Stroke Diagnosis With Machine Learning

CAE407 Data Science
Bachelor of Science in Computer Science
(Year III, Semester I)

RESEARCHER(S)

Leela Balav Sharma (12200011)
Pema Tshering (12200013)
Rinchen Pelzang (12200015)
Thinley Wangchuk (12190030)

GUIDED BY

Mr. Yonten Jamtsho

Gyalpozhing College of Information Technology
Gyalpozhing : Mongar

1 Problem Statement

Stroke is a medical emergency that occurs due to the interruption of flow of blood to a part of brain because of bleeding or blood clots. Worldwide, it is the second major reason for deaths with an annual mortality rate of 5.5 million. Every year, more than 15 million people worldwide have a stroke, and in every 4 minutes, someone dies due to stroke. It causes the unpredictable death and damage to multiple body components. If a stroke is identified early enough, it is possible to receive the appropriate therapy and recover from the stroke.

As a result, this project work attempts to develop a stroke prediction system to assist doctors and clinical workers in predicting strokes in a timely and efficient manner. The proposed machine learning algorithm helps in predicting whether or not a person is suffering from a stroke. Our system utilizes the medical records of a patients to determine the presence or absence of stroke.

1.1 Aims

The aim of the project is to predict whether the person is suffering from a stroke or not.

1.2 Scope and Limitations

1.2.1 Scope

- We used a machine learning algorithm to predict the stroke that helps the patient to take the medical treatment and they can avoid the risk of stroke.
- Reduced morbidity results from timely stroke prediction, which enables neurologists to identify high-risk patients and direct treatment strategies.
- Prediction of a stroke is essential, and it must be treated promptly to prevent death or irreversible damage.

The "healthcare-dataset-stroke-data" is a stroke prediction dataset from Kaggle that contains 5,110 observations(rows) with 12 attributes(columns). Each observation corresponds to one patient, and the attributes are variables about the health status of each patient. The model prepared with this will be more applicable towards white men from America. Generally, a system that helps to predict Stroke can be very handy to doctors, patients, and clinical workers.

1.2.2 Limitations

- No Stroke datasets available in Bhutan
- Limited time to collect our own data.
- Lack of interpretability and reproducibility because of using others dataset

2 Literature Review

Research has been done that looks at how Stroke Prediction is done using Machine Learning algorithms as below:

The research community[1] has shown great interest in developing tools and methods for monitoring and predicting stroke disease that have a significant impact on human health. It presents the latest works that utilize machine learning techniques for stroke risk prediction. They have applied four machine learning algorithms, such as naive Bayes, J48, K-nearest neighbor and random forest, in order to detect accurately a stroke. The accuracy of the naive Bayes classifier was 85.6%, whereas the accuracy for J48, K-nearest neighbor and random forest was 99.8%. logistic regression, naive Bayes, Bayesian network, decision tree, neural network, random forest, bagged decision tree, voting and boosting model with decision trees were applied in order to classify stroke risk levels. The experiment results showed that the boosting model with decision trees achieved the highest accuracy. Hence, various methodology is proposed in order to find out the various symptoms associated with the stroke disease and preventive measures for a stroke from social media resources. They defined a system to predict where a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status.

This research paper[2] provides a study of various risk factors to understand the probability of stroke. It used a regression-based approach to identify the relation between a factor and its corresponding impact on stroke. For our study, we use a dataset of electronic health records released by McKinsey Company as a part of their healthcare hackathon challenge.2 The dataset is available from Kaggle, a public data repository for datasets. The dataset contains the EHR records of 29072 patients. They demonstrate that identifying the important features impacts the final performance of machine learning framework. It is important for us to identify the perfect combination of features, instead of using all the available features in the feature space. A redundant attributes and/or totally irrelevant attributes to a class should be identified and removed before the use of a classification algorithm. Therefore, it is essential for data mining practitioners in healthcare to identify how the risk factors captured in electronic health records are inter-dependent, and how they impact the accuracy of stroke prediction independently.

