Enhanced Prediction of Intensive Care Unit Length of Stay using a Stack Ensemble of Machine Learning Models

Ashok Kumar Tella
Department of Computer Applications, National Institute of Technology Tiruchirappalli, Tamil Nadu, India
ashoknittphd@gmail.com (corresponding author)

S. R. Balasundaram
Department of Computer Applications, National Institute of Technology Tiruchirappalli, Tamil Nadu, India
blsundar@nitt.edu

Received: 14 September 2024 | Revised: 31 October 2024, 13 November 2024, and 18 November 2024 | Accepted: 20 November 2024

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | DOI: https://doi.org/10.48084/etasr.8994

ABSTRACT

The Length of Stay (LoS) refers to the time between a patient's hospital admission and discharge. LoS is considered to increase as the complexity of the disease increases. A prolonged stay in the Intensive Care Unit (ICU) can consume clinical resources and be labor intensive. Models that correctly predict LoS are needed to help medical experts make better decisions. To define an ideal process system, healthcare models must consider the patient's condition, availability of beds, resources, etc. These predictions can also help insurance companies manage their budgets. Existing models deploy machines and deep learning techniques to predict LoS. However, there is a need for improvement, considering the features associated with the process. This study presents machine learning algorithms, such as SVM and a stack ensemble, with improved accuracy over existing models. Experiments were carried out on a benchmark dataset, MIMIC-III, specific to ICU patients. The SVM model achieved an accuracy of 93.88%, while the stack ensemble model showed an improved accuracy of 94.70%. The results show that combining machine learning models achieves better prediction rates, which helps healthcare professionals make better decisions.

Keywords-length of stay; intensive care unit; machine learning models; stack ensemble models; healthcare analytics; ICU patient management

I. INTRODUCTION

The Length of Stay (LoS) is an important hospital criterion, according to the WHO. Hospitals tend to diagnose patients quickly and allocate resources efficiently. Beds are a key hospital resource [1]. Therefore, hospitals must always have beds available to diagnose patients quickly. LoS is the main hospital performance indicator and its prediction helps ICU physicians and nurses prescribe medications and assess patient health [2]. Since ICUs have more medical staff, analyzing the physiological status of patients is crucial for resource allocation. Furthermore, the safety of patient treatment should be improved [3]. Accurate LoS prediction is an important approach to managing personnel, resources, and capacity, as it substantially optimizes resource usage and reduces healthcare costs [4]. Hospitals work hard to reduce LoS and readmissions [5]. To increase accuracy, machine learning (ML) models employ demographics, medical history, test results, and periodic vitals of ICU patients. This study examined SVM, RF, XGBoost, KNN, and a stack ensemble for ICU LoS prediction. Previous studies have shown that XGBoost_ improves prediction on structured data by reducing error rates at each stage using an ensemble of decision trees. Multiple base models in an ensemble improve forecast accuracy. This study advances ML in critical care by stressing the importance of applying ML algorithms and optimization methods to address complex prediction issues to help healthcare professionals make data-driven decisions that improve patient care and ICU efficiency. The primary objectives are to create an improved ICU LoS predictive model using a stacked ensemble, combining Random Forest (RF), SVM, and KNN. The proposed model was evaluated on the MIMIC-III dataset and compared with other ML approaches.

II. RELATED WORKS

Many different ML models have been used in this vital healthcare sector. In [6], Decision Trees (DT), RF, and Gradient Boosting (GB) methods were examined. Model hyperparameters and feature selection increase prediction accuracy. These models predicted LOS accurately. This study showed that feature engineering and data preprocessing improve LoS prediction. In [7], clinical statistics and the efficacy of ML prediction were examined. Some ML models reached good accuracy, depending on the dataset and technique. In [8], model performance was improved using cross-validation and SVM, RF, and XGBoost. These models were accurate, with XGBoost being the best. To increase accuracy, model validation must continue across patient groups. In [9], ICU outcomes were examined using Logistic Regression (LR) and neural networks. This study used a multicenter dataset to train models for patient demographics and clinical factors. Patient characteristics and treatment methods vary by region, making it difficult to predict pandemic ICU outcomes. In [10], the utilization of ICU resources was predicted using COVID-19 pandemic patient data with RF and GB.

