Analyzing the Performance of Stroke Prediction Using Different Models

Bhargav Patel,
B.E Student,
Department of Computer Engineering,
Aditya Silver Oak Institute of Technology,
Ahmedabad, India

Mr. Singh Manish S.,
Associate Professor,
Department of Computer Engineering,
Aditya Silver Oak Institute of Technology,
Ahmedabad, India

Abstract - A stroke is a medical condition in which the blood arteries in the brain rupture, leading brain damage. Symptoms may appear if the brain's flow of blood and other nutrients is disrupted. Stroke is the leading cause of death and disability worldwide, according to the World Health Organization (WHO). Early awareness of the numerous stroke warning symptoms can assist to lessen the severity of the stroke. To forecast the chance of a stroke happening in the brain, many machine learning (ML) models have been created. Random forest, support vector machines, decision trees and classifiers, and neural networks were the most often utilised approaches.

Key Words: **Stroke prediction, Machine learning approaches, Neural Network, Random Forest, Comparison Analysis.**

1. Introduction

When blood flow to different parts of the brain is interrupted or reduced, the cells in those areas of the brain don't get the nutrients and oxygen they need, and they die. A stroke is a life-threatening medical condition that need immediate medical intervention. To avoid additional damage to the afflicted area of the brain, as well as effects in other parts of the body, early identification and effective therapy are necessary. According to the World Health Organization (WHO), fifteen million people worldwide suffer from strokes each year, with one person dying every four to five minutes. According to the Centers for Disease Control and Prevention(CDC), stroke is the sixth greatest cause of death in the United States . Approximately 795,000 persons in the United States suffer from the devastating effects of strokes on a regular basis. It is the fourth largest major cause of death in India. Ischemic and hemorrhagic strokes are the two types of strokes. Clots hinder drainage in a chemical stroke, whereas a weak blood artery breaks and bleeds into the brain in a hemorrhagic stroke. Stroke can be avoided by living a healthy and balanced lifestyle that includes quitting smoking and drinking, maintaining a reasonable BMI and glucose level, and having great heart and kidney function.

Detecting a stroke is essential, because it must be treated quickly to avoid permanent damage or death.

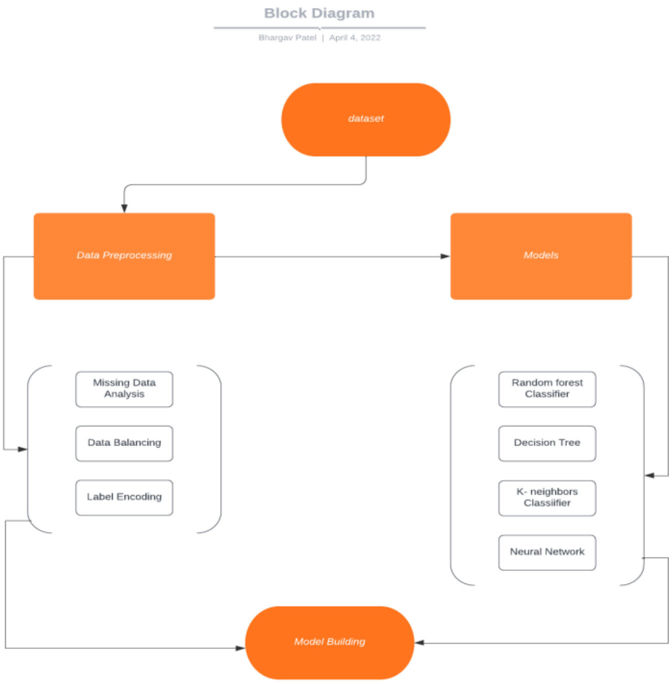
The bulk of trials showed a 94-95 percent accuracy rate, which was deemed to be pretty excellent. The most effective models in the majority of trials were Random Forest(RF), Decision Tree(DT), K-neighbors classifier, and Neural Network, with 95 percent, 90 percent, 89 percent, and 93 percent F1-Score, respectively. They implemented the Smote (Over-sampling) technique to balance the data in all of the studies, which added duplicate data to the minority class.

The major contribution of this research is that we have used NearMiss (Under-Sampling) technique for Balancing the data which removes data from majority class and Implemented using Neural Network.

2. Procedure and Experimental Methodology

This section presents a summary of the study's approach and methodology, as well as a block diagram and assessment matrices.

2.1 Block Diagram:



2.2 DataSet :

The investigation was conducted using a globally accessible stroke prediction dataset. This dataset has 5110 rows and 12 columns. The output column stroke value is either 1 or 0. A score of 0 indicates that no stroke risk has been identified, whereas a value of 1 indicates that a risk of stroke has been identified. In this dataset, the chance of 0 in the output column (stroke) is greater than the likelihood of 1 in the same column.

