Jurnal Sistem Informasi Bisnis 04(2025)

On-line: http://ejournal.undip.ac.id/index.php/jsinbis/article/view/80147

Early Detection of Patient Surge Anomalies in Hospitals: A Comparative Analysis of Gradient Boosting, Random Forest, and SVM

Masparudin[1*], Marfuah[2], Abdullah[3]

[1] Department of Software Engineering, Universitas Universal, Batam, Indonesia
[2] Department of Information Systems Universitas Universal, Batam, Indonesia
[3] Department of Information Systems Universitas Islam Indragiri, Tembilahan, Indonesia

Submitted: October 15th, 2025; Revised: November 17th, 2025; Accepted: December 8th, 2025; Available Online: December 29th, 2025

DOI: 10.21456/vol15iss4pp488-495

Abstract

Unpredictable fluctuations in patient visits often lead to resource unpreparedness and decreased service quality in hospitals. This study aims to develop an early warning system for patient surges across 110 healthcare service units. Unlike conventional approaches utilizing static thresholds, this study proposes a Statistical Anomaly Detection method based on Z-Score for dynamic labeling and applies Synthetic Minority Over-sampling Technique (SMOTE) to address extreme data imbalance. Three classification algorithms—Gradient Boosting Classifier (GBC), Random Forest (RF), and Support Vector Machine (SVM)—were compared using time-series lag features and volatility trends. Experimental results demonstrate that Gradient Boosting outperformed other methods, achieving the highest F1-Score of 37.35% and a Recall of 48.98%. Although the F1-Score reflects the extreme nature of the data imbalance, achieving high recall is explicitly prioritized in healthcare operations to minimize the critical risk of missed surge events. This study concludes that integrating statistical anomaly-based labeling with ensemble boosting algorithms effectively mitigates noise in heterogeneous hospital visit data, thereby serving as a reliable basis for proactive managerial decision-making.

Keywords: Data Mining; Gradient Boosting; Patient Surge; Z-Score; Time-Series Prediction.

1. Introduction

Uncertainty in patient visit and admission volumes poses a critical challenge to modern hospital operational management. Patient surges—particularly in Emergency Departments (ED) or inpatient care—can lead to resource overload involving medical staff, beds, and treatment facilities. This condition has been specifically linked to increased waiting times, patient boarding, and higher safety risks (Tuominen et al., 2024) To anticipate these surges and optimize resource allocation, recent studies have increasingly adopted data-driven approaches and machine learning (ML).

Previous research has demonstrated the effectiveness of ML in various forecasting scenarios within healthcare. For instance, XGBoost has been utilized to predict short-term ED admissions based on real-time records (King et al., 2022), while Support Vector Machine Regression (K-SVR) has shown low error rates in forecasting weekly inpatient bed demands (Tello et al., 2022). Furthermore, integrating calendar variables, weather data, and time-series feature engineering into non-linear models like Random Forest (RF) has significantly enhanced the accuracy of daily patient arrival predictions (Peláez-rodríguez et al., 2024; Porto & Fogliatto, 2024; Tuominen et al., 2022).

However, the majority of the current literature focuses heavily on aggregate predictions, such as total hospital arrivals or overall bed demand. The critical aspect of detecting abnormal surges at the individual service unit level (e.g., specific clinics or wards) remains underexplored, despite its vital role in granular capacity management. A systematic review indicates that classical statistical methods and "one-size-fits-all" static thresholds are often less sensitive to the high heterogeneity of data across different service units (Jiang et al., 2023).

Addressing this specific gap, this study introduces a dynamic labeling approach based on statistical anomalies (Z-Score) to detect patient surges relative to specific service unit baselines, providing a more precise alternative to static global thresholds. To handle the inherent extreme data imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied. Finally, this research performs a comparative analysis of three classification algorithms—Gradient Boosting Classifier (GBC), Random Forest (RF), and Support Vector Machine (SVM)—incorporating time-series lag features and volatility trends. This study aims to identify the most robust algorithm to serve as a reliable early warning system, thereby strengthening proactive resource allocation and minimizing overload risks in multi-service hospitals.

