IJETIR

ISSN (Online) 2583-0554

International Journal of Emerging Technologies and Innovative Research (IJETIR)

Volume 5, Issue 5, May 2025

Impact Factor: 5.731

# Predicting Hospital Stay Length Using KNN Regressor Optimized with Gridsearchcv: An Explainable Machine Learning Approach

**K. Vigneshwar[1], A. Niranjan Reddy[2], B. Pallavi[3], V. Sravika[3],G. Vishwa Teja[4]**

Assistant Professor, Department of Computer Science & Engineering[1]

Students, Department of Computer Science & Engineering[2,3,4]

Guru Nanak Institute of Technology, India

**Abstract:** *Efficient bed management is essential for minimizing hospital costs, improving efficiency, and enhancing patient outcomes. This study introduces a predictive framework for forecasting hospital-ICU length of stay (LOS) at admission, leveraging hospital EHR data. Unlike prior work, which heavily relied on advanced tree-based models, this research proposes a K-Nearest Neighbors (KNN) model with hyperparameter optimization using GridSearchCV for predicting ICU patients' LOS. The KNN model effectively classifies patients into short and long LOS categories by learning patterns in clinical information systems (CIS). To ensure robustness, we evaluated the model using various performance metrics, including Accuracy, AUC, Sensitivity, Specificity, F1-score, Precision, and Recall. The optimized KNN model demonstrated competitive predictive performance with improved interpretability compared to traditional complex models. Additionally, explainable artificial intelligence (xAI) techniques were incorporated to provide transparent insights into the decision-making process, further enhancing the trustworthiness of the predictions. This work highlights the potential of using machine learning models like KNN for reliable, interpretable, and efficient ICU LOS prediction, aiding hospitals in improving resource allocation and patient care outcomes.*

**Keywords:** *hospital*

## I. INTRODUCTION

Predicting ICU length of stay (LOS) is vital for efficient healthcare management, influencing resource use, costs, and patient outcomes. Accurate LOS prediction helps hospitals optimize bed occupancy, streamline admissions, and maintain care quality. Electronic health records (EHR) now provide rich clinical and demographic data for predictive modeling. While tree-based models like random forests and gradient boosting machines are widely used for their accuracy, they often lack interpretability. This project proposes a K-Nearest Neighbors (KNN) regression model as a more interpretable alternative. KNN is non-parametric, flexible, and suited for both classification and regression tasks. The model is optimized using GridSearchCV to enhance performance metrics such as accuracy, sensitivity, specificity, precision, and F1-score. The study classifies ICU patients into short or long LOS groups using clinical and demographic features. It also integrates explainable AI (xAI) techniques for transparent decision-making. This ensures that healthcare professionals understand key prediction drivers. The framework balances accuracy and interpretability, offering a practical solution for ICU LOS prediction.

## II. LITERATURE REVIEW

B. Alsinglawi, O. Alshari, M. Alorjani, O. Mubin, F. Alnajjar, M. Novoa, and O. Darwish dicussed that the machine learning-based framework for predicting hospital length of stay (LOS) for lung cancer patients. The study aims to support hospital resource management and improve patient care by offering accurate and interpretable predictions. The framework utilizes advanced machine learning models to analyze EHRs and identify patterns influencing LOS. A key focus of the research is explainability, ensuring that healthcare professionals can understand the factors driving the

predictions. This is achieved by integrating explainable artificial intelligence techniques, which provide insights into the relationships between input features (such as patient demographics, clinical data, and treatment details) and the predicted LOS. [1]..

B. Alsinglawi, titled "Predictive Analytics Framework for Electronic Health Records with Machine Learning Advancements: Optimizing Hospital Resources Utilization with Predictive and Epidemiological Models", presents a comprehensive framework aimed at enhancing the utilization of hospital resources through advanced predictive analytics. Conducted at Western Sydney University, the research explores how electronic health records (EHRs) can be leveraged using machine learning techniques to predict patient outcomes, optimize resource allocation, and improve healthcare delivery. The study delves into the integration of predictive models and epidemiological analysis to address critical challenges in hospital resource management, such as bed allocation, staff scheduling, and patient flow optimization.[2].

