HOURLY FORECASTING OF EMERGENCY DEPARTMENT ARRIVALS FOR DIFFERENT ESI LEVELS

by

Shaghayegh Mashinkarjavan

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Engineering Management

Charlotte

2024

Approved by:

_____
Dr. Vishnunarayan Girishan Prabhu

_____
Dr. Tao Hong

_____
Dr. Churlzu Lim

ABSTRACT

SHAGHAYEGH MASHINKARJAVAN. Hourly Forecasting of Emergency Department Arrivals for Different ESI Levels. (Under the direction of DR. VISHNUNARAYAN GIRISHAN PRABHU)

Emergency Departments (ED) play a crucial role in the healthcare system, acting as the primary gateway for most hospital admissions. As primary entry points to hospital services, it is essential that EDs receive focused attention to ensure patients experience a smooth and uninterrupted healthcare journey. However, EDs face considerable challenges, with overcrowding being a major issue. Various solutions have been proposed to tackle this challenge, among which forecasting ED arrivals stands out as a foundational approach. By accurately predicting the number of patients arriving at the ED, healthcare providers can better prepare and manage resources, aiming to reduce the impact of crowding effectively.

This study advances ED arrivals forecasting by predicting hourly patient arrivals for one-hour ahead, focusing on ESI level forecasts to improve resource allocation decisions. It introduces a dynamic, rolling base method for model training, a notable improvement over the traditional static approach. The research compares the performance of widely used forecasting models with more accurate yet straightforward proposed models. The proposed forecasting framework applies Multiple Linear Regression (MLR) and develops a Hierarchical forecasting approach, with MLR as the forecasting method for top-level and three different top-down reconciliations. Proposed models are compared with some state-of-the-art models. Model accuracy is assessed using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Among all the models, the proposed model performs better for most of the ESI levels. Following this, the Diebold-Mariano test (DM test) is applied to determine if there is a significant difference in accuracy between forecasting models.

Keywords: Patient Flow Management, Emergency Department Arrivals, ED Arrivals, Forecasting, Hierarchical Forecasting, Multiple Linear Regression, Diebold-Mariano test (DM test)

DEDICATION

To my heroes in my life, my parents, from whom I learned patience,

to my dear brother and sister, Arash and Banafshe, whose unwavering love and kindness continually inspire me to accept and conquer every challenge,

and to my beloved Masoud, whose patience always strengthens me during difficult times.

ACKNOWLEDGMENTS

I am deeply grateful to my advisor, Dr. Vishnunarayan Girishan Prabhu, whose expertise and understanding profoundly enriched my master's thesis. His recent work inspired the concept for this research. He supplied valuable resources that helped me resolve any issues that arose. His unwavering support played a crucial role in the success of my research.

I profoundly appreciate the guidance and support provided by my Co-Advisor, Dr. Tao Hong. His comprehensive knowledge and meticulous attention to detail were indispensable throughout my research. However, my gratitude extends beyond this to encompass all that I learned from him during my master's program and in his classes. I am thankful to Dr. Hong for the invaluable lessons I learned from him on a personal level. Without his teachings, I believe navigating through various complexities would have been considerably more challenging.

I also wish to thank Dr. Churlzu Lim, my committee member, whose insightful questions and attention to detail greatly contributed to the improvement of the thesis. Dr. Lim was also one of my most influential professors during my master’s program, and the knowledge I gained from his classes proved invaluable in this research. Additionally, I extend my thanks to Dr. Simon Hsiang, the chair of the Industrial and Systems Engineering department, for his support throughout my master’s studies.

Finally, I want to convey my heartfelt gratitude to my husband, Masoud Sobhani, whose unwavering support and guidance were my rock during the most challenging moments of this journey. Without his love, nothing would have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

The Emergency Department (ED) is a crucial part of the US healthcare system, handling most hospital admissions. Since EDs act as the primary entry point to other hospital departments, they should receive more focused attention to ensure smooth patient flow through their patient journey. Despite being the healthcare safety net, the ED faces challenges due to the high number of visits, exceeding 151 million per year, leading to crowding [1]. According to the American College of Emergency Physicians (ACEP), ED crowding is defined as: "Crowding occurs when the identified need for emergency services exceeds available resources for patient care in the emergency department (ED), hospital, or both. The causes of crowding are multifactorial and span the entire healthcare delivery system [2].

The primary contributors to crowding include a lack of medical staff, beds, and equipment and more patients arriving than expected, resulting in very long waiting times [3]. The crowding, directly and indirectly, impacts patients' safety and the well-being of all ED staff by heightening frustration, increasing workload, intensifying stress, and leading to higher burnout rates, all of which directly influence patient safety.

Addressing ED crowding requires multifaceted approaches to resource management, as identified in various research studies [4],[5],[1]. Effective resource management encompasses not just the efficient use of medical supplies and equipment but also the smooth functioning of ED operations [6]Managing patient flow, which is heavily impacted by the number and timing of patient arrivals, is central to streamlining ED operations. Accurate forecasts of incoming patient volume are vital for optimal resource allocation and operational planning. Moreover, by predicting

patient arrivals accurately, it's possible to reduce waiting times and improve the overall efficiency of the ED, thereby enhancing patient and provider satisfaction.

Forecasting patient arrivals at the ED is essential, as distinct patterns are observed daily, weekly, and monthly. These trends significantly influence the ED's operational planning and decision-making [1]. Many studies have traditionally focused on long-term forecasts, like daily or weekly volumes [7]. However, hourly forecasting can significantly assist in addressing dynamic ED crowding, as unpredictable patient arrivals are one of the primary contributors to ED crowding.

In most cases except acute scenarios (e.g., ST-elevation myocardial infarction (STEMI)), after a patient arrives at the ED, the subsequent step is triage, where they are evaluated and assigned an Emergency Severity Index (ESI) level, which represents the patient’s severity level. Patients are then directed to a specific area within the ED based on this classification: ESI level 1 indicates a critical need for immediate attention and requires particular resources. In contrast, ESI level 5 denotes a less urgent case that can afford to wait when beds and other resources are not immediately available. The ESI levels of patients are a critical factor considered during the ED resource allocation plan, as resource requirements are significantly different for each ESI level. However, most prior studies focusing on forecasting patient arrivals to the ED have not considered ESI levels.

In summary, adopting comprehensive forecasting strategies at all levels and including patients' ESI levels in forecasts enables emergency departments to refine their planning processes, optimize staff and resource allocation, and minimize the risk of crowding. This approach is particularly advantageous in managing challenges like bed shortages, extended waiting periods, increased likelihood of medical errors, and improved performance and patient safety [1].

This research aims to predict one hour ahead of patient arrivals to the PRISMA Health Greenville Memorial Hospital ED in Greenville, SC, along with their ESI levels. For this study, we utilized three years of historical data to predict ED arrivals one hour ahead for each ESI level, focusing on forecasts covering an entire year ahead. Our approach involved comparing various time series forecasting models. There are traditional models, which are univariate, that perform as our benchmarks: 1- Autoregressive Integrated Moving Average (ARIMA), 2- Exponential Smoothing (ES). We also add two other benchmarks with a hierarchical forecasting approach and different top-level forecast methods: 3- Hierarchical forecasting with the ARIMA top-level forecast and 4- Hierarchical forecasting with the exponential smoothing top-level forecast. These two latter models (hierarchical forecasting) include three different reconciliation methods. To enhance our predictions, we introduced calendar variables into two new models. The first is a Linear Regression model that selects variables from a mix of different types.

The second is hierarchical forecasting models using linear regression for the top-level forecast with three different reconciliation methods. The final model we propose is the ensemble model, the simple average of all eight benchmarks and three proposed models. We then evaluate how well each model could predict patient arrivals one hour ahead across each ESI level, comparing their accuracy to find the most effective approach. We calculate the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) to assess accuracy. Additionally, we apply the Diebold-Mariano (DM) test to compare the models comprehensively.

In Chapter 2 of this thesis, we present a comprehensive review of recent and significant works, focusing on methodologies referenced within this study. Chapter 3 introduced the forecasting techniques we used. Chapter 4 provides an in-depth examination of the dataset employed in this research. Following this, Chapter 5 reveals the findings from applying both the

benchmark and proposed models to our case study. Finally, Chapter 6 concludes with a series of conclusions derived from the analyses conducted in Chapter 5. Additionally, it suggests options for future research to enhance the scope and depth of this study further.

2. LITERATURE REVIEW

Researchers have extensively applied a range of forecasting techniques to estimate patient arrivals at the ED over various forecasting horizons, including hourly, daily, or monthly. Forecasting methodologies such as ARMA, VARMA, Holt-Winters, linear regression, multiple linear regression (MLR), ARIMA, SARIMAX, Artificial Neural Networks (ANNs), and Recurrent Neural Networks (RNNs) have been widely utilized for predicting ED patient arrivals. The forecasting results were analyzed through various evaluation methods in each research such as Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), or Root Square Mean Error (RMSE).

This chapter summarizes some of the critical studies and research papers related to time series forecasting and techniques used to forecast emergency arrivals for EDs. We aim to explore the main studies and critical studies in these fields. We start with a review of patient flow management to identify the key factors that highlight the importance of forecasting ED arrivals. Next, we explore studies focusing on a variety of predicted variables, forecasting horizons, and independent variables. These elements are crucial in a literature review as they directly influence the outcomes and accuracy of forecasts. Finally, we concentrate on the forecasting methods applied in the literature to predict ED arrivals.

2.1. Patient Flow Management

Hoot et al. presented a review of the role of factors that may cause crowding in the literature of emergency departments, the adverse effect of crowding for both patients and hospitals and the solution for this problem. Factors that cause ED crowding are input factors (such as Influenza

season and non-urgent visits), thoughtful factors (such as inadequate staffing), and output factors (such as hospital bed shortage and inpatient boarding). They divided the effect of crowding into four significant effects: adverse outcomes (which related to the patient's situation), reduced quality (like treatment delays), impaired access (including ambulance diversion and patient elopement), and provider loser (financial effect). The solutions for Ed crowding collected by this paper are resource management, demand management, and operation research [4].

Wiler et al. reviewed different modeling approaches for managing patient flow and crowding in EDs. It investigated simulation models, queueing theory, and statistical models, assessing their ability to tackle ED operations challenges such as patient arrivals resource management [5]. Gul & Celik's review paper provides a detailed survey on statistical forecasting in emergency departments. They presented a broad literature survey on different subjects, such as methods, variables, and main challenges in ED forecasting [8].

In 2022, Prabhu conducted a significant study focused on enhancing patient flow and ED management through three innovative approaches that benefit both patient care and staff welfare. This research introduced forecasting models for estimating both daily (long-term) for 90 days ahead and hourly (short-term) for one week ahead of arrivals, including ESI levels. These predictions were utilized to optimize physician scheduling and ED shift structures using a Mixed Integer Linear Programming (MILP) model. Additionally, the study analyzed physician stress using data analytics and machine learning to identify early signs of burnout. The findings offer actionable insights for decision-making, aiming to boost ED operational efficiency, patient safety, and overall staff contentment [9].

2.2. Predicted Variable

The predicted variable refers to the target variable, the dependent variable, or the response variable. The classification of papers differs depending on the predicted variable of forecasting models. Gul & Celik represent an exhaustive review and categorized the theme of the predicted variable in hospital emergency departments into nine primary application areas: (1) ED patient demand, arrivals, visits, volume, and presentations, (2) ED patient admissions, (3) ED Length of Stay (ED LOS), (4) ED crowding, (5) utilization of ED resources, (6) ED patient wait times, (7) ambulance diversions, (8) inpatient admissions, and (9) other topics such as anomaly detection, forecasting posttraumatic stress disorder, triage forecasting, forecasting communication risks, charge prediction, forecasting patient dispositions, predicting 90-day mortality, and forecasting blood product transfusion needs in trauma patients [8]. Among all of them, the definition of ED crowding is not the same in all papers. For instance, several studies have looked at ED crowding or ED occupancy [10]. Schweigler. et al. focused on creating precise short-term predictions of bed occupancy in ED using time series modeling techniques. They labeled queuing models in their studies and concentrated on "changes" in EDs instead of on arrivals to the EDs [11]. However, ED crowding forecasting has some limitations. Green et al. 's model assumes that the arrival and service rates of the system remain constant over time, which doesn't align with the dynamic nature of EDs [12]. Hoot et al. demonstrated that patient crowding mainly depends on patient arrival, bed occupancy, acuity level, and duration of evaluation and treatment [13]. They suggested a framework for patient crowding (or patient flow). By forecasting each aspect, they aimed to achieve an overall prediction of ED crowding.

The ED arrival forecast result is considered input for models to forecast the ED occupancy. Some studies for ED occupancy forecasting used other study models for ED arrivals and just

focused on ED crowding forecasting. However, others propose methods for forecasting both ED arrivals and ED crowding forecasting.

Hertzum conducted a study to predict hourly ED arrivals and ED occupancy. The findings indicated that the models predicting patient arrivals were more precise compared to those for ED occupancy. Regression was identified as the most accurate method for forecasting hourly patient arrivals, while the ARIMA model proved to be more effective for estimating ED occupancy [14].

Most of the studies in the literature have been focused on total ED arrivals, and a few of them targeted the ED arrivals based on the stage of their severity. Jiang et al.'s study is one of the types of research that explores a deep neural network (DNN) framework for forecasting daily arrival flows and hourly arrival flows under different triage levels for a hospital in Hong Kong. The variables for their model were selected by genetic algorithm (GA), and the forecast horizon for their study was 28 days ahead. Their proposed DNN model achieved high accuracy in terms of MAPE and RMSE [15].

2.3. Forecasting Horizon

The forecast horizon is related to the distance between the last available data in historical data and the furthest point we are forecasting for each forecasting step. The forecast horizon is different from the resolution of the data. The forecast horizon for hourly data can be one day. This means that with hourly data, the forecast produces predictions for 24 hours for the future. Therefore, we can say the forecast horizon is one day. Forecasting patient arrivals has been done on different time scales in the literature, including monthly, daily, and hourly predictions. The influence of the forecast horizon on accuracy measurements is a crucial factor to consider when evaluating the performance of forecast models.

Schweigler et al. tested their proposed model in two different sizes for the forecasting horizon. The proposed model outperformed the benchmark (historical average) for both 4-hours and 12-hours ahead. However, the performance of the same model in 4-hours ahead resulted in more accurate forecasts compared to the 12-ahead forecasts, which is expected behavior in time-series models [11].

2.4. Independent Variables

Independent variables refer to variables that are used to predict the target variable. They are the input of forecasting models. As far back as 1996, Holleman et al. employed calendar variables, incorporating seasons (daily, weekly, monthly) and holidays. In 2001, Batal et al. utilized day-before and after-holiday date indicators in applying a stepwise linear regression method for daily [16]. McCarthy et al. considered holidays and a day after holidays in their Poisson model [17]. Carvalho-Silva et al. [18] and Whitt & Zhang [10] examined the fact that daily arrival totals exhibit a decrease just before and, on the holidays, while they tend to be higher than usual on the days immediately following holidays. Other studies also consider holidays and near holidays, the interaction of the month of the year and holidays, and the interaction of the month of the year and near holidays as independent variables [19].

Some studies considered exogenous variables, including weather and temperature, precipitation [10],[19],[18] snowfall [19], and features in measuring air pollution [20] for ED arrivals prediction. A recent study considered a wide range of specific variables such as average cloud coverage, average wind speed, daily change in SFC pressure, average humidity, average solar radiation, average air quality score, max Ozone concentration, air quality, and influenza rates [21]. Gafni et al. showed that most of their defined variables are more significant than calendar

variables like month or week of the year in their study [21]. Batal et al. demonstrated that high daily temperatures and snowfall significantly impact their regression model to predict daily arrivals [16].

2.5. Forecasting Method

Wiler et al. showed in their review research that to compare the models, we should first consider the forecasting output [5]. Method-wise, in the case of forecasting hospital ED's arrivals, the methods we considered are generally classified into three main types:

1- Statistical Forecasting methods include autoregressive integrated moving average (ARIMA), linear regression (LR), and exponential smoothing (SE),
2- Hierarchical forecasting with different top-level forecasting and reconciliation methods.
3- Artificial Intelligence Forecasting Techniques include artificial neural networks (ANN), Decision Trees and Random Forests, and Deep Learning. The latter have increasingly become famous for non-linear models due to their ability to capture complex patterns in data.

2.5.1. Statistical Forecasting Methods

Carvalho-Silva et al. conducted a study about forecasting one week and one month ahead daily ED arrivals for total. They compared the accuracy (MAPE) of Seasonal ARIMA (SARIMA), ARIMA, Moving Average, Holt-Winter, and Multiplicative Winter methods. The study found that the SARIMA model with weekly seasonality demonstrated significantly better accuracy [18].

One of the critical studies that considers acuity level, or the Emergency Severity Index in ED arrivals, is Sun et al.'s study in 2009. They categorized patients into three "patient acuity

categories-PAC": P1, P2, and P3, where P1 is the most acute, and P3 is the least acute. This research considers public holidays, ambient temperature, humidity, and pollution standard index (PSI), but the specific variables used varied for each PAC. For instance, public holidays were omitted for P1, and pollution was not included as a variable for P2. The research applied the ARIMA and SARIMA methods to each acuity category and the total number of patient attendances. They utilized 24 months of data for training, the subsequent six months for validation, and the following three months for testing to forecast daily ED attendance [20].

Choudhury & Urena's study rigorously evaluated a proposed SARIMA model for normality, stationarity, and autocorrelation. They also compare the result of their model with the Holt-Winters, neural network and TBATS model, which led to the best performance of the SARIMA model. However, their study did not involve various models and different parameters for each step of their forecasting dataset, and like many other studies, they applied a unique ARIMA (3,0,0) (2,1,0)[24] model for forecasting hourly data [22].

Côté et al. investigated linear regression models, including forecasting annual, monthly, daily, and hourly arrivals for two years (2007–2009). For hourly arrivals, they implemented Fourier regression to describe wavelike patterns observed in hourly ED arrivals[23].

Whitt & Zhang employed a Seasonal Autoregressive Integrated Moving Average model with external factors (SARIMAX), incorporating variables such as holidays and temperature, to forecast daily ED arrivals. Additionally, they utilized SARIMA models, regression analyses incorporating calendar and weather variables, and the Multilayer Perceptron (MLP) model, a type of artificial neural network, as part of their machine learning methodology. The SARIMAX model surpassed the performance of all other models. Then, they implemented a model combining the result of the ED arrivals forecasting, neural networks, and a periodic doubly stochastic

inhomogeneous Poisson process to predict occupancy level one hour to six hours ahead. The essential advantage of their approach was the dynamic adaptation of the training data, rather than relying on a single fixed model to forecast all the testing data. The result showed that with at least 10 weeks of rolling average training data, predicting one-hour ahead had the best accuracy (the lowest mean square error -MSE) [10].

McCarthy et al. employed a Poisson log-linear regression model to analyze low-acuity and high-acuity arrivals hourly forecasting for one year [17]. Their model's accuracy was assessed through 50% and 90% prediction intervals.

Exponential smoothing is one of the well-established methods for forecasting hourly data for ED arrivals. Morzuch and Allen (2006) applied a double exponential smoothing model for the double multiplicative seasonal data to hourly forecast for 168 steps ahead. They compared the results with their previous study, applying the standard Holt-Winters exponential smoothing method to evaluate its performance. Despite anticipating a more accurate outcome with the new approach regarding RMSE, the improvement was minimal (RMSE: 2.403 with double seasonal, compared to 2.413 with Holt-Winters). They attributed this slight difference to the stability of their ED arrival data over time [23].

As highlighted in Section 2.1, Prabhu's study partly concentrated on forecasting ED arrivals, serving as a foundation for subsequent analysis. The research employed methods such as ARIMA, SARIMA, Extreme Gradient Boosting (XGBoost), and Random Forest Regression to predict arrivals and ESI levels for both 90 days in advance using daily data and one week ahead using hourly data. The findings demonstrated that XGBoost outperformed the other models in both long-term and short-term forecasting scenarios.

2.5.2. Hierarchical Forecasting

A topic that has not been extensively studied in emergency department (ED) research is the importance of achieving consistent forecasts across different ESI levels. This involves ensuring that the forecasts for individual ESI levels align with each level's total number of patients. Inconsistent forecasts at different levels can lead to conflicting decisions and a lack of overall coherence in managing ED crowding. Hierarchical forecasting techniques aim to address this issue and make forecasts more consistent and coherent [21]. Many studies prove that reconciliation is guaranteed to improve base forecasts [24].

