PREDICTIVE MODELLING OF EMERGENCY DEPARTMENT PATIENT ARRIVALS USING MACHINE LEARNING AND EXPLAINABLE ARTIFICIAL INTELLIGENCE

Daad M. Alhassan, Hala M. Alsuabeyl, Rana A. Almashari, Rand A. Altareefi, Layan A. AlRushaid and Fahima Hajjej

Information Systems department, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University (PNU), Saudi Arabia

ABSTRACT

Emergency department (ED) overcrowding is a persistent operational challenge that degrades patient care quality, lengthens waiting times, and strains hospital resources. Accurate forecasting of patient arrivals enables proactive staffing, bed allocation, and capacity planning. This paper presents a comparative study of machine learning and deep learning models for forecasting daily ED visit volume, developed within a modified CRISP-DM framework. Using de-identified records from the MIMIC-IV-ED (v2.2) database enriched with historical weather observations and calendar features, we engineer temporal predictors and evaluate five models: a Long Short-Term Memory (LSTM) network, a Random Forest Regressor (RFR), Extreme Gradient Boosting (XGBoost), Neural Basis Expansion Analysis for Time Series (N-BEATS), and the Temporal Fusion Transformer (TFT). Models are assessed with Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) on a chronologically held-out test set. The TFT substantially outperformed all baselines, achieving an MAE of 4.46, an RMSE of 5.02, and a MAPE of 41.5%, compared with MAE values of 12.8-14.2 for the remaining models. To promote transparency, SHapley Additive exPlanations (SHAP), Partial Dependence Plots (PDP), and the TFT attention weights are used to interpret model behaviour. The best model is embedded in an interactive forecasting dashboard, "Marsad," that produces daily and short-horizon hourly forecasts to support proactive ED operations in alignment with Saudi Vision 2030.

KEYWORDS

deep learning, Emergency department forecasting,explainable AI,healthcare operations,machine learning, temporal fusion transformer, time-series prediction.

1. INTRODUCTION

Emergency departments (EDs) are a cornerstone of timely acute care, yet they are repeatedly destabilised by unpredictable surges in patient arrivals. When demand exceeds available capacity, the result is overcrowding longer waiting times, boarding of admitted patients, increased risk of adverse outcomes, and elevated staff burnout. The phenomenon is global and is driven by population growth, the rising prevalence of chronic disease, and external factors such as weather, seasonal illness, and national events. In the Kingdom of Saudi Arabia, the Vision 2030 Health Sector Transformation Program explicitly promotes data-driven solutions that improve service quality and reduce waiting times [1], [2], [3]. Predictive analytics for patient flow aligns naturally

David C. Wyld et al. (Eds): AIFU, BIoT, NLCAI, BDML, CCSEA, SIPRO, NCOM, CLOUD, – 2026
pp. 29-45, 2026. CS & IT - CSCP 2026 DOI: 10.5121/csit.2026.1601203

with these goals and with the United Nations Sustainable Development Goal 3 (Good Health and Well-Being) by enabling more efficient and equitable use of clinical resources. Despite the maturity of forecasting research, many hospitals continue to manage patient flow reactively, responding to surges only after congestion has occurred rather than anticipating it. This reactive posture is most costly precisely when demand is most variable - during weekends, public holidays, and periods of extreme weather - when ED volumes can spike beyond planned capacity. A forecasting system that anticipates daily ED demand from temporal, meteorological, and calendar signals would allow administrators to plan staffing and bed capacity in advance, closing a critical operational gap.

The objective of this work is to design, train, and compare a set of forecasting models for daily ED visit volume, to identify the most accurate and operationally useful approach, and to deliver the result as an interpretable tool that hospital decision-makers can trust. We frame the task as a univariate-dominant, multivariate-capable time-series regression problem and follow a structured data-science methodology. The principal contributions of this paper are as follows:

(1) A reproducible data-integration pipeline that merges four MIMIC-IV-ED tables with historical weather and calendar features and aggregates them to a daily forecasting target.
(2) A controlled empirical comparison of five forecasting models - LSTM, Random Forest, XGBoost, N-BEATS, and the Temporal Fusion Transformer - under an identical chronological train/test protocol and a common set of error metrics.
(3) An explainability layer combining SHAP, Partial Dependence Plots, and TFT attention weights that exposes the drivers of each forecast and supports clinical trust.
(4) A deployed, interactive dashboard ("Marsad") that converts model outputs into daily and short-horizon hourly forecasts with uncertainty and interpretation, demonstrating end-to-end operational feasibility.

Beyond accuracy, the value of such a system is operational. A forecast that anticipates a high-demand day even a few shifts in advance allows managers to adjust rosters, pre-position beds, and brief staff, converting a reactive scramble into a planned response. The same forecasts, aggregated over weeks, support longer-range scheduling and budgeting. By pairing predictive accuracy with transparent explanations and a usable interface, this work aims to make forecasting an everyday decision aid rather than an offline analytical exercise.

The remainder of the paper is organised as follows. Section II reviews related work in ED forecasting and the underlying modelling techniques. Section III describes the data, preprocessing, models, and experimental design. Section IV reports the empirical results, and Section V discusses their interpretation and implications. Section VI presents the deployment system, Section VII details limitations and ethical considerations, and Section VIII concludes with directions for future work.

2. RELATED WORK

Time-series forecasting exploits historical, time-ordered observations to predict future values, capturing trends, seasonality, and stochastic fluctuation. Incorporating exogenous variables such as weather, holidays, and special events can further improve accuracy. Modelling strategies for ED demand span a spectrum from classical statistical methods such as ARIMA and ARIMAX to modern machine learning and deep learning architectures, the latter of which have repeatedly improved forecasting accuracy in healthcare applications [4], [5].

