INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Forecasting Patient Demand in Portuguese Emergency Departments

Armando Milhais Ferreira

Mestrado em Ciência de Dados

Orientadora:

Doutora Diana Elisabeta Aldea Mendes, Professora Associada,

ISCTE - Instituto Universitário de Lisboa

Outubro, 2025

BUSINESS SCHOOL

TECNOLOGIAS E ARQUITETURA

Departamento de Métodos Quantitativos para Gestão e Economia

Departamento de Ciências e Tecnologia da Informação

Forecasting Patient Demand in Portuguese Emergency Departments

Armando Milhais Ferreira

Mestrado em Ciência de Dados

Orientadora:

Doutora Diana Elisabeta Aldea Mendes, Professora Associada,

ISCTE - Instituto Universitário de Lisboa

Outubro, 2025

Resumo

Os Serviços de Urgência (SU) são componentes essenciais dos sistemas de saúde, mas enfrentam frequentemente problemas crónicos de sobrelotação e limitações de recursos. Antecipar as flutuações na procura de doentes é, por isso, fundamental para melhorar a eficiência operacional e garantir um acesso atempado aos cuidados. Esta dissertação avalia e compara abordagens estatísticas e de aprendizagem supervisionada para a previsão mensal de atendimentos no Serviço de Urgência da Unidade Local de Saúde Santa Maria, em Lisboa, Portugal. Utilizando dados do portal SNS Transparência, referentes ao período de 2013 a 2025, foram implementados vários modelos, incluindo Exponential Smoothing (ETS), AutoRegressive Integrated Moving Average (ARIMA), Prophet e eXtreme Gradient Boosting (XGBoost) assim como modelos de *ensemble*. As previsões foram avaliadas para horizontes de um, três e seis meses através de validação cruzada com origem móvel (rolling-origin cross-validation), utilizando métricas como o Mean Absolute Scaled Error (MASE), o symmetric Mean Absolute Percentage Error (sMAPE) e o Root Mean Squared Error (RMSE).

Os resultados mostram que as abordagens baseadas em aprendizagem supervisionada e combinações de modelos alcançaram a maior precisão preditiva, com o XGBoost e os modelos combinados ponderados a superarem os modelos estatísticos tradicionais, sobretudo em horizontes de curto prazo. No entanto, modelos estatísticos como o ARIMA e o ETS ofereceram bases interpretáveis e desempenho consistente no médio prazo. O estudo demonstra que a integração de modelos complementares aumenta a robustez e que as ferramentas de previsão podem fornecer informação acionável para o planeamento de pessoal e de capacidade hospitalar. Estas conclusões contribuem para o corpo crescente de investigação sobre previsão em saúde baseada em dados, salientando o potencial da análise preditiva para apoiar a tomada de decisão operacional nos hospitais públicos**.**

Abstract

Emergency Departments (EDs) are essential to healthcare systems but often face chronic overcrowding and resource constraints. Anticipating fluctuations in patient demand is therefore critical for improving operational efficiency and ensuring timely access to care. This thesis evaluates and compares statistical and machine learning approaches for forecasting monthly patient attendances at the Emergency Department of *Unidade Local de Saúde Santa Maria* in Lisbon, Portugal. Using open administrative data from *SNS Transparência* covering 2013 to 2025, several models were implemented, including Exponential Smoothing (ETS), AutoRegressive Integrated Moving Average (ARIMA), Prophet, and eXtreme Gradient Boosting (XGBoost), alongside benchmark naïve methods. Forecasts were assessed across one-, three-, and six-month horizons using rolling-origin cross-validation and accuracy metrics such as Mean Absolute Scaled Error (MASE), symmetric Mean Absolute Percentage Error (sMAPE), and Root Mean Squared Error (RMSE).

Results show that machine learning and ensemble approaches achieved the highest predictive accuracy, with XGBoost and weighted model combinations outperforming traditional statistical models, particularly for short-term horizons. However, statistical models such as ARIMA and ETS provided interpretable baselines with consistent medium-term performance. The study demonstrates that integrating complementary models enhances robustness and that forecasting tools can provide actionable insights for hospital staffing and capacity planning. These findings contribute to the growing body of research on data-driven healthcare forecasting, highlighting the potential of predictive analytics to support operational decision-making in public hospitals.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1. Context and Motivation

1.1.1. Field Overview

Emergency Departments (EDs) are a vital component of healthcare systems, delivering urgent and often life-saving care to patients (Nelson et al., 1998). Despite their critical role, EDs worldwide face persistent challenges such as overcrowding, long waiting times, and constrained resources (Derlet & Richards, 2000). In Portugal, these pressures are particularly visible. Public hospitals registered more than 29 million ED visits in 2024, making them the most used healthcare entry point (Serviço Nacional de Saúde, 2025).

Forecasting patient demand has emerged as a key approach to addressing these issues (Afilal et al., 2016). By anticipating the volume and characteristics of patient arrivals, hospitals can better allocate staff, beds, and diagnostic resources. Accurate forecasts help mitigate bottlenecks, reduce delays, and improve both the quality of care and the working conditions of healthcare professionals (Kadri et al., 2014). Data science has strengthened this capability by enabling the analysis of historical patterns and the development of predictive models that adapt to changing demand trends (Mehta et al., 2019).

1.1.2. Challenges

Managing patient flows in EDs remains complex due to several interrelated factors. Resource limitations, including inadequate staffing and restricted bed capacity, frequently fail to match fluctuating demand, resulting in overcrowding and extended waiting times (Rasouli et al., 2019). Overcrowding increases the likelihood of medical errors, reduces care quality, and heightens stress for healthcare workers (Maninchedda et al., 2023). Inefficient processes such as delays in triage, diagnostics, or patient discharge further contribute to systemic bottlenecks that compromise throughput (Perelman et al., 2015).

These challenges underscore the importance of developing robust forecasting models tailored to healthcare contexts. By enabling more informed planning and efficient allocation of resources, such models represent a crucial step toward improving the resilience and sustainability of emergency care systems.

These systemic pressures have motivated a growing body of research focused on forecasting ED attendances as a means of improving planning and decision-making.

1.2. Objectives and Research Questions

The main objective of this thesis is to evaluate and compare forecasting methodologies for Emergency Department (ED) attendances in Portugal. By applying both statistical and machine learning approaches to hospital-level time series, the study aims to assess their predictive performance and operational relevance for healthcare planning.

This objective is structured around the following research questions:

- **RQ1:** Which forecasting models provide the most accurate predictions of ED attendances in Portuguese hospitals?
- **RQ2:** How do statistical approaches (e.g., ARIMA, ETS, Prophet) compare with machine learning methods (e.g., gradient boosting) in terms of forecast accuracy and robustness?
- **RQ3:** To what extent can forecasting outputs be translated into operational insights, particularly for staffing allocation and capacity management in EDs?

CHAPTER 2

Literature Review

2.1. Review introduction

This chapter reviews the main body of research on forecasting patient demand in Emergency Departments (EDs). It aims to synthesize methodological approaches, data sources, and performance outcomes from recent studies and to identify existing gaps that guide the methodological design of this thesis.

The analysis presented here is based on a systematic literature review (SLR) of eight high-quality peer-reviewed studies published between 2013 and 2024. The findings provide the foundation for the model comparison and evaluation framework developed in subsequent chapters.

2.2. Review Methodological Approach

The literature review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and adopted the PICOC framework (Population, Intervention, Comparison, Outcome, Context) to ensure transparency and reproducibility (Carrera-Rivera et al., 2022).

2.2.1. Databases and search period

Four scientific databases were used (PubMed, Scopus, IEEE Xplore, and Google Scholar) covering publications from 2013 to 2024.

2.2.2. Selection criteria

Only peer-reviewed articles or conference papers written in English or Portuguese, with a minimum Q3 impact factor, and addressing forecasting of ED patient flows were included. Studies without full-text access, limited methodological detail, or unrelated to healthcare forecasting were excluded.

2.2.3. Screening process.

Out of 61 initial records, 51 were excluded based on eligibility criteria, and 3 duplicates were removed. The final dataset comprised eight studies of high methodological quality, each scored

between 15 and 20 on a five-point Likert scale assessing reporting, rigor, credibility, and relevance.

2.3. Selected Studies

The eight selected studies (see Table 1) span diverse healthcare contexts, including Portugal, the United Kingdom, France, the United States, and New Zealand. Forecasting horizons range from hourly to monthly, although most studies emphasize daily predictions. Data sources generally comprise historical records of ED admissions or hospital administrative datasets extending from two to ten years.

Table 1. Studies includes in the SLR.

Year	Authors	Setting	Model(s)	Forecast Horizon	Best Metric
2018	Carvalho-Silva et al.	Portugal	ARIMA	Daily	MAPE 5.9 %
2019	Jilani et al.	UK	FTS, ARIMA	Weekly / Monthly	MAPE ≈ 2–3 %
2020	Harrou et al.	France	VAE, LSTM	Hourly	$R^2 = 0.93$
2021	Rocha & Rodrigues	Portugal	RNN, XGBoost	Hourly	sMAPE 4.3 %
2023	Susnjak & Maddigan	New Zealand	Voting Ensemble	Daily (91 days)	MAPE ≈ 13 %
2023	Girishan et al.	USA	XGBoost + MILP	Daily (90 days)	MAPE 5.9 %
2024	Elvas et al.	Portugal	Prophet	Weekly	MAPE 6 %

2.3.1. Forecasting Approaches

Early research predominantly used statistical models, including AutoRegressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and Exponential Smoothing (ETS). These methods provided interpretable baselines but were limited in handling nonlinear or multi-seasonal patterns.

Since 2019, there has been a marked shift toward machine learning and deep learning models such as Extreme Gradient Boosting (XGBoost), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Variational Autoencoders (VAE). These techniques exploit complex temporal relationships and demonstrate improved accuracy across most settings.

Recent studies integrate ensemble and hybrid approaches, combining statistical and ML models to capitalize on complementary strengths (Rocha & Rodrigues, 2021; Susnjak & Maddigan, 2023). Such combinations typically result in better performance in multi-step forecasts.

2.3.2. External variables

The incorporation of exogenous variables varies considerably. Most studies rely solely on historical admissions, while a few integrate calendar effects (day of week, month, holidays) or weather variables such as temperature and precipitation. Socioeconomic indicators remain rarely considered despite their potential influence on ED demand.

Portuguese studies, (Carvalho-Silva et al., 2018; Elvas et al., 2024; Rocha & Rodrigues, 2021) are particularly relevant to this thesis, demonstrating that even simple calendar-based covariates can significantly improve forecast accuracy.

2.3.3. Model Evaluation

A limitation across the reviewed studies is the lack of standardized evaluation metrics. Reported measures include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), symmetric MAPE (sMAPE), Mean Absolute Scaled Error (MASE), and Coefficient of Determination (R^2). While percentage-based metrics facilitate interpretation, they complicate cross-study comparison due to differences in data scale and forecast horizon. Despite this heterogeneity, deep-learning and ensemble approaches consistently achieve the lowest errors, typically MAPE ≈ 3–7 %.

2.3.4. Operational Impact

Only one study (Girishan Prabhu et al., 2023) explicitly linked forecasting to operational outcomes, coupling XGBoost predictions with a Mixed-Integer Linear Programming model to optimize physician shift scheduling. This integration yielded measurable improvements in staff utilization and reduced idle time.
Other works (Rocha & Rodrigues, 2021) visualized forecasts through dashboards for hospital administrators but stopped short of quantitative evaluation. Overall, the literature reveals a gap between predictive performance and actionable decision-support implementation in clinical operations.

2.4. Key Insights and Research Gaps

The comparative analysis of these studies reveals several insights and unresolved issues:

1. **Shift to Data-Driven Methods:** Since 2019, research has increasingly adopted ML and deep learning, emphasizing flexibility and scalability.

2. **Short Forecast Horizons:** Most work targets daily or weekly predictions with few extending to monthly horizons that are more relevant for strategic staffing and resource planning.
3. **Heterogeneous Metrics:** The absence of unified performance measures limits comparability and meta-analysis.
4. **Data Accessibility:** Publicly available hospital datasets remain scarce, limiting replication and external validation.
5. **Limited Use of External Variables:** Contextual drivers such as weather, demographics, and socioeconomic indicators may be underexplored.
6. **Lack of Operational Evaluation:** Very few studies translate predictive results into measurable improvements in ED throughput or staff efficiency.

2.5. Research Implications

The reviewed literature demonstrates a clear evolution in forecasting methodologies applied to Emergency Department (ED) demand. Traditional statistical models have progressively been complemented or replaced by machine learning and deep learning approaches. These developments reflect a broader shift toward data-driven, flexible, and scalable forecasting techniques capable of capturing complex temporal and nonlinear patterns in healthcare demand.