Reconciliation plays a crucial role in hierarchical time series forecasting. It refers to a technique to ensure consistency across different levels of a hierarchical or grouped series of forecasts. Numerous approaches have been explored in the literature to achieve data consistency and accuracy. Among these, temporal reconciliation is a major technique in the field. This approach is based on combining short-term and long-term forecasts that were introduced by Andrawis et al. They proposed a method that merged short-term and long-term forecasts, leveraging diverse information from different time scales [25]. Temporal hierarchies were introduced by Athanasopoulos et al. [26]. The structure for aggregating data addressed how to combine more frequent data points (for example, data collected monthly) into less frequent ones (for example, on a quarterly or annual basis), focusing on the progression of time. Recently, a new approach was proposed by Di Fonzo and Girolimetto named "cross-temporal" [27]. This involves ensuring that forecasts are consistent across different levels of aggregation, both in terms of cross-section (such as different regions) and temporal (like monthly to annual forecasts). The paper illustrates this with a case study on Australian GDP, demonstrating improved accuracy over traditional single-dimension reconciliation methods.

Hierarchical forecasting works for point forecasting and probabilistic forecasting. Athanasopoulos et al., using Australian GDP data, demonstrated how these methods not only produce coherent forecasts but also improve overall accuracy in both point and probabilistic forecasting frameworks [28]. Their study assumed that the top level referred to Gaussian probabilistic forecasts and the bottom levels referred to the non-parametric bootstrap method.

In hierarchical time series forecasting, disaggregation is key. The most traditional disaggregation was studied by Gross and Sohl, for sales forecasting. They examined twenty-one different proportion disaggregation methods with district approach, named from A to I, and some of them have numbers such as “I” that contain “I1”, “I2”, “I3”. Methods A and F were based on the simple average of historical proportions. They introduced three disaggregation techniques that appeared most promising in sale forecasting applications (A, F, I). Two methods, (A) and (F), are based on historical data and weighted sales averages, and (I) assigns different weights based on the correlation between lagged and current sales proportions [29]. The problem with their disaggregation methods was that they didn't account for changes at different levels during forecasting. Addressing this, Hyndman and Athanasopoulos introduced a new statistical reconciliation method for both top-down and bottom-up approaches, overcoming this limitation [30], [31], [32]. Their method is notable for its ability to use different types of initial forecasts, including those based on expert opinions. The approach delivered reconciled point forecasts at different hierarchy levels and considered the correlations and interactions among series within each level.

In recent studies, Panagiotelis et al. introduced a novel reconciliation approach using orthogonal projections. They highlighted the significance of linear aggregation constraints in reconciliation and presented its solution through a multidimensional geometric perspective, similar

to a multivariate setting [33]. Hollyman et al. mentioned that the mathematics of this approach closely aligns with forecast combination, as both depend on equations for summing random variables. This method inherits the advantages of combined forecasts, demonstrating how to compute statistically unbiased reconciled forecasts at any hierarchical level, effectively transforming simple reconciliation problems into forecast combinations [34].

A recent study focused on forecast reconciliation methods for both point and probabilistic forecasts is Rostami-Tabar and Hyndman's. They concentrated on forecasting daily ambulance demand for 84 days (12 weeks) ahead using a hierarchical approach with different forecasting methods for the top level, including Stationary (naïve model that future days similar to past days), Exponential Smoothing State Space model (ETS), Generalized Linear Model or GLM (linear regression model for non-Gaussian distributions), Poisson Regression Using TSGLM (GLM models which include auto-regression component for considering dependencies in series), and Ensemble model that contain a simple average of all the previous models. They applied the Minimum Trace (MinT) reconciliation method for the bottom-up hierarchical forecasting. The structure of the hierarchy time series contained three main groups: nested hierarchical structure based on control area, health board, and nature of the incident, which means there were 1530 time series in total. Since they considered multiple grouped attributes in their hierarchical structure and there was no unique way to disaggregate top forecasts, they did not use the top-down approach. Their research findings indicated that at a higher level of aggregation, forecast improvement with reconciliation surpasses that of bottom-level series, which are characterized by noise and minimal to no systematic patterns. Among all the proposed forecasted models, the ensemble model had better accuracy than each individual model for higher levels of hierarchy. For the bottom level, ETS outperformed.

2.5.3. Artificial Intelligence Forecasting Techniques

The neural network model is highly effective for time series forecasting. However, choosing the wrong network parameters can cause over-fitting, as pointed out by Choudhury and Urena [20]. This issue can result in the model performing well with the in-sample data it was trained on but poorly in forecasting data, resulting in less accurate forecasting [22].

Jones et al. examined different methods, including seasonal autoregressive integrated moving averages, time series regression, exponential smoothing, and artificial neural network models, to predict the number of daily patients at EDs. Data was collected from three different hospital EDs located in different locations, and they named them "facility_1" to "facility_3 in their study and did not consider the level of each ED. They made models for forecast horizons ranging from 1, 7, 14, 21, and 30 days ahead based on training data from January 1, 2005, through March 31, 2007 [19]. The authors found that sophisticated models such as artificial neural networks only slightly improve forecast accuracy (MAPE) compared to multiple linear regression with calendar variables. This obtained an acceptable accuracy for predicting the number of daily patients at EDs.

Gafni et al. analyzed that machine learning models, including random forests and gradient, boosted machines (GBM), and a hybrid model using the boosted Prophet algorithm, led to better model accuracy (RMSE) for predicting daily ED arrivals than univariant time series models [21].

In their 2013 study, Xu et al. employed a Non-linear Least Square Regression (NLLSR) framework, Artificial Neural Network (ANN), and multiple linear regression to forecast daily arrivals of two patient types (categories 3 and 4 patients, which are less critical[1]). They considered various contributing variables, including climate factors (rainfall, wind speed, temperature, humidity), weekdays, holidays, and influenza outbreaks. The results showed that the ANN method

[1] Based on triage system in Hong Kong accident and Emergency Departments, which is similar to ESI level in U.S health system.

outperformed both multiple linear regression and NLLSR approaches in terms of performance [35].

Sudarshan et al. proposed three models, Conventional Neural Network (CNN), Long Short-Term Memory (LSTM), and Random Forest (RF), to forecast ED arrivals 3-days ahead and 7-days ahead. The prediction utilized 3.5 years of ED arrivals, calendar variables, and weather information. The outcomes revealed that CNN achieved the smallest MAPE for the 3-day forecast, while LSTM performed better for the 7-day forecast [36].

To address the use of machine learning algorithms that have become prevalent in recent years, Zhang et al. introduced the application of diverse machine learning algorithms, such as LSTM. The MIC was employed to analyze intricate non-linear relationships between multiple variables in datasets. At first, they applied the Maximal Information Coefficient (MIC) for feature selection and the kernel principal component analysis (KPCA) to reduce the dimension of all the selected variables. Among all the models evaluated, LSTM demonstrated the highest performance [37]. Interestingly, the linear regression model exhibited greater accuracy in this research than ARIMA and several machine learning models.

2.6. Summary and Conclusions

Table 2-1 summarizes some of the critical reviewed studies about forecasting the number of ED arrivals.

This study addresses gaps in previous research by aiming to predict ED arrivals with three key objectives. Firstly, it focuses on forecasting one hour ahead to reduce crowding in ED in the short term and better resource management, such as estimating the number of beds that are not occupied in EDs. Secondly, forecast ED arrivals based on ESI levels of patient arrivals, which is

essential for effective triage planning, resource allocation, predicting ED occupancy, and managing patient flow. Third, this research adopts a rolling base method for model training, a more flexible approach compared to the static models often seen in other studies. We propose a forecast framework and compare the results obtained by default-setting benchmarks commonly used in healthcare with the results obtained by these accurate yet uncomplicated models.

Table 2-1: The summary of the notable reviewed studies about ED arrival forecasting.

Year	Citation	Severities Levels or Total	Forecasting Horizon	Forecast Interval	Tested Models	Independent Variables*	Best Model
2022	Prabhu	Severities Levels and Total	90 days ahead for daily data / one week ahead for hourly data	Daily / Hourly	ARIMA/ SARIMA/ XGBoost/ Random Forest/Regression	Hour of the day	XGBoost
2022	Zhang et al.,	Total	3 months ahead	Daily / Hourly	Linear regression, ARIMA, LSTM,	Hour of the day, day of the week, season of the year, holiday, temperature variables, Mean wind speed, Air quality index	LSTM
2020	Choudhury & Urena	Total	-	Hourly	SARIMA, TBATS, Holt-Winters, Neural network,	Hour of the day	SARIMA
2018	Whitt & Zhang	Severities Levels and Total	One day ahead for total Two hours ahead for severity levels	Daily / Hourly	SARIMAX, SARIMA, Regression, MLP	Hour of the day, day of the week, month of the year, holiday[2], temperature, precipitation	SARIMAX
2018	Carvalho-Silva et al.,	Total	one week ahead / one month ahead	Daily	ARIMA, SARIMA, Moving Average, Holt-Winter, Neural Network	Day of the week, month of the year	SARIMA
2017	Hertzum	Total[3]	one month ahead	Hourly	Regression for ED arrivals and ARIMA for ED occupancy	Hours of the day, Day of the week, Month of the year	ARIMA for ED arrivals
2013	Xu et al.,	Severities Levels[4]	37 days ahead	Daily	Non-linear Least Square Regression (NLLSR), ANN, and multiple linear regression	Temperature, rainfall, wind speed, humidity, day of the week, month, holidays, and influenza outbreaks,	ANN
2009	Sun et al.,	Severities Levels[5]	3 months ahead	Daily	SARIMA, ARIMA	Day of the week, Month of the year, holidays, temperature, humidity, and PSI	For P1: ARIMA, for P2 and P3: SARIMA
2008	McCarthy et al.,	Severities Levels[6]	one year ahead	Hourly	Poisson log-linear regression model	Hour of the day, day of the week, season, calendar year, holidays,	-
2008	Jones et al.,	Total	Daily forecast for 1, 7, 14, 21, and 30 days ahead for one month	Daily	Regression, SARIMA, Exponential Smoothing, ANN, Multiple linear regression	Day of the week, month of the year, holidays, temperature, precipitation	Multiple linear regression
2001	Batal et al.,	Total	3 months ahead	Daily	Regression	Day of the week, month of the year, season, holidays, snowfall (inches), temperature	-

[2] Including 1 day and 2 days after holiday, 1 day and 3 days before holiday.

[3] Different EDs but they are not categorized based on the severity levels.

[4] categories 3 and 4 patients, which are less critical.

[5] P1, P2, and P3, where P1 is the most acute, and P3 is the least acute.

[6] High-acuity (ESI_1 and ESI_2) and low-acuity (ESI_3, ESI_4, ESI_5)

3. BACKGROUND OF FORECASTING METHOD

This chapter provides an overview of time series decomposition and then some background on the techniques and methods we used in this study, including:

1. Exponential Smoothing
2. Autoregressive Integrated Moving Average
3. Multiple Linear Regression
4. Hierarchical Forecasting
5. Evaluation techniques include mean absolute error and mean absolute percentage error.

3.1. Time Series Decomposition

To study the behavior of a time series meticulously, we need to decompose it into several elements, including trend, cycle, seasonal, and error. Trend (*T*) is related to the long-term direction of the time series. Cycle (*C*) is a repeating pattern characterized by regularity but unknown and not accrued in fixed, such as frequency a business cycle. Typically, any cyclic element will be included within the trend component unless specified otherwise. Season (*S*) is a repeating pattern with a recognized periodicity, such as every 12 months per year or every seven days per week. Seasonality is fixed and has known frequency. Irregular or error (*E*) is an unanticipated and unpredictable series element [38].

Using this clarification, we can transform each time series *y* into one of the following.

$y = T + S + E$, where each component is added together; or, $y = T \times S \times E$, where the time series is the product of components; or, $y = (T + S) \times E$, which means seasonal and trend are considered additive, while the error component is treated as a multiplicative factor.

3.2. Exponential Smoothing (ES)

Exponential smoothing has been found to have extensive application in forecasting. This method is used to predict the behavior of a time series based on the weighted average of the behavior of past data where a smaller weight is put on older data. The nearer the data is to the forecasting target, the more weight will be placed on this method. Equation 3-1 shows the one step ahead forecasting for the time *t+1* for the time series y, which equals a weighted average of all the observations in series $y_1, y_2, \ldots, y_t$.

$$\hat{y}_{t+1|t} = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \cdots + \alpha(1-\alpha)^{t-1} y_1 \qquad 3\text{-}1$$

Where $\hat{y}_{t+1|t}$ is the forecasted value of one-step ahead for the time series y for the time *t+1*, and $0 \leq \alpha \leq 1$ is the smoothing parameter. The formula shows that the weight for the time series in time *t* has the largest value, and the oldest observation (y_1) has the smallest weight value [38].

Hyndman et al. structured exponential smoothing methods and demonstrated the State Space Models model as a unique solution based on the characteristics of Error (*E*), Trend (*T*), and Seasonality (*S*), known as the ETS model, that underlie exponential smoothing methods.

ETS model considers five different time series types based on the nature of their trend: 1- time series without any trend, or time series with a fixed level of trend without any growth; 2- additive trend; 3- damped additive; 4- multiplicative trend; 5- damped multiplicative trend. Similarly, based on the seasonality of the time series, three types of seasonality for the ETS model are 1- non-seasonal, 2-additive, and 3- multiplicative. Finally, based on the types of error time, aeries can have additive or multiplicative errors [39]. This leads to 30 different models for ETS. The point forecasts will be identical when comparing the point forecasts produced by additive and

multiplicative error components using the same smoothing parameter values. Table 3-1 shows all the ETS models for additive error.

Table 3-1: ETS models for additive error [39]

Trend component	Seasonal component		
	N (None)	A (Additive)	M (Multiplicative)
N (None)	N, N	N, A	N, M
A (Additive)	A, N	A, A	A, M
A_d (Additive damped)	A_d, N	A_d, A	A_d, M
M (Multiplicative)	M, N	M, A	M, M
M_d (Multiplicative damped)	M_d, N	M_d, A	M_d, M

Here, we describe one of the models in Table 3-1 used in this research: ETS (A, N, A) or Holt-Winters' Seasonal method. This model is applied for time series with additive error, non-trend, and additive seasonal components. The general ETS model for this time series has the following formula:

$$y_t = f(l_{t-1}, s_{t-m}, \varepsilon_t) \quad 3\text{-}2$$

Where f is an additive function, l_{t-1} is the level component at time *t-1*, s_t is the seasonal component at time *t*, and *m* donates the number of the seasonality and ε_t is the error term. A quick description of the level implies an average value per time period (which is different from the trend that shows the change in the value). We can write the formula 3-2 in another way:

$$y_t = l_{t-1} + s_{t-m} + \varepsilon_t \quad 3\text{-}3$$

$$l_t = l_{t-1} + \alpha\varepsilon_t \quad 3\text{-}4$$

$$s_t = s_{t-m} + \gamma\varepsilon_t \quad 3\text{-}5$$

Both α and γ are constants; α is the smoothing parameter for the level, and γ is the smoothing parameter for the seasonality.

Forecast for h-step ahead for time *t+1* with *ETS (A, N, A)* can be written in the following:

$$\hat{y}_{t+h|t} = l_t + s_{t+h-m(k+1)} \quad 3\text{-}6$$

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)l_{t-1} \quad 3\text{-}7$$

$$s_t = \gamma(y_t - l_{t-1}) + (1-\gamma)s_{t-m} \quad 3\text{-}8$$

Where *k* is the integer part of *h-1/m*, l_t denotes the series level at time *t*.

All the constants and smoothing parameters (for *ETS (A, N, A)* smoothing parameters are α and γ) are estimated by maximizing the "likelihood" (MLE). By MLE, the probability of obtaining the training data is maximized. Maximizing the likelihood for an additive error model has a similar result as minimizing the sum of squared errors (SSE) [40].

Model selection for the ETS model can be done using all the information criteria (AIC, AICc, and BIC). For ETS models, Akaike's Information Criterion (AIC) is defined as:

$$AIC = -2log\ (L) + 2k \quad 3\text{-}9$$

L is the likelihood of the model, and *k* is the number of parameters and initial states we need to be estimated. The AIC_c or corrected AIC is considered for small sample sizes and defined as:

$$AIC_c = AIC + \frac{2k(k+1)}{N-k-1} \quad 3\text{-}10$$

Where *N* is the number of observations used for estimation (training data).

The Bayesian Information Criterion (BIC) is another statistical criterion used in model selection and statistical modeling. The BIC aims to find the model that maximizes the likelihood while penalizing complex models. It is defined as:

$$BIC = AIC + k[log\ (N) - 2] \quad 3\text{-}11$$

For all AIC, AIC_c, and BIC, lower values indicate a better trade-off between fit and complexity, making models with lower AIC, AICc, and BIC referable [40].

3.3. Autoregressive Integrated Moving Average (ARIMA)

One univariate time series forecasting model is ARIMA, which applies to stationary or non-stationary time series. The ARIMA model is a combination of three main components. This is achieved by selecting the appropriate values for its three parameters (*p*, *d*, *q*). Accurate choices for *p*, *d*, and *q* are crucial for effectively capturing cyclic patterns and seasonality in the data. Here is a brief overview of the parameters in ARIMA modeling:

1- AutoRegressive component (AR) shows the relationship between current and lagged observations. The AR with *pth* order autoregressive component displays the current observation as a linear function of p-lagged values. Therefore, *p* determines the order of AutoRegressive components in the ARIMA model.

An autoregressive model of order *p* can be written as follows.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \quad 3\text{-}12$$

Where *c* is a constant term or drift, ϕ_i (*j=1, …, p*) is the coefficient of the autoregressive term for lag *i*, and ε_t is white noise.

2- —Integrated component (I): This refers to differencing the time series data to achieve stationarity,y where the statistical properties of time series (such as mean or variance) do not change over time. The number of differencing needed to make time series stationary is shown by (the *d*).

3- Moving Average component (MA) represents the relationship between the current observation and a residual error from the past parameter. (q) indicates the number of lagged residuals used in the model. The MR model of order q is a regression model that uses past forecast errors to forecast y_t.

$$y_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} \qquad 3\text{-}13$$

Where c is a constant, θ_k is the moving average coefficient for lag k, and ε_t is white noise.

The initial stage in constructing an ARIMA model involves assessing data for stationarity and checking the normality of its distribution. In the case of a non-stationary time series, the initial step involves eliminating the trend and seasonality, which represent systematic patterns to transform the time series into one without recognizable and predictable patterns. It will be achievable by doing differencing. This shows us that for estimating the parameters of the ARIMA model, we need to determine the number of differencing (d) before finding p and q. Therefore, ARIMA (p, d, q) modeling integrates differencing with autoregressive (AR) and moving average (MA) components, and we can write the ARIMA (p, d, q) for y'_t as the differenced time series. As a result, ARIMA (p, d, q) incorporates both past values of y_t and lagged errors.

$$y'_t = c + \underbrace{\phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p}}_{Auto\ Regressive\ (AR)\ component} + \underbrace{\theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}}_{Moving\ Average\ (MA)\ component} + \varepsilon_t \qquad 3\text{-}14$$

The Seasonal ARIMA model (SARIMA) is used when the time series shows seasonal patterns. SARIMA combines non-seasonal and seasonal factors in a multiplicative approach, represented as SARIMA (p, d, q)(P, D, Q)$_m$, where p, d, q, are the same as ARIMA model and:

P: indicates the seasonal AR order.

D: represents seasonal differencing.

Q: stands for the seasonal MA order.

m: represents the number of periods per season.

For seasonal time series, first, we determine the order of *D*. Finding the "Seasonal Strength" determines the appropriate number of seasonal differences. A high variance in the seasonal component of a time series (S_t) indicates that seasonality strongly affects the non-stationary nature of the time series.

$$F_S = max\left(0{,}1 - \frac{(R_t)}{(S_t + R_t)}\right) \qquad 3\text{-}15$$

When the variance of the seasonal component is large (ratio=0), it means that the seasonal component (F_S) is fully active with a value of one. For time series with F_s less than 0.6, the seasonal difference is insignificant; therefore, it does not account for the seasonal difference (*D=0*). Otherwise, one seasonal difference is suggested (*D=1*).

If the time series is non-seasonal but is non-stationary, then only the number of differences (*d*) is needed to obtain a stationary time series[40]. With the Unit Root test, the number of differencing will be accessible. The most essential unit root test techniques are the Augmented Dickey-Fuller Test, the Augmented Dickey-Fuller Test (ADF), the Phillips-Perron (PP) test, and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test. For our analyses, we applied the KPSS test where the null hypothesis or H_0 is that the time series is stationary. If H_0 is rejected, then we need differencing [41].

After choosing *d* and *D*, initial values for *p*, *q*, *P*, and *Q* are considered by different variations of them. These values can be determined by ACF and PACF plots. After that, the

parameters of c, ϕ_1, …, ϕ_p and also, θ_1, …, θ_q need to be determined for different models. This parameters estimation can be done by MLE, which is similar to minimizing the least square estimation. After estimating all the parameters, model selection can be done using an information criterion such as AIC:

$$AIC = -2\,log\,(L) + 2(p + q + P + Q + k + 1) \quad 3\text{-}16$$

K is the constant. If in formula 3-4, $c = 0$, then *k=0*, if not ($c \neq 0$), then *k=1*.

