Predictive prognostic factors for stroke mortality in Thailand

Somchai Towanabut, Chidapha Traicharoenwong

MedicalNeurological Institute of Thailand (NIT), Bangkok, Thailand

Abstract

Stroke is the second leading cause of death worldwide and a significant public health concern in Thailand. This article presents predictive factors for in-hospital stroke mortality, utilizing data mining techniques from the Neurological Institute of Thailand's stroke database. Initially, 41 predictive variables were considered. However, after employing the CFS Subset Evaluator method for variable selection, five key predictive variables emerged: age, first diagnosis (IH, SH), the need for ventilator support, inability to receive a rehabilitation assessment, and the occurrence of pressure sores. Using these selected predictive variables, a classification model was created. The top three classifiers, with the highest F-Measure value of 0.971, were Na◉ve Bayes, Na◉ve Bayes Updateable, and Bayesian Network. The knowledge gained from this analysis can be valuable in enhancing the care provided to stroke patients and predicting high-risk stroke-related mortality.

Keywords: Death; Factor; Predict; Thailand; Data Mining

Received: 23 October 2023, Revised: 25 January 2024, Accepted: 1 February 2024

INTRODUCTION

Cerebrovascular Disease, which includes stroke and paralysis, is a prevalent neurological condition. In 2019, out of a global population of 7,637.7 million people[1], the World Health Organization[2] reported a total of 55.4 million deaths attributable to cerebrovascular disease, ranking it as the second leading cause of death and accounting for 11 percent of all global mortality. Stroke, secondary only to ischemic heart disease, is a significant public health concern in Thailand. It affects more males than females and exhibits an increasing trend. In 2019, the death rate from cerebrovascular disease in Thailand stood at 53.0 per 100,000 people[3].

Data mining is the process of discovering hidden knowledge, patterns, guidelines, and relationships within large datasets. It relies on statistics, recognition, machine learning, and mathematics.

Correspondence: Somchai Towanabut, Neurological Institute of Thailand (NIT), 312 Ratchawithi Rd, Thung Phaya Thai, Ratchathewi, Bangkok 10400, Tel: 02 306 9899, E-mail: s_towanabut@yahoo.com

Our previous studies on stroke patients involved a limited sample size, making it challenging to discern patterns among multidimensional variables. To address this, we employ data mining techniques to uncover hidden patterns using various algorithms developed in machine learning. We then validate these patterns using our dataset.

The Neurological Information Center at the Neurological Institute of Thailand has maintained a stroke database for over a decade, compiling data from hospitals across the network, resulting in a dataset exceeding 400,000 entries. Consequently, we are keen to investigate factors predicting stroke patient mortality in Thailand using this extensive dataset, aiming to advance our understanding of cerebrovascular disease.

LITERATURE REVIEW

Stroke Prognosis

Several factors influence the prognosis of stroke, including age and stroke severity, stroke mechanism, infarct location, co-morbidities, clinical symptoms, and

complications. Additionally, interventions, such as thrombolysis, mechanical thrombectomy, stroke unit, and rehabilitation, can play crucial roles in determining the treatment outcome for stroke patients. Understanding these significant factors that affect prognosis is essential for healthcare professionals to make accurate prognoses for each patient. It also allows for the delivery of appropriate patient care and helps patients and their families better comprehend the course of the disease[4].

Data Mining Classification

Data mining classification is a crucial task in the field of data analysis and machine learning. It involves categorizing data into predefined classes or labels based on the characteristics or features of the data. Classification models are trained using historical data with known labels to predict the class of new, unseen data points[5].

Supervised Learning: Classification is a type of supervised learning, where the algorithm learns from labeled training data and then predicts the class labels for unseen data.

Types of Classification Algorithms: There are various classification algorithms available, including:

- Decision Trees
- Random Forests
- Support Vector Machines (SVM)
- k-Nearest Neighbors (k-NN)
- Naive Bayes
- Neural Networks
- Logistic Regression

Model Evaluation: Classification models are evaluated using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).

Applications: Classification is used in a wide range of applications, including spam email detection, sentiment analysis, disease diagnosis, fraud detection, image recognition, and more.