Despite these advances, several persistent challenges remain. Studies often rely on limited or proprietary datasets, employ heterogeneous evaluation metrics that hinder comparability, and focus predominantly on short-term forecasting horizons. Moreover, although many models achieve high predictive accuracy, few explore the integration of forecasts into hospital decision-making processes, leaving a gap between predictive capability and operational application.

In response to these gaps, the present thesis is designed to:

- compare statistical and machine-learning models within a unified cross-validation and evaluation framework;
- extend the forecast horizon to one month or more, in line with planning cycles used by hospital administrators;
- investigate how forecasting outputs can support resource allocation and staffing strategies.

By addressing these objectives, the thesis seeks to contribute to the development of robust and interpretable forecasting models suited to the Portuguese healthcare context, offering practical insights for the management and planning of Emergency Department resources.

CHAPTER 3

Methodology

3.1. Research Design

This study adopts a quantitative, comparative forecasting design to evaluate methods for predicting patient attendances in Emergency Departments (EDs). The primary focus is Lisbon 's *Unidade Local de Saúde (ULS) Santa Maria*.

The forecasting task is defined as predicting monthly total ED attendances. Forecasts are generated for one-, three-, and six-month periods ahead, covering both short- and medium-term decision-making needs.

A set of statistical, semi-parametric, and machine learning models are compared. Statistical benchmarks include the Exponential Smoothing (ETS), and AutoRegressive Integrated Moving Average (ARIMA) models, following established principles of time-series analysis (Hyndman & Athanasopoulos, 2018). Prophet is incorporated as a modern additive model capable of handling multiple seasonalities (Taylor & Letham, 2018). Finally, a machine learning approach based on eXtreme Gradient Boosting (XGBoost) is applied, using time-based features to capture nonlinear dynamics (Chen & Guestrin, 2016).

Model performance is assessed through rolling-origin cross-validation, which reflects realistic forecasting conditions. Evaluation metrics include the Mean Absolute Scaled Error (MASE), symmetric Mean Absolute Percentage Error (sMAPE), and Root Mean Squared Error (RMSE), ensuring comparability with prior studies and interpretability for practitioners.

3.2. Data Collection and Preprocessing

The dataset used in this study was obtained from *SNS Transparência*, the official open data portal of the Portuguese public healthcare service (Serviço Nacional de Saúde, 2025). Specifically, the analysis relies on the dataset *"Atendimentos por Tipo de Urgência Hospitalar"*, which reports monthly emergency attendances disaggregated by hospital entity and type of urgency (general, paediatric, obstetric, and psychiatric). The period covers January 2013 to June 2025. The dataset schema is displayed in Table 2.

Table 2. Dataset schema.

Variable	Type	Description	Usage in Thesis
Período	Date	Year and month of record (YYYY-MM)	Used (time index)
Região	Categorical	\Health region of Portugal	Not used (focus on Lisbon hospital)
Instituição	Categorical	Hospital/ULS name	Used (to filter CHU Lisboa Norte / ULS Santa Maria)
Localização Geográfica	String (Lat/Long)	Coordinates of institution	Not used
Urgências Geral	Numeric	Number of general emergency attendances	Not used (subtype)
Urgências Pediátricas	Numeric	Number of paediatric emergency attendances	Not used (subtype)
Urgência Obstetricia	Numeric	Number of obstetric emergency attendances	Not used (subtype)
Urgência Psiquiátrica	Numeric	Number of psychiatric emergency attendances	Not used (subtype)
Total Urgências	Numeric	Total emergency attendances per institution per month	Main variable

The institutional framework of the Portuguese healthcare system underwent a reform in January 2024, through which hospitals and primary care units (ACES) were integrated into *Unidades Locais de Saúde (ULS)* (República Portuguesa, 2023). As part of this process, the *Centro Hospitalar Universitário de Lisboa Norte, EPE,* which included *Hospital de Santa Maria* and *Hospital Pulido Valente,* was reorganized into *ULS Santa Maria*, EPE. The *Centro Hospitalar Universitário de Lisboa Norte (CHULN)* data (2013 to 2023) was linked to the new *ULS Santa Maria* data (2024–2025). Continuity was assumed after verifying that the CHULN series ended in December 2023, the ULS series began in January 2024, and the levels were consistent across the transition. A visual inspection confirmed it with no structural break other than expected seasonal fluctuations (see Figure 1).

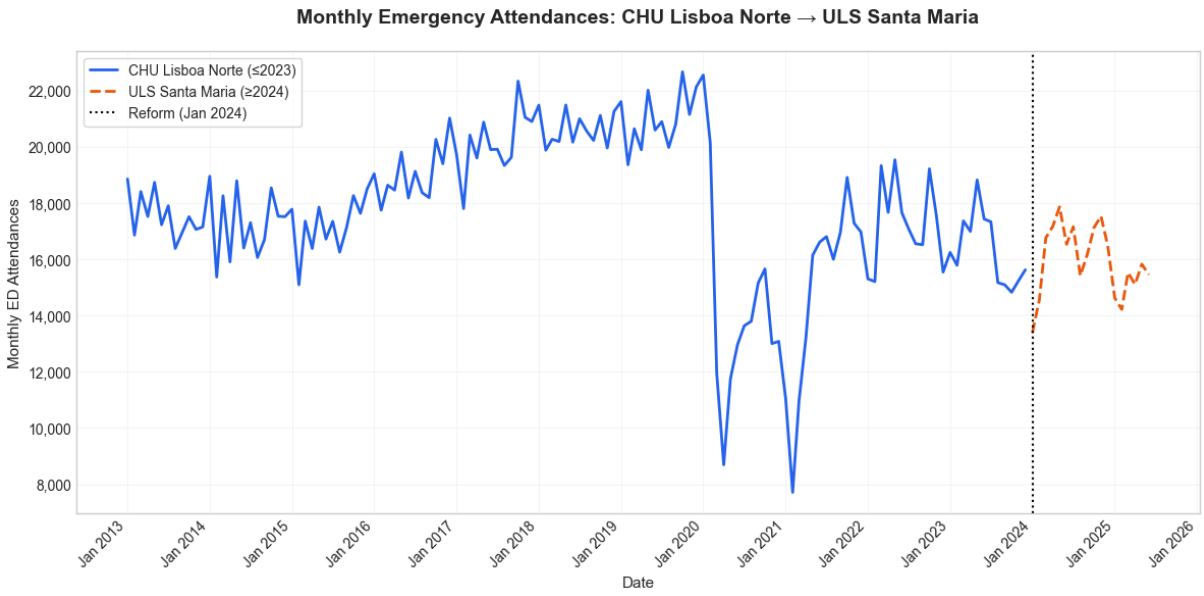

Figure 1. Trend in ED Attendances in CHU Lisboa Norte and ULS Santa Maria between 2013 and 2025.

ULS Santa Maria was selected as the case study for this thesis because it is a tertiary hospital complex, serving as a national referral centre with a consistently high patient volume. Its Emergency Department remains open continuously, unlike some smaller hospitals in Portugal that face periodic closures of specific emergency services due to staffing shortages. This ensures that the time series reflects a stable and uninterrupted demand, which is essential for developing and evaluating forecasting models.

The primary series considered in this thesis is the total number of ED attendances per month for *ULS Santa Maria*. This measure was selected to capture overall demand for emergency care, providing direct operational relevance for workforce allocation and capacity management. While the dataset also provides disaggregation by type of ED, the focus on total attendances ensures stability of volumes and avoids issues of sparsity in lower-frequency categories.

Although the dataset covers the period from January 2013 to June 2025, this study retained the full series to preserve long-term seasonal structure and to ensure sufficient data for model estimation and validation. Earlier observations may reflect different operational and reporting contexts, including the progressive digitalization of hospital information systems and changes in administrative procedures prior to 2016, and the disruption caused by the COVID-19 pandemic (2020 to 2022). Using rolling-origin cross-validation methods mitigates potential bias from older data, as recent observations receive greater weight in model fitting. Preprocessing steps included standardizing date formats and aligning institutional labels across the 2023 to

2024 reform boundary. No imputation was required, as the series presented complete monthly coverage throughout the study period.

Despite its utility, the dataset presents certain restrictions. First, attendances are reported only at monthly frequency, which limits the analysis of daily or hourly fluctuations in ED demand. Second, the information is provided at an aggregate institutional level, without patient-level variables such as age, diagnosis, or Manchester triage, which constrains the use of clinical covariates. Third, no staffing data were available for integration, which prevents a direct assessment of how forecasted demand translates into workforce requirements. Finally, the dataset is limited to hospital activity indicators and does not incorporate external contextual drivers (e.g., seasonal epidemics, weather conditions, or holiday effects). While such factors could potentially improve forecast accuracy, their integration would require additional data sources and falls outside the scope of this study. These constraints are revisited in the Discussion when considering the practical application of forecasts to hospital operations.

3.3. Exploratory Data Analysis

Prior to model development, an exploratory analysis was performed to evaluate the statistical properties, temporal consistency, and structural components of the time series. This step aimed to ensure the reliability of the dataset and to inform the specification of appropriate forecasting models.

Data quality and temporal continuity were first verified. The monthly sequence was confirmed to be complete and free of missing values or duplicate entries, and institutional identifiers were standardized to ensure continuity across the 2024 reorganization into *ULS Santa Maria*. Descriptive statistics, including mean, median, standard deviation, and coefficient of variation, were computed to characterize central tendency and dispersion, while histograms were used to assess the presence of outliers or asymmetry (tables and figures present in subchapter 4.1.1).

The temporal dynamics of the series were explored using line and seasonal subseries plots to visually identify long-term trends and recurring patterns. A Seasonal-Trend decomposition using Loess (Cleveland et al., 1990) was then applied to separate the series into trend, seasonal, and remainder components, using a 12-month periodicity.

Autocorrelation and partial autocorrelation functions (ACF and PACF) were computed up to 36 lags to detect serial dependence at both short and seasonal intervals. Complementary Ljung–

Box tests were applied at multiple lag lengths (6, 12, 18, and 24) to formally evaluate the independence of residuals (Hyndman & Athanasopoulos, 2018). Variance stabilization through logarithmic or Box–Cox transformation was evaluated by inspecting changes in dispersion across time. No transformation was applied since the variance remained approximately constant and additive seasonality was evident (see subchapter 4.1).

All exploratory analyses and visualizations were implemented in Python (using *pandas*, *numpy*, *statsmodels* and *matplotlib*).

3.4. Feature Engineering

For statistical models such as ETS, ARIMA, and Prophet, trend and seasonality were inferred internally within the model structure. In contrast, the machine-learning approach (XGBoost) required explicit construction of temporal features to encode serial dependence and calendar effects.

To represent the annual seasonal cycle, eleven-month dummy variables were added, along with a linear time index capturing the long-term trend. To account for temporal autocorrelation, lagged values of one and twelve months were included, together with rolling means over three and twelve months (computed on y_{t-1} to prevent information leakage) to represent short- and medium-term momentum.

In addition, three binary holiday indicators were created for periods known to influence healthcare utilization in Portugal:

- March to April (*Easter window*),
- August (summer vacation), and
- December to January (*Christmas/New Year*).

All features were aligned at the monthly start frequency and shifted by one period where necessary to ensure that only past information was available at each prediction step. This design produced a total of 19 exogenous predictor**s** (11 monthly dummies, 1 trend index, 3 holiday flags, 2 lags, and 2 rolling averages). To interpret the model, feature-importance analysis was performed using the gain metric, which measures each predictor's average contribution to reducing forecast error across all trees.

3.5. Forecasting Models

The forecasting framework combines statistical, semi-parametric, and machine learning approaches, all benchmarked against naïve methods to ensure that improvements are meaningful. The model selection was guided by best practices in forecasting (Hyndman & Athanasopoulos, 2018) and reflects the methodologies most frequently applied in healthcare demand prediction as stated in Chapter 1.

3.5.1. Benchmark Models

Following best practices, two simple methods are used as benchmarks. The Naïve model assumes that the next observation equals the most recent value, while the Seasonal Naïve (SNaïve) model repeats the value from the same month of the previous year.

3.5.2. Exponential Smoothing (ETS)

Exponential smoothing methods forecast future values by weighting past observations with exponentially decreasing importance. The ETS framework (Error, Trend, Seasonality) generalizes simple exponential smoothing into a family of state-space models that can capture level, trend, and seasonal patterns. Automatic model selection ensures the most appropriate form is chosen based on information criteria. ETS models are interpretable and computationally efficient, and they often perform competitively with more complex approaches in practice (Hyndman & Athanasopoulos, 2018).