(p+q+P+Q+k+*1*) is the number of parameters in the model, including *c*. This is the penalty part for comparing different models. *L* is the maximized likelihood of the model fitted to the differenced data set. The corrected AIC and BIC can be written as:

$$AICc = AIC + \frac{2(p + q + k + 1)(p + q + k + 2)}{N - p - q - k - 2} \quad 3\text{-}17$$

$$BIC = AIC + [log\,(N) - 2](p + q + k + 1) \quad 3\text{-}18$$

AIC, AIC_c, or BIC minimizing will lead to proper models.

Adding a "drift" term to an ARIMA model means including a linear trend. For instance, consider the model ARIMA (1,1,0) with drift. This model has one autoregressive term, one differencing operation, no moving average term, and a linear trend. This choice indicates the model's capability to capture a long-term trend observed in the training data (related to forecasting data points). According to Equation 3-14, this model can be shown mathematically as the following equation:

$$Y_t' = c + \emptyset_1.(Y_{t-1}' - c) + \epsilon_t \quad 3\text{-}19$$

Where:

Y_t' is the differenced time series at time *t*,

c is the constant term (drift),

$\emptyset_1$ is the coefficient of the autoregressive term for lag1,

Y_{t-1}' is the differenced time series at the previous time step,

ϵ_t is the white noise error term at time t.

3.4. Multiple Linear Regression

Multiple Linear Regression (MLR) is a fundamental and widely used model in forecasting practices. MLR is a statistical technique that models the relationship between more than one independent variable or predictor and the dependent variable or response variable by fitting a linear equation to the observed data. Independent variables can be both quantitative and categorical variables. Correlation between variables can include both main effect and cross effect.

The general K-variables linear regression is:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots + \beta_k X_{kt} + u_t \qquad 3\text{-}20$$

Where the intercept β_1 to β_k are unknown parameters, X_1 to X_k are known independent variables, and u_t is the random or nonsystematic component with zero mean and unknown variance σ^2 or independent normally distributed random variable $N(0, \sigma^2)$. Unknown variables are named Regression Coefficients and estimated using the Ordinary Least Squares (OLS). OLS minimizes the sum of squared errors (SSE). This general MLR is linear in the parameters (β) [42]. We interpret β_k as the average effect on Y_t of one unit increase in X_k, holding all other variables fixed.

Finding the right fit for the model involves a trade-off between how well it performs and how complex it becomes. Therefore, the number of regressors plays an essential role in MLR. The dependent variable is often only associated with a subset of the predictors. There are various methods to select predictors. One of the methods is "stepwise regression," which can be done backward, forward, or hybrid. Backward stepwise regression modeling starts with the model with all the potential predictors and removes one predictor each time, keeping the model if it improves the forecasting accuracy. Forward stepwise regression starts with modeling with limited predictors and adds one predictor each time. The model is kept if it has improvement in accuracy. These steps continue until no further improvement. Hybrid stepwise regression modeling adds a predictor backward and drops the predictor in forward stepwise regression when considered together.

To select the best model to forecast, we can directly estimate the error of the validation set. For this purpose, predicted y for validation data can be generated by applying the estimated coefficients within the regression equation for training data and assuming the error term is zero. Then, based on best-estimated coefficients and setting the error term to zero, y for validating the data set is predicted.

3.5. The Hierarchical Time Series Forecasting

A hierarchical time series is a collection of time series organized in a hierarchical aggregation structure. Time series is arranged into various levels based on factors such as variable categories, time periods, or regions.

Hierarchical forecasting is the methodology used to generate predictions for such hierarchical time series. Figure 3.1 shows a single-level hierarchical time series structure. Each

row is one level, and the node at the top level is the aggregation of child nodes. Therefore, we have the following equations for this single-level time series.

$$y_{total} = y_A + y_B + y_C \tag{3-21}$$

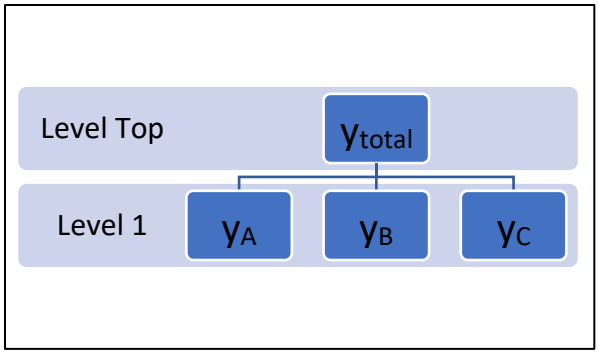

Figure 3.1: A schematic diagram of a single-level hierarchical time series

The matrix form for equation3-21 can be written as:

$$\begin{bmatrix} y_{total} \\ y_A \\ y_B \\ y_c \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} y_A \\ y_B \\ y_C \end{bmatrix} \tag{3-22}$$

The goal of hierarchal forecasting is to enhance the forecasting accuracy for a specific level or all levels within the hierarchy. A method employed to forecast an aggregation level or different levels ensures that the forecasts remain consistent and "coherent," meaning predictions for aggregates should align with the sum of the corresponding disaggregated forecasts [24].

Hierarchal forecasting methods are broadly categorized into two major approaches: "bottom-up" and "top-down." Bottom-up methods commence by generating forecasts independently for the lowest-level or bottom-level components within the hierarchy. These individual forecasts are then

aggregated upwards, leading to the forecast for higher-level series, ultimately resulting in the "Total" series (y_{total}) at the top level of the hierarchy. In "top-down" approaches, the initial step is to create a forecast for the highest-level series, typically denoted as the "Total" series (y_{total}), at the top of the hierarchy. Then, these forecasts are disaggregated and broken down into lower-level components. These methods ensure the forecasts are consistent and aligned at each hierarchy level.

The mathematical form of the hierarchy forecasting with the bottom-up approach for single level hierarchy in Figure 3.1 is described in the following. Considering the h-step ahead forecast for each bottom-level time series at time *t*: $\hat{y}_{A(t+h|t)}$, $\hat{y}_{B(t+h|t)}$, $\hat{y}_{C(t+h|t)}$, then the coherent linear h-step ahead forecast for the time *t* for the "Total" time series ($\tilde{y}_{t+h|t}$) will be:

$$\tilde{y}_{t+h|t} = \hat{y}_{A(t+h|t)} + \hat{y}_{B(t+h|t)} + \hat{y}_{C(t+h|t)} \qquad 3\text{-}23$$

For a top-down approach, let $\hat{y}_{(t+h|t)}$ be the vector of the initial h-step ahead forecast at time *t* for the "total" time series. Then, we have the following equation for each bottom-level time series in Figure 3.1.

$$\tilde{y}_{A(t+h|t)} = p_1\hat{y}_{t+h|t} \text{ , } \tilde{y}_{B(t+h|t)} = p_2\hat{y}_{t+h|t} \text{ , } \tilde{y}_{C(t+h|t)} = p_3\hat{y}_{t+h|t} \qquad 3\text{-}24$$

Where p_j (here, $j=1,2,3$) are the proportion that distribute the forecast of the top level to the bottom level, and $\tilde{y}_{A(t+h|t)}$, $\tilde{y}_{B(t+h|t)}$, $\tilde{y}_{C(t+h|t)}$ are the h-step ahead forecast for time series A, B, and C in the bottom level.

The main question for the top-down approach is how to efficiently conduct the proportions of forecast from the top to the bottom level of hierarchical data. We previously mentioned the Gross-Sohl reconciliation in Section 2.5.2, and now we will explain two methods, A and F, in more detail for single level hierarchical data.

1- An average of historical proportions, labeled as "top-down Gross-Sohl method A" or "tdgsa"[7], for j=1,…,m, number of time series in the bottom level, this method calculates the average historical proportions by the formula,

$$p_j(t) = \frac{1}{N}\sum_{t=1}^{N}\frac{y_{j,t}}{y_{j,total}} \qquad 3\text{-}25$$

Where *N* is the size of the historical data set, $y_{j,t}$ is the value of the bottom level series at time *t*, y_{total} is the aggregate value of all the time series for the time t in the bottom level, which for single-level hierarchical time series is equal to top level time series.

2- The proportion of the historical average, labeled as "top-down Gross-Sohl method F" or "tdgsf"[8], for j=1,…,n number of time series in the bottom level, this method determines the proportion of the historical average by the formula,

$$p_j(t) = \frac{\sum_{t=1}^{N}\frac{y_{j,t}}{N}}{\sum_{t=1}^{N}\frac{y_{j,total}}{N}} \qquad 3\text{-}26$$

Where *N* is the size of the historical data set, $y_{j,t}$ is the value of the bottom level series at time *t*, y_{total} is the aggregate value of all the time series for the time t in the bottom level, which for single level hierarchical time series is equal to top level time series.

These two approaches for determining the proportion lead to different results for the bottom level. However, since these two approaches do not consider the change in the bottom level of the time series, it may be less accurate to forecast for the lower level, especially if data have significant changes in the forecasting window [32]. To resolve this challenge, calculating the proportions based on the forecasted proportions of the bottom level series is considered. Against two other

[7] This method is labeled as “Top-down HP1” in [32]
[8] This method labeled as “Top-down HP2” in [32]

approaches in which we do not need the forecasting for the bottom level, this method starts with the initial forecast that refers to making h-step-ahead forecasts for all series (total time series in the top level including all the time series in the bottom level). For one-level hierarchy, this process is obtained by the following equation:

$$p_j(t) = \frac{\hat{y}_{j,t,h}}{\sum_{j=1}^{m} \hat{y}_{j,t,h}} \quad 3\text{-}27$$

Where $\hat{y}_{j,t,h}$ is the forecasted value for h-step ahead of the bottom at time *t*, for *j=1,…,m* number of time series in the bottom level. This method is considered "top-down forecast proportions" or "tdfp"[9].

For top-down approaches to single level hierarchy, the final top-level forecasts are equal to the initial top-level forecasts. Therefore, we can summarize the forecasting for time series for h-step ahead ($\tilde{y}_{j,h}$) in Figure 3.1 with tdfp- top-down forecast proportions- according to the .following equations

$$\tilde{y}_{A(t)} = \left(\frac{\hat{y}_{A(t+h|t)}}{\hat{S}_h}\right)\hat{y}_{t+h|t} \quad 3\text{-}28$$

$$\tilde{y}_{B(t)} = \left(\frac{\hat{y}_{B(t+h|t)}}{\hat{S}_h}\right)\hat{y}_{t+h|t} \quad 3\text{-}29$$

$$\tilde{y}_{C(t)} = \left(\frac{\hat{y}_{C(t+h|t)}}{\hat{S}_h}\right)\hat{y}_{t+h|t} \quad 3\text{-}30$$

Where $\hat{y}_{t+h|t}$ is the h-step ahead forecasted for the Total time series, $\hat{S}_h$ is equal to the sum of all the forecasted time series at the bottom level. If the bottom level includes m time series, then:

[9] This method is labeled as “Top-down FP” in [32]

$$\hat{S}_h = \sum_{j=1}^{m} \hat{y}_{j,h} \quad 3\text{-}31$$

3.6. Ensemble Model

An Ensemble is a composite model that combines multiple individual models to improve forecast accuracy over any single model in the ensemble. Different models can identify unique patterns, trends, and connections within the data. Combining these models allows us to take advantage of their strengths and minimize their weaknesses, improving overall performance. There are several methods to produce ensemble models, including simple averages and weighted averages. Simple average is one of the models of ensemble methods commonly used for forecasting. A simple average improves the accuracy of each forecasted method by reducing individual models' bias in many cases [43]. A simple average model is a straightforward approach that averages the forecasts from multiple models, giving each model's forecast equal weight. This method uses a linear formulation for a simple average, where a vector of n forecasts, denoted as f, is combined with the same weights.

$$F_{ensemble} = \frac{1}{n}\left(f_1 + f_2 + \cdots + f_n\right) \quad 3\text{-}32$$

3.7. Evaluation Techniques

There is no “best” forecasting model. To find a better forecasting model, we can see the forecast error loss differential and then compare the accuracy of different models. Various measurement scales are employed in point load forecasting to assess forecasting accuracy and find the better model by comparing their accuracy. Standard evaluation measures include Absolute Error (AE), Percentage Error (PE), Mean Absolute Error (MAE), Root Square Mean Error (RSME), and Mean Absolute Percentage Error (MAPE). MAE provides results in the same units as the data, not percentages, offering a more precise measure of error magnitude. It works by calculating the absolute differences between actual and predicted values, known as absolute errors (AE), and then averaging them across the dataset.

$$MAE = \frac{1}{n}\sum_{t=1}^{n} | Actual_t - Predict_t | \quad (3\text{-}33)$$

Where n is the size of the time series we forecast.

RMSE is another scale-dependent measurement that measures the average magnitude of the errors between predicted and actual values, considering both the size and direction of the errors. A lower RMSE suggests that the model's predictions are, on average, closer to the actual values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (Actual_t - Predicted_t)^2} \quad (3\text{-}34)$$

Models exhibiting lower MAE and RMSE values are preferred because they embody fewer errors. Considering Formula 3-33 and 3-34, it is clear that when AE is large, or there are outliers in the data set, the RMSE impacts more than MAE [44]. This means that when errors are

relatively small, the RMSE and MAE yield similar results. MAE and RMSE are evaluation metrics that pose no issues, even if the actual data contains zero values. Therefore, for this research, we calculate both RMSE and MAE to evaluate the forecasting method.

3.8. Diebold-Mariano (DM) test

We can compare the accuracy of two forecasting models by standard evaluation metrics. For instance, the model with a smaller MAE or RMSE is considered more precise. But suppose the differences between the two models are not too much. In that case, we need to compare forecasts and find if the differences in forecast errors between the two models hold statistical significance [45]. In this context, the Diebold-Mariano (DM) test is used to find if the results of the two forecasts are significantly different. The test allows different loss functions to measure forecasting errors, such as squared or absolute errors. The DM test can apply to various error metrics, including MAE or RMSE. To apply the test, the DM statistic is computed by 3-35:

$$\mathrm{DM} = \frac{\bar{d}}{\sqrt{\frac{\hat{\sigma}_{\bar{d}}^2}{N}}} \tag{3-35}$$

Where $\bar{d}$ is the mean difference in the forecasting errors of the two models over N forecast horizons or number of observations, $\hat{\sigma}_{\bar{d}}^2$ is the estimated variance of the differences in forecasting errors.

The Diebold-Mariano (DM) statistic is assumed to follow a normal distribution under the null hypothesis due to the central limit theorem. This means that with a large sample size, the distribution of the sample mean will be approximately normal, regardless of the original distribution of the population. Therefore, the null hypothesis for the DM test is considered: there

is no difference in the predictive accuracy of the two models. This means that the mean difference in their forecasting errors equals zero. Therefore, we can reject the null hypothesis at the 5% level if $|DM| > 1.96$.

4. DATA

The data used for the thesis include the number of emergency arrivals to the five-level emergency department (ED) of PRISMA Health Greenville Memorial Hospital (GMH), Greenville, SC. A five-level ED is following up on the Emergency Severity Index (ESI). ESI for PRISMA includes a triage algorithm that plans clinical classification for patients into five groups from the most urgent situation (ESI_1) to low severity situation (ESI_5). ESI_1 refers to patients in critical condition requiring immediate intervention. ESI_2 refers to patients who may be unstable and need to be seen quickly by a physician. ESI_3 refers to patients who are stable and require treatment within 30 minutes. Patients categorized under ESI_4 are stable and do not need immediate, urgent care. Patients in ESI_5 are highly stable and can be treated non-urgently; they usually do not need tests and are often discharged on the same day.

Hourly emergency arrivals for four years from January 2017 to the end of December 2020 for each ESI and totals were exported from the Electronic Health Record (EHR) database of PRISMA hospital and used as input for this research. Variables used for this research are 1-time, which includes the year, month of the year, day of the week, and hour of the day; 2-historical ED arrival, an integer number.

Table 4-1, including the Total arrivals time series, represents all the ESIs for emergency arrivals to the five-level emergency department at PRISMA Health Greenville Memorial Hospital. For each ESI_j (where j=1, 2, 3, 4, 5), the ED arrivals for the hour t from 2017 to December 2020 are represented as $A_j(t)$. A time series, labeled as $A_{total}(t)$, is formed by aggregating the "Total" values for all ESIs from 2017 to December 2020, with *t* representing the time points (t=1, 2, ..., T).

Table 4-2 shows some basic statistics for each time series.

Table 4-1: Data availability and severity level of each emergency department.

ESI	Time Span	Description
ESI_1	Jan 2017- Dec 2020	Hourly emergency arrivals to ESI_1
ESI_2	Jan 2017- Dec 2021	Hourly emergency arrivals to ESI_2
ESI_3	Jan 2017- Dec 2022	Hourly emergency arrivals to ESI_3
ESI_4	Jan 2017- Dec 2023	Hourly emergency arrivals to ESI_4
ESI_5	Jan 2017- Dec 2024	Hourly emergency arrivals to ESI_5
Total	Jan 2017- Dec 2025	Total hourly emergency arrivals

Table 4-2: Statistics of the data for each ESI.

Time series	Minimum Value	Maximum Value	Rounded Mean	Standard Deviation
$A_1(t)$	0	5	0	1
$A_2(t)$	0	11	2	2
$A_3(t)$	0	19	4	3
$A_4(t)$	0	13	2	1
$A_5(t)$	0	7	0	0
$A_{total}(t)$	0	30	9	5

As we see, $A_3(t)$ has the highest average ED arrivals, while $A_1(t)$and $A_5(t)$ have the lowest averages during this time. However, the emergence of COVID-19 in 2019 and 2020 led to variations in ED arrivals compared to previous years.

Having a more comprehensive understanding of data, we first studied the Total time series. Since the target of this research is forecasting each ESI, five time-series $A_1(t)$, $A_2(t)$, $A_3(t)$, $A_4(t)$, $A_5(t)$ for four years are also reviewed separately.

4.1. Hour of the day

The arrival data at different times of the year might have potential seasonality in different blocks. We aim to identify seasonality by plotting the data in potential seasonal blocks. Figure 4.1 shows the boxplot grouped by the hour of the day for the total time series between 2017 and 2020.

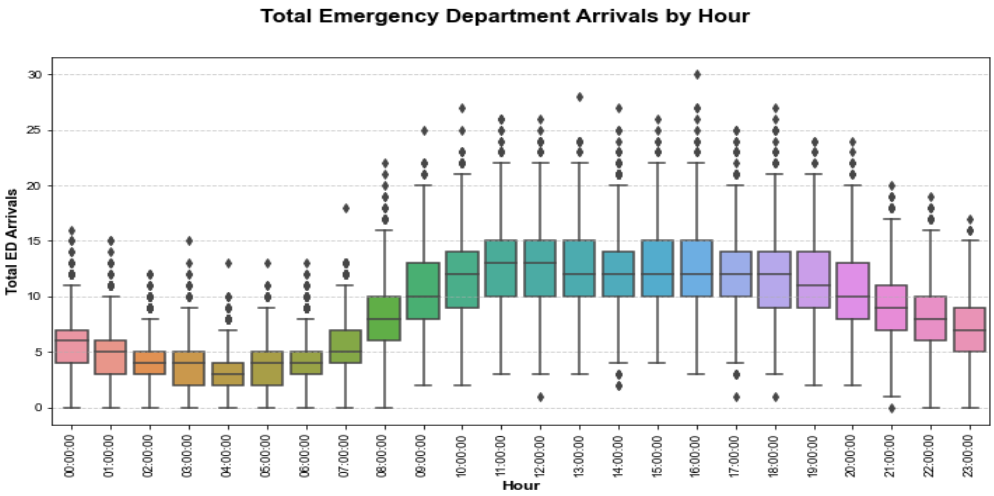

Figure 4.1: Boxplot of the total arrival grouped by hour of the day.

The box plot of total arrivals shows a daily seasonality, which means that the average arrivals are concentrated between 9:00 a.m. and 8:00 p.m. We will use this information to select the model for our proposed solution. However, this seasonal behavior is unclear when visually checking the same boxplot for individual ESI. Besides, we observed more fluctuations and less regular patterns at this hierarchy level. Figure 4.2 shows the average hourly arrival for each ESI_1 to ESI_5. ESI_1 and ESI_5 exhibit the highest variability in average arrivals.

