Indonesian Journal of Engineering, Science and Technology
IJENSET
Volume 03 - Issue 01, June 2026
e-ISSN: 3089-2392
Journal Link: https://jurnal.umla.ac.id/index.php/ijenset/index

STROKE RISK PREDICTION USING CATBOOST WITH AN EXPLAINABLE ARTIFICIAL INTELLIGENCE APPROACH

Khairul Umam[1]*, M Sadewa Wicaksana Wibowo[2]

[1] *Computer Engineering; Faculty of Science Technology and Education; Universitas Muhammadiyah Lamongan; Lamongan (62218), Indonesia*
[2] *Medical Informatics; Faculty of Health Sciences; Universitas Muhammadiyah Lamongan; Lamongan (62218), Indonesia*

*Corresponding Author Email: khairulumam744@gmail.com

Article Information

Submitted	:	April, 25 2026
Revised	:	May, 18 2026
Accepted	:	May, 20 2026
Paper page	:	12-22
DOI	:	Xxx

ABSTRACT

Stroke is among the main causes of death worldwide. According to the World Health Organization (WHO), strokes, including ischemic and hemorrhagic, account for around 11% of global mortality. Therefore, early prediction is crucial as part of efforts to prevent the risk of stroke and to assist healthcare professionals in clinical decision-making. This work aims to develop a stroke risk prediction model using the CatBoost algorithm, and to interpret the prediction results using an Explainable Artificial Intelligence (XAI) approach through the SHAP method. The CatBoost model's evaluation results demonstrate strong performance, with AUC = 0.98, an F1-score = 0.91, precision = 0.92, recall = 0.90, and accuracy of 0.93. Furthermore, the XAI analysis utilizing SHAP showed that the CatBoost model not only delivers highly accurate predictions but also successfully identifies the most relevant features leading to stroke risk, namely age, body mass index (BMI), and mean level of glucose. Finally, a comparative examination with various different machine learning models demonstrates that the CatBoost model obtains the best performance and is extremely useful in predicting stroke risk.

Keywords: *CatBoost; Explainable Artificial Intelligence (XAI) and SHAP (Shapley Additive Explanations); Machine Learning; Stroke Risk Prediction.*

I. INTRODUCTION

Stroke is one of the leading causes of death worldwide. The World Health Organization (WHO) reports that approximately 11% of global deaths are attributable to this condition, making it the second leading cause of death globally (World Health Organization, 2023). The increasing incidence of stroke is generally driven by unhealthy lifestyle factors, including smoking, alcohol consumption, an unbalanced diet, the use of certain medications, and poorly controlled blood glucose levels (Feigin et al., 2025). Stroke is a medical condition that arises when the blood flow to a region of the brain is blocked or substantially diminished. As a result, oxygen and essential nutrients cannot be adequately delivered to brain cells, leading to cellular death within minutes. This condition may cause permanent damage to brain tissue over time and, in severe cases, can result in death (Campbell, B. C. V., Khatri, 2020).

Stroke is classified into two main types: ischemic stroke, caused by a blood clot blocking blood flow to the brain, and hemorrhagic stroke, resulting from the rupture of a weakened blood vessel leading to bleeding in the brain tissue. According to the WHO, approximately 87% of stroke cases are attributed to these two types (Johnson et al., 2016). Therefore, preventive efforts are of paramount importance. One effective approach is the early prediction of stroke risk, enabling the timely identification of symptoms and facilitating appropriate medical intervention.

In recent years, technological advancements have accelerated the adoption of machine learning in the healthcare domain, particularly for disease prediction. Recent research indicates that algorithms based on machine learning can improve the accuracy of stroke prediction when compared to conventional methods, because of their capacity to capture complex patterns within clinical data. A study by Ding and Chen demonstrates that machine learning–based models, such as Random Forest and other algorithms, are capable of delivering strong predictive performance in identifying stroke risk (Ding & Chen, 2024).