*) Corresponding author: masparudin.mahmud@gmail.com

On-line: http://ejournal.undip.ac.id/index.php/jsinbis/article/view/id_paper

2. Theoretical Framework

2.1 Capacity Management and the Patient Surge Phenomenon

Patient surge is defined as a sudden increase in patient volume that exceeds the standard operational capacity of a healthcare service unit (Hick et al., 2013). This phenomenon frequently leads to overcrowding, which negatively impacts patient safety and clinical efficiency. Theoretically, hospital capacity management seeks to balance demand (patient visits) and supply (staff, beds). However, the stochastic nature of visit data, coupled with seasonal influences, complicates the use of static capacity planning methods (Hoot & Aronsky, 2020). Therefore, predictive approaches are necessary to anticipate workload anomalies before they occur.

2.2 Data Mining and Imbalanced Learning

Data mining is the process of extracting non-trivial patterns from large datasets to yield useful knowledge. In the context of medical prediction, a primary challenge often encountered is class imbalance. Statistically, "Surge" events are rare (minority) occurrences compared to "Normal" conditions (majority) (Haibo He & Garcia, 2009). Standard algorithms often exhibit a bias toward the majority class. Consequently, appropriate preprocessing strategies and the selection of algorithms robust to data imbalance, such as Ensemble Learning methods, are crucial in this study.

2.3 Statistical Anomaly Detection

In contrast to static thresholding approaches, which assign a fixed limit across all units, statistical approaches detect outliers based on the distribution of historical data. One of the most robust methods employed is the Z-Score (Standard Score).

The Z-Score quantifies the distance a data point deviates from its population mean, expressed in units of standard deviation. The mathematical formula for the Z-Score is presented in Equation (1) as follows:

$$Z_i = \frac{x_i - \mu_s}{\sigma_s} \quad (1)$$

As represented in Equation (1), x_i denotes the volume of patient visits during a specific period, while μ_s and σ_s represent the mean and standard deviation, respectively, of the service unit's historical data.

In this study, the Z-value serves as the basis for dynamic class labeling. If the Z-value exceeds a specific threshold (e.g., > 0.85), the data point is classified as an anomaly or 'Surge' (Aggarwal, 2017). This approach enables scale normalization across service units characterized by heterogeneous data volumes.

2.4 Classification Algorithms

This study compares three prominent machine learning algorithms to determine the optimal model

2.4.1 Random Forest (RF)

Random Forest is an ensemble algorithm based on the Bagging (Bootstrap Aggregating) technique. This algorithm constructs multiple decision trees in parallel and aggregates their prediction outcomes through a majority voting mechanism.

The primary advantage of RF lies in its capability to handle high-dimensional data and mitigate the risk of overfitting often associated with single decision trees. (Breiman, 2001).

2.4.2 Support Vector Machine (SVM)

SVM is a supervised learning algorithm that operates by identifying the optimal hyperplane to separate two data classes with the maximum margin. For non-linearly separable data, SVM employs the Kernel Trick technique (such as the RBF Kernel) to project data into higher-dimensional spaces. SVM is renowned for its robust generalization performance, particularly on datasets with limited yet complex sample sizes (Cortes & Vapnik, 1995).

2.4.3 Gradient Boosting Classifier (GBC)

Gradient Boosting is an ensemble technique that constructs models sequentially. Unlike RF, which builds trees independently, GBC constructs new trees to correct the residual errors of their predecessors.

Mathematically, the model aims to minimize a loss function using the gradient descent method. This algorithm frequently yields higher prediction accuracy compared to RF, particularly on datasets featuring subtle anomaly patterns or class imbalance (Friedman, 2001).

2.5 Time-Series Feature Engineering

Given the time-series nature of patient visit data, the incorporation of temporal features is a fundamental aspect of this modeling framework. This study implements feature engineering techniques utilizing Lag Features and volatility analysis. Specifically, Lag Features are leveraged to capture historical dependencies by using visit data from prior time steps (t-1, t-2) as predictors for the current state. Concurrently, service stability within a specific time window is quantified using Rolling Standard Deviation (volatility). The integration of these features enables classification algorithms to go beyond merely processing isolated data points; it allows them to learn trend patterns and the momentum of data fluctuations in depth (Hyndman & Athanasopoulos, 2018).

