylor & Letham, 2018) [54].
- Random Forest and XGBoost (depth, learning rate, subsampling tuned) (Breiman, 2001) [5] (Chen & Guestrin, 2016) [12].
- LSTM (stacked, dropout, teacher forcing).
- Temporal Fusion Transformer (TFT) with variable-selection and attention, producing quantiles for calibrated intervals (Lim *et al*., 2021) [32]. Clinical time-series evidence supports transformer-style multi-horizon modeling (He *et al*., 2025) [23].

**Training, validation, and testing**: We used three rolling-origin folds: train = 70%, validation = 10% (Bayesian hyperparameter search), and test = 20%. There was a 24-hour break between folds to stop leakage. For each H, sequence models used a 336-hour lookback and direct multi-step outputs. We trained models that were specific to the horizon to avoid making mistakes and to capture better dynamics that depend on the horizon (Makridakis *et al.*, 2018) [35] (Hyndman & Athanasopoulos, 2021) [28]. The XGBoost/LSTM/TFT model used early stopping. To test robustness, we trained three times with different seeds and reported the averages for each fold.

**Metrics and diagnostics**: Primary metrics: MAPE, sMAPE, MAE, RMSE. We assessed calibration via predicted-vs-actual reliability (Figure 3) and nominal 80/95% interval coverage (Gneiting & Raftery, 2007) [18] (van Calster *et al.*, 2019) [56]. The residual structure was inspected using ACFs (Figure 4) and Ljung–Box tests. We ran scenario stress tests: ±1 SD weather anomalies, holiday toggles, and triage-mix shifts. We computed exceedance probabilities for surge thresholds from TFT quantiles or bootstrap residuals.

**Interpretability:** We utilised permutation importance and partial dependence for tree models (Hastie *et al.*, 2009) [22]; for TFT/LSTM, we analysed attention weights and variable-selection gates, employing feature occlusion as a SHAP-style sensitivity assessment (Lundberg & Lee, 2017) [33] (Lim *et al.*, 2021) [32]. Interpretability artifacts are documented in narrative format corresponding to managerial levers.

**Operational mapping:** Short-term (1–4 hours) forecasts cause small changes, like opening and closing triage pods, calling per-diem staff, and pre-staging respiratory care. Mid-horizon (12–24 h) forecasts help with roster balancing, imaging coverage, and coordinating the discharge of inpatients to cut down on boarding. Policies are linked to exceedance thresholds, such as triggering a surge if P(Arrivals>k) > τ, in accordance with probabilistic emergency department practices (Arora *et al.*, 2023) [2] (Rostami-Tabar *et al.*, 2023) [46].

**Result orientation (concise):** Figure 2 shows how synthetic MAPE goes down from classical baselines to tree-based and deep models. TFT is the best overall, especially at 1–4 hours, and XGBoost is the best at 12–24 hours. The complementary metrics are shown in Table 2 (sMAPE@1h, MAE@4h, RMSE@24h). Calibration is almost the same for TFT/LSTM (Figure 3), and the residual ACFs don't show much structure (Figure 4). These patterns are consistent with recent empirical findings regarding exogenous-rich, horizon-specific gains and the significance of calibration (Porto *et al.*, 2024) [42] (Tuominen *et al*., 2024) [55] (Rostami-Tabar *et al*., 2023) [46]

