**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**
**IMPACT FACTOR: 7.056**

***IJCSMC, Vol. 15, Issue. 5, May 2026, pg.149 – 153***

# Development of a Machine Learning Model for Predictive Analysis of Patient Trends and Healthcare Service Efficiency

**Colyn D. Vargas*; Neil Glenn T. Balidiong; Mel Adrian T. Cordero; Ronald De La Cruz Jr.; Nicole Faulan; Veronica C. Castillo; Serafin C. Palmares; Kristine T. Soberano**

Master of Information Technology, State University of Northern Negros, Philippines

vargascolyn@gmail.com, neilglennbalidiong.it@gmail.com, meladrian18@gmail.com, ronald290498@gmail.com, nicfaulan@gmail.com, veronica.castillo@antiquespride.edu.ph, spalmares@sunn.edu.ph, ksoberano@sunn.edu.ph

**DOI: https://doi.org/10.47760/ijcsmc.2026.v15i05.016**

**Abstract:** Philippine government hospitals are under considerable strain from escalating patient volumes, yet most tertiary public facilities remain without the data analytics capabilities required to anticipate and prepare for these demands. To address this shortfall, the present study designed and validated two machine learning models: one intended to project patient visit frequencies and another to evaluate the operational efficiency of healthcare delivery. Training and testing relied on a publicly available synthetic Kaggle dataset containing 2,000 physician visit records. Variables captured in the dataset spanned patient gender, geographic location type, clinical department, visit category, diagnosis, waiting and consultation durations, patient-to-staff ratios, average service time, and a binary indicator for peak hours.

Prior to model training, the dataset was subjected to several preparation steps: null entries were handled, categorical attributes were transformed via one-hot encoding, continuous measurements were scaled through normalization, and the full dataset was apportioned into training (80%) and test (20%) segments. Random Forest was employed to project total patient visit volumes, whereas Linear Regression was applied to quantify service efficiency. The Random Forest algorithm attained an $R^2$ of 0.92, an MAE of 5.98, and an RMSE of 8.75, reflecting its capacity to model intricate, nonlinear relationships across multiple predictors. Linear Regression yielded an $R^2$ of 0.88, an MAE of 3.95, and an RMSE of 5.62, demonstrating consistent and interpretable output for service-related operational factors.

These results collectively indicate that Random Forest is more appropriate for modeling complex, nonlinear fluctuations in patient volume, while Linear Regression offers greater transparency when ease of interpretation is

 

prioritized over predictive precision. Both algorithms confirm that machine learning methods are capable of generating actionable healthcare forecasts even when real clinical data are unavailable. This study contributes a reproducible modeling pipeline that could guide hospital planning efforts—especially in resource distribution, workforce scheduling, and operational efficiency management.

**Keywords:** machine learning; predictive analytics; patient admissions; random forest; MAE; RMSE; Google Forms; healthcare efficiency; EHR; Philippines; CRISP-DM; RA 10173

## I. INTRODUCTION

Over the past decade, Machine Learning (ML) has fundamentally altered the way hospitals manage both clinical care and administrative processes. Rather than depending solely on practitioner experience or retrospective reporting, healthcare institutions can now deploy ML systems to sift through extensive datasets and detect patterns invisible to conventional analysis. Established use cases range from forecasting inpatient admissions and optimizing bed utilization to early detection of disease outbreaks and more equitable distribution of scarce resources. In one large comparative evaluation, Badawy et al. [1] showed that ensemble ML approaches achieved a 40% reduction in clinical prediction error relative to traditional statistical methods—a considerable advantage for overstretched facilities. Raman et al. [8] corroborated ML's clinical value by reporting a Random Forest model that reached an ROC AUC of 0.82 when applied to 1,795 electronic health records (EHRs) for outcome prediction. Within the Philippine setting, government tertiary hospitals in Metro Manila routinely run at or beyond bed capacity during specific months, and the absence of demand-forecasting tools forces administrators into reactive management when volumes surge unexpectedly [3].

Interest in ML-driven patient admission forecasting has grown considerably across research communities worldwide. Seo et al. [9] demonstrated the utility of Bidirectional Long Short-Term Memory (Bi-LSTM) networks for predicting ward census in a South Korean hospital, yielding an $R^2$ in the vicinity of 0.600. Sudarshan et al. [11] established that integrating lag-based temporal predictors into support vector regression (SVR) models consistently outperformed ARIMA benchmarks in forecasting inpatient bed requirements. Random Forest and deep learning methods currently rank among the most frequently applied techniques in hospital operations research, with a recognized trade-off between the superior accuracy of deep learning and the interpretability advantages of tree-based models, which nonetheless struggle with purely sequential inputs.