Furthermore, recent studies have emphasized that boosting methods demonstrate superior performance compared to traditional models. A systematic review reports that boosting-based algorithms can achieve accuracy levels exceeding 90% in stroke prediction (Adekunle et al., 2025). This evidence suggests that boosting-based approaches constitute an effective solution for addressing the complexity inherent in medical data.

One of the widely used boosting algorithms for tabular data classification is CatBoost. CatBoost is a gradient boosting–based algorithm designed to handle categorical features directly without requiring complex encoding processes, while also mitigating overfitting through the use of ordered boosting (Prokhorenkova et al., 2018). Furthermore, CatBoost employs a symmetric tree structure, which increases model stability and efficiency during the training phase. It has been shown to outperform other boosting algorithms, such as Gradient Boosting and XGBoost, particularly in handling categorical data and reducing prediction bias. These advantages make CatBoost an effective algorithm for a wide range of classification tasks, including medical datasets that are often characterized by complexity and non-linearity.

Despite its strong predictive performance, CatBoost still has limitations in terms of interpretability. Therefore, an Explainable Artificial Intelligence (XAI) approach is necessary to gain insights into the model's prediction process. One of the most commonly employed methods is SHapley Additive exPlanations (SHAP), which measures and explains the contribution of each feature to the

model's prediction output. Recent studies indicate that features such as age, glucose level, and hypertension are the primary factors influencing stroke risk based on SHAP analysis (Tang, 2026). This aligns with the demands in the medical field, where it is essential not only to produce accurate predictions but also to ensure that the results are interpretable by healthcare professionals. In addition, other studies have demonstrated that machine learning models combined with techniques for handling imbalanced data, such as SMOTE, along with SHAP-based interpretation, can enhance predictive performance and provide more meaningful clinical insights in identifying high-risk patients (Haidar et al., 2026).

Based on the background, this study aims to develop a stroke risk prediction model using CatBoost and to interpret the prediction results through an XAI approach employing the SHAP method.

It is expected that this research will contribute to the development of disease prediction systems that not only achieve high performance but are also transparent and interpretable.

II. METHOD

In this study, several stages are undertaken to predict stroke occurrence. These stages include the collection of a stroke dataset, data preprocessing, the development of a CatBoost model, and the training and testing of the model to classify whether a patient is at risk of stroke or not. Subsequently, the model's performance is evaluated, and a comparative analysis is conducted between the CatBoost model and other machine learning algorithms. The overall workflow of this research is illustrated in Figure 1, which presents the methodological flowchart adopted in this study.

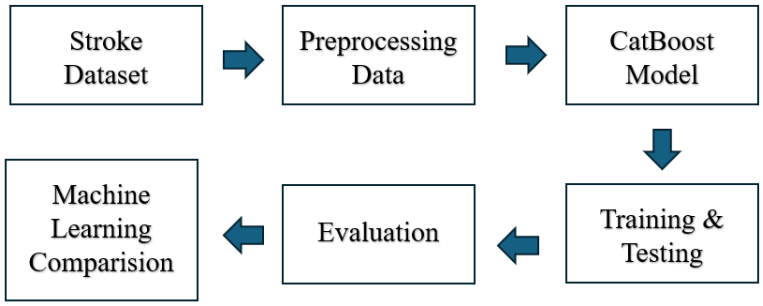

Figure 1. Block Diagram

A. Stroke Dataset

In this study, the stroke dataset was obtained from Kaggle, where the data were originally collected by a medical clinic in Bangladesh. This dataset contains a set of relevant patient information used to analyse whether an individual exhibits symptoms of stroke or not. Figure 2 presents an example of the dataset, illustrating patient information records.

The dataset contains 5,110 patient records and comprises 12 attributes, namely: id, gender, age, hypertension, heart disease, ever married, work type, residence type, average glucose level, BMI, smoking status, and stroke status. The stroke variable is used as the target label, where a value of 1 represents the occurrence of stroke, while a value of 0 indicates no stroke.

Based on the overall observation, 249 patients are identified as having a history of stroke, while 4,861 patients are classified as non-stroke cases.

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
2	51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
3	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
4	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
5	1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
6	56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
7	53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
8	10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
9	27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
10	60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
11	12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12	12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
13	12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
14	8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
15	5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
16	58202	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1
17	56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
18	34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1

Figure 2. Picture of Stroke Dataset

B. Data Preprocessing

The prepared dataset undergoes a data preprocessing stage prior to model training, with the objective of identifying and correcting underlying patterns and potential issues within the data. Based on an initial examination of the stroke dataset, several preprocessing steps are required, including the following:

1. Handling Missing Values

In the stroke dataset, several instances contain missing values (NaN), which may adversely affect model quality and reduce performance during the learning process. Therefore, an appropriate strategy for handling missing data is required. In this study, missing values are addressed using an imputation method based on the mean value of each column with incomplete data. This approach is selected due to its simplicity and effectiveness in preserving the overall data distribution while minimizing the loss of information from the dataset.

2. Data Encoding (Label Encoding)

In the stroke dataset, several features are represented as string or categorical variables. Such data types cannot be directly utilized by machine learning models; therefore, a transformation process is required. In this study, categorical features are converted into numerical (integer) values using an encoding technique. This process ensures that the data can be effectively processed by the selected algorithms without losing the essential information contained within the original features.

3. Synthetic Minority Over-sampling Technique (SMOTE)

The stroke dataset is characterized by class imbalance, in which the number of patients diagnosed with stroke is considerably lower than the number of patients without stroke. This class imbalance may cause the model to be biased toward the majority class, which can subsequently reduce its capability to correctly identify stroke cases.

To solve this issue, this study uses SMOTE, which generates synthetic samples for the minority class by interpolating between current data points and their nearest neighbours, resulting in a more balanced data distribution.

The SMOTE procedure begins by calculating the distance between instances in the minority class, followed by determining the desired oversampling percentage and selecting the number of nearest neighbours.

Subsequently, synthetic data are generated using the following formulation (Ifls, 2021):

$$x_{syn} = x_i + (x_{knn} - x_i)\delta \quad (1)$$

where:

- x_{syn} = Synthetic data
- x_i = Minority class instance
- x_{knn} = Nearest neighbor
- δ = A random value between 0 and 1

C. CatBoost Model

CatBoost is a gradient boosting–based machine learning algorithm developed by the Russian technology company Yandex. The model is specifically designed to enhance performance on tabular data, particularly datasets with many categorical features. CatBoost operates through a sequential learning process, in which each new model is constructed to correct the errors made by its predecessors. This approach employs gradient descent–based optimization on a defined loss function, enabling the model to iteratively minimize prediction errors and improve overall performance.

In the study by Liudmila Prokhorenkova et al., CatBoost introduces the ordered boosting technique, which effectively reduces prediction shift and mitigates overfitting commonly observed in traditional boosting algorithms. Furthermore, CatBoost includes a specific strategy for managing categorical features through target statistics, removing the requirement for manual encoding approaches such as one-hot encoding. The study demonstrates that CatBoost achieves competitive, and in some cases superior, performance compared to other boosting algorithms such as XGBoost and LightGBM, particularly when applied to datasets with a large number of categorical features (Prokhorenkova et al., 2018).

D. Training and Testing

In this study, the development of the CatBoost model is conducted through two main stages: training and testing. These stages aim to enable the model to learn underlying patterns within the data and perform accurate predictions. Furthermore, they serve as a basis for evaluating the model's performance on previously unseen data.

1. Training stage

This stage involves training the model on the training dataset, during which it learns the underlying patterns and relationships between the input features and the target variable, specifically the occurrence of stroke.

2. Testing stage

This stage involves evaluating the model using a testing dataset that was not utilized during the training process. The objective is to assess the model's ability to generalize and perform accurately on unseen data.

E. Evaluation

The evaluation stage aims to measure and analyze the performance of the developed model. This process is conducted during the testing phase using a testing dataset to assess the model's ability to accurately and consistently predict whether a patient is at risk of stroke.

The evaluation is carried out by comparing the model's predictions to the actual values (ground truth) in the test data. This step evaluates the model's ability to extrapolate information to previously unseen data.

In this study, several evaluation metrics are employed to assess the model's performance in predicting stroke, as follows:

1. Confusion Matrix

The confusion matrix is an evaluation tool used to assess the performance of a classification model by comparing the predicted results with the actual values (ground truth) (Moulaei et al., 2024). It provides a

comprehensive overview of how well the model performs across different classes.

The confusion matrix is made up of four major components: True Positive (TP), which represents correctly predicted positive instances; True Negative (TN), which represents correctly predicted negative instances; False Positive (FP), which represents negative instances that are incorrectly classified as positive; and False Negative (FN), which represents positive instances that are incorrectly classified as negative.

2. Accuracy

Accuracy is an evaluation metric used to measure the overall correctness of a model's predictions (Dritsas & Trigka, 2022). The formula used to calculate accuracy is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

3. Precision

Precision is used to measure the proportion of correctly predicted positive instances among all instances predicted as positive (Dritsas & Trigka, 2022). The formula used to calculate precision is as follows:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

4. Recall

Recall is an evaluation metric used to measure a model's ability to correctly identify all positive instances. In the medical domain, recall is particularly important as it reflects the model's capability to detect patients who are truly at risk of a disease (Dritsas & Trigka, 2022). The formula for calculating recall is as follows:

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

5. F1-Score

This evaluation metric represents the harmonic mean of precision and recall, and is used to assess the balance between the two (Dritsas & Trigka, 2022). The formula used to calculate this metric is as follows:

$$F1_{Score} = 2 * \frac{Precision\,.\,Recall}{Precision + Recall} \quad (5)$$

6. AUC (Area Under The Curve)

AUC (Area Under the Curve) is used to measure a model's ability to distinguish between positive and negative classes (Dritsas & Trigka, 2022). It is determined as the area under the Receiver Operating Characteristic (ROC) curve, which shows the relationship between the true positive (TP) and the false positive (FP) across various threshold values.

7. Explainable Artificial Intelligence (XAI)

In this study, in addition to employing conventional evaluation metrics, an Explainable Artificial Intelligence (XAI) approach is utilized to enhance model transparency. The method applied is SHAP (SHapley Additive Explanations), a model interpretation technique grounded in game theory, which is used to quantify the contribution of each feature to the prediction outcome. SHAP computes feature contributions by considering all possible combinations of features, thereby providing a consistent and fair interpretation of the model's predictions (Moulaei et al., 2024).

F. Machine Learning Model Comparison

At this stage, several machine learning models are introduced as benchmarks to compare their performance with the CatBoost model employed in this study. The models used for comparison are as follows:

1. Logistic Regression

Logistic Regression is a statistical classification method designed for binary classification tasks, which has also been extended to handle multiclass problems (Tazin et al., 2021).

2. AdaBoost (*Adaptive Boosting*)

AdaBoost is a boosting algorithm first introduced by Yoav Freund and Robert Schapire. It aims to build a model iteratively by adjusting the weights of training data at each iteration, allowing the model to focus more on instances that are difficult to classify (Rahman et al., 2023).

3. K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is a non-parametric method that classifies data based on the majority class among its nearest neighbors (Rahman et al., 2023). This model has been widely applied in various previous studies.

4. Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) is a type of Artificial Neural Network (ANN) consisting of an input layer, one or more hidden layers, and an output layer. This model can capture complex and non-linear relationships within the data. It is trained using the backpropagation algorithm to minimize prediction error (Heart et al., 2023).

5. Extreme Gradient Boosting (XGBoost)

XGBoost is an advanced implementation of the gradient boosting algorithm that incorporates optimization and regularization techniques to enhance model performance (Heart et al., 2023). It is widely used in machine learning competitions due to its ability to produce accurate and efficient models.

III. RESULT AND DISCUSSION

The research process follows the block diagram presented in Figure 1. In practice, the initial stage involves collecting data from the Kaggle platform, specifically the Stroke Dataset. The dataset is then analysed and proceeds to the next stage, namely data pre-processing. At this stage, several techniques are applied before the data are fed into the model. The first technique is handling missing values to address incomplete data. The second technique is data encoding, which converts categorical (string) variables into numerical (integer) representations. The third technique is the application of SMOTE to solve the class imbalance between the two classes in the stroke dataset. These pre-processing steps ultimately result in a correlation heatmap, as illustrated in Figure 3.

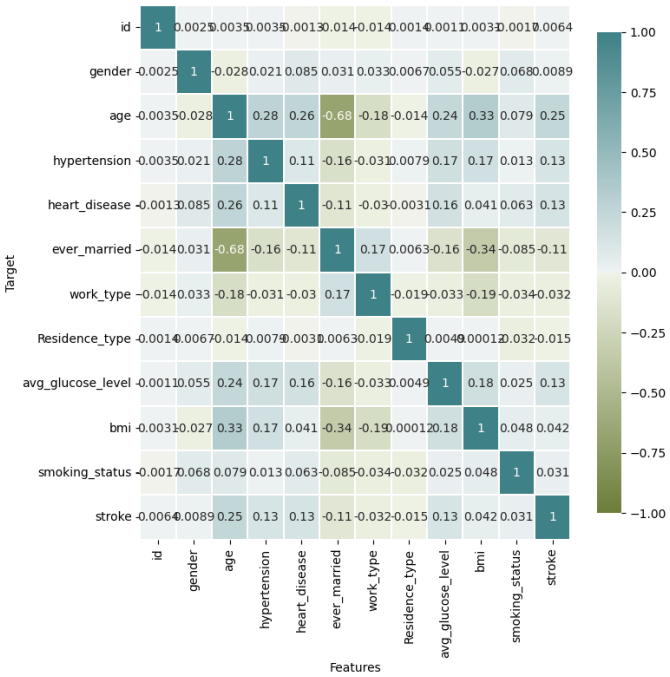

Figure 3. Correlation Heatmap

Figure 3 presents the correlation heatmap generated for each feature after the data preprocessing stage. The correlation values range from −1 to +1, where positive values indicate a direct relationship and negative values indicate an inverse relationship. Values closer to 1 or −1 signify stronger relationships between variables, while values near zero indicate weak correlations.

Subsequently, the dataset is divided into two subsets: 80% for training and 20% for testing. The training data are used to train the model to learn underlying patterns and characteristics for predicting stroke, whereas the testing data are used to evaluate the trained model's performance in predicting stroke risk in patients.

Table 1. Parameter Model

MODEL	ITERATIONS	LEARNING RATE	DEPTH
CatBoost	500	0.05	6

The third stage of the block diagram involves preparing the CatBoost model, which is trained using the training dataset with parameters specified as shown in Table 1.

Subsequently, the predefined parameters and the CatBoost model are utilized in the next stage, namely the training process. During training, the model is configured with 500 iterations, a learning rate of 0.05, and a depth of 6. The implementation is carried out using Python 3 and executed on a system equipped with an Intel® Core™ i5-10500H processor and an NVIDIA GTX GPU with 4GB memory.

After the training phase, the model is evaluated using the testing dataset to assess its performance on unseen data. The testing process produces several evaluation metrics, as presented in Table 2. The evaluation is conducted using AUC, F1-score, precision, recall, and accuracy. The CatBoost model achieves an AUC of 0.98, an F1-score of 0.91, a precision of 0.92, a recall of 0.90, and an accuracy of 0.93.

Table 2. Catboost Model Evaluation Results

MODEL	AUC	F1_SCORE	PRECISION	RECALL	ACCURACY
CatBoost	0,98	0,91	0,92	0,90	0,93

Based on Table 2, all evaluation metrics are computed from the confusion matrix, which represents the prediction results of the CatBoost model, as illustrated in Figure 4.

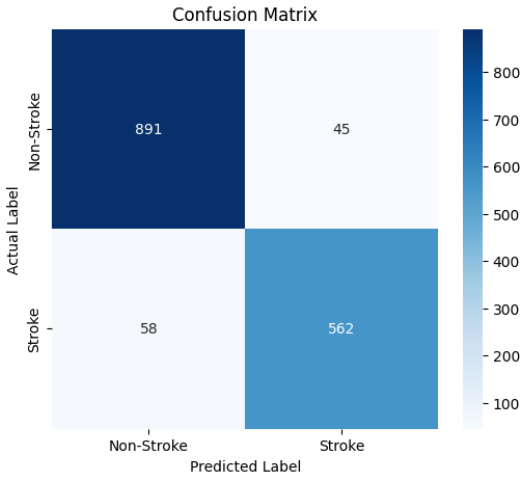

Figure 4. The Confusion Matrix of the Catboat Classification Model

The results of the confusion matrix indicate that the model correctly classified 891 non-stroke instances as True Negatives and 562 stroke instances as True Positives. This demonstrates that the model has strong performance in recognizing both classes.

However, several misclassifications were also observed. There were 45 non-stroke cases incorrectly predicted as stroke (False Positives), indicating that the model tends to produce a certain degree of over-alerting in some instances. In addition, 58 stroke cases were incorrectly predicted as non-stroke (False Negatives). This type of error is particularly critical in the medical context, as false negatives represent cases where patients who

experienced a stroke are not detected by the system, which may lead to serious consequences.

In addition to the evaluation metrics described above, this study also employs XAI using the SHAP method. This approach is utilized to provide insights into how each factor contributes to the model's predictions in determining stroke risk. Figure 5 presents a summary graph of the SHAP method, which shows the distribution of each feature. In this plot, the horizontal axis represents SHAP values, where positive values indicate that a feature increases the likelihood of predicting stroke, while negative values indicate a decrease in that likelihood. The color of each point represents the feature value, with red indicating high values and blue indicating low values.

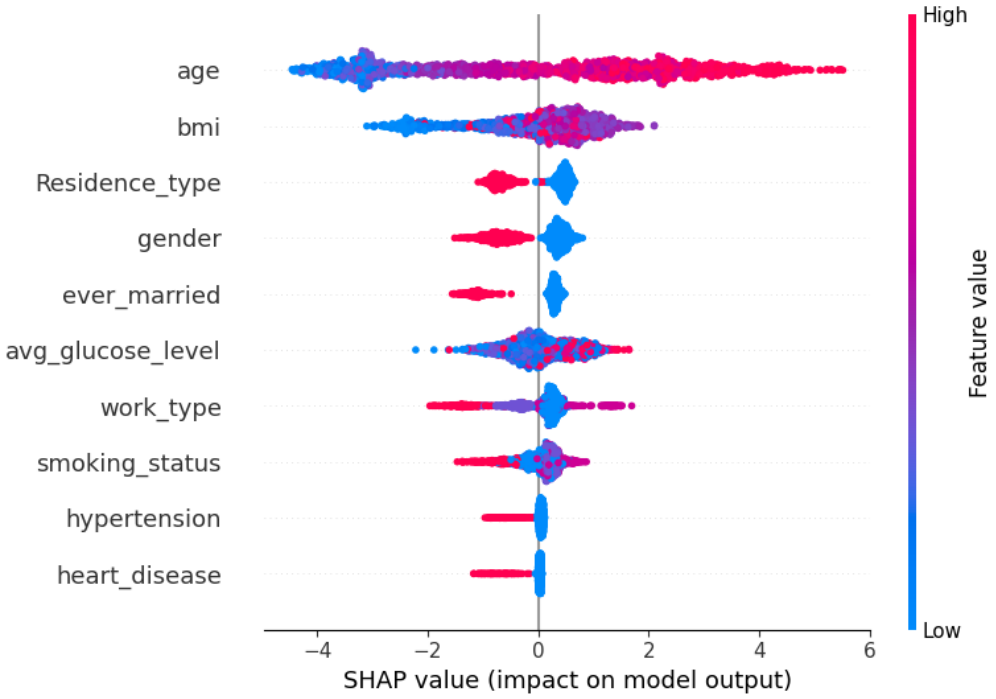

Figure 5. SHapley Additive exPlanations (SHAP) Summary Plot for the CatBoost Classification Model

The analysis results indicate that several features have varying degrees of influence on stroke prediction within the CatBoost model.

The most influential feature is age. This variable represents the dominant factor affecting the model's predictions. High SHAP values for this feature indicate that increasing age is strongly associated with a higher risk of stroke. This is evidenced by the concentration of red-coloured points on the right side of the plot, which represents higher feature values contributing positively to stroke prediction.

Another important feature is Body Mass Index (BMI), which also shows a notable influence on stroke prediction. Higher BMI values tend to contribute positively to the likelihood of stroke, although its impact is less pronounced compared to age.

Furthermore, average glucose level (avg_glucose_level) also contributes to the model's predictions. Elevated glucose levels are associated with an increased probability of stroke, as reflected by the positive SHAP value distribution.

In contrast, several features exhibit relatively minor contributions, including Residence type, gender, ever_married, work_type, hypertension, heart_disease, and smoking_status. These variables show comparatively small SHAP value distributions, indicating limited individual impact on the model's predictions compared to the three dominant features discussed above.

Nevertheless, these features may still provide meaningful contributions in specific cases and contribute to the model's overall predictive performance when considered collectively.

The final stage of this study involves comparing the performance of the proposed model with other machine learning models used in previous research. The objective is to identify the most effective model in terms of predictive performance. The comparison results are presented in Table 3.

Table 3. Comparative Evaluation of Machine Learning Models

NO	MODEL	AUC	F1-SCORE	PRECISION	RECALL	ACCURACY
1	Logistic Regression	0.88	0.72	0.72	0.72	0.79
2	AdaBoost	0.89	0.73	0.72	0.75	0.80
3	MLP	0.90	0.79	0.75	0.84	0.84
4	KNN	0.95	0.86	0.78	**0.97**	0.88
5	XGBoost	**0.98**	0.90	0.90	0.90	**0.93**
6	CatBoost	**0.98**	**0.91**	**0.92**	0.90	**0.93**

Table 3 presents a comparative evaluation of several machine learning models for stroke risk prediction. The models compared with CatBoost include Logistic Regression, AdaBoost, Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), and XGBoost.

The results indicate that boosting-based models generally outperform the other approaches. Logistic Regression achieved AUC = 0.88, F1-score = 0.72, accuracy = 0.79, and both precision and recall of 0.72.

Meanwhile, AdaBoost demonstrated improved performance with an AUC of 0.89 and an accuracy of 0.80. The neural network model (MLP) further improved the results, achieving an AUC of 0.90, an F1-score of 0.79, and a relatively high recall of 0.84.

The KNN model showed strong performance with an AUC of 0.95 and an F1-score of 0.86. Notably, it achieved the highest recall value of 0.97, indicating its strong capability in identifying nearly all stroke cases. This is particularly important in the medical domain, as it helps reduce false negative cases. In addition, KNN is computationally efficient in terms of execution time.

Finally, the best-performing models are the boosting algorithms, namely XGBoost and CatBoost. Both models achieved the highest AUC value of 0.98. XGBoost obtained an accuracy of 0.93, while CatBoost also reached an accuracy of 0.93, with the highest F1-score of 0.91 and a precision of 0.92. This indicates that CatBoost provides a well-balanced trade-off between precision and recall.

Overall, the CatBoost model demonstrates the best performance in stroke prediction, as evidenced by key evaluation metrics such as AUC, F1-score, precision, and accuracy.

IV. CONCLUSION

This study aims to predict stroke risk using the CatBoost model combined with a performance evaluation framework. Several stages were conducted, including data preprocessing, model development using CatBoost, training, testing, and evaluation. In addition, an Explainable Artificial Intelligence (XAI) approach was applied using the SHAP (SHapley Additive exPlanations) method. Finally, a comparative analysis with other machine learning models was performed.

The results show that the CatBoost model achieved AUC = 0.98, F1-score = 0.91, a precision = 0.92, a recall = 0.90, and an accuracy of 0.93. Furthermore, the SHAP analysis indicates that the model not only provides high predictive performance but also identifies the most influential features associated with stroke risk. Age, BMI, and average glucose level

emerged as the dominant predictors in the model, which is consistent with established medical knowledge regarding stroke risk factors.

Finally, a comparison with other machine learning models for stroke risk prediction was conducted. The results demonstrate that CatBoost outperforms the other models across several evaluation metrics, including AUC, F1-score, precision, and accuracy. This indicates that CatBoost is an effective and reliable model for stroke risk prediction.

ACKNOWLEDGEMENT

The author would like to thank the team that collaborated on completing this article. We also thank the Ijenset team for giving us the opportunity to publish this article.

REFERENCES

Adekunle, A., Aderinto, N., Racheal, M., Adeyanju, I. A., Osonuga, A., & Olawade, D. B. (2025). International Journal of Medical Informatics Machine learning techniques for stroke prediction : A systematic review of algorithms , datasets , and regional gaps. *International Journal of Medical Informatics*, *203*(July), 106041. https://doi.org/10.1016/j.ijmedinf.2025.106041

Campbell, B. C. V., Khatri, P. (2020). Stroke. *The Lancet*, *396*(10244), 129–142. https://doi.org/10.1016/S0140-6736(20)31179-X

Ding, Y., & Chen, H. (2024). *An exploration on the machine-learning-based stroke prediction model*. *April*, 1–8. https://doi.org/10.3389/fneur.2024.1372431

Dritsas, E., & Trigka, M. (2022). *Stroke Risk Prediction with Machine Learning Techniques*.

Feigin, V. L., Brainin, M., Norrving, B., Martins, S. O., Pandian, J., Lindsay, P., Grupper, M. F., & Rautalin, I. (2025). *World Stroke Organization : Global Stroke Fact Sheet 2025*. *20*(2). https://doi.org/10.1177/17474930241308142

Haidar, N., Alsalman, K., Al-ghraibah, A., Maisharah, S., Ghadzi, S., & Looi, I. (2026). *Prediction of recurrent ischemic stroke using machine learning from real-world data*. *4*.

Heart, E., Prediction, D., Machine, U., & Techniques, L. (2023). *Learning Techniques*.

Ifls, C. S. (2021). *SMOTE and Nearmiss Methods for Disease Classification with Unbalanced Data*. 305–314.

Johnson, W., Onuma, O., & Sachdev, S. (2016). *Stroke : a global response is needed*.

Moulaei, K., Afshari, L., Moulaei, R., & Sabet, B. (2024). *Explainable artificial intelligence for stroke prediction through comparison of deep learning and machine learning models*. 1–16.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). *CatBoost : unbiased boosting with categorical features*. *Section 4*, 1–11.

Rahman, S., Hasan, M., & Sarkar, A. K. (2023). *Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques*.

Tang, M. (2026). *Explainable machine learning for stroke risk prediction : a comparative study with SHAP-based interpretation*. *January*, 1–12. https://doi.org/10.3389/fneur.2025.1716984

Tazin, T., Alam, N., Dola, N. N., Bari, M. S., Bourouis, S., & Khan, M. M. (2021). *Stroke Disease Detection and Prediction Using Robust Learning Approaches*. *2021*. https://doi.org/10.1155/2021/7633381

World Health Organization. (2023). *Global Health Estimates*. https://www.who.int/data/global-health-estimates