On-line: http://ejournal.undip.ac.id/index.php/jsinbis/article/view/80147

2.6 System Performance Evaluation

Given the limitations of relying solely on accuracy to measure model performance in binary classification tasks with imbalanced datasets, this study employs a comprehensive suite of evaluation metrics. Beyond calculating Accuracy as the percentage of total correct predictions, the evaluation prioritizes Precision to measure the model's exactness in predicting the positive class ("Surge"), and Recall (Sensitivity) to assess the model's capability to capture all actual surge instances. Furthermore, as the primary indicator of the model's efficacy in handling data imbalance, this study utilizes the F1-Score, which represents the harmonic mean of Precision and Recall (Sokolova & Lapalme, 2009).

3. Method

3.1 Research Framework

This research is conducted through a systematic series of stages adopting the Knowledge Discovery in Databases (KDD) standard. The process encompasses data collection, pre-processing, feature engineering, statistical anomaly-based labeling, comparative modeling, and culminates in performance evaluation. The research framework is illustrated in Figure 1.

3.2 Data Acquisition

The data utilized in this study constitutes secondary data obtained from the daily patient visit records of the Batam Concession Agency Hospital (RSBP Batam), covering the timeframe from January 2021 to September 2025.

The dataset comprises 3,825 entries spanning 122 distinct service units (e.g., General Clinic, Emergency Department, Radiology, etc.). The data attributes include the record date, service unit name, visit type, and total visit volume.

3.3 Data Preprocessing

The data preprocessing stage aims to guarantee data quality prior to the modeling phase. This process begins with data cleaning, which encompasses addressing inconsistent values and the imputation or removal of missing values. Furthermore, service entries identified as having zero variance or being inactive were eliminated to prevent potential model bias (Han et al., 2022). Subsequently, data transformation was performed by converting date formats into datetime objects, followed by chronological sorting based on Service Unit and Record Date. This step is essential to preserve the integrity of the temporal order, which is critical for time-series analysis.

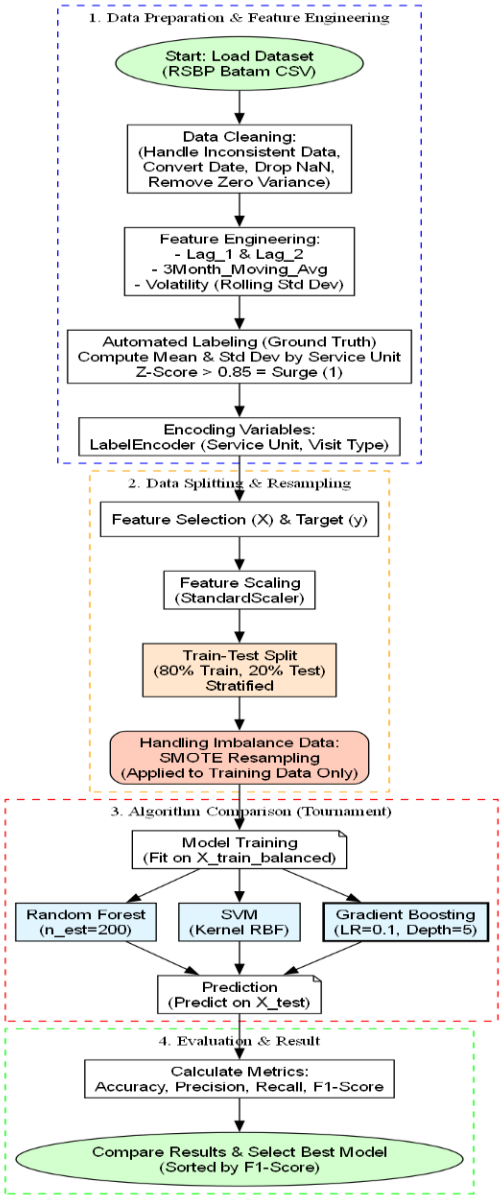

Figure 1. Research framework

3.4 Feature Engineering

Given the time-series nature of the data, the extraction of additional features is critical for effectively capturing temporal patterns and historical trends (Masini et al., 2023). Within this framework, a set of features was constructed, comprising Lag Features $(t-n)$, Moving Average (MA_3), and Service Volatility $(\sigma_{rolling})$. Specifically, Lag Features were generated by retrieving visit values from the previous one $(t-1)$ and two months $(t-2)$, enabling the model to learn short-term historical dependencies. Furthermore, general trends were captured via a Moving Average with a three-month window, while the stability of visit fluctuations for each service unit was quantified using the Volatility feature, calculated based on the rolling standard deviation over the same period.