3.5.3. ARIMA

The AutoRegressive Integrated Moving Average (ARIMA) model describes a time series through its autocorrelation structure, combining autoregressive (AR), differencing (I), and moving average (MA) components. Seasonal ARIMA extensions allow yearly seasonal patterns to be incorporated. Parameter selection is automated using information criteria and residual diagnostics. ARIMA models remain a standard reference in forecasting research and provide a robust benchmark for healthcare applications (Tunnicliffe Wilson, 2016).

3.5.4. Prophet

Developed by Meta, Prophet is a semi-parametric additive model that decomposes a time series into trend, seasonal, and holiday components. It is robust to missing data, outliers, and trend shifts, making it attractive for applied forecasting contexts. Prophet's interpretability and ability to incorporate custom holiday effects provide additional flexibility compared to ETS or ARIMA (Taylor & Letham, 2018).

3.5.5. XGBoost

Extreme Gradient Boosting (XGBoost) is a machine learning algorithm that constructs an ensemble of decision trees using gradient boosting. For time series forecasting, the model relies on engineered features such as lagged values, rolling averages, and calendar indicators to capture temporal dependencies. While XGBoost does not explicitly model autocorrelation or seasonality, its ability to capture nonlinear interactions makes it a powerful complement to statistical approaches (Chen & Guestrin, 2016).

3.5.6. Model Combination

Combining forecasts from multiple models can often improve accuracy by reducing variance and compensating for individual model biases (Hyndman & Athanasopoulos, 2018).

In this work, two ensemble specifications were considered:

- Equal-weight combination - A simple average of the four base models was computed at each forecast origin. Each model contributed equally (weight = 0.25), reflecting the assumption that all models contain complementary information about the underlying data-generating process.
- Inverse-error weighted combination - A second ensemble was built using weights inversely proportional to each model's Mean Absolute Scaled Error (MASE) measured over the training set across multiple forecast horizons as shown in Equation 1. This approach assigns larger weights to models that demonstrated higher predictive accuracy (lower MASE) during cross-validation, while still ensuring that all models contribute positively. The weights were normalized to sum to one.

Mathematically, for model i with average training error E_i, its combination weight w_i, is defined as:

$$w_i = \frac{1/E_i}{\sum_j (1/E_j)} \tag{1}$$

Both combination models were estimated recursively so that, at each forecast origin, all constituent models were re-trained using only information available up to that point in time.

3.5.7. Model Implementation and Hyperparameter Selection

Model implementation and evaluation were performed in Python 3.11 using the sktime package (Löning et al., 2019). For ETS and ARIMA, hyperparameters were selected automatically using

the corrected Akaike Information Criterion (AICc), which penalises model complexity in small samples and ensures parsimonious fits.

For the machine-learning model (XGBoost), hyperparameter optimisation was performed through a grid search with expanding-window cross-validation, using the Mean Absolute Percentage Error (MAPE) as the selection criterion. The search spanned tree depth, learning rate, regularisation, and subsampling parameters, and was executed in parallel to ensure reproducibility and computational efficiency.

3.6. Evaluation Protocol

3.6.1. Forecast Horizons

Forecasts were generated for one-, three-, and six-month horizons (h = 1,3,6). These horizons were chosen to align with both short-term operational scheduling and medium-term capacity planning requirements in ED management. The same forecast horizons were applied uniformly across all models to ensure comparability. Rolling-origin cross-validation was used within the training period to evaluate performance at each horizon.

3.6.2. Data Partitioning and Validation

The data were split chronologically into training and evaluation subsets:

- The training period spans January 2013 to December 2023, with over a decade of historical observations for model estimation (132 data points).
- The testing period covers January 2024 to June 2025, providing 18 months of out-of-sample data for forecast evaluation.

This chronological split aligns with the 2024 institutional reform boundary, allowing the models to learn from pre-reform dynamics while assessing their generalization to the new operational context of *ULS Santa Maria*. It also includes pre-pandemic, pandemic, and post-pandemic phases, ensuring that the models are trained on a representative range of demand conditions.

Within the training window, the rolling-origin cross-validation procedure described below was used to estimate forecast accuracy. After validation, models were refitted on the entire training set and used to generate forecasts for the test period.

3.6.3. Baseline Comparisons

As mentioned previously, models were benchmarked against the Naïve and Seasonal Naïve (SNaïve) approaches.

3.6.4. Accuracy Metrics

Model performance was assessed using multiple accuracy measures:

- **Mean Absolute Scaled Error (MASE):** a scale-independent metric recommended for comparing across series and against baselines. Defined as:

$$\text{MASE} = \text{mean}\left(\frac{|e_j|}{\frac{1}{T-1}\sum_{t=2}^{T}|Y_t - Y_{t-1}|}\right) \quad (2)$$

- **Symmetric Mean Absolute Percentage Error (sMAPE):** measures relative accuracy while avoiding asymmetry of traditional MAPE (Hyndman & Koehler, 2006). Defines as:

$$\text{SMAPE} = \frac{2}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{|A_t| + |F_t|} \quad (3)$$

- **Root Mean Squared Error (RMSE):** emphasizes large forecast errors, providing an operationally relevant perspective for hospital managers. Defines as:

$$RSME = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N - P}} \quad (4)$$

3.6.5. Cross-Validation Strategy

A rolling-origin time series cross-validation approach was applied. This method involves repeatedly re-estimating models using an expanding training window and producing forecasts for the specified horizons. Errors are then averaged across all forecast origins. This procedure mimics real-world forecasting conditions, where new data become available sequentially and forecasts must be updated accordingly (Hyndman & Athanasopoulos, 2018). Figure 2 explains visually the cross-validation process.

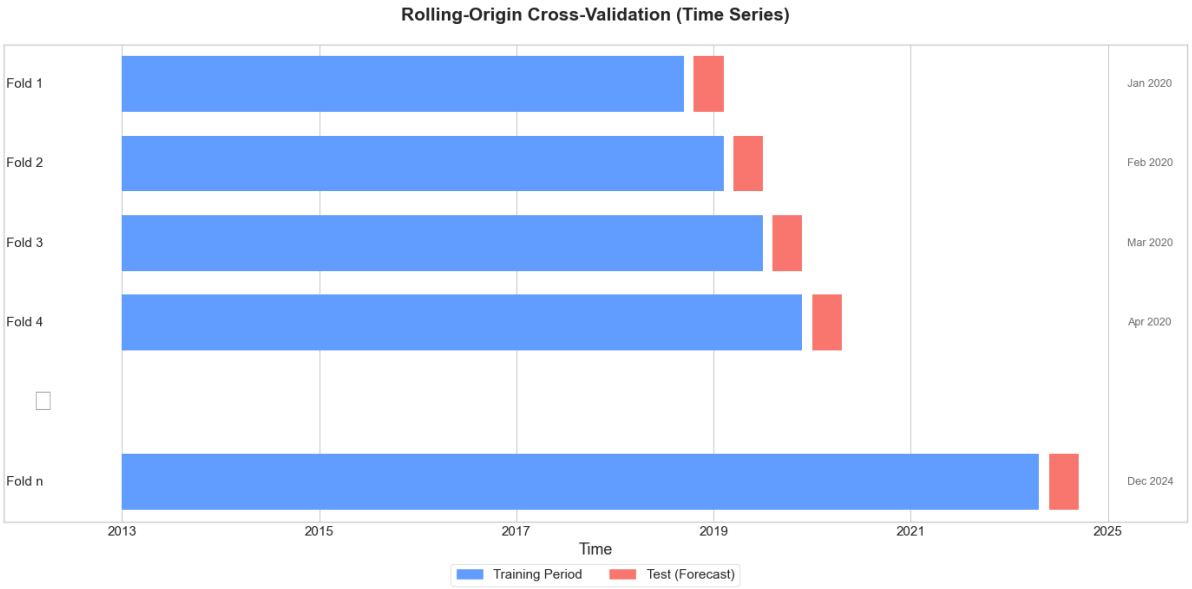

Figure 2. Rolling-origin cross-validation procedure. Each colour represents a forecast origin. The blue segment corresponds to the training period, while the red segment shows the forecast horizon. At each origin, the model is re-estimated with an expanding training window, and forecast errors are computed for the specified horizon. This process is repeated sequentially across the series, and the resulting errors are averaged to evaluate model performance.

3.6.6. Residual Diagnostics

For statistical and semi parametric models, residuals were examined to ensure model adequacy. Diagnostics included inspection of autocorrelation plots and the Ljung–Box test for independence.

CHAPTER 4

Results

4.1. Descriptive Statistics and Exploratory Data Analysis

This section provides an overview of the time-series characteristics of ED attendances for ULS Santa Maria, serving as the foundation for model specification and evaluation. Descriptive statistics and visual exploration were conducted to identify key temporal patterns, assess data quality, and verify the assumptions required for time-series forecasting.

4.1.1. Overview of the Series

The dataset comprises monthly total attendances covering more than twelve years of emergency care activity with a total of 150 time points. During this period, *ULS Santa Maria* recorded an average of approximately 17,560 attendances per month, with values ranging from 7,714 (February 2021) to 22,661 (October 2019).

Table 3. Summary statistics for monthly ED attendances (2013–2025).

Statistic	Value
Mean monthly attendances	17,560
Median monthly attendances	17,516
Standard deviation	2,659
Minimum (Feb 2021)	7,714
Maximum (Oct 2019)	22,661
Coefficient of variation	0.15

Summary statistics are displayed in Table 3. The distribution of monthly attendances is approximately symmetrical, with most months concentrated around a central range close to the median of 17,516 attendances. The mean is only slightly higher, indicating minimal skewness as you can see in Figure 3. This balanced distribution suggests that extreme high or low demand months are rare and that monthly variations remain relatively consistent over time. The absence of heavy tails or strong asymmetry supports the assumption of stable variance, justifying the use of additive seasonal models in subsequent forecasting.

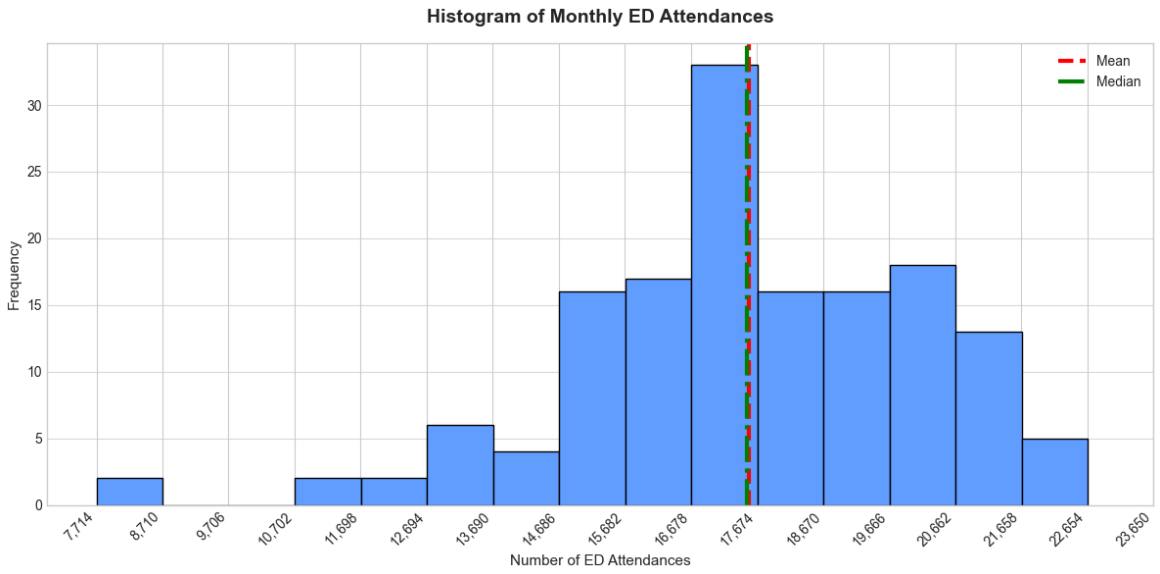

Figure 3. Histogram of monthly Emergency Department (ED) attendances at ULS Santa Maria, 2013–2025.

4.1.2. Trend and Seasonality

The monthly series from 2013 to 2025 displays moderate temporal structure with elements of both trend and seasonality (Figure 4). A gradual upward trajectory is observed from 2013 until late 2019, reflecting progressive growth in healthcare utilization in Lisbon. The series then shows a sharp contraction in 2020–2022, coinciding with the COVID-19 pandemic, followed by a partial recovery in 2022. These were treated as part of the natural historical variation rather than outliers. Retaining these periods ensures the models are trained on realistic demand variability and improve their resilience to future shocks. Since 2023, attendances exhibit a slight downward tendency, suggesting either a stabilization of demand or residual behavioural effects in the post-pandemic period.