Bayes' Theorem

Bayes' Theorem is a fundamental concept in probability theory and statistics. It describes the probability of an event based on prior knowledge of conditions that might be related to the event. In simple terms, it allows you to update your beliefs about an event when you have new evidence[6].

Mathematically, Bayes' Theorem can be expressed as:

$$P(A|B) = P(B|A)P(A)/P(B) \quad (1)$$

Where:

- $P(A|B)$ is the probability of event A happening given that event B has occurred.
- $P(B|A)$ is the probability of event B happening given that event A has occurred.
- $P(A)$ is the prior probability of event A.
- $P(B)$ is the prior probability of event B.

RESEARCH OBJECTIVES

To develop a predictive model for stroke-related mortality in Thailand.

System Overview

1. Data Collection: Gather a comprehensive dataset.
2. Data Preprocessing: Clean the collected data by exclude missing values, outliers, and inconsistencies.
3. Feature Selection: Identify relevant features that may influence stroke mortality.
4. Model Construction

Model Selection: Select appropriate data mining techniques for prediction modeling. This study conducted 44 classification model experiments.

Model Training: Employ 10-fold cross-validation.

Model Evaluation: Assess using metrics like accuracy, precision, recall, F-Measure, and area under the receiver operating characteristic curve (AUC-ROC).

5. Model Deployment and Validation: Apply the optimized model to new or unseen data to validate its generalizability and performance in real-world scenarios.
6. Interpretation and Insights

