RESEARCH PAPER

Forecasting Patient Discharge Volume in Northern Region Government Hospitals of Malaysia Using Machine Learning Techniques

Noorazimah binti Aminuddin[1], Gunasekar Thangarasu[2*] and Rosilawati Abdul Rani[3]

[1] Department of Digital Health and Health Informatics, School of Business and Technology, IMU University, Kuala Lumpur, Malaysia, noor_azimah@moh.gov.my

[2] Department of Digital Health and Health Informatics , School of Business and Technology, IMU University, , Kuala Lumpur, Malaysia, gunasekar97gmail.com

[2] Saveetha School of Engineering, Saveetha Institute of Medical and Technical Science, Chennai, India,

[3]Clinical Research Centre, Hospital Taiping, Ministry of Health Malaysia, Perak, Malaysia. dr.rosilawati@moh.gov.my

***Corresponding author:** Gunasekar Thangarasu, *Department of Digital Health and Health Informatics, IMU University, Malaysia*. Email: gunasekar97@gmail.com

Abstract: The Malaysian public healthcare system is under growing strain, which is driven by an aging population and the rising burden of chronic non-communicable diseases (NCDs). These shifts have placed added pressure on hospitals, particularly in handling inpatient discharges. To respond to this challenge, the study set out to build a machine learning-based forecasting model that can predict discharge volumes. Hospital Taiping, one of the key government hospitals in northern Malaysia, was selected as the case study. Monthly inpatient discharge data from 2019 to 2023 formed the foundation for building predictive models. To strengthen the dataset, additional temporal features were engineered, including lagged discharge counts, rolling averages and seasonal indicators. Three machine learning models, namely Random Forest (RF), XGBoost and Support Vector Regression (SVR), were then trained and fine-tuned through Bayesian Optimization. Model performance was evaluated using coefficient of determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Scaled Error (MASE) and Mean Absolute Percentage Error (MAPE). The best-performing model was prospectively validated with actual discharge data from 2024. The Random Forest model demonstrated the lowest RMSE of 177.85, a MAPE of 3.23% and the highest R^2 of 0.43. Prospective validation using 2024 data confirmed its robustness in forecasting inpatient discharges. The study developed and validated a machine learning model to forecast inpatient discharge volumes in government hospitals across the Northern Region of Malaysia. It also contributes to both methodology and practical application by providing a forecasting framework that supports more informed hospital planning and resource management.

Keywords: Machine learning, forecasting, inpatient discharge volume, healthcare resource planning

How to cite this article: Aminuddin N, Thangarasu G, Abdul Rani R. Forecasting Patient Discharge Volume in Northern Region Government Hospitals of Malaysia Using Machine Learning Techniques. Int J Drug Deliv Technol. 2026;16(53s): 449-458. DOI: 10.25258/ijddt.16.53s.48

1. Introduction

The scarcity of healthcare resources is a worldwide issue. In numerous countries, there is an increasing apprehension that the current methods of healthcare service provision are, if not already, on the brink of becoming unsustainable for the population. The increasing prevalence of chronic NCDs, coupled with an ageing population, heightened medical demand for both outpatient and inpatient services together with escalating medical costs, has exacerbated the shortage of medical resources [1].

Nonetheless, Malaysia faces the same challenges as other nations. Although the government endeavours to enhance investment in healthcare resources but these resources remain insufficient to meet the growing demand of the population. Malaysia experienced an increase in total health expenditure between 2011 and 2021, reflecting both as a percentage of Gross Domestic Product (GDP) and in per capita expenditure. This fact corresponds to the growth in population size [2]. There is also an imbalance between financing and demand for the health system which indicates that Malaysia is underinvesting in healthcare sector. Malaysia as an upper-middle-income country (UMIC),

allocates only 4.1% of GDP to health expenditures in both the public and private sectors. This figure is lower than the average of 7.4% of GDP for UMIC and 8.8% of GDP for high-income countries (HIC) [3].

In response to this issue, the Ministry of Health (MOH) Malaysia tabled the Health White Paper in Parliament in June 2023, which sets out a comprehensive proposal for systemic and structural reforms of the Malaysian health system aimed at addressing the health challenges to enhance the sustainability, equitability and resilience within the healthcare system. The paper proposes reforms founded on four pillars, one of which pertains to the provision of sustainable and equitable healthcare funding [4].

The MOH hospitals are classified into four major categories: state hospitals, hospitals with specialists, hospitals without specialists, and special medical institutes. Hospitals with specialists are split into major and minor specialized hospitals. Together with this, the Cluster Hospital (CH) concept was also implemented in MOH hospitals in 2014 to balance utilization between public hospitals [5]. The concept is grouping several hospitals within the exact geographical location into a cluster of lead hospitals (LHs) and smaller hospitals as non-lead hospitals (NLHs). The implementation of the CH concept will improve the utilization of district non-specialist hospitals and decongest the over-utilized specialist hospitals [6].

In recent years, data-driven methods have been implemented as transformative tools in other sectors including healthcare. These methods are urgently needed for precise forecasts of healthcare demand and resources, as predicting future demands and medical resources availability could guide decision-making at the higher level of healthcare management, especially in Malaysia. Data-driven methods involve the integration of artificial intelligence (AI), machine learning (ML) and data analytics to leverage the capabilities of big data in healthcare systems in order to gain insights into forecast health patterns or trends and patient behaviours [7].

The application of data-driven methods especially involving machine learning algorithms to forecast healthcare metrics, for example hospital readmission, postoperative length of stay (LOS), inpatient bed demand and inpatient hospital discharge volumes had been demonstrated the ability to make predictions for unseen input and learn temporal knowledge (8). The increasing emphasis on data-driven methods as transformative tools in healthcare sectors is urgently needed to predict future demands and availability of healthcare resources.

2. Literature Review

In the past few decades, various forecasting techniques have been developed to forecast hospital metrics like patient admissions, readmission rates, length of stay and inpatient discharge volumes. These techniques range from traditional statistical methods to advanced data-driven methods including machine learning algorithms. This chapter reviews relevant literature on various forecasting techniques implemented in hospital metrics with a specific focus on forecasting inpatient hospital discharge volumes. The first section in this chapter explores machine learning techniques used in healthcare resources while the second section examines other forecasting techniques aside from machine learning which have been widely used in hospital management worldwide. The final section focuses on time series forecasting methods that are particularly applicable in forecasting inpatient discharge volume. This comprehensive review not only highlights gaps in current studies but also establishes the foundation for the methodologies employed in this study.