In [11], the necessity for sophisticated prediction models to account for acute physiological changes in critically ill individuals was emphasized. In [12], ICU readmissions were predicted using physiological data and drug intake using DT and SVM. The models achieved good accuracy, proving that aggregated trends work. These models require further development, although physiological and pharmacological data may enhance ICU readmission predictions. In [13], a deep learning program, called DeepSOFA, evaluated patient acuity and predicted outcomes, including LoS, in real-time using clinical data. DeepSOFA predicted patient outcomes with better accuracy than general SOFA. Deep learning can provide continuous, interpretable ICU predictions, but clinical integration is problematic. ICU LoS is affected by glycemic management, which was predicted with good accuracy in [14], implying that virtual trials can enhance ICU outcome prediction, such as LoS prediction, for various patient groups. In [15], a combined model of just-in-time learning and extreme learning machines was used to predict ICU mortality and LoS, achieving better accuracy and minimal computing cost. Although just-in-time learning may provide real-time predictions, this study recommended larger datasets for validation. In [16], ML models were used to analyze EEG data to detect deep sedation. The results show that EEG-based atomic decomposition detection could predict ICU sedation.

In [17], an AutoML mortality risk adjustment model was proposed, using clinical and demographic data. This model's better mortality risk adjustment accuracy suits critical care benchmarking and quality improvement. This study suggested that AutoML could save time and resources in healthcare. In [18], locally weighted PCA and ML methods were used to predict ICU outcomes. This prediction model can be used in real-time health monitoring and early warning systems. In [19], an ANN model was used to predict hospital LoS using admission clinical, demographic, and procedural data. This ANN model could enhance the cardiology department's planning and resource use by correctly forecasting hospital LoS. However, this study advocated exploring larger and more diverse patient groups. In [20], survival analysis and regression models were used to analyze LoS in patients who died during hospitalization. Demographics and clinical history greatly affect LoS [21]. Combining clinical and demographic data in ANNs makes it possible to reliably predict LoS, supporting the idea that ML models with different types of data can predict LoS. Ensemble models can provide better predictions in complicated datasets [22-23], supporting the idea that similar ensemble techniques might make LoS predictions more accurate in ICUs.

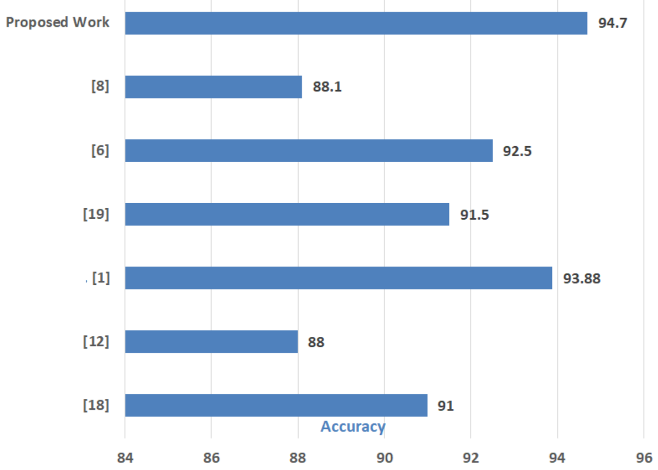

Fig. 1. Comparative analysis of the proposed and existing systems.

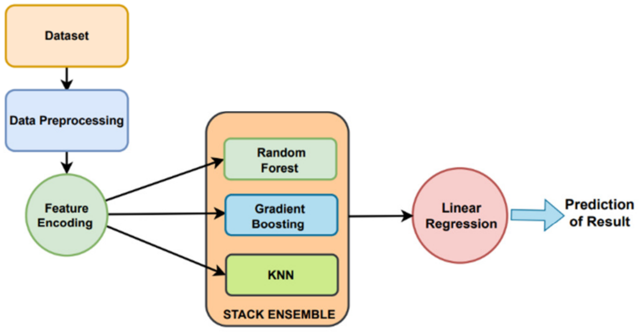

Fig. 2. Process flow for predicting LoS in the ICU.

The predictive models utilized in this research encompassed a wide range of ML techniques, such as DT, RF, and SVM. Several studies have further refined these models by implementing feature selection strategies and model optimization techniques to enhance predictive performance. The accuracy of these models varied significantly across studies, with reported values ranging from 70% to 92% (Figure 1). Although the models demonstrated considerable effectiveness, the complexity and heterogeneity of ICU data posed substantial challenges in achieving consistent performance. Limitations in previous studies include insufficient management of nonlinear interactions, restricted generalizability across patient demographics, overfitting and low interpretability in deep learning models, and increased computational demands. This study seeks to overcome these constraints by creating a stacked ensemble model that integrates many ML methods to capitalize on their strengths and improve prediction accuracy, generalizability, and clinical application.