Table 1 : Dataset

Attribute Name	Type *(Values)*	Description
1. id	Integer	A unique integer value for patients
2. gender	String literal *(Male, Female, Other)*	Tells the gender of the patient
3. age	Integer	Age of the Patient
4. hypertension	Integer *(1, 0)*	Tells whether the patient hashypertension or not
5. heart_disease	Integer *(1, 0)*	Tells whether the patient hasheart disease or not
6. ever_married	String literal *(Yes, No)*	It tells whether the patient ismarried or not
7. work_type	String literal *(children, Govt_job, Never_worked ,Private, Self-employed)*	It gives different categories forwork
8. Residence_type	String literal *(Urban, Rural)*	The patient's residence type isstored
9.avg_glucose_level	Floating pointnumber	Gives the value of averageglucose level in blood
10. bmi	Floating pointnumber	Gives the value of the patient'sBody Mass Index
11. smoking_status	String literal *(formerly smoked,never smoked, smokes, unknown)*	It gives the smoking status of thepatient
12. stroke	Integer *(1, 0)*	Output column that gives thestroke status

2.3 Data Preprocessing :

Before developing a model, data preprocessing is essential to eliminate undesirable noise and outliers from the dataset, which might cause a divergence from normal training. After doing exploratory data analysis, exclude the 'bmi', 'id', and 'Residence type' columns, as well as the row data for the Gender type 'other'.

2.4 Label Encoding :

Label encoding converts a string data column to an integer type, allowing the model to better grasp the data pattern. Label Encoding is required for the 'gender','smoking status', 'work type', and 'ever married' columns.

2.5 Handling Imbalanced Data :

In Dataset 249 rows alone in the stroke column have the value 1, where as 4861 rows have the value 0. Data preparation is used to balance the data and increase accuracy. It's obvious that this is an unbalanced dataset. The NearMiss technique has been used to balance this dataset.

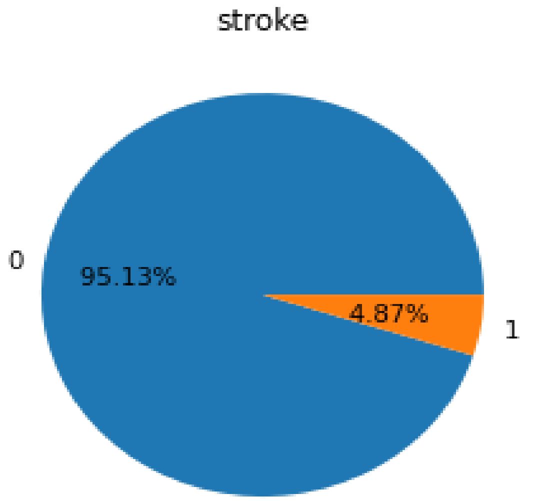


Figure 1 : Before Balancing the data

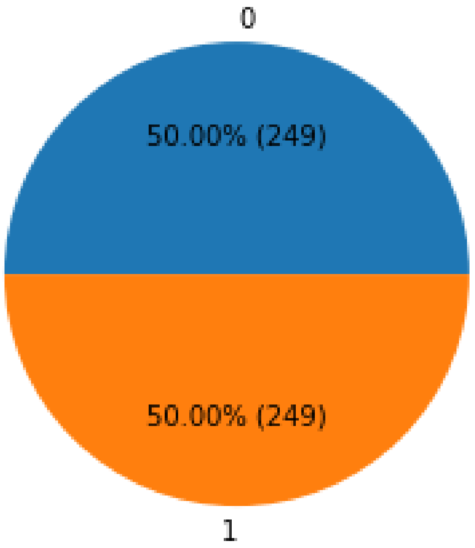
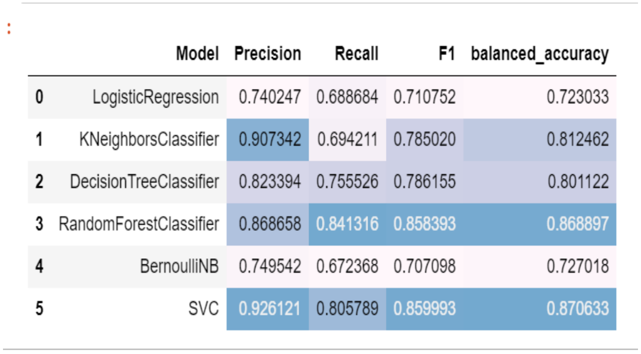


Figure 2 : After Balancing the data

After finishing data preparation and handling the unbalanced dataset, the next step is to build the model. The data is separated into training and testing data, with an 80/20 ratio of training and testing data to increase the accuracy and efficiency of this work. The model is trained using a Different Models once it has been divided. The techniques used in this study were Random Forest (RF), Decision Tree (DT), K-neighbors classifier, Nueral Network, and logistic regression.

