**A MINI PROJECT**

**ON**

**PREDICTION OF STROKE USING MACHINE LEARNING ALGORITHM**

**BY**

**NAME**

**SUBMITTED**

**TO**

**NAME**

**AT**

**ON**

**DATE**

# 1. INTRODUCTION

There are over 600 diseases of the nervous system which are stroke brain tumor epilepsy, cerebral palsy, sciatica and many more. Stroke is one of the most severe diseases globally and known to be the second leading cause of death and the third leading cause of disability in the world, and is usually directly and indirectly responsible for a considerable number of deaths. Stroke are commonly found in old people of age 65 and above.it affects the human brain which is either caused by rupture or a blockage of a blood artery in the brain. There is also various reason in which a patient got stroke which are either diet, inactivity, tobacco, alcohol, personal history, health history and complications. The World health organization claims that there are frequent deaths due to stroke. It is very difficult to predict the symptoms of stroke and because it is a very critical medical condition, it enhanced the interest of researcher in applying machine learning to diagnose stroke effectively. If stroke is detected at an early stage, the loss of death and severe damage of the brain will be avoided.

This research aims to predict stroke using machine learning algorithm and also to attain the follow objectives.

1. To build an effective model that will be able to predict the occurrence of stroke from a publicly available dataset.
2. To achieve a high accuracy

# 3. RELATED WORK

Several machine learning algorithms have been used in predicting stroke and the following techniques have been reviewed.

Sharma *et al.* (2022) proposed a model to predict stroke diseases using machine learning. The method they used were Decision Tree, Random Forest, Naïve Bayes, Multi-layer perception algorithm and JRIP algorithm. The dataset contained 5110 patients of which 59% were female and 42% were male. The model performed better using Random forest which obtained an accuracy of 98.63% compared to other models and when 10 folds were applied it performed better with an accuracy of 98.94%.

Schindler *et al*. (2020) proposed a model to predict stroke by detecting of haemorrhage in carotid plaques. The used 560 patients with symptomatic carotid stenosis and 136 patients with asymptomatic carotid stenosis. Hazards of ipsilateral ischemic stroke were compared between patients with and without IPH and it was seen that IPH was 51.6% present in patients with symptomatic carotid stenosis and 29.4% were present in patients with asymptomatic carotid stenosis.

Dritsas *et al.(*2020) proposed a model in predicting stroke using Naive Bayes (NB), Random Forest (RF), Logistic Regression (LR), K-Nearest neighbour (KNN), Stochastic Gradient Descent (SGD), Decision Tree (DT), Multilayer perception (MLP), Majority Voting and then Stacking was used to boost the accuracy of the model. The result shows that the stacking method out performs the other models with an AUC of 98.9%, F-measure, precision, recall of 97.4% and an accuracy of 98%.

Douglas, (2018) proposed a model using electronic health records to predict stroke. The dataset collected were 8000 electronic records which was splinted into training, validation, and test set at a ratio of 70:15:15 and seven (7) neural network architecture were used on the dataset. Hoe *et al.* (2020), proposed a model to predict stroke using natural language processing – bases machine learning of radiology report of brain MRI in which English text report of the brain MRI were examined during the admission of the patients the English text report was divided into a ratio of 7:3. CNN, RNN, RF and SVM were used. It was indicating that deep learning performed better than the machine learning algorithm.

Ashokan *et al.* (2020), Built an application to predict the occurrence of stroke based on the users input. The algorithms used were Random Forest (RF), Decision Tree (DT), logistic Regression (LR), and Support Vector machine (SVM). DT outperformed other algorithm with an accuracy of 0.8486.

(Malathi et al., 2022), Proposed a model for predicting of stroke using Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB) and Decision Tree (DT). The dataset consisted of 5110 records which is categorized into 2 class, the Normal class and the Stroke class. The dataset was splitted into a training and test set of ratio 7:3. Accuracy, Precision, Recall and Fi score were used to measure the performance of the model.

Minhaz *et al.* (2020), conducted a research to predict stroke, 5110 patient’s details were collected, the data was pre-processed and splitted into training and testing. In the study the authors used Logistic Regression (LR), Stochastic Gradient Descent (SGD), Decision Tree, Ada Boosting Classifier, Gaussian Classifier, Quadratic Discriminant Analysis, Multi-layer perception Classifier, K Neighbours Classifier, Gradient boosting and XGBoost classifier to predict stroke. They used AUC, accuracy cy, precision, Recall, F-1 to measure the performance of the model and finally the weight voting classifier was used by aggregating the base classifiers and achieved an accuracy of 97%.