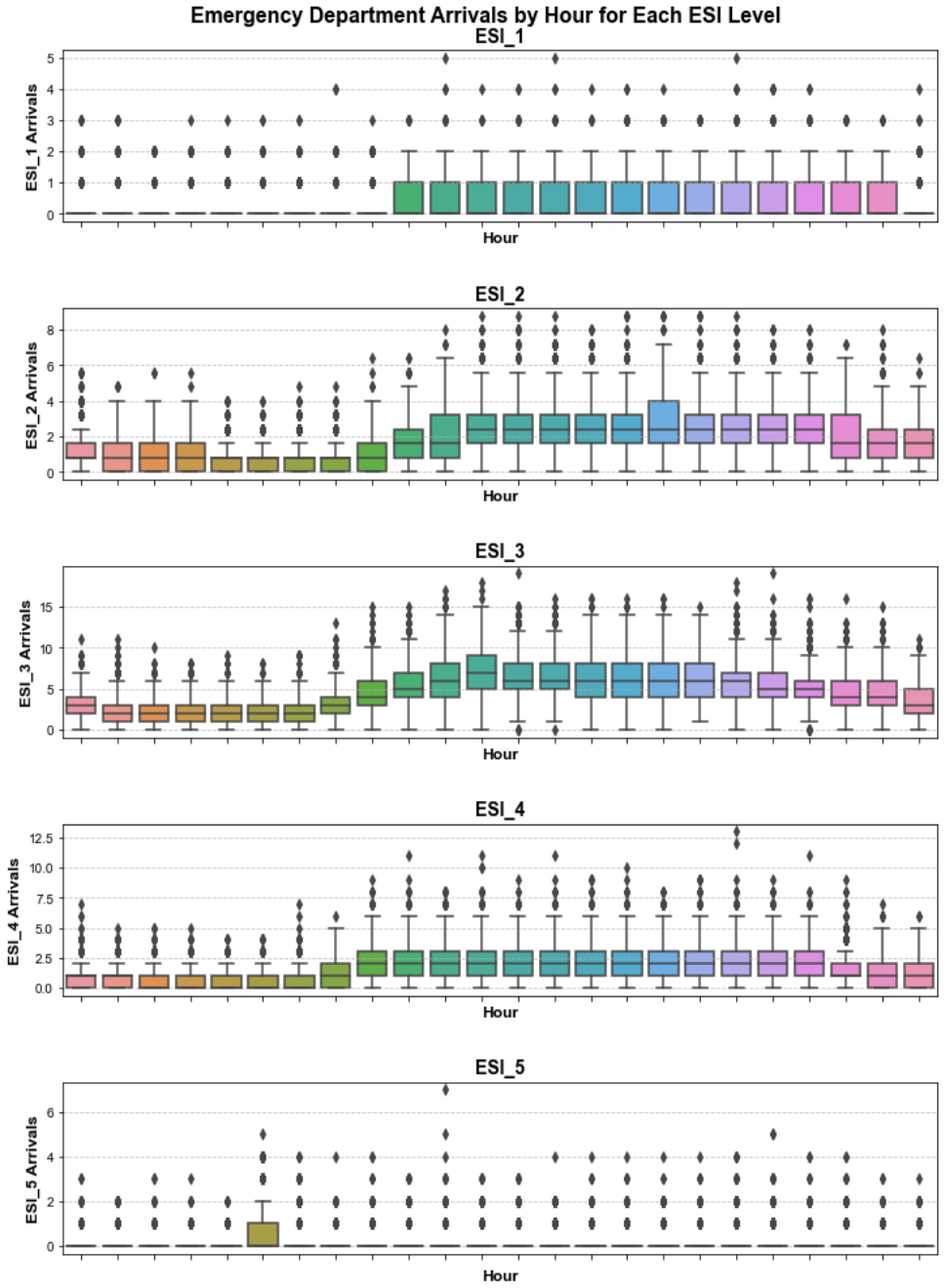

Figure 4.2: Boxplot of each ESI grouped by hour of the day.

4.2. Day of week

Another potential seasonality is the day of the week. The arrival pattern could be different on different days of the week. Figure 4.3 describes the seasonality linked to the day of the week. It shows that the average number of patient arrivals is less on weekends than on weekdays. On

weekdays, patient arrivals remain relatively stable within a specific range. However, a slight increase in patient arrivals is observed on Mondays and Tuesdays.

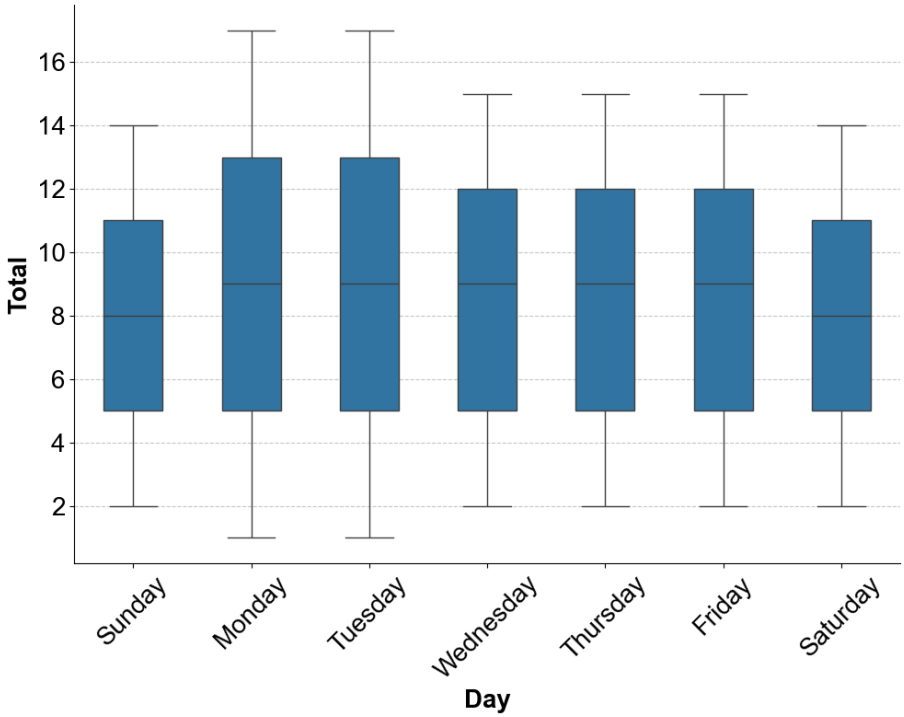

Figure 4.3: Boxplot of the total arrivals grouped by the day of the week.

Figure 4.3 shows similar boxplots for each individual ESI. We can observe the same patterns in each ESI, but they are less visible at this level. This behavior makes the day of the week a potential candidate for an expletory variable of the forecasting model to explain the weekly seasonal pattern.

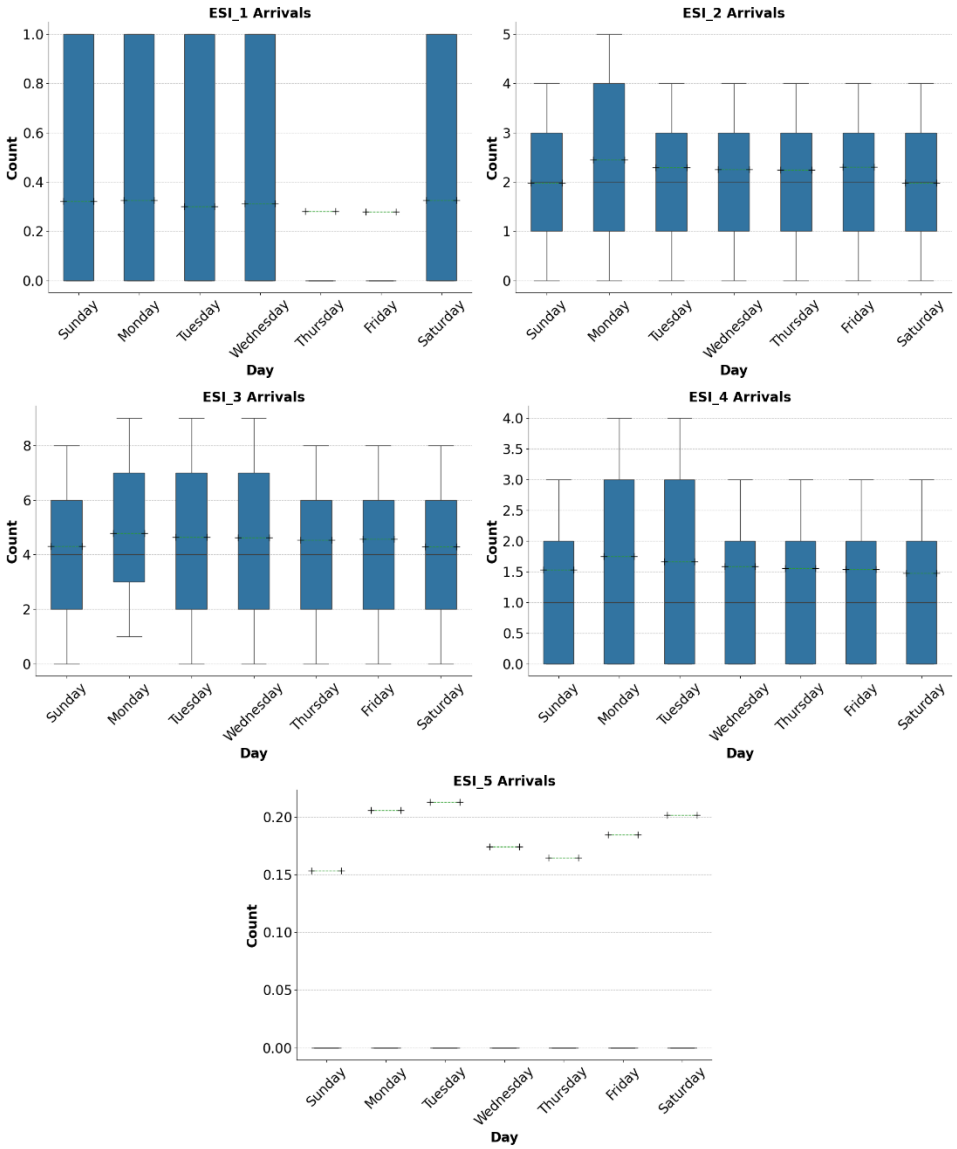

Figure 4.4: Boxplot of the arrivals in each ESI grouped by the day of the week.

4.3. Month of the year

We examined whether there were any variations in ED arrivals across different months of the year. As depicted in Figure 4.5, the total ED arrivals throughout the months exhibited seasonal behavior at some levels. We can see higher arrivals in summer and winter compared to shoulder months. Figure 4.6 shows the same boxplot for each individual ESI.

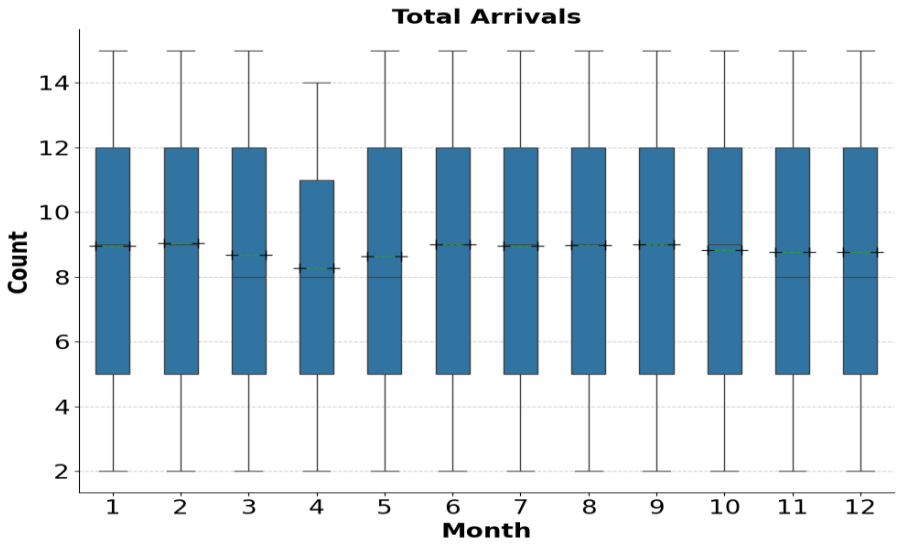

Figure 4.5: Boxplot of the total arrival grouped by month of the year.

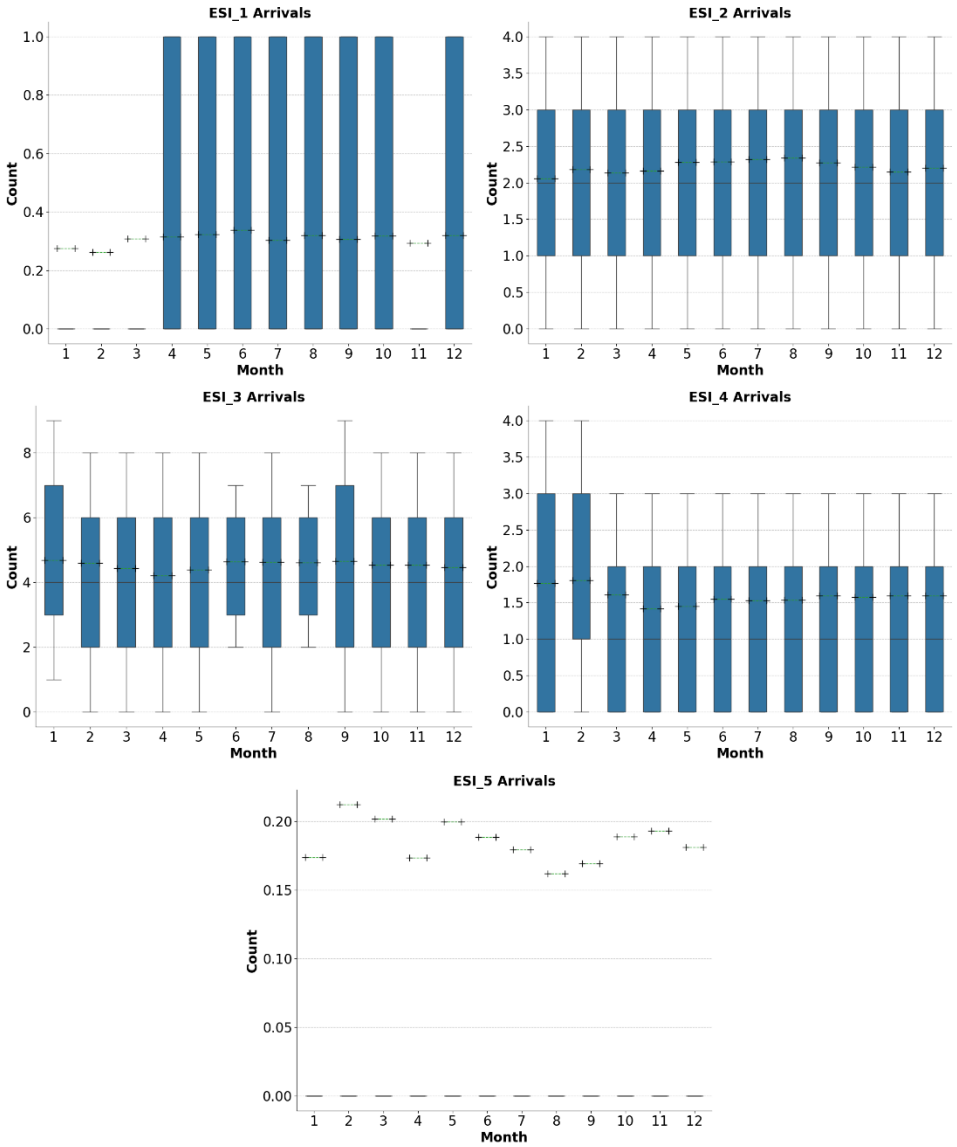

Figure 4.6: Boxplot of the arrival for each ESI grouped by month of the year.

Since the COVID-19 pandemic could affect patient arrivals in different months of the year in 2020, we investigate the average patient arrivals for the total months of 2017 to 2020 separately in Figure 4.7. Obviously, the number of arrivals dropped significantly in April 2020 due to the COVID-19 outbreak in the US. The arrivals recovered after two months, and we can even see an increase in numbers in July 2020.

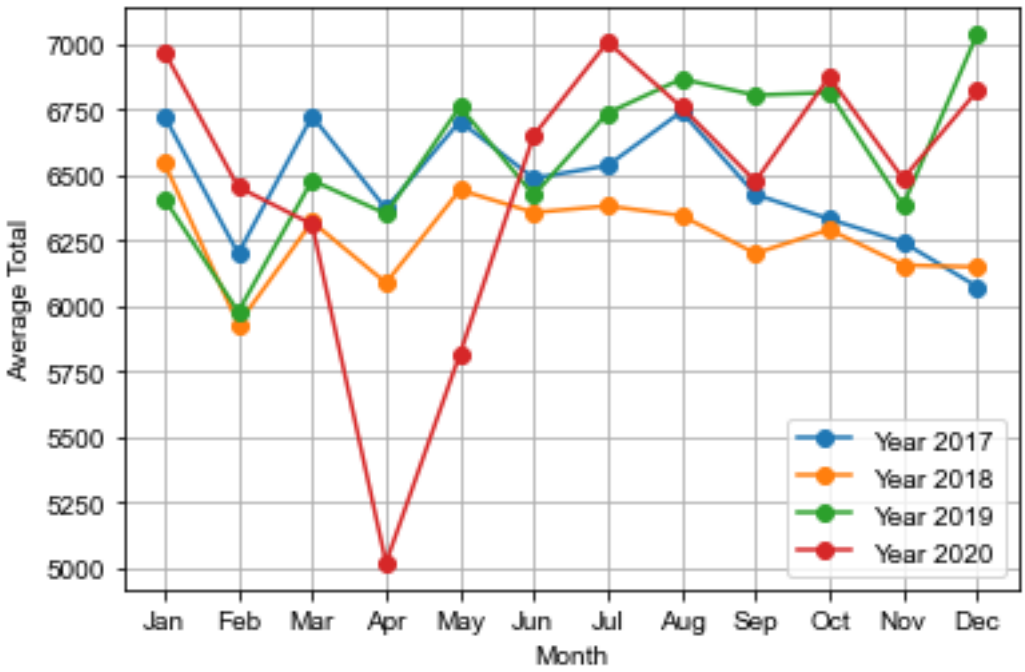

Figure 4.7: Monthly arrivals for all ESIs from 2017 to 2020.

5. METHODOLOGY

This chapter presents the methodology of all the models we apply for one-hour ahead ED arrivals corresponding to each ESI category. In this research, we proposed a solution to forecast the arrival at ED. The proposed solution is examined in a case study using real-world data introduced in Chapter 3, and it is compared to state-of-the-art models as benchmarks. This chapter elaborates on benchmark models, proposed solutions, and the case study results.

5.1. Benchmark Models

To evaluate the performance of the proposed models, we employed eight different significant models from the literature, as the benchmark models are shown in Figure 5.1 and listed as follows:

1. Exponential Smoothing
2. ARIMA

Hierarchical forecasting with exponential smoothing and three different reconciliation approaches:

3. Hierarchical forecasting with exponential smoothing and an average of historical proportions (tdgsa) reconciliation
4. Hierarchical forecasting with exponential smoothing and the proportion of the historical average (tdgsf) reconciliation
5. Hierarchical forecasting with exponential smoothing and forecasted proportions of bottom-level (tdfp) reconciliation

Hierarchical forecasting with ARIMA and three different reconciliation approaches:

6. Hierarchical forecasting with ARIMA An average of historical proportions (tdgsa) reconciliation

7. Hierarchical forecasting with ARIMA and the proportion of the historical average (tdgsf) reconciliation
8. Hierarchical forecasting with ARIMA and forecasted proportions of bottom-level (tdfp) reconciliation

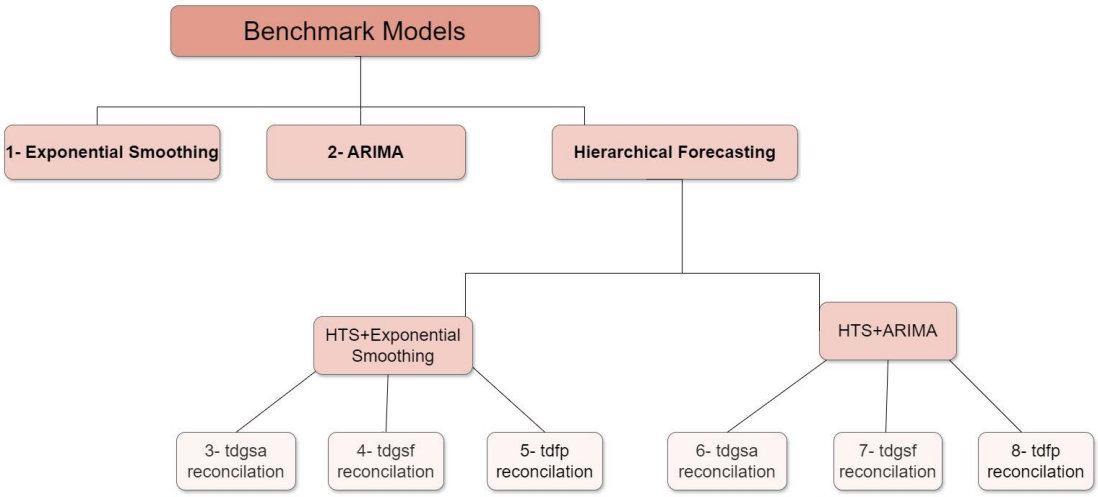

Figure 5.1: Benchmark Models.

