Health Informatics and Biomedical Engineering Applications, Vol. 78, 2023, 95–102
https://doi.org/10.54941/ahfe1003460

Detecting Stroke in Human Beings Using Machine Learning

Oni Damilola Igbagbo, Satyam Mishra, Vu Minh Phuc, Le Trung Thanh, and Pham Hai Yen

International School, Vietnam National University, Hanoi (VNU-IS), Vietnam

ABSTRACT

In developing and underdeveloped nations, stroke is a leading cause of mortality and disability. Stroke is a life-threatening condition that develops when there is a lack of blood flow to the brain from the carotid arteries and vertebral arteries. Because the brain suffers damage and can quickly expire without oxygen, stroke frequently results in death and can occasionally affect nearby body parts if the patient is not given prompt medical attention. Spasticity, contractures, paralysis, and death are among the effects. According to the World Health Organization, stroke accounts for over 137,000 fatalities per year in the United States alone and over 451,000 deaths per year in Africa. Today, stroke is a medical illness that affects people in practically every region of the world, including industrialized, developing, and undeveloped nations. In general, 1 in 4 adults over 25 will experience a stroke at some point in their lives. This year, 12.2 million people are predicted to experience their first stroke, and 6.5 million of them will pass away as a result. The number of stroke victims worldwide exceeds 110 million. What if this global endemic could be stopped? The world will be safer and life expectancy will rise if accurate stroke prediction technology is developed. We have proposed our research study to develop a solution to predict strokes in people using machine learning. We have employed four models/classifiers to check the accuracy on each of them with same dataset of people and we have achieved great results. The two models gave 98% and 98.29% successful accuracy results which is very close to state-of-the-art methods (99%).

Keywords: SDG, Stroke, Naïve-bayes, Random forest, Decision tree, Neural network, Data training

INTRODUCTION

Ensuring healthy lives and promoting well-being at all ages is essential to sustainable development goals (SDG), yet the world focus attention on covid-19 but there is a global endemic that is expected to affect over 110 million people in the world (*THE 17 GOALS | Sustainable Development*, n.d.). The Global Endemic is stroke, Stroke is a major cause of death and disability in developing and under-developed countries. Stroke occurs when there is shortage of blood from the carotid arteries and the vertebral arteries to the brain this is a life threaten disease, the brain get damage and can die within minutes if left without oxygen, therefore stroke usually causes death and sometimes it affects immediate part of the body if prompt medical attention is not given to the patient. The effect includes Spasticity, contractures, paralysis, and death.(*Learn about Stroke*, n.d.) With stroke causing over 137000 peoples death annually in the united states of America alone and also over 451000 deaths in annually in Africa according to the world health organization. There are some works being done for image processing as well using Canny Edge Detection Algorithm (Mishra & Thanh, 2022). Also, some neural network approach training for object detection etc. for driverless vehicles to take covid patients (Mishra et al., 2022). Stroke is now a medical condition that is occurring in almost all part of the world, in developed countries, developing countries and underdeveloped countries. Generally, 1 in 4 adults over the age of 25 will have stroke in their lifetime. 12.2 million people worldwide are expected to have their first stroke this year and 6.5 million will die as a result. Over 110 million people in the world have experienced stroke.(*WHO EMRO | Stroke, Cerebrovascular Accident | Health Topics*, n.d.) What if there is a way to prevent this global endemic? If there is a technology to predict stroke accurately this will increase life expectancy worldwide and the world will be a lot safer. Therefore, we have proposed a machine learning strategy to accurately predict the stroke in human beings and we have achieved 98.2% of successful results which is very close to state-of-the-art methods (99%).

METHODOLOGY

We have developed this solution in several stages. The stages are Data description, Data Processing and Modelling and Predicting Data.