Conceptual Framework

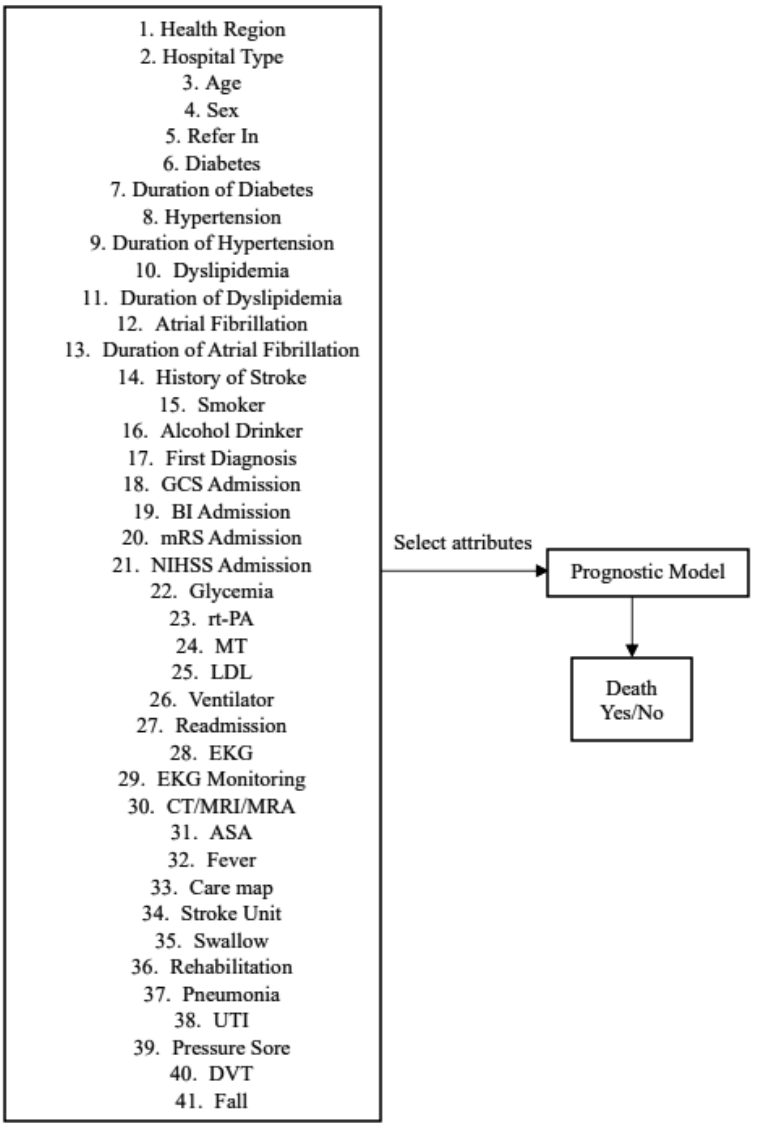

Figure 1. Conceptual Framework

This study is based on the stroke database from 2011 to 2022, following the conceptual framework in Figure 1. It includes a total of 41 predictive variables, which are as follows: Health Region, Hospital Type, Age, Sex, Refer In (Receive referrals from other hospitals), Diabetes, Duration of Diabetes, Hypertension, Duration of Hypertension, Dyslipidemia, Duration of Dyslipidemia, Atrial Fibrillation, Duration of Atrial Fibrillation, History of Stroke, Smoker, Alcohol Drinker, First Diagnosis, GCS Admission, BI Admission, mRS Admission, NIHSS Admission, Glycemia (mg%), rt-PA (Thrombolysis), MT (Mechanical Thrombectomy), LDL (Low-Density Lipoprotein cholesterol mg/dl), Ventilator (the need for ventilator support), Readmission (Patients readmitted for ischemic stroke or recurrent stroke within 28 days without prior planning.), EKG (The patient has an EKG exam.), EKG Monitoring (Patients received an EKG monitor within the first 24 hours after admission to the hospital.), CT/MRI/MRA (Patients received a CT brain scan and/or MRI/MRA within 24 hours.), ASA (The patient received antiplatelet (Aspirin) treatment within 48 hours of symptom onset.), Fever (The patient had a fever of >= 37.5^{o}C upon hospital admission.), Care map (Patients receive care following a Care Map/ Pathway.), Stroke Unit (The patient received care in the Stroke Unit.),

TABLE I. DESCRIPTIVE STATISTICS OF 99,973 STROKE PATIENTS

Variables	Mean	S.D.
Age (Years)	63.9	13.7
Duration of DM (Years)	2.3	5.1
Duration of HT (Years)	4.0	5.9
Duration of DLD (Years)	2.0	4.5
Duration of AF (Years)	0.3	1.7
GCS Admission	14.3	1.8
BI Admission	66.1	29.5
mRS Admission	2.8	1.5
NIHSS Admission	5.9	6.4
Glycemia (mg%)	138.6	66.3
LDL (mg/dl)	113.3	37.4

TABLE I. DESCRIPTIVE STATISTICS OF 99,973 STROKE PATIENTS (CONT.)

Variables	Groups	Percentage
Health Region	1	7.7
	2	7.5
	3	0.8
	4	4.9
	5	11.6
	6	8.1
	7	10.5
	8	6.5
	9	10.2
	10	8.4
	11	4.4
	12	12.2
	13	7.2
Hospital Type	Regional hosp	47.6
	General hosp	29.5
	Small gen hosp	10.4
	Community hosp	0.9
	Institute/U hosp	9.4
	Private hosp	2.2
Sex	Male	56.9
	Female	43.1
Refer In	Yes	46.0
	No	54.0
Diabetes (DM)	Yes	30.8
	No	69.2
Hypertension (HT)	Yes	61.7
	No	38.3
Dyslipidemia (DLD)	Yes	49.5
	No	50.5
Atrial Fibrillation (AF)	Yes	8.7
	No	91.3
History of Stroke	Yes	19.1
	No	80.9
Smoker	Yes	34.0
	No	66.0
Alcohol Dinker	Yes	31.7
	No	68.3

TABLE I. DESCRIPTIVE STATISTICS OF 99,973 STROKE PATIENTS (CONT.)

