Journal of Cardiovascular Disease Research

ISSN:0975 -3583,0976-2833 VOL14, ISSUE 11, 2023

Use of Robust Machine Learning Approach in Prediction of Stroke

Mr. Dilesh Yuvraj Bagul[1], Dr. P. B. Bharate[2], Dr. Aarti Sahasrabuddhe[3]

1Statistician, Index Medical College, Hospital & Research Centre, Indore, Madhya Pradesh;
2Professor, Department of Statistics, Malwanchal University, Indore, Madhya Pradesh;
3Professor, Index Medical College, Hospital & Research Centre, Indore, Madhya Pradesh

Corresponding Authors: Mr. Dilesh Yuvraj Bagul
Email id: - dileshbagul@gmail.com.

ABSTRACT

Background: Machine Learning transforms healthcare by analyzing varied data to predict conditions like stroke, aiding in prevention. Stroke, a major cause of disability and death, stems from abrupt brain blood flow interruption, underscoring the need for early risk detection. This study aimed to explore the correlation between socio-demographic, clinical, and lifestyle elements with stroke while employing machine learning techniques to predict stroke occurrence based on assessed risk factors.

Methods: A case-control study involving 1360 individuals (50% stroke patients) gathered patient data using medical records. Statistical methods (Chi-square, correlation, t-test, Wilcoxon rank) were used to explore stroke associations. Four machine learning algorithms (Decision Tree, Naïve Bayes, Random Forest, and Logistic Regression) were applied to build a predictive stroke model, evaluated by measures like sensitivity, specificity, and F1 score for performance assessment.

Results: The study identified stark variations between stroke and non-stroke groups in various health indicators: BMI, fasting blood glucose, triglycerides, total cholesterol, LDL Chol/HDL ratio (6.00 vs. 3.00), and LDL/HDL ratio (3.34 vs. 1.56). Conversely, HDL and VLDL levels were notably lower in stroke cases: (43.28 vs. 59.93) and (8.50 vs. 42.68), respectively, with no significant differences observed in age and HbA1C. Among the Machine Learning algorithms employed, the random forest displayed exceptional performance, achieving accuracy, precision, sensitivity, specificity, F1-score, and area under the curve of 92.09%, 92.11%, 90.21%, 93.64%, 91.15%, and 91.93%, respectively, considering all attributes.

Conclusion: This study indicates that Machine Learning models can predict stroke occurrence using patient data, including sociodemographic, medical history, and lifestyle factors, potentially enabling future stroke probability predictions based on risk factors and medical consultations.

Keywords: Machine learning, stroke prediction, health care, risk assessment, accuracy, intervention.

