Machine Learning Approaches To Stroke Prediction Based On The Framingham Cardiovascular Study Dataset

Jonas Chirindza
School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg
South Africa
ChirindzaJonas424@gmail.com

Ritesh Ajoodha
School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg
South Africa
Ritesh.Ajoodha@wits.ac.za

***Abstract*—Stroke is the second biggest cause of death and long-term paralysis in the world. 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. Our article considers common problems of data imputation, class imbalance, feature selection, and prediction in the clinical dataset. In addition, we present a Bayesian 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 ROC curve than the Multilayered Perceptron Backpropagation Neural network and the SVM proposed in this study for stroke prediction. Finally, our work can be used to identify potential risk factors for disease without clinical trials methods.**

I. 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 over the age of 60 is currently projected to be a key priority in South Africa to reduce the burden of stroke in the future. Early and accurate prediction 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.

Additionaly, 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. Neuroscientists are looking closely at methods and systems that can manipulate and optimise patient health data to predict the time frame of future strokes accurately. 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 for stroke prediction.

Furthermore, there has been many attempts to predict stroke and identify its risk factors using patients medical data [8]. 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 [10], [17]. 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 [1], [2], [11] 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 be used to increase the accuracy of stroke prediction and discover new risk factors for stroke. These machine learning algorithms excel at identifying risk factors closely related to stroke. This paper investigates the risk factors for stroke and machine learning methods for improving stroke prediction accuracy and our approach considers common problems of data imputation, class imbalance, feature selection, and prediction in the clinical dataset. We have extended our approach for stroke prediction to multi-layered perceptron neural network backpropagation and other non-regressive neural network models such as SVM and Bayesian naive Bayes.

We used 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 Per-

ceptron and Bayesian Naive Bayes for stroke prediction. We trained our prediction models using ten-fold cross-validation and assessed their performance using the area under ROC curve and the confusion matrix. We found 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 of 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 prediction stroke and identifying its risk factors observations. It describes different methods used for which some 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 .

II. 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.

A. Feature Selection

Most of these researchers used the Cardiovascular Health Study (CHS) [3] dataset with 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 chosen 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 [5], [7], [10]. The main reason is that only a few features are selected; even though some of these features may be significant, they may not be 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.

The authors of this paper [8] 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, with the L1 regularised logistic regression producing an average performance rate of 0.764 and the conservative mean selection-producing the highest performance rate of 0.754 for feature selection. The Data imputation accuracy for missing data was 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 of the mentioned methods for data imputation. The decision tree algorithm for feature selection used by the authors of this paper [14] produced a higher performance accuracy of more than 0.9 than all other feature selection methods proposed by the authors of this paper [8].

B. Learning Algorithms and Performance

Various studies, including [7], [8] , used the Support Vector Machine for stroke prediction. They applied SVM on the training and testing samples obtained from the International Stroke trial Database (ISTB) [13]. The Support vector machine is optimised using four kernel functions: the Radial basis function (RBF), polynomial, and linear and quadratic functions. From all these optimiser functions of the support vector machine, a linear kernel function produced the highest prediction accuracy of 0.91, with polynomial at 0.879, quadratic at 0.81, and RBF producing the lowest accuracy of 0.59. The linear SVM kernel used by the authers of this paper [7] also produces higher accuracy than the one produced by the linear SVM kernel used by [8]. The following results indicate that the type of features which are selected have a great impact on models capabilities for stroke prediction.

The Framingham model [9] and the CHS model were considered as the baseline for stroke prediction [10]. Features were manually selected from the CHS dataset: systolic blood pressure, age, cardiovascular disease, hypertension, diabetes, and atrial fibrillation. The CHS model's exponential and logarithmic forms were implemented, and the models are evaluated using the area under ROC curve performance metric. The CHS model observed the survival curves, which act as evidence that the model made an accurate stroke prediction in various groups of participants. The Framingham model is inferior to the CHS model for stroke prediction. However, considering both genders, these two models obtained an AUC score of 0.77, indicating that model performance is good. The Framingham model obtained an AUC score of 0.69 and 0.73 for men and women respectively, whilst the CHS model obtained an AUC score of 0.65 and 0.77 for men and women respectively. The results show that these models are more effective in predicting stroke for women than for men.

The authors of this paper [14] used Backpropagation neural network as the baseline method for stroke prediction using the selected features from CHS dataset. Backpropagation neural

network for stroke prediction produced an accuracy of more than 0.95. Other recent studies considered logistics regressions, random forest algorithms, and deep neural network for stroke prediction [5]. The features used to train these models consisted of patient demographics, medical history, previous disease, clinical variables, and laboratory values. For the classification of these models, the ASTRAL [12] score is used as the reference and likelihood of 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 [5], [14] shows that neural networks seem to be producing better outcomes for stroke prediction compared to other machine learning methods proposed for stroke prediction.