Variables	Groups	Percentage
First Diagnosis	Ischemic Stroke (IS)	93.4
	Intracerebral Hemorrhage (IH)	3.0
	Transient Ischemic Attack (TIA)	3.4
	Subarachnoid Hemorrhage (SH)	0.1
	Cerebral Venous Thrombosis (CVT)	0.2
rt-PA	Yes	10.7
	No	89.3
MT	Yes	0.5
	No	99.5
Ventilator	Yes	4.7
	No	95.3
Readmission	Yes	0.8
	No	99.2
EKG	Yes	99.7
	No	0.3
EKG Monitoring	Yes	94.0
	No	6.0
CT/MRI/MRA	Yes	99.9
	No	0.1
ASA	Yes	85.5
	No	14.5
Fever	Yes	10.9
	No	89.1
Care map	Yes	98.0
	No	2.0
Stroke Unit	Yes	93.2
	No	6.8
Swallow	Yes	99.3
	No	0.7
Rehabilitation	Yes	97.5
	No	2.5
Pneumonia	Yes	2.9
	No	97.1
UTI	Yes	1.6
	No	98.4
Pressure Sore	Yes	0.3
	No	99.7
DVT	Yes	0.0
	No	100.0
Fall	Yes	1.2
	No	98.8
Death	Yes	2.3
	No	97.7

Swallow (Patients had their swallowing assessed within 72 hours of admission.), Rehabilitation (Patients receive rehabilitation evaluation or care within 72 business hours of admission.), Pneumonia (The patient has

developed pneumonia as a complication.), UTI (The patient has developed complications from a urinary tract infection.), Pressure Sore (The patient has developed complications from pressure sores or skin break.), DVT (The patient has complications of deep vein thrombosis (DVT) in the legs.) and Fall (The patient had an accident and fell.). These variables were used to predict in-hospital stroke deaths, with a total of 99,973 cases, descriptive statistics as shown in TABLE I."

In data mining[7], feature selection, known by various terms like variable selection, feature reduction, attribute selection, or variable subset selection, comprises a set of techniques used to choose a subset of relevant features while eliminating irrelevant or redundant ones. The primary goals of feature selection are threefold: enhancing the performance of data mining models, expediting the learning process, and gaining a deeper understanding of the data generation process. Feature selection algorithms commonly fall into two categories: feature ranking and subset selection. Feature ranking involves ranking all features using a specified metric and discarding those that don't meet a threshold. Subset selection, on the other hand, seeks the optimal subset of features without ranking them individually. It's important to note that different feature selection methods can yield different reduced feature sets.

In this study, we employed the subset selection approach. Variables were selected using the Attribute Subset Evaluator method, focusing on predicting death as a nominal class. The selected attributes, including Age, First diagnosis, Ventilator, Rehabilitation, and Pressure sore, totaled 5

Model Construction

The model building process begins with the features selected during the Feature Selection step. A total of 44 classifiers are tested in the model selection phase using a dataset containing 99,973 cases. The model is trained using a 10-fold cross-validation method, which involves dividing the dataset into 10 sections. In each iteration, nine sections are used for training the model, and one section is reserved for testing. This process is repeated for 10 rounds, allowing us to evaluate the model comprehensively.

TABLE II. COMPARING THE PERFORMANCE OF TOP 3 MODELS BASED ON THE CLASSIFIER USING 10-FOLD CROSS-VALIDATION

Classifier	Precision	Recall	F-Measure	AUC
Naïve Bayes	0.968	0.976	0.971	0.864
Naïve Bayes Updateable	0.968	0.976	0.971	0.864
Bayesian Network	0.969	0.976	0.971	0.859

Evaluation metrics including accuracy, precision, recall, F-Measure, and area under the receiver operating characteristic curve (AUC-ROC) are considered. The top-performing models, as indicated in Table II, are selected based on these metrics.

A Naive Bayes classifier[8] is a type of probabilistic machine learning model that's used for classification tasks. It's based on Bayes' theorem and the assumption of naive independence among features, which is where the "naive" in its name comes from. Despite its simplifying assumption, Naive Bayes classifiers often perform surprisingly well in practice, especially in text classification and other tasks where the independence assumption is a reasonable approximation.

A Na◉ve Bayes Updateable classifier is an extension of the Naive Bayes classifier that is designed to be incrementally updated as new data becomes available. In other words, it can learn and adapt to new examples without needing to retrain the entire model from scratch. This can be particularly useful in situations where you have a continuous stream of data and need to make real-time predictions or keep the model up-to-date with new information.

