Machine Learning Approaches to Stroke Prediction based on Framingham Cardiovascular Study Dataset*

Jonas Chirindza[1] and Ritesh Ajoodha[2]

[1] School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa
Chirindzajonas424@gmail.com
[2] School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za
https://riteshajoodha.co.za/

Abstract. Stroke is the second biggest cause of death and long-term paralysis globally. It continues to be a significant health burden for both the elderly and national healthcare systems. Hypertension, heart illness, atrial fibrillation, diabetes, and other aspects of one's lifestyle are all potentially modifiable risk factors for a stroke. Then, putting machine learning concepts into practice over an existing health study dataset to effectively and accurately predict the occurrence of stroke will help with early intervention and treatment. In this study, we propose various machine learning methods for stroke prediction and compare them to available methods or approaches from other similar studies. Furthermore, we present a Naive Bayes probabilistic method that combines the concept of data imputation, class imbalance, and feature selection for stroke prediction, which achieves a greater area under the ROC curve than the Multilayered Perceptron neural network and the SVM proposed as the baseline methods for stroke prediction. In addition, neurologists can use our work to identify potential risk factors for stroke without any clinical trials methods. Finally, our methods can be applied to the clinical prognosis of other diseases, where data are often lacking and risk factors are poorly understood.

Keywords: Machine Learning, Naive Bayes, SVM, SMOTE, Feature selection, Stroke, Prediction.

1 Introduction

Stroke is the second leading cause of death worldwide and causes organ paralysis and sensory impairment in people over 60 in South Africa. Stroke prevention among people aged 60 and above is currently projected to be a key priority in South Africa to reduce its burden in the future. Early and accurate prediction

* Supported by National Research Foundation of South Africa

of stroke is essential and can aid in early treatment and intervention. About 6 million people die each year from stroke, and 57% of stroke-related deaths occur in people over the age of 60. Men account for at least 52% of deaths recorded each year. Hemorrhagic stroke accounts for 51% of deaths, while ischemic stroke accounts for the remainder. Ischemic and hemorrhagic strokes also cause the loss of more than 115 million healthy lives each year due to their prevalence. Additionally, in theory, predicting the occurrence of a stroke may seem easy, but in practice, accurate prediction of stroke requires effort and some machine learning skills. Neurologist are looking closely at methods and systems that can manipulate and optimise patient health data to predict the time frame of future strokes occurrences. As a result, this increases the need for algorithms and models to be used or developed for stroke predictions. Addressing this issue requires extensive research to define these algorithms and models.

Furthermore, there has been many attempts to predict stroke and identify its risk factors using patients medical data [13]. Recent studies indicate that stroke risk factors include diabetes, cardiovascular diseases, hypertension, anti-hypertensive drug use, systolic blood pressure, age, left ventricular hypertrophy, and atrial fibrillation, which is suggested as an independent risk factor for stroke [15,22]. Major risk factors for stroke incidence and prevalence are mostly due to diabetes and atrial fibrillation, with diabetes reporting more than 20 million diagnostics annually and atrial fibrillation reporting about 4 million diagnostics. Over the years similar studies [3,6,16] have been carried out, and the authors have further discovered more risk factors of stroke such as dyslipidemia, sickle cell disease, familial amyloid angiopathy, alcohol consumption, substance abuse, and smoking. The majority of prior models have chosen risk factors observations verified by research studies conducted by medical experts.

Machine learning algorithms and models can increase the accuracy of stroke prediction and discover new risk factors for stroke. Furthermore, these machine learning algorithms excel at identifying risk factors related to stroke. This paper investigates the risk factors for stroke and machine learning methods for improving stroke prediction accuracy. Our approach considers standard data imputation, class imbalance, feature selection, and prediction problems in the clinical dataset. Furthermore, we have extended our approach for stroke prediction to Multilayered perceptron neural networks, Support vector machine (SVM) and Naive Bayes.

We use feature selection methods to identify several features which are closely related to stroke. Information gain ranking algorithm is considered our default method for feature selection since it achieves the highest average area under ROC of 0.994 when evaluated through SVM, Multi-layered Perceptron and Bayesian Naive Bayes for stroke prediction. We train our prediction models using ten-fold cross-validation and assess their performance using the area under ROC curve and the confusion matrix. We discover that Multi-layered Perceptron perform well enough to predict stroke but inferior to the more common machine learning models described in this study, such as SVM or Bayesian Naive Bayes.