III. METHOD

The steps followed involved importing the dataset, rectifying missing values, normalizing features, and partitioning it into training and testing sets. Subsequently, one-hot encoding or feature scaling converted categorical and numerical information into a format suitable for ML models. Training data are used to train different basic models, such as RF, GB, and KNN, to predict test data. The stacking ensemble approach combines base model predictions and trains a meta-model. The meta-model makes final test data predictions. Final predictions were compared with test labels to evaluate the accuracy of the ensemble model or other criteria.

IV. EXPERIMENTAL SETUP

The experimental setup for the proposed stacking ensemble prediction model involves the following steps.

A. Dataset

More than 40,000 patients in critical care were admitted to Beth Israel Deaconess Medical Centre (BIDMC) between 2001 and 2012. The Medical Information Mart for Intensive Care III (MIMIC-III) [24] is a large publicly available dataset that contains extensive clinical data on these patients. The MIT Lab for Computational Physiology, in collaboration with BIDMC, developed the MIMIC-III dataset to facilitate the research and development process in the field of healthcare analytics, particularly in settings that include critical care. The dataset includes several patient characteristics and results for analysis, such as:

- Patient demographics: Age, gender, ethnicity, etc.
- ICU admission details: Admission time, discharge time, type of ICU, etc.
- Vital signs: Heart rate, blood pressure, respiratory rate, etc.
- Laboratory results: Blood tests, urine tests, and others.
- Medications: Lists of drugs administered during ICU stay.
- Procedures: Surgical and non-surgical procedures performed.
- Outcomes: LoS, mortality, discharge status, etc.

B. Data Cleaning

Data cleaning involved:

- Handling missing values.
- Imputation: For numerical or categorical characteristics, the mean, median, or the most prevalent category was used for missing values.
- Flagging missing data: Binary features indicated whether a feature was missing.
- Outlier detection: Outliers were managed by normalizing their values to reduce distortion in model training. Additionally, to ensure that every feature contributes equally to the model, numerical characteristics were standardized or normalized to have a mean of 0 and a standard deviation of 1 or to a predetermined range (e.g., 0-1).

C. Feature Engineering

Feature engineering is a critical step in preparing a subset of the MIMIC-III dataset (a subset of the MIMIC databases) for predictive modeling and ML tasks. This process involves creating new features from raw data to improve the performance of models. The following steps were used to create and encode features.

1) Temporal Features

- LoS: Calculate the duration between ICU admission and discharge.
- Time since Admission: Calculate how long a patient has been in the ICU at any given time point.
- Time of Day/Day of Week: Create features indicating the time of day or the day of the week when events occur (e.g., admission, procedures) to capture patterns related to staffing or other time-dependent factors.

2) Aggregated Statistics

- Vital Signs: Mean, median, standard deviation, minimum, and maximum values of key vital signs (e.g., heart rate, blood pressure, respiratory rate) were calculated over specific time intervals.
- Laboratory Results: Aggregated statistics, such as mean, median, and variance, were calculated for lab test results (e.g., blood glucose levels, and white blood cell counts).
- Temporal Features: Aggregated values of temporal features, such as ICU stay duration and time since admission, were derived to capture trends over time

3) Binary Indicators

- Medication and Procedures: Binary flags indicating whether a patient received specific medications or underwent certain procedures during their ICU stay.
- Comorbidities: Binary indicators for the presence or absence of comorbid conditions based on diagnosis codes (e.g., diabetes, hypertension).
- Missing data: Flags were used to indicate whether a specific feature had missing values, ensuring the model accounted for incomplete data.
- Critical Events: Indicators capturing the occurrence of critical events (e.g., ventilator usage, dialysis).

D. Feature Encoding

1) Categorical Variables

- One-Hot Encoding: Categorical variables (e.g., ethnicity, type of ICU) were converted into binary columns.
- Label Encoding: If there was an ordinal relationship (e.g., severity levels), label encoding was used to map categories to integer values.

2) Textual Data

Natural Language Processing (NLP) was used to extract meaningful features from clinical notes, which were then incorporated into the dataset. Keywords related to diagnoses, treatments, and critical events were identified using TF-IDF and Named Entity Recognition (NER). Emotions were detected using sentiment analysis tools like VADER to capture sentiment polarity, and Latent Dirichlet Allocation (LDA) was employed for topic modeling to uncover underlying themes such as interventions or health conditions. These were added as binary indicators (presence of specific keywords or topics), numerical features (emotion intensity or topic probabilities), and categorical labels (dominant topics), enriching the dataset to improve model predictions.

E. Machine Learning Models

This study employed several machine-learning techniques to improve prediction accuracy, as seen in Figure 3. LR is a fundamental model that elucidates the relationship between factors and ICU LoS. RF is an ML technique that elucidates intricate relationships and connections by aggregating several DTs. GB and XGBoost enhance predictions by sequentially correcting errors, with XGBoost proving particularly adept at managing large datasets. The KNN method identifies patterns by juxtaposing patient data with analogous examples, providing significant support in local pattern recognition. This study employs a stacked ensemble of RF, SVM, and KNN, each selected for its ability to elucidate distinct aspects of ICU data patterns. The ensemble's results are then enhanced by a meta-model that consolidates the predictions from the basic models into a definitive prediction.

RF, SVM, and KNN were used to capture varied ICU data patterns. SVM establishes unambiguous decision limits, RF finds complicated, nonlinear associations using ensemble DTs, and KNN finds local patterns by evaluating patient commonalities. These models collaborate by addressing distinct data characteristics, allowing the ensemble to generalize across ICU situations. LR was used to build the metamodel using base model predictions. The meta-model may then weigh each base model's output based on dependability and combine their strengths to predict ICU LoS.

The experiments were carried out using a workstation with an Intel Core i7-9700K CPU, 32 GB RAM, and an NVIDIA GeForce RTX 2080 Ti GPU. Python 3.8 and scikit-learn 0.24 were used to develop the model. A 70:30 train-test split was implemented to provide an independent performance validation and fulfill the evaluation requirements.

Grid search hyperparameter tuning enhanced RF (estimators, depth), SVM (kernel, regularization), and KNN. This approach gave the stacked ensemble model optimum accuracy and generalizability for ICU LoS prediction.

V. RESULTS

This study focused on predicting how long a patient would stay in an ICU. The accuracy scores of each model, which were obtained from a stacked ensemble of ML models, were as follows. The 84.74% accuracy achieved using LR might serve as a standard for comparison. With an accuracy percentage of 93.88%, the RF algorithm performed quite well, as it faithfully depicted intricate relationships between variables and feature interactions. By repeatedly fixing prediction errors, gradient boosting was able to attain an accuracy of 93.18%. However, XGBoost, which is well-known for its efficacy with large datasets, obtained an accuracy of 93.63%. By using local pattern recognition, the K-Nearest Neighbors (KNN) method achieved an accuracy of 93.13%. The voting ensemble achieved a 94.05% accuracy rate by combining predictions from many models to increase dependability.

The most effective approach in the study, as seen in Figure 3 and Table I, was the stacking ensemble, which integrates the benefits of all models by training a meta-model on their outputs. This ensemble model attained the highest accuracy, 94.70%. This model improves generalization and resilience by combining RF, GB, and KNN base models. With this combination, the model can adjust to different ICU circumstances and data distributions without overfitting. The ensemble technique also balances performance, ensuring consistent predictions in diverse clinical scenarios.

TABLE I. PERFORMANCE METRICS FOR DIFFERENT ML ALGORITHMS TO PREDICT ICU LOS

ML model	MAE	MSE	R^2	Accuracy
LR	4.35278	74.3704	0.47615	0.84732
RF	1.5938	15.8184	0.88858	0.93888
GB	3.04147	63.3165	0.55401	0.93184
XGBoost	1.75283	18.0156	0.8731	0.93633
KNN	3.08019	65.954	0.53543	0.93133
Voting ensemble	3.29011	75.4314	0.51395	0.93059
Stacking ensemble	1.78449	19.4956	0.86268	0.94702

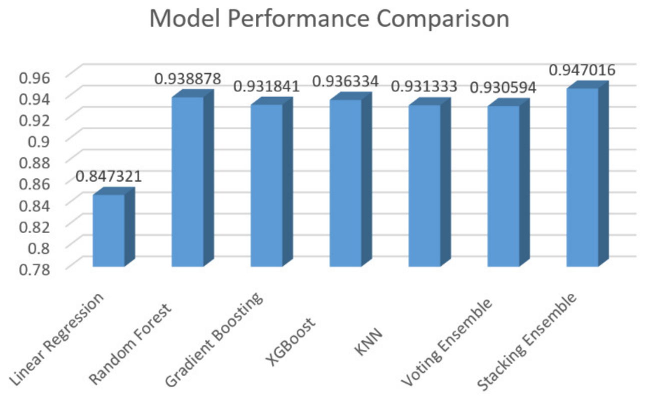

Fig. 3. Comparative analysis of different ML algorithms used to predict ICU LoS.