The study [8] developed accurate predictive models using machine learning techniques to forecast bed occupancy in mental health hospitals. This is a retrospective study that used historical data from 2008 to 2024 collected from the Central Institute of Psychiatry, Ranchi in India. The study analyzed 866 weeks of bed occupancy records to identify patterns and trends. Six machine learning models were applied in this study which included eXtreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), SVR, Gradient Boosting (GB), RF and Decision Tree (DT). This study was evaluated using statistical test and performance metrics including Mean Absolute Error (MAE), RMSE, Diebold–Mariano (DM) and MAPE. The results showed that RF and Decision Tree performed the best compared to others with Random Forest achieving the lowest MAPE of 3.57% which made it the most reliable model for forecasting bed occupancy.

The authors [9] done a study that used Random Forest (RF) algorithm to identify factors associated with Prolonged Hospital Length of Stay (PLOS) in elderly patients with hip fractures who are undergoing surgery. Efficient prediction of PLOS can optimize healthcare resource allocation and improve patient outcomes. This is a retrospective cohort study analyzing 360 patients aged 65 years and above who underwent surgical intervention at West China Hospital between October 2021 and November 2023 which included 360 elderly patients, of whom 103 (28.61%) experienced extended hospitalization. The RF model that has been developed using the training dataset was able to identify 10 key predictive variables and demonstrated an outstanding performance on the training dataset as it achieved balanced accuracy, an Area Under the Curve (AUC), an F1 score and Kappa value of 1.000. When evaluated on the test dataset, it achieved an F1-score of 0.606, an AUC of 0.846, Kappa value of 0.4325 and balanced accuracy of 0.7294, demonstrating its effectiveness in predicting prolonged hospital stays while maintaining generalizability.

The research [10] proposed ML development to predict weekly inpatient bed demand which could enable the improvement of hospital resource planning and utilization over the overcrowding issues especially in emergency departments (ED) and post-anesthesia care unit (PACU). The study analyzed five years of data from adult inpatients attended at Geisinger Medical Centre (GMC) that focused on observation status, surgical overnight recovery (SORU) or type of inpatient encounters. The proposed ML strategy integrates K-means clustering with Support Vector Machine Regression (K-SVR) techniques. Results indicate that the K-SVR model achieved a MAPE between 0.49% and 4.10% during the test period, which demonstrated reduced variability and stable forecasting outcomes. These findings

highlight the effectiveness of ML techniques, particularly K-SVR in forecasting inpatient bed demand.

The study [11] compared the performance of ML techniques between Support Vector Machine (SVM) and RF in forecasting hospital readmission. The study developed forecasting models for hospital readmissions by using four ML models which include support vector machines with linear and RBF kernels, weighted random forests and balanced random forests. The analysis was based on 11,172 hospitalization records from the General Hospital of Komotini "Sismanogleio." The results demonstrated that the balanced RF model performed the best as it achieved an AUC value of 0.78 and a sensitivity of 0.70. .

The framework [12] developed a forecasting model to identify patients at high risk of unplanned readmission (UPRA) within 14 days after discharge due to pneumonia. Currently, no effective indicators are available during hospitalization to help clinicians identify patients at high risk of UPRA. The study used 21,892 data on pneumonia cases collected from three hospitals in Taiwan between 2016 and 2018 with 1208 cases involving UPRA. The study evaluated two forecasting models: Convolutional Neural Network (CNN) and Artificial Neural Network (ANN). The ANN model achieved a higher AUC (0.75) than the CNN model which demonstrated superior accuracy and stability.

The study [13] explored the development of ML as a predictive tool for predicting the risk of intensive care unit (ICU) transfer within 24 hours for hospitalized COVID-19 patients by using electronic medical record (EMR) data from the selected hospital. The study utilized a retrospective cohort study of 1,987 COVID-19 patients admitted to non-ICU units between February 26 and April 18, 2020, at an extensive acute care health system. The study chose the RF algorithm to develop a predictive model by using time-series as input variables to train the model. The model demonstrated strong performance with a sensitivity of 72.8%, specificity of 76.3%, accuracy of 76.2% and an AUC-ROC of 79.9%.

The study [14] conducted a study to address the underexplored issue in the healthcare system of predicting hospital readmissions using machine learning techniques particularly for diabetic patients by comparing five ML techniques: Multi-Layer Perceptron (MLP), Logistic Regression (LR), DT, Naïve Bayesian (NB) classifier and SVM. This comparative study used 3090 diabetic patients from many hospitals in the United States. The study found that SVM delivered the best performance, while LR and NB Classifier were the worst. .

The study [15] explored a study to develop a machine learning model in forecasting ICU readmissions within 30 days of discharge from hospital. The study used clinical data from the MIMIC-III database that employed within the Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM). Results of the study showed that the LSTM model demonstrated better performance than traditional methods with an AUC of 0.791 and a sensitivity of 0.742.

The research [16] explored on developing a predictive tool using machine learning techniques for predicting the LOS of cardiac patients who had been admitted in hospital. This study retrospectively extracted electronic medical records (EMR) from King Abdulaziz Cardiac Center (KACC) in Riyadh, Saudi Arabia that covered from 2008 to 2016. Patients were categorized into three groups based on LOS and relevant attributes were selected using the information gain algorithm. Four ML techniques were evaluated: RF, ANN, SVM, and Bayesian Network (BN). Results of the study showed that RF model outperforming other models by achieving a sensitivity and accuracy of 80% and an AUROC of 0.94.

The article [17] focused on predicting hospital remissions using a hybrid ML model that combines a boosted C5.0 decision tree and a SVM. The model integrated a boosted C5.0 decision tree as the base classifier with a SVM as a secondary classifier. This hybrid approach was evaluated using 20,321 anonymized records of inpatient admissions during fiscal years between 2006 and 2014 at Pittsburgh Veterans Health Administration hospitals where 4,840 Congestive Heart Failure (CHF) patients were treated. The result of the study showed that SVM predictions achieved higher sensitivity across a broader range of ROC curve cut-off values which was guided by a predefined confidence threshold for the C5.0 classifier with the hybrid model achieving an accuracy range of 81% to 85%.