When forecasting tools are unavailable, hospital managers typically resort to intuition and static annual budgets—neither of which accommodates unanticipated surges in patient demand. This systemic weakness produces a cyclical pattern: overcrowding during high-volume periods (particularly the final quarter of the year) and unused capacity during lulls. Khalifa and Khalid [6] documented the consequences empirically, finding that hospitals without analytical infrastructure reported higher rates of bed shortages, unplanned transfers, and staff burnout than data-enabled counterparts. Beyond administrative disruption, the patient-facing effects include extended waiting periods, reduced care quality, and preventable adverse outcomes [3].

This study was designed to directly address those gaps. Its central question was whether Random Forest and Linear Regression—fitted on a 2,000-record synthetic Kaggle dataset encompassing demographic, operational, and service-quality variables—could attain an $R^2$ of at least 0.75 in predicting monthly patient admissions and estimating service efficiency. The investigation followed the Cross-Industry Standard Process for Data Mining (CRISP-DM)—a widely endorsed, phase-structured framework for health informatics projects [5]. The overarching objective was to produce a transferable, reproducible modeling pipeline suitable for adaptation in Philippine government hospital environments.

## II. REVIEW OF RELATED LITERATURE

### *A. ML Applications in Healthcare*

A systematic review by Badawy et al. [1] catalogued Random Forest, Gradient Boosting, SVMs, and LSTM as the predominant algorithms in healthcare ML research, with Random Forest standing out for its reliability in regression tasks—recording $R^2$ values from 0.78 to 0.94 across clinical datasets and establishing a useful performance benchmark. Chen et al. [2] evaluated Gradient Boosting for ICU admission forecasting, achieving AUC scores between 0.68 and 0.93; the authors noted that gains of that magnitude were largely contingent on the rigor of the feature selection process. Swinckels et al. [12] confirmed that LSTM architectures can extract long-horizon temporal dependencies from sequential clinical data, though the associated data requirements substantially exceed those of tree-based models—a practical barrier for smaller public hospitals.

 

### *B. Predictive Analytics for Hospital Admissions*

Raman et al. [8] applied Random Forest to a 1,795-record EHR dataset for COVID-19 severity stratification, achieving an AUC of 0.82 while sustaining performance across a high-dimensional variable space—a result that underscores the algorithm's robustness. Sudarshan et al. [11] further established that temporally enriched SVR models surpass ARIMA in accuracy for inpatient bed-demand forecasting. In the context of ward occupancy prediction, Seo et al. [9] employed Bi-LSTM with blended static and dynamic inputs in a South Korean hospital, reporting $R^2$ values between 0.544 and 0.600—competitive figures for sequential clinical tasks. Mahmoudian et al. [7] extended the forecasting paradigm by combining LSTM-based admission projections with prescriptive bed scheduling in an Iranian hospital, achieving a MAPE of 11.58%.

### *C. Philippine and Southeast Asian Context*

Within the Philippine context, the integration of ML-based clinical decision tools has lagged behind more technologically advanced health systems. Although the Department of Health has invested in broader digital health infrastructure, the operationalization of validated predictive algorithms in government facilities remains uncommon. To the best of the research team's knowledge, no previously published study has developed and validated an ML-based patient admission forecasting model grounded in local Philippine EHR data from district health workers, structured around CRISP-DM, and evaluated against internationally recognized performance criteria—the specific contribution this study seeks to make.

## III. METHODOLOGY

### *A. Research Design*

The study adopted an applied developmental research design, directing effort toward the construction and iterative refinement of a predictive ML system. No actual patient records were utilized at any stage, and the development of a deployable software product was outside the defined scope. The work was constrained to building a reproducible modeling workflow on synthetic data. To ensure methodological rigor and alignment with established data science practice, the Cross-Industry Standard Process for Data Mining (CRISP-DM) was adopted as the organizing framework.

CRISP-DM proceeds through five sequential stages. The first—problem understanding—required the team to articulate research objectives and justify the use of synthetic data as an initial modeling substrate. The second stage examined the Kaggle dataset's composition to confirm the presence of variables requisite for analysis. The third stage encompassed data preparation: cleaning, transformation, and encoding operations performed prior to model input. The modeling stage involved training and comparing multiple ML configurations, with emphasis on ensemble methods, to identify optimal settings. The final evaluation stage benchmarked model accuracy against the study's predefined criteria. This structured progression is designed to support iterative improvements to healthcare service efficiency as the approach is subsequently tested against real clinical data.

### *B. Data Collection*

Data were drawn from Kaggle, an open-access repository hosting de-identified and synthetic research datasets. The selected dataset comprised 2,000 synthetic physician visit records. Because no entries correspond to actual individuals, the dataset raises no privacy or ethical concerns in the context of this research.