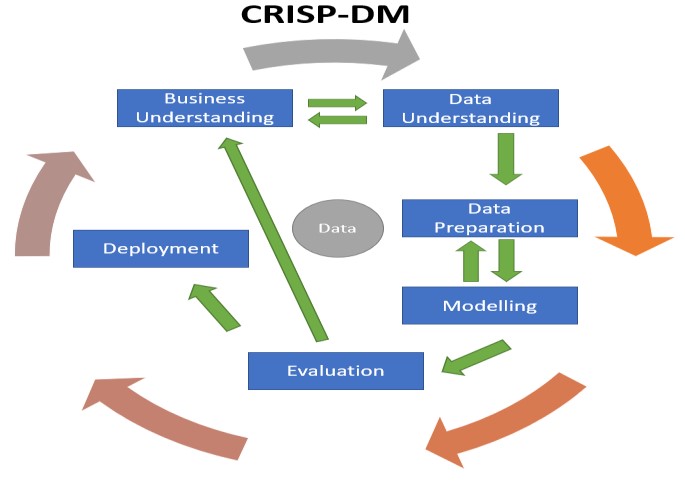
Baddam *et al.* (2022), proposed a model for heart stroke prediction using gradient boosting and bootstrap aggregation and got an accuracy of 90.2% and 80.9% respectively.

Paul *et al.* (2022), proposed a model using Gradient Boosting, Extra tree, Logistic Regression, Random Forest, Naive Bayes, support Vector machine, decision tree, Ensemble random forest to predict stroke using Advanced Random Forest Ensemble. The model was evaluated using classification accuracy, sensitivity, F1-Score, specificity, ROC-AUC and precision.

# 1. MATERIAL AND METHODOLOGY

The methodology used for this study is CRISP-DM frame work will be used in this project, it is a well knows and widely used methodology it consists of six phases, Business understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment.

Fig1: Diagram of CRISP-DM



1. Business Understanding:

Stroke is a very curial medical condition and surviving stroke is a starting point of a patient fighting for their life and the sooner a patient receives an adequate medical care. Also ensuring an accurate medication in the shortest period after the stroke onset. Predicting of stroke is very important because it helps to prevent the rate of mortality. the criteria in this investigation is to predict stroke with the following features: hypertension, BMI, Heat disease, average blood glucose level, age, ever worked, residence type. Furthermore, it can be suggested that machine learning can play a very vital role in predicting the occurrence of stroke.

1. Data Understanding:

The data was collected and analyzed based on the patient’s characteristics. Which will be useful for the prediction of stroke in the future. The dataset consists of a CSV file which was collected from kaggle. It contains over 5100 in which 2994 are female and 2115 are male with 12 variables consisting of ID, Gender (male = 0,female=1, others=3), Hypertension (hypertension = 1, No hypertension = 0), age , Heart disease(heart disease = 1, No heart disease = 0), ever married (married =yes, not married = no), work type (Private, Self-employed, Govt\_job, children, Never worked), Residence(rural = 0, urban =1), avg\_glucose\_level, body mass index (BMI), Smoking Status(formally smoked, never smoked, smokes, unknown, ), Stroke( stroke =0 , stroke =1).