The online resource by J. Brownlee, titled "How to Report Classifier Performance With Confidence Intervals", provides practical guidance on effectively evaluating and presenting the performance of classification models in machine learning. The article emphasizes the importance of reporting classifier performance metrics, such as accuracy, precision, recall, and F1-score, along with confidence intervals to provide a more robust and reliable assessment of model effectiveness. Brownlee explains that while metrics like accuracy provide a single point estimate, they do not account for variability or uncertainty inherent in the data. Confidence intervals address this limitation by quantifying the range within which the true performance of the classifier is likely to fall, given a specific confidence level (e.g., 95%).[3].

A. Gupta, T. Liu, and S. Shepherd, published in the Health Informatics Journal in June 2020, presents a novel approach to early detection and risk assessment of sepsis in healthcare settings. The study leverages electronic medical records (EMRs) and advanced probabilistic modeling to develop a clinical decision support system (CDSS). The proposed system is built using Tree Augmented Bayesian Networks (TABNs), a graphical probabilistic model that captures dependencies among clinical variables while maintaining computational efficiency. TABNs enable the integration of various patient-specific data, such as vital signs, lab results, and demographic information, to estimate the risk of sepsis in real-time. [4].

## III. METHODOLOGY

This project predicts ICU Length of Stay (LOS) using a K-Nearest Neighbors (KNN) model optimized with GridSearchCV. It leverages hospital EHR data to classify patients into short or long ICU stays based on historical trends. KNN works by finding similar past cases and predicting outcomes based on those neighbors. GridSearchCV fine-tunes parameters like neighbor count, distance metrics, and weighting to boost model accuracy. This enhances the model's adaptability to varied patient data. The approach is interpretable, offering transparency that aids clinical trust and informed decision-making. xAI tools further clarify prediction reasoning for healthcare professionals. Unlike complex models like XGBoost, KNN remains understandable and efficient. It supports real-time use in hospitals, helping manage ICU beds and resources effectively.

**Disadvantages of existing system:**

- Lack of Interpretability
- High Computational Cost
- Overfitting Risk
- Dependency on Hyperparameter Tuning

## IV. PROPOSED SYSTEM

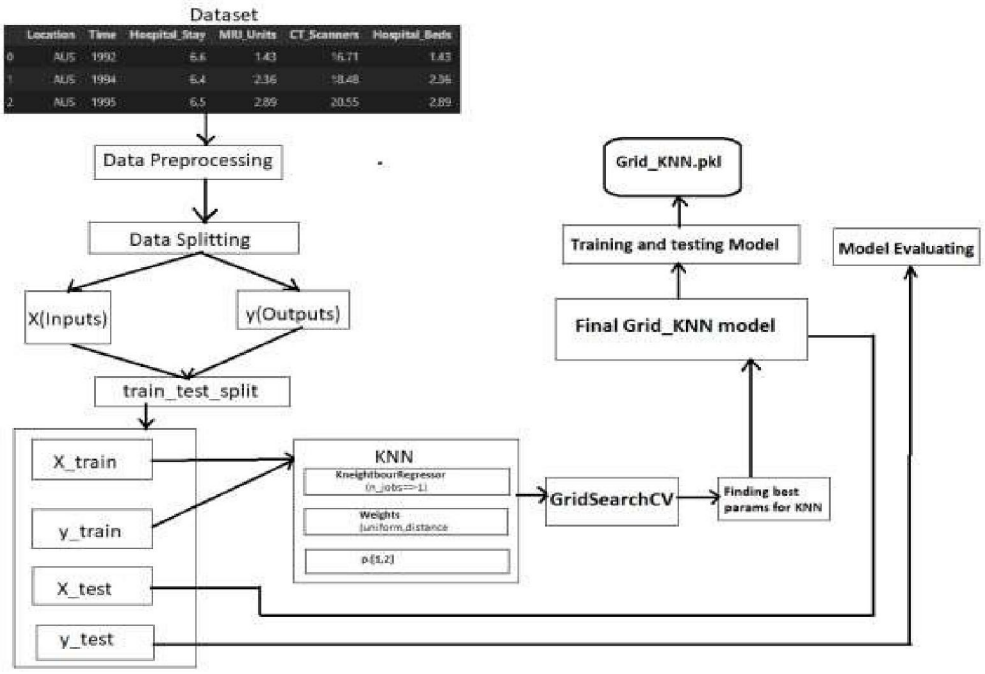
The proposed ICU length of stay (LOS) prediction system uses the K-Nearest Neighbors (KNN) algorithm, optimized with GridSearchCV for hyperparameter tuning. This approach offers a simpler, more interpretable alternative to complex models like XGBoost. KNN classifies patients based on the proximity of their features to those of previously observed patients, making predictions more understandable for healthcare professionals.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 5, May 2025
Impact Factor: 5.731