2.1. Theoretical Foundations

ED forecasting draws on a layered set of computational ideas. At the base are artificial neural networks (ANNs), composed of interconnected neurons that apply weighted inputs, bias terms, and non-linear activations such as ReLU, sigmoid, and tanh, and are trained by forward propagation and backpropagation [6], [7]. Fully connected deep neural networks (FCDNNs) stack many such layers so that earlier layers capture simple fluctuations and deeper layers model higher-order non-linear patterns; their representational power is balanced against a tendency to overfit, which is mitigated through dropout and weight decay [8]. The Extreme Learning Machine (ELM) offers a fast alternative for single-hidden-layer networks by fixing random input weights and solving only the output weights in closed form, trading some flexibility for training speed [9].

For sequential data, recurrent neural networks (RNNs) maintain a hidden state that carries information across time steps, which suits the daily and weekly rhythms of ED arrivals. Plain RNNs, however, struggle with long-range dependencies because of vanishing and exploding gradients during backpropagation through time, a limitation that motivated the gated LSTM architecture [10]. Beyond single models, ensemble forecasting - combining predictions through averaging, stacking, or Bayesian model averaging - can improve robustness during regime changes and sudden demand shocks, and hybrid schemes that pair classical methods such as ARIMA with deep models can capture linear trend and complex seasonality simultaneously [5], [11], [12].

2.2. Machine Learning and Deep Learning for ED Forecasting

Tree-based ensembles remain strong baselines for tabular healthcare data. Random Forest [13] aggregates many decision trees trained on bootstrap samples, reducing variance and capturing non-linear interactions, while XGBoost [14] builds trees sequentially so that each corrects the residual errors of its predecessors. Both externalise temporal structure through engineered lag and rolling features. Deep architectures, by contrast, model temporal dependencies intrinsically. Recurrent neural networks, and in particular the Long Short-Term Memory (LSTM) network [10], use gated memory to mitigate vanishing gradients and to learn short- and long-range patterns. Purpose-built forecasting networks such as N-BEATS [15] decompose a signal into interpretable trend and seasonal components through stacked residual blocks, and the Temporal Fusion Transformer (TFT) [16] fuses recurrent encoding with multi-head attention and variable-selection networks to deliver interpretable, probabilistic, multi-horizon forecasts.

2.3. Empirical Studies of ED Demand Prediction

Several studies have established the value of feature-rich machine learning for ED demand. Tuominen et al. [17] forecast daily arrivals at Tampere University Hospital using 386,579 visits (2015-2019) and 158 explanatory variables, combining feature selection with adaptive models and statistical benchmarks; ARIMAX augmented with weather variables achieved the best accuracy (MAPE 6.6%), with calendar variables and hospital capacity emerging as the most influential predictors. Porto and Fogliatto [4] evaluated six machine learning algorithms across eleven EDs in three countries and found that XGBoost consistently outperformed alternatives, with MAPE between 5.0% and 14.1%, while noting performance degradation over longer horizons and heterogeneity across sites. Other studies target related prediction tasks. Safaripour and Lim [18] compared classifiers on Korean Health Panel data (2008-2015) to identify frequent ED users, with Random Forest achieving the strongest performance (AUC = 0.96), although the task was classification of frequent users rather than volume forecasting.

Lee et al. [19] reviewed fourteen machine learning studies predicting unscheduled return visits, finding that XGBoost frequently led (median AUC approximately 0.73-0.76) and that most models struggled to surpass AUC = 0.75, motivating the integration of external drivers.

Pelaez-Rodriguez et al. [20] forecast ED visits up to seven days ahead at two Spanish hospitals; neural networks (extreme learning machines and fully connected deep networks) outperformed other models, continuous retraining reduced MAE by 8-20%, and day-of-week, recent visit history, and weather were the most influential predictors. More recently, Neshat et al. [21] introduced an explainable meta-learning gradient-boosting approach and quantified the impact of exogenous variables, while Vural et al. [22] applied deep learning to forecast boarding patient counts and Brossard et al. [23] demonstrated a machine-learning proof of concept for predicting ED admissions. Complementing these, Tuominen et al. [24] showed that multivariable inputs improve the forecasting of ED crowding when paired with advanced machine learning models.

2.4. Explainability in Forecasting

Because ED forecasts inform staffing and resource allocation, interpretability is a practical requirement rather than a luxury. SHapley Additive exPlanations (SHAP) [25], grounded in cooperative game theory, attribute each prediction to its input features through global summary plots and local explanations. Partial Dependence Plots (PDP) [26] visualise how a prediction changes as a single feature varies while others are held constant, revealing non-linear thresholds. Together with the intrinsic attention weights of the TFT, these techniques expose the drivers of demand, support bias detection, and enable what-if analysis for administrators.

2.5. Research Gap

The reviewed literature confirms that machine learning improves over classical baselines, that feature engineering and exogenous variables matter, and that performance can degrade across horizons and heterogeneous sites. However, comparatively few studies place modern attention-based forecasters such as the TFT side by side with strong tree-based and recurrent baselines under a single, leakage-free protocol, and fewer still pair such a comparison with an interpretable, deployable tool intended for real hospital operations. This work addresses that gap. Table I summarises representative studies and situates the present contribution among them.