Our main contribution for this study are as follows:

1. Evaluation of pre-existing problems in data imputation, class imbalance, feature selection and prediction in Framingham cardiovascular study dataset.
2. Automated feature selection models, Information Gain Ranking and Correlation Attribute Evaluation which identifies the risk factors for stroke.
3. Oversampling technique to address the issue of class imbalances in medical datasets by increasing the data samples of the minority class.
4. Identifying new probable risk factors for stroke
5. A Predictive models to predict stroke based on patients pre-existing medical data.

This paper is organized as follows; section 2 highlights the contributions in the domain of predicting stroke and identifying its risk factors observations. It describes different methods used for which some of them are the starting point of our research. Section 3 describes various ways we considered in our approach such as data imputation, class imbalance, feature selection and prediction. Section 4 describes our experimental results and outlines major findings of our study. Lastly, section 5 provides summary and outlines our contribution in this study, and also provides future consideration on this study .

2 Related Work

Machine learning methods are widely used in medical studies and heavily in stroke prediction. Existing literature on stroke prediction and risk factors is extensively studied to learn more about numerous ideas connected to our current study. The following are major contributions for predicting and identifying stroke risk factors, and key conclusions are reached as a results of these studies.

2.1 Feature Selection

The authors of [13] used Forward feature selection method to reduce the susceptibility to over-fitting and produced an average performance rate of 0.75 using a linear kernel function of SVM. The Data imputation accuracy for missing data is obtained using the root-mean-square, mean absolute deviation, column median, and bias. The column median produces the highest imputation accuracy of 0.774 compared to 0.768 of the other three data imputation methods proposed. The decision tree algorithm for feature selection proposed by the authors of [19] produces a higher performance accuracy of more than 0.9 when compared to feature selection methods proposed by the authors of [13].

Most of these researchers used the Cardiovascular Health Study (CHS) [7] dataset which consist of more than 1000 features. Out of all studies presented in this paper, those who used machine learning algorithms for feature selection produced high accuracy in selecting components that are highly related to stroke as compared to those who manually selects features. Manually selecting features can produce poor selection accuracy because it is cumbersome to manually select the correct and significant features in a dataset of about 1000 features. In

many instances, the prediction accuracy of stroke is low when manually selected features are trained on prediction models [11,15,10]. The main reason is that only few features are selected; even though some of these features may be significant, they may not be well enough to improve the prediction accuracy. Machine learning methods for selecting features ensure that all selected features are substantial enough to help with stroke prediction. As a result, high prediction accuracy is guaranteed.

2.2 Learning Algorithms and Predictions

The authors of [14] present Automated hyper-parameter optimization based on Deep Neural Network (DNN) for stroke prediction on imbalanced health dataset. The method proposed produced a false accuracy of 0.191 and 0.716 for overall performance in stroke prediction. As compared to other available studies, the approach proposed performs better on imbalanced dataset than any other available traditional machine learning methods.

The authors of [19] uses Backpropagation neural network as the baseline method for stroke prediction based on features selected from the CHS dataset. Backpropagation neural network for stroke prediction produced an accuracy of more than 0.95. Other recent studies [10] considered logistics regressions, random forest algorithms, and deep neural network for stroke prediction . The features used to train these models consisted of patient demographics, medical history, previous disease, clinical variables, and laboratory values. For classification of these models, the ASTRAL [17] score is used as the reference and likelihood of the prevalence of stroke. The deep neural network produced an accuracy score of 0.888 higher than the ASTRAL score. Random forest algorithm and Logistics regression produced 0.839 and 0.849, respectively, which were lower than the ASTRAL score. The results from this papers [19,10] shows that neural networks seem to be producing better outcomes for stroke prediction compared to other machine learning methods proposed for stroke prediction.