The models from Na◉ve Bayes and Na◉ve Bayes Updateable, as shown in Table III, are identical, confirming the results of this study.

$$\text{Posterior(death)}=P(\text{death})P(\text{age}|\text{death})P(\text{fstdx}|\text{death})P(\text{ventilator}|\text{death})P(\text{rehab}|\text{death})P(\text{p\_sore}|\text{death})/\text{Evidence} \quad (2)$$

$$\text{Posterior(}\sim\text{death)}=P(\sim\text{death})P(\text{age}|\sim\text{death})P(\text{fstdx}|\sim\text{death})P(\text{ventilator}|\sim\text{death})P(\text{rehab}|\sim\text{death})P(\text{p\_sore}|\sim\text{death})/\text{Evidence} \quad (3)$$

$$\text{Evidence}=P(\text{death})P(\text{age}|\text{death})P(\text{fstdx}|\text{death})P(\text{ventilator}|\text{death})P(\text{rehab}|\text{death})P(\text{p\_sore}|\text{death}) + P(\sim\text{death})P(\text{age}|\sim\text{death})P(\text{fstdx}|\sim\text{death})P(\text{ventilator}|\sim\text{death})P(\text{rehab}|\sim\text{death})P(\text{p\_sore}|\sim\text{death}) \quad (4)$$

TABLE III. NAÏVE BAYES MODEL AND NAÏVE BAYES UPDATEABLE MODEL

	Death		
	Yes	No	
Attribute	(0.02)	(0.98)	**% death**
Age			
Mean	69.1428	63.7612	
Std. dev.	14.2532	13.6905	
Weight sum	2,332	97,641	
precision	1	1	
First Diagnosis			
Ischemic Stroke (IS)	1,968	91,375	2.1
Intracerebral Hemorrhage (IH)	344	2,684	11.4
Transient Ischemic Attack (TIA)	4	3,364	0.1
Subarachnoid Hemorrhage (SH)	10	77	11.5
Cerebral Venous Thrombosis (CVT)	11	146	7.0
[total]	2,337	97,646	2.3
Ventilator			
Yes	1,319	3,395	28.0
No	1,015	94,248	1.1
[total]	2,334	97,643	2.3
Rehabilitation			
Yes	2,074	95,352	2.1
No	260	2,291	10.2
[total]	2,334	97,643	2.3
Pressure Sore			
Yes	80	213	27.3
No	2,254	97,430	2.3
[total]	2,334	97,643	2.3

In this paper, we select the Naive Bayes classifier with the best F-Measure value and a higher AUC value than the Bayesian Network classifier.

Predictive Equation[9]:

Model Deployment and Validation

TABLE IV. DESCRIPTIVE STATISTICS OF 42,092 STROKE PATIENTS IN 2023

Variables	**Mean**	**S.D.**	**Percentage**
Age	63.7	13.6	
First Diagnosis	Ischemic Stroke (IS)		92.9
	Intracerebral Hemorrhage (IH)		3.6
	Transient Ischemic Attack (TIA)		3.1
	Subarachnoid Hemorrhage (SH)		0.2
	Cerebral Venous Thrombosis (CVT)		0.2
Ventilator	Yes		5.9
	No		94.1
Rehabilitation	Yes		98.1
	No		1.9
Pressure Sore	Yes		0.2
	No		99.8
Death	Yes		2.5
	No		97.5

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     Test1
Instances:     unknown (yet). Reading incrementally
Attributes:   6

=== Summary ===

Correctly Classified Instances       40980               97.3582 %
Incorrectly Classified Instances      1112                2.6418 %
Kappa statistic                          0.2717
Mean absolute error                      0.0413
Root mean squared error                  0.1458
Total Number of Instances            42092

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.993    0.789    0.980      0.993    0.987      0.290    0.860     0.994     N
                 0.211    0.007    0.432      0.211    0.284      0.290    0.860     0.271     Y
Weighted Avg.    0.974    0.770    0.967      0.974    0.969      0.290    0.860     0.976