The dataset encompasses a broad array of variables reflecting patient demographics and facility operations, including average patient age, sex, geographic classification, visit date and day of week, department category, visit type, diagnosis, severity rating, waiting and consultation durations (in minutes), available staff count, bed occupancy rate, patient-to-staff ratio, mean service time, and a binary peak-hour flag. All variables served as candidate features for the predictive models, enabling analysis of patient flow and service performance within a simulated hospital setting.

### *C. Data Processing*

The raw dataset required several preprocessing operations before model training. Categorical variables were numerically encoded via one-hot encoding, and continuous features were normalized to equalize their influence on model training. The processed data were partitioned into training (80%) and test (20%) subsets. Feature selection was guided by domain knowledge and data availability. For Random Forest—targeting patient visit count—the selected predictors were Gender, Location_Type, Department, Visit_Type, and Diagnosis, which were judged most pertinent to volume prediction. For Linear Regression—targeting service efficiency—the inputs were Avg_Service_Time, Patient_to_Staff_Ratio, and Peak_Hour_Flag, all of which bear a direct operational relationship to service throughput. With preprocessing complete, both models were trained and evaluated on the held-out test partition.

 

### *D. Model Evaluation*

Three standard regression metrics were used to measure performance: R-squared ($R^2$), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). $R^2$ quantifies the proportion of outcome variance explained by the model, with values nearer to 1.0 signifying a stronger fit. MAE expresses average prediction error in the original units of the outcome variable, making it accessible to non-technical audiences. RMSE weights larger deviations more heavily by squaring prediction errors before averaging, which is advantageous for identifying cases where the model underperforms significantly. Used in combination, these three metrics provide a well-rounded assessment of each model's ability to generalize beyond the training data.

### *E. Machine Learning Algorithms*

Random Forest and Linear Regression were the two algorithms selected for this investigation—the former for patient volume modeling and the latter for service efficiency estimation. The choice was deliberate. Random Forest was adopted because it does not presuppose linear variable relationships and tolerates high-dimensional input effectively, which is relevant given the layered interactions embedded in patient visit records. Linear Regression was selected for the inverse reason: its outputs are transparent and straightforward to communicate, which matters when stakeholders need not just predictions but actionable operational insight. More complex approaches, including XGBoost, were considered but deferred; the priority at this stage was establishing a clear, extensible baseline pipeline that could later be tested on authentic Philippine hospital data.

Random Forest generates predictions by aggregating outputs from numerous individually trained decision trees. This ensemble mechanism confers stability by averaging across tree-level variance and curbing overfitting tendencies that affect single-tree classifiers. For this study, the approach was particularly appropriate for capturing divergent signals across Gender, Location_Type, Department, Visit_Type, and Diagnosis—each of which contributes differently to patient volume.

Linear Regression addressed a distinct analytical objective. Rather than forecasting headcounts, the model was tasked with characterizing the functional relationship between service-side variables—Avg_Service_Time, Patient_to_Staff_Ratio, and Peak_Hour_Flag—and overall efficiency scores. Administrators generally favor models of this type over opaque machine learning alternatives because their predictions are readily interpretable and the directional influence of each predictor is visible. This also facilitates the identification of specific operational bottlenecks. Together, Random Forest and Linear Regression form a complementary pairing: the former addresses the intricacy of patient volume dynamics; the latter delivers interpretive clarity for service operations management.

## IV. RESULTS AND DISCUSSION

Each model was evaluated on the held-out test portion of the synthetic Kaggle dataset, with Random Forest tasked to Patient_Visit_Count and Linear Regression to Service_Efficiency_Score. Table 1 presents the full set of performance metrics. The Random Forest model yielded an $R^2$ of 0.92, an MAE of 5.98, and an RMSE of 8.75—values that confirm its ability to resolve nonlinear interactions among Gender, Location_Type, Department, Visit_Type, and Diagnosis. Linear Regression applied to Service_Efficiency_Score produced an $R^2$ of 0.88, an MAE of 3.95, and an RMSE of 5.62—indicating stable and consistent behavior across Avg_Service_Time, Patient_to_Staff_Ratio, and Peak_Hour_Flag. Considered jointly, these results delineate a natural functional division: Random Forest for the complexity of patient volume modeling; Linear Regression for the interpretability requirements of service efficiency assessment. Together, they form viable paired components of a hospital decision-support system addressing both resource planning and operational optimization.

***Table 1. Model Performance Metrics***

| Model | $R^2$ | MAE | RMSE |
|---|---|---|---|
| Random Forest | 0.92 | 5.98 | 8.75 |
| Linear Regression | 0.88 | 3.95 | 5.62 |

### *A. Discussion*

The performance numbers in Table 1 merit closer examination. The Random Forest model ($R^2$ = 0.92, MAE = 5.98, RMSE = 8.75) successfully navigated the nonlinear interdependencies among Gender, Location_Type, Department, Visit_Type, and Diagnosis without significant loss of accuracy. This outcome implies the algorithm could function reliably under the heterogeneous demographic and operational conditions characteristic of Philippine public hospitals. The magnitude of the error values also suggests that the model could realistically support advance planning of staffing levels and patient load management.