The authors of [11,13] , proposes the Support Vector Machine as their baseline method for stroke prediction. The SVM proposed is applied on the training and testing samples obtained from the International Stroke trial Database (ISTB) [18]. The Support vector machine is optimised using the Radial basis function (RBF), polynomial, and linear and quadratic functions. The linear kernel function produces the highest prediction accuracy of 0.91 whilst RBF produces the lowest accuracy of 0.59. The linear SVM kernel proposed by the authors of [11] has proven to be superior to the one proposed by the authors of [13]. The following results indicate that the type of features which are selected have a great impact on models capabilities to predict stroke.

The Papers presented above demonstrated that machine learning methods can help us to accurately predict a stroke. They also indicate that neural networks, in particular, Backpropagation networks provide better stroke prediction than any other available methods. In addition, most researchers have shown that the most relevant features (risk-factor observations) are required to predict stroke accurately and that machine learning methods are more efficient in feature

selection. This paper extends the provided literature by showing how pre-existing medical data can improve stroke prediction accuracy and identify closely related features when the issues of class imbalances are addressed. In the next section we presents considered approaches explored by this paper.

3 Considered approaches

We describe a stroke prediction machine learning-based methods . Following steps are considered:

1. Impute the missing entries in the Cardiovascular study dataset using methodical techniques.
2. We apply the oversampling technique that increases the data-points of the minority class since class imbalance exist in our dataset.
3. We use an automatic approach is used to select the relevant feature subset.
4. We train several machine learning algorithms to assess the accuracy of their predictions.

3.1 Data Collection and pre processing

We are going to use the pre-existing Cardiovascular study dataset [6] from ongoing cardiovascular research on the citizens of Framingham, Massachusetts.This data contains medical information such as systolic blood pressure, diastolic blood pressure, glucose level, BMI, prevalent hypentension and total cholesterol. This medical information is based on the third-generation cohort consisting of about 4238 male and female enrolled participants of Framingham Cardiovascular study from 2002 to 2015.

3.2 Data Imputations Methods for Missing Data

- Mean: For each variable, calculate the mean of the observed values and use that mean to impute missing values for that variable.
- Median: For each variable, calculate the Median of the observed values and use that Median to impute missing values for that variable.

3.3 Class Imbalance and Feature Selection

Our dataset has an issue of class imbalance, we use Synthetic Minority Oversampling Technique (SMOTE) to address the issue of class imbalance. This approach generates the synthetic samples of class with minority data points. It adopts feature space, that interpolates between positive instances close to each other to generate new samples. The formal procedure of SMOTE is carried out in this manner. We set the total number of oversampling to be an integer, which will assist us in obtaining approximated distributions of classes as proposed by the authors of this paper [4]. Then, through a series of steps, an iterative

process is carried out. First, from the training set, a minority class instance is chosen at random. The next step is to find its K closest neighbours. Finally, K instances of oversampling are picked at random to determine additional examples via interpolation. The difference between the feature vector (sample) under examination and each of the K neighbours is employed, and before this difference can be added to the initial feature vector, we multiply it by an integer between 0 and 1. As a result of the above approach, we select a random location along the interpolation line segment connecting the features.

The are many features mentioned above that are closely related to stroke and contributes to stroke prediction. However, some of these features may not be help full for this particular task. We consider Correlation Attribute Evaluation (CAE) as one of our proposed feature selection method. This method selects features by simply assessing their importance concerning the target variable using Pearson's correlation method. It evaluates nominal properties on a value basis, with each value serving as an indication. Furthermore, we propose Information Gain Ranking method as one of our feature selection methods. Information Gain Ranking selects features by calculating their entropy. The entropy lies between a closed interval of 0 and 1, where 1 represents maximum information gain and 0 no information gain.

3.4 Classification Models and Evaluation Metrics

The Multi-Layered Perceptron (MLP). The MLP is a backpropagation neural network consisting of three layers, namely input, hidden and output. It measures the gradient of the cost function for each weight using the chain rule. Unlike a native direct calculation, it efficiently calculates one layer at a time. It computes the gradient but doesn't specify how it'll be used. Instead, it generalises the delta rule's computation, which computes derivatives using a simple back-propagation application of the chain rule [2].