INTRODUCTION

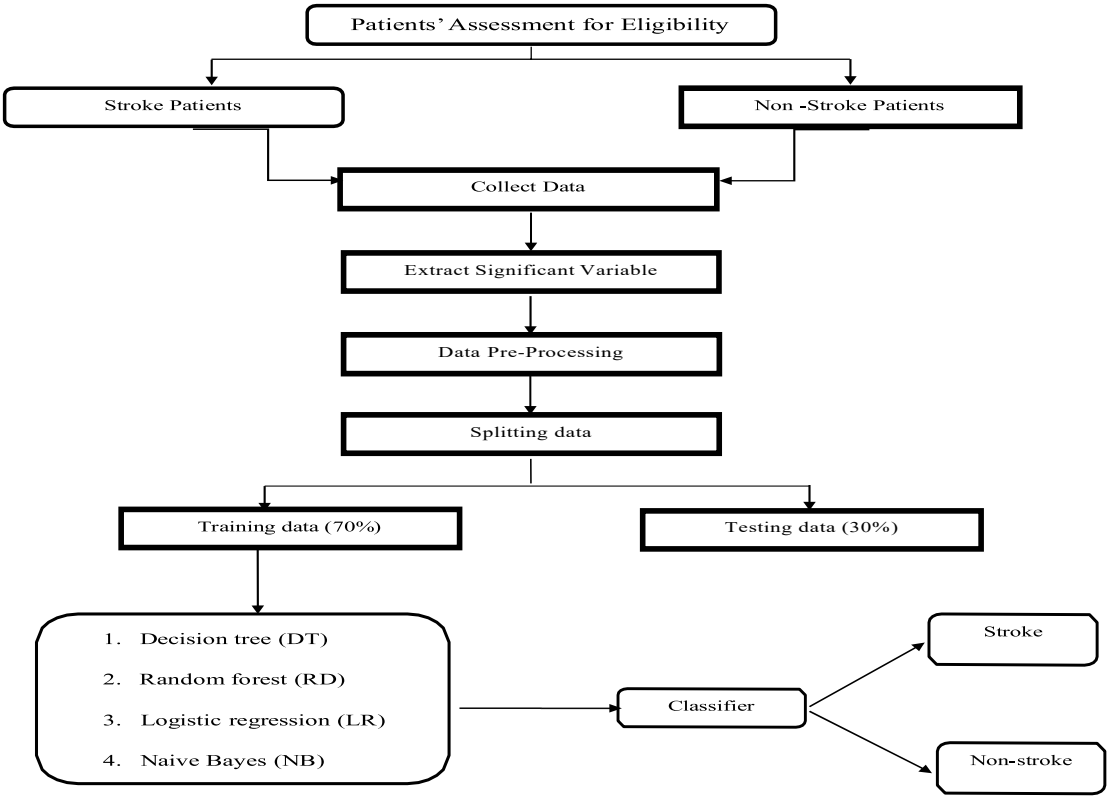
Machine Learning (ML) has Significantly impacted the fields of statistics and healthcare by providing powerful tools to analyze complex data and make accurate predictions, based on their medical history, and sociodemographic and lifestyle factors. Algorithms included in ML are beneficial in processing and analyzing large datasets to make predictions of various lifestyle diseases including stroke. Stroke is a critical neurological condition that occurs due to sudden interruption of blood flow to the brain and is one of the main causes of disability and death worldwide. It is crucial to identify individuals who're at risk to plan preventive measures and provide timely intervention thus this study is an attempt to predict the occurrence of stroke by application of Machine Learning Techniques.

Stroke or cerebrovascular accident (CVA) is an acute compromise of the cerebral blood supply. The American Heart Association/American Stroke Association (AHA/ASA) provides a basic definition of Stroke as: “An acute episode of focal neurologic dysfunction lasting longer than 24 hours”. There are two types of strokes: Ischemic and Hemorrhagic. Ischemic strokes are due to blood vessel blockage limiting the blood supply to the brain. In contrast, hemorrhagic strokes are due to blood vessel rupture leading to blood spilling into the intracranial cavity. [1] According to World Health Organization (WHO) around 15 million people globally suffer from strokes each year, with one person dying every four to five minutes. India is among the top five countries in terms of death & disability due to stroke. [2,3]

Early detection and appropriate management are crucial to prevent brain damage and other complications. Stroke occurrence relates to various modifiable and non-modifiable risk factors, including age, gender, hypertension, diabetes, smoking, dyslipidemia, and family history.[4] Advancements in medical technology enable the anticipation of stroke onset through Machine Learning (ML) techniques. Various ML methods like Decision Trees, Random Forest, Logistic Regression, and Naïve Bayes aid in reliable stroke prediction based on diverse parameters.[5] These methods offer the capability to integrate multiple variables and complex patterns to enhance predictive accuracy. By utilizing advanced Machine Learning techniques, this study aims to provide a proactive and precise approach to identify individuals at risk of stroke before the onset of symptoms.

ISSN:0975 -3583,0976-2833 VOL14, ISSUE 11, 2023

MATERIALS AND METHODOLOGY



Flow chart of study

A case-control study was employed as the study design conducted at a tertiary care hospital of Madhya Predesh, India. The population of this study was patients above 18 years old both male and female who attended tertiary care hospitals. The Stroke patients diagnosed with standard clinical criteria were selected and the non-stroke group included suspected patients whose final diagnosis was observed as negative for stroke. The Purposive sampling technique for the selection of study participants was used. The sample size of 1360 was calculated by ensuring a shrinkage factor of 0.9 and at 10% Anticipated R^2 for 13 Predictors [10].