3.5 Z-Score-Based Dynamic Labeling (Proposed Method)

The primary contribution of this study is the implementation of a dynamic labeling method utilizing Statistical Anomaly Detection. Rather than relying on static thresholds, the "Surge" status is determined based on the statistical deviation from the specific service unit's historical mean (Blázquez-García et al., 2022). The calculation of the Z-Score is performed for every data point with reference to Equation (1). Based on the calculation results, the data is classified into two primary categories, namely Class 1 ($Z > 0.85$) to represent a Surge condition or Class 0 ($Z \leq 0.85$) to denote a Normal condition.

A threshold of 0.85 was selected based on empirical experimentation, which isolates approximately 20% of the most extreme occurrences as anomalies warranting close attention, thereby yielding a dataset significantly reduced in noise.

3.6 Modeling Scenarios

This study evaluates the comparative performance of three distinct Supervised Learning algorithms. First, the Gradient Boosting Classifier (GBC) was employed with a configuration of a learning rate of 0.1 and a maximum tree depth of 5, aimed at effectively handling data imbalance through an iterative learning approach. As a comparator within the ensemble category, Random Forest (RF) was established as a bagging-based baseline model utilizing 200 estimators. To complement this analysis, Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel was also implemented to assess the performance of a non-tree-based model against the existing data characteristics, thereby providing a comprehensive comparative perspective.

Data partitioning into training and testing sets was performed with an 80:20 proportion using the Stratified Sampling technique to ensure that the ratio of the surge class remained balanced across both datasets (Kohavi & Edu, 1995).

In addition to employing class weighting within the algorithms, this study applied the Synthetic Minority Over-sampling Technique (SMOTE) to address extreme imbalance in the training data (Chawla et al., 2002). SMOTE operates by synthesizing new samples for the minority class (surge) based on k-nearest neighbors, thereby enabling the model to learn a clearer decision boundary without majority bias. Crucially, this technique was applied exclusively to the training set, while the testing set was kept unaltered to preserve the validity of the evaluation (Nießl et al., 2022).

3.7 Performance Evaluation

Given that this study addresses an imbalanced classification problem, reliance solely on the accuracy metric is deemed insufficient to accurately represent the model's true performance (Grandini et al., 2020). Consequently, performance evaluation is conducted based on the Confusion Matrix, focusing on three critical indicators. First, Precision is calculated to quantify the accuracy of surge predictions in order to minimize false alarms, as presented in Equation (2).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Second, Recall (Sensitivity) is evaluated via Equation (3) to ensure the model's capability to capture all actual surge occurrences, thereby minimizing the risk of missed detections.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

As a balanced metric, Equation (4) presents the F1-Score, which represents the harmonic mean of Precision and Recall, establishing it as the primary indicator of model success.

$$F1 = \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

In these mathematical formulations, the variable TP represents True Positives (correctly detected surges), while FP and FN refer to False Positives (false alarms) and False Negatives (missed surges), respectively.

4. Results and Discussion

4.1 Results

This subsection outlines the empirical findings derived from a series of experiments, ranging from the analysis of data characteristics and the application of dynamic labeling, to the evaluation of classification model performance following the handling of data imbalance.

4.1.1 Exploratory Data Analysis (EDA)

The research data encompasses a recapitulation of patient visits spanning from January 2021 to September 2025.

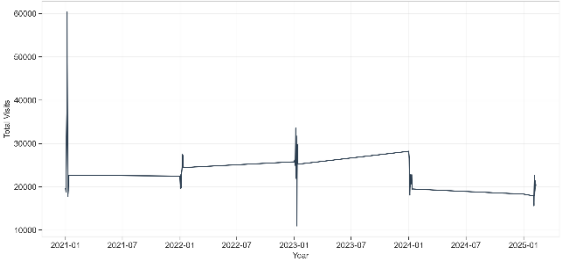

Figure 2. Daily patient visit trends (2021–2025).

Figure 2 presents a visualization of daily patient visit trends throughout the study period (2021–2025). The graph exhibits highly dynamic fluctuations,

On-line: http://ejournal.undip.ac.id/index.php/jsinbis/article/view/80147

characterized by distinct seasonal patterns and sporadic extreme spikes.