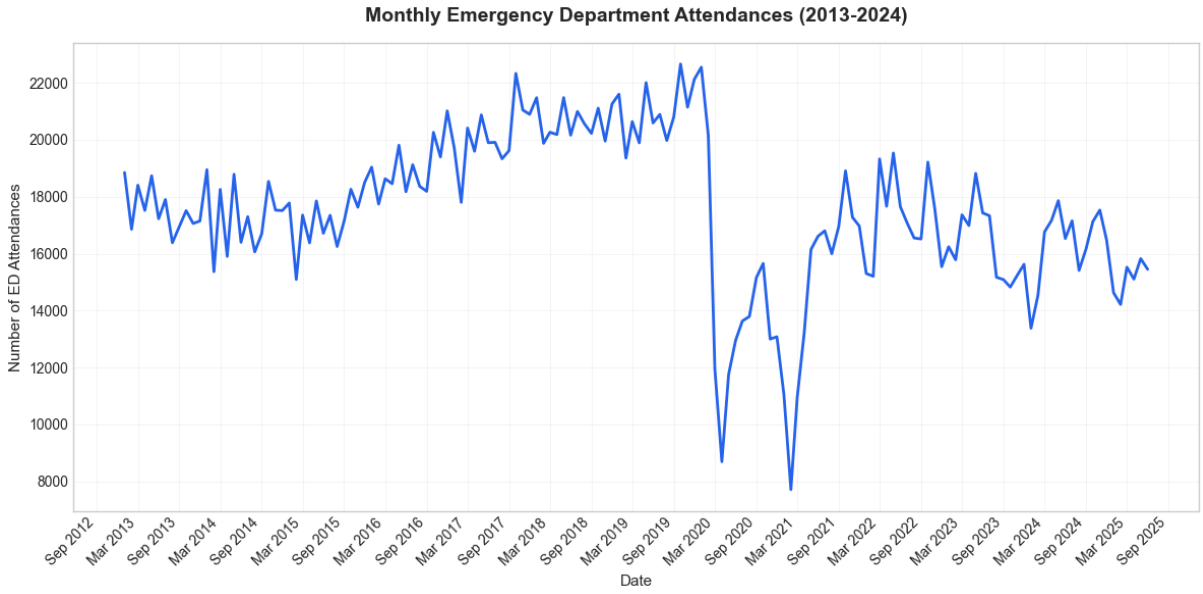

Figure 4. Monthly Emergency Department attendances at ULS Santa Maria (2013–2025).

Seasonal variation is present but relatively subtle and irregular. Modest peaks typically occur in fall (mainly October) and winter months (December to January) and lower volumes in summer (July to August), patterns that align with seasonal respiratory illnesses and a drop in utilization during vacation times (Figure 5). However, the amplitude of these cycles is limited and varies between years, indicating weak additive seasonality rather than a strong recurrent signal.

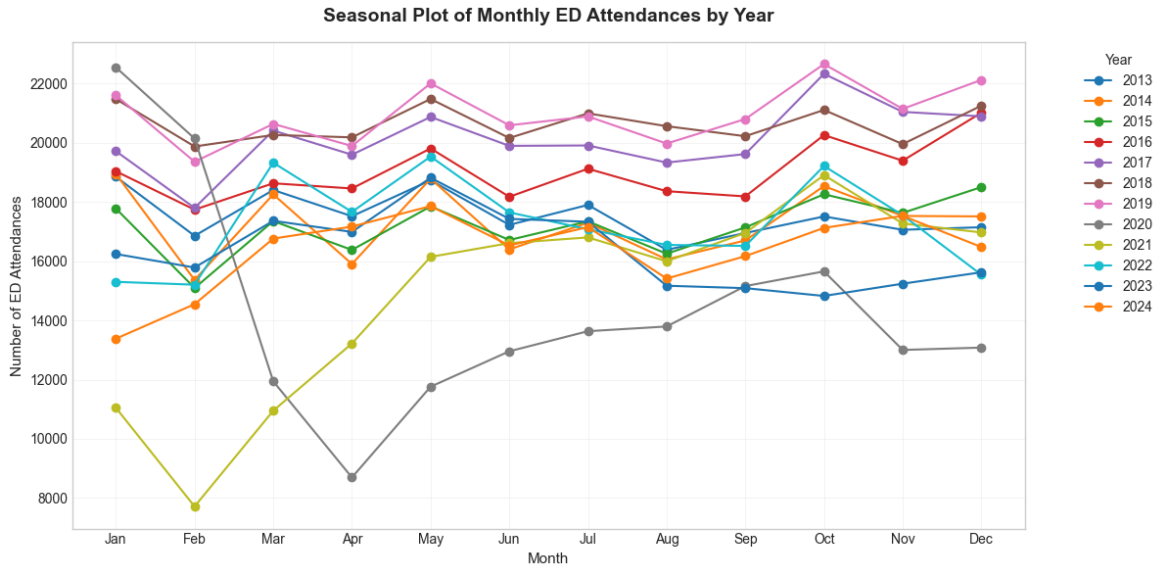

Figure 5. Seasonal plot of monthly attendances by year (2013 to 2025).

To formally assess the presence and strength of seasonal components, subsequent analysis applies time-series decomposition and autocorrelation diagnostics to quantify periodic behaviour and verify the stability of seasonal effects across the observation window.

4.1.3. Decomposition of Components

The series was decomposed using the STL method (Seasonal-Trend decomposition using Loess). This procedure separates the time series into three additive components (trend, seasonality, and remainder) to better understand its underlying temporal structure and assess the adequacy of subsequent forecasting models (see Figure 6).

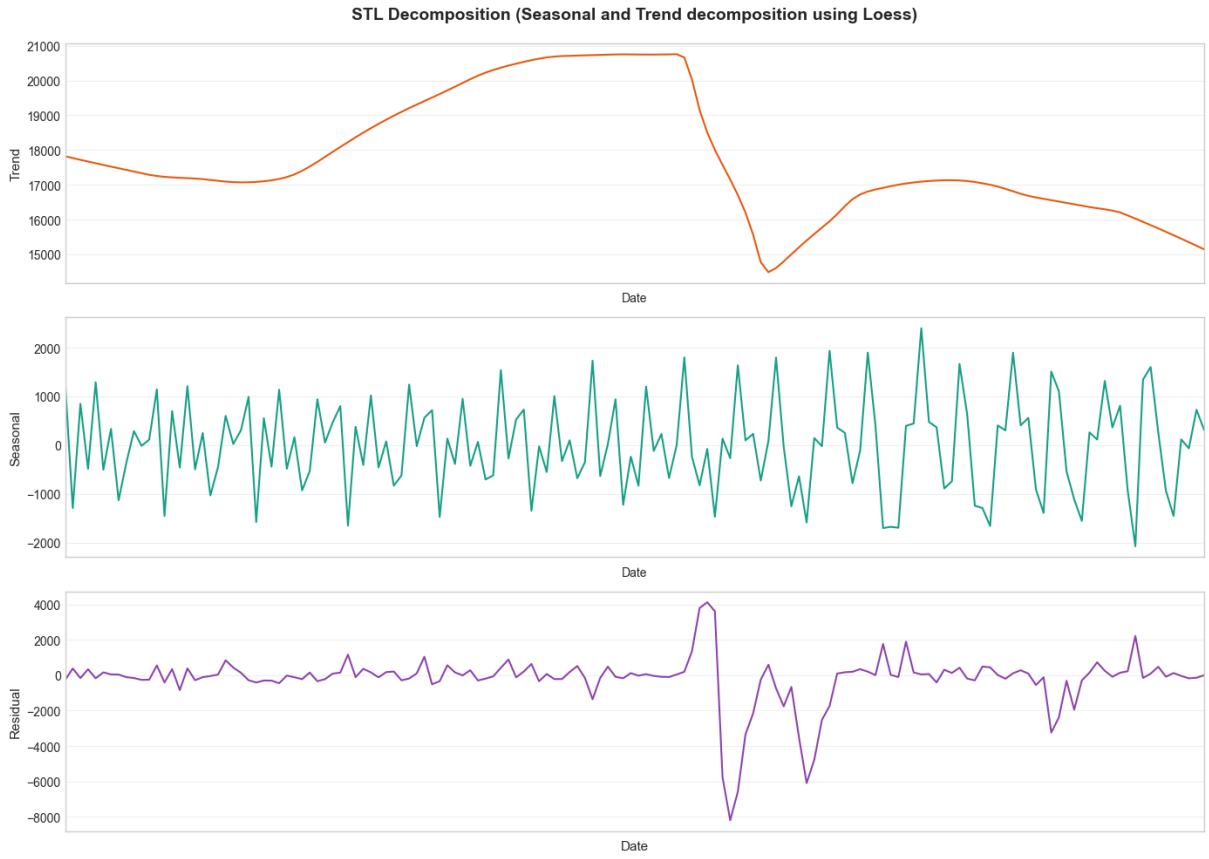

Figure 6. STL decomposition of monthly Emergency Department attendances (2013 to 2025), showing the trend, seasonal, and remainder components.

Like we have observed previously, the trend component indicates that patient attendances increased gradually between 2013 and 2019, reaching a peak in September 2019, followed by a sharp decline during 2020 and 2021 that corresponds to the COVID-19 pandemic. Since 2022, attendances have stabilized but show a mild downward trajectory, suggesting a potential shift in post-pandemic healthcare utilization patterns.

The seasonal component exhibits a modest and stable amplitude of approximately 16 % of the series mean, confirming the presence of a recurring annual pattern. The limited variation in seasonal magnitude across years supports the assumption of additive seasonality. The monthly average of the seasonal component (see Table 4) shows that attendances typically fall below the long-term mean in February (-1 495) and August (-851), while above-trend peaks occur in May and October (+1 364 each). This pattern suggests a bi-modal seasonal structure rather than a single winter surge, possibly reflecting local factors such as spring allergies and respiratory conditions and the resumption of activity after the summer period.

Table 4. Mean Monthly Seasonal Component from STL Decomposition.

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
69.5	-1494.9	293.4	-234.6	1364.3	-61.0	286.6	-851.0	-597.6	1364.1	217.9	-275.2

The remainder component captures short-term irregularities not explained by trend or seasonality. Residual diagnostics show that the decomposition explains approximately 68.6 % of the total variance in the series (see Table 5). The Ljung–Box test results ($p < 0.001$ for lags 12 and 24) indicate that residuals retain some autocorrelation, implying that additional temporal structure remains.

Table 5. STL Decomposition Diagnostics. * The residual variance explained is a pseudo-R^2 metric calculated as *1 − var(resid)/var(total).*

Component	Statistic	Result
Trend	Peak (Sep 2019); Trough (Sep 2020)	Pre-pandemic maximum followed by COVID-19 collapse
	% Change 2013 → 2019	-1.4 %
	% Change post-2022	-10.7 %
Seasonality	Peak months	May and October
	Trough months	February and August
	Amplitude ratio	16.3 % of mean
Residuals	Ljung–Box p-values (12, 24 lags)	< 0.001
	Variance explained*	68.6 %

4.1.4. Autocorrelation and Partial Autocorrelation Analysis

To further assess temporal dependence in the series, the residual component obtained from the STL decomposition was analysed using autocorrelation (ACF) and partial autocorrelation (PACF) functions. These diagnostics evaluate whether serial structure remains in the data after removing trend and seasonal components, which would indicate the need for additional autoregressive or moving-average terms in subsequent models.

As shown in Figure 7, the ACF displays a strong positive correlation at lag 1, followed by a gradual decay over the first six to eight months. Additional spikes at lags 9 to 10 and 13 to 15

suggest the presence of secondary cycles, while the small oscillations around lag 12 reflect residual annual seasonality that the STL decomposition did not fully capture. The PACF exhibits significant coefficients at lags 1 and 2, and an isolated weaker spike near lag 19, indicating that a low-order autoregressive process could explain part of the remaining structure.