Advantages Proposed System Advantages

- Simplicity
- Interpretability
- Efficient Hyperparameter Tuning
- Transparency

SYSTEM ARCHITECTURE




This system architecture illustrates the workflow for predicting hospital stay length using a KNN regressor. It involves preprocessing the dataset, splitting data, applying GridSearchCV to optimize KNN parameters, and evaluating the final trained model.

**MODULES:**

**1) Data Collection:**The first step in the project involves collecting Electronic Health Records (EHR) from hospitals, which include critical clinical information such as patient demographics, medical history, test results, and previous admissions. This data is crucial for accurately predicting ICU Length of Stay (LOS).

**2) Data Preprocessing:**After data collection, the next step is preprocessing, which involves preparing the raw data for analysis. This stage includes data cleaning to address missing values, remove duplicates, and correct any inconsistencies in the dataset. For missing data, techniques such as mean imputation or median imputation are used to fill gaps, ensuring that no valuable information is discarded. Feature engineering follows, where relevant features are selected or derived from the original dataset.

**3) Model Selection and Training:**With the dataset prepared, the next stage involves selecting the right machine learning model for the task. The proposed model for this project is K-Nearest Neighbors (KNN), a non-parametric and simple algorithm that works well for classification tasks like predicting ICU LOS. KNN classifies patients by comparing their features with the closest neighbors in the training set.

**4) Model Saving and Serialization:**After the KNN model is trained and evaluated, it is important to save the model for future use, especially in a production environment. This is done by serializing the trained model using libraries like

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 5, May 2025
Impact Factor: 5.731

joblib or pickle. Serialization allows the model to be stored as a file, which can be reloaded later without the need to retrain the model from scratch.

**5) Model Deployment and Prediction via Flask:**The next step is deploying the saved KNN model into a Flask web application. Flask provides an easy-to-use framework for creating web-based interfaces, allowing healthcare professionals to interact with the system. The web application is designed to take input from users, such as new patient data (e.g., age, medical history, vital signs), and feed it to the saved model to generate predictions. When a user submits the data, the Flask app loads the saved model and performs the prediction in real-time.

**6) User Interface and Visualization:**The user interface is an essential component of the system, designed with ease of use in mind. The Flask app's front end is built using HTML, CSS, and JavaScript to ensure that it is visually appealing and user-friendly. It allows healthcare professionals to input patient data through a form, with fields for age, medical history, and other relevant information. Once the user submits the data, the Flask application triggers the model to generate the prediction.

## V. IMPLEMENTATION

**KNN using GridSearchCV:**

The proposed ICU Length of Stay (LOS) prediction system uses the K-Nearest Neighbors (KNN) algorithm, chosen for its simplicity and interpretability. KNN predicts a patient's LOS by comparing their data to similar past cases and classifying the stay as short or long based on neighbors' outcomes. To improve accuracy, GridSearchCV fine-tunes key hyperparameters like neighbor count, distance metric, and weighting scheme through cross-validation. This ensures the model generalizes well and avoids overfitting. The optimized KNN model is deployed via a Flask-based web application, offering a user-friendly interface for real-time predictions. This setup supports transparent, efficient decision-making in clinical settings.

## EXPERIMENTAL RESULTS

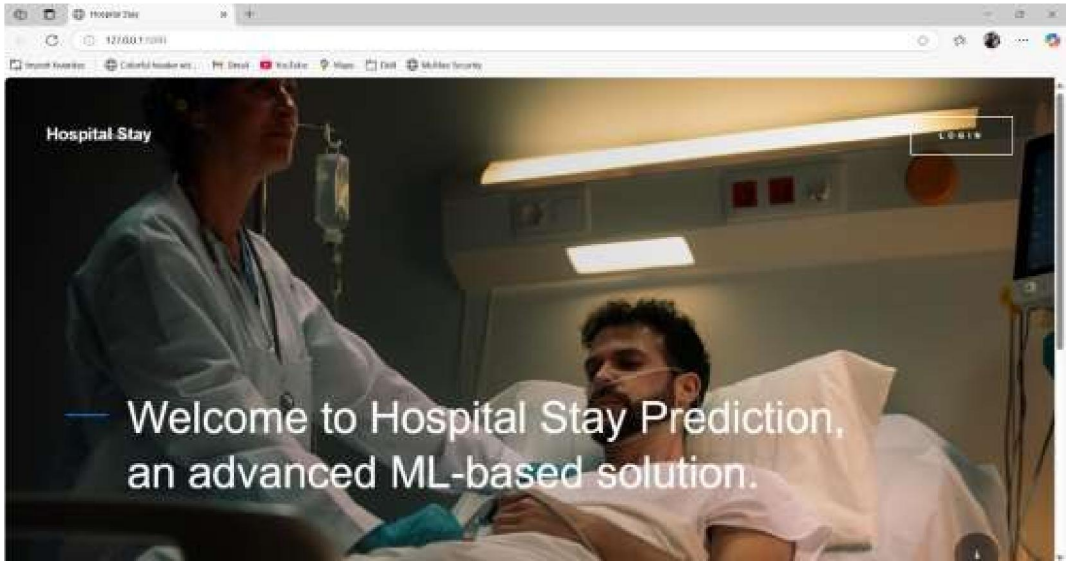



*Figure. Home Page*

The image shows a student homepage with the title "Welcome to Hospital Stay Prediction an advanced ML-based solution" prominently displayed.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 5, May 2025
Impact Factor: 5.731