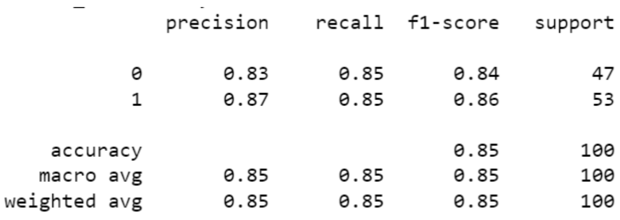
3. Evaluation of models

Figure 3 : Evaluation of Classification Models



	Model	Precision	Recall	F1	balanced_accuracy
0	LogisticRegression	0.740247	0.688684	0.710752	0.723033
1	KNeighborsClassifier	0.907342	0.694211	0.785020	0.812462
2	DecisionTreeClassifier	0.823394	0.755526	0.786155	0.801122
3	RandomForestClassifier	0.868658	0.841316	0.858393	0.868897
4	BernoulliNB	0.749542	0.672368	0.707098	0.727018
5	SVC	0.926121	0.805789	0.859993	0.870633

Figure 4 : Classification report of RandomForest



```
              precision    recall  f1-score   support

           0       0.83      0.85      0.84        47
           1       0.87      0.85      0.86        53

    accuracy                           0.85       100
   macro avg       0.85      0.85      0.85       100
weighted avg       0.85      0.85      0.85       100
```

Figure 5 : Confusion Matrics of RandomForest

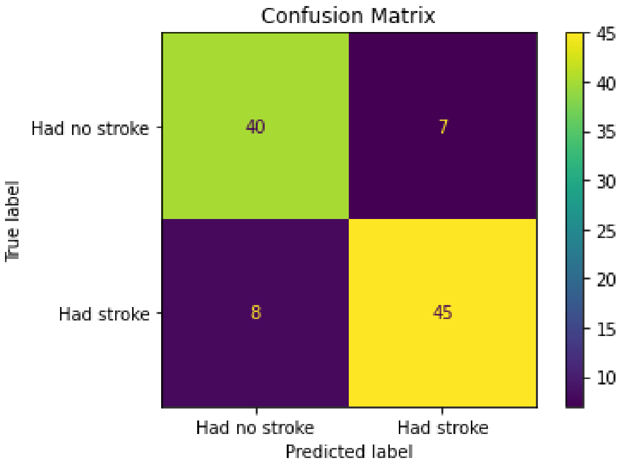
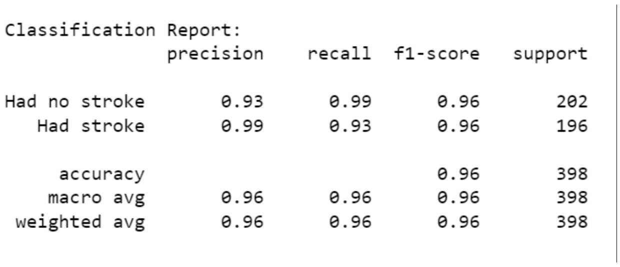


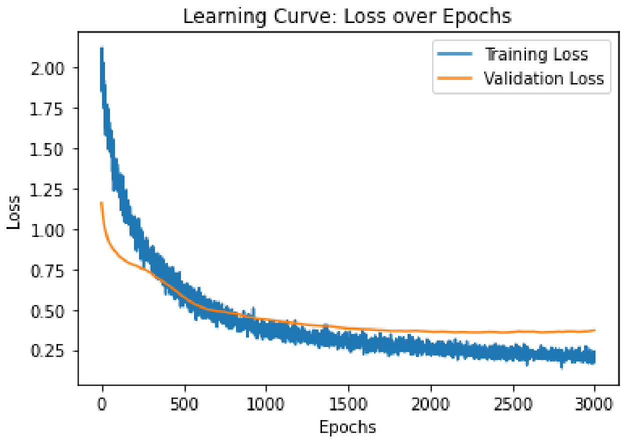
Figure 6 : Evaluation of Neural Network



```
Classification Report:
               precision    recall  f1-score   support

Had no stroke       0.93      0.99      0.96       202
   Had stroke       0.99      0.93      0.96       196

     accuracy                           0.96       398
    macro avg       0.96      0.96      0.96       398
 weighted avg       0.96      0.96      0.96       398
```