The Support Vector Machine (SVM). SVM classifies the classification points with a hyper-plane. SVM also ensures that when a hyper-plane is created, two margin lines are also produced. SVM maximises the margin lines from both tags of the classification point. Points that are classified above the hyper-plane are considered the positive points, and those that are below the hyper-plane are considered the negative points. SVM chooses the hyperplane, which maximises the marginal distance. Support Vectors in SVM are points passing through the marginal planes that have been created parallel to the hyper-plane. Support vectors help us to determine the maximised distance of the marginal planes. In the case of non -linear separable, the SVM uses SVM kernels functions. The SVM used in this study follows the implementation by the authors of this paper [13].

Naive Bayes Classifier. Bayesian Naive Bayes is a classification method that use strategy Bayes theorem [20] and the assumption of predictor independence [1]. A Bayesian Naive Bayes classifier, inessential words, posits no features in a class are dependent on each other. Naive Bayes classifier is very effective in stroke prediction and it is widely used from classification problems. The Naive

Bayes used in this study follows similar implementation used by the authors of this paper [9].

We evaluate these three classification models using area under ROC curve and confusion matrix [5], and these model are trained and tested using a 10 fold cross validation [8].

3.5 Limitations

The Cardiovascular study dataset has a sample size of about 4000; thus, adequate sample size is required to produce more accurate and valid research results. It might be not easy to identify or recognise essential relationships from the Cardiovascular study dataset sample size. Having a larger sample size can assure that the sample considered is representative of the population and that the statistical outputs can be generalised to a larger population. If the sample size negatively impacts the research, we will consider choosing a more suitable sample size before doing similar research in future.

3.6 Ethical clearance

The proposed research does not require ethical clearance because it will not use any personal data or involve participants with limited capacity to consent. However, this research will make use of a publicly available dataset.

4 Experimental Results

In this section we presents the results obtained, as a result of applying our considered approaches for identifying risk factors and predicting stroke. This section is structured as follows: we present the results of data imputation, class imbalance, feature selection and lastly, we present the results of stroke prediction and list top ten core features identified as risk factors for stroke.

4.1 Dataset and Imputation

The Framingham Heart Study (FHS) [12] studies cardiovascular disease risk factors in the elderly. The cardiovascular research dataset obtained from this study is a valuable resource for learning risk factors and predicting stroke. According to [21] the success of the original cohort, which began in 1948, paved the way for epidemiological studies involving 5,209 men and women at the age of 28 to 62. The FHS participants were examined with 2-6 years follow up with over 15 attributes collected via questionnaires and medical examinations from 1948 until 1972 when their offspring cohort began. The cardiovascular study dataset used for our research is based on the third-generation cohort consisting of about 4238 male and female enrolled participants. The FHS investigators examined the third generation cohort from 2002 until 2015 with 2-6 years follow up. The quality of the Framingham cardiovascular study dataset makes it one

of the most used data for identifying risk factors and stroke prediction after the Cardiovascular Heart Disease (CHS) dataset [7].

No records were removed because the dataset had a small subset of missing values and records logged as unknown. We looked at suggested imputation methods to fill in missing values in a dataset. The imputation models were evaluated using 5-fold cross-validation, and the data set is split into an 8:2 ratio for the training and test sets. Performance metrics were calculated by comparing the actual and imputed values in the validation dataset. We performed the same process for all missing attributes and averaged the results. Out of the two proposed imputation methods, the imputation through column means was superior to imputation through column medians as it produced the smallest MAD, bias values and higher area under ROC curve.

To obtain the area under the ROC, we used Information gain ranking method to select features and Bayesian naive bayes for stroke predictions which achieves an area of 0.748 under the ROC curve. The overall predictive performance of stroke using the column mean imputation is the best. As a result, we present our results going forward using the column mean as the default imputation method. After preprocessing and imputation, the final dataset consists of 15 features and 4238 records with only 25 occurrence of stroke.

Table 1: A table describing the performance of data imputation methods when tested through Bayesian Naive bayes method for stroke prediction and Information gain ranking method for feature selection.

Method	Mean Absolute Deviation	Biases	ROC area
Column Mean	9.83	0.018	**0.748**
Column Median	9.87	0.065	0.728

4.2 Class Imbalance

From our dataset, we only have 25 class instances for the occurrence of stroke; thus, this affects the ability of our models to correctly classify the cases of the minority class, even though the accuracy of the model is kept at a high value. From imputation results, we used Bayesian Naive Bayes for stroke prediction to evaluate the performance of our imputation methods using the area under the ROC curve, and it produced an area of 0.748 with an overall accuracy of 0.982. Though the accuracy of the model is high, its ability to classifier instances of the minority class (Stroke occurrence) is poor as it misdiagnosis over 80% of them, as can be seen in the confusion matrix below, where class 1 represents the diagnosis of stroke.