The first benchmark model is the exponential smoothing model. For the exponential smoothing benchmark model, we use the *ETS()* function in the (*fable)* package in *R* software with default setting and without specifying the right-hand side of the formula in order to make a model to fit the training data [46]. For point forecasts, multiplicative trend methods are not considered since they produce poor forecasts. Therefore, in the first step, *ETS()* applies all combinations of (Error, Trend, Seasonal) models that determine the optimized values by using MLE for the smoothing parameters and initial state variables. Then, the model selection for choosing the best models (more accurate and less complicated) according to AICc is made.

Finally, the best model is applied to point forecast with the *forecast()*[10] function in the *fable* package for one step ahead or as many steps as required.

The second benchmark model in this research is ARIMA. In R software, *auto.arim()* function in the *forecast* package fits the ARIMA model for any univariate time series to use for point forecast [47]. The algorithm for this *auto.arima()* is determined by Hyndman and Khanddakar (JSS, 2008):

1. Identify whether the training data is stationary or not. If not, select the number of seasonal differences (D) using the seasonal strength test[11] and the number of non-seasonal differences (d) via KPSS[12].
2. Choose p, q, P, Q, and c by minimizing the AIC_c.
3. Use stepwise search to create different models with different p,d,q, P, D, Q.
4. Obtain the best model by minimizing the AIC.

The authors suggest that making as few differences as possible would be better because differencing can decrease the accuracy of predictions [48]. After training the model with the selected ARIMA model, with the *forecast()* function, the forecasting for one step (or any other number of steps) will be done.

Hierarchical forecasting is another approach for this study. The first step is creating the hierarchical time series from the training dataset. For the hierarchical forecasting model, the *hts()*

[10] The version of *fable* package used for this research is 0.3.3 and for the *forecast* is 8.21.1
[11] By nsdiffs() function automatically.
[12] By ndiffs() function automatically.

function in the (*hts)* [13] package is applied to create hierarchical and grouped time series. We identify the bottom and top levels for a one-level hierarchical time series. Once we set up the hierarchical structure, we use the *forecast()* function to generate the top-level forecast. The standard method in the *hts()* function for this top-level or total time series forecast is ETS models, but we have the option to use other methods, such as ARIMA by argument (*fmethod)*. Following that, we can apply specific reconciliation methods outlined in Section 3.5. We specify the reconciliation method using the argument (*method)* [49].

5.2. Proposed Models

We propose five primary models for predicting one-hour-ahead ED arrivals: The Multiple Linear Regression (MLR) models and hierarchical forecasting with MLR for top-level forecast and three different reconciliation models for the top-down approach, Ensemble model. Figure 5.2 shows all the proposed models.

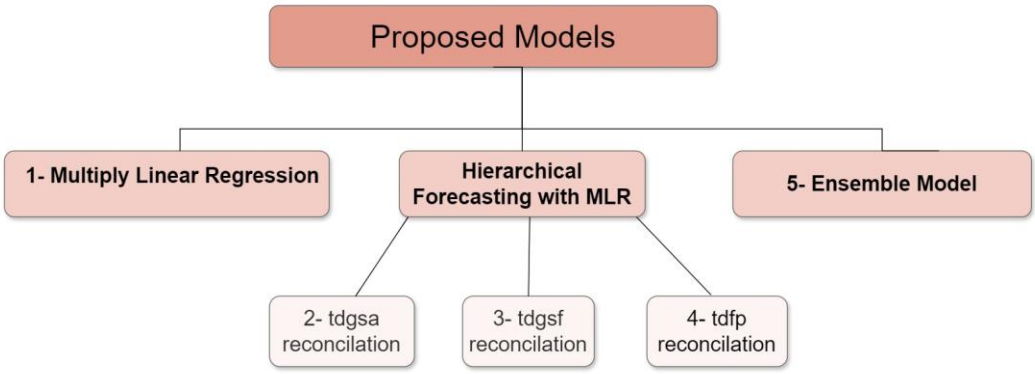

Figure 5.2: Proposed Models.

[13] hts package used in this study is version 6.0.2

5.2.1. Multiple Linear Regression models (MLR)

Figure 5.3 shows the steps of developing the MLR models.

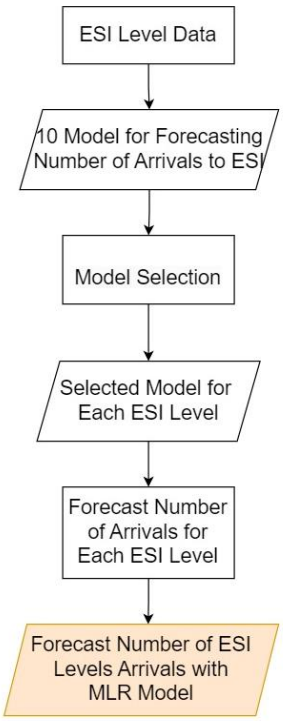

Figure 5.3: Forecasting Number of ESI level arrivals with MLR model.

According to some literature, we assume that holidays can impact the volume of ED arrivals. Therefore, we consider this feature an independent variable to predict the number of EDs' arrivals. To capture the potential holiday effects on ED arrivals, we introduced two holiday-related variables, including a binary variable indicating the federal holidays (F). The federal holidays include New Year's Day, the Birthday of Martin Luther King, Presidents Day, Memorial Day, Independence Day, Labor Day, Columbus Day, Veterans Day, Thanksgiving, and Christmas. This variable is categorical, including two classes, 0 for non-holiday dates and 1 for holiday dates.

Another calendar variable is days after federal holidays (AF). The latter variable is categorical, with zero values if the day before is not a holiday and one, two, or three for one, two, and three days after each federal holiday, respectively.

We use a trend variable (T_t) as a quantitative variable to capture potential trends in the hourly arrival. A trend variable is a natural number for each hourly data point in ascending order. It is the same as an index for the data set. Since our data contains three years, we can identify the long-term trend in arrivals.

We also consider some interaction effects of class variables in our models. Interaction terms help the model understand the complex ways variables relate to each other. We considered two types of interaction effects in MLR models, including:

1- Hour of the day×Day of the week that leads to 24×7 cross effect ($H_t \times D_t$)

2- Hour of the day×Days after a holiday that leads to 24×4 cross effect ($H_t \times (AF)_t$).

Combining the variables defined above, we developed ten models listed in Table 5-1. We use Python, and functions are employed from the *sklearn* library to derive the results for each model.

Table 5-1: The models developed for the proposed solution.

Name	**Model**
M_1	$A_j(t)=\beta_0 + \beta_1 T_t+\beta_2 H_t$
M_2	$A_j(t)=\beta_0 + \beta_1 T_t+\beta_2 H_t+\beta_3 D_t$
M_3	$A_j(t)=\beta_0 + \beta_1 T_t+\beta_2 H_t+\beta_3 D_t+\beta_4 (H_t \times D_t)$
M_4	$A_j(t)=\beta_0 + \beta_1 T_t+\beta_2 H_t+\beta_3 F_t$
M_5	$A_j(t)=\beta_0 + \beta_1 H_t+\beta_2 D_t+\beta_3 (H_t \times D_t)+\beta_4 F_t$
M_6	$A_j(t)=\beta_0 + \beta_1 T_t+\beta_2 H_t+\beta_3 D_t+\beta_4 M_t$
M_7	$A_j(t)=\beta_0 + \beta_1 T_t+\beta_2 H_t+\beta_3 D_t+\beta_4 M_t+\beta_5 (H_t \times D_t)$
M_8	$A_j(t)=\beta_0 + \beta_1 H_t+\beta_2 D_t+\beta_3 M_t+\beta_4 (H_t \times D_t) + \beta_5 F_t$
M_9	$A_j(t)=\beta_0 + \beta_1 H_t+\beta_2 D_t+\beta_3 M_t+\beta_4 (H_t \times D_t) + \beta_5 F_t+\beta_6 (AF)_t$
M_10	$A_j(t)=\beta_0 + \beta_1 H_t+\beta_2 D_t+\beta_3 M_t+\beta_4 (H_t \times D_t)+\beta_5 F_t+\beta_6 (AF)_t+\beta_7 (H_t \times (AF)_t)$

5.2.2. Hierarchical forecasting with MLR model

The second solution proposed in this research is hierarchical forecasting with an MLR model for the top-level forecast and using three different top-down reconciliations. Figure 5.4 shows the process of this forecasting model.

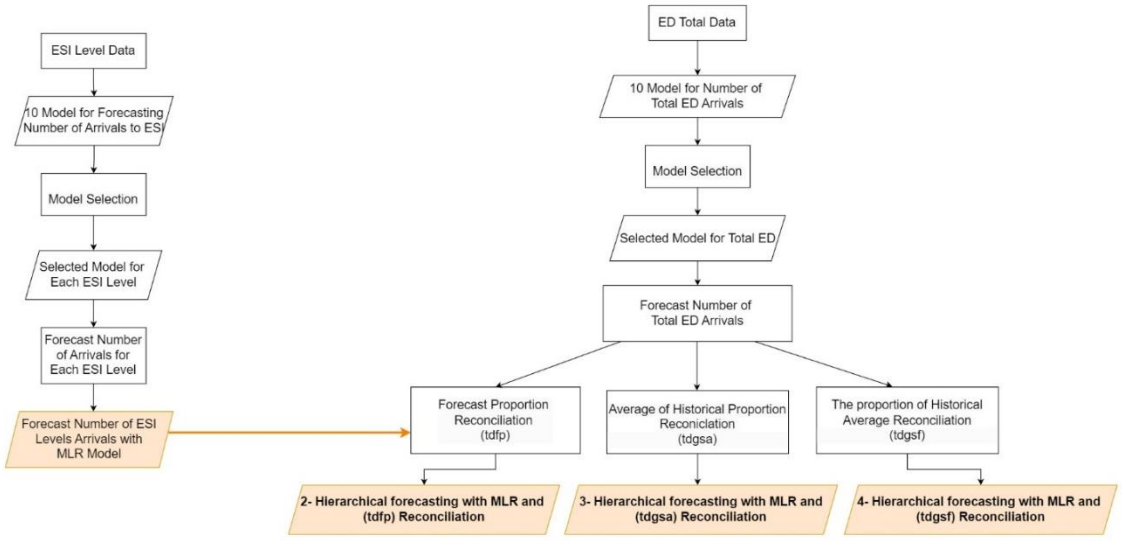

Figure 5.4: Hierarchical forecasting with MLR model

The following steps describe this method:

1. Finding the best MLR model for the top-level time series using a validation period (The model selection is explained in Section 5.2.4)
2. Finding the proportion of top-level forecast for each bottom-level time series of hierarchical data.
3. Reconcile the forecasts using three methods (tdgsa), (tdgsf), and (tdfp), which are explained in Section 3.5.

To forecast the top level, we cannot use the same models we developed for the ESI level (Section 5.2.1) because, as we explored the data in Chapter 4, the patterns are different at the aggregate level, and using the same model as we have for each level will make the reconciliation

useless. Therefore, we developed different versions of MLR models for $A_{total}(t)$. We employed the same variables we used for the models for each level except the hour of the day (H) variable. Instead, we introduced a new categorical variable named (HG) to represent seasonal blocks of the day. This variable, (HG), is divided into three categories as follows:

- Group_1 (00:00 a.m. to 8:00 a.m.): Characterized by the lowest number of total ED arrivals during the day.
- Group_2 (9:00 a.m. to 8:00 p.m.): This includes most ED arrivals, indicating peak activity during this period.
- Group_3 (9:00 p.m. to 11:00 p.m.): Marked by fewer ED arrivals.

The interaction effects we considered are $(HG)_t \times D_t$, which leads to 3×7 cross effect, and $(HG)_t \times (AF)_t$, which leads to 3×4 cross effect.

We developed ten different models for the total time series. The model functions are listed in Table 5-2.

Table 5-2: The models for the Total time series.

Name	**Model**
MT_1	$A_{total}(t)=\beta_0 + \beta_1 T_t+\beta_2(HG)_t$
MT_2	$A_{total}(t)=\beta_0 + \beta_1 T_t+\beta_2(HG)_t+\beta_3 D_t$
MT_3	$A_{total}(t)=\beta_0 + \beta_1 T_t+\beta_2(HG)_t+\beta_3 D_t+\beta_4 M_t$
MT_4	$A_{total}(t)=\beta_0 + \beta_1 T_t+\beta_2(HG)_t+\beta_3 F_t$
MT_5	$A_{total}(t)=\beta_0 + \beta_1 T_t+\beta_2(HG)_t+\beta_3 D_t+\beta_4 F_t$
MT_6	$A_{total}(t)=\beta_0 + \beta_1 T_t+\beta_2(HG)_t+\beta_3 D_t+\beta_4 M_t+\beta_5((HG)_t \times D_t)$
MT_7	$A_{total}(t)=\beta_0 + \beta_1 T_t+\beta_2(HG)_t+\beta_3 D_t+\beta_4 F_t+\beta_5((HG)_t \times D_t)$
MT_8	$A_{total}(t)=\beta_0 + \beta_1 T_t+\beta_2(HG)_t+\beta_3 D_t+\beta_4 M_t+\beta_5 F_t+\beta_6((HG)_t \times D_t)$
MT_9	$A_{total}(t)=\beta_0 + \beta_1 T_t+\beta_2(HG)_t+\beta_3 D_t+\beta_4 M_t+\beta_5 F_{t+}+\beta_6((HG)_t \times D_t) + \beta_7(AF)_t$
MT_10	$A_{total}(t)=\beta_0 + \beta_1 T_t+\beta_2(HG)_t+\beta_3 D_t+\beta_4 M_t+\beta_5 F_t+\beta_6((HG)_t \times D_t)+\beta_7(AF)_t+\beta_8((HG)_t \times (AF)_t)$

We must notice that the reconciliation by (tdgsa) and (tdgsf) methods don't need the bottom-level component individual forecasts; they use the proportion of the historical actual values to do the reconciliation. On the other hand, for the forecast proportion or (tdfp) method, we need a base forecast at the bottom level. We use the MLR selected model, explained in Section 5.2.1, as the base forecast for the hierarchical forecasting with MLR and (tdfp) reconciliation method.

5.2.3. Ensemble model

This model is the average of all twelve models (the benchmarks and the four other proposed models). We want to see if the performance of the average forecasting results of all the models is better than each individual model.

5.2.4. Model Selection for MLR Models

The model selection for both the MLR proposed model and hierarchical with MLR models is based on a rolling forecast approach for a validation period. Figure 5.5 shows the process for model selection.

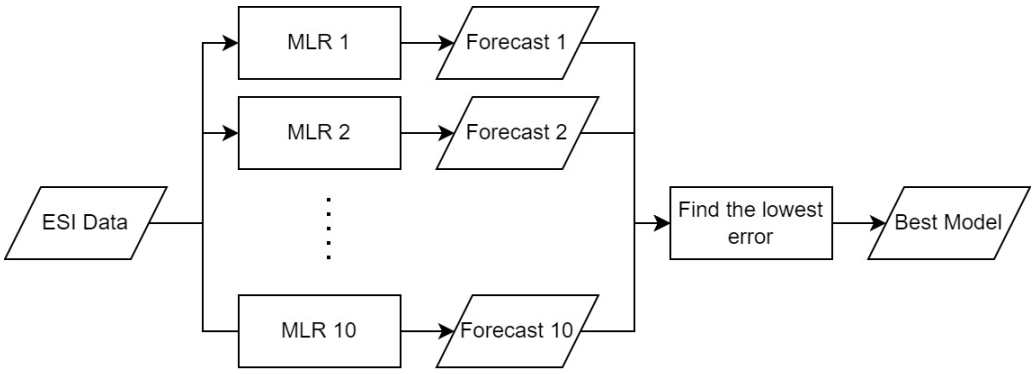

Figure 5.5: Model Selection for MLR models.

To forecast on a rolling basis, we need to move the training and forecasting windows h-step forward to cover the entire validation period. The training for each forecasting data point is shifted by size h. Figure 5.6 shows an example of a rolling forecast with a fixed training period.

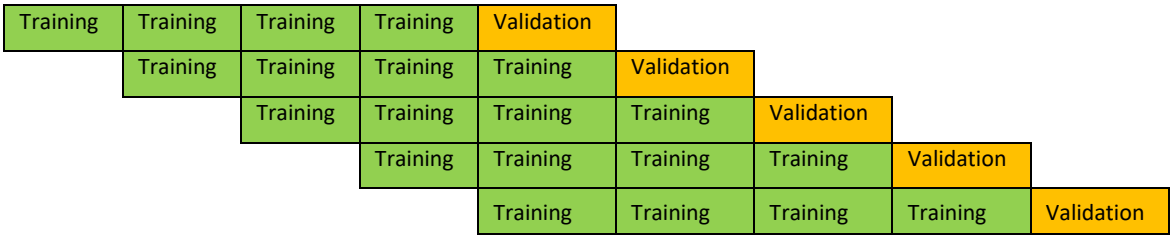

Figure 5.6: The rolling basis forecast.

In the case study section, we used two fixed years of data to train the models to forecast one-hour ahead for a validation period. We repeat the forecast for all hours of one validation year. Then, we calculate the average of the evaluation metrics explained in Section 3.6. The model with the smallest average MAE is selected as the best model to forecast the testing period.

5.3. Post-processing of forecasting results

Consequently, we will incorporate eight distinct benchmark models derived from representative studies in the literature and four proposed models in our study. Our objective is to predict numerical values as integers. To improve our understanding of these forecasted values and the precision of their evaluation, we will consider rounded values for all forecasts for the year 2020. As a result, a post-processing step will be applied, wherein all forecasted values across the selected models will be rounded to the nearest integer before assessing each model's forecasting performance.

5.4. Case Study

To evaluate the proposed models' performance and compare them with the benchmark models, we conducted a case study using the real-world data described in Chapter 4. In this case study, we use 2020 as the testing year. We will predict the one-hour ahead values for each time series $A_j(t)$ for j=1,…,5 at the time t for the test year. In simpler terms, we aim to forecast the values for each hour of the day, starting from the first hour on January 1, 2020, and continuing for the entire year.

For this purpose, we consider a rolling-basis approach to creating a training dataset for each point forecast. Each training dataset has a three-year window. This means that for each data point to forecast, the training data includes three years of hourly data, approximately 3*365*24=26280 hourly data points. For instance, to forecast $A_2(t)$ emergency arrivals at 1:00 p.m. on October 10, 2020, the training data starts from 1:00 p.m. on October 10, 2017, to 12:00 p.m. on October 10, 2020.

In addition, for the model selection parts of the proposed solution, we used 2019 as the validation year and two years of data as the training window. We used this domain to evaluate the performance of all the proposed MLR models and the MLR model for the $A_{total}(t)$ time series.

The first benchmark model for our experiment is the exponential smoothing model. We create the training data set for each ESI_1 to ESI_5 time series. Then, we fit a model by the *ETS()* function by default arguments for each data point to forecast. The function selects the best model based on minimizing AICc. Among all 30 models described in Section 3.2, the one that is selected by the *ETS()* function for most of the training data set, for all the time series and various data points, is *ETS(A, N, A)*. This model is an additive error ETS model with additive seasonality and non-trend. For each forecast, the initial states for the level (l_0) and seasonality-related parameters

(s_0 through s_{23}) are estimated automatically with the *ETS()* function. We refer $\tilde{f}_{1,j}(t)$ as the forecasted value of the time series $A_j(t)$ obtained through the exponential smoothing method.

The second benchmark model we applied for this research is the ARIMA model. A training model for each data point is created to forecast one-hour ahead ED arrivals for each time series $A_j(t)$, j=1,…,5. ARIMA models that the *auto.arima()* function selected to forecast each $A_j(t)$, j=1,…,5 are shown in Table 5-3. The best ARIMA model for each hour in the forecasting window is selected separately. Therefore, for a given $A_j(t)$, we could have different models in different forecasting hours.

Table 5-3: The ARIMA models used for different ESIs.