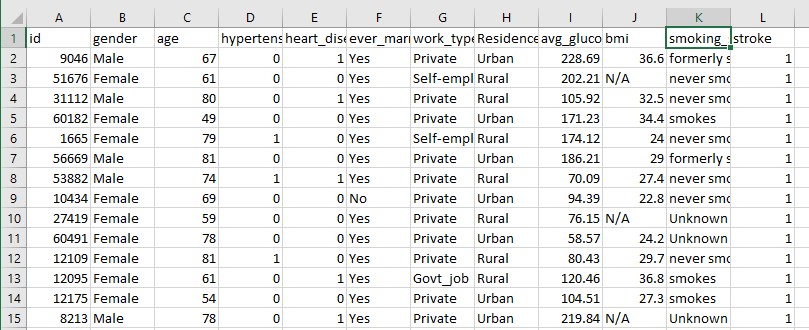
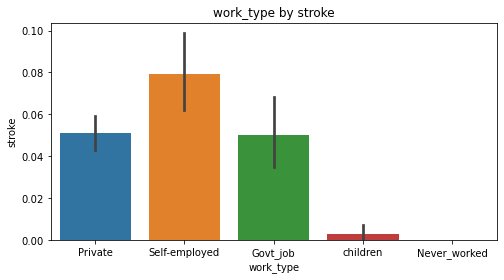
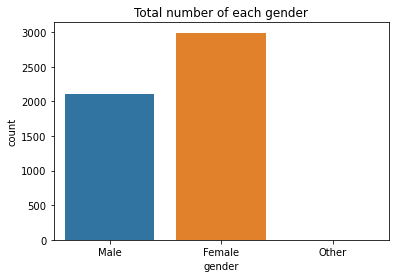


Fig2: Stroke Dataset

Fig 3,4,5,6,7,8 shows a visualization of an Exploratory Data Analysis of the dataset.



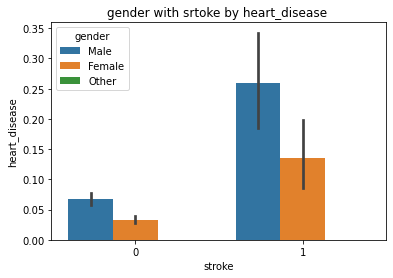
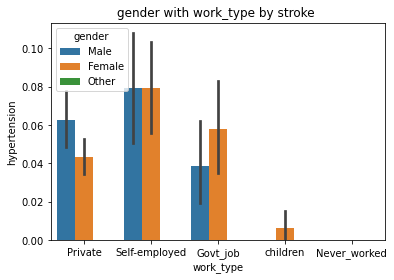
Fig4: Work type value count Fig 3: Gender Value count

Fig 6: Hypertension vs work type

Fig

5

:

Heart Disease vs

Stroke

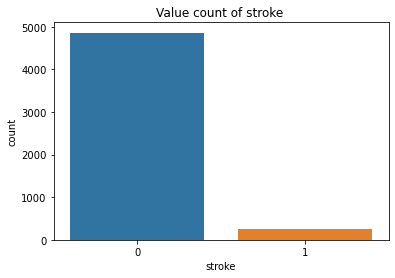
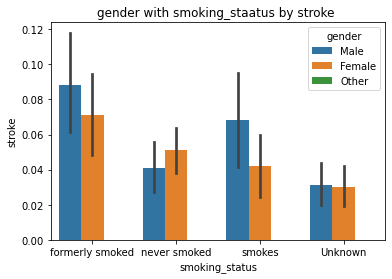


Fig8: Gender vs stroke count Fig7: Gender with smoking status vs stroke

The count plot in fig.3 showed that there are three gender type (male, female and other), and the highest number of count are the female with a total value of 2994, the male has a total value of 2115, and other as a count of 1. Fig.4 shows the sum of each work type. the visualization shows that those who are self-employed has a high tendency of having stroke compared to private and government.it also shows that children have lowest tendency of having stroke which is very rare. Fig 5: displays a visualization of gender with stroke vs heart disease, the 0 represent those without stroke and the 1 represent those with stroke. The male has a high stroke count and also heart disease. Also those with heart disease out stroke and has a higher count in male that female. Fig 6: is a visualization of hypertension with work type, showing that both male and female who are self-employed has a high tendency of having hypertension which can be either due to not stress, in the private sector the males has a high tendency of having hypertension compare to the females and also in the government sector the female has a higher tendency of having hypertension compare to male, which can either be due to the stress of taking care of the home and working for the government. Fig 7: shows gender with smoking status vs stroke, the fig shows that both male and female who formally smoked has a higher tendency of having stroke and also those who smokes also has a high tendency of having stroke compared to those who never smoked. Those who never smoked can also have stroke which can be due to other factors. Finally, fig: 8 shows the total value of those who have stroke compare to those without stroke and it can be seen from the visualization that those who do not have stroke has a higher value count compare to those that have.