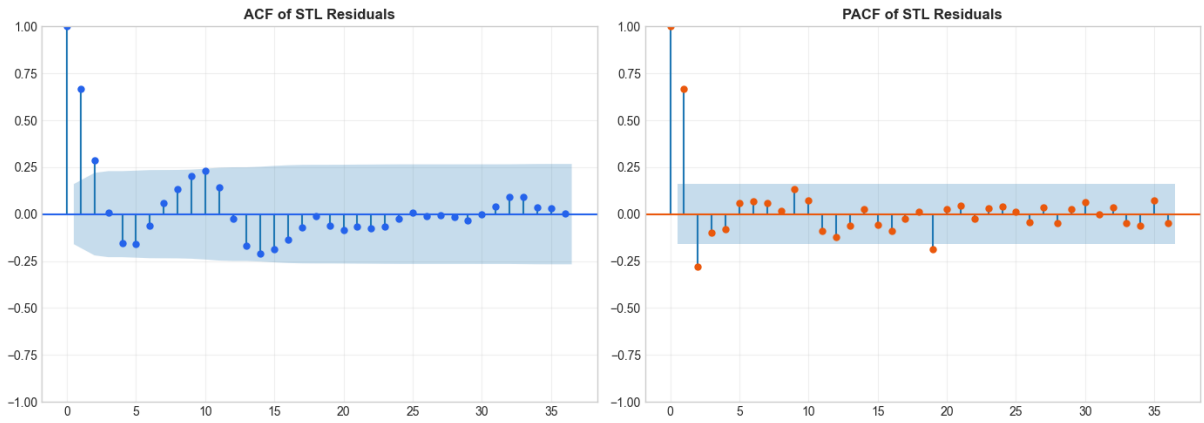

Figure 7. Autocorrelation function (ACF) and Partial autocorrelation function (PACF) of STL residuals.

Formally, the Ljung–Box test confirms these graphical findings. The test was highly significant at all evaluated horizons ($p < 0.001$ for lags 6, 12, 18, 24), rejecting the null hypothesis that residuals are uncorrelated. Again, these results imply that the residual series is not white noise and retains both short-term serial dependence and weak seasonal persistence.

This evidence supports the inclusion of autoregressive and seasonal terms in the subsequent ARIMA models.

4.2. Model specification

Building on the methodological framework established in Chapter 2, four forecasting models were estimated to predict monthly Emergency Department attendances. All models were trained using the dataset described in Section 2.2, with the training period covering January 2013 to December 2023 and forecasts validated on the January 2024 to June 2025 out-of-sample horizon defined in the evaluation protocol.

4.2.1. Exponential Smoothing (ETS)

Model selection based on the corrected Akaike Information Criterion (AICc = 2301.3), selected a ETS(A, A_d, A) specification for the exponential smoothing model. This model comprises an additive error, additive damped trend, and additive seasonal components. The damped trend parameter ($\phi = 0.80$) indicates a gradual attenuation of long-term growth, consistent with the slowdown observed after 2019. Level smoothing was complete ($\alpha \approx 1$), implying that the model

responds strongly to the most recent observations, while the near-zero trend and seasonal smoothing parameters ($\beta, \gamma \approx 0$) suggest that those components were largely governed by the estimated states rather than by continued adaptation.

Visual inspection of fitted versus actual values shows that the model captures the medium-term evolution of attendances, reproducing the pre-pandemic upward trend and the pronounced COVID-19 drop, followed by the moderate post-2022 decline. Seasonal peaks around May and October and troughs in February and August are well represented, consistent with the additive seasonal component.

Residual analysis supports an adequate but imperfect fit. The mean residual (-24 attendances) is close to zero and variance remains stable, indicating unbiased estimation of the level component. However, the Ljung–Box tests at 12 and 24 lags ($p < 0.01$) reveal some remaining short-term autocorrelation, implying that limited temporal dependence was not fully captured by the ETS dynamics. The Jarque–Bera test ($p < 0.00$) confirms slight non-normality of residuals. Despite these departures, residuals display homoscedasticity and no systematic patterns over time.

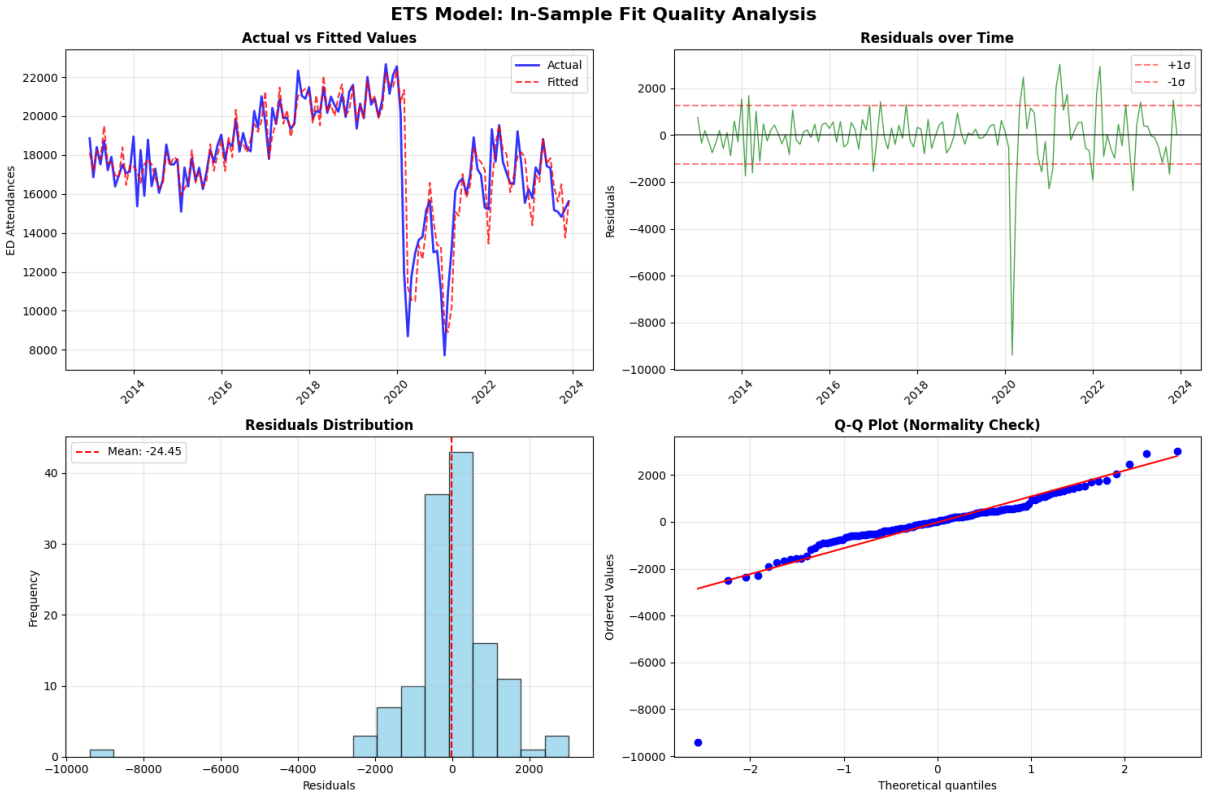

Figure 8. In-sample fit and residual diagnostics for the ETS(A, A_d., A) model.

4.2.2. AutoRegressive Integrated Moving Average (ARIMA)

The automatic ARIMA modelling procedure selected a SARIMA(1, 0, 2) × (2, 0, 0)[12] specification, corresponding to one non-seasonal autoregressive term, two non-seasonal moving-average terms, and two seasonal autoregressive components with a monthly seasonality (m = 12). This structure implies that short-term persistence and annual dependencies dominate the dynamics of monthly emergency attendances, while differencing was unnecessary (d = 0), confirming that the series is stationary in level.

Information-criterion values (AICc = 2312.9) indicate a model of comparable parsimony to ETS. The estimated non-seasonal AR(1) coefficient (0.67) denotes moderate temporal persistence, whereas the moving-average parameters (0.25 and 0.21) capture short-term shock adjustments. The seasonal AR coefficients at lags 12 and 24 (0.29 and 0.17) highlight a recurring yearly pattern consistent with the observed seasonality in attendances.

In-sample fitting reproduces the overall trajectory of the series, including the pre-2020 increase, the abrupt pandemic-related collapse, and the partial recovery thereafter. Residuals are approximately centred around zero (mean = -21 attendances) with homogeneous variance and no visible serial correlation. The Ljung–Box test (lag 10, $p = 0.28$) confirms that remaining autocorrelation is statistically insignificant, supporting the adequacy of the ARIMA specification. The residual distribution approximates normality apart from a few extreme values associated with the COVID-19 period, as reflected in the Jarque–Bera statistic ($p < 0.00$).

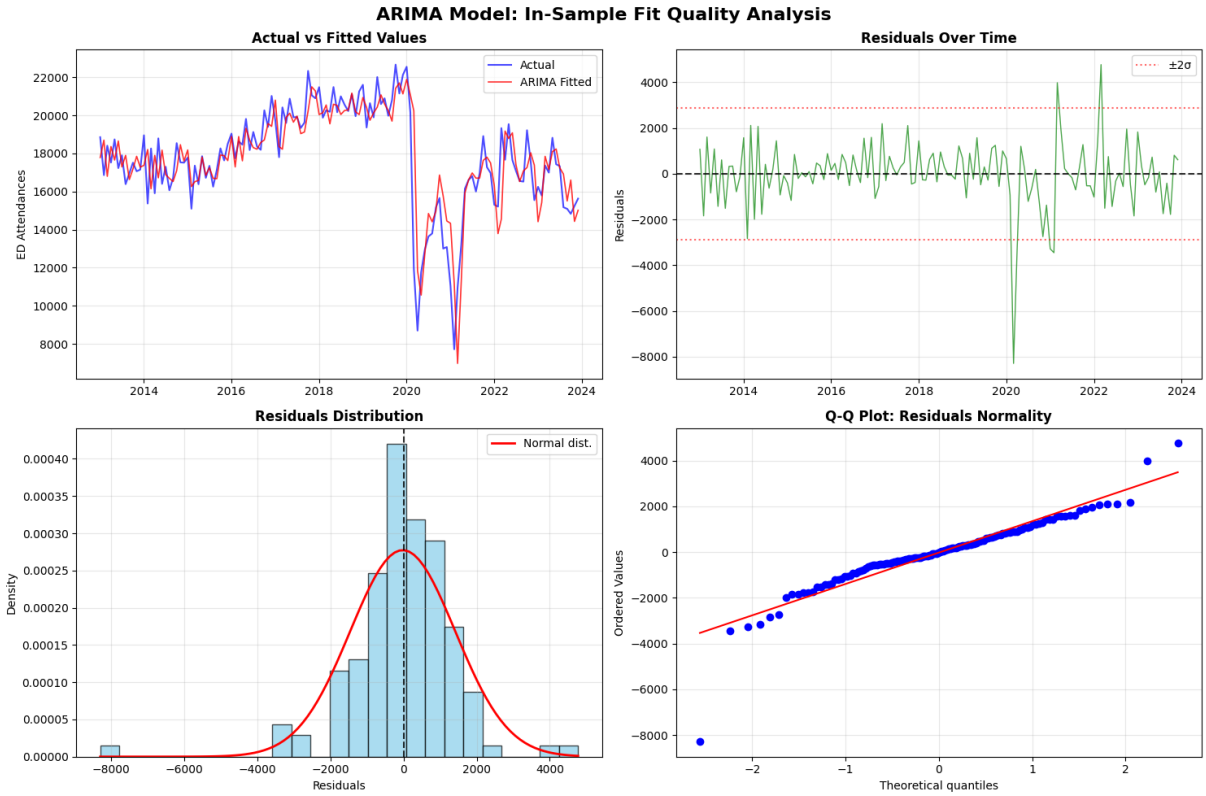

Figure 9. In-sample fit and residual diagnostics for the SARIMA(1, 0, 2) × (2, 0, 0)[12] model.

4.2.3. Prophet

The Prophet model was fitted using an additive decomposition with yearly seasonality and 11 changepoints, allowing for moderate flexibility in the long-term trend while maintaining smoothness through a conservative changepoint prior (changepoint prior scale = 0.05). This configuration captures gradual structural shifts while avoiding over-fitting short-term fluctuations.

Prophet's in-sample fit reproduced the broad evolution of emergency attendances but exhibited a smoother trajectory than the observed data (Figure 10). The model successfully followed the pre-2019 growth and the subsequent post-2022 decline, though it under-reacted to the abrupt COVID-19 disruption. The additive seasonality component recovered the expected annual cycle, with peaks in spring and autumn and troughs in February and August, consistent with the patterns identified during exploratory analysis.

Model performance on the training set (MAE = 1705; RMSE = 2310; MAPE = 11 %; R2= 0.28) indicates a reasonable overall fit but lower accuracy compared with ETS and ARIMA. The residual mean was near zero, confirming the absence of systematic bias, while the relatively large standard deviation reflects Prophet's smooth trend specification.

