BRAIN. Broad Research in Artificial Intelligence and Neuroscience
e-ISSN: 2067-3957 | p-ISSN: 2068-0473
Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science Evaluation Center, neredataltics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu)
2025, Volume 16, Issue 2, pages: 296-310.
Submitted: March 5th, 2025 | *Accepted for publication:* May 13th, 2025

Binary Classification Models for Stroke Outcome Prediction

Virgilijus Sakalauskas
Kauno kolegija Higher Education Institution, Pramones pr. 20, LT-50468 Kaunas, Lithuania.
virgilijus.sakalauskas@go.kauko.lt
https://orcid.org/0000-0002-5572-8889

Dalia Kriksciuniene
Vilnius University, Universiteto St. 3, LT-01513 Vilnius, Lithuania.
dalia.kriksciuniene@knf.vu.lt
https://orcid.org/0000-0002-0730-3763

Abstract: *Stroke is found to be a leading cause of mortality and long-term disability worldwide, forcing effective predictive models to identify at-risk individuals and optimise treatment plans. In this study, we evaluate the performance of various machine learning (ML) algorithms in predicting stroke-related mortality. Five binary classification models—Logistic Regression (LR), Random Forest (RF), Gradient Boosting Machines (XGBoost), Support Vector Machine (SVM), and Neural Networks (MLPClassifier)-were applied to a dataset containing clinical and demographic features of stroke patients registered by the neurology department of the Clinical Centre of Montenegro. Each model was trained and evaluated using standard classification metrics: accuracy, precision, recall, and F1-score. Also, the importance of the feature was analysed to find the key predictors of stroke mortality across different models. The research shows the Random Forest and XGBoost performance over simpler models, proposing superior accuracy and interpretability. By analysing how precision, recall, and accuracy changes across a range of classification thresholds, we gained deeper insight into the model's reliability under different clinical conditions. This analysis revealed clear trade-offs: lower thresholds improve recall (reducing the risk of missed death predictions), while higher thresholds enhance precision (minimising false positives). The findings support the selection of threshold values tailored to specific clinical priorities, such as early warning, balanced risk assessment, or high-confidence decision-making.*

Keywords: *artificial intelligence; data mining; healthcare data; stroke; machine learning algorithms; threshold value.*

How to cite: Sakalauskas, V. & Kriksciuniene, D. (2025). Binary classification models for stroke outcome prediction. *BRAIN: Broad Research in Artificial Intelligence and Neuroscience, 16*(2), 296–310. https://doi.org/10.70594/brain/16.2/22

©2025 Published by EduSoft Publishing. This is an open-access article under the CC BY-NC-ND license

1. Introduction

Stroke is the most complex public health challenge and is responsible for millions of deaths. It is not just the second leading cause of death globally, but it also complicates lives for individuals and their families and raises problems for the healthcare system. Despite high achievements in medical treatment and preventive strategies, predicting survival following stroke-related complications remains a challenging endeavour. This uncertainty highlights the necessity for predictive models to offer clinicians a more precise understanding of patient outcomes.

The rise of machine learning (ML) algorithm applications also helps healthcare facilities predict stroke patients' physical and mental well-being. Machine learning models have the potential to detect patterns and correlations that escape traditional statistical methods. In stroke management, such models could reveal robust methods for early identification of patients at risk of mortality, leading to faster interventions and tailored treatments. Nevertheless, despite the progress, not all ML training algorithms work correctly. The task of choosing the right model that balances accuracy with its interpretability is a critical challenge that must be comprehensively addressed.

Numerous researchers have explored various ML algorithms and techniques to predict clinical outcomes in stroke patients (see the next section), with models such as Logistic Regression and Random Forest, frequently cited for their performance. However, with the emergence of more sophisticated algorithms, such as XGBoost and Neural Networks, there is a growing interest in comparative evaluation, employing a comprehensive dataset and adequate evaluation metrics. Moreover, as the complexity of these models increases, it is essential to determine the most influential predictors shaping their outputs. Knowing which features are most important in determining a patient's prognosis can be just as valuable as the prediction itself.

In this study, we aim to evaluate the performance of five prominent ML classifiers: Logistic Regression, Random Forest, Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Neural Networks (MLPClassifier), for predicting stroke-related mortality.

The experimental research was conducted using a database of registered stroke cases from the Neurology Department of the Clinical Centre in Montenegro. The initial database consists of 944 structured patient records, encompassing 58 variables. We selected 10 variables for our research and applied specific coding to their value. A detailed description of the study of stroke patients' database used is provided in section 4.

By utilising this clinical and demographic data from stroke patients, we have explored the accuracy of these models and the key features that drive their predictions. By providing a full-scale comparison of these approaches, we aim to elucidate the performance of the methods and offer the most significant promise for improving stroke care in terms of predictive power, accuracy, and interpretability.

The next section is intended for a literature review of ML applications for stroke-related datasets. Section 3 provides a brief overview of ML techniques, while Section 4 introduces the stroke database used in our research. The most crucial section of the article – section 5 - presents the results of our study. Finally, the article concludes with a dedicated section for discussion and final remarks.