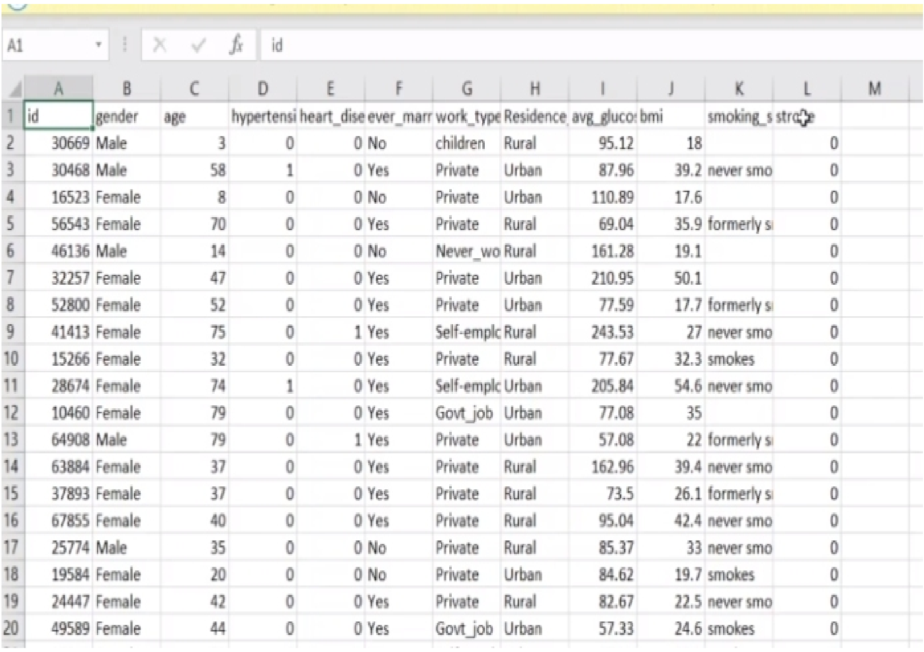
Data Description

We have collected various data of people including patients affected by stroke with the aim of predicting whether a patient will suffer from stroke. The dataset (***Stroke Prediction Dataset*, n.d.**) contains 1111 features, and the 12th column is our target which predicts.

Figure 1 shows the dataset of people including patients affected by stroke. In there, the first column is the ID of the patient, and we have the gender, follow by the age column and the hypertension column, then we have the heart disease where 0 stands for no disease and 1 stand for having heart disease, the marital status, occupation, Area of residency, Average glucose level, Body max Index (BMI) and smoking status are the remaining columns while the stroke column is the last column where 0 stand for no stroke and 1 stand for suffering from stroke. Most of the data set are in small quantity except for the age, average glucose, and BMI.

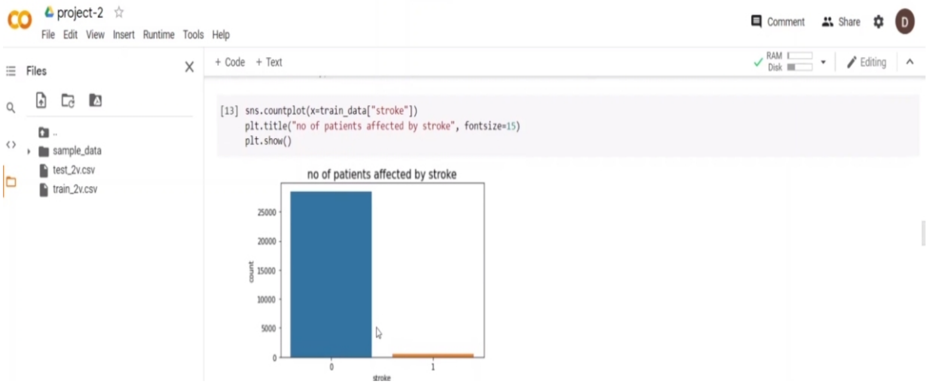
Data Processing

We loaded the data set on Google Collaboratory and used the necessary functions to analyze the dataset, the dataset was first cleaned where the empty or null dataset were eliminated after the dataset was reduced. We plotted 13 graphs with the features that was given and explore different charts and bar graph, and from our data the numbers of people suffering from stroke were