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, Back-propagation 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.

III. 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.

A. Data Collection and pre processing

We are going to use the pre-existing Cardiovascular study dataset [2] from ongoing cardiovascular research on the citizens of Framingham, Massachusetts. This data. 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.

B. Data imputations:

- 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.

C. 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 [18]. 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.

Input: Minority data $\mathfrak{D}^{(t)} = \{\boldsymbol{x}_i \in X\}$ where $i = 1,2,\dots,T$
Number of minority instances (T), SMOTE percentage (N), number of nearest neighbors (k)

for $i = 1,2,\dots,T$ **do**
1. Find the k nearest (minority class) neighbors of $\boldsymbol{x}_i$
2. $\widehat{N} = \lfloor N/100 \rfloor$

while $\widehat{N} \neq 0$ **do**
1. Select one of the k nearest neighbors, call this $\overline{\boldsymbol{x}}$
2. Select a random number $\alpha \in [0,1]$
3. $\widehat{\boldsymbol{x}} = \boldsymbol{x}_i + \alpha(\overline{\boldsymbol{x}} - \boldsymbol{x}_i)$
4. Append $\widehat{\boldsymbol{x}}$ to $\mathcal{S}$
5. $\widehat{N} = \widehat{N} - 1$

end while
end for
Output: Return synthetic data $\mathcal{S}$

Fig. 1. Shows a simple example of the SMOTE application

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.

D. Classification Models and Evaluation Metrics

The Multi-Layered Perceptron (MLP). The MLP is a back-propagation 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. To calculate the slopes of the cost function, firstly, an error denoted by δ^l is introduced to relate it to the gradients. The neural network for MLP is based on four equations, and those equations provide a way of calculating the slope and error of the loss function.

1) The equation of computing error:

$$\delta^l = (a^l - y) \odot \sigma'(z^l)$$

The first term represents the rate of change of the cost function, and the second term represents the change of activation functions with z^l and and calculates how fast the activation function would change at $\sigma'(z^l)$ would change at z^l.

2) The equation for computing error using the error of the next layer:

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$$

The first term denotes the weight matrix transpose, the second term denotes the error in the layer of $(l+1)^{th}$ iteration. Lastly, the third term represents the change of activation functions with z^l and and calculates how fast the activation function would change at z^l.

3) The equation for the derivative of the cost function with respect to any bias in the neural network:

$$\frac{\partial C}{\partial b^l_j} = \delta^l_j$$

Where C on the left hand side represents the cost function, b represent the bias in the network, and the right hand side represent the error δ^l at j^{th} iteration.

4) The equation for the derivative of the cost function with respect to weights of the neural network

$$\frac{\partial C}{\partial w^l_{jk}} = a^{l-1}_k \delta^l_j$$

here C on the left hand side represents the cost function, w^l_{jk} represent the weight in the network, a^{l-1}_k on the right hand side represent the activation, and the second term represent the error δ^l at j^{th} iteration.

The MLP Algorithm is as follows:

1) Input our training set, find the corresponding activation with the input layer.
2) We calculate how fast the activation function $\sigma'(z^l)$ would change at z^l.
3) Compute the error using equation (1).
4) We compute the backpropagation error using an equation (2).
5) We compute the gradient of the cost function using equation (3) and (4).

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 [8].

Bayesian Naive Bayes Classifier. Bayesian Naive Bayes is a classification method that use strategy Bayes rule and the assumption of predictor independence [24]. A Bayesian Naive Bayes classifier, inessential words, posits no features in a class are dependent on each other. The Bayes' Theorem can be represented using the equation below:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

1) $P(y|X)$ is the probability of the target variable given our prediction variables.
2) $P(X)$ is the prior or weight probability of our predictor.
3) $P(X|y)$ is the probability of predictors.
4) $P(y)$ is the prior probability of target variable.

Bayesian Naive Bayes classifier is very effective in stroke prediction and it is widely used from classification problems. The Bayesian Naive Bayes used in this study follows similar implementation used by the authors of this paper [25].

We are going to evaluate all of these three classification models using area under ROC curve and confusion matrix [26], and these model will trained and tested using a 10 fold cross validation [27] .

E. 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 Cardio-vascular 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.

F. 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.

IV. Experimental Results

In this section we presents the results obtained, as a results 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 features identified as risk factors for stroke.