The study 18] explored in developing a statistical-based predictive model as the study aimed to predict patient inflow and hospital admissions at Liaquat University of Medical and Health Sciences (LUMHS) located in Jamshoro, Pakistan. The study used secondary data that been collected from LUMHS inpatient records and employed the data collected with the Autoregressive Integrated Moving Average (ARIMA) (1,0,1) model in MATLAB software to analyze the data. The dataset consisted of 859,000 patient inflow records and 109,011 patient admissions over a period of one year. In addition, the best ARIMA model was selected based on Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and p-values. The results of the study showed that the ARIMA (1,0,1) model achieved high accuracy with an R^2 of approximately 95% and low forecasting errors.

The authors [19] explored a study related to the relationship of patterns of outpatient hospital visits with meteorological environmental factors. This study utilized the ARIMA model to examine these relationships and predict future outpatient visits. It collects data from January 2015 to July 2021. The collected data were related to monthly outpatient hospital visits and employed within the ARIMA model. This study later developed an ARIMAX model which was then evaluated using forecasting error metrics like coefficient of determination (R2), stationary R2, normalized Bayesian information criterion (BIC), and MAPE. The results of the study showed that The ARIMA model was the most effective model which successfully predicted outpatient visits for 2019 with a relative error of just 2.77%.

The research [20] did a systematic review to compare effectiveness of ML techniques against conventional statistical modeling (CSM) in predicting myocardial infarction (MI) readmission and risk of mortality. This study followed PRISMA guidelines which is a comprehensive literature review that was conducted across multiple databases. These guidelines covered studies that been

published from January 2000 to March 2020 with a total of 24 studies involving 374,365 patients who met the inclusion criteria and the data were extracted from Medline, Cochrane Central, Embase, and other sources. The ML techniques studied included artificial neural networks (n=12), random forests (n=11), decision trees (n=8), SVM (n=8) and Bayesian methods (n=7) while CSM primarily comprised logistic regression (n=19), risk scores (n=12) and Cox regression (n=2). The results showed that ML models generally exhibited slightly higher C-indexes for predicting mortality and readmission in Myocardial Infarction patients compared to CSM.

The study [21] did a study with the aim of identifying factors affecting LOS in the ICU and comparing statistical models between Gamma and Log-normal for predicting these associated factors. This study was conducted as a cross-sectional design of 565 patients admitted in ICU from Imam Khomeini Hospital in Ahvaz. The data from the year 2015 were retrospectively analyzed using SPSS 21 and STATA 7 software. The results of the study showed that the average LOS was 8.16 days, with a mean patient age of 58.61 years while the statistical analysis stated that the Gamma regression model provided a better fit than the Log-normal model, with the type of disease diagnosis emerging as the most significant predictor of LOS.

3. Research Methodology

The methodological approaches employed in the study to develop a machine learning model for forecasting inpatient discharge volumes in government hospitals, with a focus on Hospital Taiping in the northern region of Malaysia. The methodology begins with a clear articulation of the study design and population, followed by a detailed description of the data collection, preprocessing steps, and feature engineering methods used in the study. The chapter also summarizes the model development and validation procedures, with each step carefully structured to ensure methodological rigor, reliability, and applicability of the forecasting model within the Malaysian healthcare setting.

3.1. Research design

This study applied a cross-sectional research design with a quantitative method of analysis. The cross-sectional approach was appropriate as it enabled the examination of historical discharge data across a defined time frame from 2019 to 2023. Quantitative methods were used to develop a forecasting ML model based on numerical inpatient hospital discharge data. The flowchart of the proposed machine learning pipeline of this study consists of several sequential stages.

Figure 1

Locations of the five major specialist hospitals in the northern region of Peninsular Malaysia

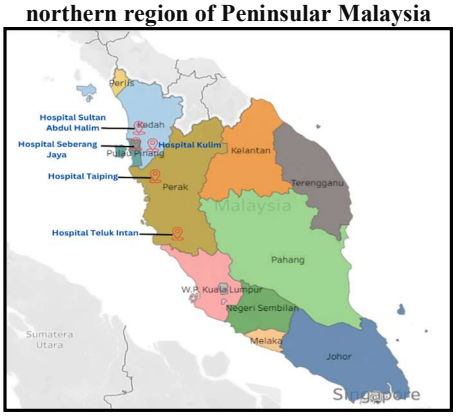
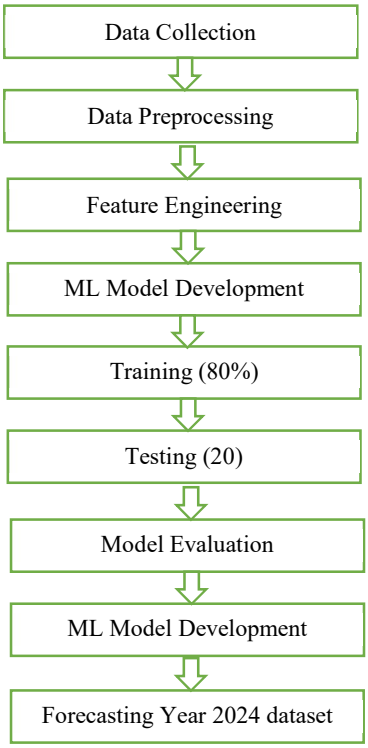


Figure 2
Research Design Steps



3.2. Study Population

This study was conducted using data from Hospital Taiping, a major specialist government hospital located in northern Perak, Malaysia. As one of the pioneering public hospitals in the region, it serves as the lead facility within a cluster of five hospitals, with a total bed allocation of 608 beds. This study utilized two datasets collected from the same source. The primary dataset was used to develop and train ML models, consisting of total inpatient discharge data spanning five years, from 1st January 2019 to 31st December 2023. The data set utilized in this study comprises monthly

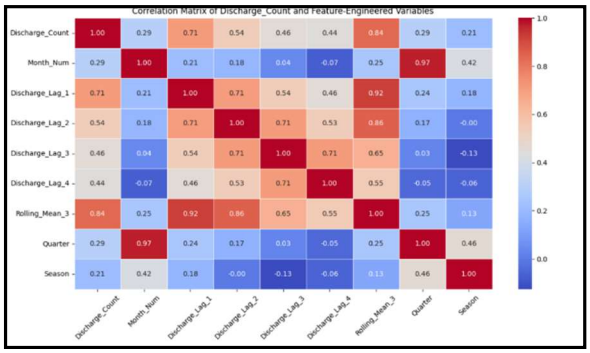
inpatient discharge records collected from January 2019 to December 2023, encompassing 60 observations. Each observation represents aggregated inpatient discharge counts for a specific month. The dataset comprises 60 rows and four columns, representing monthly inpatient discharge data across five years. It includes four primary variables as demonstrated in Table 1.