This high variability confirms that the employment of static thresholds (e.g., setting a fixed limit of >50 patients across all days) renders them ineffective. Consequently, the Z-Score approach proposed in this study is particularly relevant, as it adapts surge detection boundaries based on the historical standard deviation observed in the data.

Initial exploratory analysis revealed significant variability across service units, characterized by extreme volume heterogeneity. Major departments such as the Laboratory and Emergency Department (ED) recorded thousands of monthly visits, standing in sharp contrast to specialized services like Pediatric Surgery, which exhibited a visit intensity of fewer than 50 patients.

This striking volume disparity underscores the urgency of implementing the Z-Score-based normalization method as a more objective approach compared to absolute thresholding. Furthermore, the temporal trend visualization demonstrates consistent seasonal patterns with recurrent fluctuations, alongside the emergence of surge anomalies closely correlated with specific periods (e.g., pandemic waves or holiday seasons).

4.1.2 Dynamic Labeling Results (Z-Score Labeling)

The implementation of the Statistical Anomaly Detection method, utilizing the specific threshold $Z > 0.85$ successfully mapped workload status in a more proportional manner compared to conventional quantile methods.

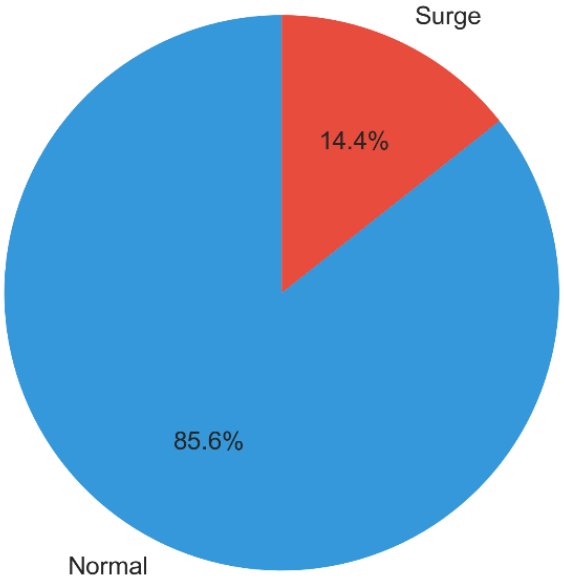

Figure 3. Proportion of Normal and Surge Classes

Figure 3 illustrates the results of the dynamic labeling process utilizing the Z-Score method. From the total clean data entries, a 'Surge' event proportion of 14.4% was identified, while the remainder was categorized as 'Normal'. This proportion reflects real-world hospital conditions, where surges constitute anomalous events rather than routine occurrences.

This labeling method proved to be adaptive to the heterogeneous characteristics of the service units. For instance, in low-volume units such as the Dental Clinic, an increase of 15 patients is categorized as a surge. Conversely, in high-volume units like the Laboratory, an increase of 100 patients may still be considered a normal fluctuation if the standard deviation is high.

Nevertheless, this class distribution disparity (class imbalance) poses a distinct challenge for the classification process. This necessitates the application of the SMOTE Resampling technique during the training phase to prevent model bias toward the majority class.

4.1.3 Algorithm Performance Comparison

Three algorithms (Gradient Boosting, Random Forest, and SVM) were evaluated using cross-validation scenarios with an 80:20 training-testing data split. Given the class imbalance, the evaluation prioritized the F1-Score as the primary metric, rather than relying solely on Accuracy.

Table 4.1. Algorithm Performance Evaluation Matrix (Post-SMOTE)

Algorythm	Acc.	Precision	Recall	F1-Score
GBC	76.36%	30.19%	48.98%	37.35%
RF	77.53%	29.32%	39.80%	33.77%
SVM	63.44%	17.03%	39.80%	23.85%

The experimental results exhibit interesting dynamics. Although **Random Forest** achieved the highest **Accuracy (77.53%)**, the algorithm tended to be **conservative**, exhibiting a lower Recall. Conversely, the **Gradient Boosting Classifier (GBC)** was identified as the **superior method**, securing the highest **F1-Score (37.35%)** and the highest **Recall (48.98%)**. This indicates that GBC is significantly more sensitive in detecting **actual surge threats** compared to the other models.

4.2 Discussion

This subsection interprets the significance behind the numerical findings, analyzes the impact of resampling techniques, and discusses the implications for hospital operational management.