The goal of this paper[3] is to review the contribution of ML in solving some stroke-related problems (prevention/risk factor identification, diagnosis, treatment and prognostication). There are several sub types of machine learning algorithms, however in this research they have focused on supervised learning, unsupervised learning and deep learning are focused types in the current study. Supervised learning trains a model that maps an input to an output based on observations. It considers the most systematic approach because it reviews best of studies and writes a result and discussion with some interesting insights, get from all studies which are definitely valuable for further investigation. In this paper, the researchers are focusing mainly to determine where a particular person has stroke or not using logistion regression. Symptoms can sometimes develop slowly and sometimes it can develop quickly. It is even possible for someone to wake up while sleeping with symptoms. So, the classification model has been built to predict the risk of stroke or to automatically diagnose stroke, using features such as lifestyle factors or radiological imaging to minimize the risk of a patient by giving them awareness in advance.

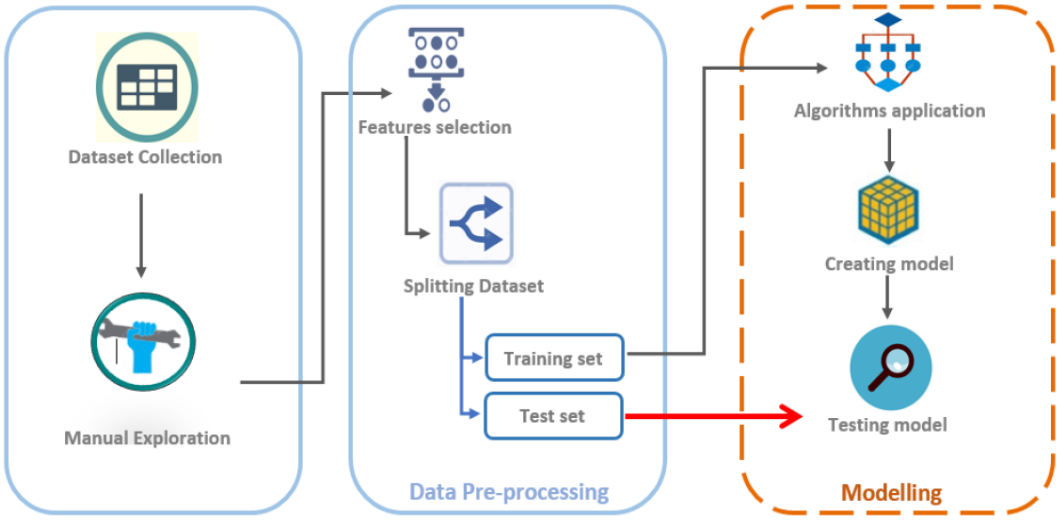
This study[4] investigated the applicability of machine learning techniques to predict long-term outcomes in ischemic stroke patients. This was a retrospective study using a prospective cohort that enrolled patients with acute ischemic stroke. Favorable outcome was defined as modified Rankin Scale score 0, 1, or 2 at 3 months. They developed three machine learning models (deep neural network, random forest, and logistic regression) and compared their predictability to choose a best one. To evaluate the accuracy of the machine learning models, comparision was done with the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score to evaluate their effectiveness. A total of 2604 patients were included in this study, and 2043 (78%) of them had favorable outcomes. The area under the curve for the deep neural network model was significantly higher than that of the ASTRAL score (0.888 versus 0.839; P¡0.001), while the areas under the curves of the random forest (0.857; P=0.136) and logistic regression (0.849; P=0.413) models were not significantly higher than that of the ASTRAL score. Hence, the paper concluded that machine learning algorithms, particularly the deep neural network, can improve the prediction of long-term outcomes in ischemic stroke patients.

This conferencing[5] focused on assessing the predictive value of classification schemes that estimate stroke risk in patients with Atrial fibrillation. Patients who have atrial fibrillation (AF) have an increased risk of stroke, but their absolute rate of stroke depends on age and comorbid conditions. The CHADS2 was formed by assigning 1 point each for the presence of congestive heart failure, hypertension, age 75 years or older, and diabetes mellitus and by assigning 2 points for history of stroke or transient ischemic attack. Main outcome measure was hospitalization for ischemic stroke, determined by Medicare claims data. They concluded that the 2 existing classification schemes and especially a new stroke risk index, CHADS2, can quantify risk of stroke for patients who have atrial fibrillation and may aid in selection of antithrombotic therapy.