We used SMOTE to address the issue of class imbalance by oversampling the minority class with 10000%. After applying SMOTE method, our final dataset

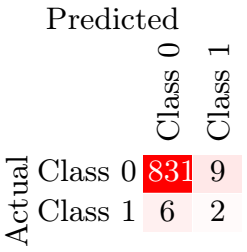


Fig. 1: A Confusion Matrix for describing the performance Bayesian Naive Bayes on a set of test data without application of SMOTE to address class imbalances. For class 1 which is the case of prevalence stroke, there are 2 correctly classified instances and 9 incorrectly classified ones.

consists of 6738 records with 2500 occurrences of stroke. To evaluate the performance SMOTE method, we used Bayesian Naive Bayes for stroke prediction, information gain ranking for feature selection and our default imputation method. The SMOTE produced area under the ROC curve of 0.999 with more than 90% correctly classified instances for stroke occurrence.

4.3 Feature Selection

For this study, we use two feature selection methods to help us identify important features to use for the training and testing of our proposed models.

1. The Information Gain Ranking for feature selection reduced our features from 15 to 8 based on a 0.5 selection threshold information gain ranking. We evaluated our method performance through SVM using 10-fold cross-validation. For prediction performance, we found an average area under the ROC curve of 0.995, which is an improvement from what the authors of this paper [15] obtained using SVM on 16 manually selected features. In our study, this method only selected features that are important and invaluable to stroke prediction.
2. Correlation Attribute Evaluation feature selection method reduced our features from 15 to 11 based on a 0.5 selection threshold. We then used the same approach used on point 1 above to evaluate our method performance. We found an area under the ROC curve of 0.994, slightly lower than the one we obtained when Using the Information gain ranking above. Furthermore, we notice that the Information gain ranking for feature selection technique yields the best results for SVM, Multilayered Perceptron and Bayesian Naive Bayes. Furthermore, we notice that the Information gain ranking for feature selection technique yields the best results for SVM, Multilayered Perceptron and Bayesian Naive Bayes.

Table 2: Average AUC for our proposed feature selection method with prediction models.

Feature Selection Method	Multi-Layered Perceptron	SVM	Bayesian Naive Bayes
Information Gain Rank	0.988	0.995	**0.999**
Correlation Attribute Evaluation	0.986	0.994	0.997

4.4 Stroke Prediction and Risk factors

Firstly, our model's performance were evaluated using the confusion matrix and area under the ROC curve. We discover that all of our feature selection methods produces excellent results when evaluated using our proposed metrics for for stroke prediction. Overall, the Information Gain Ranking for feature selection outperformed the Correlation Attribute Evaluation feature selection method for all of our prediction methods. We also found that the Bayesian Naive Bayes is superior to SVM and Multilayered perceptron for all feature selection methods described in section 3. From table 3 it is significant to note that SVM without SMOTE gives poor results than any other considered approaches in our study. This tells us that class imbalance affects the ability of models to classify instances of the minority class correctly.

Table 3: Average AUC for our best performing approaches with comparison to approaches from other similar studies using 10 fold cross validation.

Approach	Average AUC
SVM, SMOTE, Information Gain ranking	0.995
Naive bayes, Information Gain ranking	0.748
Naive bayes, SMOTE, Information Gain ranking	**0.999**
Multi-Layered peceptron, SMOTE, Information Gain ranking	0.988
SVM, Conservative mean feature selection (used by [13])	0.774
Margin-based censored algorithms(MCR), Conservative mean feature selection (used by [13])	0.774
SVM, 16 Manually Selected features (used by [13])	0.753

In addition to getting better results, our method can immediately identify potential risk factors without the need for lengthy medical studies to understand them fully. This will provide a fast way to characterise new diseases and determine predictors before further studies are identified. This approach can also be used to identify risk factors that were previously unnoticed. To obtain the best performance

in our study, we ranked the average merits obtained from information gain rank in descending order over. We used feature selection techniques to identify ten core features. As a result, we found tremendous agreement between the ten core features and those identified in clinical trials.