4.2.1 Effectiveness of Addressing Data Imbalance (SMOTE)

A critical finding of this study is the pivotal importance of addressing data imbalance. In preliminary experiments conducted without resampling, all models exhibited a pronounced bias toward the majority class ("Normal"), yielding Recall values below 10%. The implementation of SMOTE (Synthetic Minority Over-sampling Technique) on the

On-line: http://ejournal.undip.ac.id/index.php/jsinbis/article/view/id_paper

training dataset proved effective in rectifying the model's decision boundaries.

Specifically for Gradient Boosting, SMOTE drastically enhanced surge detection capability (Recall) from approximately 9% to 48.98%. Although this resulted in a slight decline in global Accuracy (due to an increase in False Alarms), this trade-off is deemed highly acceptable within a medical context. In such settings, the failure to detect a surge (False Negative) carries operational risks that are far more detrimental than those associated with false warnings.

Regarding the observed reduction in Precision (and the consequent rise in False Alarms), this phenomenon of metric shift aligns with the findings of Fernandez et al., who state that oversampling algorithms operate by expanding the decision regions of the minority class. While this inherently increases the risk of false positives, it remains crucial for minimizing missed detections (Fernandez et al., 2018).

4.2.2 Analysis of Gradient Boosting Superiority

The superior performance of the Gradient Boosting Classifier (GBC) compared to Random Forest (RF) and SVM in terms of the F1-Score can be attributed to two fundamental factors.

First, regarding the algorithmic mechanism, GBC employs a boosting approach that constructs trees sequentially, wherein each subsequent tree focuses on correcting the residual errors of its predecessor. This stands in contrast to RF, which utilizes the bagging technique characterized by independent tree construction. This boosting approach frequently outperforms bagging on datasets with high noise levels due to its capacity for incremental bias reduction, particularly regarding hard-to-classify samples (Hastie, 2009).

Second, concerning model adaptability to data characteristics, the underperformance of SVM—which achieved only 63% accuracy and a 23% F1-Score—indicates that patient visit patterns possess complex and overlapping non-linear decision boundaries. These patterns prove difficult to disentangle even when utilizing the RBF kernel, yet they can be mapped more effectively via the hierarchical "if-then" rule structure inherent in GBC decision trees.

This finding corroborates the observations of Ben-Hur and Weston (2010), who noted that SVM often encounters difficulties in identifying an optimal hyperplane on datasets characterized by extreme class imbalance and high noise without highly specific kernel tuning (Ben-Hur & Weston, 2010). .

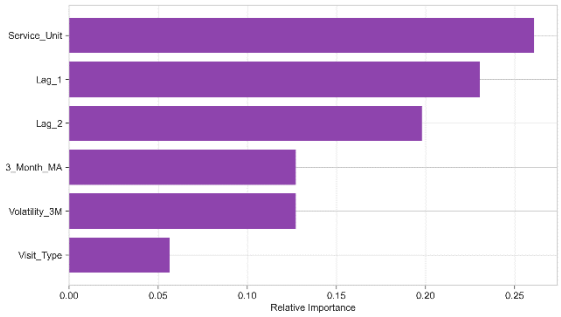

Figure 4. Feature Importance of the GBC Model

To elucidate the factors influencing the model's predictions, Figure 4 presents the Feature Importance analysis derived from the best-performing model (GBC). It is evident that the 3_Month_MA (3-Month Moving Average) and Lag Features (visits from the preceding 1-2 months) exhibit the most significant contributions.

This finding validates the hypothesis that patient surges possess strong autocorrelation properties, indicating that short-term historical trends serve as the most reliable predictors for future events.

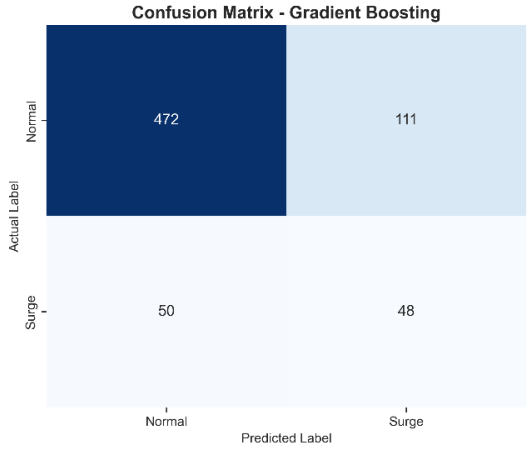

Figure 5. Confusion Matrix of the GBC Model

The detailed prediction performance of the Gradient Boosting model is visualized via the Confusion Matrix in Figure 5. This matrix illustrates the model's efficacy in minimizing False Negatives (missed surges).

The substantial count of True Positives within the 'Surge' class confirms that this early warning system possesses sufficient sensitivity to serve as an operational decision support tool for the hospital, albeit with a certain number of False Positives that require mitigation through management policies.

On-line: http://ejournal.undip.ac.id/index.php/jsinbis/article/view/80147

4.2.3 The Role of Historical Features and Volatility

Model analysis confirms that the engineered Time-Series features play a pivotal role in enhancing prediction performance. Specifically, Lag Features proved effective in capturing short-term autocorrelation, where current surge phenomena are strongly influenced by the visit momentum of the preceding one to two days (Wilson, 2016).

Furthermore, the incorporation of the Volatility feature (Rolling Standard Deviation) successfully assisted the model in differentiating characteristics across service units. This feature enables the algorithm to discriminate between units with low standard deviations, which tend to be stable, versus highly volatile units such as the Emergency Department (ED).

Without the inclusion of this volatility feature, the model would encounter significant difficulty in determining whether a 10% volume increase represents a critical anomaly or merely a normal operational variation.

4.2.4 Managerial Implications

The proposed early detection system serves as a vital Decision Support System (DSS), particularly in facilitating proactive anticipation and operational risk mitigation.

With a Recall rate approaching 50%, the system possesses the capability to identify nearly half of all total extreme surge incidents before they occur, providing management with crucial lead time for resource preparation.

In the context of risk mitigation, although Precision resides in the 30% range—implying that out of ten alerts, three represent actual surges while seven constitute precautionary alerts—this performance is far superior to a purely reactive approach. These alerts can be operationalized as a "Yellow Light" mechanism or standby status to ensure the availability of human resources and logistics, thereby preventing service collapse due to overcrowding (Mullainathan & Obermeyer, 2017).

5. Conclusion

This study demonstrates that the integration of Z-Score-based dynamic labeling and the SMOTE data balancing technique effectively addresses the imbalance issue in patient visit data. Based on the evaluation, the Gradient Boosting Classifier (GBC) emerged as the best-performing model, outperforming Random Forest and SVM, achieving an F1-Score of 37.35% and a Recall of 48.98%. The application of SMOTE proved crucial in improving anomaly detection sensitivity by up to fivefold compared to the unbalanced scenario.

In terms of implications, this model enables hospital management to shift from a reactive to a proactive approach through an early warning system capable of anticipating nearly half of extreme surge events. This facilitates efficient medical resource allocation before a surge occurs. Although the False Positive rate remains moderate, the risk of false alarms is considered more tolerable for the sake of patient safety compared to the risk of failing to detect actual surges.

45 The limitations of this study lie in the model's reliance on internal historical features, the relatively high rate of false alarms, and the use of a fixed empirical threshold ($Z > 0.85$) without an exhaustive sensitivity analysis or non-ML baselines. Therefore, future work may evaluate adaptive threshold optimization strategies. Additionally, future research is suggested to explore Deep Learning algorithms, such as LSTM or GRU, and to integrate external variables (e.g., epidemiological or weather data) to capture more complex temporal patterns and improve prediction precision.

References

Aggarwal, C. C., 2017. *Outlier Analysis*. Springer International Publishing. https://doi.org/10.1007/978-3-319-47578-3

Ben-Hur, A., & Weston, J., 2010. *A User's Guide to Support Vector Machines* (pp. 223–239). https://doi.org/10.1007/978-1-60327-241-4_13

Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A., 2022. A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Computing Surveys*, *54*(3), 1–33. https://doi.org/10.1145/3444690

Breiman, L., 2001. Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Cortes, C., & Vapnik, V., 1995. Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V., 2018. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/10.1613/jair.1.11192

Friedman, J. H., 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Grandini, M., Bagli, E., & Visani, G., 2020. *Metrics for Multi-Class Classification: an Overview*. 1–17. http://arxiv.org/abs/2008.05756