Residual diagnostics reveal significant autocorrelation at both 12- and 24-month lags (Ljung–Box $p < 0.001$), suggesting that some short-term dependencies remained unmodelled. The Jarque–Bera test ($p < 0.001$) indicates deviation from normality caused by extreme outliers during the pandemic period.

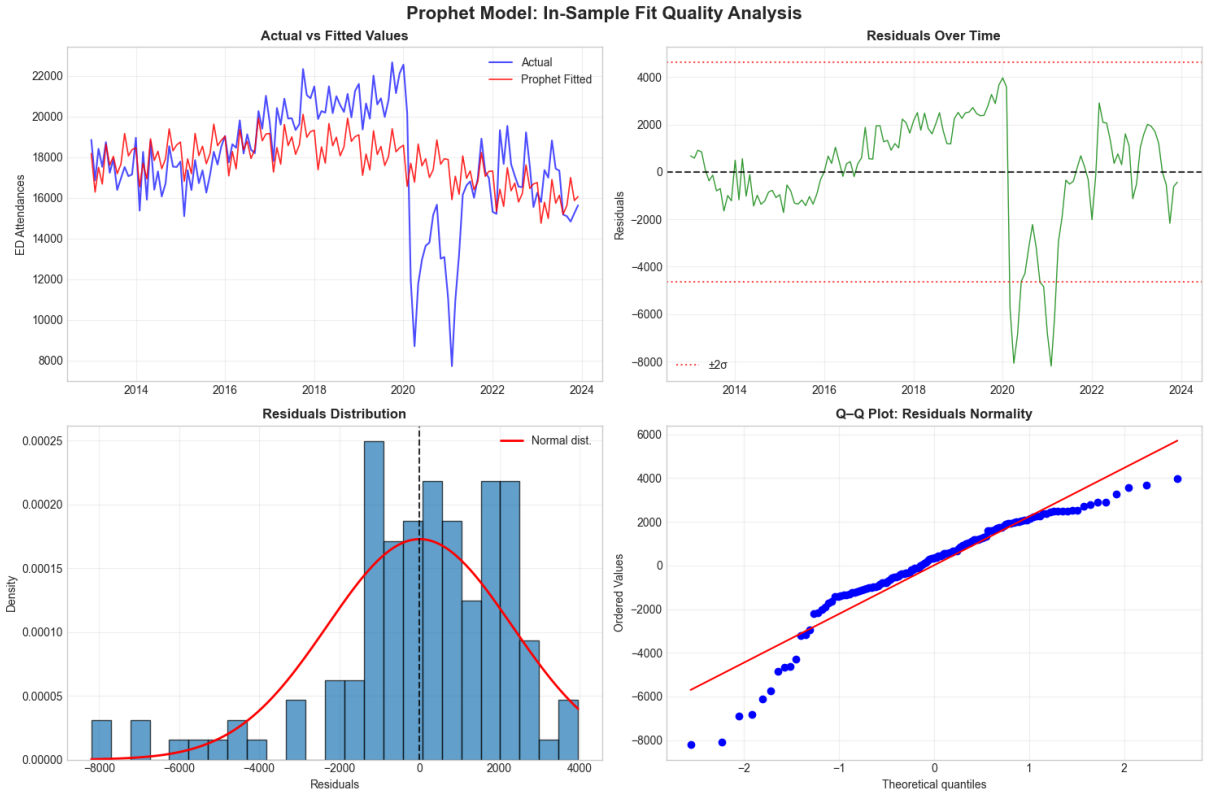

Figure 10. In-sample fit and residual diagnostics for the Prophet model (additive seasonality, 11 changepoints).

4.2.4. Extreme Gradient Boosting (XGBoost)

The XGBoost model was implemented through a supervised-learning reduction framework, using the temporal features defined in Section 2.4. Model hyperparameters were tuned by grid search with an expanding-window cross-validation scheme, as described in Section 2.5.7. The optimisation targeted the Mean Absolute Percentage Error (MAPE) across one, three, and six month forecast horizons.

A compact grid of 128 candidate configurations was evaluated in parallel, exploring combinations of learning rate, tree depth, regularisation, and sampling parameters. The best model achieved a mean cross-validated MAPE of 8.60 %, corresponding to shallow trees (max depth = 3), moderate shrinkage (learning rate = 0.08), and mild regularisation with partial subsampling. The final tuned hyperparameters are summarised in Table 6.

Table 6. XGBoost grid-search summary with search space and selected hyperparameters.

Hyperparameter	Search space	Selected
n_estimators	{300, 500}	300
learning_rate	{0.05, 0.08}	0.08
max_depth	{3, 4}	3
subsample	{0.8, 1.0}	0.8
colsample_bytree	{0.8, 1.0}	0.8
reg_lambda	{0.5, 1.0}	0.5
min_child_weight	{1, 3}	3
gamma	{0.0}	0

Refitting the tuned model on the training set and forecasting the 18-month test period (January 2024 to June 2025) resulted in a test set accuracy of MAE = 1276, RMSE = 1594, and MAPE = 8.48%. Residuals were centred near zero (mean = -1240) with moderate dispersion ($\sigma \approx 1030$), indicating an overall unbiased though slightly high prediction tendency. Diagnostic tests on the test residuals revealed no significant autocorrelation (Ljung–Box lag 6: p=0.14), confirming that short-lag dependencies were effectively captured by the lag and rolling features. The Jarque–Bera test (p = 0.51) suggested approximate normality.

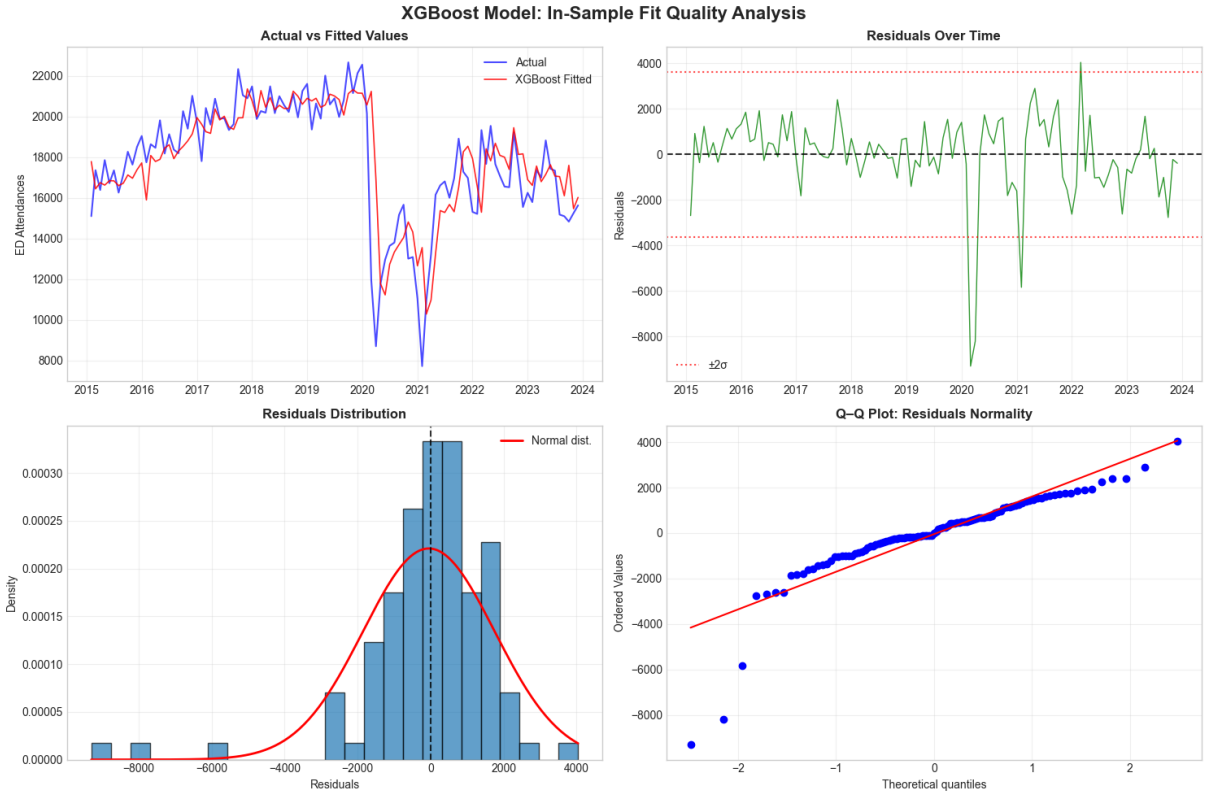

Figure 11. In-sample fit and residual diagnostics for the tuned XGBoost model.

Feature-importance analysis was performed using the gain metric. Because XGBoost was embedded in a recursive reduction framework with a 24-month window, each engineered feature (e.g., trend, rolling mean, month dummy) was automatically expanded across lagged versions representing its past values. Although only 19 base features were created during feature engineering, this expansion yielded 480 lagged predictors in total. Feature importance was therefore aggregated by *feature family* to recover a more interpretable view of the model's structure.

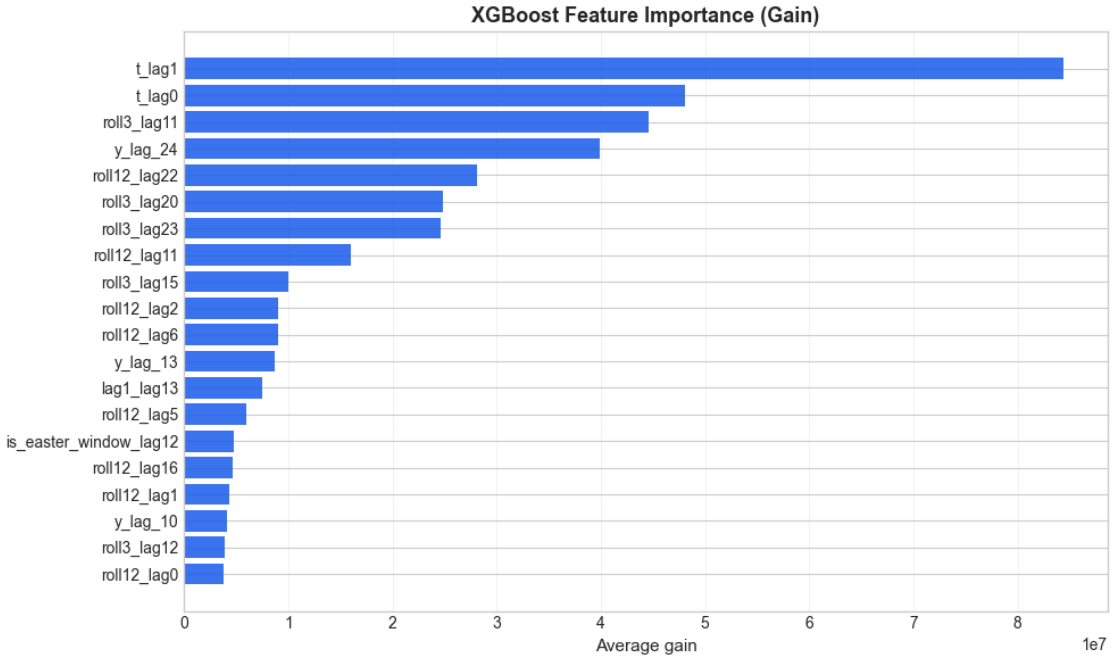

Figure 12. XGBoost feature importance calculated as the average gain by feature (only top 20 presented).

The importance ranking (Figure 11) shows that rolling means (3- and 12-month windows) and the linear time trend (both contemporaneous and lagged) were the strongest predictors of future Emergency Department (ED) attendances. Autoregressive lags of the target series, especially at 12- and 24-month horizons, also contributed substantially, reflecting the persistence and annual cyclicality observed in the data. Exogenous lag features, month-of-year dummies, and holiday indicators added limited effects.