Models	**$A_1(t)$**	**$A_2(t)$**	**$A_3(t)$**	**$A_4(t)$**	**$A_5(t)$**
ARIMA (1,1,0)	×				
ARIMA (0,1,1)		×		×	
ARIMA (1,1,1)		×	×		
ARIMA (1,1,0) with drift				×	×
ARIMA (0,1,1) with drift		×		×	
ARIMA (1,1,2)		×			
ARIMA (0,1,2)			×		
ARIMA (2,1,1)			×		
ARIMA (2,1,0)					×
ARIMA (0,1,3)			×		
ARIMA (3,1,0)				×	×
ARIMA (5,1,0)	×			×	×
ARIMA (3,1,3) with drift			×		
ARIMA (4,0,1) with non-zero mean		×			
ARIMA (1,0,3) with non-zero mean		×		×	
ARIMA (5,0,1) with non-zero mean		×		×	

Table 5-3 shows that the selected models that *auto.arima()* function select as the best models to forecast each ESI time series data for the most point forecast. We denote $\tilde{f}_{2,j}(t)$ as the forecasted value of the time series $A_j(t)$ using ARIMA model.

For hierarchical forecasting, we organize our data to match the hierarchical structure. The hierarchy's top level, labeled "Total" captures the overall measure of all ESIs combined. This

aggregated "Total" series is then broken down into five distinct series, which form the bottom level of our hierarchical data structure. This hierarchical time series structure for the case study is single-level. Figure *5.7* illustrates the arrangement of the hierarchical tree diagram for the data. It consists of a single level, with five nodes in the bottom and top levels containing the total number of ED arrivals to all the ESIs. The total number of time series for the hierarchal time series is six.

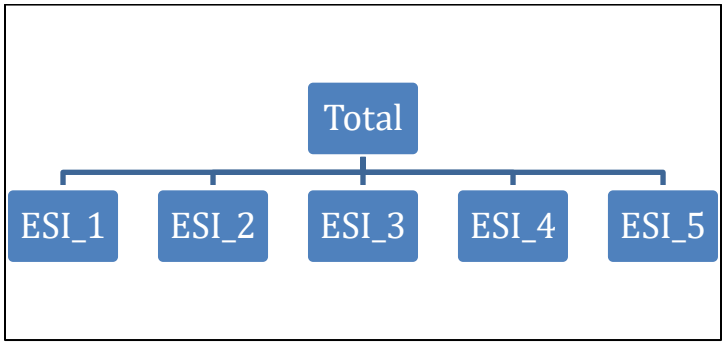

Figure 5.7: Hierarchical tree diagram.

Table 5-4 lists the names of six benchmark models that use the *hts()* function for the time series $A_j(t)$ at the specific time t.

Table 5-4: All HTS with ETS and ARIMA models

$\tilde{f}_{3,j}(t)$	HTS with ETS and (tdfp) reconciliation
$\tilde{f}_{4,j}(t)$	HTS with ETS and (tdgsa) reconciliation
$\tilde{f}_{5,j}(t)$	HTS with ETS and (tdgsf) reconciliation
$\tilde{f}_{6,j}(t)$	HTS with ARIMA and (tdfp) reconciliation
$\tilde{f}_{7,j}(t)$	HTS with ARIMA and (tdgsa) reconciliation
$\tilde{f}_{8,j}(t)$	HTS with ARIMA and (tdgsf) reconciliation

In benchmark hierarchical forecasting models denoted as HTS_with ETS and HTS_with ARIMA, we utilize ETS and ARIMA forecasting for the ESI_Total time series, respectively. For HTS_with ETS, three top-down reconciliation methods (tdgsa, tdgsf, and tdfp) described in Section3.5 are applied, resulting in three forecasting models for the year 2020, predicting one hour

ahead for each of the five time series $A_j(t)$, for j=1,…,5. Similarly, in HTS_with ARIMA, the same three top-down reconciliation methods are considered, generating three forecasting models for the $A_j(t)$.

The first proposed model is the MLR model we developed for each ESI level. To select the best model for each ESI, we used the year 2019 as the validation year and then ran all the ten models proposed in Table 5-1 the models on a rolling basis for one-hour ahead forecasting with two years of training data. Each model was applied to forecast one hour ahead of ED arrivals for 2019 for total ED arrivals with two years of rolling base hourly training data (starting from 2017). Then, we calculate MAE and RMSE as evaluation metrics to rigorously assess each model's performance. The model that exhibits the best outcomes based on these metrics is selected to forecast ED arrivals for $A_j(t)$, time series for 2020. The results of the model selection are shown in Table *5-5*.

Table 5-5: The result of MAE for each ESI using the year 2019.

Models	**$A_1(t)$**	**$A_2(t)$**	**$A_3(t)$**	**$A_4(t)$**	**$A_5(t)$**
M_1	0.4546	1.2044	1.6770	1.0352	0.3198
M_2	0.4544	1.2055	1.6740	1.0329	0.3195
M_3	0.4542	1.2008	1.6683	1.0315	0.3159
M_4	0.4545	1.2044	1.6771	1.0352	0.3198
M_5	0.4492	1.2011	1.6677	1.0342	0.3019
M_6	0.4494	1.2088	1.6764	1.0373	0.3099
M_7	0.4526	1.2002	1.6715	1.0323	0.3221
M_8	0.4493	1.2032	1.6708	1.0355	0.3063
M_9	0.4492	1.2035	1.6709	1.0353	0.3064
M_10	0.4494	1.2073	1.6740	1.0366	0.3066

We select models with the lowest MAE to forecast each time series. For $A_1(t)$, since M_5 and M_9 models yield equal MAE values, we use the simple average of these models to forecast for the ED arrivals of ESI_1 by MLR method. We identify $\tilde{f}_{9,j}(t)$ as the forecast of the time

series $A_j(t)$, j=1,…,5 at time t, produced by the MLR model. The selected models to apply for forecasting each time series for 2020 are listed in Table 5-6.

Table 5-6: Selected best MLR models for each ESI using the validation year.

Forecasted value	**Selected MLR Model**
$\tilde{f}_{9,1}(t)$	Average of M_5 and M_9 M_5: $A_1(t)=\beta_0+\beta_1 H_t+\beta_2 D_t+\beta_3 (H_t\times D_t)+\beta_4 F_t$ M_9: $A_1(t)=\beta_0+\beta_1 H_t+\beta_2 D_t+\beta_3 M_t+\beta_4 (H_t\times D_t)+\beta_5 F_t+\beta_6 (AF)_t$
$\tilde{f}_{9,2}(t)$	M_7: $A_2(t)=\beta_0+\beta_1 T_t+\beta_2 H_t+\beta_3 D_t+\beta_4 M_t+\beta_5 (H_t\times D_t)$
$\tilde{f}_{9,3}(t)$	M_5: $A_3(t)=\beta_0+\beta_1 H_t+\beta_2 D_t+\beta_3 (H_t\times D_t)+\beta_4 F_t$
$\tilde{f}_{9,4}(t)$	M_3: $A_4(t)=\beta_0+\beta_1 T_t+\beta_2 H_t+\beta_3 D_t+\beta_4 (H_t\times D_t)$
$\tilde{f}_{9,5}(t)$	M_5: $A_5(t)=\beta_0+\beta_1 H_t+\beta_2 D_t+\beta_3 (H_t\times D_t)+\beta_4 F_t$

The proposed hierarchical forecasting model, with a multiple linear regression model, has some differences from HTS_with ETS and HTS_with ARIMA. The initial forecasting with regression and a top-down reconciliation method involves identifying an optimal linear regression model to predict the $A_{total}(t)$ time series. To achieve this, we need to implement model selection. Similar to model selection for forecasting with MLR for each $A_j(t)$, for j=1,…,5 time series, we conduct a systematic approach comprising ten distinct models for $A_{total}(t)$ time series. Each model, as determined in Table 5-2, was applied to forecast one hour ahead of ED arrivals for 2019 for total ED arrivals with two years of rolling base hourly training data (starting from 2017). The model that is outperformed based on MAE and RMSE metrics is selected to forecast ED arrivals for the $A_{total}(t)$ time series for 2020.

As mentioned, we execute all the MLR forecasting models in Python for this research. Once we collect all the results for forecasting the one-hour ahead values of $A_{total}(t)$ for 2019, we evaluate the forecasting results. Table 5-7 provides a summary of MAE for all the models.

Table 5-7: The MAE result for model selection for Total time series using the year 2019.

Models	MAE
MT_1	2.5292
MT_2	2.5176
MT_3	2.5157
MT_4	2.5289
MT_5	2.5155
MT_6	2.4956
MT_7	2.4969
MT_8	2.4949
MT_9	2.4945
MT_10	2.495

While all the models to forecast the total ED arrival have a promised performance, models MT_8 and MT_9 have the lowest MAE. We consider RMSE for this case. Table 5-8 shows the RMSE values for the $A_{total}(t)$ for all the models for the year 2019.

Table 5-8: The RMSE result for model selection for Total time series using the year 2019.

Models	RMSE
MT_1	3.2538
MT_2	3.229
MT_3	3.2247
MT_4	3.2535
MT_5	3.2261
MT_6	3.203
MT_7	3.2046
MT_8	3.2008
MT_9	3.201
MT_10	3.2019

This makes MT_8 the preferred choice for forecasting the $A_{total}(t)$ time series for 2020. Where $\tilde{A}_{total}(t)$ is the one-hour ahead forecast for the ESI_Total time series for the time t with MT_8, MLR method, and obtained by:

$$\tilde{A}_{total}(t) = T_t+(HG)_t+D_t+M_t+F_t+D_t\times(HG)_t \quad 5\text{-}1$$

After we obtain the forecast for $A_{total}(t)$ time series through MT_8, to implement hierarchical forecasting with the MLR model (HTS_with MLR), we consider three top-down reconciliation methods (tdgsa, tdgsf, and tdfp).

First, we consider the hierarchical model with the MLR method produced by (tdfp) reconciliation. Equation 5-2 determines the forecasted value for each $A_j(t)$, j=1, 2, ..., 5.

$$\tilde{f}_{10,j}(t) = P_j(t) \times \tilde{A}_{total}(t) \qquad 5\text{-}2$$

Where $\tilde{f}_{10,j}(t)$ is the one-hour ahead forecast for $A_j(t)$ time series for the time t with hierarchical with MLR by (tdfp) reconciliation. As mentioned in Section 3.5, we also require base forecasts for the bottom level for hierarchical with MLR using (tdfp) reconciliation. We use the MLR forecasting models selected for each time series according to Table 5-5 as the base forecasts. Therefore, we establish the forecast proportion following Equation 3-26 in Section 3.5. Here, we can show the formula to calculate forecasting for each $A_j(t)$, j=1, 2, ..., 5 with hierarchical with MLP using (tdfp) reconciliation:

$$P_j(t) = \frac{\tilde{f}_{9,j}(t)}{\tilde{f}_{9,1}(t) + \tilde{f}_{9,2}(t) + \tilde{f}_{9,3}(t) + \tilde{f}_{9,4}(t) + \tilde{f}_{9,5}(t)} \qquad 5\text{-}3$$

Where $\tilde{f}_{9,j}(t)$ is the one-hour forecast for the $A_j(t)$ j=0,1,…,5 time series for the time t with MLR model.

For the hierarchical forecasting model with MLR and (tdgsa)/(tdgsf), since we deal with single-level hierarchical data, we easily determine the proportions with (tdgsa) and (tdgsf) using the formula outlined in Section 3.5. Consequently, if $\tilde{f}_{11,j}(t)$ is the forecast for all the time series

of bottom level represented $A_j(t)$, j=1, 2, ..., 5 by hierarchical forecasting model with MLR for top-level forecast and (tdgsa) reconciliation, then we have:

$$\tilde{f}_{11,j}(t) = P_j(t) \times \tilde{A}_{total}(t) \quad 5\text{-}4$$

For each $A_j(t)$, we will have corresponded $P_j(t)$ according to Equation 3-25. Since the historical data for each time series is produced on a rolling basis for our case study, the training time series for each $A_j(t)$ at time t is different. Therefore, the proportion for each time series needs to be calculated separately. Equation 5-5 shows how the proportion for (tdgsa) reconciliation is determined:

$$P_j(t) = \frac{1}{N} \sum_{s=(t-1)}^{(t-1)-26280} \frac{A_j(s)}{A_{total}(s)} \quad 5\text{-}5$$

Where *N=26280* is the size of the historical data set, $A_j(s)$ is the value of the bottom level (j) series at time *s* of historical data, $A_{total}(s)$ is the top level of training time series at time s.

Similarly, $\tilde{f}_{12,j}(t)$ is the forecasted value with hierarchical and MLR forecast for top-level and (tdgsf) reconciliation for time t.

$$\tilde{f}_{12,j}(t) = P_j(t) \times \tilde{A}_{total}(t) \quad 5\text{-}6$$

The (tdgsf) reconciliation method or proportion of the historical average is determined by Equation 3-26 in Section 3.5. Equation 5-7 shows how the proportion for (tdgsf) reconciliation for our case study is calculated:

$$P_j(t) = \frac{\sum_{s=t-1}^{(t-1)-26280} \frac{A_j(s)}{N}}{\sum_{s=(t-1)}^{N(t-1)-26280} \frac{A_{total}(s)}{N}} \quad 5\text{-}7$$

Where *N=26280* is the size of the historical data set, $A_j(s)$ is the value of the bottom level series at time *s* of historical data, $A_{total}(s)$ is the top level of training time series at time s.

Ultimately, for every ED arrival at each ESI level and for each one-hour-ahead forecast for the year 2020, we find the simple average of all the models as an ensemble model. Thus, we have:

$$\tilde{f}_{ensemble,j}(t) = \frac{1}{12}(\tilde{f}_{1,j}(t) + \tilde{f}_{2,j}(t) + \cdots + \tilde{f}_{12,j}(t)) \quad 5\text{-}8$$

Where $\tilde{f}_{1,j}(t)$ to $\tilde{f}_{12,j}(t)$ are the forecasted value of each benchmark and proposed models for the time series $A_j(t)$, at time t.

5.5. Results

This section presents the evaluation of the results for both the benchmark and proposed forecasting models, as detailed in the preceding sections. The results of all thirteen models that we applied to forecast one-hour ahead of the year 2020 of our case study, are post-processing by rounding to the nearest integer number, as we mentioned in Section 5.2. Then, the methods are evaluated by determining the MAE and RMSE values. Table 5-8 and Table 5-9 show the MAE and RMSE values respectfully for all the forecasted models for all the $A_j(t)$ time series.

Table 5-9: The MAE values of the forecasts using all models.

		MAE				
		$A_1(t)$	$A_2(t)$	$A_3(t)$	$A_4(t)$	$A_5(t)$
Benchmark Models						
1	$\tilde{f}_{1,j}(t)$: ETS	0.3322	1.4029	2.3545	1.1098	0.2704
2	$\tilde{f}_{2,j}(t)$: AUTOARIMA	0.4165	1.3251	2.085	1.0938	0.3178
3	$\tilde{f}_{3,j}(t)$: HTS+ETS+tdfp	0.3322	1.4002	2.3487	1.1101	0.2688
4	$\tilde{f}_{4,j}(t)$: HTS+ETS+tdgsa	**0.3309**	1.4802	2.2843	1.0744	0.2214
5	$\tilde{f}_{5,j}(t)$: HTS+ETS+tdgsf	0.3323	1.4768	2.2928	1.0728	0.2214
6	$\tilde{f}_{6,j}(t)$: HTS+ARIMA+tdfp	0.4107	1.2858	1.9297	1.0741	0.3293
7	$\tilde{f}_{7,j}(t)$: HTS+ARIMA+tdgsa	0.3547	1.224	1.8763	1.0092	0.2216
8	$\tilde{f}_{8,j}(t)$: HTS+ARIMA+tdgsf	0.7148	1.7207	2.5367	1.4461	0.7033
Proposed Models						
9	$\tilde{f}_{9,j}(t)$: MLR	0.3385	**1.1762**	**1.716**	0.9798	0.2043
10	$\tilde{f}_{10,j}(t)$: Hierarchical+MLR+tdfp	0.3384	1.1873	1.8021	1.0053	0.1994
11	$\tilde{f}_{11,j}(t)$: Hierarchical+MLR+tdgsa	**0.3309**	1.1854	1.8005	1.0115	0.2214
12	$\tilde{f}_{12,j}(t)$: Hierarchical+MLR+tdgsf	0.3314	1.1854	1.8023	1.0011	0.2214
13	$\tilde{f}_{13,j}(t)$: Ensamble Model	0.3334	1.2181	1.8176	**0.9405**	0.2225

Table 5-10: The RSME values of the forecasts using all models.

		RMSE				
		$A_1(t)$	$A_2(t)$	$A_3(t)$	$A_4(t)$	$A_5(t)$
Benchmark Models						
1	$\tilde{f}_{1,j}(t)$: ETS	0.6809	1.8351	2.99	1.4718	0.6798
2	$\tilde{f}_{2,j}(t)$: AUTOARIMA	0.7163	1.7576	2.7034	1.4706	0.6944
3	$\tilde{f}_{3,j}(t)$: HTS+ETS+tdfp	0.6834	1.8297	2.9786	1.4713	0.677
4	$\tilde{f}_{4,j}(t)$: HTS+ETS+tdgsa	0.6817	1.906	2.9071	1.4193	0.6197
5	$\tilde{f}_{5,j}(t)$: HTS+ETS+tdgsf	0.6825	1.9041	2.9139	1.421	0.6197
6	$\tilde{f}_{6,j}(t)$: HTS+ARIMA+tdfp	0.7148	1.7207	2.5367	1.4461	0.7033
7	$\tilde{f}_{7,j}(t)$: HTS+ARIMA+tdgsa	0.687	1.6279	2.4699	1.3544	0.6198
8	$\tilde{f}_{8,j}(t)$: HTS+ARIMA+tdgsf	0.6843	1.6297	2.4731	1.361	0.6197
Proposed Models						
9	$\tilde{f}_{9,j}(t)$: MLR	0.6813	1.5786	2.2539	1.3295	0.5584
10	$\tilde{f}_{10,j}(t)$: Hierarchical+MLR+tdfp	0.6859	1.5809	2.3302	1.337	0.5497
11	$\tilde{f}_{11,j}(t)$: Hierarchical+MLR+tdgsa	0.6817	1.5969	2.3396	1.3422	0.6197
12	$\tilde{f}_{12,j}(t)$: Hierarchical+MLR+tdgsf	0.6817	1.5969	2.3313	1.3362	0.6197
13	$\tilde{f}_{13,j}(t)$: Ensamble Model	0.6808	1.6098	2.3605	1.263	0.6205

From the tables above, for $A_2(t)$, $A_3(t)$, $A_4(t)$ both MAE and RMSE conclude the same methods. However, for $A_1(t)$ and $A_5(t)$ each evaluation technique suggested different forecasting methods. On the other hand, the variations among the majority of the models are relatively minimal. The negligible difference in MAE between the two models with the lowest MAE makes it challenging to determine whether this difference is significant. Therefore, we determine the statistical significance of forecasting performance differences between models with minor variations in MAE by the Diebold-Mariano (DM) test. As described in Section 3.8, the DM statistic corresponds to a normal distribution under the null hypothesis, which considers that the two models have equal forecasting accuracy. This means that the mean differences in the forecasting errors of the two methods follow the normal distribution, so the differences between the two methods are insignificant. The level of significance for this case study is 5%. Table 5-11 to Table 5-20 illustrates the outcomes of the p-value for DM statistics for all the comparisons for each pair of 13 models (78 comparisons). Then we explain the result for the top pairs of models in terms of MAE and also RMSE.

Table 5-11: p-value for DM statistics for MAE of $A_1(t)$.

$\tilde{f}_{1,1}(t)$	-												
$\tilde{f}_{2,1}(t)$	<0.01	-											
$\tilde{f}_{3,1}(t)$	0.26	<0.01	-										
$\tilde{f}_{4,1}(t)$	0.90	<0.01	0.31	-									
$\tilde{f}_{5,1}(t)$	0.16	<0.01	0.93	0.02	-								
$\tilde{f}_{6,1}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	-							
$\tilde{f}_{7,1}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-						
$\tilde{f}_{8,1}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-					
$\tilde{f}_{9,1}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-				
$\tilde{f}_{10,1}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.92	-			
$\tilde{f}_{11,1}(t)$	0.90	<0.01	<0.01	≈1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-		
$\tilde{f}_{12,1}(t)$	0.49	<0.01	0.59	0.13	0.25	<0.01	<0.01	<0.01	<0.01	<0.01	**0.13**	-	
$\tilde{\boldsymbol{f}}_{\boldsymbol{13,1}}(\boldsymbol{t})$	0.16	<0.01	0.71	0.13	0.72	<0.01	<0.01	<0.01	<0.01	<0.01	**0.13**	0.29	-
	$\tilde{\boldsymbol{f}}_{\boldsymbol{1,1}}(\boldsymbol{t})$	$\tilde{f}_{2,1}(t)$	$\tilde{f}_{3,1}(t)$	$\tilde{f}_{4,1}(t)$	$\tilde{f}_{5,1}(t)$	$\tilde{f}_{6,1}(t)$	$\tilde{f}_{7,1}(t)$	$\tilde{f}_{8,1}(t)$	$\tilde{f}_{9,1}(t)$	$\tilde{f}_{10,1}(t)$	$\tilde{\boldsymbol{f}}_{\boldsymbol{11,1}}(\boldsymbol{t})$	$\tilde{f}_{12,1}(t)$	$\tilde{f}_{13,1}(t)$

Table 5-12: p-value for DM statistics for RMSE of $A_1(t)$.