Haibo He, & Garcia, E. A., 2009. Learning from Imbalanced Data. *IEEE Transactions on*

Knowledge and Data Engineering, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Han, J., Pei, J., & Tong, H., 2022. Data Mining: Concepts and Techniques, Fourth Edition. In *Data Mining: Concepts and Techniques, Fourth Edition*. https://doi.org/10.1016/C2013-0-18660-6

Hastie, T. et. all., 2009. Springer Series in Statistics The Elements of Statistical Learning. *The Mathematical Intelligencer*, *27*(2), 83–85. http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

Hick, J. L., Barbera, J. A., And, & Kelen, G. D., 2013. Refining Surge Capacity: Conventional, Contingency, and Crisis Capacity. *Disaster Medicine and Public Health Preparedness*, *3*(S1), S59–S67. https://doi.org/https://doi.org/10.1097/DMP.0b013e31819f1ae2

Hoot, N. R., & Aronsky, D., 2020. *HHS Public Access*. *52*(2), 126–136. https://doi.org/10.1016/j.annemergmed.2008.03.014.Systematic

Hyndman, R. J., & Athanasopoulos, G., 2018. *Forecasting : Principles and Practice Chapter 1 Getting started*. 291.

Jiang, S., Liu, Q., & Ding, B., 2023. *A systematic review of the modelling of patient arrivals in emergency departments*. *13*(3), 1957–1971. https://doi.org/10.21037/qims-22-268

King, Z., Farrington, J., Li, K., & Crowe, S., 2022. *Machine learning for real-time aggregated prediction of hospital admission for emergency patients*. 1–12. https://doi.org/10.1038/s41746-022-00649-y

Kohavi, R., & Edu, S., 1995. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. 1–7. papers://5e3e5e59-48a2-47c1-b6b1-a778137d3ec1/Paper/p2015

Masini, R. P., Medeiros, M. C., & Mendes, E. F., 2023. Machine learning advances for time series forecasting. *Journal of Economic Surveys*, *37*(1), 76–111. https://doi.org/10.1111/joes.12429

Mullainathan, S., & Obermeyer, Z., 2017. Does Machine Learning Automate Moral Hazard and Error? *American Economic Review*, *107*(5), 476–480. https://doi.org/10.1257/aer.p20171084

Nießl, C., Herrmann, M., Wiedemann, C., Casalicchio, G., & Boulesteix, A., 2022. Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. *WIREs Data Mining and Knowledge Discovery*, *12*(2). https://doi.org/10.1002/widm.1441

Peláez-rodríguez, C., Torres-lópez, R., Pérez-aracil, J., & López-laguna, N., 2024. An explainable machine learning approach for hospital emergency department visits forecasting using continuous training and multi-model regression. *Computer Methods and Programs in Biomedicine*, *245*(January), 108033. https://doi.org/10.1016/j.cmpb.2024.108033

Porto, B. M., & Fogliatto, F. S., 2024. Enhanced forecasting of emergency department patient arrivals using feature engineering approach and machine learning. *BMC Medical Informatics and Decision Making*, *6*. https://doi.org/10.1186/s12911-024-02788-6

Sokolova, M., & Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Tello, M., Reich, E. S., Puckey, J., Maff, R., Arce, A. G., Bhattacharya, B. S., & Feijoo, F., 2022. Machine learning based forecast for the prediction of inpatient bed demand. *BMC Medical Informatics and Decision Making*, *9*, 1–13. https://doi.org/10.1186/s12911-022-01787-9

Tuominen, J., Lomio, F., Oksala, N., Palomäki, A., & Peltonen, J., 2022. Forecasting daily emergency department arrivals using high - dimensional multivariate data : a feature selection approach. *BMC Medical Informatics and Decision Making*, *7*, 1–12. https://doi.org/10.1186/s12911-022-01878-7

Tuominen, J., Pulkkinen, E., Peltonen, J., & Kanniainen, J., 2024. Forecasting emergency department occupancy with advanced machine learning models and multivariable input ☆. *International Journal of Forecasting*, *40*(4), 1410–1420. https://doi.org/10.1016/j.ijforecast.2023.12.002

Wilson, G. T., 2016. Time Series Analysis: Forecasting and Control, 5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1. *Journal of Time Series Analysis*, *37*(5), 709–711. https://doi.org/10.1111/jtsa.12194