=== Confusion Matrix ===

     a     b   <-- classified as
 40760   289 |     a = N
   823   220 |     b = Y
```

Figure 2. Validation Stroke Death in Thailand Model

Validate the model's generalizability and performance in real-world scenarios by applying it to new or unseen data. Utilize the Na◉ve Bayes Classifier to predict stroke deaths in hospitals using data from 2023, comprising 42,092 instances with characteristics outlined in Table

The Stroke Death in Thailand model, used to forecast data in 2023, showed a slight reduction in accuracy (97.4%), sensitivity (96.7%), F-Measure (96.9%) and AUC (0.9) values compared to those during model creation. Nevertheless, this model remains effective for predicting stroke patient mortality in Thailand.

DISCUSSION

Our network dataset comprises over 400,000 entries from more than 100 hospitals across Thailand. The average age of patients is 63.9 years, with Ischemic Stroke being the most common diagnosis (93.4%). Other conditions include Intracerebral Hemorrhage, TIA, SH, and CVT, with only 4.7% requiring ventilator support. Furthermore, 93.2% of patients are admitted to stroke units, and 98% follow care maps. Rehabilitation assessments are received by 97.5% of patients, and pressure sores occur in only 0.3% of cases. The total stroke-related death rate in this dataset is 2.3%.

From the model presented in Table III, we predict five factors related to stroke patient mortality:

1. Patient age averages 69 years.
2. The presence of Intracerebral Hemorrhage (IH) or Subarachnoid Hemorrhage (SH)
3. The need for ventilator support.
4. The patient inability to receive a rehabilitation evaluation. *

5. Occurrence of pressure sores.

* After reviewing the factors, it was observed that some patients were not evaluated for rehabilitation. Subsequently, these patients experienced unfavorable outcomes, including unconsciousness, depression, fatigue, instability, coma, surgical procedures, large infarction, brain edema, and the need for palliative care. These factors contributed to their adverse condition.

This analysis enhances the quality of care provided to stroke patients and aids in identifying those at high risk of stroke-related mortality.

CONCLUSION

In the literature review, several factors influencing the prognosis of stroke were identified, including age, stroke severity, and complications. In our predictive model, we consider factors related to stroke-related mortality, such as age, stroke type (IH & SH), stroke severity (as indicated by the need for ventilator support), and the inability to receive a rehabilitation assessment. An additional factor we include is the occurrence of pressure sores. These predictive factors enhance the care provided to stroke patients and assist healthcare providers in identifying those at high risk of stroke-related mortality.

ACKNOWLEDGMENT

Thank you to the Stroke Network Hospitals and the Neurological Institute of Thailand for collaborating in the development of stroke services and providing access to anonymized patient data. This research was approved by the IRB at the Neurological Institute of Thailand. We hope that the knowledge gained can be applied to improve stroke patient care in the future.

REFERENCES

[1] World Bank. "World Development Indicators." World Bank. [Online]. Available: http://wdi.worldbank.org/table/2.1. Accessed: April 28, 2021.

[2] World Health Organization (WHO)."The top 10 causes of death." WHO. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed: April, 2021.

[3] Strategy and Planning Division Ministry of Public Health, "Public Health Statistics A.D. 2019," Nonthaburi, 2020.

[4] "Overview of Ischemic Stroke Prognosis in Adults - UpToDate," UpToDate, https://www.uptodate.com/contents/overview-of-ischemic-stroke-prognosis-in-adults. Accessed: Sep. 12, 2023.

[5] P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining," Addison-Wesley, 2005.

[6] Wikipedia Contributors, “Bayes’ theorem,” Wikipedia, Sep.21,2023.https://en.wikipedia.org/wiki/Bayes%27_theorem (accessed Sep. 22, 2023).

[7] M. Kantardzic, "Data Mining: Concepts, Models, Methods, and Algorithms," 2nd ed. Wiley, 2011.

[8] S. Ray,“Naive Bayes Classifier Explained: Applications and Practice Problems of Naive Bayes Classifier,” Analytics Vidhya, Sep. 11, 2017. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ (accessed Sep. 19, 2023).

[9] Wikipedia Contributors, “Naive Bayes classifier,” Wikipedia, Sep. 03, 2023. https://en.wikipedia.org/wiki/Naive_Bayes_classifier (accessed Sep. 19, 2023).