$\tilde{f}_{1,1}(t)$	-												
$\tilde{f}_{2,1}(t)$	<0.01	-											
$\tilde{f}_{3,1}(t)$	0.24	<0.01	-										
$\tilde{f}_{4,1}(t)$	0.06	<0.01	0.59	-									
$\tilde{f}_{5,1}(t)$	0.01	<0.01	0.78	0.09	-								
$\tilde{f}_{6,1}(t)$	<0.01	0.64	<0.01	<0.01	<0.01	-							
$\tilde{f}_{7,1}(t)$	0.02	<0.01	0.40	0.09	0.15	<0.01	-						
$\tilde{f}_{8,1}(t)$	0.08	<0.01	0.82	0.31	0.49	<0.01	0.12	-					
$\tilde{f}_{9,1}(t)$	0.12	<0.01	0.89	0.50	0.83	<0.01	0.24	0.65	-				
$\tilde{f}_{10,1}(t)$	0.11	<0.01	0.91	0.47	0.79	<0.01	0.26	0.67	0.93	-			
$\tilde{f}_{11,1}(t)$	0.06	<0.01	0.59	≈1	0.08	<0.01	0.09	0.30	0.50	0.47	-		
$\tilde{f}_{12,1}(t)$	0.04	<0.01	0.64	0.49	0.34	<0.01	0.10	0.35	0.60	0.56	0.49	-	
$\tilde{\boldsymbol{f}}_{\boldsymbol{13,1}}(\boldsymbol{t})$	**0.39**	<0.01	0.43	0.49	0.20	<0.01	0.04	0.17	0.34	0.31	0.49	0.39	-
	$\tilde{\boldsymbol{f}}_{\boldsymbol{1,1}}(\boldsymbol{t})$	$\tilde{f}_{2,1}(t)$	$\tilde{f}_{3,1}(t)$	$\tilde{f}_{4,1}(t)$	$\tilde{f}_{5,1}(t)$	$\tilde{f}_{6,1}(t)$	$\tilde{f}_{7,1}(t)$	$\tilde{f}_{8,1}(t)$	$\tilde{f}_{9,1}(t)$	$\tilde{f}_{10,1}(t)$	$\tilde{f}_{11,1}(t)$	$\tilde{f}_{12,1}(t)$	$\tilde{f}_{13,1}(t)$

Table 5-13: p-value for DM statistics for MAE of $A_2(t)$.

$\tilde{f}_{1,2}(t)$	-												
$\tilde{f}_{2,2}(t)$	<0.01	-											
$\tilde{f}_{3,2}(t)$	0.14	<0.01	-										
$\tilde{f}_{4,2}(t)$	<0.01	<0.01	<0.01	-									
$\tilde{f}_{5,2}(t)$	<0.01	<0.01	<0.01	0.02	-								
$\tilde{f}_{6,2}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	-							
$\tilde{f}_{7,2}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-						
$\tilde{f}_{8,2}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-					
$\tilde{f}_{9,2}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-				
$\tilde{\boldsymbol{f}}_{\boldsymbol{10,2}}(\boldsymbol{t})$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.01	-			
$\tilde{\boldsymbol{f}}_{\boldsymbol{11,2}}(\boldsymbol{t})$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	**0.10**	**0.71**	-		
$\tilde{\boldsymbol{f}}_{\boldsymbol{12,2}}(\boldsymbol{t})$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	**0.10**	**0.71**	**≈1**	-	
$\tilde{f}_{13,2}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.33	0.09	<0.01	<0.01	<0.01	<0.01	–
	$\tilde{f}_{1,2}(t)$	$\tilde{f}_{2,2}(t)$	$\tilde{f}_{3,2}(t)$	$\tilde{f}_{4,2}(t)$	$\tilde{f}_{5,2}(t)$	$\tilde{f}_{6,2}(t)$	$\tilde{f}_{7,2}(t)$	$\tilde{f}_{8,2}(t)$	$\tilde{\boldsymbol{f}}_{\boldsymbol{9,2}}(\boldsymbol{t})$	$\tilde{\boldsymbol{f}}_{\boldsymbol{10,2}}(\boldsymbol{t})$	$\tilde{f}_{11,2}(t)$	$\tilde{f}_{12,2}(t)$	$\tilde{f}_{13,2}(t)$

Table 5-14: p-value for DM statistics for RMSE of $A_2(t)$.

$\tilde{f}_{1,2}(t)$	-												
$\tilde{f}_{2,2}(t)$	<0.01	-											
$\tilde{f}_{3,2}(t)$	0.01	<0.01	-										
$\tilde{f}_{4,2}(t)$	<0.01	<0.01	<0.01	-									
$\tilde{f}_{5,2}(t)$	<0.01	<0.01	<0.01	0.21	-								
$\tilde{f}_{6,2}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	-							
$\tilde{f}_{7,2}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-						
$\tilde{f}_{8,2}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.31	-					
$\tilde{\boldsymbol{f}}_{\boldsymbol{9,2}}(\boldsymbol{t})$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-				
$\tilde{\boldsymbol{f}}_{\boldsymbol{10,2}}(\boldsymbol{t})$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.04	-			
$\tilde{f}_{11,2}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.12	-		
$\tilde{f}_{12,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.12	**≈1**	-	
$\tilde{f}_{13,2}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.01	0.01	-
	$\tilde{f}_{1,2}(t)$	$\tilde{f}_{2,2}(t)$	$\tilde{f}_{3,2}(t)$	$\tilde{f}_{4,2}(t)$	$\tilde{f}_{5,2}(t)$	$\tilde{f}_{6,2}(t)$	$\tilde{f}_{7,2}(t)$	$\tilde{f}_{8,2}(t)$	$\tilde{\boldsymbol{f}}_{\boldsymbol{9,2}}(\boldsymbol{t})$	$\tilde{\boldsymbol{f}}_{\boldsymbol{10,2}}(\boldsymbol{t})$	$\tilde{f}_{11,2}(t)$	$\tilde{f}_{12,2}(t)$	$\tilde{f}_{13,2}(t)$

Table 5-15: p-value for DM statistics for MAE of $A_3(t)$

$\tilde{f}_{1,3}(t)$	-												
$\tilde{f}_{2,3}(t)$	<0.01	-											
$\tilde{f}_{3,3}(t)$	0.04	<0.01	-										
$\tilde{f}_{4,3}(t)$	<0.01	<0.01	<0.01	-									
$\tilde{f}_{5,3}(t)$	<0.01	<0.01	<0.01	<0.01	-								
$\tilde{f}_{6,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	-							
$\tilde{f}_{7,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-						
$\tilde{f}_{8,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.12	-					
$\tilde{\boldsymbol{f}}_{\boldsymbol{9,3}}(\boldsymbol{t})$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-				
$\tilde{f}_{10,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-			
$\tilde{f}_{11,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.79	-		
$\tilde{f}_{12,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	≈1	0.68	-	
$\tilde{f}_{13,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.06	0.05	0.08	-
	$\tilde{f}_{1,3}(t)$	$\tilde{f}_{2,3}(t)$	$\tilde{f}_{3,3}(t)$	$\tilde{f}_{4,3}(t)$	$\tilde{f}_{5,3}(t)$	$\tilde{f}_{6,3}(t)$	$\tilde{f}_{7,3}(t)$	$\tilde{f}_{8,3}(t)$	$\tilde{\boldsymbol{f}}_{\boldsymbol{9,3}}(\boldsymbol{t})$	$\tilde{f}_{10,3}(t)$	$\tilde{f}_{11,3}(t)$	$\tilde{f}_{12,3}(t)$	$\tilde{f}_{13,3}(t)$

Table 5-16: p-value for DM statistics for RMSE of $A_3(t)$

$\tilde{f}_{1,3}(t)$	-												
$\tilde{f}_{2,3}(t)$	<0.01	-											
$\tilde{f}_{3,3}(t)$	<0.01	<0.01	-										
$\tilde{f}_{4,3}(t)$	<0.01	<0.01	<0.01	-									
$\tilde{f}_{5,3}(t)$	<0.01	<0.01	<0.01	0.02	-								
$\tilde{f}_{6,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	-							
$\tilde{f}_{7,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-						
$\tilde{f}_{8,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.28	-					
$\tilde{\boldsymbol{f}}_{\boldsymbol{9,3}}(\boldsymbol{t})$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-				
$\tilde{f}_{10,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-			
$\tilde{f}_{11,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.70	-		
$\tilde{f}_{12,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.22	0.01	-	
$\tilde{f}_{13,3}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.02	<0.01	-
	$\tilde{f}_{1,3}(t)$	$\tilde{f}_{2,3}(t)$	$\tilde{f}_{3,3}(t)$	$\tilde{f}_{4,3}(t)$	$\tilde{f}_{5,3}(t)$	$\tilde{f}_{6,3}(t)$	$\tilde{f}_{7,3}(t)$	$\tilde{f}_{8,3}(t)$	$\tilde{\boldsymbol{f}}_{\boldsymbol{9,3}}(\boldsymbol{t})$	$\tilde{f}_{10,3}(t)$	$\tilde{f}_{11,3}(t)$	$\tilde{f}_{12,3}(t)$	$\tilde{f}_{13,3}(t)$

Table 5-17: p-value for DM statistics for MAE of $A_4(t)$

$\tilde{f}_{1,4}(t)$	-												
$\tilde{f}_{2,4}(t)$	0.25	-											
$\tilde{f}_{3,4}(t)$	0.82	0.24	-										
$\tilde{f}_{4,4}(t)$	<0.01	0.19	<0.01	-									
$\tilde{f}_{5,4}(t)$	<0.01	0.01	<0.01	0.28	-								
$\tilde{f}_{6,4}(t)$	<0.01	<0.01	<0.01	0.97	0.89	-							
$\tilde{f}_{7,4}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-						
$\tilde{f}_{8,4}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-					
$\tilde{\boldsymbol{f}}_{\boldsymbol{9,4}}(\boldsymbol{t})$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-				
$\tilde{f}_{10,4}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.53	<0.01	<0.01	-			
$\tilde{f}_{11,4}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.63	<0.01	<0.01	0.14	-		
$\tilde{f}_{12,4}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.11	<0.01	<0.01	0.28	<0.01	-	
$\tilde{\boldsymbol{f}}_{\boldsymbol{13,4}}(\boldsymbol{t})$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.46	<0.01	<0.01	<0.01	-
	$\tilde{f}_{1,4}(t)$	$\tilde{f}_{2,4}(t)$	$\tilde{f}_{3,4}(t)$	$\tilde{f}_{4,4}(t)$	$\tilde{f}_{5,4}(t)$	$\tilde{f}_{6,4}(t)$	$\tilde{f}_{7,4}(t)$	$\tilde{f}_{8,4}(t)$	$\tilde{\boldsymbol{f}}_{\boldsymbol{9,4}}(\boldsymbol{t})$	$\tilde{f}_{10,4}(t)$	$\tilde{f}_{11,4}(t)$	$\tilde{f}_{12,4}(t)$	$\tilde{f}_{13,4}(t)$

Table 5-18: p-value for DM statistics for RMSE of $A_4(t)$

$\tilde{f}_{1,4}(t)$	-												
$\tilde{f}_{2,4}(t)$	0.91	-											
$\tilde{f}_{3,4}(t)$	0.75	0.94	-										
$\tilde{f}_{4,4}(t)$	<0.01	<0.01	<0.01	-									
$\tilde{f}_{5,4}(t)$	<0.01	<0.01	<0.01	0.25	-								
$\tilde{f}_{6,4}(t)$	0.01	<0.01	0.01	<0.01	<0.01	-							
$\tilde{f}_{7,4}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-						
$\tilde{f}_{8,4}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-					
$\tilde{\boldsymbol{f}}_{\boldsymbol{9,4}}(\boldsymbol{t})$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-				
$\tilde{f}_{10,4}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.01	<0.01	0.10	-			
$\tilde{f}_{11,4}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.02	<0.01	0.01	0.25	-		
$\tilde{f}_{12,4}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.18	0.87	<0.01	-	
$\tilde{\boldsymbol{f}}_{\boldsymbol{13,4}}(\boldsymbol{t})$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-
	$\tilde{f}_{1,4}(t)$	$\tilde{f}_{2,4}(t)$	$\tilde{f}_{3,4}(t)$	$\tilde{f}_{4,4}(t)$	$\tilde{f}_{5,4}(t)$	$\tilde{f}_{6,4}(t)$	$\tilde{f}_{7,4}(t)$	$\tilde{f}_{8,4}(t)$	$\tilde{\boldsymbol{f}}_{\boldsymbol{9,4}}(\boldsymbol{t})$	$\tilde{f}_{10,4}(t)$	$\tilde{f}_{11,4}(t)$	$\tilde{f}_{12,4}(t)$	$\tilde{f}_{13,4}(t)$

Table 5-19: p-value for DM statistics for MAE of $A_5(t)$

$\tilde{f}_{1,5}(t)$	-												
$\tilde{f}_{2,5}(t)$	<0.01	-											
$\tilde{f}_{3,5}(t)$	0.03	<0.01	-										
$\tilde{f}_{4,5}(t)$	<0.01	<0.01	<0.01	-									
$\tilde{f}_{5,5}(t)$	<0.01	<0.01	<0.01	≈1	-								
$\tilde{f}_{6,5}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	-							
$\tilde{f}_{7,5}(t)$	<0.01	<0.01	<0.01	0.32	0.32	<0.01	-						
$\tilde{f}_{8,5}(t)$	<0.01	<0.01	<0.01	≈1	≈1	<0.01	0.32	-					
$\tilde{\boldsymbol{f}}_{\boldsymbol{9,5}}(\boldsymbol{t})$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-				
$\tilde{\boldsymbol{f}}_{\boldsymbol{10,5}}(\boldsymbol{t})$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-			
$\tilde{f}_{11,5}(t)$	<0.01	<0.01	<0.01	≈1	≈1	<0.01	0.32	≈1	<0.01	<0.01	-		
$\tilde{f}_{12,5}(t)$	<0.01	<0.01	<0.01	≈1	≈1	<0.01	0.32	≈1	<0.01	<0.01	≈1	-	
$\tilde{f}_{13,5}(t)$	<0.01	<0.01	<0.01	0.01	0.01	<0.01	0.32	0.01	<0.01	<0.01	0.01	0.01	-
	$\tilde{f}_{1,5}(t)$	$\tilde{f}_{2,5}(t)$	$\tilde{f}_{3,5}(t)$	$\tilde{f}_{4,5}(t)$	$\tilde{f}_{5,5}(t)$	$\tilde{f}_{6,5}(t)$	$\tilde{f}_{7,5}(t)$	$\tilde{f}_{8,5}(t)$	$\tilde{\boldsymbol{f}}_{\boldsymbol{9,5}}(\boldsymbol{t})$	$\tilde{\boldsymbol{f}}_{\boldsymbol{10,5}}(\boldsymbol{t})$	$\tilde{f}_{11,5}(t)$	$\tilde{f}_{12,5}(t)$	$\tilde{f}_{13,5}(t)$

Table 5-20: p-value for DM statistics for RMSE of $A_5(t)$

$\tilde{f}_{1,5}(t)$	-												
$\tilde{f}_{2,5}(t)$	0.02	-											
$\tilde{f}_{3,5}(t)$	0.07	<0.01	-										
$\tilde{f}_{4,5}(t)$	<0.01	<0.01	<0.01	-									
$\tilde{f}_{5,5}(t)$	<0.01	<0.01	<0.01	≈1	-								
$\tilde{f}_{6,5}(t)$	<0.01	<0.01	<0.01	<0.01	<0.01	-							
$\tilde{f}_{7,5}(t)$	<0.01	<0.01	<0.01	0.32	0.32	<0.01	-						
$\tilde{f}_{8,5}(t)$	<0.01	<0.01	<0.01	≈1	≈1	<0.01	0.32	-					
$\boldsymbol{\tilde{f}_{9,5}(t)}$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-				
$\boldsymbol{\tilde{f}_{10,5}(t)}$	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-			
$\tilde{f}_{11,5}(t)$	<0.01	<0.01	<0.01	≈1	≈1	<0.01	0.32	≈1	<0.01	<0.01	-		
$\tilde{f}_{12,5}(t)$	<0.01	<0.01	<0.01	≈1	≈1	<0.01	0.32	≈1	<0.01	<0.01	≈1	-	
$\tilde{f}_{13,5}(t)$	<0.01	<0.01	<0.01	0.01	0.01	<0.01	0.32	0.01	<0.01	<0.01	0.01	0.01	-
	$\tilde{f}_{1,5}(t)$	$\tilde{f}_{2,5}(t)$	$\tilde{f}_{3,5}(t)$	$\tilde{f}_{4,5}(t)$	$\tilde{f}_{5,5}(t)$	$\tilde{f}_{6,5}(t)$	$\tilde{f}_{7,5}(t)$	$\tilde{f}_{8,5}(t)$	$\boldsymbol{\tilde{f}_{9,5}(t)}$	$\boldsymbol{\tilde{f}_{10,5}(t)}$	$\tilde{f}_{11,5}(t)$	$\tilde{f}_{12,5}(t)$	$\tilde{f}_{13,5}(t)$

In the following, we describe the result of the DM test for top models with better performance in terms of MAE and RMSE for each ESI level.

ESI_1: Table 5-11 indicates that the null hypothesis for most of the cases that compare $\tilde{f}_{4,1}(t)$ and $\tilde{f}_{11,1}(t)$ in terms of MAE is not rejected. It turns out there are no significant differences between the distribution of errors in these two models, meaning the forecasting with either of these two models leads to an accurate forecast. Furthermore, Table 5-12 reveals no significant differences between $\tilde{f}_{13,j}(t)$ which is selected models with $\tilde{f}_{1,j}(t)$, the second best model to forecast $A_1(t)$ due to its lowest RMSE according to in Table 5-10. In Fact, RMSE comparisons between $\tilde{f}_{13,j}(t)$ and almost all other models for forecasting ESI_1 reveal that the differences between models are not significant. This lack of substantial difference is attributed to two main reasons. First, according to

Table 4-2, the number of arrivals in this ESI level is small since the average number of arrivals is zero. It also shows that in the most crowded case, the number of arrivals is not bigger than 5. Secondly, as Figure 4.2 illustrates that the randomness of the number of arrivals at this ESI level is inevitable. Overall, because of this characteristic of arrivals to ESI_1, the results for most

of the comparisons of the other two forecasting models also show that the differences are between the following the normal distribution, and the differences are not significant.

ESI_2: According to Table 5-6, $\tilde{f}_{9,2}(t)$ is forecasted value for $\mathrm{A}_2(t)$ that selected variables of trend, hour of the day, day of the week, month of the year, and the interaction of hour and day of the week. However, according to Table 5-13, there are no significant differences between the MAE of $\tilde{f}_{9,2}(t)$ and $\tilde{f}_{11,2}(t)$ which is forecasted value of $\mathrm{A}_2(t)$ by hierarchical with MLR and (tdgsa) reconciliation. Also, the null hypothesis in the DM test for comparing the MAE of $\tilde{f}_{9,2}(t)$ and $\tilde{f}_{12,2}(t)$ is rejected. It reveals that there is no difference between the distribution of results of MAE of forecasting one-hour ahead if we select either of these three models. Additionally, Table 5-14 shows that in terms of RMSE, for $\tilde{f}_{9,2}(t)$ and $\tilde{f}_{10,2}(t)$ does not differ significantly, suggesting that choosing models based on the smallest RMSE does not highlight a significant difference between the hierarchical with MLR and (tdfp) and the MLR model for forecasting $A_2(t)$.

ESI_3: Table 5-15 and Table 5-16 demonstrates that the selected MLR model for forecasting $A_3(t)$ performs better than other models in terms of both MAE and RMSE. It reveals that $\tilde{f}_{9,3}(t)$ which is the forecasted value of $A_3(t)$ using the M_5 model that contains variables of hour of the day, day of the week, holiday, and interaction hour of the day and day of the week yields superior outcomes in terms of both MAE and RMSE.

ESI_4: From Table 5-17 and Table 5-18, for forecasting $A_4(t)$, ensembled model is selected due to the lowest RMSE and lowest MAE. This model is significantly different from all other models evaluated by RMSE. However, when comparing the MAE for the forecasted $A_4(t)$, the difference between the ensemble model and the M_5 model is not significant.