TABLE I. Summary of Representative ED Forecasting Studies

Study	**Models**	**Best result**	**Key insight**
Tuominen et al. [17]	ARIMA/ARIMAX, RF, adaptive	MAPE 6.6% (ARIMAX)	Calendar & capacity dominate
Porto &Fogliatto [4]	6 ML algorithms	MAPE 5.0-14.1% (XGBoost)	Feature engineering helps; horizon decay
Safaripour & Lim [18]	RF, LR, others	AUC 0.96 (RF)	Classifies frequent users
Lee et al. [19]	XGBoost, RF (review)	AUC approx. 0.73-0.76	External drivers needed
Pelaez-Rodr. et al.[20]	ELM, FCDNN, others	MAE 8.6-10.1	Continuous retraining helps
This work	LSTM, RF, XGBoost, N-BEATS, TFT	MAE 4.46 (TFT)	Attention model + XAI + dashboard

3. MATERIALS AND METHODS

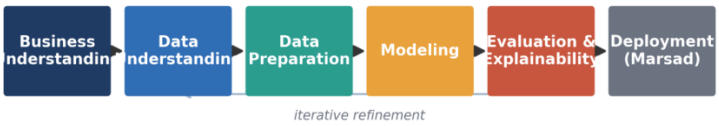

Fig. 1. Modified CRISP-DM methodology adopted for ED visit forecasting.

The project follows a modified CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology tailored to ED forecasting, comprising business understanding, data understanding, data preparation, modelling, evaluation with explainability, and deployment. Fig. 1 illustrates the pipeline, including the iterative refinement that links evaluation back to earlier phases.

3.1. Problem Formulation and Business Understanding

The forecasting task is defined over a univariate target series of daily ED visit counts, optionally conditioned on exogenous covariates. Let y(t) denote the number of ED visits on day t, and let x(t) denote the vector of contextual features for that day (calendar and weather variables, together with engineered lag and rolling terms). Given a history of length L, the objective is to learn a function f that predicts the next H daily values:

$$\hat{y}_{t+1:t+H} = f(y_{t-L+1:t}, x_{t-L+1:t+H}) \quad (1)$$

with parameters chosen to minimise forecast error on unseen future days. For the primary comparison reported here, the emphasis is on the next-day and short-horizon daily forecast, which is the quantity most relevant to shift planning and staffing. From a business perspective, the requirement is not only low error but also timeliness and interpretability: forecasts must be available early enough to act upon, and decision-makers must be able to understand why a given level of demand is expected. These requirements, derived from the operational context of ED management and the goals of the Vision 2030 Health Sector Transformation Program, informed both the choice of models and the design of the deployed system.

3.2. Data Sources and Integration

The primary data source is the MIMIC-IV-ED database, version 2.2 [27], a large, de-identified collection of emergency department records released through PhysioNet under a credentialed data-use agreement. Four core tables - edstays, triage, vitalsign, and diagnosis - were extracted and merged on the shared key stay_id so that each ED visit is represented as a single record containing demographic, clinical, and diagnostic information, including visit timestamps, triage vital signs, and ICD-coded diagnoses.

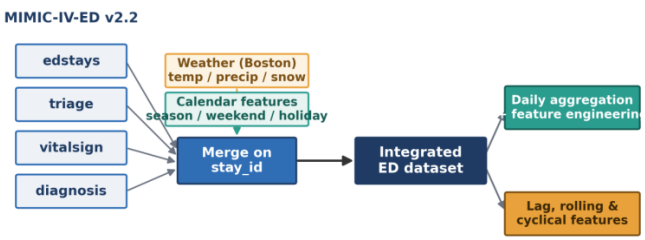

Fig. 2. Integration of MIMIC-IV-ED tables with weather and calendar features, followed by daily aggregation and feature engineering.

To capture external drivers of demand, the visit records were enriched with two additional sources. Average monthly weather data for Boston, Massachusetts - temperature, precipitation, and snowfall - were obtained from a historical weather archive [28] and mapped to each record by the month of admission. Calendar-based and seasonal features were then engineered with Python's datetime and holidays libraries to encode temporal effects, yielding the following variables: season (Winter, Spring, Summer, Fall); is_weekend; is_holiday (U.S. federal holidays); and the monthly averages avg_weather_C, avg_precip, and avg_snow. Fig. 2 summarises the integration of clinical, environmental, and calendar sources into a single dataset that is subsequently aggregated to a daily forecasting target.

Table II describes the principal fields available after integration. Patient-level identifiers and clinical measurements support quality checks and exploratory analysis, while the environmental and calendar fields provide the exogenous context used during modelling.

TABLE II. Selected Fields of the Integrated ED Dataset

Feature	Type	Description
intime / outtime	Datetime	Timestamps of ED entry and discharge/transfer.
gender, race	Categorical	Patient demographics.
arrival_transport	Categorical	Arrival mode (e.g., walk-in, ambulance).
disposition	Categorical	Visit outcome (admitted, home, etc.).
acuity	Numerical	Triage acuity level (1 = most severe, 5 = least).
temperature, heartrate	Numerical	Triage vital signs.
o2sat, sbp, dbp, pain	Numerical	Oxygen saturation, blood pressure, pain score.
icd_title	Text	Diagnosis description (ICD-9/ICD-10).
avg_weather_C	Numerical	Average monthly temperature (degrees C), Boston.
avg_precip, avg_snow	Numerical	Average monthly precipitation and snowfall.
season, is_weekend, is_holiday	Engineered	Seasonal and calendar indicators.

3.3. Data Quality and Exploratory Analysis

Data quality was verified along several dimensions: completeness (missing-value assessment), consistency and validity (range and type checks on vital signs and timestamps), duplicate detection, outlier detection, and integrity of relationships across the merged tables. Exploratory data analysis was performed both before and after preprocessing using univariate, bivariate, and multivariate techniques - bar charts, histograms, box plots, line charts, scatter plots, and a correlation matrix. The analysis revealed clear weekly and seasonal regularities in daily ED demand, with an average of approximately 94 visits per day, but only weak linear correlation between the weather variables and visit counts. This early finding - that exogenous meteorological signals carry limited predictive information for this dataset - shaped the interpretation of later modelling results.