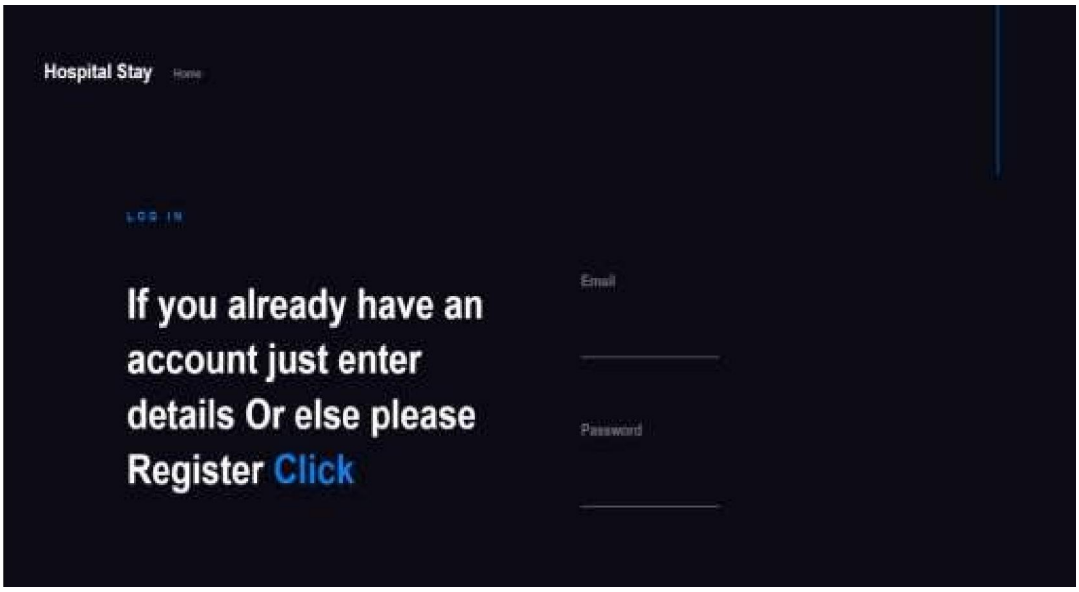



*Figure. Login Page*

The second image shows the login interface for the "Hospital Stay Prediction" web application.

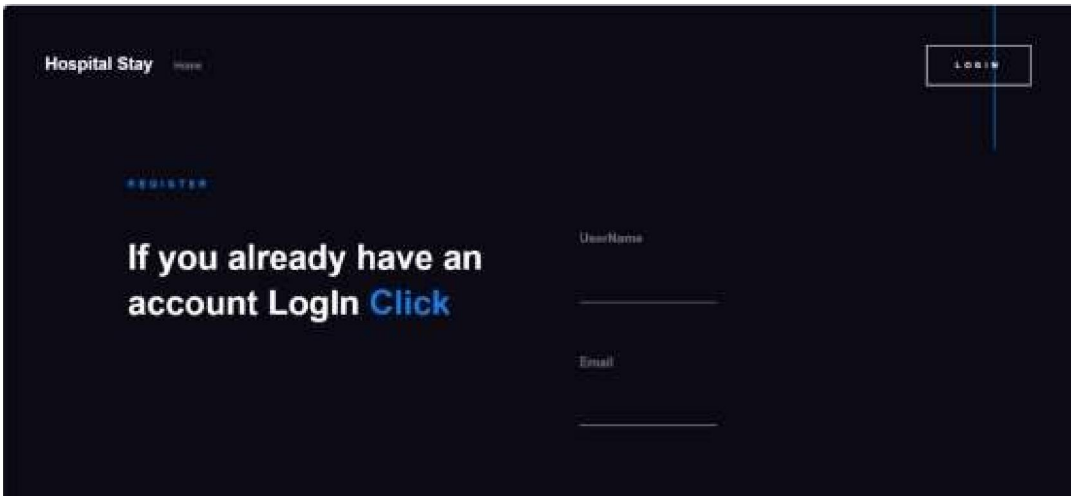



**Figure. Registration Page**

This image presents the registration page of the Hospital Stay Prediction web application, maintaining the same modern dark-themed interface as the login screen.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 5, May 2025
Impact Factor: 5.731


**Figure. About Page**

This is the "About" page of the Hospital Stay Prediction web application.

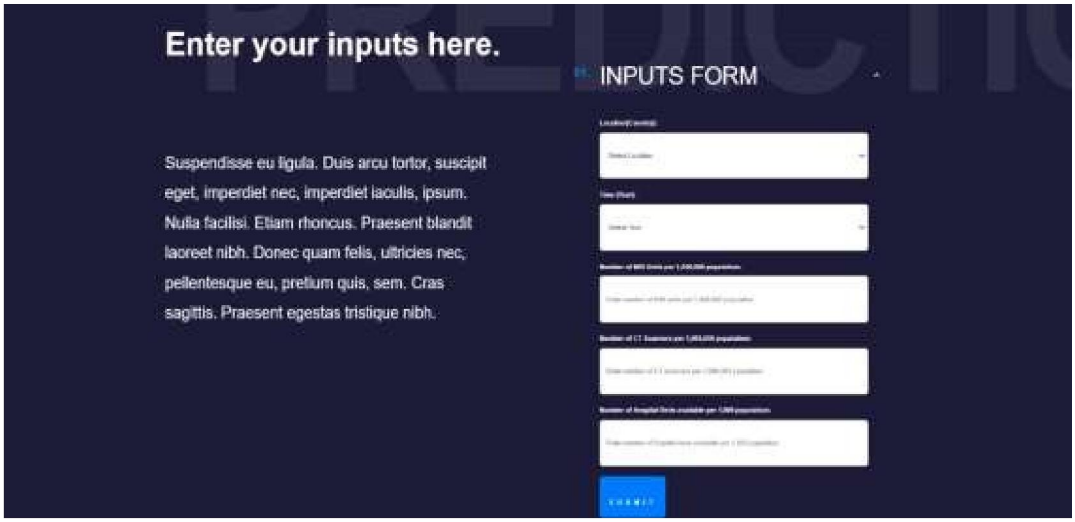
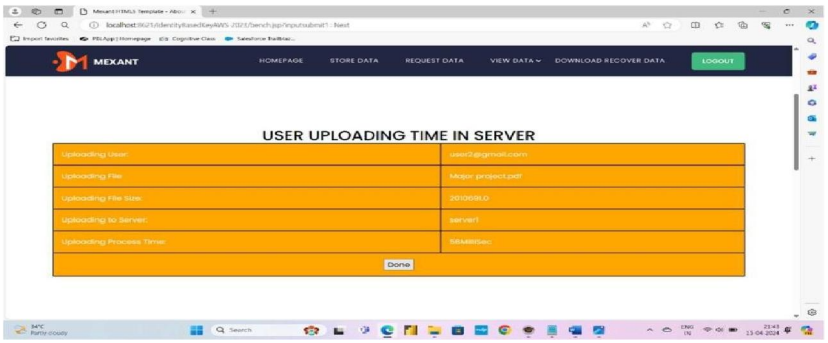



***Figure. Input Page***




Please enter the relevant healthcare data to predict hospital stay durations more accurately.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 5, May 2025
Impact Factor: 5.731


**Figure. Result Page**

To get an accurate prediction of hospital stay duration, please provide key healthcare data.

## VI. CONCLUSION

In conclusion, the proposed ICU Length of Stay (LOS) prediction system utilizing the K-Nearest Neighbors (KNN) algorithm optimized with GridSearchCV offers a promising solution for improving hospital resource management and patient care. By leveraging hospital Electronic Health Records (EHRs), this system classifies patients into short or long ICU stays based on patterns observed in previous patient data. The simplicity and interpretability of KNN, combined with the power of hyperparameter optimization, make the system both effective and transparent, ensuring that healthcare professionals can trust the model's predictions. The integration of Explainable AI (xAI) techniques further enhances the system's transparency, allowing clinicians to understand the rationale behind predictions and fostering confidence in the system's recommendations. The use of Flask for deployment ensures that the model is accessible through a user-friendly interface, facilitating real-time predictions and aiding healthcare professionals in making informed decisions about resource allocation.

## VII. FUTURE ENHANCEMENT

Future enhancements to the proposed ICU Length of Stay (LOS) prediction system could focus on improving its accuracy, scalability, and clinical applicability. One potential direction is the integration of more complex machine learning models, such as Random Forests or Neural Networks, which can capture more intricate patterns within the data. By combining the simplicity of KNN with the complexity of advanced algorithms, hybrid models could be developed to further enhance predictive performance while retaining interpretability.

## REFERENCES

**[1].** Awad, M. Bader-El-Den, and J. McNicholas, ''Patient length of stay and mortality prediction: A survey,'' Health Services Manage. Res., vol. 30, no. 2, pp. 105–120, May 2017.

**[2].** OECD. (2020). Length of Hospital Stay (Indicator). Accessed: Jul. 21, 2021. [Online]. Available: https://data.oecd.org/healthcare/lengthof-hospital-stay.htm

**[3].** Australian Institute of Health and Welfare, Canberra, ACT, Australia. (2011). Hospital Performance: Length of Stay in Public Hospitals in 2011–12. [Online]. Available: https://www.aihw.gov.au/reports/ hospitals/hospital-performance- length-of-stay-in-2011-12

**[4].** F. Pecoraro, F. Clemente, and D. Luzi, ‘‘The efficiency in the ordinary hospital bed management in Italy: An in-depth analysis of intensive care unit in the areas affected by COVID-19 before the outbreak,’’ PLoS ONE, vol. 15, no. 9, Sep. 2020, Art. no. e0239249.

**[5].** M. Hassan, H. P. Tuckman, R. H. Patrick, D. S. Kountz, and J. L. Kohn, ‘‘Hospital length of stay and probability of acquiring infection,’’ Int. J. Pharmaceutical Healthcare Marketing, vol. 4, no. 4, pp. 324–338, Nov. 2010.

**[6].** M. C. Blom, K. Erwander, L. Gustafsson, M. Landin-Olsson, F. Jonsson, and K. Ivarsson, ‘‘The probability of readmission within 30 days of hospital discharge is positively associated with inpatient bed occupancy at discharge—A retrospective cohort study,’’ BMC Emergency Med., vol. 15, no. 1, pp. 1–6, Dec. 2015.

**[7].** E. Rocheteau, P. Liò, and S. Hyland, ‘‘Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit,’’ 2020, arXiv:2007.09483.

**[8].** W. Hanson, C. S. Deutschman, H. L. Anderson, P. M. Reilly, E. C. Behringer, C. W. Schwab, and J. Price, ‘‘Effects of an organized critical care service on outcomes and resource utilization: A cohort study,’’ Crit. Care Med., vol. 27, no. 2, pp. 270–274, Feb. 1999.

**[9].** S. Siddiqui, S. Ahmed, and R. Manasia, ‘‘Apache II score as a predictor of length of stay and outcome in our ICUs,’’ J. Pakistan Med. Assoc., vol. 55, no. 6, p. 253, 2005.

**[10].** W. A. Knaus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, and D. E. Lawrence, ‘‘APACHE—Acute physiology and chronic health evaluation: A physiologically based classification system,’’ Crit. Care Med., vol. 9, no. 8, pp. 591–597, Aug. 1981.

**[11].** A Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, A. Damiano, and F. E. Harrell Jr., ‘‘The APACHE III prognostic system: Risk prediction of hospital mortality for critically III hospitalized adults,’’ Chest, vol. 100, no. 6, pp. 1619–1636, 1991.

**[12].** M. T. Keegan, O. Gajic, and B. Afessa, ‘‘Comparison of APACHE III, APACHE IV, SAPS3, and MPM0III and influence of resuscitationstatus on model performance,’’ Chest, vol. 142,no. 4, pp. 851–858, Oct. 2012.

**[13].** C.-C. Yeh, Y.-A. Chen, C.-C. Hsu, J.-H. Chen, W.-L. Chen, C.-C. Huang, and J.-Y. Chung,‘‘Quick-SOFA score ≥ 2 predicts prolonged hospital stay in geriatric patients with influenza infection,’’ Amer. J. Emergency Med., vol. 38, no. 4, pp. 780–784, Apr. 2020.

**[14].** Li, L. Chen, J. Feng, D. Wu, Z. Wang, J. Liu, and W. Xu, ‘‘Prediction of length of stay on the intensive care unit based on least absolute shrinkage and selection operator,’’ IEEE Access, vol. 7, pp. 110710–110721, 2019.

**[15].** M. M. Islam, T. N. Poly, and Y.-C. Li, ‘‘Recent advancement of clinical information systems: Opportunities and challenges,’’ Yearbook Med. Informat., vol. 27, no. 1, pp. 83–90,Aug. 2018.