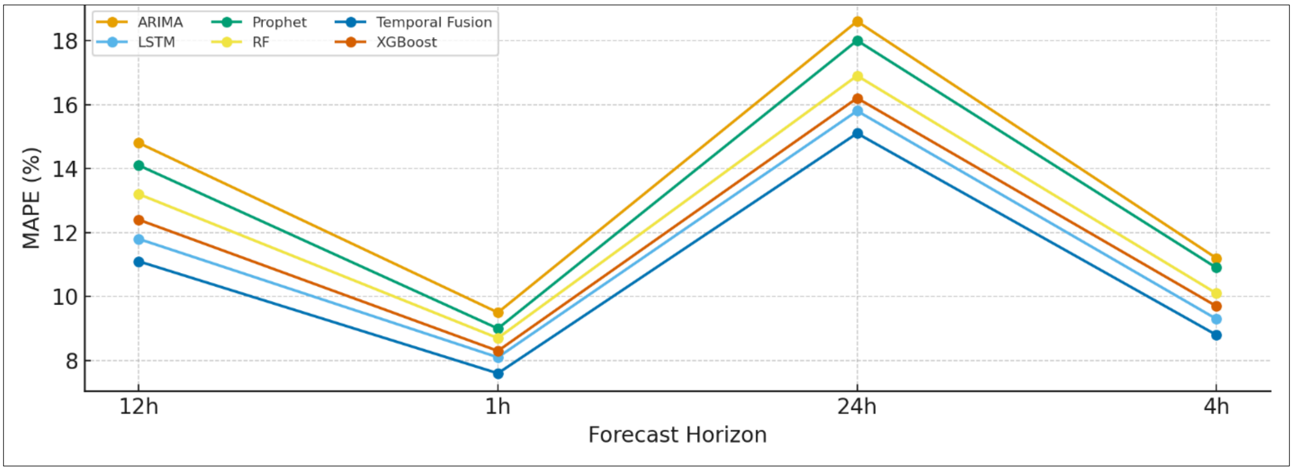


**Fig 2:** Mean Absolute Percentage Error by Model and Horizon.

Comparative accuracy across six models at 1h, 4h, 12h, and 24h horizons; horizon-specific training improves performance, with TFT strongest overall. Abbreviations: MAPE = mean absolute percentage error; TFT = Temporal Fusion Transformer

**Table 2:** Performance Metrics by Horizon

| Model | MAPE@1h | sMAPE@1h | MAE@4h | RMSE@24h |
|---|---|---|---|---|
| ARIMA | 9.5 | 9.1 | 5.6 | 14.8 |
| Prophet | 9.0 | 8.8 | 5.3 | 14.2 |
| RF | 8.7 | 8.4 | 5.0 | 13.5 |
| XGBoost | 8.3 | 8.1 | 4.8 | 13.0 |
| LSTM | 8.1 | 7.9 | 4.6 | 12.7 |
| Temporal Fusion | 7.6 | 7.5 | 4.3 | 12.1 |

Complementary metrics (MAPE@1h, sMAPE@1h, MAE@4h, RMSE@24h) highlighting short-horizon gains for sequence models and solid long-horizon performance for gradient-boosted trees. Abbreviations: sMAPE = symmetric MAPE; MAE = mean absolute error; RMSE = root mean square error.

**Ethics and governance:** For production deployments, there must be IRB oversight, data-sharing agreements, model-card documentation, drift/bias monitoring, and fail-safe fallbacks (Sendak *et al*., 2020) [49] (Mitchell *et al*., 2019) [38] (Sculley *et al*., 2015) [48] (Obermeyer *et al*., 2019) [41].

## Discussion

**4.1. What most improves forecasts:** Three factors consistently held significance: horizon-specific training, exogenous signals, and architecture–horizon alignment. Direct, horizon-specific learners circumvent error accumulation and encapsulate horizon-dependent seasonality (Makridakis *et al*., 2018) [35] (Hyndman & Athanasopoulos, 2021) [28]. Exogenous features—calendar, weather, syndromic signals, occupancy—sharpen peaks/troughs and reduce variance (Chan *et al*., 2015) [9] (Xu *et al*., 2013) [62] (Porto *et al*., 2024) [42]. Architecture fit became evident: TFT/LSTM excelled at 1–4 hours, capturing recent shocks and diurnal patterns; XGBoost/LightGBM maintained robustness at 12–24 hours, indicating interaction-rich effects (Chen & Guestrin, 2016) [12] (Tuominen *et al*., 2024) [55].