A. Dataset and Imputation

The Framingham Heart Study (FHS) [22] 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 [23] 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 [3] .

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 model was evaluated using 5-fold cross-validation, and the data set was 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 and bias values.

Table 2: 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

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.

B. 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 misclassifies over 80% of them, as can be seen in the confusion matrix below, where class 1 represents the diagnosis of stroke.

Table 3: A Confusion Matrix for describing the performance of the SMOTE through Bayesian Naive Bayes on a set of test data.For the class 1 which is the case of prevalence stroke, there are 2 correctly classified instances and 9 incorrectly classified ones

Class 0	Class 1
831	9
6	2

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

C. 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 [10] 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.

Table 4: 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

Furthermore, we notice that the Information gain ranking for feature selection technique yields the best results for SVM, Multilayered Perceptron and Bayesian Naive Bayes..

D. Stroke Prediction and Risk factors

First, our model's performance was evaluated using the confusion matrix and area under the ROC curve. We discovered that all of our feature selection methods produced excellent results when evaluated using our proposed metrics for performance. 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 III. From table 5 it is significant to note that SVM without SMOTE gave poor results than 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.

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.

Table 5: Average AUC for our best performing approaches with comparison to approaches from other similar studies

Approach	Average AUC
SVM+ SMOTE + Information Gain ranking	0.995
Bayesian Naive Bayes+ Information Gain ranking	0.748
Bayesian Naive Bayes + SMOTE + Information Gain ranking	**0.999**
Multi-Layered Peceptron + SMOTE + Information Gain ranking	0.988
SVM + Conservative mean feture selecton (used by [8])	0.774
Margin-based censored algorithms (MCR) + Conservative mean feture selecton (used by [8])	0.774
SVM+ 16 Manually Selected features (used by [8])	0.753

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.

Table 6: 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

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.

V. 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. Furthermore, some of these features of CHS are highly correlated with each other, and this feature selection method evaluates features individually. 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 III.

Table 7: 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

Class 0	Class 1
855	0
5	488

Table 8: 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

Class 0	Class 1
855	1
6	486

Table 9: 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

Class 0	Class 1
821	22
7	498

We also note from the confusion matrix above on table 7 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 neuroscientists 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 you to apply machine learning methods to medical data analysis.

A. Contribution

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 neuroscientists or medical experts could use to identify if whether a patient is more likely to have stroke or not. This will help with early treatment and intervention. As we explained that stroke is the second biggest cause of death and long-term paralysis in the world. It 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.

B. Recommendations and future research

The implication of this recommendation is that an early intervention and treatment for people who are more likely to be affected by stroke would carried out because aspects of one's lifestyle are all potentially modifiable risk factors for a stroke. Future consideration of this research can (1) incorporate our proposed approaches to application; (2) explore the top ten risk into depth the factors identified by Information Gain Ranking in table 6; and (3) use the CHS dataset from Cardiovascular Heart Study [3] to predict and identify the risk factors of stroke since it has more than 1000 features available.

ACKNOWLEDGEMENTS

Sincere gratitude Kaggle for providing Framingham cardiovascular study dataset.

REFERENCES

[1] Amelia K Boehme, Charles Esenwa, and Mitchell SV Elkind. Stroke risk factors,genetics, and prevention.Circulation research, 120(3):472–495, 2017.

[2] Thomas R Dawber, Gilcin F Meadors, and Felix E Moore Jr. Epidemiologicalapproaches to heart disease: the framingham study.American Journal of Public Health and theNations Health, 41(3):279–286, 1951.

[3] Linda P Fried, Nemat O Borhani, Paul Enright, Curt D Furberg, Julius M Gardin,Richard A Kronmal, Lewis H Kuller, Teri A Manolio, Maurice B Mittelmark, Anne Newman,et al. The cardiovascular health study: design and rationale.Annals of epidemiology, 1(3):263–276, 1991.

[4] N Gayatri, S Nickolas, AV Reddy, S Reddy, and AV Nickolas. Feature selectionusing decision tree induction in class level metrics dataset for software defect predictions. InProceedings of the world congress on engineering and computer science, volume 1, pages 124–129. Citeseer, 2010.

[5] JoonNyung Heo, Jihoon G Yoon, Hyungjong Park, Young Dae Kim, Hyo Suk Nam,and Ji Hoe Heo. Machine learning–based model for prediction of outcomes in acute stroke.Stroke,50(5):1263–1265, 2019.

