Critical Analysis of Machine Learning and Deep Learning Models for Predicting Medical Equipment Demand in Hospitals

MSc Research Practicum- Part II
MSc in Data Analytics

Raghul Sureshbabu
Student ID: x23377755

School of Computing
National College of Ireland

Supervisor: Prof.Vikas Sahani

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Raghul Sureshbabu		
Student ID:	x23377755		
Programme:	MSc in Data Analytics	**Year:**	2026
Module:	Research Practicum-Part II		
Supervisor:	Prof. Vikas Sahani		
Submission Due Date:	28/01/2026		
Project Title:	Critical Analysis of Machine Learning and Deep Learning Models for Predicting Medical Equipment Demand in Hospitals		
Word Count:	7385	**Page Count:**	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Raghul Sureshbabu

Date: 26/01/26

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Critical Analysis of Machine Learning and Deep Learning Models for Predicting Medical Equipment Demand in Hospitals

Raghul Sureshbabu
x23377755

Abstract

Accurate forecasting of medical equipment demand is a major challenge in hospital management. This paper addresses this gap by conducting a critical analysis of statistical, machine learning, and deep learning models for forecasting Medicare Durable Medical Equipment (DME) demand from 2014 to 2022. This analysis incorporates SARIMAX as a statistical baseline, and other algorithms such as XGBoost, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and N-BEATS in order to predict the complex temporal dependencies with SHAP based explainability for interpretation.

In the primary experiments, XGBoost achieves the highest accuracy (R2 = 0.9812) and N-BEATS produces a similar result (R2 = 0.9801), where both clearly outperform other models. An additional experiment with recently released 2023 data confirms that XGBoost continued to show strong predictive accuracy (R2 = 0.9883), while N-BEATS performance decreased to 97% indicating that gradient boosting is more accurate and stable than other algorithms. This study is useful for the healthcare operations area by recommending a data-driven and interpretable forecasting approach that can assist in trustworthy and ethically consistent decision-making in the planning and procurement of medical equipment in hospitals.

1 Introduction

Accurate demand prediction of medical equipment is an important factor in the management of healthcare operations, as it directly influences the hospital budgeting, allocation of resources, and efficiency in patient care. Inaccurate predictions may lead to either a shortage of necessary equipment, including ventilators and wheelchairs or a surplus of expensive devices with both having serious financial and clinical implications (Zhou et al., 2020; Jin et al., 2025). The growing use of electronic health records and open-access healthcare data has offered new possibilities for hospitals and policymakers to use data analytics and predictive modelling in order to make informed decisions (Pampa et al., 2025). Moreover, the COVID-19 pandemic revealed the vulnerability in the medical supply chain, making the process of intensive and data driven forecasting systems critically important in enhancing timely access to necessary medical supplies (Koc & Turkoglu, 2023).

The Medicare Durable Medical Equipment (DME) data provided by the Centres for Medicare and Medicaid Services (CMS) in the United States is an important source of information on long-term consumption patterns of medical equipment between 2014 and 2022. The evaluation of these time and seasonal trends can inform healthcare administrators to make more evidence-based procurement and planning decisions.

1.1 Motivation

Although the importance of predictive analytics in healthcare is increasing, demand forecasting of medical equipment is a relatively under-researched field compared to patient admission forecasting or disease outbreak forecasting (Tuominen et al., 2024; Menebo, 2025). This lack of focus has created a knowledge gap in the understanding of how advanced forecasting models can be used to assist in the long-term resource planning of durable medical equipment (DME), which is an essential component of hospital logistics and patient rehabilitation. Medical equipment demand forecasting is particularly critical to ensure that resources are distributed equally and the financial system remains stable. However, modern healthcare systems do not usually have the tools to perform such accurate forecasting (Walter et al., 2024).

1.2 Limitations

ARIMA and SARIMAX are traditional statistical models commonly applied to healthcare forecasting due to their interpretability and simplicity (Tamatta, 2021; Zhou et al., 2020). These models are capable of capturing linear relationships, but cannot capture non-linear time-based dynamics of healthcare data (Tuominen et al., 2024). Alternatively, machine learning and deep learning models such as XGBoost, LSTM, GRU, and N-BEATS are more flexible and predictive than the traditional models (Jin et al., 2025; Guo et al., 2025; Pampa et al., 2025) Nevertheless, extensive comparative studies of these models within the framework of DME forecasting are still limited, and minimal advances have been achieved in solving the explainability issues that prevent their use in the context of real-world healthcare. (Telesca & Rondinone, 2025; Hu et al., 2025). This research assists in addressing these gaps by integrating Explainable Artificial Intelligence (XAI) technique such as SHAP into the ML and DL models and develop an effective, transparent, and ethical model for medical equipment demand forecasting.

1.3 Research Question and Objectives

This research aims to address these research gaps by answering the following question:

"To what extent can advanced machine learning and deep learning models like XGBoost, LSTM, GRU and N-Beats outperform traditional statistical approaches like SARIMAX in forecasting the medical equipment demand, and how can such improvements enhance health care resource planning and operational efficiency?"

Objectives:

1. Examine the Medicare DME data to determine a time and seasonal demand trend.
2. Design and implement statistical (SARIMAX), machine learning (XGBoost) and deep learning (LSTM, GRU and N-BEATS) forecasting models.

https://data.cms.gov/provider-summary-by-type-of-service/medicare-durable-medical-equipment-devices-supplies/medicare-durable-medical-equipment-devices-supplies-by-geography-and-service

3. Measure and compare the outcome of these models by the RMSE (Root Mean Square Error), MAE (Mean Absolute Error) and R2 (co- coefficient of determination).
4. Explainable AI techniques such as SHAP will be used to improve the model transparency and its explainability.
5. Evaluate both ethical and operational concerns of deploying AI based forecasting software in healthcare environments.

1.4 Structure of the Report

Section 2-Related Work: Summarizes the past contributions in the field of statistical, machine learning (ML), and deep learning (DL) forecasting models in medical care, focuses particularly in the research of medical equipment demand and Explainable Artificial Intelligence (XAI).
Section 3-Research Methodology: Describes the dataset, preprocessing techniques, experimental design, and evaluation metrics employed in developing the forecasting models.
Section 4-Design Specification: Deploys the framework and the technical contents of the constructed forecasting framework, including the integration of the XAI techniques.
Section 5-Implementation: Presents implementation, model training, parameter adjustment, and computing environment.
Section 6-Evaluation: Analyzes the performance result of the models, compares the traditional and the advanced models and discusses the result based on interpretability and practical relevance.
Section 7-Conclusion and Future Work: Summarizes the key findings, limitations and directions on future research.