2. Literature Review

Machine learning (ML) is a widely recognised tool, frequently referenced in healthcare literature, including those examining the aetiology and possible complications of stroke. Usually, stroke mortality has been predicted using statistical models, like Cox proportional hazards and logistic regression. The popularity of these methods is due to their simplicity and ease of interpretation. However, their performance declines when modelling complex relationships in large datasets, which force researchers to explore more advanced machine-learning methods. The following section reviews relevant literature concerning the application of ML models for predicting stroke-related mortality.

2.1. Logistic Regression

In stroke research, the most widely used model is logistic regression. Wang (2023) and Krikščiunienė & Sakalauskas (2022) found that logistic regression could effectively predict stroke mortality, using clinical variables such as age, hypertension, and stroke severity. Although this method is straightforward, it exhibits notable limitations when dealing with more complicated datasets. Logistic regression's assumption of a linear connection between variables may not accurately represent the nature of medical data, especially in large, diverse patient groups.

2.2. Random Forest

Random Forest is an ML method that constructs ensembles of decision trees for predictive modelling and data-driven inference. It aggregates the predictions of multiple decision trees, at the same time reducing the risk of overfitting. Random Forest handles missing data better than many traditional methods. Egegamuka et al. (2024) utilised the Random Forest method to predict stroke outcomes, introducing a novel outlier detection technique to eliminate irrelevant features. Fernandez-Lozano et al. (2021) investigated the use of Random Forest for predicting mortality and morbidity at three months post-admission. Although this model was more accurate than logistic regression, it comes with a trade-off: as the number of trees increases, it becomes harder to interpret the model's decisions.

2.3. XGBoost and Gradient Boosting Techniques

Gradient-boosting algorithm (XGBoost) is recognised for its performance in classification tasks. Wang et al. (2022) identified high-risk aSAH (aneurysmal subarachnoid haemorrhage) patients by using XGBoost prognostic model. Chung et al. (2023) demonstrated the model's performance for identifying the patients receiving different AIS (Acute Ischemic Stroke) treatments and provided clinical evidence for feature optimisation of AIS treatment strategies.

XGBoost's advantage lies in its ability to minimise bias and variance while handling missing data effectively. It also provides valuable insights into the importance of features, helping to identify the factors that most influence patient outcomes. However, its complexity can be a barrier for those looking for transparent and interpretable models.

2.4. Support Vector Machines (SVMs)

Support Vector Machines (SVMs) are helpful in cases where datasets are smaller but high-dimensional. Feng (2023) and Zhang (2023) showed that SVMs could accurately predict stroke mortality, particularly when paired with feature selection techniques. Despite this, SVMs can be less intuitive for clinicians, making their use in clinical practice more challenging, especially compared to models that provide more precise insights into how predictions are made.

2.5. Neural Networks and Deep Learning

Neural networks, particularly multi-layer perceptrons (MLP), are increasingly used to predict medical outcomes because of their ability to model complex and nonlinear relationships between variables. Cheon, Kim, & Lim (2019) applied Principal Component Analysis (PCA) with quantile scaling to extract relevant background features from medical records and predict stroke occurrence. To predict stroke mortality and identify the most significant risk factors, the neural network was also employed in the research of Someeh et al. (2023). They reported that stroke mortality is most strongly influenced by the following features: smoking, lower education, age, lack of physical activity, diabetes, and body weight. However, neural networks are often seen as "black boxes", because they do not readily provide interpretable explanations for their predictions, making it difficult for clinicians to trust and act upon the results without additional interpretative tools.

2.6. Interpretable Machine Learning

The challenge of balancing accuracy with interpretability is a common theme in healthcare applications of machine learning. Lundberg & Lee (2017) introduced SHAP (Shapley Additive explanations), a technique that makes machine learning models more interpretable by attributing prediction outcomes to individual features. SHAP has proven helpful in increasing the transparency of complex models like XGBoost and neural networks, helping to make their predictions more understandable and applicable in clinical practice.

Fernandes et al. (2024) conducted a comprehensive review of 25 review papers published between 2020 and 2024 on machine-learning and deep-learning applications in brain stroke diagnosis, focusing on classification, segmentation, and object detection. The analysis shows that advanced ML models, such as Random Forest, XGBoost, and neural networks offer significant improvements in prediction accuracy over traditional methods, but lack interpretability. Logistic regression remains a favoured method for its ease of use, but it lacks the predictive power of these newer models.

As machine learning in healthcare continues to evolve, researchers must focus on developing models that are both accurate and interpretable for medical professionals.

3. A Brief Overview of ML Techniques

For this research, we employ five machine learning (ML) techniques to predict stroke-related mortality. Each method adopts a unique approach to learning patterns from the data, and we outline them below with the key mathematical formulations.

3.1. Logistic Regression

Logistic regression is a linear model used for binary classification. It estimates the probability that a given input belongs to a certain class based on a linear combination of input features.

The model is defined as (1):

$$P(y = 1|X) = \frac{1}{1+e^{-(\propto_0+\propto_1 X_1+\ldots+\propto_n X_n)}} \quad (1)$$

where $P(y = 1|X)$ is the probability of the target class; $\propto_0, \propto_1, \ldots, \propto_n$ are the coefficients (weights) and $X_1, \ldots, X_n$ are the input features.

3.2. Random Forest Classifier

Random Forest is an ensemble-learning method that combines multiple decision trees to improve predictive performance. It averages the predictions of several decision trees to reduce overfitting and variance.