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	gender	age	hypertensi	heart_dise	ever_marr	work_type	Residence	avg_gluco:	bmi	smoking_s	stroke	
2	30669	Male	3	0	0	No	children	Rural	95.12	18		0	
3	30468	Male	58	1	0	Yes	Private	Urban	87.96	39.2	never smo	0	
4	16523	Female	8	0	0	No	Private	Urban	110.89	17.6		0	
5	56543	Female	70	0	0	Yes	Private	Rural	69.04	35.9	formerly si	0	
6	46136	Male	14	0	0	No	Never_wo	Rural	161.28	19.1		0	
7	32257	Female	47	0	0	Yes	Private	Urban	210.95	50.1		0	
8	52800	Female	52	0	0	Yes	Private	Urban	77.59	17.7	formerly si	0	
9	41413	Female	75	0	1	Yes	Self-emplc	Rural	243.53	27	never smo	0	
10	15266	Female	32	0	0	Yes	Private	Rural	77.67	32.3	smokes	0	
11	28674	Female	74	1	0	Yes	Self-emplc	Urban	205.84	54.6	never smo	0	
12	10460	Female	79	0	0	Yes	Govt_job	Urban	77.08	35		0	
13	64908	Male	79	0	1	Yes	Private	Urban	57.08	22	formerly si	0	
14	63884	Female	37	0	0	Yes	Private	Rural	162.96	39.4	never smo	0	
15	37893	Female	37	0	0	Yes	Private	Rural	73.5	26.1	formerly si	0	
16	67855	Female	40	0	0	Yes	Private	Rural	95.04	42.4	never smo	0	
17	25774	Male	35	0	0	No	Private	Rural	85.37	33	never smo	0	
18	19584	Female	20	0	0	No	Private	Urban	84.62	19.7	smokes	0	
19	24447	Female	42	0	0	Yes	Private	Rural	82.67	22.5	never smo	0	
20	49589	Female	44	0	0	Yes	Govt_job	Urban	57.33	24.6	smokes	0	

Figure 1: Data of people including patients affected by stroke.

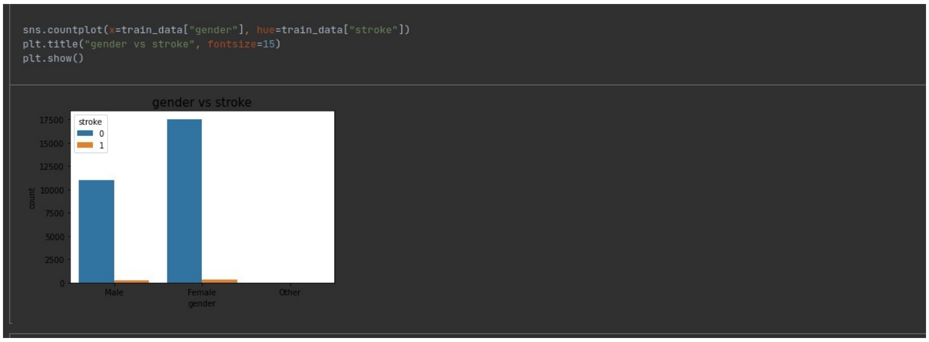


```
[13] sns.countplot(x=train_data["stroke"])
     plt.title("no of patients affected by stroke", fontsize=15)
     plt.show()
```

Figure 2: Bar graph showing number of patients affected by stroke.

548 while 28524 were people with no stroke, we plotted a bar chat of people having stroke versus people not having stroke. Figure 2 shows the graph.

A graph was plotted between gender and stroke to know the relation between both, and numbers of female stroke patients are more than male. The author went further to check the number of smokers versus nonsmokers and the correlation between gender and smoking status was checked, we observed from our dataset that our dataset contains numerical, strings and object type of values. Therefore, we combined our dataset into type of columns that contains numerical values, columns that contains string values and columns that contain string values because we can't feed the string directly into our dataset and the string data was converted into dataset where it was later divided into various classes. Figure 3 shows the graph plotted between gender and stroke.



```
sns.countplot(x=train_data["gender"], hue=train_data["stroke"])
plt.title("gender vs stroke", fontsize=15)
plt.show()
```

Figure 3: Graph plotted between gender and stroke.

Modeling and Predicting Data

We extracted out the Y and X train, the X train contains 11 columns whereas the 12th column is the Y train that has true columns after which we used Train-Test Split functionality (Karthik et al., 2018) where we split our data into training and test datasets where 75% was allocated as the training data while 25% was used as the test data.

After splitting data, we build our own model where we use:

- Naive-Bayes classification models
- Decision tree
- Random forest classification
- Neural networks

After training our datasets in each model we use exhaustive cross-validations methods to which learn and test on all possible to divide the original sample into a training and a validation set.