2 Related Work

The development of healthcare forecasting has been transformed into non-linear and temporal dependencies that are modeled by more advanced machine-learning (ML) and deep-learning (DL) techniques compared to more traditional statistical models. This section critically analyzes previous studies related to the medical-equipment demand prediction and how statistical, Machine learning and Deep learning methods are evaluated and placed in this evolving research environment.

2.1 Statistical Forecasting Models

2.1.1 Baseline Statistical Models

The classical time-series models like ARIMA, SARIMAX, and Exponential Smoothing are still basic in the short-term healthcare prediction. Zhou et al. (2020) showed the ability of ARIMA to capture the trend of hospital admission with coefficients that can be interpreted and with a low computational cost. Tamatta (2021) applied this framework to hospital waiting lists, finding that it worked better on a daily basis, but struggled with adjusting to regime changes. On the same note, Elgohary (2019) employed ARIMA and ETS to predict the number of hospital visits involving outpatients in Egypt, and the findings still confirm that statistical methods can work well when data are stationary and seasonal variations are constant. These

works have presented credible foundations with respect to which transparent parameters are useful in decision making at a managerial level.
Nevertheless, the models presuppose a linearity and a stable variance that results in the poor performance of volatile or nonlinear medical demand. They can be rarely applied to exogenous variable and lose accuracy over long durations, which restricts their application to many-factor settings like emergency departments.

2.1.2 Early Hybrid Statistical-Machine Learning Models

Hybrid paradigm was first proposed by Zhang (2003) who combined the linear aspect of ARIMA with a feed forward neural network to model nonlinear residuals. This theoretical breakthrough led to the development of hundreds of ARIMA-ANN and ARIMA-SVM implementations in fields and provided error decreases of moderate magnitude relative to purely statistical models. Levine et al. (2008) in healthcare had addressed data-driven demand forecasting of critical medical technologies where hybrid predictive models could prove useful in inventory management. However, the majority of early hybrids were manually tuned, were case-specific and had no theoretical generalisation.

2.1.3 Limitations of Statistical and Early Hybrid Models

Early-hybrid and traditional models are also restricted to strict linear models, lack scalability and poor multivariate processing. They have a disadvantage in that their ability to interpret is poor in high-dimensional data or non-stationary data. The study aims to fill these research gaps by combining SARIMAX (as a structured temporal explainability) with XGBoost and with deep architectures, such that nonlinear learning may be achieved without losing statistical information specified by SHAP-based interpretability.

2.2 Traditional Machine Learning Models

2.2.1 Tree and Boosting Ensembles

Ensemble learners have been shown to have great predictive power in terms of multivariate forecasting. XGBoost was used by Menebo (2025) where the authors used meteorological parameters to predict medication shortages, but they found higher accuracy than ARIMA, though there is also a lack of visibility in the features of the prediction. The review of ML and DL logistics forecasting conducted by Pampa et al. (2025) shows that gradient-boosting optimization is better suited in tabular demand forecasting. The findings reported by Menebo confirmed that, including external data, i.e. web searches or weather enhances responsiveness to real-world factors (Zhou, 2025). The study by Xu and Chan (2019) affirmed the previously mentioned claim. Although Walter et al. (2024) highlighted the forecasting ability of AI in supply-chain, they also cautioned against the issue of lack of explainability that obstructs trust in operations.

Research by Pahlevani et al. (2024) and Lin et al. (2019) indicated that random forests have the ability to minimize the error of wait-time forecasting in hospitals, but this improvement is largely determined by the quality of features. There are weak yet topical works that demonstrates that pure ML overfits in cases when it is time-correlated as in Elgohary (2019) and Cortez and Morais (2007).

Some of the studies highlight the importance of integrating exogenous variables. Both Menebo (2025) and Xu and Chan (2019) showed that weather- and search-engine-based indicators of online searches improve medical demand models. Similar integration of medical-device forecasting was proposed earlier by Levine et al. (2008). All of this demonstrates that multivariate designs increase flexibility, but not all of them quantify the contribution of each predictor, which is again bridged in the study through SHAP interpretability to visualise the impact of each variable.

2.2.2 Limitations of Machine Learning Models

Ensemble methods are useful at dealing with nonlinear, multivariate information, although they have low transparency and short-horizon bias. The black-box nature makes them be unable to be accepted in sensitive fields like healthcare. This research aims to minimize this limitation by the correlation of SHAP-based layers to the ML pipeline to generate explainable attributions on time.

2.3 Deep Learning Models

2.3.1 Recurrent Architectures (LSTM, GRU)

Recurrent networks are good at long-term modelling. Tuominen et al. (2024) used multivariable LSTM models to predict emergency-department occupancy which outperformed ARIMA in terms of a significant improvement in accuracy. Predicting Valley Fever infections, Jin et al. (2025) revealed the advantage of using memory-cells structures to predict epidemiological data in seasonal data. Shafiekhani et al. (2022) created an ANFIS-LSTM hybrid, which better predicted COVID-19 hospital-resource predictions but with limited interpretability. The implementation of structure-aware temporal modelling of chronic-disease progression by Hu et al. (2025) introduced the indication that the architectures design is a fundamental concern to the degree of reliability. All of these studies demonstrate the superiority of recurrent networks in temporal dynamics with limited transparency.

2.3.2 Non-Recurrent and Transformer Architectures (N-BEATS, TFT)

Recurrent models that appeared recently have performed better than LSTMs on complex time series. Yang et al. (2025) compared ARIMA and the Temporal Fusion Transformer in the prediction of national incidence of depression, finding significant improvements in accuracy but with a higher computing cost and poor accuracy interpretation. Kong et al. (2025) also gave a complete overview of Deep-learning forecasting, noting only the scalability but pointing to a strong demand of explainable mechanisms.

2.3.3 Limitations of Deep Learning Models

In spite of effective accuracy, deep models possess the drawbacks of being confusing, consumptive of data, and prone to overfitting. Their performance requires big, labeled datasets which are not easily available in healthcare. The study aims to address these questions by using SHAP to come up with deep interpretable forecasting.

2.4 Hybrid and Explainable AI Frameworks

2.4.1 Model Combination and Integrative Forecasting

Integrating different model approaches makes the system more robust and adaptable. Koc & Turkoglu (2023) followed a deep LSTM on demand of medical-equipment in times of COVID-19 ascertaining that combined model adds adaptability in the crisis data condition. Guo et al. (2025) suggested ensemble combinations to demand in emergency-department, and it was shown that it was stable to the single-model variation. Collectively, such studies approve of multi-modelling in healthcare predictions.

2.4.2 Explainable AI in Healthcare Forecasting

Competence is required in ethical deployment. The interpretable ML framework created by Telesca & Rondinone (2025) to establish the relationship between cardiopulmonary disease and pollution is one example of model-agnostic explainability. The topic of AI governance and risk in supply chains of Industry 4.0/5.0 remained approached by Rane et al. (2025), whilst Walter et al. (2024) emphasised the importance of explainability when it comes to establishing organisational trust. These findings support the incorporation of SHAP into the hybrid forecasting model of this study to promote responsible AI in medical settings.

2.4.3 Limitations of Hybrid and XAI Frameworks

The current hybrid and XAI research are still disjointed, domain-focused, and hardly tested on real-world hospital data. Computational consistency and cross-model scalability are not taken into consideration. The given study fills these gaps with the integrated SARIMAX-XGBoost-DL framework that was tested on multivariate healthcare data.

Table 1: Summary of Key Studies relevant to Healthcare and DME Forecasting

S.No	Author (Year)	Model	Domain	Key Findings	Limitations	Current Research Contribution
1	Zhou et al. (2020)	SARIMAX vs ETS	Hospital admission forecasting	Accurate and interpretable short-term dependence.	Assumes linearity and stationarity	Routine statistical benchmark of SARIMAX component.
2	Tamatta (2021)	ARIMA, TBATS, NN	Hospital waiting list	Seasonality is captured effectively	Non-stationary data, which is weak at regime shifts.	Highlights need for non-linear adaptation
3	Zhang (2003)	ARIMA–ANN hybrid	General time-series	Combined linear and nonlinear dynamics	Manually tuned, task-specific	Modern hybridisation conceptual basis.
4	Menebo (2025)	XGBoost with weather & policy variables	Medication demand forecasting	Good predictive ability on the multivariate data.	Minimal interpretability, the short-horizon problem.	Motivates integration of XGBoost with SHAP explainability

5	Tuominen et al. (2024)	LSTM, GRU	Emergency department occupancy	Multivariate time dependencies that are captured.	Black box nature, data intensive nature.	Temporal dynamics Supported by integration of DL.
6	Jin et al. (2025)	LSTM vs ARIMA	Infectious-disease forecast	Temporal accuracy was enhanced by base on LSTM.	Lack of interpretability	Deep-learning compares the information.
7	Koç & Türkoğlu (2023)	Deep LSTM	COVID-19 equipment demand	Efficient pandemic-time prediction.	No explainability mechanism.	Authenticates requirements of hybrid and XAI frameworks.
8	Telesca & Rondinone (2025)	Interpretable ML (XAI)	Cardiorespiratory health	Attained transparent and ethical inference.	Domain-specific model	Facilitates SHAP based-ethical interpretation.

2.5 Summary and Research Gap

The literature reviewed indicates that there has been gradual evolution from interpretable yet linear models of statistical inference (e.g. ARIMA, SARIMAX) to more predictive and opaque machine learning and deep learning models. Statistical models use is still useful in short-term predictions and transparency in parameters, but its nonlinear and multivariate characteristics are not observed in the actual healthcare data. Tree-based and boosting ensembles, including the XGBoost provide high predictive accuracy on multi-varied datasets, yet they do not follow the dynamics of the sequence of time and cannot offer much explainability. RNN-based and transformer-based deep learning models (e.g., LSTM, GRU, N-BEATS) are a well-established way of modelling the temporal dependencies but are known to be highly uninterpretable, have substantial data demands and have ethical consequences.

Current literature pays limited attention to the long-term hospital or durable medical equipment (DME) demand forecast, and very few studies combine interpretability or ethical considerations into model construction.

This indicates a two-fold gap:

(1) there is no overall, comparative structure that can bring together the paradigms of statistical, machine learning, and deep learning and

(2) there is a lack of transparency and ethically controlled AI-based decision-making in healthcare prediction

To overcome these limitations, the current study aims to create a unified hybrid forecasting model that involves the combination of SARIMAX, XGBoost, and deep learning models with SHAP-based explainability. The goal of the approach is to be able to make sound long-term DME demand forecasting, but in a way that is interpretable, ethically transparent and operationally reliable.

3 Research Methodology

The research followed a well-defined scientific methodology that was designed to provide reproducibility, validity, and consistency to all forecasting models. The process began with

the data acquisition where all Medicare Durable Medical Equipment (DME) Utilization Public Use Files between 2014 and 2022 were collected directly via the official CMS data portal. These nine files of annual utilisation were then compiled into a single dataset to form an analytical foundation after acquisition. The compilation step involved cleaning and preprocessing operations that involved the elimination of redundant administrative fields, the verification of numeric fields and the standardisation of schema variants between years.
The modelling process was then reinforced by performing feature engineering, which involved developing lag-based temporal indicators and maintenance of important structural healthcare utilisation attributes. The development proceeded with the implementation and training of five forecasting models comprising statistical, machine-learning and deep-learning families. The evaluation was conducted under the same testing conditions and multiple error measures to ensure accurate predictions. Finally, explainable AI techniques were used to help with the interpretation by determining which factors had the strongest influence on the model outputs. A combination of these steps developed an entire end-to-end scientific process, where raw administrative data were transformed into interpretable results of forecasting.

3.1 Data Collection, Scenario Setup, and Sample Preparation

3.1.1 Data Gathering and Compilation

The research dataset was collected from the CMS portal entitled Medicare Durable Medical Equipment, Devices & Supplies by Geography and Service. As the data show full utilisation of Medicare at a population level, there was no need to carry out any sampling or randomisation. File cleaning was performed on each year of the data which was then merged into a consistent master dataset following the schema alignment and consistency checks.

Scenario / Case Study Definition

The forecasting scenario was defined to predict the utilisation of DME for each combination of state, HCPCS code, and year. This framework developed a panel time-series forecasting environment, which enabled the study to explore both time sequence and interstate variation in equipment usage.

3.1.2 Data Preparation Steps

A number of preprocessing operations were performed to prepare the dataset for modelling. The standardisation of variable names and units was done across years, and utilisation values were aggregated at the State-HCPCS-Year level. Financial and utilisation fields were validated and converted, and non-critical administrative identifiers were eliminated. To ensure that time series was preserved, chronological order was maintained.

Train–Test Split

A chronological train-test split was used to guarantee forecasting validity. Data from 2014 to 2020 were used for training, and data from 2021 to 2022 were strictly used for testing. This approach was used for SARIMAX and XGBoost Model which operate on tabular aggregated data.

Deep learning models (LSTM, GRU, N-BEATS) were generated on each of the State-HCPCS time series and the sequences were divided into training and testing sets with an 80/20 random split. This was done due to the variation in length of sequences across groups and a constant approach of using number of years would have given inconsistent number of sequences. The split was performed after the generation of the sequences, to make sure that the deep learning models were trained and tested on similar sequence structures.

3.2 Techniques Applied and Model Development

The research developed and compared five forecasting approaches. SARIMAX was used as a classical statistical baseline to capture the linear time series behaviour. XGBoost offered a nonlinear gradient boosted machine learning algorithm, which is suited for structured tabular data. LSTM and GRU models represented recurrent neural network architectures that could model the sequential and temporal dependencies and N-BEATS represented a modern deep learning architecture that was optimised for the multi-step prediction tasks.

The feature structures differ across models, where XGBoost is trained on lagged features and engineered covariates and sequence-based inputs were used to train the deep learning models The train/test structure differed across models where SARIMAX and XGBoost used a chronological split whereas, the deep learning models used an 80/20 sequence-based split. Its application served as the underlying architecture, which was then heavily customised for the CMS DME data, feature engineering operations and explainability methods.

3.3 Additional experiment with latest year dataset

The base of the experimental design for this study is CMS DME data from 2014-2022. During the course of the project, the 2023 utilisation file has been released. Rather than reconsidering the original study window this dataset was used as an additional experiment (Analysis 2). Models that were developed and evaluated in Analysis 1 were tested on a extended testing period, which also involved observations of 2023, without modifying the used training setup. This experiment was intended to test which of the best performing models from Analysis 1 achieved strong predictive accuracy for the most recent year's data without further modifications to the model.

3.4 Data Analysis and Calculations Performed

To transform raw CMS utilisation records to a dataset that is ready for forecasting, multiple analytical procedures were performed. These included aggregating HCPCS level utilisation metrics, computing lag based temporal predictor, scaling of features to make them compatible with Neural networks, converting sequences into numeric matrices and 3D tensors and aligning previous forecasts for consistent evaluation across models. These operations ensured that common and standardised inputs were used for all forecasting approaches.

3.5 Statistical Techniques and Evaluation Methodology

All forecasting models were evaluated using a common test procedure and a set of widely used statistical indicators. Error magnitude was captured by RMSE, MAE, and MSE while the proportion of variance explained by the predictions was captured by R2. The models were tested on the same period (2021-2022) to ensure that comparisons were fair and based on the same methodology.

To make the model more interpretable and transparent, SHAP explainability was applied to the best-performing model. This enabled the study to determine the features that contributed most to the utilisation predictions across states and HCPCS categories, and it gave a deeper understanding of model behaviour and the underlying patterns of utilisation.

4 Design Specification

The system consists of three major components: the Data Processing Framework, the Modelling Framework and the Training and Evaluation Pipeline. Together, these layers provide the technical foundation for consistent preparation, execution and comparison of the types of forecasting models that are used in this study.

4.1 Data Processing Framework

The data processing layer is the foundation layer of the system architecture. Its main task is to convert raw CMS Durable Medical Equipment (DME) utilisation files into formats that are suitable for modelling. This involves ensuring the standardisation of yearly datasets, schema harmonisation, lag-based temporal feature generation and validation of numerical variables for accuracy.

For statistical and machine learning models such as SARIMAX and XGBoost, the framework prepares the clean data in the form of two-dimensional tabular datasets. For deep learning models, such as LSTM, GRU and N-BEATS, it converts the inputs to three-dimensional tensors, which represent samples, time steps and features. This design ensures that the correct input format is given to each model and the chronological integrity for time series forecasting is maintained.

4.2 Modelling Framework

The modelling framework, which is on top of the data layer provides a unified interface used to train and compare the prediction models. It standardises how engineered features are to be passed to each model, manages model-specific parameters and ensures consistency in configuration of the training and prediction output formatting.

Statistical, machine-learning and deep-learning models are easily incorporated into this framework. It also supports scalability as it allows for additional forecasting algorithms to be added to the system without having to modify the underlying pipeline.

4.3 Training and Evaluation Pipeline

The training and evaluation pipelines structure the implementation of the whole experimental workflow. It is responsible for chronologically splitting the training and testing data, batching and scaling operations, model execution routines, forecast generation and calculation of evaluation metrics. The general structure of this pipeline, particularly the processing of the multi-variable sequence data is inspired by the open - source forecasting architecture proposed by Tuominen et al. (2024). Their framework guided the design of the training and testing workflow and it influenced several aspects of the multivariate input handling design used in this study.

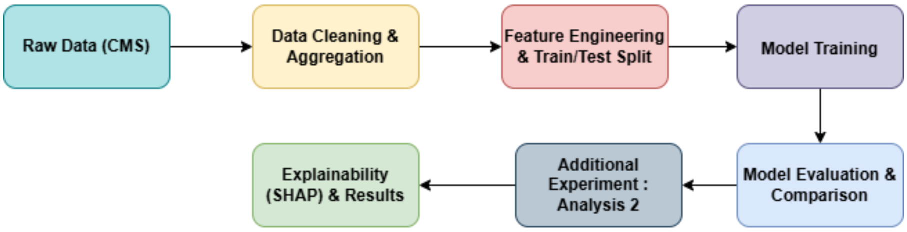

Figure 1: Data Pipeline

5 Implementation

The implementation stage is dedicated to the transformation of developed forecasting architecture into fully operational system that performs compiled dataset processing, training the prediction models and generating the final predicted outputs.

5.1 Final Dataset Transformation

The final phase of data preparation was to convert these multi-year raw CMS DME utilization files into a single aggregate data source for modelling. This involved standardising variable formats across years, cleaning and validating all numeric fields and eliminating fields that are not predictable. The dataset was further aggregated to the State-HCPCS-Year level to have a uniform analytic structure. Feature engineering was also incorporated as additional predictors in order to enhance the accuracy of the model. These included lag-1 and lag-2 utilisation variables and structural utilisation variables that included the counts of suppliers, beneficiaries, referring providers, allowed charges, and Medicare payments. Deep-learning models involved additional procedures, such as MinMax scaling of numerical variables and transformation of the dataset into 3D tensors organised as samples x time steps x features. The result of this stage was a complete cleaned and structured dataset ready for the final modelling stage.

5.2 Model Implementation and Forecast Generation

Five forecasting models were trained and implemented using the prepared dataset, SARIMAX, XGBoost, LSTM, GRU, and N-BEATS. SARIMAX and XGBoost were trained on the full chronological split (2014-2020) training, (2021-2022) testing. whereas, the deep learning models were trained on an 80/20 random split of constructed sequences. This methodological decision was required because the per-group time-series lengths vary among states and HCPCS codes and a consistent chronological cut-off cannot be used with sequence-based training.

All the forecasting models were trained on the historical part of the data to learn the underlying patterns and the annual variations of DME utilisation. After training, the models were applied to the reserved test period to make predictions of their utilisation forecasts. The outcomes of this stage were the final trained model artefacts; the output of the corresponding predictions and the accuracy measures are needed to compare the performance. All the models were trained and tested using a regular procedural pipeline, which was fair and consistent throughout the modelling process.

5.3 Explainability Outputs

Following the forecasts, an explainability layer was added to facilitate the interpretation of forecasts. The model that had the best performance was analysed using SHAP to determine which features it relied on the most to make predictions. This generated a series of interpretive outputs, such as feature-influence summaries and visual explanations, that were used for the clarification of the process through which the model understood utilisation patterns among states and procedure codes.

6 Evaluation

This section presents a detailed analysis of the forecasting outcomes obtained from the statistical, machine learning, and deep learning models. The assessment is based on the findings that address the research goal, namely to identify the most accurate modelling approach to predict Durable Medical Equipment (DME) utilisation patterns by state and HCPCS code. The evaluation of model behaviour and reliability is assessed using statistical performance metrics, visual diagnostics and explainability tools.

The evaluation was based on the forecasting outcomes of each model, where SARIMAX and XGBoost model metrics are based on the chronological split, whereas the deep learning model metrics are based on the sequential-based split
To facilitate interpretation, a set of plots and charts is included at critical steps of the analysis so that the trends, errors, and performance differences are well illustrated.

6.1 Experiment 1 - SARIMAX Model Evaluation

The SARIMAX model was implemented as the statistical baseline. It was trained on the aggregated annual data and tested with the full pattern of exogenous predictors.

Performance Results

RMSE: 1,345,961.80; MAE: 1,321,119.05; R^2: 0.0931

Interpretation

SARIMAX predicted 9.31% of the variance in DME utilisation, which suggests that a traditional linear time-series model does not work with highly granular, state- and service-level claims data. The large errors indicate non-stationarity, wide variance across HCPCS codes, and nonlinear interactions, which SARIMAX is unable to model.

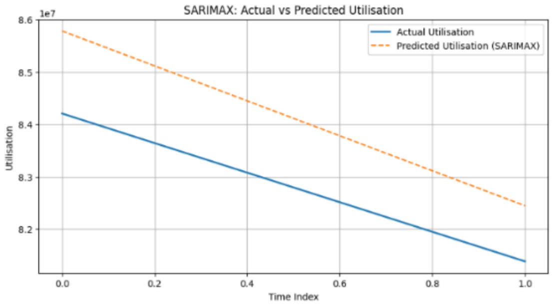

Figure 2: Actual vs Predicted Utilisation (SARIMAX)

SARIMAX is a required baseline but it is not suitable for high-dimensional healthcare utilisation forecasting.

6.2 Experiment 2 - XGBoost Model Evaluation

XGBoost was trained on the panel data using lag features (t-1, t-2), the count of suppliers, beneficiaries, service counts and payment variables

Performance Results

RMSE: 1,755.28; MAE: 211.84; R^2: 0.9812

This has the **highest accuracy among all models**.

Interpretation

The model predicts **98.12%** of the variation in DME utilisation, which is drastic when compared to SARIMAX. The predictive power of lag features was observed to be far more significant, which proves that past utilisation has a strong influence on future patterns.

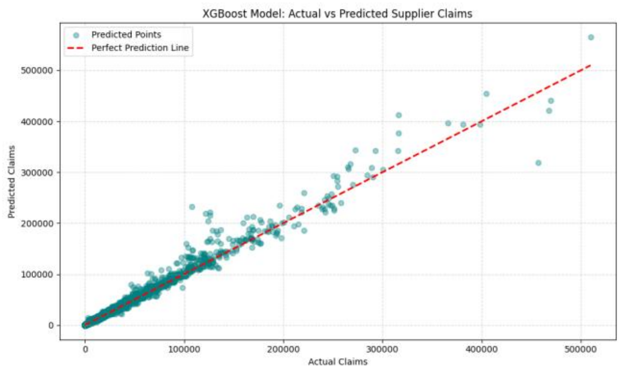

Figure 3: Actual vs Predicted supplier claims (XGBoost)

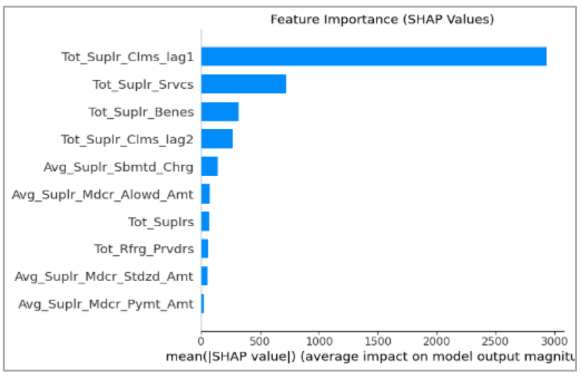

Figure 4: Feature Importance (XGBoost)

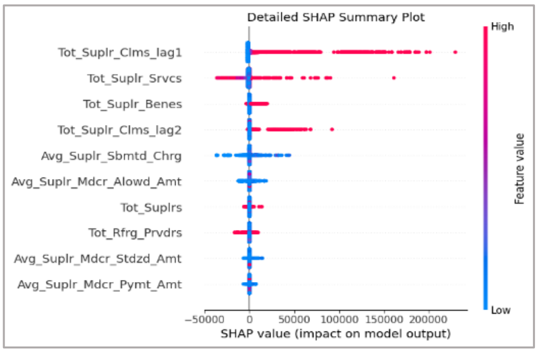

Figure 5: SHAP explainability AI (XGBoost)

6.3 Experiment 3 – LSTM Model Evaluation

The LSTM model was trained using multivariate sequences on a three-year lookback window

Performance Results

RMSE: 2,593.93; MAE: 541.49; R^2: 0.9650

Interpretation

LSTM was also effective and its variance was 96.5%. Even though it was very accurate, it could not outperform XGBoost since it was vulnerable to data scaling, magnitude variance and irregular spikes in HCPCS-level claims

6.4 Experiment 4 – GRU Model Evaluation

The GRU model was trained using a lookback window of three years and the same sequence structure and input features as LSTM. GRU gating mechanism is much simpler compared to that of the LSTM, where it results in faster computation while retaining the ability to model the sequential dependencies

Performance Results

RMSE: 2,216.66; MAE: 675.35; R^2: 0.9744

Interpretation

The GRU model showed good prediction performance, with a variance of 97.44% in DME utilisation. Its accuracy is slightly lower than XGBoost (0.9812) but slightly better than LSTM (0.9650) indicating that the GRU architecture generalises well to the irregular and highly variable nature of DME claims data.
Compared to LSTM, the GRU displayed:

- Similar ability to capture temporal patterns
- Fewer parameters,
- More efficient training which makes it an efficient deep learning candidate for healthcare time-series forecasting

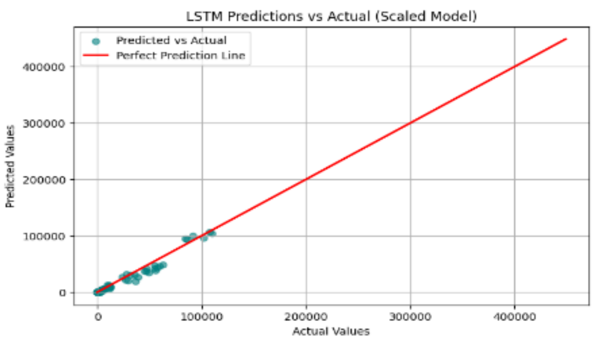

Figure 6: Actual vs Predicted (LSTM)

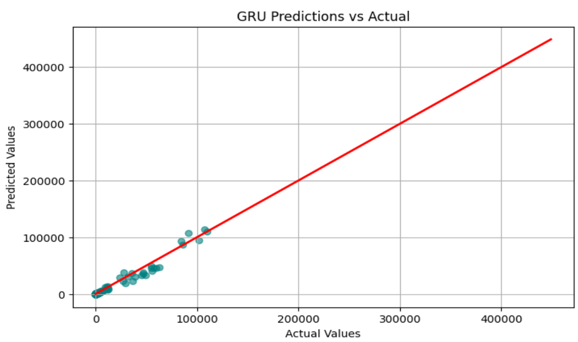

Figure 7: Actual vs Predicted (GRU)

GRU balances accuracy and computational efficiency and is better than LSTM but not as good as the XGBoost model, which is the most robust model overall.

6.5 Experiment 5 – N-Beats Model Evaluation

N-BEATS architecture was applied as the final deep learning model in this research. In comparison to LSTM and GRU models, which operate on the principle of sequential gating, N-BEATS model applies fully connected blocks, which explicitly learn the aspects of trend and seasonality of historical data. This renders it as a promising architecture to predict the activities with complex nonlinear and long-term time patterns

Performance Results

RMSE: 1,953.51; MAE: 517.92; R^2: 0.9801

Interpretation

The N-BEATS model performed quite well, with the variance of 98.01%, which ranks it extremely close to XGBoost and a little higher than LSTM and GRU. The error measures also indicate high predictive accuracy with lower values of RMSE and MAE compared to the recurrent neural network models. This means that N-BEATS could generate meaningful temporal patterns using the multivariate sequences without recurrent or convolutional operations. N-BEATS seemed to be especially useful in terms of smooth temporal dynamics and capturing non-linear relationships between the utilisation history and service-level characteristics. However, it was moderately sensitive to the large variation in the magnitude of DME claims, which probably led to a slight increase in error rates compared to XGBoost. Nevertheless, the performance shows that N-BEATS is an efficient healthcare predictive model, particularly when trained on extensive historical windows.

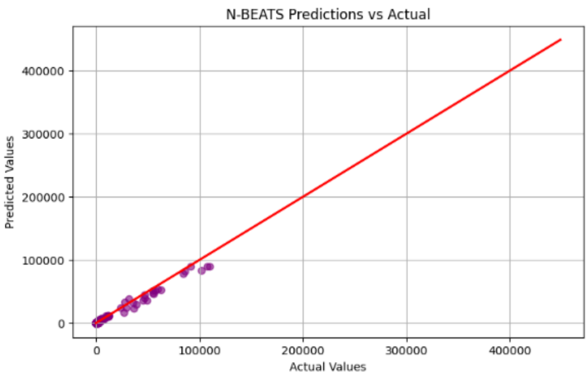

Figure 8: Actual vs Predicted (N-BEATS)

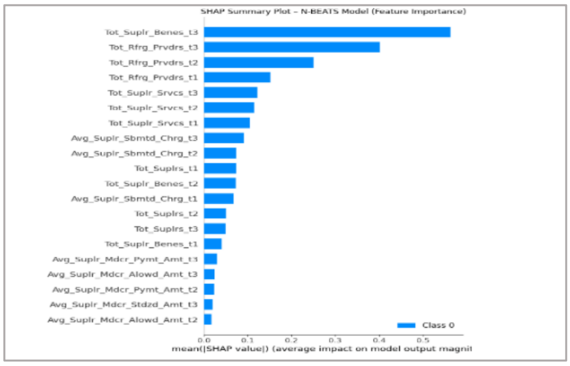

Figure 9: SHAP explainability AI (N-BEATS)

Compared to LSTM and GRU, N-BEATS is more powerful because of the combined features of deep learning and the interpretability of decomposition-based models, and it almost matches the performance of XGBoost. Its structure makes it highly applicable in forecasting operations involving long term historical patterns as long as there are adequate computational resources and training time.

6.6 Experiment 6 – XGBoost Extended Dataset Evaluation (2023-2024)

This additional experiment was performed after the main analysis with the latest year data (2023) on CMS DME utilisation. The purpose of this experiment was to analyze how strong the best performing model from the main analysis, XGBoost, is when the timeframe is extended to a full 10 years (2014 - 2023). The architecture of the models and the set of features that was used remained the same while the extended dataset was tested on 80/20 split on the combined dataset from 2014 to 2023.

Performance Results (2014–2023)

RMSE: 1,461.68; MAE: 256.33; R^2: 0.9883

Interpretation:

The extended experiment of XGBoost demonstrates that the model is still very accurate and improves a little bit when 2023 data is added, and R2 goes from 0.9812 to about 0.9883, RMSE and MAE are still at a similar scale. This indicates that XGBoost generalises well to the most recent year without any further tuning, and continues to capture the underlying demand structure quite well through a variety of features such as lagged claims and supplier variables, indicating that it be considered both the most accurate and the most stable model when new data becomes available.

6.7 Experiment 7 – N-BEATS Extended Dataset Evaluation (2023–2024)

In parallel with the XGBoost robustness check an extended experiment for the N-BEATS model was conducted for the same dataset. The objective was to explore what a deep learning model that performed close to the effectiveness of XGBoost in the main analysis does when given an additional year of claims data. As in the case of XGBoost, the architecture and training procedure of N-BEATS followed the same procedure as the previous analysis, only the dataset was expanded and tested with an 80/20 split on the entire 2014-2023 sequence set.