[6] Sun Ha Jee, Ji Wan Park, Sang-Yi Lee, Byung-Ho Nam, Hwang Gun Ryu, Su YoungKim, Youn Nam Kim, Ja Kyoung Lee, Sun Mi Choi, and Ji Eun Yun. Stroke risk predictionmodel: A risk profile from the korean study.Atherosclerosis, 197(1):318–325, 2008.

[7] R. S. Jeena and S. Kumar. Stroke prediction using svm. In2016 International Conference on Control, Instrumentation, Communication and Computational Technologies(ICCICCT), pages 600–602, 2016.

[8] Aditya Khosla, Yu Cao, Cliff Chiung-Yu Lin, Hsu-Kuang Chiu, Junling Hu, andHonglak Lee. An integrated machine learning approach to stroke prediction. InProceedings ofthe 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages183–192, 2010.

[9] Dominique Laurier, Nguyen Phong Chau, Bernard Cazelles, Patrick Segond, andPCV-METRA Group12. Estimation of chd risk in a french working population using a modifiedframingham model.Journal of clinical epidemiology, 47(12):1353–1364, 1994.

[10] Thomas Lumley, Richard A Kronmal, Mary Cushman, Teri A Manolio, andSteven Goldstein. A stroke prediction score in the elderly: validation and web-based applica-tion.Journal of clinical epidemiology, 55(2):129–136, 2002.

[11] Teri A Manolio, Richard A Kronmal, Gregory L Burke, Daniel H O'Leary, and Thomas R Price. Short-term predictors of incident stroke in older adults: the cardiovascular healthstudy.Stroke, 27(9):1479–1486, 1996.

[12] G Ntaios, M Faouzi, J Ferrari, W Lang, K Vemmos, and P Michel. An integer-based score to predict functional outcome in acute ischemic stroke: the astral score.Neurology,78(24):1916–1922, 2012.
[13] Peter AG Sandercock, Maciej Niewada, and Anna Członkowska. The interna-tional stroke trial database.Trials, 12(1):1–7, 2011.
[14] M Sheetal Singh and Prakash Choudhary. Stroke prediction using artificial intelligence. In2017 8th Annual Industrial Automation and Electromechanical EngineeringConference (IEMECON), pages 158–161. IEEE, 2017.
[15] Gary Stein, Bing Chen, Annie S Wu, and Kien A Hua. Decision tree classifier fornetwork intrusion detection with ga-based feature selection. InProceedings of the 43rd annualSoutheast regional conference-Volume 2, pages 136–141, 2005.
[16] Jon M Sutter and John H Kalivas. Comparison of forward selection, back-ward elimination, and generalized simulated annealing for variable selection.Microchemical jour-nal, 47(1-2):60–66, 1993.
[17] Philip A Wolf, Robert D Abbott, and William B Kannel. Atrial fibrillation as anindependent risk factor for stroke: the framingham study.stroke, 22(8):983–988, 1991.
[18] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer.Smote: synthetic minority over-sampling technique.Journal of artificial intelligence research,16:321–357, 2002.
[19] Alberto Fernandez, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla.Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniver-sary.Journal of artificial intelligence research, 61:863–905, 2018.
[20] S Gnanambal, M Thangaraj, VT Meenatchi, and V Gayathri. Classificationalgorithms with attribute selection: an evaluation study using weka.International Journal ofAdvanced Networking and Applications, 9(6):3640–3644, 2018.
[21] William B Kannel, Daniel McGee, and Tavia Gordon. A general cardiovascularrisk profile: the framingham study.The American journal of cardiology, 38(1):46–51, 1976.
[22] WILLIAM B Kannel and T Gordan. Evaluation of cardiovascular risk in theelderly: the framingham study.Bulletin of the New York Academy of Medicine, 54(6):573, 1978.
[23] Connie W Tsao and Ramachandran S Vasan. Cohort profile: The framinghamheart study (fhs): overview of milestones in cardiovascular epidemiology.International journalof epidemiology, 44(6):1800–1813, 2015.
[24] Ritesh Ajoodha and Benjamin Rosman. 2018. Learning the influence structure between partially observed stochastic processes using IoT sensor data. In Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence.
[25] Joseph C Griffis, Jane B Allendorfer, and Jerzy P Szaflarski. Voxel based gaussian naıve bayes classification of ischemic stroke lesions in individual t1-weighted mri scans.Journalof neuroscience methods, 257:97–108, 2016.
[26] Jesse Davis and Mark Goadrich. The relationship between precision-recalland roc curves. InProceedings of the 23rd international conference on Machine learning, pages233–240, 2006.
[27] Tadayoshi Fushiki. Estimation of prediction error by using k-fold cross-validation.Statistics and Computing, 21(2):137–146, 2011.