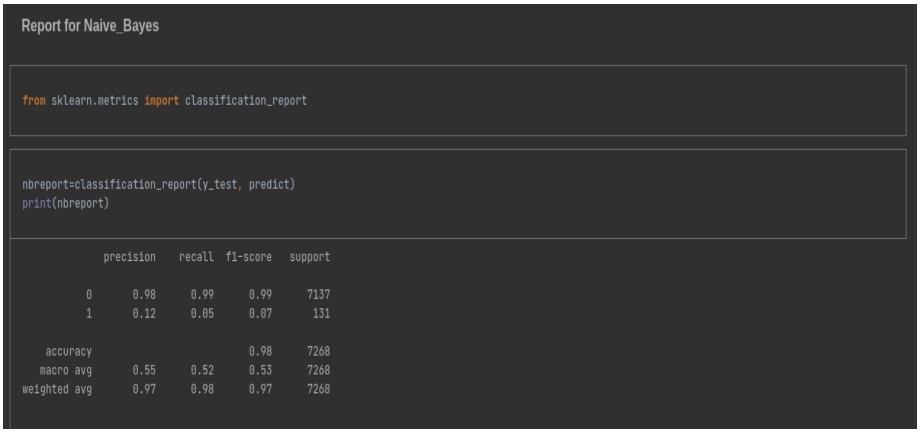
Naïve-Bayes Classification Models

We conducted the Y test and X test, and our model was fit into the training set and the accuracy was tested using model score which gives 98% accuracy on our given datasets. (*1.9. Naive Bayes*, n.d.) The Confusion matrix or Error matrix was carried out and this was used to check the direct comparison, accuracy, and precision. The confusion matrix predicted that 7112 don't have stroke and 32 people are expected to have stroke. We went further to use classification report to measure the quality of predictions from the classification algorithm. Figure 4 below shows the report for Naïve Bayes.

Decision Tree

We went further to use decision tree model to analyze the datasets after fitting in our training set and the accuracy was 96% which was lower compared to the accuracy of Naive-Bayes Classification, which was 98%. Figure 5 below shows the accuracy of decision tree model. (*1.10. Decision Trees*, n.d.)

The Confusion matrix was also carried out and it displayed that 7122 people don't have stroke and 22 people have strokes compare to naive-Bayes,



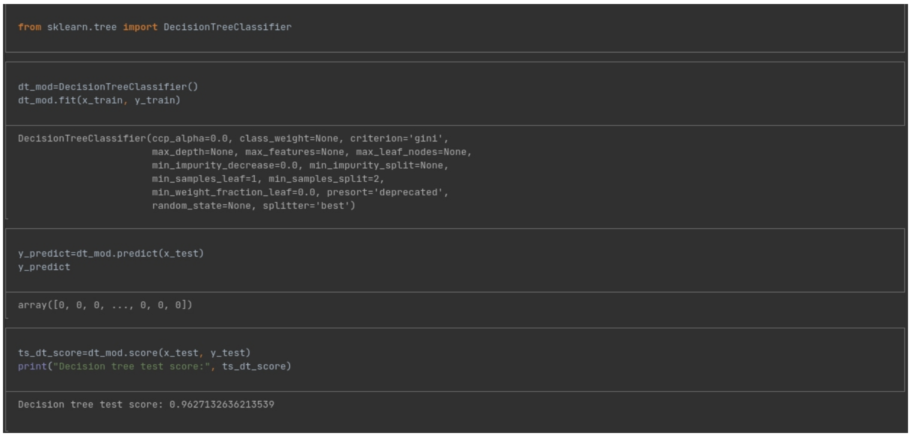
Report for Naive_Bayes

```
from sklearn.metrics import classification_report
```

```
nbreport=classification_report(y_test, predict)
print(nbreport)
```

```
              precision    recall  f1-score   support

           0       0.98      0.99      0.99      7137
           1       0.12      0.05      0.07       131

    accuracy                           0.98      7268
   macro avg       0.55      0.52      0.53      7268
weighted avg       0.97      0.98      0.97      7268
```

Figure 4: Report for Naive Bayes.



```
from sklearn.tree import DecisionTreeClassifier
```

```
dt_mod=DecisionTreeClassifier()
dt_mod.fit(x_train, y_train)
```

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=None, splitter='best')
```

```
y_predict=dt_mod.predict(x_test)
y_predict
```

```
array([0, 0, 0, ..., 0, 0, 0])
```

```
ts_dt_score=dt_mod.score(x_test, y_test)
print("Decision tree test score:", ts_dt_score)
```