We found that some of these highly-rated features have not yet been identified as risk factors for stroke in clinical trials. Nevertheless, our results indicate that our method is accurate and effective in determining possible risk factors for stroke. Further studies of these characteristics may lead to more accurate predictions of stroke.

Table 4: Top ten identified features Information Gain Ranking method for feature selection

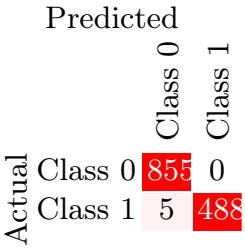
Feature Name	Average Merit
BMI	**0.921 +- 0.002**
Total Cholesterol	0.812 +- 0.002
Systolic Blood Pressure	0.75 +- 0.002
Diastolic Blood Pressure	0.68 +- 0.002
Glucose	0.652 +- 0.002
AGe	0.624 +- 0.003
Heart rate	0.583 +- 0.003
Prevalent Hypertension	0.341 +- 0.003
Cigarettes per Day	0.26 +- 0.002
Is smoking	0.19 +- 0.003

5 Discussions and Conclusion

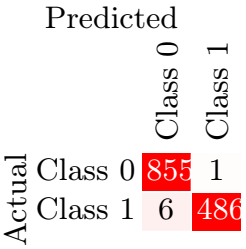
As we have seen in this study that the Information Gain Ranking performs well on Framingham cardiovascular study dataset. However, this feature selection method may not work with other data sets such as CHS because more than 1000 features are available. Our article presents several machine learning methods and combines data imputation, class imbalance, feature selection, and prediction elements. We also provide a detailed comparison of our approach with other similar studies available. Furthermore, we propose Information Gain Ranking for feature selection, which provides better performance than other proposed feature selection methods described in Section 3.

We also note from the confusion matrix (a) on figure 2 that Bayesian naive Bayes has proven to outperform other prediction methods proposed in this study by simply achieving a better area under the ROC curve. Furthermore, medical experts can use our work to identify potential risk factors for stroke without any clinical trials. As a result, our work can help neurologists accurately predict the future occurrence of stroke by simply optimising patient medical data using our approaches and will help with early intervention and stroke treatment, the health

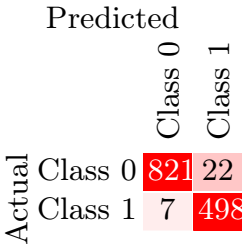
burden of stroke in elderly and national healthcare systems will also decline. We hope that this article will inspire many researchers to apply machine learning methods in medical data analysis.



(a) A Confusion Matrix for describing the performance of the Bayesian Naive Bayes model on a set of test data. For the class 1 which is the case of prevalence stroke, there are 488 correctly classified instances and 0 incorrectly classified ones



(b) A Confusion Matrix for describing the performance of the SVM on a set of test data. For the class 1 which is the case of prevalence stroke, there are 486 correctly classified instances and 1 incorrectly classified ones



(c) A Confusion Matrix for describing the performance of the Multi-Layered Perceptron on a set of test data. For the class 1 which is the case of prevalence stroke, there are 498 correctly classified instances and 22 incorrectly classified ones

Fig. 2: Confusion matrices assessing the performance of our 3 classification models.

5.1 Contribution, Recommendations, and Future Consideration

We provide machine learning methods to predict stroke by simply optimizing patients medical data. The main aim for this is to create a platform from our considered approaches that neurologists or medical experts could use to identify if whether a patient is more likely to have stroke or not. As we explained that stroke continues to be a significant health burden for both the elderly

and national healthcare systems. Using our proposed approaches will help with treatment on affected patients ,and the burden of stroke in elderly people and national healthcare systems will be reduced.

The implication of the above recommendation is that an early intervention and treatment for people who are affected or at risk of being affected by stroke would be carried out at an early stages. Future consideration of this research can (1) incorporate our proposed approach to application; (2) explore the top ten risk factors identified by Information Gain Ranking on table 4 into depth; and (3) use the CHS dataset from Cardiovascular Heart Study [7] to predict and identify the risk factors of stroke since it has more than 1000 features available.