1. Data Preparation:

The data collected was dirty and needed to be preprocessed in order to successfully predict the occurrence of stroke. There were 201 missing values on the BMI column in the dataset of 5110 rows which were filled using mean and also the column containing ID is dropped because it wasn’t relevant. The dataset was encoded categorically as machine learning algorithms do support strings as input also to improve the predictions done by the models. The column containing stroke was unbalanced, which means that the positives and negatives of stroke outcome are no close from being equitable. The Synthetic Minority Oversampling Techniques (SMOTE) was used to balance the Stroke outcome. After balancing the data, the data was splitted into training and testing of which 70% was for training and 30% was for testing. And finally feature scaling was carried out before the modelling.

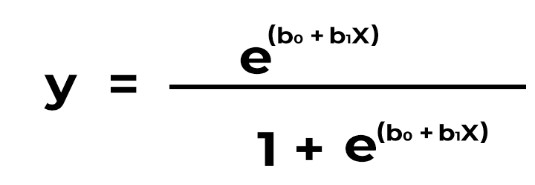
Fig:9 Balanced stroke outcome



1. Modelling:

Various models are used in building machine learning, but for the purpose of this work, machine learning algorithms such as Random forest classifier, Logistic regression, Neural Network and XGBoost classifier and at the end the four models were stacked.

Logistic Regression uses a category of independent factors to predict a categorical dependent variable.it is a predictive analysis algorithm and based on the concept of probability.



where,

* x = input value
* y = predicted output
* b0 = bias or intercept term
* b1 = coefficient for input (x)

Random Forest Classifier is a collection of prediction trees. Every tree is dependent on random vectors sampled independently, with similar distribution with every other tree in the random forest.

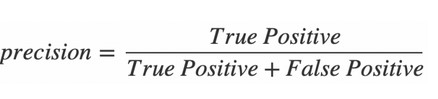
Xgboost Classifier is usually precise and adapts well to all types of data and problems. It combines the result of the model used, called based learners to make prediction.

Neural Network whichwork better at predictive analytics because of the hidden layers was used in prediction of stroke because it ‘learns’ the way a human does.

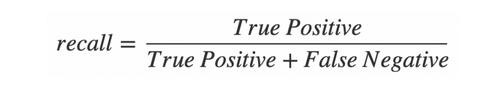
* 1. Evaluation:

The performance of the model should be assessed in order to know the best classifier. There were 4 different models used to evaluate the result which are precision, Recall, F1-Score, accuracy and AUC. Performance evaluation metrics are shown below:

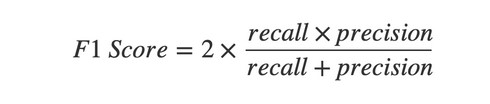
Precision: Is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.



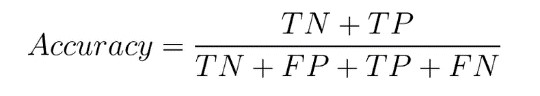
* Recall: is used to measure the fraction of positive patterns that are correctly classified.



* F1-score:



* Accuracy: Accuracy represents the number of correctly classified data instances over the total number of data instances.



* 1. Deployment:

Deployment is the process of placing a finished machine learning model into a live environment where it can be used for its intended purpose. In this project deployment of the model wasn’t carried out.

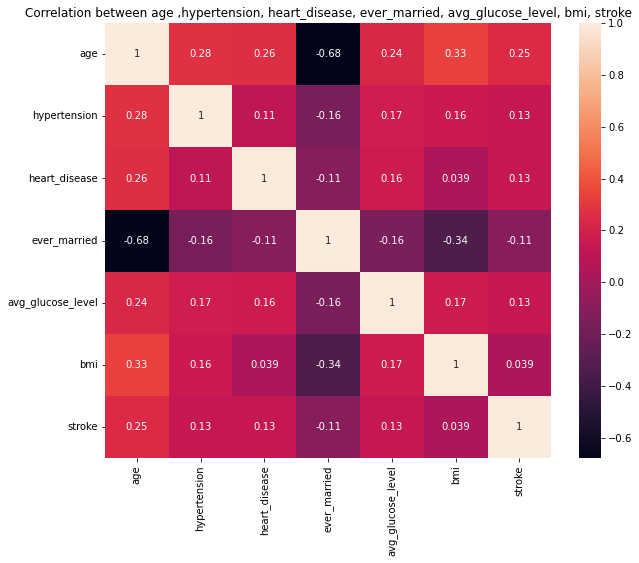
# 2. RESULT/DISSCUSION

The project was carried out using four (4) different models, Random Forest, Logistic Regression, Neural Network, XGBoost and the end the models were stacked. Among this base algorithm some showed better accuracy compare to others. Random Forest gave an accuracy of 99.5% in the training and 96.67% in the testing dataset which shows that it did well in both the training and testing dataset. Also logistic regression got an accuracy of 77.57% in the training and 77.27% in the testing dataset. Neural network used a batch size of 128 and 100 epochs and got an accuracy of 95.12%. Xgboost was also used and gave an accuracy of 96.98% in the testing dataset. Finally, after using the base models and got different accuracy, they were stacked in order to boost the accuracy score and got an accuracy of 97.25% compare to the base algorithms. Accuracy of each model is shown below.

Table 1: Models Accuracy

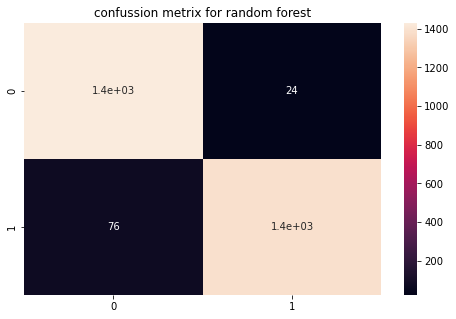
|  |  |
| --- | --- |
| **MODELS** | **ACCURACY** |
| Logistic Regression | 77.57% |
| XGBoost | 96.98% |
| Random Forest | 96.67% |
| Neural Network | 95.12% |
| Stacked Models | 97.25% |

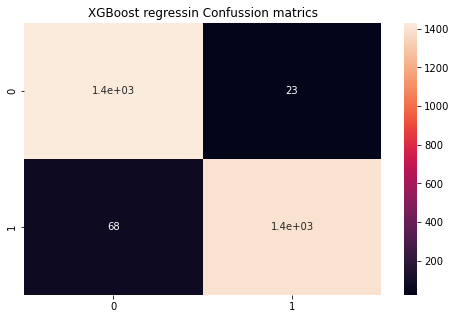
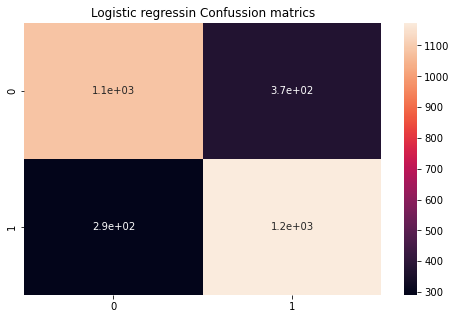
Fig 10: Correlation Matrix between Features



Correlation Matrix is an analysis that brings together correlation coefficients between an x axis and y axis. Fig 10 above shows that it is highly positively correlated because the two variables move in the same direction and it helps to predict the evolution of the relationship between the variables. Among the metrics, gender, age, hypertension, heart disease, avg\_glucose\_level, bmi, smoking status have respectably high effect on stroke. In correlation metrics values closer to 1 indicates a positive linear correlation, and also value closer to negative one indicates a negative linear correlation. As seen in the correlation metrics that Ever married shows a negative sign, meaning it is a negative linear correlation.

Fig 8: Confusion Matrix of logistic Regression, Random Forest, Gradient Boosting.





The confusion metrics shows that logistic regression model predicts a total of 291 false negatives and 372 false positives, while random forest predicts a total of 76 false negative and 24 false positive. For gradient boosting it predicted a false negative of 68 and a false positive of 23. And finally for neural network it predicted a false negative of 291 and false positive. A Receiver Operating Characteristics (ROC) curve was also built and area under the ROC curve (AUC) was calculated. The higher the AUC, the better is the model at correctly classifying instance. Below shows the AUC and ROC of each model, where Gradient Boosting shows a higher AUC of 99.33% followed by Random forest with AUC of 99.14% and logistic regression has the lowest with 85.89%.

Fig 9: ROC curve and AUC score.