For each tree, predictions are made based on majority voting. The general structure for decision trees is based on recursively splitting the dataset according to feature values that minimise the Gini impurity (2):

$$Gini(D) = 1 - \sum_{i=1}^{C} p_i^2 \quad (2)$$

where *D* is a dataset, pi is the proportion of class I in *D, and C* is the number of classes.

The Random forest takes an average of all decision trees by voting.

3.3. Gradient Boosting Machines (XGBoost)

XGBoost (Extreme Gradient Boosting) is an advanced machine learning method that builds multiple small models (decision trees) in sequence, each correcting the previous model's errors. Instead of training all models at once, XGBoost builds one tree at a time, with each subsequent tree aiming to reduce the residual errors of the preceding ones. Initially, XGBoost minimises an objective function, comprising the logistic loss and a regularisation term that penalises tree complexity. After adding each tree, XGBoost calculates the gradient, which guides the model in adjusting its predictions to get closer to the actual values. The next tree is built to minimise the difference between the actual values and the predictions (using these gradients). The final prediction is obtained by aggregating the results of all the individual trees.

3.4. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning method used for classification tasks, aiming to separate data into two different groups (like predicting "survival" or "non-survival"). SVM aims to find the best boundary (called a hyperplane) that separates the two classes as clearly as possible.

SVM tries to find a line (in 2D) or a plane (in higher dimensions) that divides the data points into two groups. The best hyperplane is the one that creates the most significant "gap" or margin between the two groups.

SVM doesn't just draw any boundary between the groups, it looks for the hyperplane that leaves the most space between the closest points of each group. These closest points are called support vectors. The wider the gap, the better the model.

Sometimes, the data can't be separated in a straight line. In these cases, SVM employs a technique known as the kernel trick. This trick transforms the data into a higher-dimensional space where separating the groups with a hyperplane becomes easier.

Once the hyperplane is found, new data points are classified based on which side of the hyperplane they fall on. If a new point is on one side of the line, it belongs to class A. If it's on the other, it belongs to class B.

3.5. Neural Networks (MLPClassifier)

A neural network consists of layers of interconnected neurons. Each neuron applies a weighted sum of inputs followed by a non-linear activation function.

For a single neuron, the output is given by (3):

$$Output = \sigma\cdot(\sum_{i=1}^{n} \omega_i X_i + b) \qquad (3)$$

Where ω_i are the weights, b is the bias term, and $\sigma(\cdot)$ is an activation function, often the sigmoid function for binary classification.

The neural network is trained by minimising a loss function, typically the binary cross-entropy for classification, using backpropagation (4):

$$L = -\sum_{i=1}^{N} (y_i \log log \left(\hat{y}_i\right) + (1 - y_i) \log log (1 - \hat{y}_i)) \qquad (4)$$

In models like Random Forest and XGBoost, feature importance is derived from the contribution of each feature to splits in the decision trees. For models like SVM and Neural Networks, permutation importance can be used, which measures the change in model performance when a feature's values are randomly shuffled.

4. Stroke Clinical Data Description

The database used in this experimental research consists of clinical data records of stroke patients registered by the neurology department of the Clinical Centre of Montenegro, operating in Podgorica, Montenegro. The original dataset contains 944 structured patient records and 58 variables. Of these, 50 are categorical variables encoded, using a {1, 2, 3} scale representing 'Yes', 'No',and 'Unspecified', while the remaining eight include demographic information, admission date, and discharge date. The data were collected between 02/25/2017 and 12/18/2019. The demographic data of stroke patients vary by age (13 to 96 years) and gender (485-male, 427-female). For our research, we have cleansed the initial stroke database, re-encoded selected variables, and finally, based on a combination of domain knowledge from clinical neurologists at the data source, we have selected an 11-variable database for research (Sakalauskas et al., 2022).

The example of database structure and data records is presented in Table 1. The full database can be downloaded from https://github.com/Virgilijus11/StrokeData.git.

Table 1. Sample of data records and variables of the stroke cases database

Days at Hospital	Vital Status	Stroke Type	Treatment methods	Health Status	Age	Gender	Past Stroke	Stroke Symptoms	Health Com-plications	Smoke
2	1	2	4	9	79	2	1	12	4	3
2	0	2	4	0	79	2	1	13	0	2
6	0	1	24	0	79	2	0	2	0	2
16	0	1	24	2	78	2	0	12	0	2
13	0	1	24	0	77	2	0	23	1	2
45	1	3	14	1	77	2	0	12	0	1
9	0	1	24	3	78	2	0	123	0	2
5	0	1	24	0	77	2	0	23	4	1
6	1	1	14	0	77	1	0	23	0	2
1	0	1	24	0	76	1	0	2	0	2
2	0	1	24	0	76	2	0	3	0	2
8	0	1	24	0	77	2	1	23	0	2
3	0	1	24	0	75	1	0	2	0	2
7	1	1	24	0	77	1	0	23	2	2
1	0	1	24	0	77	2	0	123	0	1

The variables of the stroke database are coded for applying the survival modelling methodologies; their values are explained in Table 2.