Table 1
Descriptive Analysis of Inpatient Discharge Volume by Year (2019-2023)

Year	Total Inpatient Discharge	Mean	Standard Deviation (SD)	Median	IQR
2019	48,315	4026.25	182.55	4051.0	122.50
2020	44,999	3749.92	333.32	3840.0	211.75
2021	48,542	4045.17	355.65	4058.5	591.75
2022	53,566	4463.83	224.64	4471.0	320.25
2023	52,288	4357.33	177.76	4328.5	245.25
Total	247,710	4128.50	363.34	4150.5	477.75

The dataset was confirmed to be normally distributed based on the Kolmogorov–Smirnov test, where the p-value > 0.05. Consequently, Pearson correlation analysis was employed to examine the strength and direction of the linear associations between the dependent variable, Discharge Count, and the set of feature-engineered independent variables. This led to the creation of a heatmap, which displays the correlation coefficients, as shown in Figure 4. The corresponding numerical values of the correlation coefficients are presented in Table-1 to support and complement the visual interpretation.

Figure 2
Heatmap of Pearson Correlation Coefficients Between Discharge Count and Feature-Engineered Variables



It was revealed that Rolling_Mean_3 exhibited the strongest positive correlation with the dependent variable, recording a coefficient of 0.84. It was followed by Discharge_Lag_1 and Discharge_Lag_2, with correlation values of 0.71 and 0.54, respectively. Moderate associations were also observed for Discharge_Lag_3 and Discharge_Lag_4, which showed coefficients of 0.46 and 0.44. These findings confirm the predictive relevance of lag-based temporal features in modelling inpatient discharge volume. In contrast, the calendar-based features, including Month Num, Quarter, and Season, displayed relatively weak correlations with the dependent variable, with correlation coefficients of 0.29, 0.29, and 0.21, respectively.

High correlations were also observed among several independent variables, notably between Rolling_Mean_3 and Discharge_Lag_1 with a correlation coefficient of 0.92 and between Month_Num and Quarter with a coefficient of 0.97. These findings suggest some degree of multicollinearity, particularly among time-based features. In addition to correlation analysis, multicollinearity diagnostics were performed using the Variance Inflation Factor (VIF) to assess the degree of redundancy among the predictors yet complement the Pearson correlation analysis. The results, as illustrated in Table 5, indicated that several features exhibited high collinearity, notably Month_Num and Quarter, which recorded VIF values above 21 due to their structural dependency. Likewise, Rolling_Mean_3 and Discharge_Lag_1 showed elevated VIF values of 16.36 and 8.70, respectively, consistent with their strong pairwise correlation.

However, no feature was removed, as the tree-based ML algorithms applied in this study. Namely, RF and XGBoost are not inherently sensitive to multicollinearity. This is supported by [21], who found that Random Forest maintained consistent and high predictive performance even when applied to datasets with substantial multicollinearity, highlighting its robustness in such conditions. Although XGBoost showed slightly lower accuracy in some instances, it still demonstrated reliable performance, suggesting that it can reasonably handle multicollinearity. While Support Vector Regression was also evaluated in this study, the selection of features across all models was guided by a performance-based evaluation approach rather than automated feature selection techniques.

3.3. Feature Engineering

This study was conducted using data from Hospital Taiping, a major specialist government hospital located in northern Perak, Malaysia. As one of the pioneering public hospitals in the region, it serves as the lead facility within a cluster of five hospitals, with a total bed allocation of 608 beds. This study utilized two datasets collected from the same source. The primary dataset was used to develop and train ML models, consisting of total inpatient discharge data spanning five years, from 1st January 2019 to 31st December 2023.

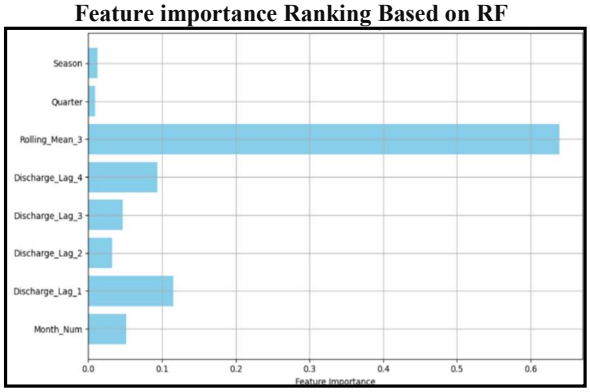
Table 2
Pearson Correlation Coefficients Between Discharge Count and Feature-Engineered Variables

Feature	Correlated Coefficient	P-Value	VIF
Rolling_Mean_3	0.840	4.575×10^{-17}	3571
Discharge_Lag_1	0.711	1.935×10^{-10}	2987
Discharge_Lag_2	0.537	9.745×10^{-6}	3478
Discharge_Lag_3	0.461	2.106×10^{-4}	4130
Discharge_Lag_4	0.443	3.943×10^{-4}	4043
Month_Num	0.294	2.245×10^{-2}	2987

The study to capture temporal relationships within the dataset, lag features were engineered using historical discharge counts. Specifically, four lag variables, including Discharge_Lag_1, Discharge_Lag_2, Discharge_Lag_3,

and Discharge_Lag_4, were created to represent discharge volumes from one to four months prior. These features enable the model to detect patterns of autocorrelation, a common characteristic in time series data, where recent observations influence future outcomes. By incorporating these lagged inputs, the model is better equipped to learn short-term trends and fluctuations in inpatient discharge. In addition to the lag features, a rolling mean feature called Rolling_Mean_3 was constructed by calculating the average discharge count over three months. This feature helps reduce short-term variability and highlight local trends, thereby improving the model's ability to learn stable and recurring patterns in the data.

