

Original Article

In-hospital mortality predictors of stroke patients with diabetes mellitus

Mawaddah Ar Rochmah¹, Dhite Bayu Nugroho², Abdul Gofir¹, Muhammad Robikhul Ikhsan³, Faishal Hanif⁴, Atitya Fithri Khairani¹, Lukman Ade Chandra², Ismail Setyopranoto¹

Departments of ¹Neurology, ²Pharmacology, and ³Internal Medicine, Faculty of Medicine, Public Health, and Nursing, Universitas Gadjah Mada, Yogyakarta, ⁴Department of Neurology, Prof. Dr. Margono Soekarjo Hospital, Central Java, Indonesia.

ABSTRACT

Objectives: The incidence of stroke is higher among type 2 diabetes mellitus (T2DM) patients with a higher mortality rate. Prognostic scores for stroke patients can assist with treatment planning and counseling. The objective of this study was to create a machine-learning-based prognostic score to estimate in-hospital mortality in acute stroke with T2DM.

Materials and Methods: This study used data from claims-based diabetes registry at Dr. Sardjito General Hospital, Yogyakarta, Indonesia, to identify patients diagnosed with acute stroke and T2DM between January 2016 and December 2020. Four machine learning algorithms were trained and evaluated based on standard performance metrics. Important features were selected from the best-performing model and implemented in a web-based in-hospital mortality prediction scoring system.

Results: Of the 18,652 patients in the registry, the final analytic dataset comprised 749 patients (557 survivors and 192 non-survivors). The random forest showed superiority compared to other models. The six most important features were length of stay, sepsis, pneumonia, age, dyslipidemia, and hemiplegia. Using these features, the web-based system estimates the probability of in-hospital death for an individual patient.

Conclusion: Machine learning analysis may support an in-hospital mortality prediction score for patients with acute stroke and T2DM patients by leveraging the key features identified by the random forest model.

Keywords: Machine learning, Mortality predictor, Scoring, Stroke, Type 2 diabetes

INTRODUCTION

Globally, stroke is the leading contributor to disability and the second-leading contributor to mortality.^[1,2] In the INTERSTROKE case-control study across 32 countries, type 2 diabetes mellitus (T2DM) was among the ten risk factors associated with 90% of stroke cases.^[3] Similarly, a community-based survey in Sleman, Yogyakarta, Indonesia, found that T2DM, age, and hypertension showed a significant association with stroke.^[4] T2DM not only increases stroke risk but is also linked to higher in-hospital and long-term mortality, stroke recurrence, multiple comorbidities, and in-hospital complications across stroke types.^[5,6]

Prognostic scores can assist clinicians with treatment planning and counseling. For example, the stroke subtype, Oxford Community Stroke Project classification, age, and

pre-stroke modified Rankin Scale (mRS) score predict in-hospital and 7-day mortality in acute stroke.^[7,8] Another study identified total anterior circulation stroke, estimated glomerular filtration rate <15, 1-year increments in age, liver disease, 1-point increments in pre-stroke mRS, atrial fibrillation, coronary heart disease, chronic obstructive pulmonary disease, and hypertension as predictors of 10-year mortality.^[9] Severity indices such as the acute physiology and chronic health evaluation II (APACHE II) and the sequential organ failure assessment (SOFA) have also been used to predict stroke mortality using clinical and laboratory data.^[10,11] However, these scores may not generalize to all populations because of differences in racial/ethnic composition, hospital type, healthcare systems, and available treatments;^[12] and the required variables are not always available in routine practice.

*Corresponding author: Ismail Setyopranoto, Department of Neurology, Faculty of Medicine, Public Health, and Nursing, Universitas Gadjah Mada, Yogyakarta, Indonesia. ismail.setyopranoto@ugm.ac.id

Received: 15 March 2025 Accepted: 29 October 2025 Epub ahead of print: 06 February 2026 Published: 12 March 2026 DOI: 10.25259/JNRP_85_2025

This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-Share Alike 4.0 License, which allows others to remix, transform, and build upon the work non-commercially, as long as the author is credited and the new creations are licensed under the identical terms. ©2026 Published by Scientific Scholar on behalf of Journal of Neurosciences in Rural Practice

A systematic review of machine-learning methods for predicting stroke mortality showed that many studies used detailed clinical and laboratory variables as explanatory features.^[13] Most models addressed binary outcomes (survival vs. mortality) and commonly employed support vector machines, logistic regression, random forests, k-nearest neighbors (KNN), artificial neural networks, and eXtreme gradient boosting (XGBoost).^[13] Compared with conventional statistical methods, machine learning has demonstrated superior performance in predicting mortality in acute coronary syndrome and ischemic stroke,^[14,15] and in some cases, has outperformed traditional stroke prognostic scores.^[13] Nevertheless, reliance on granular clinical or laboratory data can limit implementation. Claims-based registries derived from electronic medical records offer a scalable alternative by capturing comorbidities and in-hospital complications without requiring detailed laboratory measurements.

Accordingly, we aimed to develop a machine-learning-based scoring system to predict in-hospital mortality among acute stroke patients with T2DM using demographic and clinical information, comorbidities, and in-hospital complications available from a claims-based registry. We also evaluated whether these routinely available variables could predict in-hospital mortality using real-world data as effectively as models that depend on detailed clinical and laboratory data.

MATERIALS AND METHODS

Data collection

This study used a claims-based diabetes registry data at Dr. Sardjito General Hospital, a tertiary hospital in Yogyakarta, Indonesia, covering patients diagnosed with acute stroke and T2DM between January 2016 and December 2020. The dataset consisted of demographic features, diagnosis information, and discharge status, including comorbidity data derived from International Classification of Diseases, 10th Revision (ICD-10) codes. The inclusion criteria specified hospitalized adult patients (18 years old or older) with either acute hemorrhagic or ischemic stroke and T2DM requiring treatment in a tertiary care hospital. The exclusion criteria ruled out patients without a confirmed T2DM diagnosis, those with sequelae of stroke or post-stroke diagnosis, incomplete data, or patients with any other type of diabetes. The data cleaning process involved categorization of comorbid diseases, data transformation, and the translation of ICD-10 codes into disease names.

Data pre-processing

Feature/variable selection was conducted based on clinical judgments and a descriptive analysis of the top 10 comorbid diseases in the stroke and T2DM population. The data

were divided at random into a training set (75%) and a testing set (25%). This split was performed using the initial split function from the *tidymodels* package in R,^[16] which allows for a random allocation of data and preserves the distribution of variables between the training and testing sets. The models were developed based on the training dataset and evaluated for performance using the testing dataset. This helped to ensure a balanced representation of the dataset in both training and testing phases. The case of missing data was addressed using multivariate imputation by the chained equation technique, where appropriate. This technique iteratively imputes each variable with missing values using predictive models built from the remaining variables. The cycle repeats until the imputations stabilize, preserving multivariable relationships under a missing-at-random assumption.

The tidy models framework in R was utilized for data pre-processing. This included a series of steps for feature selection, data transformation, and handling missing values. The same pre-processing steps were consistently applied to both the training and test data. A workflow was developed to combine the pre-processing recipe and model specification. This workflow incorporated a 10-fold cross-validation to yield a reliable performance estimate.

Machine learning model development and identification of important features

Four machine learning algorithms were implemented: Random Forest, Logistic Regression, XGBoost, and KNN. These were chosen based on their known performance and interpretability in similar prediction tasks.^[13] A 10-fold cross-validation resampling procedure was used to ensure robust model validation. This method trained and validated each algorithm multiple times on different subsets of the data to provide a comprehensive overview of performance.

The model was evaluated with standard metrics – accuracy, precision, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). Following the model evaluation, the best-performing model was selected and applied to the test data for an unbiased estimate of its generalizability. To identify a concise, clinically usable set of predictors, we derived model-based importance scores from the selected model for all candidate variables. These scores reflect each variable's relative contribution to distinguishing the binary outcome between survivors and non-survivors. Variables were ranked by their importance, and the top 10 variables or features were selected as key predictors. In addition, we used SHapley Additive exPlanations (SHAP) to identify the ten features with the greatest influence on the model's predictions. These features will be used to develop a scoring system to aid in patient outcome predictions.

Prediction scoring system user interface development

For practical use in the clinical setting, the predictive scoring system model was embedded into an interactive Shiny web application and then deployed online, allowing real-time predictions based on individual patient data. The application visualizes the SHAP values, providing a personalized interpretation of prediction results.

Validation with real-world data

A secondary dataset was used, derived from the stroke registry of Dr. Sardjito General Hospital in Yogyakarta, Indonesia, covering January to December 2024. This dataset included all patients diagnosed with acute stroke and T2DM to evaluate the performance of the in-hospital mortality prediction scoring system. The required variables for input into the scoring system were prepared based on patients' discharge data and aligned with the identified critical features. The prediction scores were collected and analyzed for comparison between patients who survived with those who died.

Ethical consideration

Approval for this study was obtained from the Ethics Committee, Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada, Ref. No.: KE/FK/0388/EC/2024.

RESULTS

Patients' selection and characteristics

The patient selection procedure for this study began with the initial extraction of data from the claims-based diabetes registry of the Dr. Sardjito General Hospital main database for the years 2016–2020. This data included 18,652 patients of all ages diagnosed with diabetes using either the primary or secondary ICD-10 codes, regardless of whether they were hospitalized or treated as outpatients. A series of inclusion and exclusion criteria was applied to the dataset after the initial data extraction. Specifically, only adults 18 years or older with T2DM who had any form of stroke (either hemorrhagic or ischemic) requiring hospitalization were included. The goal of focusing on inpatient care was to target patients with potentially severe conditions requiring hospitalization. Following this thorough selection procedure, a final dataset comprised 749 patients [Figure 1].

The characteristics and comorbidities of patients who were discharged as survivors ($n = 557$) and deceased/non-survivors ($n = 192$) among the total number of participants in the study ($n = 749$) are presented in Table 1. Overall, patients had a median length of stay of 8 (interquartile range [IQR] 5, 13) days. The average length of stay was significantly shorter in non-survivors compared to survivors. The overall

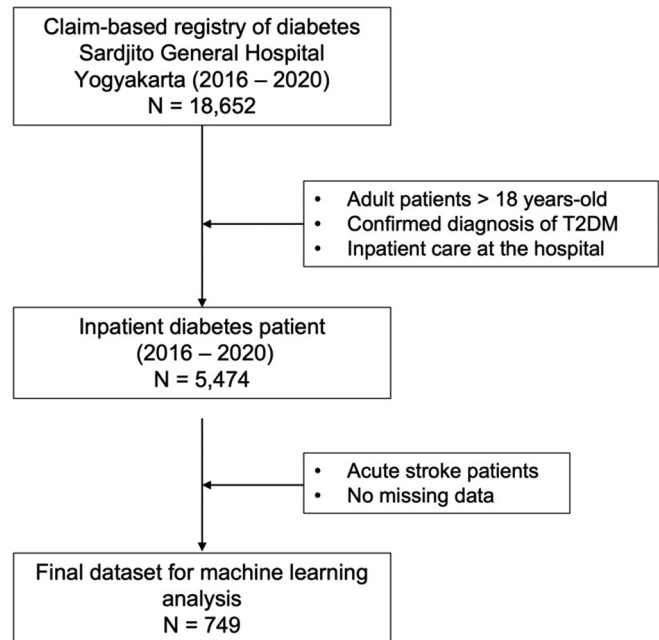


Figure 1: Schematic workflow of the patient selection process. T2DM: Type 2 diabetes mellitus.

mean age of the subjects was 62 (standard deviation ± 11) years old. The proportion of male was similar in survivors and non-survivors. The majority of patients were diagnosed with hypertension (62%). Hypertension, sepsis, kidney failure, pneumonia, anemia, hypoalbuminemia, leukocytosis, dyslipidemia, and coronary artery disease were significantly different between survivors and non-survivors using the bivariate analysis.

Machine learning models and their performance metrics

After training each machine learning algorithm, we assessed its performance using the test set and standard metrics. Figure 2 illustrates the mean accuracy, precision, sensitivity, and specificity of the four predictive models. Overall, the random forest model demonstrated superior performance compared to the other evaluated algorithms, particularly in terms of sensitivity, precision, and accuracy. These results demonstrate the model's robust ability to correctly identify patients who were deceased on hospital discharge and its overall accuracy of prediction.

As presented in Figure 3, the AUC-ROC values for Logistic Regression, Random Forest, KNN, and XGBoost values were 0.821, 0.884, 0.756, and 0.874. Primarily based on the AUC-ROC metric, the random forest model emerged as the top-performing model. The AUC-ROC is commonly used for binary classification because it integrates the sensitivity-specificity trade-off across thresholds. The random forest model demonstrated a robust balance between sensitivity and specificity.

Table 1: Patients' characteristics.

| Characteristic | Overall n=749 ¹ (%) | Survive on hospital discharge n=557 ¹ (%) | Non-survivor on hospital discharge, n=192 ¹ (%) | p ² |
|-------------------------|-----------------------------------|---|---|----------------|
| Hypertension | | | | 0.003 |
| No | 283 (38) | 193 (35) | 90 (47) | |
| Yes | 466 (62) | 364 (65) | 102 (53) | |
| Sepsis | | | | <0.001 |
| No | 655 (87) | 539 (97) | 116 (60) | |
| Yes | 94 (13) | 18 (3) | 76 (40) | |
| Kidney failure | | | | <0.001 |
| No | 574 (77) | 457 (82) | 117 (61) | |
| Yes | 175 (23) | 100 (18) | 75 (39) | |
| Pneumonia | | | | <0.001 |
| No | 543 (72) | 447 (80) | 96 (50) | |
| Yes | 206 (28) | 110 (20) | 96 (50) | |
| Urinary Tract Infection | | | | 0.3 |
| No | 610 (81) | 458 (82) | 152 (79) | |
| Yes | 139 (19) | 99 (18) | 40 (21) | |
| Anemia | | | | 0.003 |
| No | 621 (83) | 475 (85) | 146 (76) | |
| Yes | 128 (17) | 82 (15) | 46 (24) | |
| Hypoalbuminemia | | | | <0.001 |
| No | 524 (70) | 415 (75) | 109 (57) | |
| Yes | 225 (30) | 142 (25) | 83 (43) | |
| Leukocytosis | | | | 0.002 |
| No | 691 (92) | 524 (94) | 167 (87) | |
| Yes | 58 (8) | 33 (5.9) | 25 (13) | |
| Hypokalemia | | | | 0.6 |
| No | 583 (78) | 436 (78) | 147 (77) | |
| Yes | 166 (22) | 121 (22) | 45 (23) | |
| Hyponatremia | | | | 0.8 |
| No | 648 (87) | 481 (86) | 167 (87) | |
| Yes | 101 (13) | 76 (14) | 25 (13) | |
| Dyslipidemia | | | | <0.001 |
| No | 551 (74) | 379 (68) | 172 (90) | |
| Yes | 198 (26) | 178 (32) | 20 (10) | |
| Coronary artery disease | | | | 0.023 |
| No | 635 (85) | 482 (87) | 153 (80) | |
| Yes | 114 (15) | 75 (13) | 39 (20) | |
| Hemiplegia | | | | <0.001 |
| No | 538 (72) | 367 (66) | 171 (89) | |
| Yes | 211 (28) | 190 (34) | 21 (11) | |
| Age (years) | 62 (11) | 62 (11) | 63 (11) | 0.3 |
| Sex | | | | >0.9 |
| Male | 400 (53) | 297 (53) | 103 (54) | |
| Female | 349 (47) | 260 (47) | 89 (46) | |
| Length of stay (days) | 8 (5, 13) | 8 (6, 13) | 6 (3, 13) | <0.001 |

¹n (%); Mean (Standard deviation); Median (interquartile range). ²Pearson's Chi-squared test; Welch Two Sample t-test; Wilcoxon rank-sum test. P<0.05 is considered significant.

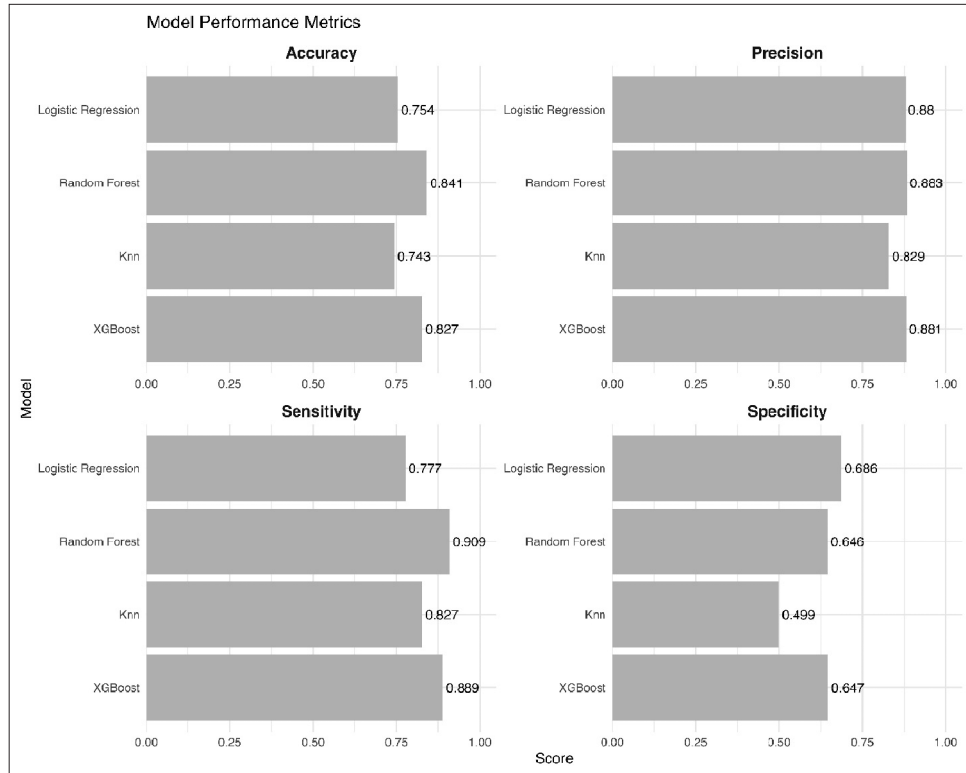


Figure 2: Bar diagram of model performance.

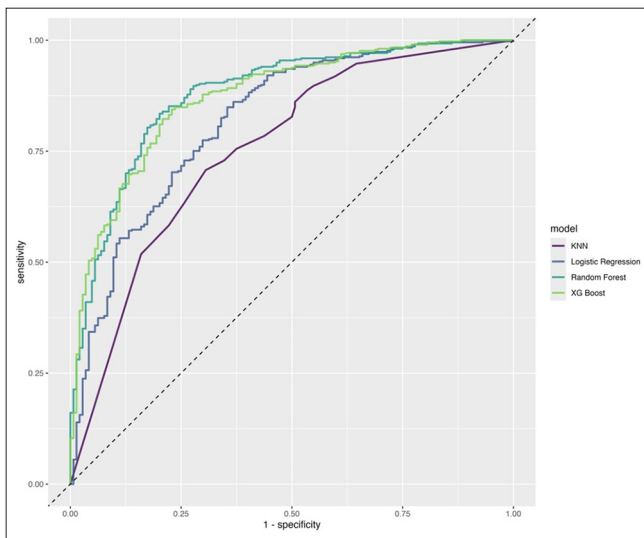


Figure 3: Receiver operating characteristic curve of model performance. KNN: K-nearest neighbors, XG Boost: eXtreme Gradient Boosting.

Important features from random forests model

The top ten characteristics, as determined by the Random Forests model, are displayed in Figure 4. Length of stay, sepsis, pneumonia, age, dyslipidemia, hemiplegia, anemia, urinary tract infection, kidney failure, and hypertension were the top ten most influential variables. The y-axis displays these

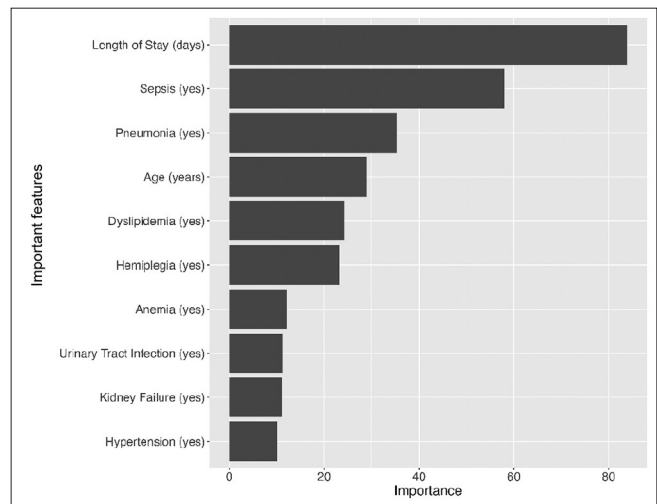


Figure 4: Feature importance in the random forest model.

variables, and the x-axis displays their importance scores. Length of stay emerged as the most significant factor for predicting the outcomes of patients with stroke and T2DM. Comorbidities such as sepsis, pneumonia, dyslipidemia, hemiplegia, anemia, urinary tract infection, kidney failure, and hypertension were also found to be important. Age was also an important factor in the prediction.

Web-based in-hospital mortality prediction tool

A web-based in-hospital mortality prediction tool, developed using the R Shiny framework and accessible at ugm.id/dmstroke, enables users to input predictor variables and generates predictions that are visualized as SHAP plots [Figure 5]. The tool includes a predictive model based on the random forest algorithm using variables selected for their clinical significance and the results of our variable importance analysis. The variables used by this model were those with an importance score >20%: Length of stay, sepsis, pneumonia, age, dyslipidemia, and hemiplegia. The variables of sepsis, pneumonia, dyslipidemia, and hemiplegia were inputted as nominal variables, choosing between yes/no. The variables of length of stay and age can be inputted as numeric variables by entering their respective representative values.

The SHAP plot is then generated to illustrate the contribution of each input variable to the predicted outcome. The influence of each variable on the prediction is quantified using SHAP values, which are a robust unified measure of feature importance. We used the SHAP method to interpret the random forest model's predictions because it accounts for complex interactions between variables and provides consistent attribution of each feature's contribution.

For illustration, we present four example cases of patients with acute stroke and T2DM, each with different clinical features during hospitalization. All cases had a confirmed acute stroke diagnosis based on clinical presentation and head computed

tomography scan. In case 1, Henry, a 59-year-old male, presented with hemiplegia and urinary tract infection during his 12-day hospital stay. In case 2, Bobby, a 67-year-old male, suffered from sepsis, pneumonia, kidney failure, hemiplegia, urinary tract infection, and hypertension on his second day of hospitalization. In case 3, Sussie, a 55-year-old female, presented with loss of consciousness and was diagnosed with sepsis, pneumonia, kidney failure, urinary tract infection, and hypertension by the fourth day of her hospitalization. In case 4, Annie, a 53-year-old female, had weakness of her left extremities on her 6th day of hospitalization. Each patient had unique SHAP plots, highlighting how their distinct characteristics could influence the model's prediction [Figure 5]. The green SHAP bars indicate the features that increased the probability of in-hospital mortality, meanwhile the red SHAP bars decrease it. Each feature affected the in-hospital mortality probability to varying degrees.

The in-hospital mortality probabilities predicted for Henry, Bobby, Sussie, and Annie were 0.032, 0.906, 0.972, and 0.01, respectively. Henry had a 0.032 probability, meaning a 3.2% predicted chance of in-hospital mortality. In his case, the positive contributors to mortality (shown in green) were the absence of dyslipidemia; whereas the negative contributors to mortality (shown in red) were the absence of sepsis, a 12-day hospitalization, the presence of hemiplegia, the absence of pneumonia, and age 59. As shown in Figure 5, in SHAP contributions for Bobby, Sussie, and Annie varied, reflecting the complex interaction between features.



Figure 5: Web-based prediction system and SHapley Additive exPlanations analysis.

Validation with real-world data

Between January and December 2024, a total of 451 patients with acute stroke were identified, comprising 323 cases of acute ischemic stroke and 128 cases of intracerebral hemorrhage. Among these, 121 patients (105 with acute ischemic stroke and 16 with intracerebral hemorrhage) were also diagnosed with T2DM and included in the validation of the scoring system.

The median probability scores for in-hospital mortality were 0.01 (IQR: 0.00–0.09) in the survivor group ($n = 105$) and 0.63 (IQR: 0.18–0.81) in the deceased group ($n = 16$), with the difference being statistically significant ($p < 0.05$) according to the Wilcoxon rank-sum test. In the subgroup analysis of acute ischemic stroke patients, the median probability scores for in-hospital mortality were 0.01 (IQR: 0.00–0.09) for survivors ($n = 94$) and 0.66 (IQR: 0.45–0.83) for those who died ($n = 11$), showing a statistically significant difference ($p < 0.05$) by the Wilcoxon rank-sum test. Similarly, in the subgroup analysis of intracerebral hemorrhage patients, the median probability scores for in-hospital mortality were 0.00 (IQR: 0.00–0.02) for survivors ($n = 11$) and 0.26 (IQR: 0.03–0.64) for those who died ($n = 5$), also demonstrating a statistically significant difference ($p < 0.05$) by the Wilcoxon rank-sum test. Figure 6 presents a visual comparison of the prediction score.

DISCUSSION

Supervised machine-learning approaches enable flexible prediction from heterogeneous clinical data and have been applied to stroke outcome prediction for more than a decade.^[13] Prior studies have used machine learning for 30-day mortality prediction in subarachnoid hemorrhage patients using random forest,^[17] for 10-day mortality model using an artificial neural network plus multivariable statistics,^[18] for post-stroke survival in post-stroke patients using Support Vector Machine,^[19] for mortality rate prediction in stroke patients using deep learning,^[20] for survival prediction in stroke patients using logistic regression,^[21] and for 1-year mortality prediction using XGBoost.^[22] Some of these studies also compared several machine learning models to identify the best prediction model for their datasets, an approach we also adopted for predicting in-hospital mortality among patients with acute stroke and T2DM.

We employed Logistic Regression, Random Forest, KNN, and XGBoost which represent various types of classifiers to identify the best model for mortality prediction. Logistic regression is a statistical classification method, originally designed for binary tasks, with model output a binary variable (for example: “yes” or “no”). KNN is a distance-based classifier method that measures similarity or dissimilarity between two observations in the provided dataset.^[23] In

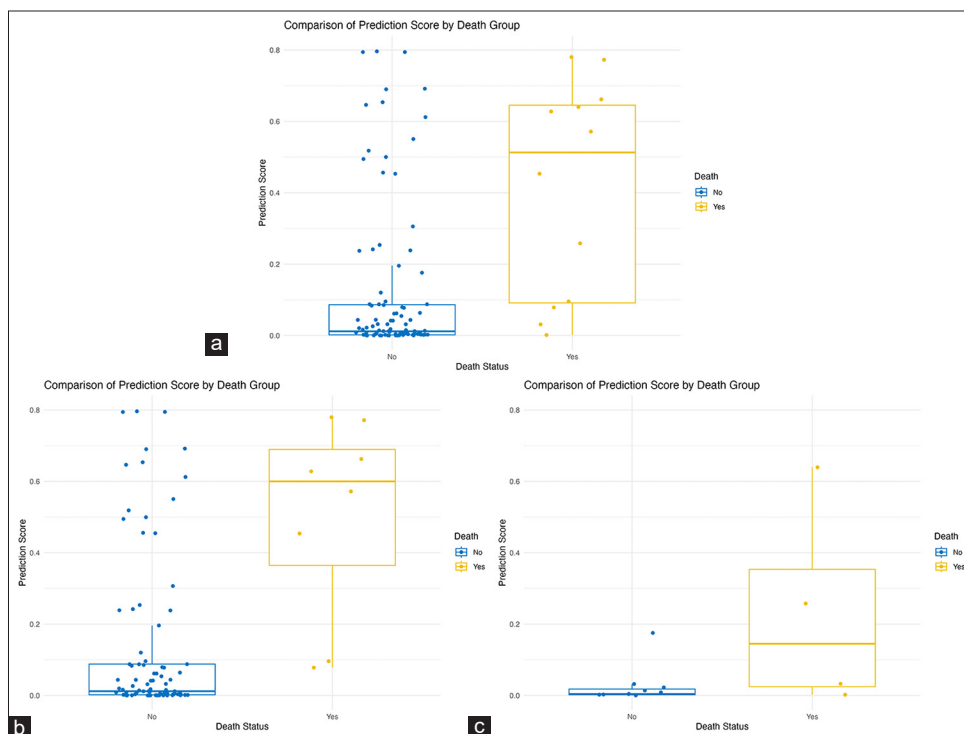


Figure 6: Comparison of in-hospital mortality prediction score in acute stroke and type 2 diabetes mellitus patients using secondary data. (a) All stroke patients, (b) Sub-analysis of acute ischemic stroke patients, (c) Sub-analysis of intracerebral hemorrhage patients.

a random forest, multiple decision trees are built and their votes are combined, typically by majority, to determine the class.^[24] XGBoost is a fast and scalable implementation of a gradient-boosting library with a linear and a decision-tree algorithm.^[25,26]

In this study, random forest outperformed the others in sensitivity, precision, accuracy, and AUC-ROC. The AUC-ROC curve of random forests was 0.884, indicating an acceptable level of discrimination.^[27] We used all features or parameters presented in Table 1 as the data input since our goal was an in-hospital mortality prediction scoring system that could be used throughout the dynamic course of hospitalization of acute stroke patients with T2DM. As a result, the random forests suggested length of stays, sepsis, pneumonia, age, dyslipidemia, hemiplegia, anemia, urinary tract infection, kidney failure, and hypertension as the top 10 important variables to predict the in-hospital mortality. Of the 10 variables, only the important variables that exceeded 20% importance values were employed in the prediction system. Our in-hospital mortality prediction model demonstrates strong discriminatory ability. Compared to previously reported mortality predictors in acute stroke, such as SOFA, APACHE II, and other machine learning studies utilizing clinical and laboratory data for features, our model achieves a similarly high level of discrimination.^[10,11,13] However, the currently available mortality predictors may not be interchangeable due to the differences in the data features used by each prognostic score. Therefore, providing a prognostic score with different data features could benefit clinicians and users in clinical practice, enabling them to tailor predictions based on the available data.

The validation using an independent dataset demonstrated that our prediction scoring system effectively distinguishes between survivors and non-survivors in acute stroke patients with T2DM. Moreover, the findings were supported by subgroup analyses, highlighting its applicability to both major types of stroke: Acute ischemic stroke and intracerebral hemorrhage. Notably, this validation relied solely on data collected at the time of patients' discharge. However, the prediction scoring system has the potential to be utilized continuously from admission to discharge, adapting to the dynamic changes in each patient's condition throughout their hospital stay.

This work illustrates the practical value of machine learning in routine data environments. A lightweight, web-based predictor can support clinical, managerial, and counseling needs; for example, facilitating goals-of-care discussions with patients and families and informing resource allocation – without requiring laboratory or imaging results, unlike most models reported in the previous studies.^[13,17-22]

Our study is limited by a restricted dataset comprising demographic, clinical, and comorbidity information extracted

from a claims-based diabetes registry at a tertiary hospital in Indonesia; detailed laboratory and imaging data were not captured. In addition, the real-world data validation relied solely on information collected at the time of patients' discharge. Nevertheless, the prediction scoring system is dynamic and can be utilized throughout the hospital stay, adapting to changes in patients' conditions as new input features become available. These data limitations underscore the need for further exploration and validation of the scoring system using real-world data across diverse geographical and clinical settings to ensure its reliability and broader applicability, as also noted by Abujaber *et al.*^[22] in their report.

Data availability statement

Data will be made available upon request.

CONCLUSION

Using a random forest model, our machine learning approach offers an in-hospital mortality scoring system for patients with acute stroke and T2DM in a tertiary hospital setting. The key features incorporated into the web-based in-hospital mortality scoring system were length of stay, sepsis, pneumonia, age, dyslipidemia, and hemiplegia. The web-based in-hospital mortality prediction scoring system can be utilized dynamically during a patient's hospitalization to account for changing conditions.

Acknowledgment: We thank all the staff in the Division of Cerebrovascular, Department of Neurology, and the Division of Endocrinology, Department of Internal Medicine, Faculty of Medicine, Public Health, and Nursing, Universitas Gadjah Mada, for the insightful discussion on the research topic.

Authors' Contributions: MAR: Writing – original draft, review and editing, methodology, data curation, conceptualization, funding acquisition; DBN: Methodology, data curation, formal analysis; AG: Writing – review and editing, conceptualization, supervision; MRI: Writing – review and editing, data curation, supervision; FH: Writing – original draft, methodology, formal analysis; AFK: Methodology, formal analysis; LAC: Methodology, formal analysis; IS: Writing – review and editing, conceptualization, supervision.

Ethical approval: The research/study was approved by the Institutional Review Board at the Ethical Committee of Faculty of Medicine, Public Health, and Nursing Universitas Gadjah Mada, approval number KE/FK/0388/EC/2024, dated 15th March 2024.

Declaration of patient consent: The authors certify that they have obtained all appropriate patient consent forms. In the form, the patients have given their consent for clinical information to be reported in the journal. The patients understand that their names and initials will not be published and due efforts will be made to conceal their identity, but anonymity cannot be guaranteed.

Financial support and sponsorship: Academic Excellence Scheme B Program of Universitas Gadjah Mada 2024 (No.6529/UN1.P1/PT.01.03/2024)

Conflicts of interest: There are no conflicts of interest.

Use of artificial intelligence (AI)-assisted technology for manuscript preparation: The authors confirm that there was no use of artificial intelligence (AI)-assisted technology for assisting in the writing or editing of the manuscript and no images were manipulated using AI.

REFERENCES

- World Health Organization. Noncommunicable diseases: Progress monitor 2020. World Health Organization; 2020. Available from: <https://iris.who.int/handle/10665/330805> [Last accessed on 2024 Dec 23].
- World Health Organization. Global health estimates 2019: Estimated deaths by age, sex, and cause. Geneva: World Health Organization; 2020. Available from: <https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/gh-leading-causes-of-death> [Last accessed on 2024 Dec 23].
- O'Donnell MJ, Xavier D, Liu L, Zhang H, Chin SL, Rao-Melacini P, *et al.* Risk factors for ischemic and intracerebral hemorrhagic stroke in 22 countries (the INTERSTROKE study): A case-control study. *Lancet* 2010;376:112-23.
- Setyopranoto I, Bayuangga HF, Panggabean AS, Alifaningdyah S, Lazuardi L, Dewi FS, *et al.* Prevalence of stroke and associated risk factors in Sleman district of Yogyakarta special region, Indonesia. *Stroke Res Treat* 2019;2019:2642458.
- De Miguel-Yanes JM, Jiménez-García R, Hernández-Barrera V, Méndez-Bailón M, De Miguel-Díez J, Lopez-de-Andrés A. Impact of type 2 diabetes mellitus on in-hospital-mortality after major cardiovascular events in Spain (2002-2014). *Cardiovasc Diabetol* 2017;16:126.
- Szlachetka WA, Pana TA, Tiamkao S, Clark AB, Kongbunkiat K, Sawanyawisuth K, *et al.* Impact of diabetes on complications, long term mortality and recurrence in 608,890 hospitalised patients with stroke. *Glob Heart* 2020;15:2.
- Myint PK, Clark AB, Kwok CS, Davis J, Durairaj R, Dixit AK, *et al.* The SOAR (stroke subtype, oxford community stroke project classification, age, prestroke modified rankin) score strongly predicts early outcomes in acute stroke. *Int J Stroke* 2014;9:278-83.
- Kwok CS, Potter JF, Dalton G, George A, Metcalf AK, Ngeh J, *et al.* The SOAR stroke score predicts inpatient and 7-day mortality in acute stroke. *Stroke* 2013;44:2010-2.
- Szlachetka WA, Pana TA, Mamas MA, Bettencourt-Silva JH, Metcalf AK, Potter JF, *et al.* Predicting 10-year stroke mortality: Development and validation of a nomogram. *Acta Neurol Belg* 2022;122:685-93.
- Moon BH, Park SK, Jang DK, Jang KS, Kim JT, Han YM. Use of APACHE II and SAPS II to predict mortality for hemorrhagic and ischemic stroke patients. *J Clin Neurosci* 2015;22:111-5.
- Qin W, Zhang X, Yang L, Li Y, Yang S, Li X, *et al.* Predictive value of the sequential organ failure assessment (SOFA) score for prognosis in patients with severe acute ischemic stroke: A retrospective study. *J Int Med Res* 2020;48:300060520950103.
- Matsumoto K, Nohara Y, Soejima H, Yonehara T, Nakashima N, Kamouchi M. Stroke prognostic scores and data-driven prediction of clinical outcomes after acute ischemic stroke. *Stroke* 2020;51:1477-83.
- Schwartz L, Anteby R, Klang E, Soffer S. Stroke mortality prediction using machine learning: Systematic review. *J Neurol Sci* 2023;444:120529.
- Hernesniemi JA, Mahdiani S, Tynkkynen JA, Lyytikäinen LP, Mishra PP, Lehtimäki T, *et al.* Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome - the MADDEC study. *Ann Med* 2019;51:156-63.
- Abedi V, Avula V, Razavi SM, Bavishi S, Chaudhary D, Shahjouei S, *et al.* Predicting short and long-term mortality after acute ischemic stroke using EHR. *J Neurol Sci* 2021;427:117560.
- R Core Team. R: A language and environment for statistical computing; 2021 Available from: <https://www.r-project.org> [Last accessed on 2025 Mar 08].
- Peng SY, Chuang YC, Kang TW, Tseng KH. Random forest can predict 30-day mortality of spontaneous intracerebral hemorrhage with remarkable discrimination. *Eur J Neurol* 2010;17:945-50.
- Çelik G, Baykan ÖK, Kara Y, Tireli H. Predicting 10-day mortality in patients with strokes using neural networks and multivariate statistical methods. *J Stroke Cerebrovasc Dis* 2014;23:1506-12.
- Ho KC, Speier W, El-Saden S, Liebeskind DS, Saver JL, Bui AA, *et al.* Predicting discharge mortality after acute ischemic stroke using balanced data. *AMIA Annu Symp Proc* 2014;2014:1787-96.
- Cheon S, Kim J, Lim J. The use of deep learning to predict stroke patient mortality. *Int J Environ Res Public Health* 2019;16:1876.
- Hadianfard Z, Lotfnezhad Afshar H, Nazarbaghi S, Rahimi B, Timpka T. Predicting mortality in patients with stroke using data mining techniques. *Acta Inform Prag* 2022;11:36-47.
- Abujaber AA, Albalkhi I, Imam Y, Nashwan A, Akhtar N, Alkhalwaldeh IM. Machine learning-based prognostication of mortality in stroke patients. *Heliyon* 2024;10:e28869.
- Cunningham P, Delany SJ. K-nearest neighbour classifiers - a tutorial. *ACM Comput Surv* 2021;54:128.
- Liu Y, Wang Y, Zhang J. New machine learning algorithm: Random forest. In: Liu B, Ma M, Chang J, editors. *Information computing and applications*. Berlin: Springer; 2012. p. 246-52.
- Friedman J, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Ann Statist* 2000;28:337-407.
- Chen T, He T. Xgboost: eXtreme gradient boosting; 2017. Available from: <https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf> [Last accessed on 2025 May 15].
- Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol* 2010;5:1315-6.

How to cite this article: Ar Rochmah M, Nugroho DB, Gofir A, Ikhsan MR, Hanif F, Khairani AF, *et al.* In-hospital mortality predictors of stroke patients with diabetes mellitus. *J Neurosci Rural Pract.* 2026;17:64-72. doi: 10.25259/JNRP_85_2025