Several operational patterns emerged from the analysis. Demand exhibited strong day-of-week structure, with systematic differences between weekdays and weekends, and a milder seasonal component across the year. Holidays were associated with shifts in volume relative to comparable ordinary days, and the distribution of triage acuity levels and arrival-transport modes confirmed that the series reflects a realistic mix of urgent and non-urgent presentations. The strong temporal regularity, combined with the weak influence of weather, suggested early on that recent history would be the dominant predictor and that models able to exploit autocorrelation would have an advantage.

3.4. Preprocessing and Feature Engineering

After integration, records were cleaned and standardised: categorical fields were normalised and encoded, implausible vital-sign values were treated as missing, and the visit-level data were aggregated to a daily series of total ED visits. All numerical inputs were scaled with Min-Max normalisation, which is particularly important for the scale-sensitive neural network models. To expose temporal structure to models that do not natively process sequences, three families of features were engineered: lag features (recent historical visit counts), rolling statistics (moving averages capturing local trend), and differencing terms. In addition, cyclical calendar variables such as day-of-week and month were encoded with sine and cosine transformations so that their periodic continuity is preserved - for example, so that December and January are represented as adjacent rather than maximally distant. These cyclical encodings were supplied to all models to provide a consistent representation of recurring temporal patterns.

3.5. Forecasting Models

Five models spanning the methodological spectrum from efficient ensembles to expressive deep architectures were implemented and compared. Each is summarised below; Table III contrasts their advantages, limitations, and overall assessment.

1) Random Forest Regressor (RFR).

RFR [13] constructs many decision trees on bootstrap samples and averages their outputs, with a random feature subset at each split to diversify the ensemble and improve generalisation. It captures non-linear relationships among temporal, environmental, and historical variables with minimal preprocessing. Temporal dependence is supplied externally through lag variables and rolling statistics, and feature-importance scores aid interpretability, making RFR a stable baseline.

2) XGBoost.

XGBoost [14] builds trees sequentially, each correcting the residual errors of its predecessors through gradient-based optimisation with L1/L2 regularisation and stochastic controls to limit overfitting. It is well suited to heterogeneous tabular data with non-linear interactions among temperature, day-of-week, and historical counts. As with RFR, temporal structure is externalised through engineered features, and SHAP provides post-hoc explanations.

3) Long Short-Term Memory (LSTM).

The LSTM network [10] introduces explicit temporal reasoning through input, forget, and output gates that regulate how information is stored, updated, and exposed across time steps, addressing the vanishing-gradient limitation of plain recurrent networks. A typical configuration ingests sequences of historical ED counts with exogenous covariates, processes them through stacked recurrent layers, and projects to a dense output. Critical hyperparameters include units per layer, look-back window, dropout, learning rate, batch size, and early-stopping patience.

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big) \quad \textbf{(2)}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \textbf{(3)}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \textbf{(4)}$$

$$h_t = o_t \cdot \tanh(C_t) \quad \textbf{(5)}$$

Here f(t), i(t), and o(t) denote the forget, input, and output gates, C(t) is the cell state, and h(t) is the hidden state passed to the next step. The forget gate determines which parts of the previous cell state are retained, the input gate controls how much new information is written, and the output gate regulates how much of the updated state is exposed, enabling the network to balance recent fluctuations against longer-range structure.

4) N-BEATS.

N-BEATS [15] is a deep architecture for interpretable time-series forecasting composed of stacked residual blocks. Each block produces a backcast that reconstructs the explained part of the input and a forecast that contributes to the prediction, with the residual passed to the next block. Projecting hidden representations onto polynomial or Fourier bases yields an explicit decomposition into trend and seasonal components without manual feature crafting. Given a look-back window x of length L, each block computes a residual r = x - x-hat, and the final forecast aggregates block- and stack-level outputs while training minimises a joint loss over backcast and forecast.

5) Temporal Fusion Transformer (TFT).

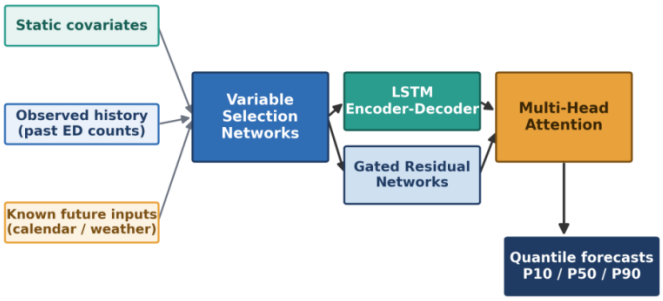

Fig. 3. Temporal Fusion Transformer architecture: variable selection, GRNs, LSTM encoder-decoder, and multi-head attention producing quantile forecasts.

The TFT [16] is a state-of-the-art architecture for multivariate, multi-horizon forecasting that fuses a recurrent encoder with transformer-style attention. It jointly processes static descriptors, known future inputs, and observed historical variables through variable-selection networks and gated residual networks (GRNs), passes them through a sequence-to-sequence LSTM encoder-decoder, and applies multi-head attention to identify the most influential historical time steps. A quantile-regression output produces calibrated prediction intervals at the 10th, 50th, and 90th percentiles, providing uncertainty estimates that are valuable for staffing decisions. The attention weights and variable-selection scores are intrinsically interpretable. Fig. 3 depicts the architecture; these properties motivated the selection of TFT as the primary model.