This paper[6] is based on the study of Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults where the aims of this study was to compare Cox models, machine learning (ML), and ensemble models combining both approaches, for prediction of stroke risk in a prospective study of Chinese adults. The research includes the sample from all age where all the participants provided a blood sample, and random blood glucose tests were conducted to screen for diabetes. All stroke cases were verified and adjusted by trained medical staff using the International Classification of Diseases. All the participants were randomly assigned to a training set (85%; 174,498 men with 16 649 strokes and 253,766 women with 20,100 strokes), a validation set (12.75%; 26,174 men with 2,467 strokes and 38,065 women with 3014 strokes), and test set (2.25%; 4,620 men with 471 strokes and 6,718 women with 533 strokes), with all subsequent analyses performed separately by sex. Among several approaches, an ensemble model combining both Gradient boosted tree(GBT) and Cox models achieved the best performance for identifying individuals at high risk of stroke in a contemporary study of Chinese adults. The results highlight the potential value of expanding the use of machine learning algorithms in clinical practice.

3 Methodoloy

3.1 Machine Learning Workflow



3.2 Stroke Dataset

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
2	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
3	51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
4	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
5	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
6	1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
7	56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
8	53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
9	10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
10	27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
11	60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12	12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
13	12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
14	12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
15	8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
16	5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
17	58202	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1
18	56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
19	34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
20	27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smoked	1
21	25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1
22	70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
23	13861	Female	52	1	0	Yes	Self-employed	Urban	233.29	48.9	never smoked	1
24	68794	Female	79	0	0	Yes	Self-employed	Urban	228.7	26.6	never smoked	1
25	64778	Male	82	0	1	Yes	Private	Rural	208.3	32.5	Unknown	1
26	4219	Male	71	0	0	Yes	Private	Urban	102.87	27.2	formerly smoked	1
27	70822	Male	80	0	0	Yes	Self-employed	Rural	104.12	23.5	never smoked	1
28	38047	Female	65	0	0	Yes	Private	Rural	100.98	28.2	formerly smoked	1
29	61843	Male	58	0	0	Yes	Private	Rural	189.84	N/A	Unknown	1
30	54827	Male	69	0	1	Yes	Self-employed	Urban	195.23	28.3	smokes	1

3.3 Steps involved in Machine Learning Project

1. **Dataset Collection**
   The process of gathering and measuring information from countless different sources to solve business problem at hand.

2. **Data Pre-processing**
   Cleaning and Organizing the raw data to make it suitable for a building and training Machine Learning models.

3. **Feature Selection**
   Selecting input variable to your model by using only relevant data and getting rid of noise in data.

4. **Train-Test Split**
   Splitting the dataset into training and test sets.

5. **Model Selection**
   Choosing one among many candidate models for a predictive modeling problem.

6. **Model Training**
   Machine learning algorithm is fed with sufficient training data to learn from for future predictions.

7. **Model Evaluation**
   Testing the trained model to see if it would operate well in real world problems.

8. **Hyperparameter Tuning**
   Processing of finding the optimized values for hyperparameters, which maximizes your model's predictive accuracy.

9. **Model Deployment**
   Integrating a machine learning model into an existing production environment to make practical business decisions based on data.

References

[1] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," *Sensors*, vol. 22, no. 13, p. 4670, 2022.

[2] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," *Healthcare Analytics*, vol. 2, p. 100032, 2022.

[3] M. S. Sirsat, E. Fermé, and J. Câmara, "Machine learning for brain stroke: a review," *Journal of Stroke and Cerebrovascular Diseases*, vol. 29, no. 10, p. 105162, 2020.

[4] I. R. Konig, A. Ziegler, E. Bluhmki, W. Hacke, P. M. Bath, R. L. Sacco, H. C. Diener, and C. Weimar, "Predicting long-term outcome after acute ischemic stroke: a simple index works in patients from controlled clinical trials," *Stroke*, vol. 39, no. 6, pp. 1821–1826, 2008.

[5] B. F. Gage, A. D. Waterman, W. Shannon, M. Boechler, M. W. Rich, and M. J. Radford, "Validation of clinical classification schemes for predicting stroke: results from the national registry of atrial fibrillation," *Jama*, vol. 285, no. 22, pp. 2864–2870, 2001.

[6] M. Chun, R. Clarke, B. J. Cairns, D. Clifton, D. Bennett, Y. Chen, Y. Guo, P. Pei, J. Lv, C. Yu *et al.*, "Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million chinese adults," *Journal of the American Medical Informatics Association*, vol. 28, no. 8, pp. 1719–1727, 2021.