A stroke is a medical condition where there is an interruption in cerebral perfusion. In general, stroke is classified into two primary types of strokes: haemorrhagic (Haemorrhage) stroke and ischaemic stroke (Ishemic) (Deljavan, Farhoudi, & Sadeghi-Bazargani, 2018). However, the definition of symptoms enables more stroke types and clinical subtypes. The aetiology of stroke is often associated with morbidities of patients, such as diabetes and heart diseases. Therefore, numerous potentially significant variables are registered in the stroke data sets.

Ischaemic strokes occur when cerebral blood vessels become obstructed, restricting blood flow to the brain. This type of stroke makes up about 87% of all stroke cases (Deljavan, Farhoudi, & Sadeghi-Bazargani, 2018).

Table 2. The definition of stroke clinical database variables

Variable name	**Meaning and coding of data**
Days at Hospital	– the number of days after stroke till hospital admission
Vital Status	– 1: Event (death), 0: Alive/censored
Stroke Type	- 1: Ischemic, 2: Haemorrhage, 3: SAH, 4: Unspecified
Treatment methods	- 0: No treatment, 1: Anticoagulation, 2: Dual Antiplatelet Therapy, 3: Thrombolysis, 4: Others, Two digit codes: mean combined treatment methods, e.g., 24: means 2 and 4 are applied.
Health Status	- Health score before stroke from 0: best to 9: worst. 0: Without symptoms; 1: Without significant disability despite symptoms; 2: Minor disability; 3: Moderate disability, but able to walk independency; 4: Moderate disability, not able to walk independency; 5: Major disability; 9: Unknown
Age	– Patient age, years
Gender	– 1: Male, 2: Female, 9: Unspecified
Past Stroke	– Stroke in the past. 1: Yes, registered in the patient health record, 0: No
Stroke Symptoms	0: No symptoms, 1: Impaired consciousness, 2: Weakness/paresis, 3: Speech disorder (aphasia), Several digit codes: 123-means all three symptoms
Health complications	0: unspecified, 1: other CV (cardiovascular) complications, 4: other complications, Several digit codes: 23: means 2 and 3
Smoke Status	1-Smokes, 2-No, 3-Smoked before

There are two types of ischemic strokes, thrombotic and embolic strokes, which differ in underlying pathophysiology.

A thrombotic stroke is caused by a clot forming in a blood vessel of the brain, often related to atherosclerosis.

An ischaemic stroke can be embolic, involving a blood clot that travels from another part of the body to the cerebral circulation. Approximately 15% of embolic strokes are due to a condition called atrial fibrillation.

A haemorrhagic stroke is caused by bleeding, which may occur within the brain parenchyma or in the subarachnoid space. This is the cause of about 20% of all strokes (World Stroke Organization, 2019). Haemorrhagic strokes are classified into two main categories, based on the location and cause of bleeding: intracerebral haemorrhage (haemorrhage) and subarachnoid haemorrhages (SAH). Intracerebral haemorrhages are caused by a broken blood vessel located in the brain. Severely elevated blood pressure can cause weakening of the small blood vessels in the brain. It may be related to anticoagulant therapy. The second category, subarachnoid haemorrhages (SAH), occurs when a blood vessel gets damaged, leading to bleeding in the area between the brain and the thin tissues that cover it. A ruptured aneurysm, AVM, or head injury can cause SAH. SAH is a less common type of haemorrhagic stroke, approximately 5–6% of all strokes. Due to its distinct aetiology, it is allocated to the separate category of stroke SAH. The other types of stroke include various cases such as Cryptogenic Stroke, Brain Stem Stroke, and others.

Recurrent stroke accounts for nearly 25% of all stroke cases. Age and modifiable lifestyle factors, such as smoking, hypertension, and obesity, are among the common predictors of stroke and its outcomes.

The variables in the research data set were prepared according to the clinical characteristics associated with the identified stroke subtypes.

5. Performance Comparison of Binary Classification Models

In this research, five widely used binary classification models—Logistic Regression (LR), Random Forest (RF), Gradient Boosting Machines (XGBoost), Support Vector Machine (SVM), and Neural Networks (MLPClassifier)—were implemented using Python to predict stroke-related mortality. The main goal was to compare their performance based on multiple evaluation metrics and select the most effective model for the given task.

The experiment includes four stages: model training and evaluation, testing the impact of the training set size on accuracy, feature importance analysis, and threshold-based evaluation of Random Forest model performance. The data preprocessing, model implementation, and all experiment stages were conducted using Python (Figure 1), enabling the generation of all results presented in this research.



Figure 1. QR code for Python program

Before applying the machine learning models, the dataset underwent several preprocessing steps: handling missing values, encoding categorical variables, splitting the initial database into training and testing sets and applying feature scaling where appropriate.

Missing data were addressed using a forward-filling procedure (*fill) to ensure the* completeness of the dataset. For all categorical features, we have applied an encoding procedure, converting them into numerical form to ensure compatibility with machine learning algorithms. The dataset was split into training and testing sets using the train_test_split() function from the scikit-learn.. The initial split ratio was set at 70% of all records for training and 30% for testing. Also, we explored the effect of changing this ratio. In Logistic Regression, SVM, and Neural Networks, we have applied feature scaling to ensure that the models converge efficiently.