ESI_5: Finally, Table 5-9 demonstrates that the hierarchical forecasting with MLR and (tdfp) model performs better than other models when evaluated using MAE. Table 5-19 indicates

that, in terms of MAE, this model significantly outperforms others used to predict$A_5(t)$. Table 5-20 shows that, regarding RMSE, the distribution for differences between forecasting errors for none models to forecast $A_5(t)$, and $\tilde{f}_{9,3}(t)$ is not normal distribution. It reveals that the selected MLR model for forecasting $A_5(t)$ M_5 leads to smaller RMSE; hence, it is the best-selected model among other presented models.

To sum up, we conclude that forecasting ED arrivals for most of the ESI levels with MLR and hierarchical forecasting with MLR outperform in this case study. The differences between the results of proposed models for ESI_2, ESI_3, and ESI_4 are more significant. These ESI levels are more crowded than ESI_1 and ESI_5. On the other hand, the ESI_1 level is less crowded; it means that the randomness in this level is higher, and it is harder to find patterns. Therefore, the accuracy of the proposed models depends on the volume of the ED arrivals in ESI levels, which are the bottom level of the hierarchy. If the number of arrivals is higher at each ESI level, the accuracy of the proposed forecasting models is developed.

6. CONCLUSION

Preparing for the volume of arrivals by severity level is crucial for efficient hospital patient flow management [50], [51]. Given that each ESI level corresponds to patients with different needs, customized management strategies are essential. Additionally, the crowding of emergency departments is significantly exacerbated by the attendance of low-acuity patients. Furthermore, when analyzing the arrivals at each ESI level as separate time series, it is evident that the characteristics of each series vary, underscoring the importance of customized forecasting and management approaches for each severity level. As Figure 6.1 illustrates, our study proposes a solution to a real-world problem. We then apply real-world data to develop our solution, which involves proposing forecasting models to predict the number of hourly ED arrivals segmented by ESI level. We compare the performance of the proposed models with the state-of-the-art benchmark models that are widely applied in the healthcare industry. Ultimately, the forecasting results for the proposed model confirm the success of the proposed models.

Figure 6.1: Research Conclusion

This research stands out for its use of the rolling base technique, allowing for hourly forecasts for one-hour ahead for the entire year. Including an entire year in the forecast makes our

evaluation more accurate, not restricted to particular weeks, months, or seasons. This approach prevents errors in selecting the best forecasting model due to calendar changes. Additionally, the rolling base method ensures that the training data, encompassing three years of hourly data, is constantly updated, leading to a more accurate assessment of forecasting performance.

This dissertation concentrated on diverse methodologies for predicting emergency arrivals at each hospital ESI level. For each one-hour ahead forecasting, all the benchmark models (ARIMA, ETS, HTS with ARIMA, and HTS with ETS) and proposed models (MLR and hierarchical forecasting with MLR models) identified a distinct model for each ESI. Moreover, the selected benchmark models for each ESI are specific for each hour; for example, the selected model for hour 14:00 on December 12th, 2020, was different from the one for 14:00 on May 12th, 2020, due to the benchmark models selected based on AIC. However, the proposed models simplified this by using a single model for each ESI level, chosen for its lowest amount of MAE or RMSE in training data, for all hourly forecasts throughout the testing data. This means the same model applied for 14:00 on December 12th as for 14:00 on May 12th for ESI_1, is the same; however, it can be different for ESI_2 (ESI_3, ESI_4, ESI_5) in that time. This made the model selection and implementation for proposed models more straightforward.

This research provided a proper case study for applying a hierarchical forecasting model using linear regression for the top-level forecast with different reconciliation methods. Hierarchical forecasting models' advantage lies in their ability to produce coherent forecasts, a feature critical to any forecasting effort where maintaining consistency across various levels is essential.

Our proposed models outperformed this case study by analyzing thirteen different models. We applied the DM test to emphasize that the differences between selected models are significant and are not only chosen by measuring the forecast evaluation metrics. The MLR models and hierarchical forecasting models with MLR exhibit superior forecasting accuracy for most ESI levels. The linear regression forecasting models are coherent, too, meaning the sum of detailed forecasts aligns with the aggregated forecast at the top level (total ED arrivals). Furthermore, despite MLR models achieving the highest accuracy among all considered models, their simplicity in theory and application is undeniable.

The better performance of MLR forecasting models is due to the ability of regression models to capture the complex relationship between the independent and dependent variables. The proposed models capture daily, weekly, and yearly seasonality by considering these variables as categorical variables. However, *ETS()* and *auto.arima()* functions with default settings that are used for exponential smoothing and ARIMA forecasting model as benchmarks are able to capture limited seasonality. Therefore, if the time series is long, there will be different seasonality patterns that *ETS()* and *auto.arima()* functions are not capable of capturing all of them. The other important reason for outperforming the MLR models relies on the method we use for model selection. We select the best MLR model with a validation year, which is not an in-sample model-selecting method. This attribute makes them exceptionally reliable for forecasting purposes. On the other hand, the model selection for *ETS()* and *auto.arima()* are based on minimizing the AIC. AIC evaluates different models based on the goodness of the in-sample fit and also simplicity. For instance, if the performance of one ARIMA is close to a seasonal ARIMA, or the improvement of the seasonal ARIMA is not too much, AIC will select the simpler model, which is ARIMA.

Therefore, when we implement forecasting one-hour ahead for entire hours of a year (28280-point forecast), the accuracy of the model will be decreased.

To further advance our findings and address existing limitations, we have identified five main areas for improvement in our research. First, finding the right forecasting time horizon is essential. It can be examined if the accuracy of the model will be changed in other forecasting horizons. A different short-term forecast allows us to adjust plans more effectively for future needs. Second, we need to precisely determine the ideal amount of training data for each model. However, we focused on identifying the most effective models using a fixed amount of data. Therefore, rather than varying the amount of training data, we consistently used as much historical data as we had available for all models (three years of training data and one year of testing). Third, our contribution is focused on implementing MLR and hierarchical forecasting with MLR models comparable to univariant benchmark models, exponential smoothing, and ARIMA. Therefore, our analysis has not included any additional exogenous independent variables such as temperature, precipitation, snowfall, or air pollution. However, future studies can explore the impact of adding some other independent variables to the accuracy of the MLR models. These improvements will help make our research more practical and quicker to use, offering valuable paths for further development. Another suggestion for further development for this research is considering different structures for hierarchical data or applying different grouped time series that have the same behaviors. Finally, instead of the three well-known reconciliation methods for the top-down reconciliation method, other reconciliation methods can be considered.

REFERENCES

[1] V. G. Prabhu, "Improving Patient Safety, Patient Flow and Physician Well-Being in Emergency Departments," *All Dissertations*, Aug. 2022, Accessed: Feb. 05, 2024. [Online]. Available: https://tigerprints.clemson.edu/all_dissertations/3147

[2] "Crowding | ACEP." Accessed: Feb. 05, 2024. [Online]. Available: https://www.acep.org/patient-care/policy-statements/crowding

[3] S. Oueida, "Modeling a New Computer Framework for Managing Healthcare Organizations: Balancing and Optimizing Patient Satisfaction, Owner Satisfaction, and Medical Resources," *Modeling a New Computer Framework for Managing Healthcare Organizations: Balancing and Optimizing Patient Satisfaction, Owner Satisfaction, and Medical Resources*, pp. 1–212, Jan. 2020, doi: 10.4324/9781003027836/MODELING-NEW-COMPUTER-FRAMEWORK-MANAGING-HEALTHCARE-ORGANIZATIONS-SORAIA-OUEIDA.

[4] N. R. Hoot and D. Aronsky, "Systematic Review of Emergency Department Crowding: Causes, Effects, and Solutions," *Ann Emerg Med*, vol. 52, no. 2, pp. 126-136.e1, Aug. 2008, doi: 10.1016/J.ANNEMERGMED.2008.03.014.

[5] J. L. Wiler, R. T. Griffey, and T. Olsen, "Review of Modeling Approaches for Emergency Department Patient Flow and Crowding Research," *Academic Emergency Medicine*, vol. 18, no. 12, pp. 1371–1379, Dec. 2011, doi: 10.1111/J.1553-2712.2011.01135.X.

[6] V. G. Prabhu, K. Taaffe, R. Pirrallo, W. Jackson, and M. Ramsay, "Physician Shift Scheduling to Improve Patient Safety and Patient Flow in the Emergency Department," *Proceedings - Winter Simulation Conference*, vol. 2021-December, 2021, doi: 10.1109/WSC52266.2021.9715398.

[7] V. G. Prabhu, K. Taaffe, R. Pirrallo, W. Jackson, M. Ramsay, and J. Hobbs, "Forecasting Patient Arrivals and Optimizing Physician Shift Scheduling in Emergency Departments," *Proceedings - Winter Simulation Conference*, pp. 1136–1147, 2023, doi: 10.1109/WSC60868.2023.10407202.

[8] M. Gul and E. Celik, "An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments," *Health Systems*, vol. 9, no. 4, pp. 263–284, Oct. 2020, doi: 10.1080/20476965.2018.1547348.

[9] V. Girishan Prabhu, K. Taaffe, R. G. Pirrallo, W. Jackson, and M. Ramsay, "Overlapping shifts to improve patient safety and patient flow in emergency departments," *https://doi.org/10.1177/00375497221099547*, Jun. 2022, doi: 10.1177/00375497221099547.

[10] W. Whitt and X. Zhang, "Forecasting arrivals and occupancy levels in an emergency department," *Oper Res Health Care*, vol. 21, pp. 1–18, Jun. 2019, doi: 10.1016/J.ORHC.2019.01.002.

[11] L. M. Schweigler, J. S. Desmond, M. L. McCarthy, K. J. Bukowski, E. L. Ionides, and J. G. Younger, "Forecasting models of emergency department crowding," *Academic Emergency Medicine*, vol. 16, no. 4, pp. 301–308, Apr. 2009, doi: 10.1111/J.1553-2712.2009.00356.X.

[12] L. V. Green, J. Soares, J. F. Giglio, and R. A. Green, "Using queueing theory to increase the effectiveness of emergency department provider staffing," *Academic Emergency Medicine*, vol. 13, no. 1, pp. 61–68, Jan. 2006, doi: 10.1197/J.AEM.2005.07.034.

[13] N. Hoot, L. LeBlanc, I. Jones, S. Levin, … C. Z.-A. of emergency, and undefined 2008, "Forecasting emergency department crowding: a discrete event simulation," *Elsevier*, Accessed: Dec. 10, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0196064407018604?casa_token=FnBAYGiI4W8AAAAA:m6erKx1ivyBhuvTO9pAKe_Y0lLqlt9v548x-BGaBpQQhskOCQJycwMWDTGeg1Tx68YVAeeiyYs8

[14] M. Hertzum, "Forecasting Hourly Patient Visits in the Emergency Department to Counteract Crowding," *The Ergonomics Open Journal*, vol. 10, no. 1, pp. 1–13, Apr. 2017, doi: 10.2174/1875934301710010001.

[15] S. Jiang, K. S. Chin, and K. L. Tsui, "A universal deep learning approach for modeling the flow of patients under different severities," *Comput Methods Programs Biomed*, vol. 154, pp. 191–203, Feb. 2018, doi: 10.1016/J.CMPB.2017.11.003

[16] H. Batal, J. Tench, S. McMillan, J. Adams, and P. S. Mehler, "Predicting Patient Visits to an Urgent Care Clinic Using Calendar Variables," *Academic Emergency Medicine*, vol. 8, no. 1, pp. 48–53, Jan. 2001, doi: 10.1111/J.1553-2712.2001.TB00550.X.

[17] M. L. Mccarthy, S. L. Zeger, R. Ding, D. Aronsky, N. R. Hoot, and G. D. Kelen, "The challenge of predicting demand for emergency department services," *Wiley Online LibraryML McCarthy, SL Zeger, R Ding, D Aronsky, NR Hoot, GD KelenAcademic Emergency Medicine, 2008•Wiley Online Library*, vol. 15, no. 4, pp. 337–346, Apr. 2008, doi: 10.1111/j.1553-2712.2008.00083.x.

[18] M. Carvalho-Silva, M. T. T. Monteiro, F. de Sá-Soares, and S. Dória-Nóbrega, "Assessment of forecasting models for patients arrival at Emergency Department," *Oper Res Health Care*, vol. 18, pp. 112–118, Sep. 2018, doi: 10.1016/J.ORHC.2017.05.001.

[19] S. S. Jones, A. Thomas, R. S. Evans, S. J. Welch, P. J. Haug, and G. L. Snow, "Forecasting Daily Patient Volumes in the Emergency Department," *Academic Emergency Medicine*, vol. 15, no. 2, pp. 159–170, Feb. 2008, doi: 10.1111/J.1553-2712.2007.00032.X.

[20] Y. Sun, B. H. Heng, Y. T. Seow, and E. Seow, "Forecasting daily attendances at an emergency department to aid resource planning," *BMC Emerg Med*, vol. 9, no. 1, pp. 1–9, Jan. 2009, doi: 10.1186/1471-227X-9-1/FIGURES/5.

[21] G. Gafni-Pappas and M. Khan, "Predicting daily emergency department visits using machine learning could increase accuracy," *Am J Emerg Med*, vol. 65, pp. 5–11, Mar. 2023, doi: 10.1016/J.AJEM.2022.12.019.

[22] A. Choudhury and E. Urena, "Forecasting hourly emergency department arrival using time series analysis," *British Journal of Health Care Management*, vol. 26, no. 1, pp. 34–43, Jan. 2020, doi: 10.12968/BJHC.2019.0067/ASSET/IMAGES/LARGE/BJHC.2019.0067_F04.JPEG.

[23] M. J. Côté, M. A. Smith, D. R. Eitel, and E. Akçali, "Forecasting emergency department arrivals: a tutorial for emergency department directors.," *Hosp Top*, vol. 91, no. 1, pp. 9–19, 2013, doi: 10.1080/00185868.2013.757962.

[24] S. L. Wickramasuriya, G. Athanasopoulos & Rob, and J. J. Hyndman, "Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace

Minimization," *J Am Stat Assoc*, vol. 114, pp. 804–819, 2019, doi: 10.1080/01621459.2018.1448825.

[25] R. R. Andrawis, A. F. Atiya, and H. El-Shishiny, "Combination of long term and short term forecasts, with application to tourism demand forecasting," *Int J Forecast*, vol. 27, no. 3, pp. 870–886, Jul. 2011, doi: 10.1016/J.IJFORECAST.2010.05.019.

[26] G. Athanasopoulos, R. J. Hyndman, N. Kourentzes, and F. Petropoulos, "Forecasting with temporal hierarchies," *Eur J Oper Res*, vol. 262, no. 1, pp. 60–74, Oct. 2017, doi: 10.1016/J.EJOR.2017.02.046.

[27] T. Di Fonzo and D. Girolimetto, "Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives," *Int J Forecast*, vol. 39, no. 1, pp. 39–57, Jan. 2023, doi: 10.1016/J.IJFORECAST.2021.08.004.

[28] G. Athanasopoulos, P. Gamakumara, A. Panagiotelis, R. J. Hyndman, and M. Affan, "Hierarchical Forecasting," 2019, Accessed: Dec. 14, 2023. [Online]. Available: http://business.monash.edu/econometrics-and-business-statistics/research/publications

[29] C. W. Gross and J. E. Sohl, "Disaggregation methods to expedite product line forecasting," *J Forecast*, vol. 9, no. 3, pp. 233–254, 1990, doi: 10.1002/FOR.3980090304.

[30] R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, and H. L. Shang, "Optimal combination forecasts for hierarchical time series," *Comput Stat Data Anal*, vol. 55, no. 9, pp. 2579–2589, Sep. 2011, doi: 10.1016/J.CSDA.2011.03.006.

[31] R. J. Hyndman and G. Athanasopoulos, "Optimally Reconciling Forecasts in a Hierarchy," *Foresight: The International Journal of Applied Forecasting*, no. 35, pp. 42–48, 2014, Accessed: Dec. 08, 2023. [Online]. Available: https://ideas.repec.org/a/for/ijafaa/y2014i35p42-48.html

[32] G. Athanasopoulos, R. A. Ahmed, and R. J. Hyndman, "Hierarchical forecasts for Australian domestic tourism," *Int J Forecast*, vol. 25, no. 1, pp. 146–166, Jan. 2009, doi: 10.1016/J.IJFORECAST.2008.07.004.

[33] A. Panagiotelis, G. Athanasopoulos, P. Gamakumara, and R. J. Hyndman, "Forecast reconciliation: A geometric view with new insights on bias correction," *Int J Forecast*, vol. 37, no. 1, pp. 343–359, Jan. 2021, doi: 10.1016/J.IJFORECAST.2020.06.004.

[34] R. Hollyman, F. Petropoulos, and M. E. Tipping, "Understanding forecast reconciliation," *Eur J Oper Res*, vol. 294, no. 1, pp. 149–160, Oct. 2021, doi: 10.1016/J.EJOR.2021.01.017.

[35] M. Xu, T. C. Wong, and K. S. Chin, "Modeling daily patient arrivals at Emergency Department and quantifying the relative importance of contributing variables using artificial neural network," *Decis Support Syst*, vol. 54, no. 3, pp. 1488–1498, Feb. 2013, doi: 10.1016/J.DSS.2012.12.019.

[36] V. K. Sudarshan, M. Brabrand, T. M. Range, and U. K. Wiil, "Performance evaluation of Emergency Department patient arrivals forecasting models by including meteorological and calendar information: A comparative study," *Comput Biol Med*, vol. 135, p. 104541, Aug. 2021, doi: 10.1016/J.COMPBIOMED.2021.104541.

[37] Y. Zhang, J. Zhang, M. Tao, J. Shu, and D. Zhu, “Forecasting patient arrivals at emergency department using calendar and meteorological information,” *Applied Intelligence*, vol. 52, no. 10, pp. 11232–11243, Aug. 2022, doi: 10.1007/S10489-021-03085-9/TABLES/5.

[38] “Forecasting: Principles and Practice (3rd ed).” Accessed: Dec. 07, 2023. [Online]. Available: https://otexts.com/fpp3/

[39] R. Hyndman, A. Koehler, K. Ord, and R. Snyder, *Forecasting with Exponential Smoothing*. in Springer Series in Statistics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. doi: 10.1007/978-3-540-71918-2.

[40] R. J. Hyndman, A. B. Koehler, R. D. Snyder, and S. Grose, “A state space framework for automatic forecasting using exponential smoothing methods,” *Int J Forecast*, vol. 18, no. 3, pp. 439–454, Jul. 2002, doi: 10.1016/S0169-2070(01)00110-8.

[41] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin, “How sure are we that economic time series have a unit root?*,” *J Econom*, vol. 54, pp. 159–178, 1992.

[42] D. N. Gujarati and D. C. Porter, “Basic Econometrics (5th ed.),” *Basic Econometrics*, p. 946, 2009.

[43] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, “An Introduction to Statistical Learning,” 2023, doi: 10.1007/978-3-031-38747-0.

[44] R. J. Hyndman and A. B. Koehler, “Another look at measures of forecast accuracy,” *Int J Forecast*, vol. 22, no. 4, pp. 679–688, Oct. 2006, doi: 10.1016/J.IJFORECAST.2006.03.001.

[45] M. Tests and F. X. Diebold, “Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests,” *Source: Journal of Business & Economic Statistics*, vol. 33, no. 1, pp. 1–9, 2015, doi: 10.1080/07350015.2014.983236.

[46] “Package ‘fable’ Title Forecasting Models for Tidy Time Series,” 2023.

[47] “Package ‘forecast’ Title Forecasting Functions for Time Series and Linear Models Description Methods and tools for displaying and analysing univariate time series forecasts including exponential smoothing via state space models and automatic ARIMA modelling,” 2023, Accessed: Dec. 08, 2023. [Online]. Available: https://orcid.org/0000-0002-3665-9021

[48] R. J. Hyndman and Y. Khandakar, “Automatic Time Series Forecasting: The forecast Package for R,” *J Stat Softw*, vol. 27, no. 3, pp. 1–22, Jul. 2008, doi: 10.18637/JSS.V027.I03.

[49] R. J. Hyndman, G. Athanasopoulos, and H. L. Shang, “hts: An R Package for Forecasting Hierarchical or Grouped Time Series”.

[50] V. G. Prabhu *et al.*, “How Does Imaging Impact Patient Flow in Emergency Departments?,” in *Proceedings of Winter Simulation Conference (WSC)*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 961–972. doi: 10.1109/WSC57314.2022.10015455.

[51] V. G. Prabhu, K. Taaffe, and R. Pirrallo, “Patient Care Management for Physicians: Reducing Handoffs in the ED,” in *Proceedings - Winter Simulation Conference*, Institute of Electrical and Electronics Engineers Inc., Dec. 2019, pp. 1126–1136. doi: 10.1109/WSC40007.2019.9004784.