VI. CONCLUSION

This study presents a stacked ensemble model that uses ML techniques such as RF, GB, and KNN to provide more accurate predictions on ICU LoS. This model was evaluated using the extensive MIMIC-III dataset, comparing it to more conventional methods such as XGBoost, standalone RF, and LR. The proposed stacked ensemble model surpassed these models with an accuracy rate of 94.70%, demonstrating its efficacy as a reliable instrument for optimizing patient care and managing ICU resources. Over-fitting in uncommon patient scenarios, bias in the MIMIC-III dataset, and model complexity that impairs interpretability are some of the drawbacks. As these features might affect the generalizability on different datasets, external validation is required to ensure reliability across a range of clinical scenarios.

To fill the current knowledge gaps in LoS prediction, such as the difficulty of striking a balance between interpretability and model accuracy, this work uses ensemble learning to maximize the benefits of individual methods while reducing their drawbacks. The design of the stacked ensemble offers a computationally efficient, accurate, and scalable model that is better at identifying intricate patterns in ICU data than standalone techniques. Future improvements may involve real-time patient data for dynamic predictions, integrate explainable AI to improve model interpretability, and validate the model on a variety of datasets for wider application. Concentrating on these improvements may increase the model's usefulness in optimizing patient care and ICU resource planning.

REFERENCES

[1] L. Y. Sun, A. Bader Eddeen, M. Ruel, E. MacPhee, and T. G. Mesana, "Derivation and Validation of a Clinical Model to Predict Intensive Care Unit Length of Stay After Cardiac Surgery," *Journal of the American Heart Association*, vol. 9, no. 21, Nov. 2020, Art. no. e017847, https://doi.org/10.1161/JAHA.120.017847.

[2] N. B. Medeiros, F. S. Fogliatto, M. K. Rocha, and G. L. Tortorella, "Forecasting the length-of-stay of pediatric patients in hospitals: a scoping review," *BMC Health Services Research*, vol. 21, no. 1, Sep. 2021, Art. no. 938, https://doi.org/10.1186/s12913-021-06912-4.

[3] B. Stocker, H. K. Weiss, N. Weingarten, K. Engelhardt, M. Engoren, and J. Posluszny, "Predicting length of stay for trauma and emergency general surgery patients," *The American Journal of Surgery*, vol. 220, no. 3, pp. 757–764, Sep. 2020, https://doi.org/10.1016/j.amjsurg.2020.01.055.

[4] S. Bacchi, Y. Tan, L. Oakden-Rayner, J. Jannes, T. Kleinig, and S. Koblar, "Machine learning in the prediction of medical inpatient length of stay," *Internal Medicine Journal*, vol. 52, no. 2, 2022, Art. no. 176–185, https://doi.org/10.1111/imj.14962.

[5] C. Neto, M. Brito, H. Peixoto, V. Lopes, A. Abelha, and J. Machado, "Prediction of Length of Stay for Stroke Patients Using Artificial Neural Networks," in *Trends and Innovations in Information Systems and Technologies*, Budva, Montenegro, 2020, pp. 212–221, https://doi.org/10.1007/978-3-030-45688-7_22.

[6] L. Hempel, S. Sadeghi, and T. Kirsten, "Prediction of Intensive Care Unit Length of Stay in the MIMIC-IV Dataset," *Applied Sciences*, vol. 13, no. 12, Jan. 2023, Art. no. 6930, https://doi.org/10.3390/app13126930.

[7] K. Stone, R. Zwiggelaar, P. Jones, and N. M. Parthaláin, "A systematic review of the prediction of hospital length of stay: Towards a unified framework," *PLOS Digital Health*, vol. 1, no. 4, 2022, Art. no. e0000017, https://doi.org/10.1371/journal.pdig.0000017.

[8] S. Iwase *et al.*, "Prediction algorithm for ICU mortality and length of stay using machine learning," *Scientific Reports*, vol. 12, no. 1, Jul. 2022, Art. no. 12912, https://doi.org/10.1038/s41598-022-17091-5.

[9] H. Magunia *et al.*, "Machine learning identifies ICU outcome predictors in a multicenter COVID-19 cohort," *Critical Care*, vol. 25, no. 1, Aug. 2021, Art. no. 295, https://doi.org/10.1186/s13054-021-03720-4.

[10] S. S. Lorenzen *et al.*, "Using machine learning for predicting intensive care unit resource use during the COVID-19 pandemic in Denmark," *Scientific Reports*, vol. 11, no. 1, Sep. 2021, Art. no. 18959, https://doi.org/10.1038/s41598-021-98617-1.

[11] S. K. Andersen, C. L. Montgomery, and S. M. Bagshaw, "Early mortality in critical illness – A descriptive analysis of patients who died within 24 hours of ICU admission," *Journal of Critical Care*, vol. 60, pp. 279–284, Dec. 2020, https://doi.org/10.1016/j.jcrc.2020.08.024.

[12] Y. Xue, D. Klabjan, and Y. Luo, "Predicting ICU readmission using grouped physiological and medication trends," *Artificial Intelligence in Medicine*, vol. 95, pp. 27–37, Apr. 2019, https://doi.org/10.1016/j.artmed.2018.08.004.

[13] B. Shickel, T. J. Loftus, L. Adhikari, T. Ozrazgat-Baslanti, A. Bihorac, and P. Rashidi, "DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning," *Scientific Reports*, vol. 9, no. 1, Feb. 2019, Art. no. 1879, https://doi.org/10.1038/s41598-019-38491-0.

[14] J. L. Dickson *et al.*, "Generalisability of a Virtual Trials Method for Glycaemic Control in Intensive Care," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 7, pp. 1543–1553, Jul. 2018, https://doi.org/10.1109/TBME.2017.2686432.

[15] Y. Ding, X. Li, and Y. Wang, "Mortality prediction for ICU patients using just-in-time learning and extreme learning machine," in *2016 12th World Congress on Intelligent Control and Automation (WCICA)*, Guilin, China, Jun. 2016, pp. 939–944, https://doi.org/10.1109/WCICA.2016.7578592.

[16] S. B. Nagaraj *et al.*, "Electroencephalogram Based Detection of Deep Sedation in ICU Patients Using Atomic Decomposition," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 12, pp. 2684–2691, Sep. 2018, https://doi.org/10.1109/TBME.2018.2813265.

[17] R. J. Delahanty, D. Kaufman, and S. S. Jones, "Development and Evaluation of an Automated Machine Learning Algorithm for In-Hospital Mortality Risk Adjustment Among Critical Care Patients*," *Critical Care Medicine*, vol. 46, no. 6, Jun. 2018, Art. no. e481, https://doi.org/10.1097/CCM.0000000000003011.

[18] Y. Ding, X. Ma, and Y. Wang, "Health status monitoring for ICU patients based on locally weighted principal component analysis," *Computer Methods and Programs in Biomedicine*, vol. 156, pp. 61–71, Mar. 2018, https://doi.org/10.1016/j.cmpb.2017.12.019.

[19] P. F. Tsai *et al.*, "Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network," *Journal of Healthcare Engineering*, vol. 2016, no. 1, 2016, Art. no. 7035463, https://doi.org/10.1155/2016/7035463.

[20] A. Lim and P. Tongkumchum, "Methods for analyzing hospital length of stay with application to inpatients dying in Southern Thailand," *Global Journal of Health Science*, vol. 1, no. 1, 2009, Art. no. 27.

[21] D. A. Huntley, D. W. Cho, J. Christman, and J. G. Csernansky, "Predicting Length of Stay in an Acute Psychiatric Hospital," *Psychiatric Services*, vol. 49, no. 8, pp. 1049–1053, Aug. 1998, https://doi.org/10.1176/ps.49.8.1049.

[22] P. Manorom, U. Detthamrong, and W. Chansanam, "Comparative Assessment of Fraudulent Financial Transactions using the Machine Learning Algorithms Decision Tree, Logistic Regression, Naïve Bayes, K-Nearest Neighbor, and Random Forest," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15676–15680, Aug. 2024, https://doi.org/10.48084/etasr.7774.

[23] M. Ivanova, V. Tsenev, and V. Mikova, "An Approach for the Evaluation of a Measurement System: A Study on the Use of Machine Learning and Predictions," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12342–12347, Dec. 2023, https://doi.org/10.48084/etasr.6450.

[24] A. Johnson, T. Pollard, and R. Mark, "MIMIC-III Clinical Database." PhysioNet, 2015, https://doi.org/10.13026/C2XW26.