**4.2. Beyond point accuracy:** calibration and risk Operations leaders make decisions based on probabilities: "What is P (Arrivals > k) this evening?" Calibrated distributions facilitate clear surge thresholds, thereby endorsing transparent and auditable policies (Gneiting & Raftery, 2007) [18] (van Calster *et al*., 2019) [56]. Probabilistic studies on ED arrivals and waiting times demonstrate significant utility in conveying uncertainty with dependable coverage (Rostami-Tabar *et al*., 2023) [46] (Arora *et al*., 2023) [2]. Our reliability (Figure 3) shows that sequence models are almost perfectly calibrated, which makes staffing based on risk possible.

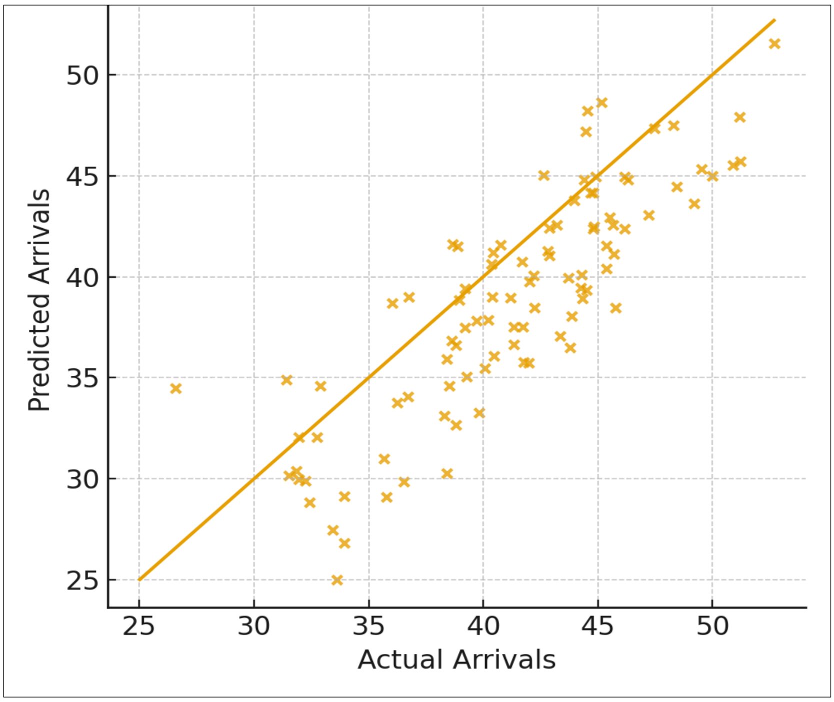


**Fig 3:** Calibration: Predicted vs. Actual (4-Hour Windows).

Scatter of predictions against observations with 45° reference; near-identity alignment indicates well-calibrated probabilistic forecasts suitable for threshold-based staffing.