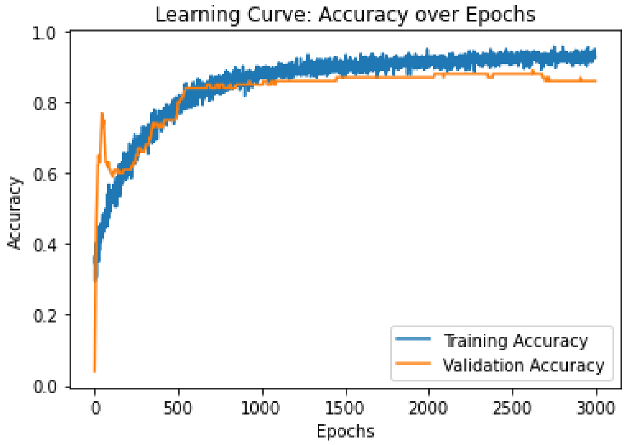
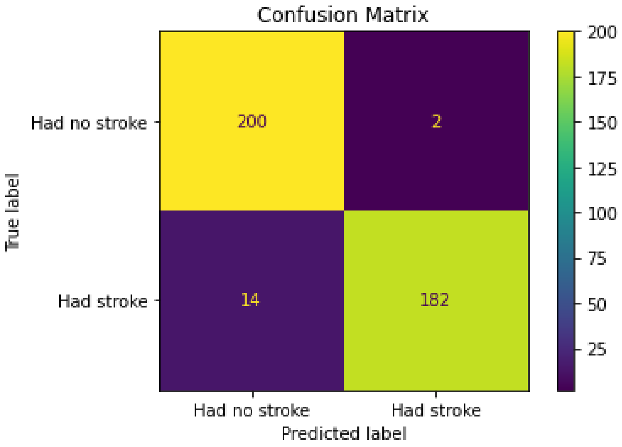


Figure 7 : Confusion Matrics of Neural network



4. Conclusion

The 'Neural Network' algorithm performs the best out of all the algorithms tested, with a 96 percent accuracy rate. The following graph shows a comparison of accuracies achieved from various methods. 'Neural Network' outperformed the others in terms of accuracy, recall, and F1 scores.