Each binary classification model described in section 3 was implemented in Python, using corresponding Python libraries. The implemented function trains and evaluates the following models:

Logistic Regression:	Logistic regression () from scikit-learn.
Random Forest Classifier:	RandomForestClassifier() from scikit-learn.
XGBoost:	XGBClassifier() from XGBoost.
Support Vector Machine (SVM):	SVC() from scikit-learn.
Neural Network (MLPClassifier):	MLPClassifier() from scikit-learn.

To optimise model performance, a grid search was conducted to identify the optimal hyperparameter values. The models were trained on the training data set and evaluated on the testing set.

To assess and compare the models' performance, we used four evaluation metrics:

- Accuracy: The percentage of correct predictions;
- Precision: The percentage of correct predictions along the predicted positive cases;
- Recall: The percentage of correct predictions along the actual positive cases;

- F1-Score: The harmonic mean of precision and recall;

Each model's performance was recorded using these metrics and compared across models to identify the best-performing one.

5.1. Performance of ML Models

The results of the research are summarised in Table 3, which compares the performance data for all binary classification models in the case of a training set size of 70% of the dataset - 188 records representing the "alive" class and 96 the "deceased" class. The random state parameter here is set equal to 42.

Table 3. Performance of various ML algorithms in predicting stroke-related mortality

		Accuracy	***Precision***	***Recall***	***F1 score***
Logistic regression	*0-alive*	76.41%	0.77	0.91	0.84
	1-dead		0.73	0.48	0.58
Random Forest	*0-alive*	79.93%	0.80	0.92	0.86
	1-dead		0.78	0.56	0.65
XGBoost	*0-alive*	75.70%	0.80	0.85	0.82
	1-dead		0.66	0.58	0.62
SVM	*0-alive*	76.41%	0.76	0.94	0.84
	1-dead		0.78	0.42	0.54
MLPClassifier	*0-alive*	76.06%	0.80	0.85	0.82
	1-dead		0.66	0.59	0.63

The table presents the results of five different binary classification models in terms of their performance for predicting stroke-related mortality, where the target classes are "0" (alive) and "1" (dead). The performance metrics employed: accuracy, precision, recall, and F1 score, provide insight into the models' predictive power and reliability.

Paired t-test analysis comparing model performances revealed statistically significant differences at the 95% confidence level.

As we see Logistic regression performs well in predicting the "alive" class, with high recall (0.91), meaning it successfully identifies a large proportion of those alive. However, it struggles with predicting the "dead" class, with a much lower recall (0.48), indicating a high rate of false negatives.

The Random Forest method shows the highest overall accuracy (79.93%) and performs well for both classes. For the "alive" class, it achieves high precision and recall, while for the "dead" class, recall (0.56) and F1-score (0.65) significantly improve compared to logistic regression, enhancing its ability to identify mortality cases.

The XGBoost model exhibits relatively balanced performance, though it demonstrates slightly lower overall accuracy, particularly in its ability to identify patients at risk of death.

SVM performance in predicting the "alive" class is notably high, with a recall of 0.94, correctly identifying the majority of surviving patients. However, the "dead" class, as seen from the low recall equal to 0.42 means it misses over half of the death cases. Consequently, the low recall reduces its effectiveness in predicting mortality.

The accuracy and F1 score of MLPClassifier are comparable with XGBoost, indicating that it may also have issues in accurately identifying mortality cases.

In summary, Random Forest emerged as the best-performing model for stroke-related mortality prediction, especially regarding overall accuracy and its ability to identify patients likely to survive. Nonetheless, the low precision and recall observed for class '1 – dead' underscore the need to improve mortality prediction across all models

5.2. Impact of the Training Set Size on Model Accuracy

During the next experiment, we evaluated the impact of the training set size on model accuracy. Figure 2 presents the variation in model accuracy as the training set size is systematically adjusted. This graph illustrates the dynamic changes in the accuracy of five machine learning models based on the proportion of the training set used, ranging from 0.9 to 0.1. The y-axis represents accuracy in percentages. The x-axis shows the training set size as a percentage proportion of the total dataset.

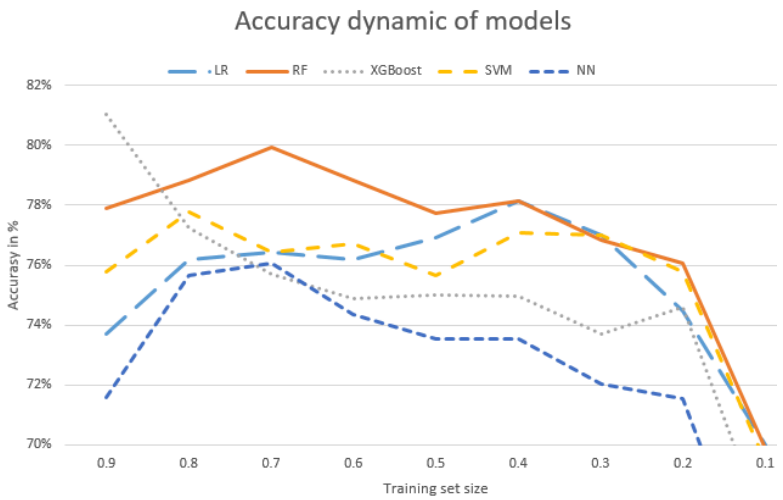

Figure 2. Impact of the training set size on model accuracy.

The Random Forest (RF) graph stands out with the highest accuracy, peaking at around 80% accuracy for a training size of 0.7. However, we see its performance sharply drop down when the training set size is below 0.4, indicating that it may not generalise well as the training set decreases further.

XGBoost linear graph only shows the best performance for the training set around 0.9. Then, it is stable enough between 0.7-0.2, with accuracy consistently around 75%. If the training set size drops below 0.2, the accuracy falls accordingly.

Logistic Regression (LR) for all set sizes performs well but maintains lower accuracy than RF. Its peak is around 78% with a training set of 0.4. For smaller training sets (between 0.4 and 0.3), the performance of LR is very similar to RF.

SVM linear graph demonstrates a performance similar to the logistic regression model. When the training set proportion is 0.3 or lower, SVM performs comparably to the best-performing RF model. Thus, under limited training data conditions, the Support Vector Machine model may offer the highest accuracy. Neural Networks (NN) exhibited the lowest overall accuracy across all training set sizes.

5.3. Feature Importance Score

Subsequently, we identify the most critical features contributing to model predictions. Based on the feature importance table (Table 4) across multiple models, the most influential predictors of stroke-related mortality were identified.

In the table, the feature with the highest impact on classification accuracy is marked by number 1, and the weakest feature influence is marked by number 9.

The models' feature importance scores were evaluated using the technique outlined below:

- Random Forest and XGBoost use the Gini index;
- Logistic Regression uses absolute coefficient values;
- SVM and MLP use permutation importance.

Table 4. Feature importance scores across models applied

	LR	*RF*	*XG Boost*	*SVM*	*MLP Classifier*	*Average score*
Health_Status	2	3	2	1	1	1.8
Age	1	1	9	4	2	3.4
Stroke_Symptoms	5	2	4	2	6	3.8
Health_complications	10	4	1	5	3	4.6
Smoke	6	6	3	3	5	4.6
Stroke_Type	4	7	5	6	4	5.2
Treatment_methods	3	5	6	9	8	6.2
Days_till_Hospital	8	9	7	8	9	8.2
Past_Stroke	7	10	8	7	10	8.4
Gender	9	8	10	10	7	8.8

The ‘Average Score’ refers to the mean rank assigned to each feature across all models. As we can derive from the column *Average score*, the *Health Status*, *Age* and *Stroke Symptoms* are the most critical features, dominating all models and suggesting that patient-specific health factors provide the most valuable information for mortality prediction.

Health Complications and Smoke features are mid-level predictors, reflecting the nuanced differences in stroke severity and related complications.

Past Stroke and *Gender* are the least important features, suggesting that historical factors and demographic information, while important, are overshadowed by more immediate clinical indicators in determining stroke-related mortality.

These insights can help guide clinical focus towards key factors influencing stroke outcomes, aiding in prioritising treatment and care for high-risk patients.

Gender and *Past Stroke* were identified as the least significant variables, suggesting that acute clinical indicators are more significant in determining stroke-related mortality than demographic data and historical determinants.

Hence, directing clinician attention toward important variables affecting stroke outcomes can assist in prioritising high-risk patients' care and treatment.

5.4. Threshold-Based Evaluation of Random Forest Model Performance

Given that the Random Forest model demonstrated superior performance in predicting stroke outcomes relative to the other models tested, its behaviour was further analysed with a focus on the impact of varying the classification threshold. In stroke mortality prediction, machine learning models do not produce categorical labels directly, like “alive” or “dead.” Instead, they generate a probability between 0 and 1 for each case. Normally, we use a threshold of 0.5, meaning if the predicted probability of death is above 0.5, the model predicts that the patient will die. However, this default threshold may not always be the best choice - especially in healthcare, where failing to identify a true mortality case (false negative) can be more detrimental than a false positive prediction. To better understand model behaviour, we evaluated the performance of the Random Forest model across a range of classification thresholds. Specifically, we looked at three metrics: precision – how many of the predicted deaths were actually correct, recall – how many of the actual deaths the model was able to catch, accuracy – the overall rate of correct predictions.

This type of threshold-based analysis is useful for several reasons:

- *Customising for clinical priorities*: in some clinical settings, it may be more important to catch every possible high-risk patient (high recall), while in others it might be better to avoid false alarms (high precision);
- *Improving model understanding*: by seeing how precision, recall, and accuracy change with the threshold, we get a clearer picture of the model’s strengths and limitations;

- *Better decision-making*: this analysis helps determine the most appropriate threshold for different clinical goals — whether we want to screen broadly or intervene selectively.

Using Python's *matplotlib.pyplot* library, we plotted a graph (see Figure 3) that shows how precision, recall, and accuracy behave as the threshold increases from 0.20 to 0.80.