Performance Results (2014–2023)

RMSE: 2,040.03; MAE: 496.48; R^2: 0.9777

Interpretation:

On the extended 2014-2023 dataset, N-BEATS still provides good predictive strength with R2 being around 0.978, slightly decreasing from 0.9801, with slight increase in RMSE and an indication of loss of accuracy with a bit of precision at higher values of utilisation. This indicates that although N-BEATS is a competitive deep learning model with the ability to capture complex temporal patterns, it is a bit more vulnerable to changes and variations that the latest year of DME claims may have introduced, and the longer experiment indeed favours the argument that XGBoost can provide a slightly more robust and stable performance when new years of data are added without reconfiguring the model.

```
Extended XGBoost Performance (2014-2023 dataset)
RMSE: 1,461.68
MAE : 256.33
R²  : 0.9883
```

```
Extended N-BEATS Evaluation (2014-2023 dataset):
R² Score: 0.9777
MAE     : 496.48
RMSE    : 2,040.03
```

6.8 Discussion

The results of this study show that different modelling families provided different levels of the predictive strength of forecasting DME utilisation, XGBoost and N-BEATS modelling performed in particular well and demonstrated significantly better performance as compared to SARIMAX.

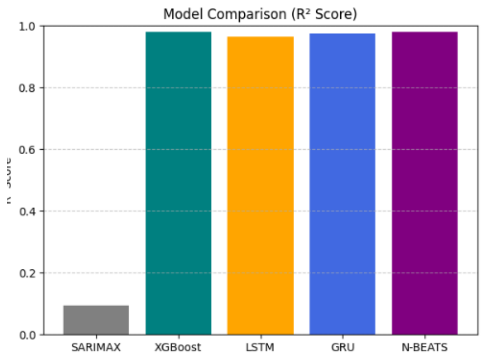

Figure 10: Model comparison by proportion of variance

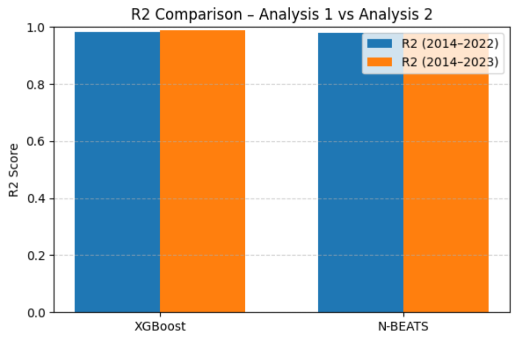

Figure 11: Analysis 1 and Analysis 2 results comparison

In the primary model comparison measured by R2, MAE and RMSE, XGBoost showed the highest performance, with the lowest accuracy and error values, closely followed by the deep learning models GRU and N-BEATS, whereas the SARIMAX was lagging behind the performance, which can be attributed to the limitations when capturing the nonlinear and heterogeneous utilisation patterns. The additional robustness experiments with the recently available data (2023) further reinforced the ranking where XGBoost variance is around 0.98 but with some small changes in the error metrics, whereas the N-BEATS value was slightly decreased around 0.97 and a small rise in RMSE. XGBoost benefited from engineered lag features and covariates, which are likely to be the reason for its accuracy and stability, and N-BEATS still proved that it can capture long-term utilisation patterns without explicit decomposition. The integration of SHAP for both models provided an important interpretability layer, which helped to identify which features were most influential, and also confirmed that meaningful behavioural signals could be obtained from the dataset.

However, there are some aspects of the modelling framework which offer opportunities for refinement. The usage of an 80/20 sequence-based split for neural models, signifies that the evaluation was not entirely in line with the principle of chronological forecasting. Additionally, hyperparameters have been chosen according to standard defaults instead of being fine-tuned with a special validation procedure, which could have restricted the full capability of the deep learning architectures. These limitations do not affect the validity of the results, but do highlight the areas for further work that would increase the stability and generalisability of the findings, such as using richer covariates, examining time-based cross validation, or using higher frequency claims data. When considered in the context of previous studies, the results of the current study are consistent with the general agreement that machine learning and modern deep learning models are more capable than traditional statistical approaches in dealing with heterogeneous and non-linear healthcare utilisation data. Overall, the findings represent an important addition to the literature that provides useful information with clear avenues for future methodological improvement.

7 Conclusion and Future Work

This research aimed to determine the most suitable forecasting method for predicting the utilisation of Durable Medical Equipment (DME) by state-level CMS data and develop a modelling framework that is able to manage the heterogeneity across states and HCPCS codes. The objective was to prepare a unified dataset, to implement a variety of forecasting models, to test the accuracy of such models, and to add an interpretability component to understand the decisions made by models. These aims were successfully accomplished. The comparative analysis showed that the modern machine learning and deep learning models, such as XGBoost and N-BEATS, are more suitable for capturing the utilisation dynamics than the classical statistical methods. The work also showed the feasibility of incorporating explainability features into the forecasting pipeline in order for the high-performing model to be interpreted. In doing so, the study effectively answered the research question and gave a clear, evidence-based understanding as to which modelling strategies are most promising for DME utilisation forecasting.

Although the framework proved reliable and sufficiently robust for the purposes of this study, the work also pointed to several meaningful directions for future development. A natural next step would be to expand the scope by including demographic, socioeconomic or policy-related variables which may identify drivers of utilisation in addition to historical patterns alone. Moreover, the study can be enhanced in the future by using a richer temporal granularity, e.g. monthly or quarterly claims, to be able to capture local variations or temporary demand fluctuations. Architectures such as transformers, hierarchical forecasting models or covariate neural networks, indeed are worth considering to improve the performance of groups with shorter or more irregular time series. In addition to the academic extensions, the framework can be applied commercially as a decision-support system for insurers, healthcare providers, and policy agencies that aim to predict the demand of DME. This applied system would utilise improved models in the future to facilitate procurement arrangements, reimbursement review, and resource utilisation. Altogether, these possibilities offer a positive way of extending this work to more sophisticated and functional forecasting tools.