TABLE III. Comparative Summary of the Forecasting Models

Model	**Type**	**Key advantages**	**Limitations**
XGBoost	Boosted trees	Fast, scalable; non-linear; SHAP-interpretable	No native sequence model; manual lag features
Random Forest	Bagged trees	Robust to overfitting; feature importance	No temporal modelling; manual features
LSTM	Recurrent NN	Learns temporal dependencies directly	Costly; tuning-sensitive; limited parallelism

N-BEATS	Deep residual	Interpretable trend/seasonal decomposition	Needs large, clearly periodic data
TFT	LSTM + attention	SOTA multi-horizon;probabilistic;interpretable	High compute; data- and tuning-intensive

3.6. Experimental Setup

Models were implemented in Python within Jupyter environments. NumPy and Pandas supported numerical operations and time-series handling, scikit-learn provided evaluation utilities and the tree-based models, and the deep models used PyTorch Forecasting with PyTorch Lightning, which provide ready implementations of advanced time-series models and manage the training loop, validation, and hardware utilisation. The dataset was divided chronologically, with the first 80% of the series used for training and the most recent 20% reserved for testing. A chronological split preserves temporal order and prevents leakage, since a random split could allow a model to learn indirectly from future observations. Early stopping was applied to the deep models to reduce overfitting, and the cyclical sine/cosine features described above were supplied across all models to encode periodicity consistently.

Each model was configured according to its characteristics. The tree-based models were tuned over the number of estimators, maximum depth, and - for XGBoost - the learning rate and regularisation strength, with lag and rolling features supplying temporal context. The LSTM was configured through the number of units per layer, the look-back window length, dropout, learning rate, batch size, and early-stopping patience. N-BEATS was trained with stacked residual blocks using trend and seasonality basis expansions over a fixed look-back window. The TFT was trained with PyTorch Forecasting using a quantile loss over the 10th, 50th, and 90th percentiles, with variable-selection networks and gated residual networks active throughout and the encoder length set to span the dominant weekly cycle. To support fair comparison, all models shared the same training and test partitions and the same engineered feature set wherever architecturally applicable.

3.7. Evaluation Metrics

Three complementary error metrics were used. The Mean Absolute Error (MAE) reports the average magnitude of errors in the original units (daily visits) and is robust to outliers. The Root Mean Squared Error (RMSE) penalises large deviations more heavily and is therefore sensitive to occasional large misses. The Mean Absolute Percentage Error (MAPE) expresses error as a percentage of the actual value, enabling scale-free comparison but becoming unstable when actual values are small. Formally, for actual values y and predictions y-hat over n test points[29]:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{8}$$

Because daily ED counts are occasionally low, MAPE can be inflated by a small number of low-volume days; MAE and RMSE therefore provide a more reliable assessment for this dataset, a point revisited in the discussion.

3.8. Explainability Methods

To make the forecasts transparent, three interpretability techniques were applied. SHAP [25] attributes each prediction to its input features, revealing global drivers such as recent visit history and day-of-week through summary plots and individual predictions through local explanations. Partial Dependence Plots [26] expose how the predicted demand responds to a single feature in isolation, surfacing non-linear thresholds. Finally, the TFT's attention weights and variable-selection scores indicate which historical time steps and which inputs the model relied upon, allowing clinicians to verify that meaningful predictors receive appropriate importance and aiding the detection of spurious or proxy signals.

4. RESULTS

All five models were trained and evaluated on the identical chronological test set. This section reports each model's performance and then compares them directly.

4.1. Per-Model Performance

The LSTM achieved an MAE of 13.82, an RMSE of 19.46, and a MAPE of 210.6%. The relatively high RMSE indicates occasional large errors, and the model proved sensitive to hyperparameters, suggesting that it required more extensive tuning and data to stabilise. The Random Forest Regressor produced an MAE of 14.19, an RMSE of 20.11, and a MAPE of 210.2%; its reliance on engineered features rather than native sequence modelling limited its ability to capture temporal dependence. XGBoost performed comparably with an MAE of 13.44, an RMSE of 19.79, and a MAPE of 185.1%, offering stable predictions with an average error of about 13 visits against a daily mean near 94. N-BEATS recorded an MAE of 12.78, an RMSE of 19.83, and a MAPE of 150.3%; despite being purpose-built for time series, it did not surpass the simpler baselines, consistent with the weak periodic signal and limited dataset size.

The Temporal Fusion Transformer markedly outperformed every baseline, achieving an MAE of 4.46, an RMSE of 5.02, and a MAPE of 41.5%. An average error of four to five visits per day is highly acceptable for hospital resource planning, enabling more precise estimation of demand and supporting the allocation of staff, beds, and operational resources. The low RMSE indicates that large misses were rare and that the model remained stable even during periods of fluctuation, although it still showed minor difficulty in fully capturing extreme spikes - a limitation attributable to the weak predictive strength of the available exogenous features.

4.2. Comparative Analysis

Table IV consolidates the results, and Figs. 4 and 5 visualise them. The absolute-error metrics (Fig. 4) show that the four baselines cluster together with MAE between 12.8 and 14.2 and RMSE near 19-20, whereas the TFT reduces both metrics by roughly two-thirds. The relative-error view (Fig. 5) tells the same story: the TFT's MAPE of 41.5% is far below the 150-211% range of the other models, even accounting for the well-known instability of MAPE on low-count days.

TABLE IV. Performance Comparison of Forecasting Models

Model	MAE	RMSE	MAPE
LSTM	13.82	19.46	210.56%
Random Forest	14.19	20.11	210.21%
XGBoost	13.44	19.79	185.09%
N-BEATS	12.78	19.83	150.34%
TFT (selected)	4.46	5.02	41.49%

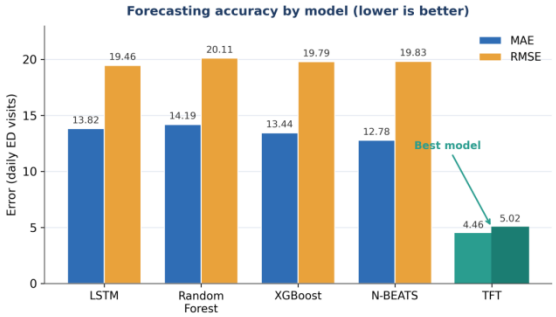

Fig. 4. MAE and RMSE by model on the held-out test set (lower is better). The TFT reduces error by roughly two-thirds relative to the baselines.