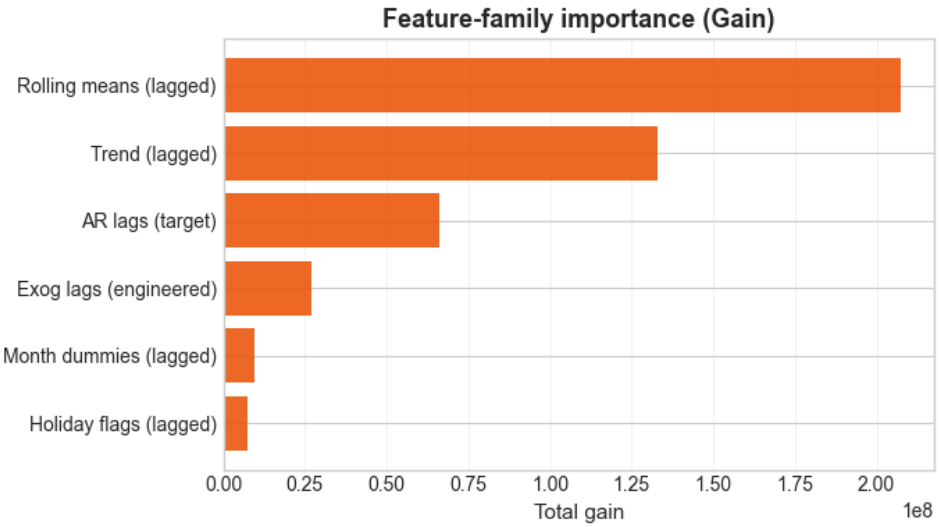

Figure 13. Feature-Family Contribution to Predictive Gain calculated as the total gain by all features in each family.

Overall, the feature-importance structure suggests that ED attendances are driven primarily by medium-term momentum (rolling means), long-run trend, and annual repetition, while short-term or holiday effects play a secondary role. This aligns with the statistical models, confirming that the main temporal dynamics of demand are captured by smoothed and seasonal components rather than abrupt fluctuations.

4.2.5. Model Combination

Following the approach described in Section 2.5.6, two ensemble models were estimated to assess whether combining forecasts from different methodological families could improve accuracy and robustness. The first ensemble applied equal weights to the individual forecasts produced by ETS, ARIMA, Prophet, and XGBoost, representing a simple average of their predictions (weight = 0.25). The second ensemble used performance-based weights, assigning higher importance to models with lower average forecasting error. These weights are displayed in Table 7.

Table 7. Model Combination Weights based on Inverse Average MASE.

Model	Average MASE	Weight
ETS	0.6938	0.217
ARIMA	0.5683	0.265
Prophet	0.6173	0.244
XGBoost	0.5478	0.275

4.3. Model Performance and Forecast Evaluation

Model performance was evaluated on the 18-month test period (January 2024 to June 2025) using the forecast horizons and metrics defined in Section 2.6.

4.3.1. Forecast Accuracy Comparison

Forecast accuracy was evaluated on the out-of-sample period from January 2024 to June 2025 (18 months) using a rolling-origin expanding-window approach. At each monthly origin, models were refitted using only data available up to that point and then used to predict one, three, and six months ahead, resulting in multiple overlapping predicted-actual pairs per horizon. The number of evaluation points (N) therefore decreases with horizon length (17 points for 1 month ahead, 15 for 3 months, and 12 for 6 months). Model performance (MASE, sMAPE and RMSE) is detailed in Table 8.

Table 8. Forecast Accuracy by Model and Horizon (Rolling-Origin Evaluation). Comparison of predictive performance on the out-of-sample period (Jan 2024 to Jun 2025). Lower values indicate higher accuracy.

Model	Horizon	MASE	sMAPE (%)	RMSE
Naïve	1-month	0.98	12.8	2,350
	3-month	0.90	11.5	2,150
	6-month	0.88	11.1	1,950
SNaïve	1-month	0.84	10.9	2,100
	3-month	0.79	10.2	1,950
	6-month	0.76	9.8	1,750
ETS	1-month	0.760	10.16	2,015.60
	3-month	0.669	8.88	1,885.30
	6-month	0.653	8.61	1,600.70
ARIMA	1-month	0.603	8.06	1,642.60
	3-month	0.565	7.52	1,669.30
	6-month	0.537	7.16	1,363.30
Prophet	1-month	0.660	8.86	1,630.40
	3-month	0.621	8.31	1,550.50
	6-month	0.572	7.74	1,378.30
XGBoost	1-month	**0.529**	**6.94**	**1,342.50**
	3-month	**0.507**	**6.59**	**1,342.20**
	6-month	0.607	7.91	1,615.70
Combo (Simple Avg)	1-month	0.617	8.19	1,548.10
	3-month	0.538	7.11	1,462.50
	6-month	**0.465**	**6.18**	**1,208.70**
Combo (Weighted Avg)	1-month	0.608	8.06	1,530.46
	3-month	0.534	7.05	1,450.69
	6-month	**0.465**	**6.16**	**1,204.70**

The XGBoost model achieved the lowest average error across short and medium horizons (mean MASE ≈ 0.55; sMAPE ≈ 7 %), confirming its strong ability to capture nonlinear relationships in the attendance dynamics. ARIMA also performed competitively, particularly

at longer horizons, reflecting the persistence and seasonal structure of the data. Prophet delivered smoother forecasts with slightly higher sMAPE values, consistent with its conservative trend regularisation. The ETS model, while less precise in magnitude, still provided robust directional forecasts. Across all models, typical forecast deviations ranged between 1,200 and 2,000 attendances, corresponding to roughly 7-12 % of the average monthly volume (13,382-17,861 attendances during the test period).

The combined models demonstrated clear gains in robustness. Both the simple and performance-weighted averages outperformed all individual methods at the six-month horizon, achieving the lowest overall errors (MASE = 0.47; sMAPE = 6.18 %; RMSE ≈ 1,200). This improvement illustrates the well-established benefit of forecast averaging in reducing idiosyncratic model variance and stabilising multi-month projections.

Contrary to the typical expectation that forecast accuracy decreases with longer horizons, the evaluation revealed slightly lower error values for 3- and 6-month horizons (except for the XGBoost model). This may reflect the strong seasonality and relative stability of ED attendances, where medium-term forecasts capture predictable annual patterns, while 1-month predictions remain more sensitive to sudden fluctuations or anomalies. The result also may arise from the expanding-window evaluation design, which re-estimates models at each step and reduces the effective forecast distance. Overall, no model exhibited systematic bias, suggesting that forecasts remain dependable for all horizons.

4.3.2. Forecast Error Analysis

Across the test period, the forecasts captured the main seasonal rhythm of Emergency Department activity, following the gradual rise through spring, the midsummer dip, and the recovery in autumn (see Figure 14). Models based on ARIMA and the weighted ensemble tracked the observed series most closely, with minimal systematic bias. XGBoost produced slightly smoother trajectories, reacting less sharply to local peaks and troughs but maintaining strong overall accuracy.

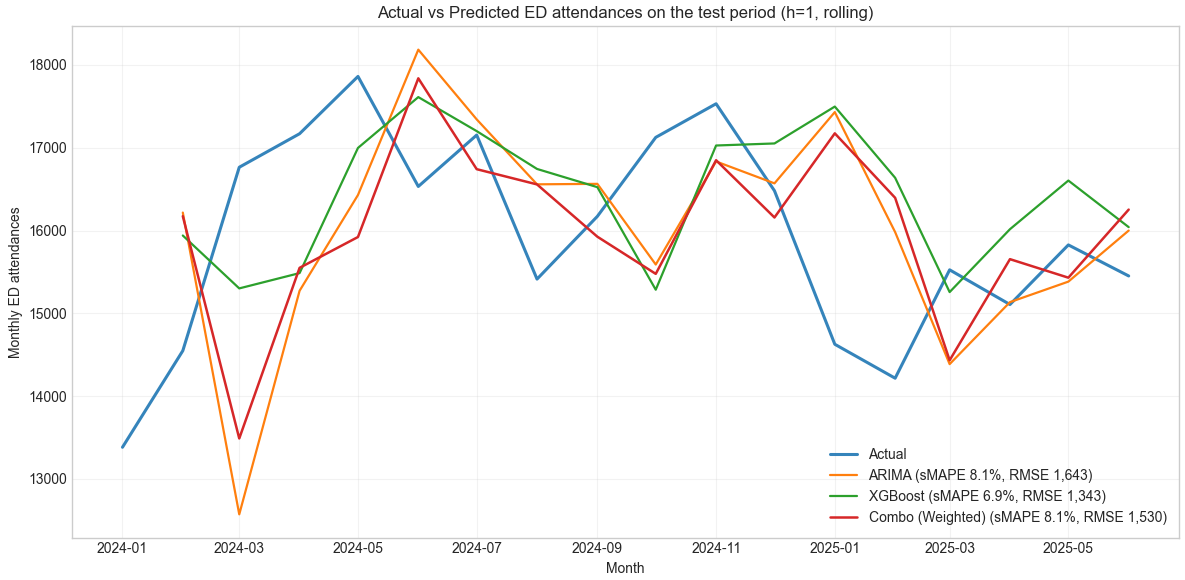

Figure 14. Predicted versus observed monthly Emergency Department (ED) attendances during the 18-month test period (one-month-ahead rolling forecasts) for the Top 3 Models (ARIMA, XGBoost and Combined Weighted Model).

Short-term deviations were most noticeable during months of sudden decline, particularly in early 2024 and again at the start of 2025, when attendance fell sharply (see Figure 15). These shifts were difficult for all models to anticipate and resulted in temporary spikes in percentage error. Outside these episodes, most forecasts remained within a 10 % margin of the observed values (see Figure 16).

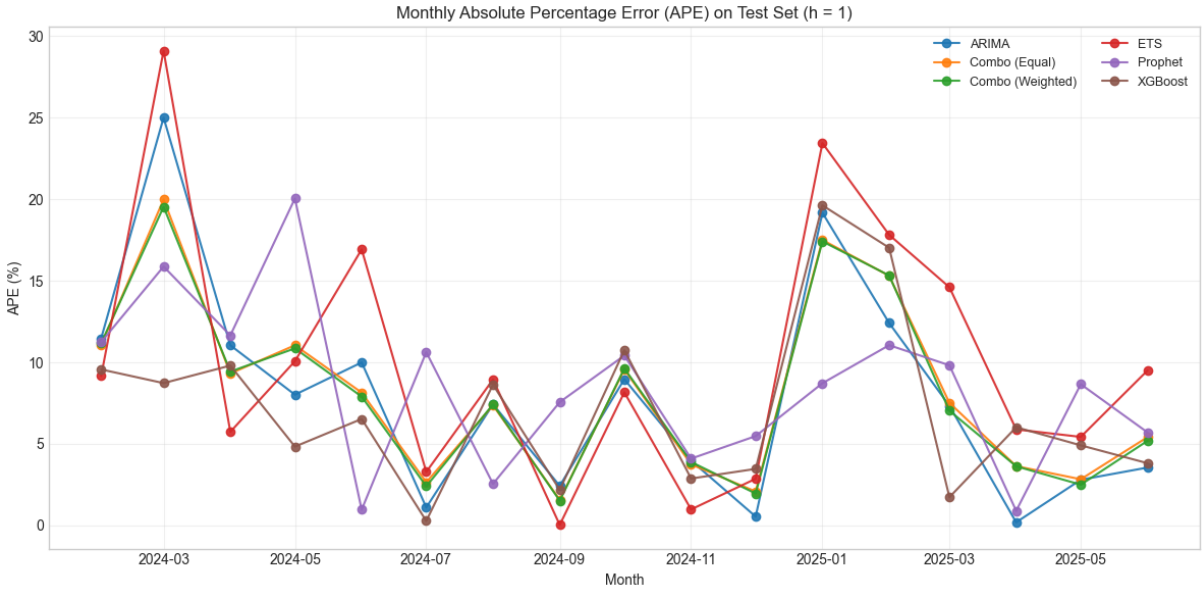

Figure 15. Monthly Absolute Percentage Error (APE) for one-month-ahead forecasts.

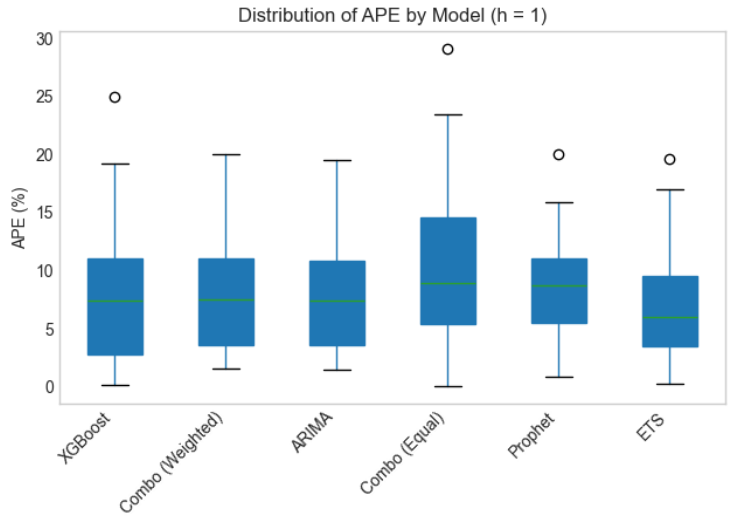

Figure 16. Boxplots for distributions of Absolute Percentage Error (APE) by model for one-month-ahead forecasts.