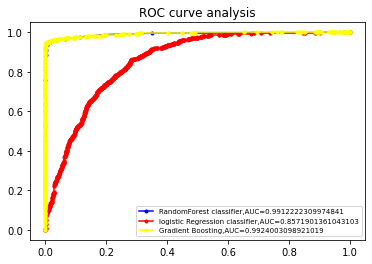


Table 2: Evaluation of the Models

The table below shows the evaluation of the three models used in this project.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Logistic**  **Regression** | **Random**  **Forest** | **Gradient**  **Boosting** |
| **Precision**  **(%)** | 0.776 | 0.967 | 0.970 |
| **Recall**  **(%)** | 0.775 | 0.966 | 0.969 |
| **F1-**  **Score(%)** | 0.775 | 0.966 | 0.969 |

# 3. CONCLUSION

Theanalysis of this study shows that intake of substance like smoking, alcohol, can increase the chances of having stroke. Marinating a balance diet can help slim the chance of have stroke by reducing cholesterol and blood pressure. According to study findings, it is possible to predict the occurrence of stroke using historical data mining approaches. Four (4) algorithms, Random forest, Logistic Regression, Neural Network and XGBoost classifier were used as the based models and gradient boosting (XGB) yielded the highest accuracy score with 96.98% compared to other models. Finally stacking was done in order to boost the accuracy score, and got an accuracy of 97.25%.

The proposed techniques can identify the critical stroke indicators from a large dataset for building effective prediction models for stroke prediction.

# REFERENCE

Asoka, S., Narayanan, S.G.S., Mandresh S., Vidhyasagar B.S. & Anand P.G. (2020). An Effective Prediction System using Predictive Models, *international Research Journal of Engineering and Technology*, 7(3), 3979-3985.

Baddam, S.S.R., Reddy, G.R., & Devi, B.V. (2022) Heart Stroke Prediction using Bagging and Boosting Classifiers, *international journal for research,*10 (5), 2300-2304.

Dritsas, E., & Trigka, M. (2022). Stroke Risk Prediction with Machine Learning Techniques. *Sensors*, *22*(13). <https://doi.org/10.3390/s22134670>

Emon, M. U., Keya, M. S., Meghla, T.I., Rahman, M, Mamun, M. S., & Kaiser M. S. A. (2020) Performant Analysis of Machine Learning Approaches in Stroke Prediction, *Fourth international conference on* electronics*, communication and Aerospace technology,*

Heo, T. S., Kim, Y. S., Choi, J. M., Jeong, Y. S., Seo, S. Y., Lee, J. H., Jeon, J. P., & Kim, C. (2020). Prediction of stroke outcome using natural language processing-based machine learning of radiology report of brain MRI. *Journal of Personalized Medicine*, *10*(4), 1–11. <https://doi.org/10.3390/jpm10040286>

Lozano, C.F., Hervella, P., Mato-Abad, V., Yanez, M. R., Garaboa, S. S., Dequidt, I. L., Gestal, A. E., Sobrino, T., Campos, F., Castillo, J., Yanez, S.R., & Rey, R.I., Random Forest-Based Prediction of Stroke Outcome, *Scientific Reports*, [https://doi.org/10.1038/s41598-021-89434-7.](https://doi.org/10.1038/s41598-021-89434-7)

Malathi, P., & Gokul, R. (2022), Stroke Prediction Using a Machine Learning Technique, *international journal of Advanced Research Trends in Engineering and Technology,* 9(8), 58-67.

Paul, D., Gain, G., Orang, S., Das, P., & Chaudhuri, A. K., (2022). Advanced Random Forest Ensemble for Stroke Prediction, *international journal of Advanced Research in computer and communication Engineering*, 11(2), 251-259.

Risk by Detection of Hemorrhage in Carotid Plaques: Meta-Analysis of Individual Patient Data.

*JACC: Cardiovascular Imaging*, *13*(2), 395–406. <https://doi.org/10.1016/j.jcmg.2019.03.028>

Schindler, A., Schinner, R., Altaf, N., Hosseini, A. A., Simpson, R. J., Esposito-Bauer, L., Singh,

N., Kwee, R. M., Kurosaki, Y., Yamagata, S., Yoshida, K., Miyamoto, S., Maggisano, R., Moody,

A. R., Poppert, H., Kooi, M. E., Auer, D. P., Bonati, L. H., & Saam, T. (2020). Prediction of Stroke

Sharma, C., Sharma, S., Kumar, M., & Sodhi, A., (2022). Early Stroke Prediction using Machine Learning, https:// [www.research.net](http://www.research.net/) /publication/359518651.