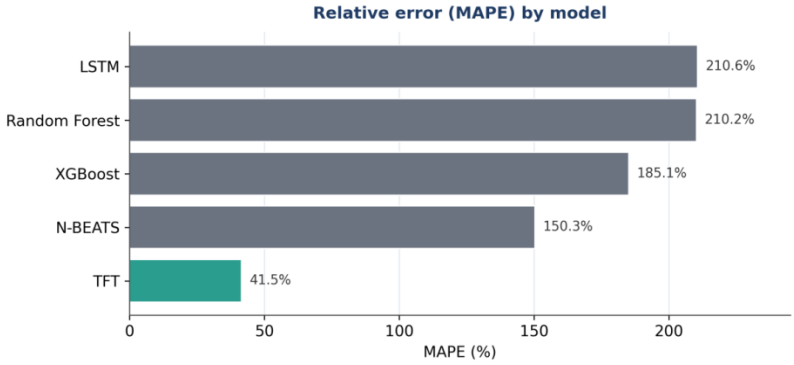

Fig. 5. MAPE by model. The TFT attains 41.5%, far below the other models, whose values are inflated by low-count days.

5. DISCUSSION

The results highlight that model choice is decisive for forecasting time-dependent healthcare demand. The TFT achieved the best MAE and RMSE by a wide margin, reflecting its capacity to model both short- and long-range temporal dependencies through the combination of recurrent encoding and attention, and to weight heterogeneous inputs adaptively through its variable-selection networks. From an operational standpoint, its low error makes it the most suitable model for staffing and capacity planning.

The tree-based models, although effective on tabular data, were constrained by their reliance on engineered lag and rolling features to represent sequence; even with these features they could not match the TFT. The LSTM, despite being designed for sequence modelling, did not surpass the TFT, plausibly because of its sensitivity to hyperparameters and the noise in the data. N-BEATS

performed weakest, which suggests that the dataset lacks the strong, clean periodicity that the architecture exploits, and that complex models are not universally superior on modest, noisy data.

Two observations recur across all models. First, MAPE values are consistently high. Because daily counts can occasionally be very low, small absolute errors translate into large percentage errors, so MAE and RMSE offer a more faithful picture of performance in this context. Second, the exogenous weather and seasonal features showed only weak correlation with ED visits, meaning the models relied chiefly on historical visit patterns. In effect, the task behaved more like a univariate time-series problem than a strongly multivariate one. This explains why the principal accuracy gains came from a model better at exploiting temporal structure (the TFT) rather than from richer exogenous inputs, and it points to data enrichment - epidemiological indicators, public events, and real-time capacity signals - as the most promising route to further improvement.

The explainability layer reinforced these conclusions. SHAP and the TFT attention and variable-selection scores consistently elevated recent visit history and day-of-week effects, while assigning limited importance to weather, corroborating the correlation analysis. Partial Dependence Plots made the model's response to individual features legible to non-specialists. Crucially, these tools also serve a governance role: by surfacing which inputs drive each forecast, they allow clinicians to validate the model's reasoning, support what-if analysis for administrators, and flag any spurious reliance on proxy variables.

Placed against the literature, the absolute accuracy of the selected model is encouraging. Reported MAPE values for daily ED forecasting commonly fall in the single digits to low tens of percent on large, multi-site datasets [4], [17], whereas percentage errors on this single-hospital series are inflated by frequent low-count days, making MAE the more meaningful comparator here. An MAE of roughly four to five visits against a daily mean near ninety-four corresponds to an error of about five percent of typical volume, which is competitive and operationally useful. The gap between the TFT and the baselines is also consistent with prior findings that attention-based, multi-horizon architectures excel when demand is governed by interacting temporal factors, while tree-based and plain recurrent models remain strong but secondary baselines.

A further consideration concerns transferability. Because MIMIC-IV-ED reflects a single United States academic medical centre, the trained models are best understood as a methodological proof of concept rather than a ready-made predictor for Saudi hospitals. The pipeline, feature engineering, and model-selection findings are expected to transfer, but the learned parameters would need to be re-estimated on local data that capture regional demand patterns, holiday calendars, and climate. The system was deliberately designed with this in mind: its data layer and retraining loop allow the same architecture to be re-fitted to a hospital-specific series, so that the demonstrated accuracy can be reproduced in the target deployment environment once local records are available.

Evaluated against the project's business objectives, the system meets its core aims. It delivers accurate daily forecasts that enable proactive shift planning, scheduling, and resource allocation; it represents real-world demand drivers in an interpretable way; and its outputs are readily visualised in a dashboard for operational use. In short, the proposed TFT-based forecasting system provides reliable and actionable predictions that support proactive hospital decision-making.

6. SYSTEM DEPLOYMENT

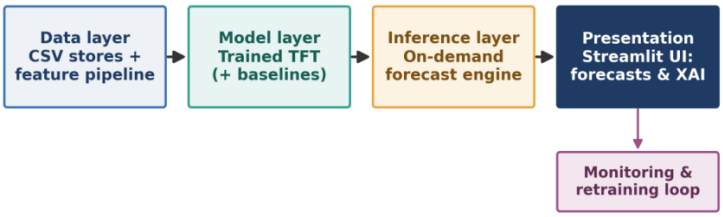

Fig. 6. Layered deployment architecture of the Marsad forecasting application, with a monitoring and retraining loop.