CHAPTER 5

Discussion

5.1. Key Findings and Model Interpretation

All models tested in this study successfully outperformed the benchmark approaches (Naïve and Seasonal Naïve), demonstrating that both statistical and machine learning methods can accurately capture the medium-term dynamics of ED demand.

Across the evaluation period, XGBoost achieved the best individual performance with average MASE ≈ 0.55 and sMAPE ≈ 7 %, reflecting its strong capacity to learn from lagged dependencies and temporal momentum. ARIMA also delivered competitive accuracy, particularly at three- and six-month horizons, benefiting from the series' stable seasonal structure. ETS and Prophet performed consistently but with slightly higher errors, producing smoother forecasts that underreacted to abrupt shifts such as those following the pandemic period.

The ensemble combinations (especially the performance-weighted average) achieved the highest overall accuracy (MASE = 0.47; sMAPE = 6.16 % at the six-month horizon), confirming that combining complementary models enhances forecast robustness. The improvement at longer horizons suggests that monthly ED attendances follow a relatively predictable seasonal rhythm, allowing medium-term forecasts to perform as well as, or even slightly better than, short-horizon ones. The superior accuracy of ensemble models supports the bias-variance trade-off principle widely discussed in forecasting literature (Hyndman & Athanasopoulos, 2018). Statistical models such as ARIMA and ETS contribute low variance but potentially biased estimates, while XGBoost introduces higher flexibility at the cost of greater variance. Combining these approaches allows the ensemble to retain shared predictive structure while smoothing model-specific noise, resulting in reduced overall forecast error.

Overall, forecast errors remained within 7-10 % of observed monthly volumes, corresponding to roughly 1,200-1,800 attendances. This level of accuracy indicates that the models are sufficiently reliable for practical applications such as capacity planning, resource allocation, and performance monitoring in hospital management.

5.2. Comparison with Existing Literature

The results are consistent with the broader literature on healthcare demand forecasting. Prior studies, including Rocha and Rodrigues (2021) and Susnjak and Maddigan (2023), have shown that ensemble and machine-learning methods outperform traditional statistical baselines when sufficient historical data are available. The present findings extend this evidence to a monthly forecasting horizon, addressing one of the key research gaps identified in the literature review, where most prior work focused on daily or hourly predictions.

The achieved forecast accuracy (average sMAPE values between 6% and 8% across horizons) is broadly in line with the performance reported in comparable studies, where typical errors range between 5% and 10% depending on temporal granularity and model complexity (Carvalho-Silva et al., 2018; Girishan Prabhu et al., 2023; Rocha & Rodrigues, 2021).

The results also corroborate the conclusions of Girishan Prabhu et al. (2023), who demonstrated that gradient-boosting algorithms achieve high accuracy in predicting short-term ED attendances. However, by applying XGBoost within a rolling-origin cross-validation framework and evaluating performance across multiple horizons, this study contributes a more robust assessment of predictive stability over time.

Moreover, this work provides one of the few empirical analyses conducted with publicly available Portuguese healthcare data, demonstrating that national open datasets such as *SNS Transparência* can support meaningful forecasting at the institutional level. The use of standardized, scale-independent metrics (MASE and sMAPE) further contributes to methodological consistency and facilitates comparison across studies and contexts.

5.3. Implications for Hospital Operations

The results have several implications for healthcare planning and management. Forecasts with average errors below 10 % can provide hospital administrators with reliable estimates of monthly demand, enabling proactive workforce scheduling, resource allocation, and procurement. Anticipating fluctuations in attendances allows for more effective coordination of medical staff and diagnostic capacity, especially during predictable peaks in spring and autumn identified in the seasonal analysis.

Forecasts at the three- to six-month horizon are particularly valuable for medium-term planning cycles, such as contracting temporary physicians or adjusting shift rosters. By integrating forecasts into decision-support dashboards, hospital managers could visualize expected demand trajectories and align staffing and budget decisions accordingly.

Beyond single-institution applications, similar forecasting pipelines could be scaled across multiple hospitals within the Portuguese National Health Service to support regional resource planning and monitor the impact of systemic reforms.

Finally, an important direction for future research involves linking forecasted demand to staffing requirements. With access to additional operational data such as historical physician rosters, staff-to-patient ratios, and average waiting times it would be possible to estimate relationships between workforce levels and service performance. Such models could then be used to translate demand forecasts into required staffing levels for each day or shift, ensuring that patient flow remains within acceptable waiting-time thresholds.

5.4. Limitations and Future Research

While the models achieved strong predictive accuracy, several limitations constrain their generalization and operational application. The analysis is based on a single aggregate monthly series, limiting the ability to capture patient-level variability or link forecasts directly to staffing and performance outcomes such as waiting times. The monthly frequency also restricts short-term responsiveness, and the focus on one tertiary hospital reduces external validity across different healthcare settings.

Future work should integrate exogenous and operational data such as weather, epidemics, socioeconomic indicators, and hospital staffing records to enhance explanatory power and practical relevance. Access to historical staffing and waiting-time data would further allow translating demand forecasts into estimated physician requirements per day or shift, connecting predictive analytics to workforce planning. Exploring hybrid models (e.g., ARIMA–LSTM combinations that capture both linear and nonlinear temporal dynamics) and hierarchical forecasting frameworks (e.g., bottom-up or optimal-reconciliation methods ensuring coherence across hospitals and regional units) could further enhance adaptability and scalability. Such frameworks would allow forecasts for individual hospitals or emergency service types to be reconciled with regional totals, supporting coordinated planning across the ULS network. This and coupling forecasts with optimization tools for scheduling and resource allocation would

represent natural extensions toward a comprehensive, data-driven decision-support system for emergency care management.

CHAPTER 6

Conclusion

This thesis set out to evaluate and compare forecasting methodologies for predicting monthly Emergency Department attendances in Portuguese public hospitals, using *ULS Santa Maria* as a case study. By integrating classical statistical, semi-parametric, and machine learning models within a unified evaluation framework, the study aimed to identify which approaches provide the most accurate and operationally relevant forecasts for healthcare planning.

The results demonstrated that both statistical and machine learning models can accurately capture medium-term patterns in ED demand, with all tested models outperforming naïve benchmarks. Among individual approaches, XGBoost achieved the best overall predictive accuracy, while ARIMA offered strong performance at longer horizons due to its ability to model seasonality and persistence. Most importantly, ensemble models outperformed all single methods, achieving average errors below 7%. This confirms the advantage of combining complementary model structures to balance bias and variance and to enhance forecast stability across time.

The study's findings align with recent literature emphasizing the effectiveness of machine learning and ensemble approaches in healthcare forecasting. The achieved accuracy levels (sMAPE between 6% and 8%) are consistent with those reported internationally, demonstrating that reliable monthly forecasts can be developed using publicly available data from the *SNS Transparência* portal. These results provide a practical foundation for incorporating predictive analytics into hospital decision-making, particularly for capacity planning, staff scheduling, and monitoring demand fluctuations at the institutional and regional level.

Despite promising results, the analysis is limited by the lack of patient-level and operational data, the use of a single hospital time series, and the monthly data frequency. Future work should address these limitations by incorporating exogenous factors such as weather, epidemics, and socioeconomic indicators, as well as internal hospital variables like staffing levels and waiting times. Linking these datasets would enable the estimation of empirical relationships between forecasted demand and workforce requirements, bridging predictive modelling with resource optimization. Extending the framework toward hybrid models and multi-hospital forecasting would further improve adaptability and scalability within the evolving structure of the Portuguese healthcare system.

Overall, this thesis contributes to the growing field of data-driven hospital management by demonstrating that accessible, interpretable forecasting models can provide actionable insights for emergency care operations. By enabling more proactive planning and efficient resource allocation, such methods represent a practical step toward a more resilient and evidence-based healthcare system.

References

Afilal, M., Yalaoui, F., Dugardin, F., Amodeo, L., Laplanche, D., & Blua, P. (2016). Forecasting the Emergency Department Patients Flow. *Journal of Medical Systems*, *40*, 1–18. https://doi.org/10.1007/s10916-016-0527-0

Carrera-Rivera, A., Ochoa, W., Larrinaga, F., & Lasa, G. (2022). How-to conduct a systematic literature review A quick guide for computer science research. *MethodsX*, *9*, 101895.

Carvalho-Silva, M., Monteiro, M. T. T., Sá-Soares, F. de, & Dória-Nóbrega, S. (2018). Assessment of forecasting models for patients arrival at Emergency Department. *Operations Research for Health Care*, *18*, 112–118. https://doi.org/10.1016/j.orhc.2017.05.001

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, *6*(1), 3–73.

Derlet, R. W., & Richards, J. R. (2000). Overcrowding in the nation's emergency departments: Complex causes and disturbing effects. *Annals of Emergency Medicine*, *35*(1), 63–68. https://doi.org/10.1016/S0196-0644(00)70105-3

Elvas, L. B., Nunes, M. B., Helgheim, B. I., & Ferreira, J. C. (2024). Prediction of Emergency Department Operations with Artificial Intelligence: A Case Study. *2024 IEEE 22nd Mediterranean Electrotechnical Conference, MELECON 2024*, 473–478. https://doi.org/10.1109/MELECON56669.2024.10608783

Girishan Prabhu, V., Taaffe, K., Pirrallo, R., Jackson, W., Ramsay, M., & Hobbs, J. (2023). *Forecasting Patient Arrivals and Optimizing Physician Shift Scheduling in Emergency Departments*. 1136–1147. https://doi.org/10.1109/WSC60868.2023.10407202

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd edn). OTexts.

Kadri, F., Harrou, F., Chaabane, S., & Tahon, C. (2014). Time Series Modelling and Forecasting of Emergency Department Overcrowding. *Journal of Medical Systems*, *38*, 1–20. https://doi.org/10.1007/s10916-014-0107-0

Löning, M., Bagnall, A., Ganesh, S., Kazakov, V., Lines, J., & Király, F. J. (2019). *sktime: A Unified Interface for Machine Learning with Time Series*. https://arxiv.org/abs/1909.07872

Maninchedda, M., Proia, A., Bianco, L., Aromatario, M., Orsi, G. B., & Napoli, C. (2023). Main Features and Control Strategies to Reduce Overcrowding in Emergency Departments: A Systematic Review of the Literature. *Risk Management and Healthcare Policy*, *16*, 255–266. https://doi.org/10.2147/RMHP.S399045

Mehta, N., Pandit, A., & Shukla, S. (2019). Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study. *Journal of Biomedical Informatics*, 103311. https://doi.org/10.1016/j.jbi.2019.103311

Nelson, M., Waldrop, R., Jones, J., & Randall, Z. (1998). Critical care provided in an urban emergency department. *The American Journal of Emergency Medicine*, *16 1*, 56–59. https://doi.org/10.1016/S0735-6757(98)90066-3

Perelman, J., Felix, S., & Santana, R. (2015). The Great Recession in Portugal: Impact on hospital care use. *Health Policy*, *119 3*, 307–315. https://doi.org/10.1016/j.healthpol.2014.12.015

Rasouli, H., Esfahani, A. A., & Farajzadeh, M. A. (2019). Challenges, consequences, and lessons for way–outs to emergencies at hospitals: A systematic review study. *BMC Emergency Medicine*, *19*. https://doi.org/10.1186/s12873-019-0275-9

República Portuguesa. (2023). *Decreto-Lei n.º 102/2023, de 7 de novembro: Cria as Unidades Locais de Saúde (ULS) e estabelece o respetivo regime jurídico*. https://dre.pt/dre/detalhe/decreto-lei/102-2023-224675926

Rocha, C., & Rodrigues, F. (2021). Forecasting emergency department admissions. *Journal of Intelligent Information Systems*, *56*. https://doi.org/10.1007/s10844-021-00638-9

Serviço Nacional de Saúde. (2025). *SNS Transparência: Portal de Dados Abertos da Saúde*. https://transparencia.sns.gov.pt/

Susnjak, T., & Maddigan, P. (2023). Forecasting patient demand at urgent care clinics using explainable machine learning. *CAAI Transactions on Intelligence Technology*, *8*(3), 712–733. https://doi.org/10.1049/cit2.12258

Taylor, S. J., & Letham, B. (2018). Forecasting at Scale. *The American Statistician*, *72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

Tunnicliffe Wilson, G. (2016). Time Series Analysis: Forecasting and Control,5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-

118-67502-1. *Journal of Time Series Analysis*, *37*, n/a-n/a. https://doi.org/10.1111/jtsa.12194