Figure 8 : Recall Values Comparision

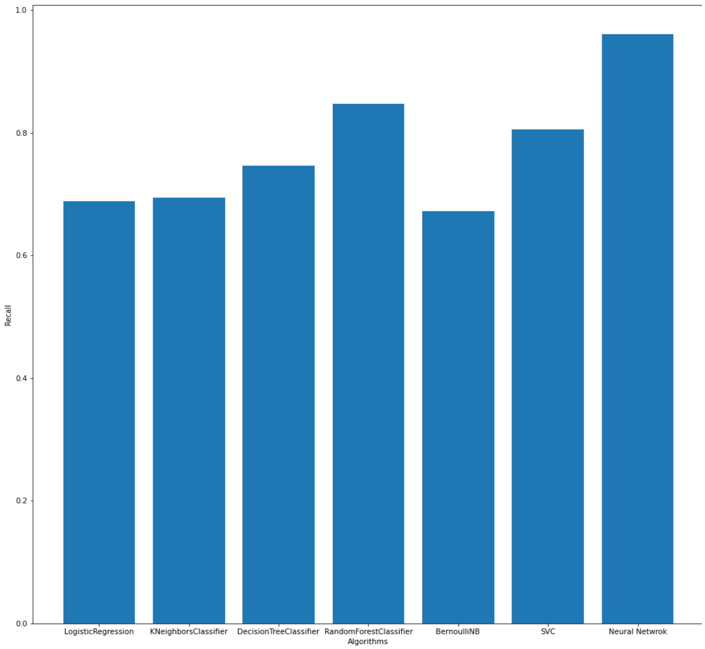


Figure 9 : Precision Values Comparision

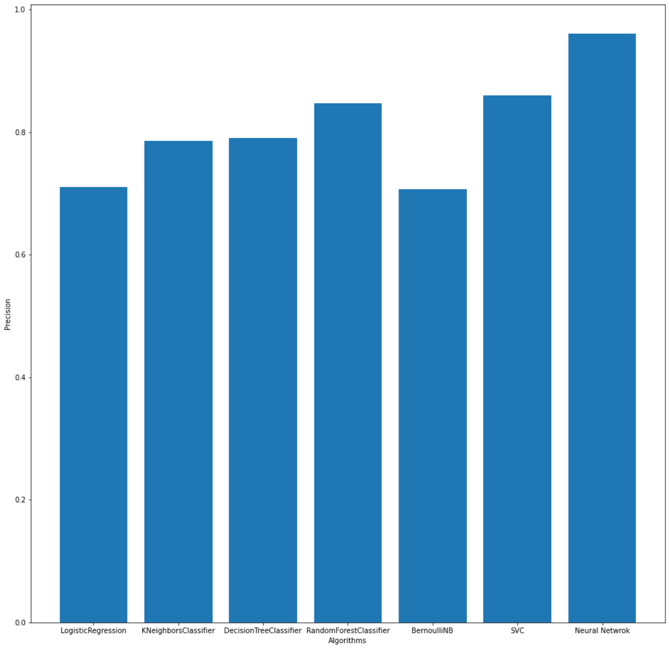


Figure 10 : Accuracy Values Comparision

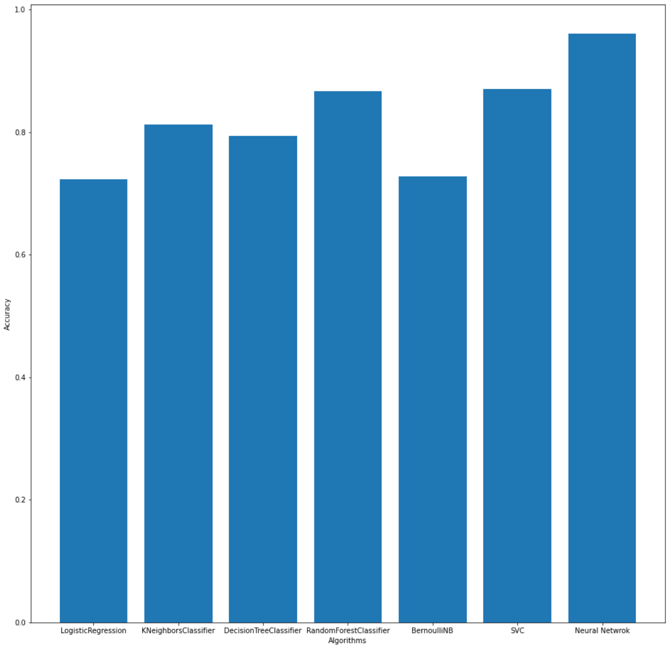
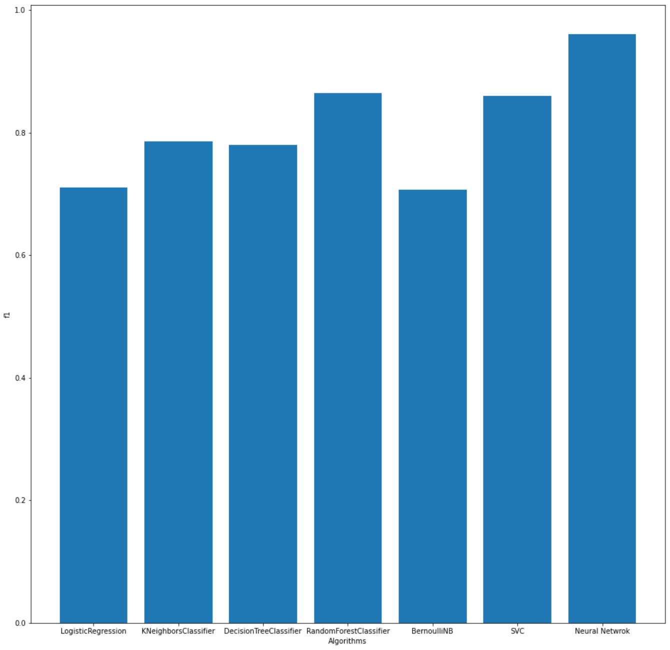


Figure 11 : F1 Score Comparision



The implementation of different Models is suggested in this paper. Delivering additional data as an input-set to neural networks and giving Brain CT Scan Image as input can increase neural network accuracy.

5. REFERENCES

[1] Nueral Network

[2] Dataset named 'Stroke Prediction Dataset' from Kaggle

[3] Singh, M.S., Choudhary, P., Thongam, K.: A comparative analysis for various stroke prediction techniques. In: Springer, Singapore (2020).

[4] 7 Techniques to Handle Imbalanced Data – Kdnuggets.

[5] Under sampling Techniques - https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/

[6] Documentation for Logistic Regression from Scikit-learn.org.

[7] Documentation for Decision Tree Classification from Scikit-learn.org.

[8] Documentation for Random Forest Classification from Scikit-learn.org.

[9] Documentation for K-Nearest Neighbor from Scikit-learn.org.

[10] Documentation for SVM from Scikit-learn.org.