To demonstrate operational feasibility, the best model was embedded in an interactive forecasting application named "Marsad." The system follows a layered architecture (Fig. 6): a data layer that maintains the feature pipeline and CSV stores; a model layer holding the trained TFT and baseline models; an inference layer that generates forecasts on demand; and a presentation layer implemented as a Streamlit interface. A lightweight monitoring and retraining loop tracks model behaviour over time.

Through the interface, a user supplies forecast parameters and receives a summary of expected demand together with a fourteen-day daily forecast and a short-horizon twelve-hour hourly forecast. Forecast results are accompanied by trend visualisations and interpretation drawn from the model's attention and variable-importance information, so that users not only see the expected demand but also understand the factors behind it. The frontend and backend are integrated so that parameter selection triggers on-demand inference, with state management and navigation handling the flow of data from user input to rendered forecast. The current prototype relies on CSV-based data and cloud hosting suitable for research, with documented model versioning and a response plan for drift, providing a foundation for future integration with live hospital information systems.

6.1. User Interface

The Marsad interface is organised around a small number of task-focused pages. A home page orients the user and provides navigation; an input page (Fig. 7) collects forecast parameters; and a results page presents the forecast (Fig. 8). The results view combines summary indicators of expected demand with a fourteen-day daily forecast chart and a twelve-hour hourly forecast, so that planners can see both the near-term shift picture and the two-week outlook in one place. A dedicated explanation view surfaces the model's interpretation - which recent periods and which inputs most influenced the forecast - translating the attention and variable-selection signals into plain-language insight for non-specialist staff.

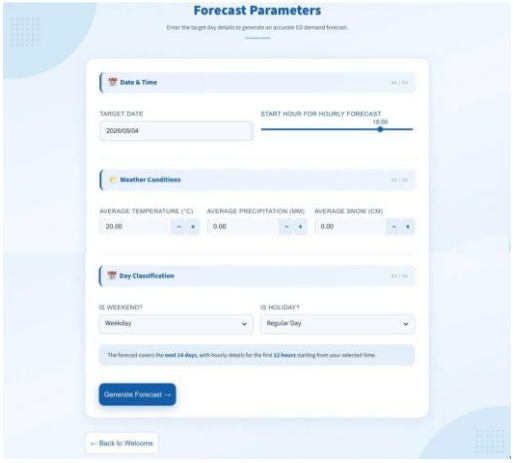

Fig. 7. User Input Interface for Forecast Parameters.

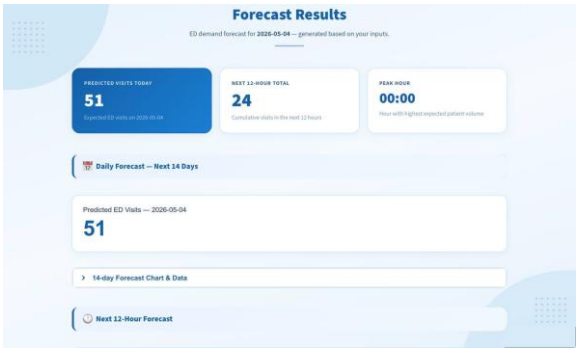

Fig. 8. Forecast summary indicators

6.2. Integration, Management, and Monitoring

The frontend and backend are integrated so that parameter selection triggers on-demand inference: the application loads the serialised models, assembles the required feature window, executes prediction, and renders the result, with state management coordinating navigation and the flow of data from input to output. For longer-term reliability, the system includes a monitoring structure that tracks forecast behaviour, a documented response plan for handling performance drift, and model versioning so that updates are auditable. These elements make the prototype not merely a demonstration of accuracy but a blueprint for a maintainable operational service.

7. Limitations and Ethical Considerations

A comprehensive assessment of the proposed approach requires consideration of both its practical limitations and the ethical factors associated with its use. The following subsections outline the key challenges, assumptions, and responsible deployment considerations relevant to this work.

7.1. Limitations

Several limitations temper the findings. The available exogenous features - weather and seasonal indicators - correlated only weakly with ED visits, so the models relied heavily on historical

patterns and could not fully capture demand variations driven by unobserved factors such as epidemics, public events, or policy changes. The data also contain noise and sudden fluctuations that are difficult for any model, and especially for data-hungry deep architectures such as N-BEATS, to learn stably. Evaluation used a single chronological train/test split rather than time-series cross-validation; while this preserves temporal order and avoids leakage, it assesses performance on only one test segment and may understate variance. MAPE proved unreliable on low-count days and should be read alongside MAE and RMSE. Finally, the deployed prototype currently forecasts only the next twelve hours at hourly resolution and relies on CSV data and cloud hosting suited to a research setting rather than large-scale production.

7.2. Ethical Considerations

Because the study uses aggregated, de-identified ED records from MIMIC-IV-ED [27], the risk of direct harm to individuals is low, and access is governed by a credentialed data-use agreement. Nevertheless, fairness, transparency, and privacy were treated as first-class concerns. Aggregation reduces the risk of individual bias, but indirect biases can persist if historical demand reflects inequitable access; the explainability layer helps detect such effects by exposing the features that drive predictions. Forecasts are intended to support, not replace, clinical and administrative judgement, and their uncertainty - made explicit through the TFT's quantile outputs - should be communicated to decision-makers. Any future deployment on real patient streams would require renewed privacy safeguards, governance review, and ongoing monitoring for drift and disparate impact.

8. Conclusion and Future Work

This paper presented an interpretable forecasting system for emergency department patient arrivals, developed within a modified CRISP-DM framework on de-identified MIMIC-IV-ED data enriched with weather and calendar features. Five models - Random Forest, XGBoost, LSTM, N-BEATS, and the Temporal Fusion Transformer - were compared under a leakage-free chronological protocol. The TFT was the clear best performer, achieving an MAE of 4.46, an RMSE of 5.02, and a MAPE of 41.5%, well ahead of baselines whose MAE ranged from 12.8 to 14.2. Explainability methods (SHAP, PDP, and attention) confirmed that recent visit history and day-of-week effects dominate demand while weather contributes little, and the selected model was deployed in the Marsad dashboard to deliver daily and short-horizon hourly forecasts with interpretation in support of proactive ED operations and Saudi Vision 2030.