Figure 2
Feature importance Ranking Based on RF



3.4. Model Development

This study employed SVR, RF and XGBoost algorithms to develop a forecasting model of inpatient discharge volume in government hospitals in Malaysia. The ML algorithms were selected based on their performance forecasting healthcare resources through a literature review. Detailed descriptions and evaluation of each algorithm will be provided in the following subsections:

3.4.1 Support Vector Regression (SVR)

In this study, SVM is applied in the form of SVR to forecast inpatient discharge volumes. SVR is a supervised ML algorithm that aims to find a function that best predicts a continuous outcome while maintaining an epsilon, which is the margin of tolerance around the predicted values (44). Instead of minimizing the overall prediction error, SVR focuses on fitting the data within this margin, where only the data points that fall outside the boundary, known as support vectors, influence the final model. This makes SVR more robust to outliers compared to traditional linear regression. To model non-linear relationships, SVR employs kernel functions such as the Radial Basis Function (RBF), which project the input data into a higher-dimensional space where linear relationships can be established. SVR handles non-linear regression problems by transforming the input variables into a high-dimensional space using a non-linear function φ(γ), known as the kernel trick (32)**.** Then the SVR tries to minimize the ϵ−loss function (45):

$$loss = C\sum_{i=1}^{n}(\xi i + \xi_i^*) + \frac{1}{2}\,||\,w\,||^2 \quad (3.1)$$

Subject to the constraints (46):

$$\psi i - \Big((\omega * \varphi(\gamma i))\Big) \le \epsilon + \xi i \quad (3.2)$$

$$\Big((\omega * \phi(\gamma i))\Big) - \psi i \le +\xi * i \quad (3.3)$$

From the equations, ξi and ξ_i^* are positive slack variables, i = 1,2,3..., N, C is the regularization parameter that controls the trade-off between flatness and tolerance for deviations, and www is the weight vector of the hyperplane. This primal problem is converted into a dual problem using Lagrange multipliers, yielding the SVR prediction function (47):

$$f(\gamma, a_i, a_i^*) = \sum_{i=1}^{n}(a_i - a_i^*)\,K(\gamma, \gamma_i) + b \quad (3.4)$$

Where K(γ, γ_i) = φ(γ) φ(γ_i) is the kernel function. The Lagrange multipliers are found by solving the following dual optimization problem (32):

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i - a_i^*)(a_j - a_j^*)\,K(\gamma_i, \gamma_j) \quad (3.5)$$

$$\sum_{i=0}^{n}\phi_i\,(a_i - a_i^*) + \epsilon\sum_{i=0}^{n}(a_i + a_i^*) \quad (3.6)$$

With the constraints

$$\sum_{i=0}^{n}(a_i + a_i^*) = 0, 0 \le a_i, a_i^* \le C, i = 1,2,\dots,N. \quad (3.7)$$

Support Vector Machine (SVM), particularly in its regression form, SVR, offers several advantages. It effectively captures nonlinear relationships with kernel functions, which makes it suitable for complex healthcare data with irregular patterns (49). SVM is also known for its strong generalization ability, especially when working with small to medium-sized datasets (50). Due to the use of the epsilon-insensitive loss function, it is relatively robust to outliers (51). In addition, SVM has a lower risk of overfitting, especially in high-dimensional spaces, by maximizing the margin between classes (52).

3.4.2 Random Forest

RF is an ensemble learning algorithm that can handle both classification and regression tasks by building multiple decision trees during training periods and combining outputs from multiple trees to generate an average prediction (56). In this study, Random Forest is applied in the form of Random Forest Regression (RFR) as a type of Decision Tree designed to predict continuous numerical outcomes. It trains multiple independent decision trees, referred to as a forest, and averages their predictions for the final output. It employs bagging and the random subspace method, where each tree is built on a different bootstrap sample and feature subset, ensuring trees are grown in parallel without interdependence (57)

$$\mathrm{RF} = \frac{1}{K}\sum_{k=1}^{K} h_k\,(x) \tag{3.8}$$

K represents the total number of independent regression trees generated for bootstrap samples for the input vector x , and h_k represents the meaning of predictions calculated from *K* regression trees (58). Out-of-bag (OOB) prediction for regression was calculated by averaging the predictions of all trees where a given instance was excluded from the bootstrap training set. Formally, for the *i*-th sample, the OOB prediction is defined as (59):

$$\hat{f}_{oob}(x_i) = \frac{1}{J_i}\sum_{j\in J_i} \hat{h}_j\,(x_i) \tag{3.9}$$

Where $\hat{h}_j(x_i)$ represents the predicted response for x_i generated by the *j*th tree. The model is also trained to minimize the MSE for out-of-bag data (OOB), which is a direct measure of prediction error for regression problems. The mathematical formulation is (60):

$$MSE_{OOB} = \frac{1}{n}\sum_{i=i}^{n}\left(y_i - \overline{y}_{i_{OOB}}\right)^2 \tag{3.10}$$

y_i denotes the prediction for the *i*th observation, while $\overline{y}_{i_{OOB}}$ is the average of the *i*th predictions from all trees. Meanwhile, the R-squared value is computed using the out-of-bag (OOB) dataset as follows, where Var_y denotes the total variance of the target variable (61):

$$R^2{}_{OOB} = 1 - \frac{MSE_{OOB}}{Var_y} \tag{3.11}$$

Random Forest, particularly in the regression form, offers several advantages. It is highly efficient in modelling complex, non-linear relationships by combining multiple decision trees to produce more accurate and stable predictions [20]. Unlike the individual decision trees, which are prone to overfitting, Random Forest Regression (RFR) mitigates this risk by employing bagging and random feature selection, thereby enhancing generalization performance [21]. It is a versatile algorithm capable of addressing both classification and regression tasks, making it suitable for a wide range of applications. Furthermore, its ensemble structure allows it to manage large datasets and high-dimensional input spaces effectively.