Statistical Analysis

Statistical analysis was conducted using R 4.3.1 software. The Pearson correlation coefficient estimated correlations between continuous variables, and the Point biserial correlation determined correlations with binary responses. The Phi correlation coefficient measured binary variable associations. The chi-square test assessed categorical variable relationships with stroke. To verify normality, the Shapiro-Wilk test was employed for continuous variables. Group differences were evaluated via t-tests for normally distributed variables such as age, BMI, fasting blood glucose levels, HbA1c, Total Cholesterol, Triglyceride, HDL, LDL, VLDL, Chol/HDL ratio, and LDL/HDL ratio. A correlation matrix assessed multicollinearity among predictors.

Model Building:

After completing the essential steps of data preparation and dataset management, the next pivotal stage in our analysis was the construction of predictive models. To ensure a robust assessment of these models, the dataset was randomly divided into two portions: 70% for training and 30% for testing. Subsequently, we proceeded to train four distinct machine learning models utilizing the training dataset, aiming to predict the occurrence of stroke.

Evaluation Matrix:

Once these models were trained, we turned our attention to evaluating their performance. To accomplish this, we employed the reserved testing dataset to assess how well each model could predict stroke. Through this rigorous evaluation process, we were able to identify the model that demonstrated the most effective predictive capabilities for stroke occurrence. This step is of paramount importance as it aids in selecting the most suitable model for our specific analysis and its future applications Evaluations of model performance were done using matrices such as Sensitivity, Specificity, and Area under the receiver operating characteristic curve.

ISSN:0975 -3583,0976-2833 VOL14, ISSUE 11, 2023

RESULTS:

In this study, there was a significant association observed between gender and stroke occurrence. It was found that approximately 63.97 % of males had a stroke, while in females, the rate was 36.03%. There was a significant association observed between residency and strokes. In rural areas, 85.59% of individuals were found to have a stroke, compared to 14.41% in urban areas. Moreover, physical activity and occupation were observed significant relationships with stroke occurrence. Almost (99.85%) of married individuals were found to have strokes, whereas only 0.15% of unmarried individuals had strokes. As far as occupation is concerned (33.09%) housewives were found to have a higher stroke rate as show in Table 1.

Table 1: Comparison of Nominal variables in stroke and Non-Stroke group

Variables	Categories	Stroke	Percentage	Non stroke	Percentage	P-value	Significance
Gender	Male	435	(63.97 %)	320	(47.06%)	0.00	Sig
	Female	245	(36.03 %)	360	(52.94%)		
Resident	Rural	582	(85.59 %)	608	(89.41%)	0.033	Sig
	Urban	98	(14.41 %)	72	(10.59%)		
Ever Married	Yes	679	(99.85 %)	677	(99.56%)	0.317	**NS**
	No	1	(0.15 %)	3	(0.44%)		
Occupation	Private	156	(22.94 %)	119	(17.5%)	0.00	Sig
	Government	135	(19.85%)	105	(15.44%)		
	Self Employed	116	(17.06%)	102	(15%)		
	House wife	225	(33.09%)	324	(47.65%)		
	No	48	(7.06%)	30	(4.41%)		
Physical activity	Yes	98	(14.41%)	605	(88.97%)	0.00	Sig
	No	582	(85.59%)	75	(11.03%)		
Alcohol	Yes	183	(26.91%)	128	(18.82%)	0.00	Sig
	No	497	(73.09%)	552	(81.18%)		
Smoking	Yes	99	(14.56%)	51	(7.5%)	0.00	Sig
	No	581	(85.44%)	629	(92.5%)		
Diabetes	Yes	560	(82.35%)	266	(39.12%)	0.00	Sig
	No	120	(17.65%)	414	(60.88%)		
Hypertension	Yes	546	(80.29%)	215	(31.62%)	0.00	Sig
	No	134	(19.71%)	465	(68.38%)		
Drug defaulter Hypertensive	Yes	56	(8.24%)	3	(0.44%)	0.00	Sig
	No	490	(72.06%)	212	(31.18%)		
Heart disease	Yes	61	(8.97%)	11	(1.62%)	0.00	Sig
	No	619	(91.03%)	669	(98.38%)		
Mental illness	Yes	2	(0.29%)	0	(0%)	0.157	NS
	No	678	(99.71%)	680	(100%)		
Renal Disease	Yes	0	(0%)	0	(0%)	-	-
	No	680	(100%)	680	(100%)		
Road traffic accident (Head injury)	Yes	4	(0.59%)	2	(90.29%)	0.413	**NS**
	No	676	(99.41%)	678	(99.71%)		
Liver Disease	Yes	2	(0.29%)	12	(1.76%)	0.007	Sig
	No	678	(99.71%)	668	(98.24%)		