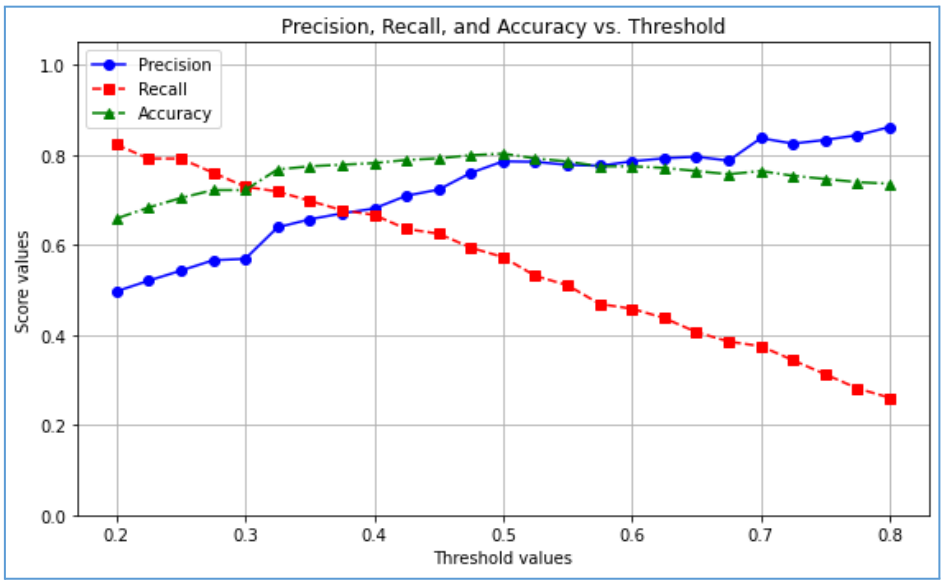

Figure 3. Random Forest model performance across different threshold values

As shown in the figure, precision increases consistently as the threshold rises from 0.20 (0.497) to 0.80 (0.862). As we raise the threshold, the model becomes more careful in predicting death.

This behaviour is useful when we want to avoid false positives, for example, before assigning intensive treatments or issuing alerts that might worry patients or families.

Recall drops steadily from 0.823 at threshold 0.20, to 0.260 at threshold 0.80. This shows that as we raise the threshold, the model becomes too cautious - and starts missing real death cases. At lower thresholds, the model captures more actual mortality cases but at the cost of increased false positive predictions. At higher thresholds, the model fails to detect a greater number of actual deaths, as it only predicts mortality when the probability is sufficiently high. This trade-off is critical in stroke care, where missing a high-risk patient can have serious consequences.

Accuracy follows a parabolic trend, increasing up to a threshold of 0.50—where it peaks at 0.803—after which it gradually declines. Up to 0.50, the model becomes more balanced between false positives and false negatives, which improves its overall accuracy. Beyond 0.50, even though precision continues to rise, recall falls too much, which also pulls accuracy down.

This means that while 0.50 gives the best balance, it might not align with clinical priorities, especially if catching all death risks is a higher priority than overall correctness.

Table 5 illustrates how this threshold behaviour translates into real-world clinical decision-making:

Table 5. Best Threshold options

Clinical Goal	Recommended Threshold	Why?
Early warning / screening	0.30 – 0.40	Higher recall (up to 0.73), catching more true deaths
Balanced decision-making	0.50	Best overall accuracy (0.803) and solid precision
High-confidence intervention	0.70 – 0.80	Very high precision (above 0.83), fewer false positives

Therefore, the optimal classification threshold should be selected according to the specific priorities of the clinical context. Using lower thresholds when not missing any deaths is the top priority. Conversely, higher thresholds are preferable when decision-making requires a high degree of certainty prior to initiating critical interventions. Use the middle ground (like 0.50) when you need a balanced, all-purpose approach.

This kind of threshold analysis doesn't just improve model performance — it helps bridge the gap between AI predictions and real medical decisions, making the model more useful in practical, patient-focused settings.

6. Conclusion

In the present study, we used a dataset of clinical and demographic features from the neurology department of the Clinical Centre of Montenegro to assess the predictive power of five binary classification machine learning models: Random Forest, XGBoost, Support Vector Machine (SVM), Neural Networks (MLPClassifier), and Logistic Regression for stroke-related mortality. By evaluating these models on important performance criteria, including accuracy, precision, recall, and F1-score, we were able to learn more about their predictive abilities and overall reliability. To determine the most important factors influencing mortality forecasts, we also looked at the significance of attributes.

In terms of accuracy and interpretability, our findings show that ensemble-based models such as Random Forest and XGBoost fared better than alternative approaches. These models emphasised the significance of health status, age, and stroke symptoms in predicting stroke outcomes by repeatedly identifying them as the most significant predictors. Although neural networks performed competitively, especially in precision and recall, the model's interpretability issues continue to be a drawback in clinical settings where knowing the underlying causes is essential.

Notably, despite their lower accuracy, simpler models such as Logistic Regression provide more accurate insights into feature relevance, which could make them useful in situations where interpretability and transparency are crucial. On the other hand, SVM struggled with generalisability across various test set sizes, although producing excellent results for particular metrics.

To further examine the practical applicability of our best-performing model, we conducted a threshold-based evaluation of the Random Forest classifier. By analysing how key metrics: precision, recall, and accuracy change across different classification thresholds, we revealed important trade-offs relevant to clinical decision-making. Lower thresholds improved recall and supported broader screening strategies, while higher thresholds significantly boosted precision, making them suitable for high-confidence decisions. The analysis confirmed that threshold selection can be tuned to align with different clinical priorities, enhancing both flexibility and usability of the model in real-world settings.

Despite the promising results, this study presents several notable limitations. The dataset was derived from a single clinical center, which may limit the generalisability of the findings to other populations or healthcare settings. Additionally, the relatively small sample size increases the risk of overfitting, particularly in complex models like neural networks. External validation on larger, multi-center datasets is essential to confirm the robustness of the proposed models.