Acknowledgements

Sincere gratitude to Kaggle for providing the Framingham cardiovascular study dataset. This work is based on the research supported in part by the National Research Foundation of South Africa (Grant numbers: 121835).

References

1. Ajoodha, R., Rosman, B.: Learning the influence structure between partially observed stochastic processes using iot sensor data. In: Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence (2018)
2. Arber, S., Hunter, J.J., Ross Jr, J., Hongo, M., Sansig, G., Borg, J., Perriard, J.C., Chien, K.R., Caroni, P.: Mlp-deficient mice exhibit a disruption of cardiac cytoarchitectural organization, dilated cardiomyopathy, and heart failure. Cell **88**(3), 393–403 (1997)
3. Boehme, A.K., Esenwa, C., Elkind, M.S.: Stroke risk factors, genetics, and prevention. Circulation research **120**(3), 472–495 (2017)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
5. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240 (2006)
6. Dawber, T.R., Meadors, G.F., Moore Jr, F.E.: Epidemiological approaches to heart disease: the framingham study. American Journal of Public Health and the Nations Health **41**(3), 279–286 (1951)
7. Fried, L.P., Borhani, N.O., Enright, P., Furberg, C.D., Gardin, J.M., Kronmal, R.A., Kuller, L.H., Manolio, T.A., Mittelmark, M.B., Newman, A., et al.: The cardiovascular health study: design and rationale. Annals of epidemiology **1**(3), 263–276 (1991)
8. Fushiki, T.: Estimation of prediction error by using k-fold cross-validation. Statistics and Computing **21**(2), 137–146 (2011)
9. Griffis, J.C., Allendorfer, J.B., Szaflarski, J.P.: Voxel-based gaussian naïve bayes classification of ischemic stroke lesions in individual t1-weighted mri scans. Journal of neuroscience methods **257**, 97–108 (2016)

10. Heo, J., Yoon, J.G., Park, H., Kim, Y.D., Nam, H.S., Heo, J.H.: Machine learning–based model for prediction of outcomes in acute stroke. Stroke **50**(5), 1263–1265 (2019)
11. Jeena, R., Kumar, S.: Stroke prediction using svm. In: 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). pp. 600–602. IEEE (2016)
12. Kannel, W.B., Gordan, T.: Evaluation of cardiovascular risk in the elderly: the framingham study. Bulletin of the New York Academy of Medicine **54**(6), 573 (1978)
13. Khosla, A., Cao, Y., Lin, C.C.Y., Chiu, H.K., Hu, J., Lee, H.: An integrated machine learning approach to stroke prediction. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 183–192 (2010)
14. Liu, T., Fan, W., Wu, C.: A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artificial intelligence in medicine **101**, 101723 (2019)
15. Lumley, T., Kronmal, R.A., Cushman, M., Manolio, T.A., Goldstein, S.: A stroke prediction score in the elderly: validation and web-based application. Journal of clinical epidemiology **55**(2), 129–136 (2002)
16. Manolio, T.A., Kronmal, R.A., Burke, G.L., O'Leary, D.H., Price, T.R.: Short-term predictors of incident stroke in older adults: the cardiovascular health study. Stroke **27**(9), 1479–1486 (1996)
17. Ntaios, G., Faouzi, M., Ferrari, J., Lang, W., Vemmos, K., Michel, P.: An integer-based score to predict functional outcome in acute ischemic stroke: the astral score. Neurology **78**(24), 1916–1922 (2012)
18. Sandercock, P.A., Niewada, M., Członkowska, A.: The international stroke trial database. Trials **12**(1), 1–7 (2011)
19. Singh, M.S., Choudhary, P.: Stroke prediction using artificial intelligence. In: 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON). pp. 158–161. IEEE (2017)
20. Swinburne, R.: Bayes' theorem. Revue Philosophique de la France Et de l **194**(2) (2004)
21. Tsao, C.W., Vasan, R.S.: Cohort profile: The framingham heart study (fhs): overview of milestones in cardiovascular epidemiology. International journal of epidemiology **44**(6), 1800–1813 (2015)
22. Wolf, P.A., Abbott, R.D., Kannel, W.B.: Atrial fibrillation as an independent risk factor for stroke: the framingham study. stroke **22**(8), 983–988 (1991)