Table 2: Comparison of continuous Variables in Stroke and Non-Stroke group

Continuous variables	Stroke (Mean ± SD)	Non-Stroke (Mean ± SD)	P-value	Sig.
Age	52.98 ± 8.87	52.81 ± 8.86	**0.69**	**NS**
BMI	32.28 ± 5.24	23.93 ± 3.60	0.00	Sig.
Fasting Blood Glucose Level	178.34 ± 48.98	130.33 ± 53.52	0.00	Sig
HbA1C	5.50 ± .59	5.49 ± 0.60	**0.71**	**NS**
Triglyceride	371.88 ± 100.89	145.06 ± 11.01	0.00	Sig
Total Cholesterol	230.24 ± 40.48	179.10 ± 12.49	0.00	Sig

ISSN:0975 -3583,0976-2833 VOL14, ISSUE 11, 2023

HDL	43.28 ± 11.96	59.93 ± 2.64	0.00	Sig
LDL	127.82 ± 27.00	90.82 ± 8.32	0.00	Sig
VLDL	8.50 ± 2.42	42.68 ± 4.92	0.00	Sig
Chol/HDL Ratio	6.00 ± 5.14	3.00 ± 0.21	0.00	Sig
LDL/HDL	3.34 ± 3.12	1.56 ± 0.50	0.00	Sig

Table 2 is showing comparison of health indicators between stroke and non-stroke group. Stroke group had significantly higher BMI (32.28 vs. 23.93), fasting blood glucose (178.34 vs. 130.33), triglycerides (371.88 vs. 145.06), total cholesterol (230.24 vs. 179.10), LDL (127.82 vs. 90.82), Chol/HDL ratio (6.00 vs. 3.00), and LDL/HDL ratio (3.34 vs. 1.56). Conversely, HDL, VLDL levels were significantly lower (43.28 vs. 59.93), (8.50 vs. 42.68) respectively in stroke cases. Age and HbA1C shows no significant difference.

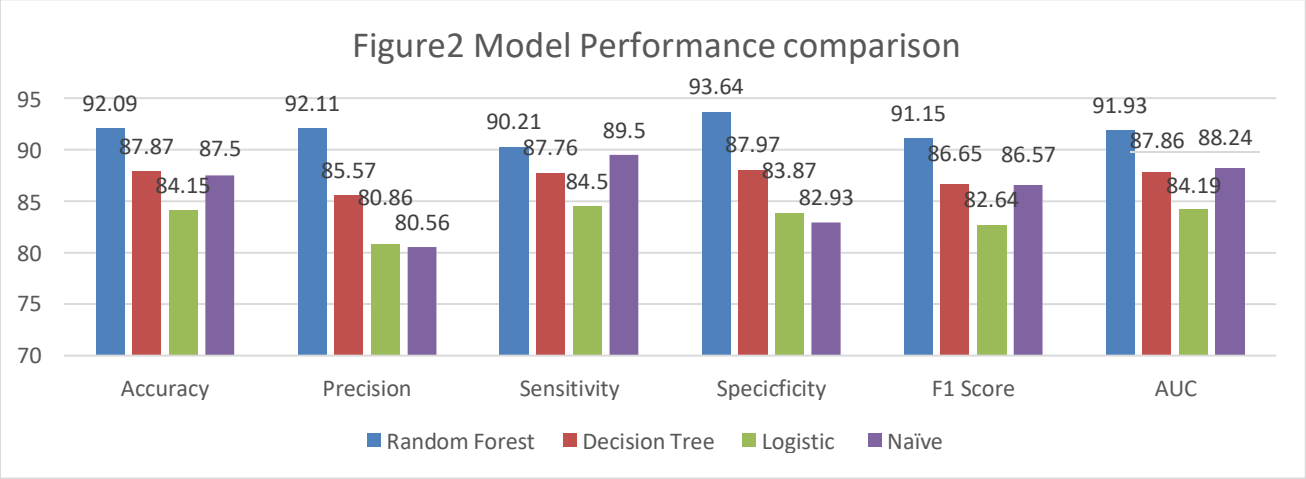


Figure 2 shows the performance comparison regarding effectiveness in predicting strokes. Random Forest method demonstrated the highest accuracy of 92.09%, with notable precision (92.11%), sensitivity (90.21%), and specificity (93.64%). The Decision Tree method was found to have slightly less accuracy of (87.87%). The Logistic Regression and Naïve Bayes method observed the accuracy of 84.15% and 87.5% respectively, highlighting their competence in stroke prediction. These insights assist in selecting the most suitable model for stroke prediction tasks.

DISCUSSION

The comparison of nominal variables in the stroke and non-stroke groups revealed substantial differences in various parameters. For instance, the analysis showed a significant difference in gender distribution, with 435 males (63.97%) in the stroke group compared to 320 males (47.06%) in the non-stroke group. Similar trends were observed in residence, where 582 (85.59%) stroke cases were from rural areas compared to 608 (89.41%) in the non-stroke group. Variables like occupation, physical activity, alcohol consumption, and health conditions (diabetes, hypertension, etc.) also showed significant variations between the groups, reinforcing their potential role in stroke occurrences.

Continuous variables, such as BMI, fasting blood glucose, and lipid profiles, demonstrated significant differences between the stroke and non-stroke groups. For instance, the mean BMI was notably higher in the stroke group (32.28 ± 5.24) compared to the non-stroke group (23.93 ± 3.60), this was similar to findings of study conducted by Shiozawa *et al*. [6] The fasting blood glucose levels were considerably elevated in the stroke group (178.34 ± 48.98) compared to the non-stroke group (130.33 ± 53.52) and this is consistent with results of study published by Zhang *et al*. [7] Other parameters like triglycerides, total cholesterol, and HDL/LDL ratios showed similar significant variations, underscoring their potential as risk indicators for stroke which is similar with study done by Liu *et al*. [8]

In evaluating the predictive models for stroke occurrences, the Random Forest model demonstrated the highest accuracy at 92.09%, outperforming other models such as Decision Tree (87.87%), Logistic Regression (84.15%), and Naïve Bayes (87.5%). Similar results were found in study done by Shobayo *et al*. [9] This illustrates the superior predictive capacity of Random Forest in accurately identifying stroke cases based on the given attributes, ensuring precise and reliable predictions.

These findings highlight the pivotal role of both categorical and continuous variables in predicting stroke occurrences. The significant variations observed in lifestyle, health conditions, and physiological markers underline their importance as potential risk factors for stroke. Moreover, the high accuracy of the Random Forest model signifies its efficacy in predicting stroke occurrences based on these factors, offering promising avenues for enhanced preventive strategies and

ISSN:0975 -3583,0976-2833 VOL14, ISSUE 11, 2023

personalized healthcare. In contrast, clirindza J *et al* (2022) achieved better predictive accuracy and AUC with Naive bayes model.
The study's comprehensive analysis suggests the criticality of considering a multitude of factors in stroke prediction. Integrating machine learning predictive models, particularly Random Forest, in clinical settings could substantially improve risk assessment, early intervention, and personalized preventive measures to mitigate the prevalence of stroke cases.

However, further validation and wider application are essential to cement these findings as essential tools for risk assessment and personalized care in stroke management. The combination of these tables underscores the complexity of stroke determinants and the potential power of machine learning in effectively predicting and preventing stroke occurrences.

CONCLUSION

This study highlights that gender and residence-based disparities in stroke prevalence. Individuals with specific health conditions, notably diabetes and hypertension, show a higher incidence of strokes. Also, Machine learning models, especially Random Forest, exhibit strong predictive performance for stroke prediction, offering valuable tools for informed healthcare decision-making. Understanding these patterns and leveraging predictive models can enhance stroke management and preventive strategies.

In this study we also evaluated the relationship of various risk factors with stroke occurrence using different machine learning models. The important predictors found in the study like Diabetes, hypertension, BMI, dyslipidemia, gender, residence etc. Apart from that a new risk factor i.e., "drug defaulter hypertensive patients" has emerged as a significant predictor. The study results indicate that the Random Forest technique demonstrated superior performance compared to other machine learning models tested.

Limitations:

Our data was sourced from a single tertiary care hospital, which potentially limited the generalizability. Notably, the dataset exhibited an imbalanced class distribution, which can potentially introduce bias in the study's outcomes.

Conflict of Interest:

The authors assert that they have no conflicts of interest related to the publication of this paper.

REFERENCES:

1. Sacco RL, Kasner SE, Broderick JP, Caplan LR, Connors JJB, Culebras A, *et al*. An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. 2013 Jul;44(7):2064–89.
2. Jones SP, Baqai K, Clegg A, Georgiou R, Harris C, Holland EJ, *et al*. Stroke in India: A systematic review of the incidence, prevalence, and case fatality. International Journal of Stroke. 2022 Feb 1;17(2):132–40.
3. Centers for Disease Control and Prevention [Internet]. 2023 [cited 2023 Jul 10]. Stroke Facts | cdc.gov. Available from: https://www.cdc.gov/stroke/facts.htm
4. Murphy SJ, Werring DJ. Stroke: causes and clinical features. Medicine (Abingdon). 2020 Sep;48(9):561–6.
5. Khan M. Stroke Disease Detection and Prediction Using Robust Learning Approaches. Journal of Healthcare Engineering. 2021 Nov 26;2021:1–12.
6. Shiozawa M, Kaneko H, Itoh H, Morita K, Okada A, Matsuoka S, Kiriyama H, Kamon T, Fujiu K, Michihata N, Jo T, Takeda N, Morita H, Nakamura S, Node K, Yasunaga H, Komuro I. Association of Body Mass Index with Ischemic and Hemorrhagic Stroke. Nutrients. 2021 Jul 9;13(7):2343.
7. Zhang Y, Gu S, Wang C, Liu D, Zhang Q, Yang M, Zhou Z, Zuo H. Association between fasting blood glucose levels and stroke events: a large-scale community-based cohort study from China. BMJ Open. 2021 Aug 18;11(8):e050234.
8. Liu X, Yan L, Xue F. The associations of lipids and lipid ratios with stroke: A prospective cohort study. J Clin Hypertens (Greenwich). 2019 Jan;21(1):127-135. doi: 10.1111/jch.13441. Epub 2018 Nov 21. PMID: 30461182; PMCID: PMC8030570.
9. Shobayo O, Zachariah O, Odusami MO, Ogunleye B. Prediction of Stroke Disease with Demographic and Behavioural Data Using Random Forest Algorithm. *Analytics*. 2023; 2(3):604-617.
10. Gangavarapu Sailasya and Gorli L Aruna Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms" International Journal of Advanced Computer Science and Applications (IJACSA), 12(6), 2021.