References

Tuominen, J., Pulkkinen, E., Peltonen, J., Kanniainen, J., Oksala, N., Palomäki, A. and Roine, A., 2024. Forecasting emergency department occupancy with advanced machine learning models and multivariable input. *International Journal of Forecasting*, *40*(4), pp.1410-1420. Available at: https://doi.org/10.1016/j.ijforecast.2023.12.002

Menebo, M., 2025. Leveraging Big Data, Weather Insights, and the XGBoost ML Model to Better Forecast Medication Demand and Manage Shortages. Available at: https://doi.org/10.21203/rs.3.rs-5759776/v1

Babanejaddehaki, G., An, A. and Papagelis, M., 2025. Disease outbreak detection and forecasting: A review of methods and data sources. *ACM Transactions on Computing for Healthcare*, *6*(2), pp.1-40. Available at: https://doi.org/10.1145/3708549

Koç, E. and Türkoğlu, M., 2022. Forecasting of medical equipment demand and outbreak spreading based on deep long short-term memory network: the COVID-19 pandemic in Turkey. *Signal, image and video processing*, *16*(3), pp.613-621.Available at: https://doi.org/10.1007/s11760-020-01847-5

Jin, X., Wei, F., Kandala, S.S., Umesh, T., Steele, K., Galgiani, J.N. and Laubichler, M.D., 2025. Time series forecasting of Valley fever infection in Maricopa County, AZ using LSTM. *The Lancet Regional Health–Americas*, *43*. Available at: https://doi.org/10.1016/j.lana.2025.101010

Pampa, I.L.P., Rosales, L.B.R. and Atúncar, G.S., 2025. Machine Learning and Deep Learning for demand forecasting in Logistics: A systematic literature review. Available at: https://doi.org/10.21203/rs.3.rs-7336582/v1

Kong, X., Chen, Z., Liu, W., Ning, K., Zhang, L., Muhammad Marier, S., Liu, Y., Chen, Y. and Xia, F., 2025. Deep learning for time series forecasting: a survey. *International Journal of Machine Learning and Cybernetics*, pp.1-34. Available at: https://doi.org/10.1007/s13042-025-02560-w

Walter, A., Ahsan, K. and Rahman, S., 2025. Application of artificial intelligence in demand planning for supply chains: a systematic literature review. *The International Journal of Logistics Management*, *36*(3), pp.672-719. Available at: https://doi.org/10.1108/IJLM-02-2024-0120

Tamatta, R.K., 2021. Time series forecasting of hospital inpatients and day-case waiting list using ARIMA, TBATS and neural network models. MSc Thesis, National College of Ireland. Available at: https://norma.ncirl.ie/3432/1/rajkumartamatta.pdf

Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, pp.159-175.Available at: https://doi.org/10.1016/S0925-2312(01)00702-0

Zhou, L., Zhao, P., Wu, D., Cheng, C. and Huang, H., 2018. Time series model for forecasting the number of new admission inpatients. *BMC medical informatics and decision making*, *18*(1), p.39. Available: https://doi.org/10.1186/s12911-018-0616-8

Levine, R., Pickett, J., Sekhri, N. and Yadav, P., 2008. Demand forecasting for essential medical technologies. *American journal of law & medicine*, *34*(2-3), pp.225-255. Available at: https://doi.org/10.1177/009885880803400206

Shafiekhani, S., Namdar, P. and Rafiei, S., 2022. A COVID-19 forecasting system for hospital needs using ANFIS and LSTM models: A graphical user interface unit. *Digital Health*, *8*, p.20552076221085057. Available at: https://doi.org/10.1177/20552076221085057

Pahlevani, M., Taghavi, M. and Vanberkel, P., 2024. A systematic literature review of predicting patient discharges using statistical methods and machine learning. *Health Care Management Science*, *27*(3), pp.458-478. Available at: https://doi.org/10.1007/s10729-024-09682-7

Elgohari, H., Bakr, A.M. and Majeed, M.A., 2019. Forecasting the number of outpatient visits in tertiary hospital using time series based on ARIMA and ES models. *Aust J Basic Appl Sci*, *13*(7), pp.70-77.

Lin, W.C., Goldstein, I.H., Hribar, M.R., Sanders, D.S. and Chiang, M.F., 2020, March. Predicting wait times in pediatric ophthalmology outpatient clinic using machine learning. In *AMIA Annual Symposium Proceedings* (Vol. 2019, p. 1121). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153152/

Xu, S. and Chan, H.K., 2019. Forecasting medical device demand with online search queries: a big data and machine learning approach. *Procedia Manufacturing*, *39*, pp.32-39. Available at: https://doi.org/10.1016/j.promfg.2020.01.225

Cortez, P. and Morais, A.D.J.R., 2007. A data mining approach to predict forest fires using meteorological data. Available at: https://hdl.handle.net/1822/8039

Guo, P., Loh, W.Y., Maung, K., Choo, E.L.W., Dickens, B.L., Tan, K.B., Abishgenadan, J., Ma, P. and Lim, J.T., 2025. High-dimensional point forecast combinations for emergency department demand. *arXiv preprint arXiv:2501.11315*. Available at: https://doi.org/10.48550/arXiv.2501.11315

Yang, D., Tang, Y., Chan, V.K.Y., Fang, Q., Chan, S.S.M., Luo, H., Wong, I.C.K., Ou, H.T., Chan, E.W.Y., Bishai, D.M. and Chen, Y., 2025. Population-Wide Depression Incidence Forecasting Comparing Autoregressive Integrated Moving Average and Vector Autoregressive Integrated Moving Average to Temporal Fusion Transformers: Longitudinal Observational Study. *Journal of medical Internet research*, *27*, p.e67156. Available at: https://doi.org/10.2196/67156

Telesca, V. and Rondinone, M., 2025. An Interpretable Machine Learning Framework for Analyzing the Interaction Between Cardiorespiratory Diseases and Meteo-Pollutant Sensor Data. *Sensors*, *25*(15), p.4864. Available at: https://doi.org/10.3390/s25154864

Hu, J., Zhang, B., Xu, T., Yang, H. and Gao, M., 2025. Structure-aware temporal modeling for chronic disease progression prediction. *arXiv preprint arXiv:2508.14942*. Available at: https://arxiv.org/abs/2508.14942

Rane, J., Amol Chaudhari, R. and Rane, N., 2025. Artificial Intelligence and Machine Learning for Supply Chain Resilience: Risk Assessment and Decision Making in Manufacturing Industry 4.0 and 5.0. *Artificial Intelligence and Machine Learning for Supply Chain Resilience: Risk Assessment and Decision Making in Manufacturing Industry*, *4*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5366932

Vollmer, M.A., Glampson, B., Mellan, T., Mishra, S., Mercuri, L., Costello, C., Klaber, R., Cooke, G., Flaxman, S. and Bhatt, S., 2021. A unified machine learning approach to time series forecasting applied to demand at emergency departments. *BMC Emergency Medicine*, *21*(1), p.9. Available at: https://doi.org/10.1186/s12873-020-00395-y

Fourkiotis, K.P. and Tsadiras, A., 2024. Applying machine learning and statistical forecasting methods for enhancing pharmaceutical sales predictions. *Forecasting*, *6*(1), pp.170-186. Available at: https://doi.org/10.3390/forecast6010010