```
Decision tree test score: 0.9627132636213539
```

Figure 5: Output of accuracy of decision tree model.

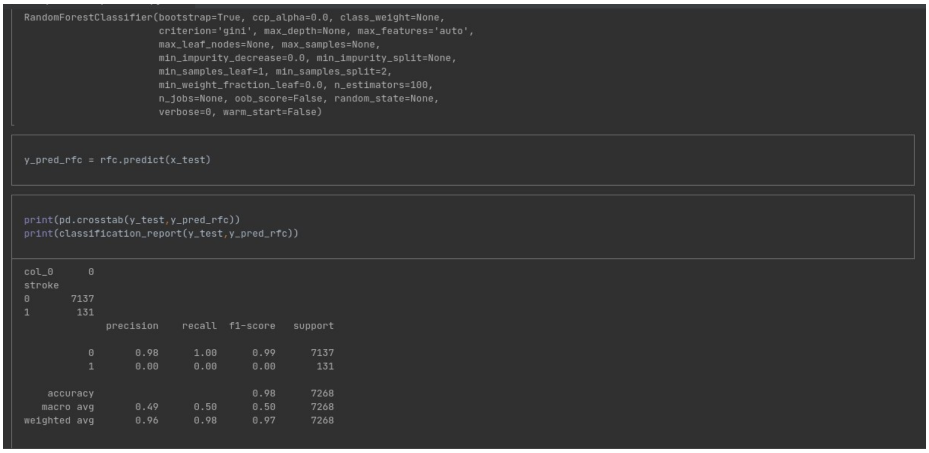
that predicted 7112 people don’t have stroke and 32 people are expected to have stroke.

Random Forest Classification

After loading our random forest classifiers, we fitted our model to the training datasets so and it predicted 97% accuracy on our given datasets. Which is a good result compared to decision tree which was 96% accuracy. Figure 6 below displays the output of the random forest classification. (*Sklearn.Ensemble.RandomForestClassifier*, n.d.)

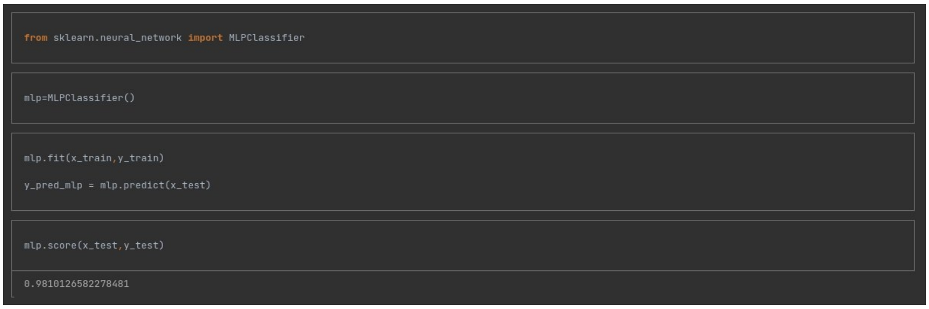
Neural Networks

We used MLP Classifier (Salama et al., 2012) to build this model and made our prediction on Xtest and the MLPclasssifier also gives 98.29%, which is a very good result compared to the previous result gotten from Random forest and decision tree. (Dreiseitl & Ohno-Machado, 2002)



```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)

y_pred_rfc = rfc.predict(x_test)

print(pd.crosstab(y_test,y_pred_rfc))
print(classification_report(y_test,y_pred_rfc))

col_0      0
stroke
0       7137
1        131
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      7137
           1       0.00      0.00      0.00       131

    accuracy                           0.98      7268
   macro avg       0.49      0.50      0.50      7268
weighted avg       0.96      0.98      0.97      7268
```

Figure 6: Random Forest Classification.



```
from sklearn.neural_network import MLPClassifier

mlp=MLPClassifier()

mlp.fit(x_train,y_train)

y_pred_mlp = mlp.predict(x_test)

mlp.score(x_test,y_test)