Finally, our results demonstrate the applicability of machine learning models in predicting stroke-related mortality, with Random Forest and XGBoost emerging as the most reliable choices. These models are appropriate for practical clinical applications due to their high performance and capacity to interpret feature importance. By enabling the early identification of high-risk patients and directing more individualised treatment plans, this study highlights the potential of data-driven approaches to enhance stroke management.

Statements and Declarations

Data Availability

The data file is uploaded to https://github.com/Virgilijus11/StrokeData.git

Funding Statement

This research received no financial support or specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Approval

This study does not involve the collection of any personal information or data. Therefore, there are no ethical issues or controversies in this study.

Contributions

The first author applied machine learning algorithms and performed stroke database analysis. With the second author's help, he wrote the manuscript and formulated the conclusions. The second author organised the database according to our needs, validated and visualised the results, and performed a literature analysis with a reference list. All authors read and approved the manuscript.

Acknowledgements

The authors would like to extend gratitude to the Clinical Centre of Montenegro, operating in Podgorica, for collecting and sharing data used for research.

Reference

Cheon, S., Kim, J., & Lim, J. (2019). The use of deep learning to predict stroke patient mortality. *International Journal of Environmental Research and Public Health, 16*(11), 1876. https://doi.org/10.3390/ijerph16111876

Chung, C. C., Su, E. C., Chen, J. H., Chen, Y. T., & Kuo, C. Y. (2023). XGBoost-based simple three-item model accurately predicts outcomes of acute ischemic stroke. *Diagnostics, 13*(5), 842. https://doi.org/10.3390/diagnostics13050842

Deljavan, R., Farhoudi, M., & Sadeghi-Bazargani, H. (2018). Stroke in-hospital survival and its predictors: The first results from Tabriz Stroke Registry of Iran. *International Journal of General Medicine, 11*, 233–240. https://doi.org/10.2147/IJGM.S158296

Egegamuka, N., Ekedebe, N., Ajoku, K., Okafor, C., & Ozor, C. (2024). Development of random forest model for stroke prediction. *International Journal of Innovative Science and Research Technology, 9*, 2783–2795. https://doi.org/10.38124/ijisrt/IJISRT24APR2566

Feng, Y. (2023). Support vector machine for stroke risk prediction. *Highlights in Science, Engineering and Technology, 38*, 917–923. https://doi.org/10.54097/hset.v38i.5977

Fernandes, J. N. D., Cardoso, V. E. M., Comesaña-Campos, A., & Pinheira, A. (2024). Comprehensive review: Machine and deep learning in brain stroke diagnosis. *Sensors, 24*(13), 4355. https://doi.org/10.3390/s24134355

Fernandez-Lozano, C., Hervella, P., Mato-Abad, V., Rodríguez-Yáñez, M., Suárez-Garaboa, S., López-Dequidt, I., Estany-Gestal, A., Sobrino, T., Campos, F., Castillo, J., Rodríguez-Yáñez, S., & Iglesias-Rey, R. (2021). Random forest-based prediction of stroke outcome. *Scientific Reports, 11*, Article 10071. https://doi.org/10.1038/s41598-021-89434-7

Krikščiunienė, D., & Sakalauskas, V. (2022). Overview of the artificial intelligence methods and analysis of their application potential. In D. Krikščiunienė & V. Sakalauskas (Eds.), *Intelligent systems for sustainable person-centred healthcare* (Vol. 205, pp. 167–183). Springer. https://doi.org/10.1007/978-3-030-79353-1_9

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv*. https://doi.org/10.48550/arXiv.1705.07874

Sakalauskas, V., Krikščiunienė, D., Ognjanovic, I., & Sendelj, R. (2022). Time-to-event modelling for survival and hazard analysis of stroke clinical case. In *Business Information Systems Workshops: BIS 2021*, 14–27. *Springer*. https://doi.org/10.1007/978-3-031-04216-4_2

Someeh, N., Mirfeizi, M., Asghari-Jafarabadi, M., Alinia, S., Farzipoor, F., & Shamshirgaran, S. M. (2023). Predicting mortality in brain stroke patients using neural networks: outcomes analysis in a longitudinal Study. *Scientific Reports, 13*(1). https://doi.org/10.1038/s41598-023-45877-8.

Wang, L. (2023). Logistic regression for stroke prediction: An evaluation of its accuracy and validity. *Highlights in Science, Engineering and Technology, 39*, 1086–1092. https://doi.org/10.54097/hset.v39i.6712

Wang, R., Zhang, J., Shan, B., He, M., & Xu, J. (2022). XGBoost machine learning algorithm for prediction of outcome in aneurysmal subarachnoid hemorrhage. *Neuropsychiatric Disease and Treatment, 18*, 659–667. https://doi.org/10.2147/NDT.S349956

World Stroke Organization. (2019). *WSO annual report 2019*. https://www.world-stroke.org/assets/downloads/WSO_2019_Annual_Report_online.pdf

Zhang, H. (2023). Stroke prediction based on support vector machine. *Highlights in Science, Engineering and Technology, 31*, 53–59. https://doi.org/10.54097/hset.v31i.4812