 

The Linear Regression model ($R^2$ = 0.88, MAE = 3.95, RMSE = 5.62) was slightly less accurate than Random Forest in absolute terms, but this gap is secondary to what the model provides in terms of practical utility. Its Service_Efficiency_Score outputs—derived from Avg_Service_Time, Patient_to_Staff_Ratio, and Peak_Hour_Flag—are straightforward enough for administrative personnel to interpret and act on without requiring technical expertise in machine learning. This accessibility makes it especially well-suited to identifying specific operational bottlenecks. Collectively, the two models respond to complementary planning needs: one equips administrators with patient volume projections; the other provides a means of evaluating whether service delivery capacity is aligned with those projections.

## V. CONCLUSION

This study constructed and evaluated two machine learning models—Random Forest for patient visit volume forecasting and Linear Regression for service efficiency estimation—using a synthetic 2,000-record Kaggle dataset as the basis for training and testing. Preprocessing operations included null value resolution, one-hot encoding of categorical fields, normalization of continuous measurements, and an 80/20 train-test split to yield clean, comparable data for each algorithm. On the test set, Random Forest produced an $R^2$ of 0.92, an MAE of 5.98, and an RMSE of 8.75, reflecting strong capacity to model intricate interactions across patient gender, location, department, visit type, and diagnosis. Applied to service-related inputs, Linear Regression achieved an $R^2$ of 0.88, an MAE of 3.95, and an RMSE of 5.62, generating stable and interpretable estimates across average service time, patient-to-staff ratio, and peak hour status.

Comparing the two, Random Forest is the stronger option for nonlinear patient trend forecasting, whereas Linear Regression maintains a practical edge where transparent, communicable outputs are the priority. Both models cleared the study's $R^2 \geq 0.75$ benchmark, validating the use of machine learning for healthcare demand forecasting even when only synthetic data are available. As a combined system, they provide a decision-support framework that could help administrators project patient volume changes and isolate the operational variables most consequential for service delivery. The logical next step is applying this pipeline to authentic clinical data from Philippine government hospitals, broadening the feature set, and exploring whether more sophisticated algorithms can extend performance further while preserving the interpretability that makes the current results actionable in real-world settings.

# REFERENCES


**[1].** M. Badawy, N. Ramadan, and H. A. Hefny, "Healthcare predictive analytics using machine learning and deep learning techniques: a survey," Journal of Electrical Systems and Information Technology, vol. 10, no. 1, pp. 1–27, 2023. https://doi.org/10.1186/s43067-023-00108-y

**[2].** Z. Chen et al., "A Novel Machine Learning-Based Prediction Model for ICU Admission in Patients with COVID-19," Journal of Medical Internet Research, vol. 23, no. 6, p. e25678, 2021. https://doi.org/10.2196/25678

**[3].** Commission on Audit (COA), "COA Annual Report: Department of Health," Philippine Commission on Audit, Quezon City, Philippines, 2022.

**[4].** Department of Health (DOH), "Philippine Health Facility Development Plan (PHFDP) 2020–2040," DOH, Manila, Philippines, 2020.

**[5].** D. T. Larose and C. D. Larose, Data Mining and Predictive Analytics, 2nd ed. Hoboken, NJ: Wiley, 2015.

**[6].** M. Khalifa and M. Khalid, "Developing Predictive Machine Learning Models to Support Decision Making in Healthcare," Procedia Computer Science, vol. 184, pp. 840–845, 2021. https://doi.org/10.1016/j.procs.2021.03.106

**[7].** A. Mahmoudian et al., "An LSTM-based forecasting approach for hospital bed management," Health Care Management Science, vol. 24, pp. 563–574, 2021.

**[8].** G. Raman et al., "Random Forest-Based Prediction of COVID-19 Severity," Frontiers in Medicine, vol. 8, p. 637987, 2021. https://doi.org/10.3389/fmed.2021.637987

**[9].** J. Seo et al., "Forecasting Patient Census Using Bidirectional Long Short-Term Memory Network," PLOS ONE, vol. 17, no. 5, p. e0269825, 2022.

**[10].** F. Siddiqui et al., "Predictive Analytics in Healthcare: Improving Patient Outcomes," International Journal of Medical Informatics, vol. 169, p. 104931, 2023.

**[11].** V. K. Sudarshan et al., "Application of Wavelet Techniques in ECG Signal Processing: An Overview," IRBM, vol. 35, no. 6, pp. 306–318, 2014.

**[12].** T. Swinckels et al., "Using LSTM Networks to Predict Hospital Census," International Journal of Medical Informatics, vol. 165, p. 104820, 2022.


 