3.4.3 Extreme Gradient Boosting (XGBoost)

XGBoost is a supervised ML algorithm, which is an extension of the traditional gradient boosted decision trees algorithm [18]. It was introduced as a unique implementation of the gradient boosting framework, designed explicitly for regression and classification tasks. The core principle behind XGBoost is "boosting," which involves combining predictions from multiple weak learners using an additive training approach to build a strong predictive model [19]. It is designed to enhance both computational efficiency and predictive accuracy, making it widely applicable across various industries. Its core mechanism is outlined as follows:

$$\hat{y}_i = \sum_{l}^{k} f_k\,(x_i), f_k \in F \tag{3.12}$$

F denotes the set of all regression trees, and *f* represents an individual tree within this set. The objective function in XGBoost comprises two main components: a loss function that quantifies the error between predicted and actual values and a regularization term that penalizes model complexity to minimize overfitting. The objective function is expressed as:

$$Obj^{(p)} = \sum_{k=1}^{n} l\,(\overline{y}_i, y_i) + \sum_{k=1}^{p} \sigma\,(f_i) \tag{3.13}$$

where l denotes the loss function, n refers to the number of observations used in the model, and σ refers to the regularization term as shown in the equation below:

$$\sigma(f) = YT + 0.5\lambda\omega^2 \tag{3.14}$$

where ω represents vector scores assigned to the leaf nodes, Y indicates the minimum loss required to justify further splitting of a leaf node, and λ defines the regularization parameters. XGBoost also approximates the loss function using a second-order Taylor expansion, incorporating both the first and second derivatives of the loss function. This technique enhances optimization efficiency by accelerating convergence and increasing the reliability of decision tree splitting. The mathematical formulation is:

$$L^{(t)} = \sum_{i=1}^{n}[\,g_i f_t(x_i) + \frac{1}{2}h_i f_t(x_i)] + \Omega(f_i) \tag{3.15}$$

where $g_i = \frac{\partial L(y_i,\hat{y}_i)}{\partial \hat{y}_i}$ is the first-order gradient of the loss function, representing the direction of error correction, while $h_i = \frac{\partial^2 L(y_i,\hat{y}_i)}{\partial \hat{y}_i{}^2}$ is the second-order gradient, which helps determine the step size for updating the model, and $f_t(x_i)$ denotes the new tree added in the boosting process. XGBoost could also identify the optimal split in a decision tree by calculating a gain function, which quantifies the loss reduction when the dataset is divided into left and right branches. This function is expressed as:

$$L_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{3.16}$$

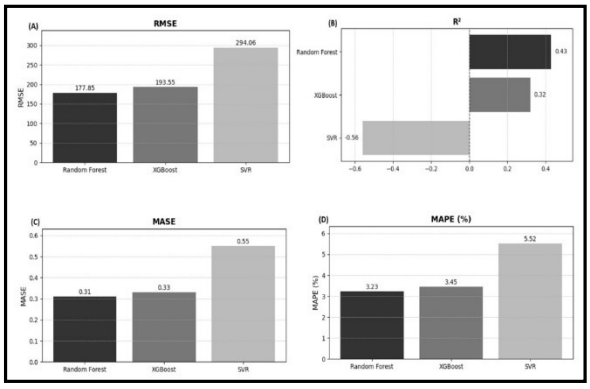
where *L* and *R* denote the left and right child nodes, g_i and h_i are the first and second derivatives of the loss function, λ is a regularization parameter that penalizes complex trees to prevent overfitting, and γ is a pruning parameter that ensures splits occur only when they improve model performance.

3.5. Model Evaluation

A comparative analysis was conducted to determine the most appropriate ML model for forecasting inpatient discharge volumes. The evaluation involved three regression

algorithms, including RF, XGBoost, and SVR. Each model was assessed using four widely accepted forecasting error metrics: RMSE, R^2, MASE, and MAPE. Figure 3 depicts the performance of the three models, offering a clear framework for comparing.

Figure 3
Comparison of Model Performance Forecasting Error Metrics: RMSE, R^2, MASE, and MAPE



These metrics offer distinct perspectives on model accuracy and generalizability. RMSE quantifies the average magnitude of forecasting errors and is especially sensitive to large deviations, where positive and lower values indicate greater accuracy [20]. In addition, the R^2 values reflect the proportion of variance in the dependent variable that is explained by the independent variables within a regression model. It ranges from 0 to 1, with values closer to 1 indicating a stronger explanatory relationship. While higher R^2 values are generally preferred, the interpretation of what constitutes an acceptable value varies depending on the domain and research context. In empirical social science and applied modeling studies, where real-world data is often noisy and influenced by multiple latent variables, lower R^2 values may still be meaningful. [21] contends that an R^2 value exceeding 0.10 can be deemed acceptable if the model includes statistically significant predictors, as the focus is not solely on maximizing explanatory power but also on understanding variable relationships. Complementing this, emphasize that R^2 remains a robust and interpretable evaluation metric in biomedical regression models, and values above 0.30 can still reflect practical predictive usefulness, especially in complex or high-variance datasets. Therefore, R^2 values in the range of 0.10 to 0.50 and above are widely accepted in applied contexts, particularly when the model contributes to informed decision-making. Meanwhile, MASE provides a scale-independent measure of forecast error, with values below one indicating performance superior to a nave forecasting method. Furthermore, the MAPE, expressed as a percentage, evaluates the average absolute forecasting error relative to actual values. Interpretation guidelines suggest that MAPE values below 10% are highly accurate, while values between 10% and 20% are regarded as accurate.

Based on the evaluation criteria, the Random Forest model demonstrated the best performance, yielding the lowest RMSE of 177.85, the highest coefficient of determination at 0.43, a MASE of 0.31, and a MAPE of 3.23 percent. All these values fall within acceptable thresholds for forecasting accuracy. The XGBoost model also produced satisfactory results, although with slightly higher errors and lower explanatory power. Specifically, it recorded an RMSE of 193.55, a coefficient of determination of 0.32, a MASE of 0.33, and a MAPE of 3.45%. In contrast, the SVR model performed poorly, as indicated by the highest RMSE of 294.06, a negative coefficient of determination of -0.56, a MASE of 0.55, and a MAPE of 5.52%.

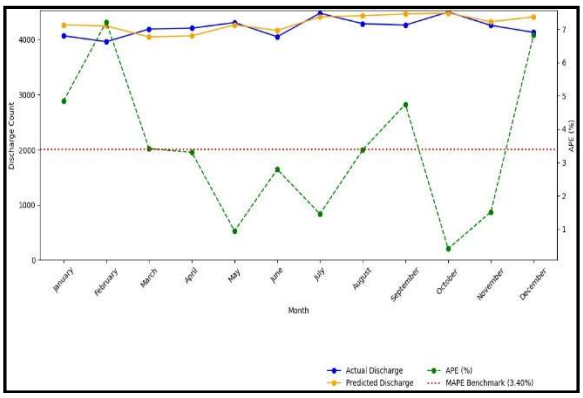
These findings indicate a weak predictive capability, suggesting that the SVR model failed to generalize effectively to the underlying discharge trends in the dataset. These results are summarized in Table 8, which presents the comparative performance metrics for each model across RMSE, R^2, MASE, and MAPE.

Figure 4
Performance Evaluation of The Developed Models

ML Model	RMSE	R^2	MASE	MAPE (%)
Random Forest	177.85	0.43	0.31	3.23
XGBoost	193.55	0.32	0.33	3.45
SVR	294.06	-0.56	0.55	5.52

As part of the prospective validation, the MAPE was used to assess the average deviation between the monthly predicted and actual discharge volumes. MAPE aggregates the monthly APE values and expresses the average forecasting error as a percentage of actual values. The RF model achieved a MAPE of 3.40% across all twelve months of 2024, indicating that its predictions deviated, on average, by only 3.4% from the actual monthly discharge figures.

Figure 4
Monthly Comparison of Actual and Predicted Inpatient Discharges for 2024



This level of precision is well within the range commonly interpreted as highly accurate for healthcare forecasting models, especially in operational environments where MAPE values below 10% are typically deemed acceptable. In addition, MASE was employed to provide a scale-independent evaluation of the model's forecasting accuracy. The resulting MASE of 0.794 indicates that the Random Forest model's error was lower than that of the baseline, reaffirming its capacity to produce accurate and reliable forecasts when applied to unseen data. The result has been summarized in Figure 4.

In this study, several ML algorithms were evaluated, and RF was identified as the best performance model based on forecast error metrics, including R^2, RMSE, MAPE, and MASE. During model development using historical data from 2019 to 2023, the Random Forest model achieved the lowest RMSE of 177.85, a MASE of 0.31, a MAPE of 3.23%, and an R-squared value of 0.43. To further assess the model's practical applicability, prospective validation was conducted using actual inpatient discharge data from 2024. In this phase, the model produced a MAPE of 3.40%, an Absolute Percentage Error (APE) of 1.74% for the total annual forecast, and a MASE of 0.794. These results demonstrate that the RF model maintained strong predictive performance when applied to unseen data, confirming its potential for real-world deployment in forecasting inpatient discharge volumes to aid healthcare management, particularly in resource allocation.

3.5. Benchmark

Most existing studies primarily concentrate on short-term and medium-term predictions like the study done by [18] which focused on forecasting monthly inpatient discharges; [20] focused on weekly inpatient discharges and [21] focused on daily inpatient discharges using time series and ML techniques. These studies offer practical insights into short-term fluctuations in hospital demand and are particularly relevant for informing operational decisions and resource management at the hospital level. Meanwhile, this study introduces a strategic forecasting model to predict annual inpatient discharge volumes which is crucial for long-term hospital planning and policymaking.

Furthermore, previous studies have explored various predictive analysis approaches for forecasting inpatient discharge volume; only a limited number have specifically evaluated ML models within this context. For example, [20] investigated short-term discharge forecasting using LSTM, ARIMA and RF. Their results showed that the Random Forest model achieved the best performance, with the lowest NMSE of 0.2637 and a MAPE of 0.2095. However, their study was confined to short-term predictions. It did not assess other advanced machine learning techniques, such as SVR or XGBoost, both of which were examined in the present study.

A key methodological difference lies in the use of feature engineering. Unlike previous studies, this research applied a deliberate feature construction process, incorporating lagged discharge values, rolling averages, and calendar-based indicators such as month and quarter. These features were designed to capture temporal discharge patterns more effectively and enhance model performance, an aspect not addressed in prior work.

The broader comparison of algorithms, combined with the use of structured feature engineering and the inclusion of prospective validation using real-time 2024 data, marks a significant methodological advancement and demonstrates the robustness of Random Forest in both retrospective and prospective forecasting scenarios.

3.6. Practical Implementation

The findings of this study present several practical implications for facilitating higher levels of healthcare management, covering both state and federal governments in Malaysia. By demonstrating that machine learning algorithms can accurately forecast inpatient discharge volumes, the study offers a data-driven tool to enhance the efficiency of healthcare service delivery and long-term resource allocation planning.

Forecasting inpatient discharge volumes over a longer time period enables hospital administrators and state health authorities to gain a clearer understanding of anticipated service demand. This capability is significant given Malaysia's shifting demographic profile, characterized by an ageing population and an increasing prevalence of chronic NCDs. By anticipating these changes, health facilities are better positioned to plan for essential resources, including bed availability, workforce requirements, and management of medical supplies.

Moreover, integrating predictive models into hospital decision-making processes fosters a transition from reactive responses to proactive planning. Instead of depending solely on retrospective data or ad hoc adjustments, healthcare institutions can utilize forecasted discharge volumes to align their operational strategies with projected demand. This approach helps mitigate the risks of overcrowding, delays in patient discharge, and disruptions in service delivery. Significantly, the inclusion of prospective validation using real-time data from 2024 strengthens the model's credibility for real-world applications. This validation approach simulates actual deployment conditions, enhancing stakeholder confidence in the model's reliability and generalizability within dynamic healthcare environments.

Therefore, the forecasting framework developed in this study can serve as a foundational model for other Malaysian government hospitals. It could also be adapted for predicting related healthcare resource metrics such as admissions, readmission, bed occupancy rates, or outpatient visits, thereby supporting more coordinated planning and enhancing operational efficiency across the healthcare system

4. Conclusion

In conclusion, this study developed and validated a machine learning model to forecast inpatient discharge volumes in government hospitals across the Northern Region of Malaysia. By applying a structured approach to feature engineering and performance evaluation, the Random Forest model emerged as the most reliable for capturing discharge patterns. The study contributes to both methodology and practical application by providing a forecasting framework that supports more informed hospital planning and resource management. These findings underscore the importance of predictive analytics in enhancing healthcare system preparedness, particularly in response to ongoing demographic shifts and the growing burden of chronic diseases.

Recommendations

This study was limited to data from a single government hospital in the Northern Region of Malaysia. To build on these findings, future research should consider involving multiple hospitals from different regions to capture a broader range of discharge patterns and operational contexts. Expanding the dataset in this way would test the model's consistency and reliability across various healthcare settings. Therefore, further validation using multi-center data across diverse settings is recommended to strengthen the model's generalizability and support its use at the national level. Future studies could also explore the inclusion of external factors, such as public health emergencies or policy changes, to make the model more responsive to the changing conditions that affect hospital demand.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

References

[1] OECD. (2023). Health at a glance 2023: OECD indicators. OECD Publishing. https://doi.org/10.1787/4dd50c09-en

[2] Planning Division, Ministry of Health Malaysia. (2023). Malaysia national health accounts health expenditure report 2011–2021. Ministry of Health Malaysia.

[3] World Health Organization. (2021). Global health expenditure database. https://apps.who.int/nha/database

[4] Ministry of Health Malaysia. Health white paper for Malaysia. https://www.moh.gov.my/moh/resources/Penerbitan/Penerbitan%20Utama/Kertas%20Putih%20Kesihatan/Kertas_Putih_Kesihatan_(ENG)_compressed.pdf

[5] Khairul Anuar, I. L., Mohd Kamil, N. D. N., Noris, N. J., Bakit, P. A., Jie, N. R., & Ibrahim, N. H. (2021). Barriers in intervention characteristics of cluster hospital (CH) implementation in Malaysia: An exploratory study. LIFE: International Journal of Health and Life-Sciences, 7(1), 10–24. https://doi.org/10.20319/lijhls.2021.71.1024

[6] Noris, N. J., Ng, R. J., Saimy, I. S., Anuar, K., & Nasir, M. (2023). Cluster hospital implementation in Malaysia's public hospitals: Barriers and boosters from healthcare providers' perspective. *Journal of Health Management,* 20(2), 199–210. https://doi.org/10.1177/09720634231154720

[7] Ajegbile, M. D., Olaboye, J. A., Maha, C. C., Igwama, G. T., & Abdul, S. (2024). The role of data-driven initiatives in enhancing healthcare delivery and patient retention. *World Journal of Biology Pharmacy and Health Sciences,* 19(1), 234–242.

[8] Avinash, G., Pachori, H., Sharma, A., & Mishra, S. (2025). Time series forecasting of bed occupancy in mental health facilities in India using machine learning. *Scientific Reports,* 15(1), 2686. https://doi.org/10.1038/s41598-025-00000-0

[9] Liu, H., Xing, F., Jiang, J., Chen, Z., Xiang, Z., & Duan, X. (2024). Random forest predictive modeling of prolonged hospital length of stay in elderly hip fracture patients. *Frontiers in Medicine,* 11. https://doi.org/10.3389/fmed.2024.1234567

[10] Tello, M., Reich, E. S., Puckey, J., Maff, R., Garcia-Arce, A., Bhattacharya, B. S., et al. (2022). Machine learning-based forecast for the prediction of inpatient bed demand. *BMC Medical Informatics and Decision Making,* 22(1), 270. https://doi.org/10.1186/s12911-022-01999-1

[11] Michailidis, P., Dimitriadou, A., Papadimitriou, T., & Gogas, P. (2022). Forecasting hospital readmissions with machine learning. *Healthcare,* 10(6), 981. https://doi.org/10.3390/healthcare10060981

[12] Tey, S. F., Liu, C. F., Chien, T. W., Hsu, C. W., Chan, K. C., Chen, C. J., et al. (2021). Predicting the 14-day hospital readmission of patients with pneumonia using artificial neural networks (ANN). *International Journal of Environmental Research and Public Health,* 18(10), 5110. https://doi.org/10.3390/ijerph18105

[13] Cheng, F. Y., Joshi, H., Tandon, P., Freeman, R., Reich, D. L., & Mazumdar, M., et al. (2020). Using machine learning to predict ICU transfer in hospitalized COVID-19 patients. *Journal of Clinical Medicine*, 9(6), 1668. https://doi.org/10.3390/jcm9061668

[14] Alajmani, S., & Elazhary, H. (2019). Hospital readmission prediction using machine learning techniques. *International Journal of Advanced Computer Science and Applications*, 10(4), 173–180. https://doi.org/10.14569/IJACSA.2019.0100423

[15] Lin, Y. W., Zhou, Y., Faghri, F., Shaw, M. J., & Campbell, R. H. (2019). Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PLoS ONE,* 14(7), e0218942. https://doi.org/10.1371/journal.pone.0218942

[16] Daghistani, T. A., Elshawi, R., Sakr, S., Ahmed, A. M., Al-Thwayee, A., & Al-Mallah, M. H. (2019). Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. *International Journal of Cardiology,* 288, 140–147. https://doi.org/10.1016/j.ijcard.2019.03.009

[17] 25. Thangarasu. G, Dominic PDD, Subramanian. K, Efficient Energy Usage model for WSN-IoT Environments, International Conference on Computational Intelligence, IEEE, 2020, 252-255

[18] 26. Krishnan.K, Jenefa, L, Kandasamy,L, Thangarasu. G, Impact of AI powered resources on students Performance, Second International Conference on Smart Technologies for Smart Nation, IEEE, 2023, 720-724

[19] 27. Subrmanian. K. Thangarasu. G, An Efficient Air Pollution Prediction model using Machine Learning Algorithms, Journal of Advanced Research in Applied Sciences and Engineering Technology, 47, 2, 2024, 68-75.

[20] Cho SM, Austin PC, Ross HJ, Abdel-Qadir H, Chicco D, Tomlinson G, et al. Machine Learning Compared With Conventional Statistical Models for Predicting Myocardial Infarction Readmission and Mortality: A Systematic Review. Canadian Journal of Cardiology. 2021 Aug;37(8):1207–14

[21] Gharacheh L, Torabipour A, Khiavi F, Malehi A, Haddadzadeh M. Comparison of Statistical Models of Predict the Factors Affecting the Length of Stay (LOS) in the Intensive Care Unit (ICU) of a Teaching Hospital. Materia Socio Medica. 2017;29(2):88