Future work follows directly from the limitations. The most promising direction is data enrichment: integrating epidemiological signals, public-event calendars, and real-time hospital capacity indicators to strengthen the exogenous information available to the models. On the methodological side, time-series cross-validation, more thorough hyperparameter optimisation, longer historical series, and ensemble combinations may improve robustness. On the systems side, extending the hourly module to a full twenty-four-hour horizon, adding automated retraining pipelines, connecting to live hospital information systems, and providing mobile-friendly interfaces would move the prototype toward production-grade deployment.

More broadly, the study illustrates a repeatable template for operational forecasting in healthcare: integrate routine administrative records with low-cost external context, compare a spectrum of models under a leakage-free protocol, select on operationally meaningful metrics, attach explanations that domain experts can audit, and deliver the result through an interface that fits existing workflows. The same template applies to related capacity problems such as inpatient admissions, theatre scheduling, and clinic demand. We therefore view the contribution less as a

single trained model and more as an end-to-end, interpretable, and transferable approach to demand forecasting that hospitals can adapt to their own data and decision processes.

ACKNOWLEDGMENT

The authors thank the College of Computer and Information Sciences at Princess Nourah bint Abdulrahman University for its support, and the PhysioNet team and the creators of MIMIC-IV-ED for making the dataset available to the research community.

REFERENCES

[1] Saudi Vision 2030, "National Transformation Program," Kingdom of Saudi Arabia, 2023.
[2] Saudi Vision 2030, "Health Sector Transformation Program," Kingdom of Saudi Arabia, 2023.
[3] Ministry of Health (MOH), "Saudi Health Sector Transformation Strategy," 2022.
[4] B. M. Porto and F. S. Fogliatto, "Enhanced forecasting of emergency department patient arrivals using a feature engineering approach and machine learning," BMC Medical Informatics and Decision Making, vol. 24, no. 377, 2024.
[5] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice. OTexts, 2018.
[6] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.
[7] T. Geetha and S. Sendhilkumar, Machine Learning: Concepts, Techniques and Applications. Chapman and Hall/CRC, 2023.
[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929-1958, 2014.
[9] G.-B. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: a survey," International Journal of Machine Learning and Cybernetics, vol. 2, no. 2, pp. 107-122, 2011.
[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
[11] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: concerns and ways forward," PLOS ONE, vol. 13, no. 3, p. e0194889, 2018.
[12] S. Smyl, "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting," International Journal of Forecasting, vol. 36, no. 1, pp. 75-85, 2020.
[13] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[14] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 785-794.
[15] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-BEATS: neural basis expansion analysis for interpretable time series forecasting," in Proc. Int. Conf. Learning Representations (ICLR), 2020.
[16] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," International Journal of Forecasting, vol. 37, no. 4, pp. 1748-1764, 2021.
[17] J. Tuominen, F. Lomio, N. Oksala, A. Palomaki, J. Peltonen, H. Huttunen, and A. Roine, "Forecasting daily emergency department arrivals using high-dimensional multivariate data: a feature selection approach," BMC Medical Informatics and Decision Making, vol. 22, no. 134, 2022.
[18] R. Safaripour and H. J. J. Lim, "Comparative analysis of machine learning approaches for predicting frequent emergency department visits," Health Informatics Journal, vol. 28, 2022.
[19] Y.-C. Lee, C.-J. Ng, C.-C. Hsu, C.-W. Cheng, and S.-Y. Chen, "Machine learning models for predicting unscheduled return visits to an emergency department: a scoping review," BMC Emergency Medicine, vol. 24, 2024.
[20] C. Pelaez-Rodriguez, R. Torres-Lopez, J. Perez-Aracil, N. Lopez-Laguna, S. Sanchez-Rodriguez, and S. Salcedo-Sanz, "An explainable machine learning approach for hospital emergency department visits forecasting using continuous training and multi-model regression," Computer Methods and Programs in Biomedicine, vol. 245, p. 108033, 2024.
[21] M. Neshat, M. Phipps, N. Jha, D. Khojasteh, M. Tong, and A. Gandomi, "Effective predictive modeling for emergency department visits and evaluating exogenous variables impact: using explainable meta-learning gradient boosting," arXiv:2411.11275, 2024.

[22] O. Vural, B. Ozaydin, J. Booth, B. F. Lindsey, and A. Ahmed, "Deep learning-based forecasting of boarding patient counts to address ED overcrowding," arXiv:2505.14765, 2025.
[23] C. Brossard et al., "Predicting emergency department admissions using a machine-learning algorithm: a proof of concept with retrospective study," BMC Emergency Medicine, vol. 25, no. 3, 2025.
[24] J. Tuominen, E. Pulkkinen, J. Peltonen, J. Kanniainen, N. Oksala, A. Palomaki, and A. Roine, "Forecasting emergency department crowding with advanced machine learning models and multivariable input," arXiv:2308.16544, 2023.
[25] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.
[26] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.
[27] A. Johnson, L. Bulgarelli, T. Pollard, L. A. Celi, R. Mark, and S. Horng, "MIMIC-IV-ED (version 2.2)," PhysioNet, 2023.
[28] Time and Date AS, "Past weather in Boston, Massachusetts, USA," accessed 2025.
[29] University of New Mexico, “Support Vector Regression to Predict Power Consumption,” 2024. [Online].

©2026 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.