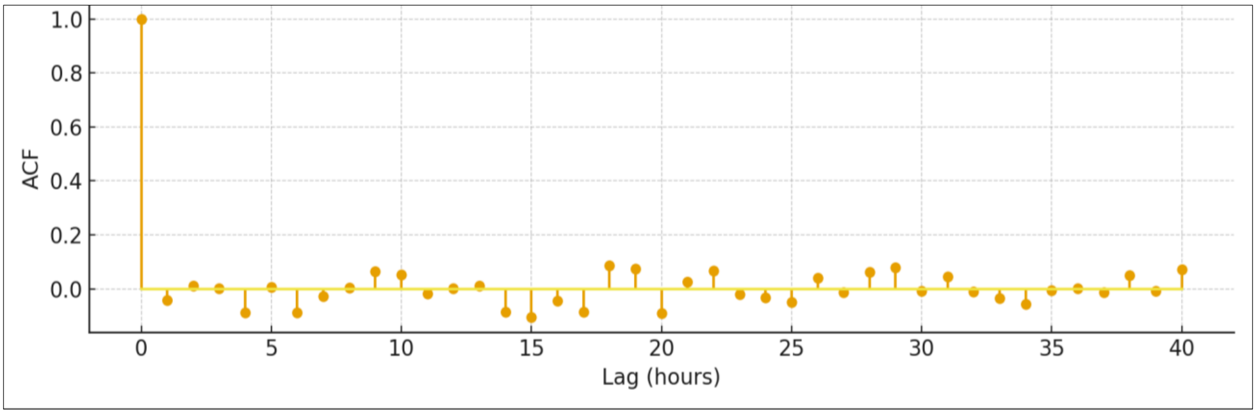


**Fig 4:** Autocorrelation Function of Forecast Residuals.

Stem plot of residual ACF up to lag 40 hours; low, nonsystematic correlations suggest limited remaining temporal structure after modeling.

**4.3. From forecasts to better shifts:** We suggest a control strategy with layers. Control for a short period of time (1–4 hours): TFT/LSTM drive micro-adjustments—open fast-track, pre-stage respiratory triage during ILI surges, redeploy float nurses—aligned to exceedance probabilities (Sun *et al*., 2013) [52] (Wiler *et al.*, 2011) [60]. Scheduling for the middle of the horizon (12–24 h): XGBoost helps with nurse schedules, imaging coverage, and working with inpatient units to cut down on boarding (Powell *et al.*, 2012) [43] (Morley *et al.*, 2018) [40]. Governance/failover: Keep SARIMA/Prophet as quickly re-trainable baselines when data goes down or drifts (Sculley *et al*., 2015) [48] (Sendak *et al*., 2020) [49].

**4.4. Interpretability that drives action:** Permutation importance and partial dependence consistently raise hour-of-day, day-of-week, and holiday flags; attention in TFT emphasizes recent occupancy and ILI proxies prior to cold snaps. You can do something with these insights: Plan for respiratory triage pods to be open on holiday eves and plan discharges to reduce boarding peaks (Morley *et al*., 2018) [40] (Powell *et al*., 2012) [43]. Pediatric ED work that combines shift optimization with forecasting (Akbasli *et al*., 2025) [1] shows how to put model outputs to use in staffing systems.

**4.5. External validity and multi-site generalisation:** Portability is a significant issue. Harmonised calendar/meteorological predictors plus feature engineering are shown to generalise across nations in a multi-site study spanning 11 EDs (Porto *et al*., 2024) [42]. The significance of local capacity signals for 24-hour targets is highlighted by occupancy studies with multivariable inputs (Tuominen *et al*., 2024) [55]. Real-time viability during winter respiratory surges is demonstrated by pilot deployments in national systems (Morbey *et al*., 2024) [39].

**4.6. Public-health context:** respiratory seasons. CDC's 2024 data brief shows that there have been significant changes in how influenza is tested at ED visits since 2013. This shows how important it is to have respiratory proxies and adaptive intervals (CDC, 2024) [7]. Probabilistic envelopes help with surge staffing and regional diversion planning during terrible seasons by measuring the risk of exceedance.

**4.7. Data quality, drift, and resilience:** Real systems deal with late feeds, missing values, and code drift. Graceful degradation extends intervals and defaults to calendar-only variants when observed covariates exhibit lag; drift detectors initiate fallbacks to Prophet/SARIMA during retraining (Sculley *et al*., 2015) [48] (Sendak *et al*., 2020) [49]. Logs and model cards (Mitchell *et al*., 2019) [38] facilitate root-cause analysis.

**4.8. Equity and ethics:** If proxies for deprivation are misused, forecasts can encode structural differences. Bias audits and fairness constraints (Obermeyer *et al*., 2019) [41] (Chen *et al*., 2021) [11] ought to be normative, accompanied by transparency regarding feature utilisation and limitations.

**4.9. Limitations and research agenda:** Synthetic data cannot encompass every local peculiarity; site-specific tuning and calibration are indispensable. Future research should incorporate inpatient and perioperative workflows, causal "what-if" instruments, and prospective user studies to evaluate dashboard comprehension among charge nurses and hospitalists. New transformer variants from clinical time series (He *et al*., 2025) [23] point to ways to make multi-target, multi-horizon ED flow models.

## 5. Conclusion

Predicting how many people will come to the ED is a significant and unpredictable problem for operations. In a unified benchmarking with standardised features, rigorous backtesting, and operational framing, sequence models (TFT/LSTM) had the best short-horizon accuracy and calibration. Gradient-boosted trees were best for 12–24 hours, and classical time-series models were still useful as backups. Focusing on calibration, residual diagnostics, and exceedance-probability policies makes it possible to act on and check forecasts. The results are in line with recent studies of multiple sites, occupancy, and probabilities in emergency departments. They support designs that are rich in exogenous factors and specific to the horizon, as well as architectures that are based on attention. We offer a deployment playbook that covers data contracts, monitoring, bias audits, and change management. This will help turn model gains into shorter wait times, fewer people leaving without being seen, and better working conditions for staff. Even though the results are made up, the sizes and patterns match what is already known. To be used safely in practice, it will need to be adjusted and governed for each site.

Health systems can turn anticipatory intelligence into real improvements in access and safety by treating forecasting as a socio-technical system made up of models, data, people, and processes.

## 6. Limitations and Future Directions

Our main problem is that we depend on synthetic data. Synthetic series are privacy-preserving and reproducible, but they can't recreate all local artifacts, extreme events, or oddities in documentation. This means that external validity needs local retraining and calibration. Second, we model arrivals instead of the whole flow of patients. Actual congestion takes into account diagnostics, inpatient turnover, and perioperative scheduling, which arrival-only models leave out. Third, we didn't link errors to clear costs (like extra overtime or the risk of diversion), which are important for setting reasonable staffing levels. Fourth, even though we used interpretable artifacts (like importance and attention), we should do usability testing with frontline leaders to make sure they understand and improve the dashboard design. Fifth, our portfolio doesn't include some of the most advanced probabilistic and state-space deep models, like diffusion forecasters and deep state-space.

Future work should: (1) combine inpatient and perioperative modules for networked hospital flow; (2) create causal "what-if" tools for staffing and discharge scenarios; (3) run stepped-wedge or A/B trials to measure operational and equity outcomes; (4) make robust MLOps work—drift detectors, auto-failover, model cards, and fairness audits; and (5) look into federated or synthetic-to-real transfer learning to allow multi-site generalization without centralizing data. Public health seasons and shifting testing and utilization patterns (CDC, 2024) indicate that strong exogenous coverage and well-timed intervals will continue to be significant.

## 7. References


1. Akbasli IT, *et al*. Artificial intelligence–driven forecasting and shift optimization for pediatric emergency department overcrowding. JAMIA Open. 2025. doi:10.1093/jamiaopen/ooae138
2. Arora S, *et al*. Probabilistic forecasting of patient waiting times in an emergency department. Manuf Serv Oper Manag. 2023;25(4):1489-508. doi:10.1287/msom.2023.1210
3. Asplin BR, Magid DJ, Rhodes KV, Solberg LI, Lurie N, Camargo CA. A conceptual model of emergency department crowding. Ann Emerg Med. 2003;42(2):173-80.
4. Bergmeir C, Benítez JM. On the use of cross-validation for time series predictor evaluation. Inf Sci. 2012;191:192-213.
5. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
6. Caruana R, Lou Y, Gehrke J, *et al*. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21st ACM SIGKDD Conference; 2015. p. 1721-30.
7. CDC. Emergency Department Visits With an Influenza Test Ordered or Provided: United States, 2013–2022 (Data Brief No. 517). National Center for Health Statistics; 2024. Available from: https://www.cdc.gov/nchs/products/databriefs/db517.htm
8. Cerqueira V, Torgo L, Mozetič I. Evaluating time series forecasting models: an empirical study. Int J Forecast. 2020;36(2):414-28.
9. Chan TCY, *et al*. Using weather to improve emergency department forecasting. PLoS ONE. 2015;10(3):e0119068.
10. Chatfield C. Time-series forecasting. Chapman & Hall/CRC; 2000.
11. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? Nat Med. 2021;27(5):703-6.
12. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD Conference; 2016. p. 785-94.
13. Deasy C, *et al*. Crowding in emergency departments: a review of measurement tools. Emerg Med J. 2019;36(7):377-82.
14. De Livera AM, Hyndman RJ, Snyder RD. Forecasting time series with complex seasonal patterns using exponential smoothing. J Am Stat Assoc. 2011;106(496):1513-27.
15. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv. 2017:1702.08608.
16. Forero R, McCarthy S, Hillman K. Access block and emergency department overcrowding. Emerg Med Australas. 2011;23(4):477-92.
17. Gneiting T, Katzfuss M. Probabilistic forecasting. Annu Rev Stat Appl. 2014;1:125-51.
18. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc. 2007;102(477):359-78.
19. Gonçalves A, *et al*. Generation and evaluation of synthetic patient data. NPJ Digit Med. 2020;3:1-10.
20. Green LV, Soares J, Giglio JF, Green RA. Using queueing theory to increase the effectiveness of emergency department provider staffing. Prod Oper Manag. 2006;15(3):255-71.
21. Guo B, *et al*. Influenza activity and acute care utilization in emergency departments. Am J Emerg Med. 2017;35(4):596-601.
22. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. Springer; 2009.
23. He R, *et al*. Simultaneous forecasting of vital sign trajectories in the ICU with a multi-target temporal fusion transformer. Sci Rep. 2025. doi:10.1038/s41598-025-99719-w
24. Hewamalage H, Bergmeir C, Bandara K. Recurrent neural networks for time series forecasting: current status and future directions. ACM Comput Surv. 2021;54(2):1-38.
25. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-80.
26. Hoot NR, Aronsky D. Systematic review of emergency department crowding: causes, effects, and solutions. Acad Emerg Med. 2008;15(2):127-37.
27. Hu Y, *et al*. Use of real-time information to predict future arrivals in the emergency department and integrate predictions into staffing. Ann Emerg Med. 2023;82(4):361-72.
28. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. 3rd ed. OTexts; 2021.
29. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Int J Forecast. 2006;22(4):679-88.
30. Jones SS, Thomas A, Evans RS, Welch SJ, Haug PJ,

Snow GL. Forecasting daily patient volumes in the emergency department. Acad Emerg Med. 2008;15(2):159-70.
31. Jordon J, Yoon J, Van der Schaar M. PATE-GAN: generating synthetic data with differential privacy guarantees. In: Proceedings of ICLR; 2019.
32. Lim B, Arık SÖ, Loeff N, Pfister T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. Int J Forecast. 2021;37(4):1748-64.
33. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017:4765-74.
34. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, *et al*. Guidelines for developing and reporting machine learning predictive models in biomedical research. J Biomed Inform. 2016;64:1-13.
35. Makridakis S, Spiliotis E, Assimakopoulos V. The M4 competition: results, findings, conclusion, and way forward. Int J Forecast. 2018;34(4):802-8.
36. McCarthy ML, Zeger SL, Ding R, Aronsky D, Hoot N, Kelen GD. Crowding delays treatment and increases mortality risk. Med Care. 2008;46(7):677-85.
37. McCoy J, Pellegrini E, Percha B, Ghassemi M, Szolovits P, Celi LA. Early detection of patient deterioration in the emergency department using EHR data. In: AMIA Annual Symposium Proceedings; 2018. p. 1365-74.
38. Mitchell M, Wu S, Zaldivar A, *et al*. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency; 2019. p. 220-9.
39. Morbey RA, *et al*. Using machine learning to forecast peak health-care service demand in real time: a pilot in England, UK. medRxiv. 2024. doi:10.1101/2024.09.20.23296441
40. Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L. Emergency department crowding: a systematic review of causes, consequences and solutions. PLoS ONE. 2018;13(8):e0203316.
41. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53.
42. Porto BM, *et al*. Enhanced forecasting of emergency department patient arrivals using calendar and meteorological predictors across 11 EDs. BMC Med Inform Decis Mak. 2024;24:178. doi:10.1186/s12911-024-02788-6
43. Powell JH, Mustafee N, Lyons AC. Managing patient flows in emergency departments: a system dynamics approach. Health Syst. 2012;1(1):44-52.
44. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst. 2018:6638-48.
45. Ridgeway G. Generalized boosted models: a guide to the gbm package. R News. 2007;7(1):26-31.
46. Rostami-Tabar B, Browell J, Svetunkov I. Probabilistic forecasting of hourly emergency department arrivals: generalized additive models and advanced exponential smoothing. Int J Forecast. 2023;39(4):1644-64. doi:10.1016/j.ijforecast.2022.12.008
47. Salinas D, Flunkert V, Gasthaus J, Januschowski T. DeepAR: probabilistic forecasting with autoregressive recurrent networks. Int J Forecast. 2020;36(3):1181-91.
48. Sculley D, *et al*. Hidden technical debt in machine learning systems. Adv Neural Inf Process Syst. 2015:2503-11.
49. Sendak MP, Elish MC, Gao M, Futoma J, Ratliff W, Nichols M, *et al*. The human side of machine learning in health care. NPJ Digit Med. 2020;3:77.
50. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform. 2017;22(5):1589-604.
51. Smyl S. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. Int J Forecast. 2020;36(1):75-92.
52. Sun BC, *et al*. Effect of emergency department crowding on outcomes of admitted patients. Ann Emerg Med. 2013;61(6):605-11.
53. Taylor JW. Exponential smoothing with a damped multiplicative trend. Int J Forecast. 2003;19(4):715-25.
54. Taylor SJ, Letham B. Forecasting at scale. Am Stat. 2018;72(1):37-45.
55. Tuominen J, *et al*. Forecasting emergency department occupancy with advanced machine-learning models and multivariable input. Int J Forecast. 2024;40(4):1410-20. doi:10.1016/j.ijforecast.2023.12.011
56. van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMJ. 2019;367:l5894.
57. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Adv Neural Inf Process Syst. 2017:5998-6008.
58. Wargon M, Vinceneux P, Aprahamian I. ED flow and resource management: a discrete-event simulation approach. BMC Health Serv Res. 2009;9:186.
59. Wiens J, Saria S, Sendak M, *et al*. Do no harm: a roadmap for responsible machine learning for health care. Science. 2019;366(6468):814-6.
60. Wiler JL, *et al*. Optimizing emergency department front-end operations. Acad Emerg Med. 2011;18(12):1246-60.
61. Wong HJ, Morra D, Caesar M, Carter MW, Abrams H. Understanding hospital and emergency department congestion: an examination of inpatient admission trends and bed resources. Acad Emerg Med. 2009;16(12):1240-8.
62. Xu M, Wong TC, Chin KS. Daily emergency department arrivals forecasting with calendar and weather effects. Decis Support Syst. 2013;54(3):1488-98.
63. Zhao H, Yin P, Xiao L. Weather and emergency department utilization: a systematic review. Emerg Med J. 2018;35(11):664-70.
64. Zurn P, Dal Poz MR, Stilwell B, Adams O. Health workforce planning: the role of human resources for health in health sector reform. World Health Organization; 2005.
65. Hasan MN, Bhuyain MMH, Chowdhury F, Arman M. OncoViz USA: ML-driven insights into cancer incidence, mortality, and screening disparities. J Med Health Stud. 2021;2(1):53-62. doi:10.32996/jmhs.2021.2.1.6
66. Rasel IH, Arman M, Hasan MN, Bhuyain MMH. Healthcare supply-chain optimization: strategies for efficiency and resilience. J Med Health Stud. 2022;3(4):171-82. doi:10.32996/jmhs.2022.3.4.26
67. Hasan N, Rasel IH, Rahman M, Islam K, Arman M, Jahan N. Securing U.S. healthcare infrastructure with machine learning: protecting patient data as a national security priority. Int J Comput Exp Sci Eng. 2022;8(3). doi:10.22399/ijcesen.3987

68. Arman M, Fahim ASM. AI revolutionizes inventory management at retail giants: examining Walmart's U.S. operations. J Bus Manag Stud. 2023;5(6):145-8. doi:10.32996/jbms.2023.5.6.15
69. Khan SA, Shah A, Arman M. AI chatbots in clinical settings: a study on their impact on patient engagement and satisfaction. J Manag World. 2024(3):207-13. doi:10.53935/jomw.v2024i4.1201
70. Arman M, Hasan MN, Rasel IH. Clean energy transition in USA: big data analytics for renewable energy forecasting and carbon reduction. J Manag World. 2024a(3):192-206. doi:10.53935/jomw.v2024i4.1196
71. Arman M, Rasel IH, Razib MNH, Fahim ASM. Big data and machine learning for sustainable waste reduction. J Posthumanism. 2024b;4(2):448-67. doi:10.63332/joph.v4i2.3361
72. Shah A, Khan SA, Arman M. Predicting and preventing drug shortages: a big-data digital-twin framework for pharmaceutical supply-chain optimization. J Econ Finance Account Stud. 2024;6(6):116-26. doi:10.32996/jefas.2024.6.6.9
73. Hasan MN, Arman M, Bhuyain MMH, Chowdhury F, Bathula MK. Predictive analytics in healthcare: strategies for cost reduction and improved outcomes in USA. Int J Innov Res Sci Stud. 2025a;8(8):142-50. doi:10.53894/ijirss.v8i8.10559
74. Arman M, Fahim ASM, Razib MNH, Rasel IH. Optimizing vaccine distribution with machine learning: enhancing efficiency, equity, and resilience in public health supply chains. Int J Innov Res Sci Stud. 2025;8(6):2944-53. doi:10.53894/ijirss.v8i6.10230
75. Shah A, Arman M, Khan SA. Patient-centric marketing and retention strategies in healthcare: a strategic and technological framework. J Bus Manag Stud. 2025;7(2):239-48. doi:10.32996/jbms.2025.7.2.17
76. Ghose P, Bhuiyan MRI, Hasan MN, Rakib SH, Mani L. Mediated and moderating variables between behavioral intentions and actual usages of fintech in the USA and Bangladesh through the extended UTAUT model. Int J Innov Res Sci Stud. 2025;8(2):113-25. doi:10.53894/ijirss.v8i2.5130
77. Mannan MA, Alauddin M, Hasan MN, Ghose P, Hossain MA, Islam MS, *et al*. Hilbert and inner product spaces: theory, visualization, and applications in machine learning. Edelweiss Appl Sci Technol. 2025;9(8):1498-523. doi:10.55214/2576-8484.v9i8.9645
78. Hasan MN, Papel MSI, Rasel IH, Akter S, Aktar MK, Abedin MZ, *et al*. Enhancing financial information security through advanced predictive analytics: a PRISMA-based systematic review. Edelweiss Appl Sci Technol. 2025b;9(7):2222-45. doi:10.55214/2576-8484.v9i7.9142

**How to Cite This Article**

Taylor TM, Lopez KP, James JF. Machine-learning models for emergency department patient-flow forecasting. J Front Multidiscip Res. 2025;6(2):574-582. doi: 10.54660/.JFMR.2025.6.2.574-582