0.9810126582278481
```

Figure 7: The output of MLP classifier.

RESULTS

After analysing all the models mentioned above, we have successfully obtained a way to predict stroke in human beings with accuracy of 98.2%. After modelling and prediction of data, we have found that using Naïve-Bayes model, which gives 98% accuracy on our given datasets. The Confusion matrix or Error matrix was carried out and this was used to check the direct comparison, accuracy, and precision. The confusion matrix predicted that 7112 don't have stroke and 32 people are expected to have stroke. When deploying our dataset through Decision Tree, the accuracy was 96% which was lower compared to the accuracy of Naive-Bayes Classification, which was 98%. When deploying through Random Forest classification, it predicted 97% accuracy on our given datasets. Which is a good result compared to decision tree which was 96% accuracy. And last but not least, when deployed through neural networks, it gave 98.29%, which is a very good result compared to the previous result gotten from Random Forest and decision tree.

Through Table 1, we can conclude that most of the time all datasets work fine with all mentioned models/classifiers, but Naïve-Bayes and Neural Networks are the most accurate ones to go with.

Table 1. Accuracy results obtained after training the people dataset through different models/classifiers.

Models/Classifiers	Accuracy Obtained
Naïve-Bayes	98%
Decision Tree	96%
Random Forest	97%
Neural Networks	98.29%

CONCLUSION

Accuracy is essential for us to understand the external world, as it is the closeness of the measured value to a standard or true value. Accuracy is obtained by taking small readings. This small reading reduces the error of the calculation, and these small readings can contribute to the life expectancy of humans, although all the forementioned four models/classifiers used, have given the great accurate results but the degree of accuracy in each cannot be neglected as it may be the defining point between the life and death of a human being. We have applied a machine learning strategy in our study to accurately predict the stroke in human beings and we have achieved 98.2% of successful results which is very close to state-of-the-art methods (99%) [Table 2 shows the comparison between our research findings and state-of-the-art methods]. Therefore, we concluded that most of the time all datasets work fine with all mentioned models/classifiers, but Naïve-Bayes and Neural Networks are the most accurate ones to go with. This will help to prevent stroke as early as possible instead of managing the patients. We wish scientists can furthermore develop it, so that it can be applied in real world to save precious human-lives.

Table 2. Comparison between our research findings and state-of-the-art methods.

Our Findings	STATE OF THE ART (SOTA)
Accuracy of 98.29% We used the following algorithm: Naïve-Bayes, Decision Tree, Random Forest, Neural Networks	Accuracy of 99% Algorithms in SOTA; Bidirectional Long Short-Term Memory
The best algorithm is Neural networks.	The best algorithm is Bidirectional Long Short-Term Memory.

ACKNOWLEDGMENT

Foremost, we would like to express our sincere gratitude to our advisor Dr. Le Trung Thanh for the continuous support for our group's study and research, for his patience, motivation, enthusiasm, and immense knowledge.

His guidance helped us in all the time of research and writing of this research. We could not have imagined having a better advisor and mentor for our work.

REFERENCES

1.9. Naive Bayes. (n.d.). Scikit-Learn. Retrieved December 31, 2022, from https://scikit-learn/stable/modules/naive_bayes.html

1.10. Decision Trees. (n.d.). Scikit-Learn. Retrieved February 2, 2023, from https://scikit-learn/stable/modules/tree.html

Learn about stroke. (n.d.). World Stroke Organization. Retrieved December 10, 2022, from https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learn-about-stroke

Sklearn.ensemble. RandomForestClassifier. (n.d.). Scikit-Learn. Retrieved December 31, 2022, from https://scikit-learn/stable/modules/generated/sklearn.ensemble.~RandomForestClassifier.html

Stroke Prediction Dataset. (n.d.). Retrieved February 2, 2023, from

THE 17 GOALS | Sustainable Development. (n.d.). Retrieved December 10, 2022, from https://sdgs.un.org/goals

WHO EMRO | Stroke, Cerebrovascular accident | Health topics. (n.d.). World Health Organization - Regional Office for the Eastern Mediterranean. Retrieved December 10, 2022, from https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5), 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Karthik, S., Srinivasa Perumal, R., & Chandra Mouli, P. V. S. S. R. (2018). Breast Cancer Classification Using Deep Neural Networks. In S. Margret Anouncia & U. K. Wiil (Eds.), *Knowledge Computing and Its Applications: Knowledge Manipulation and Processing Techniques: Volume 1* (pp. 227–241). Springer. https://doi.org/10.1007/978-981-10-6680-1_12

Mishra, S., & Thanh, L. T. (2022). SATMeas - Object Detection and Measurement: Canny Edge Detection Algorithm. In X. Pan, T. Jin, & L.-J. Zhang (Eds.), *Artificial Intelligence and Mobile Services – AIMS 2022* (pp. 91–101). Springer International Publishing. https://doi.org/10.1007/978-3-031-23504-7_7

Mishra, S., Minh, C. S., Thi Chuc, H., Long, T. V., & Nguyen, T. T. (2022). Automated Robot (Car) using Artificial Intelligence. *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, 319–324. https://doi.org/10.1109/ISMODE53584.2022.9743130

Salama, G. I., Abdelhalim, M., & Zeid, M. A. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. *Breast Cancer (WDBC)*, *32*(569), 2. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset