Stroke Prediction and Analysis

Wangsing Chan, Dayu Qin

***Abstract*—Machine learning (ML) was used to forecast the development of stoke, which was trained on 300 original variables related to stroke, obtained from an annual health survey collected by the Behavioral Risk Factor Surveillance System (BRFSS) 2021. Random Forest was utilized to perform feature selection with the new features being validated with prerequisite medical knowledge. Performances of all the learning algorithms on the training set were compared with LightGBM achieving the lowest MSE but due to its instability, XGBoost was chosen for the stroke prediction instead. This model was able to predict a stroke onset rate of 4.1% on the test set, showing its reliability given that the real stroke onset rate was around 3%. The same process was repeated with a new dataset where only samples with heart disease were included. K-Means clustering was also adopted to the dataset but there was no visual or numerical indication of possible hidden relation within the heart disease patients. Ultimately, XGBoost obtained a high F1 score of 0.84 on the heart disease patient training set and predicted an stroke onset rate of 14% for the test set.**

I. Introduction

Accurate prediction of stroke is crucial for early intervention and treatment. A stroke occurs when something blocks the blood supply to the brain or a blood vessel in the brain bursts. Given that "Time lost is brain lost", every minute counts hence it is important to develop a stroke prediction system as soon as possible for early screening while utilizing information provided by related risk factors. There has been many medical studies and data analysis attempts to classify the main predictors of stroke. Past studies have reported numerous of stroke risk factors, including age, diabetes mellitus, cigarette smoking or even creatine levels. [1]

Previous works on stroke prediction involved models adopting risk factors that were selected manually by medical experts but this usually means that their prediction models only utilized a small amount of features. Furthermore, traditional prediction methods use the Cox proportional-hazards model (a regression model used for investigating the association between the survival time of patients and one or more predictor variables) [2], but this method strongly depends on quality of pre-determined features [3], meaning the model also depends on current existing medical knowledge.

Data collected from clinical trials are often high-dimensional, censored, heterogeneous and contain missing information, presenting challenges to traditional statistical analysis. [4] To address this problem, a machine learning approach can be used instead to adapt to identify features highly related to stroke occurrences from the large data set that would otherwise be too inefficient to do manually.

A. Related Works

Most of the previous machine learning approaches for medical studies used the Cox Proportional-Hazards Model for bench-marking. Katten et.al [5] compared it with several machine learning methods including neural networks on a urology data set. Nonetheless, only five features were used within these simple machine learning methods. Khosla et al [6] also compared the Cox Hazard Model with a more advanced supervised machine learning method such as Support Vector Machines (SVM) at the same time combined with a feature selection algorithm based on conservative mean. However, this model was limited using labelled data meaning some hidden relations within the data set may be neglected.

B. Our approach

- Imputing missing data in the data set with a systematic method.
- Feature selection from the data set
- Applying the best performing supervised machine learning to perform stroke Applying the best performing supervised machine learning to perform stroke occurrence prediction on the whole data set occurrences prediction on the whole data set
- Applying the best performing supervised machine learning to perform stroke occurrence prediction on heart disease patients

II. Methodology

A. Dataset

The dataset contains 300 original variables related to stroke, obtained from an annual health survey collected by the Behavioral Risk Factor Surveillance System (BRFSS) in 2021. Furthermore, the training dataset for stroke prediction is labelled.

B. Performance Metrics

*a) **Mean Squared Error (MSE)**:* MSE is one of the most important evaluation metrics for checking any classification model's performance. It can be obtained via finding the squared difference between the actual and predicted value. This metric represents the absolute measure of goodness of fit, the lower the MSE the better the model performs.

*b) **F1 Score**:* F1 score is considered as the most all-encompassing method for assessing performances of classifiers, it can minimize both false negatives as well as false positives simultaneously, which can be accomplished by taking the harmonic mean of precision and recall. Since precision and recall is required to obtain the harmonic mean, each class possesses the same weight during the average calculation, outputting a truly balanced mean (different class size are of equal importance) so if either precision or recall is low, it affects the overall F1 Score. This is especially important

in the medical diagnosis domain as it takes into account both sensitivity. In binary classification, recall is known as sensitivity, it is the proportion of true positive elements in the actual positives. Calculated by dividing the true positives by the sum of actual positives (True positive + False negative) is crucial for deciding how well the model can conduct the binary stroke classification task.

C. Missing Data Imputation

Before classification is performed, feature engineering was conducted. The data set retrieved from the BRFSS contained a lot of missing data, which was expected as clinical data often has omission stemming from either patients lost to follow-up, partially filled-out surveys or incomplete medical records. Knowing that missing data can compromise the validity of this model, data imputation was used to remedy the missing data. [7] This was done via:

- Removing feature columns with more than 10% of missing values.
- Removing duplicated rows.
- Column mean: replacing each missing value with the mean of the feature's observed values

D. Feature Selection

Some features of the clinical data may not contribute much to the model such as phone number, data collection date or "when its a safe time to talk" and should be dropped, this helps as high-dimensional feature vectors impose a high computational cost as well as slow training time. Feature selection was introduced as it addresses the dimensionality reduction problem by selecting a subset of the input features, which is most relevant for predicting stroke. A wrapper-based technique was used for feature selection. This feature selection technique is based on backward elimination and is known as the recursive feature elimination (RFE). However, due to previous testing, where feature contributions calculation performance were similar to that of a Random Forest model, hence Random Forest was used instead of RFE to calculate feature contribution. Using the Random Forest algorithm from Sklearn, the Gini importance can be computed to view each feature contribution to stroke. For each feature, the algorithm collects how on average each feature decreases the impurity of each node and the average over all trees in the forest is the measure of the feature importance. Subsequently, a feature importance threshold was set to eliminate redundant features with the aid of manual approval based on common medical knowledge.

E. Learning Algorithms

a) ***XGBoost****:* XGBoost is a popular open-source implementation of the gradient boosted trees algorithm, it is a supervised learning algorithm that is widely used for regression as well as classification. It's called gradient boosting since it adopts a gradient descent algorithm to minimizes the loss when adding new trees, and a learning rate is also applied. The training proceeds iteratively, adding new trees that predict the residuals or errors of prior trees, these trees are then combined with previous trees to produce predictions for each sample. [8] In summary, when performing binary classification, XGBoost first calculates similarity scores and gain to determine how to split the data. Subsequently, the trees can be pruned by finding the difference between the gain values and a tree complexity parameter Gamma, then an output value, in this case, whether a patient has stroke or not, can be determined. It is also worth noting that this model contains another hyper-parameter lambda, a regularization parameter that also affects the similarity scores. More importantly, during classification, the minimum number of residuals in a leaf is related to a metric called Cover. Compared to the normal gradient boost algorithm, it has more effective tree pruning and also possess a regularization variable preventing the model from over-fitting.

b) ***K Nearest Neighbor (K-NN)****:* The K nearest neighbor (K-NN) is a classification method belonging to the family of supervised machine learning algorithms, where the sample is classified based on its k nearest neighbors. This is particularly useful as it is intuitive and simple to implement, moreover it involves no training phase since it doesn't build a model but simply labels new data entries based on historical sample class near it. Furthermore, K-NN is a non-parametric algorithm, meaning there are no assumptions to be met by the training data for implementation, making it useful for non-linear data problems, hence it was also tested for the stroke prediction. Not to mention its flexibility in allowing one to choose the distance criteria for the K-NN model, whether be Euclidean distance or Manhattan distance etc.

c) ***Support vector machines (SVM)****:* Support vector machines (SVM) are believed to be one of the best "off-the-shelf" supervised learning algorithms. SVM uses a decision boundary aka a hyperplane to separate the two classes. Its main goal is to find the best hyperplane which maximizes the distance between the classes. The official term for the distance that one is trying to maximize during SVM is the margin, and that the two marginal hyper-planes are of same distance from the optimal hyperplane. Some of the observations from the training set can be classified as support vectors and these support vectors are responsible for defining the optimal hyperplane. Furthermore, these support vectors usually lie on the marginal hyper-planes. It also indicates that the rest of the data points in the training set are irrelevant, in the sense that if the rest of the data points were to move around, it wouldn't affect the decision boundary. Kernel functions are used to systematically find support vector classifiers in the higher dimensions, and the stroke prediction can be formulated as a binary classification problem that fits into the framework of SVM.

d) ***Multi-linear Regression****:* Multi-linear Regression is an algorithm that attempts to model two or more independent variables and a response variable by fitting a linear equation to the observed data. It is a straight forward and simplistic model with low computing time. This also utilizes the ordinary least squares method and this algorithm was also used for stroke prediction.

e) ***LASSO regression****:* Some prerequisite knowledge for LASSO includes regularization. L1 and L2 Regularization

techniques can be used to address over-fitting in feature selection. Regression model that uses L1 regularization is called Lasso (Least Absolute Shrinkage and Selection Operator) Regression, it adds the absolute values of the magnitude of the coefficients as a penalty term to the loss function. It uses cross-validation to choose the penalty factor, assuring that the model will generalize well to future data samples. This method was also adopted to the stroke prediction problem for performance comparison against the aforementioned models.

f) ***Unsupervised learning****:* As mentioned before, clustering analysis will be performed on the heart disease patients. The K-means clustering was adopted to the heart disease patients data set. The K-means algorithm partitions all the samples into K clusters with each entry belonging to the cluster with the closest mean. It does this by finding the K number of centroids & allocating every point based on the centroid. Results from clustering will then be fed back into the best performing supervised learning algorithm above.

III. Experimental Results

A. Data preprocessing & feature selection

After the Removing feature columns with more than 10% of missing values and duplicated rows, the data set reduced from 299 features to 123 features, no further missing values were detected. Top 12 important features were outputted from the Random Forest model with the top 3 features being ”HeartDisease”, ”X_LLCPWT” and ”IDATE” all with importance of 0.036, 0.035 and 0.033 respectively. Skewness corrections were also performed on a few of the quantitative feature, such as “X_LLCPWT”, “X_STRWT”, and “WTKG3”. The categorical and quantitative features were then separated in order to encode the categorical features. Subsequently, winsorization based on the Tukey rule was performed on the data as well as normalization

Another feature selection was conducted again using the Random Forest on the preprocessed data set and the top 10 results can be visualized in figure 1 below.

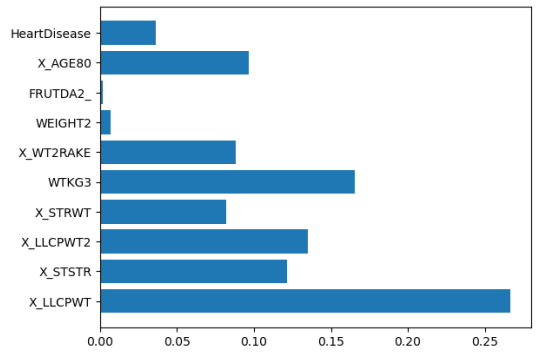

Fig. 1: Top 10 features with the importance from the original data set.

Features such as ”X_LLCPWT”, ”X_STRWT” & ”WEIGHT2” all denote variables that are weight or are variables required for weight sampling. For example, ”X_STRWT” represents stratum weight which was needed for the design weight calculation. It was demonstrated above that variables related to weight has the highest importance out of all the related features. This was expected since It is commonly known that obesity is typically associated with independent risk characteristics of stroke such as high blood pressure, high blood glucose and atherogenic serum lipids. Therefore being categorised as overweight increases risk of stroke by 22% and for obese individuals risk increases by 64%. [9]

The second most relevant feature that was not associated with weight was ‘X_AGE80’ which denotes imputed age value collapsed above 80. Aging is the most robust yet immutable risk factor relating to stroke. This can be explained by the fact that risk factors like hypertension, atrial fibrillation and coronary artery disease increases with age, in turn the risk of stroke also increases. [10] Moreover, with aging, both cerebal micro- and macro- circulations undergo functional and structural alterations. As for the ”HeartDisease” feature, it is common knowledge that heart disorders like coronary artery disease may increase one’s risk of stroke due to plague building up in the artery, reducing the oxygen-rich blood flowing to the brain.

”FRUITDA2” denotes the fruit juice intake times per day but studies have indicated that interactions between intake of fruit and vegetables and ischemic stroke are not significant. [11] Therefore, since it was one of the most contributing factors identified by the model, it may be due to the fact that individuals who consume more fruit and vegetable instead of junk food will be less prone to obesity, which also reduces the risk of stroke.

B. Task 1: stroke prediction on the ***whole dataset***

The performances of the prediction algorithms were evaluated based on the mean MSE, this is indicated in figure 2 below.

Algorithm	Mean (std Dev) of MSE
Multiple Linear Regression	0.035 (0.001)
Lasso Regression	0.037 (0.001)
Ridge Regression	0.035 (0.001)
ElasticNet Regression	0.037 (0.001)
KNN	0.042 (0.001)
Multilayer Perceptron	0.076 (0.001)
Random Forest	0.036 (0.001)
XGBoost	0.038 (0.001)
AdaBoost	0.036 (0.001)
LightGBM	0.124 (0.092)

Fig. 2: Average training MSE using different algorithms on whole dataset.

Comparing the performance amongst numerous learning algorithms, it could be seen that LightGBM had achieved the lowest MSE out of all the tested learning algorithms, 0.0124. However, its standard deviation was also the highest, 0.092 compared to the rest of the algorithms with a standard deviation of 0.001, thus it may not be suitable for stroke prediction given its instability. The lowest performing algorithm was the Multi-layer Perceptron, hence it will not be adopted for the final stroke prediction problems. The rest of the algorithms achieved similar performances, a MSE of approximately 0.04.

Due to XGboost's ability to provide a more direct route to the minimum error, allowing it to converge with fewer steps, overall lowering computing cost and time compared to other gradient boosting approaches, it was chosen for the final stroke prediction. In addition, the training set indicated that the stroke vs non-stroke classes were not balanced, hence classic modelling approaches such as Multiple Linear Regression or Random Forest were not chosen.

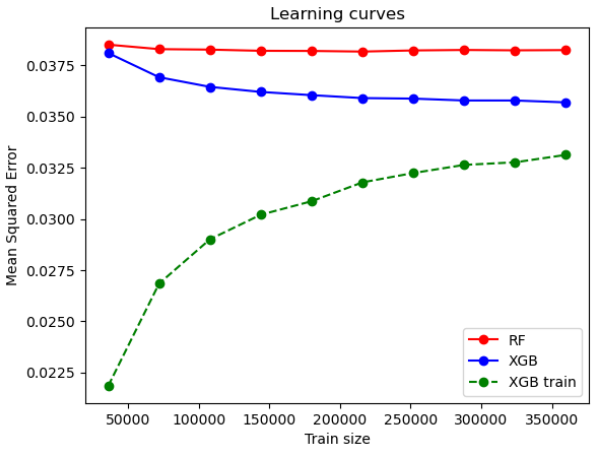

Fig. 3: Learning curve of XGBoost & Random Forest on the training and test set of the whole population.

Figure 3 above showed the learning curve of XGBoost and Random Forest on the test set, with the green line denoting the learning curve of XGBoost on the training set. Although Random Forest converged faster, the XGBoost was able to obtain lower MSE values as well as showed a more obvious converging trend. Moreover, the training learning curve and the test learning curve of XGBoost didn't overlap, showing its ability to avoid over-fitting. The XGBoost algorithm was adopted to the stroke prediction problem on the test set and was able to predict a stroke onset rate of 4.1% . This was a reasonable number as it was mentioned in the briefing that the stroke onset rate in the provided data set was around 3%.

*C. Task 2: stroke prediction on the **heart disease patients***

A similar process was repeated on the heart disease patients. Samples diagnosed with heart diseases were extracted from the original data set, and the new total samples reduced to 32,353. The same data imputation process was conducted and the new number of total features was 132. Feature selection was also performed on the heart disease patients and the most important features were identical to the previous task.

However, there was one feature that was different the previous task, which was "POTATOE1" with an importance of 0.02, as presented on the right in figure 4. This variable represented the consumption of french fry, and since french fry can be categorized as junk food, it also contributes to obesity which in turn is a relating stroke risk factor, thus validating the reliability of the Random Forest model.

Stroke prediction was then conducted on heart disease patients training set with the same algorithms used in the previous task, as demonstrated in figure 5 on the right.

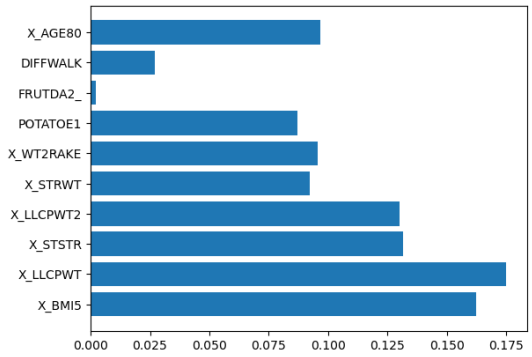

Fig. 4: Top 10 features with the importance from the heart disease patients data.

Algorithm	Mean (std Dev) of MSE
Multiple Linear Regression	0.137 (0.005)
Lasso Regression	0.136 (0.005)
Ridge Regression	0.137 (0.005)
ElasticNet Regression	0.136 (0.005)
KNN	0.161 (0.007)
Multilayer Perceptron	0.272 (0.007)
Random Forest	0.136 (0.004)
XGBoost	0.141 (0.003)
AdaBoost	0.140 (0.004)
LightGBM	0.216 (0.032)

Fig. 5: Average training MSE using different algorithms on heart disease patients.

Comparing the performance of numerous learning algorithms, it can be seen that Random Forest achieved the best result with a mean MSE of 0.136 with the lowest standard deviation of 0.004. The lowest performing algorithm was LightGBM with the highest MSE 0.216 and the highest standard deviation 0.032, which validates the previous assumption made on the LightGBM's instability. Again due to the rest of the algorithms having similar score, XGBoost was again used for stroke prediction on heart disease patients. **K-Means clustering** was then applied on the the data set, with k=2 and k=10 to try to discover possible hidden groupings within the hear disease patients. However there were no visible nor numerical indication of such relation, perhaps indicating that there are no hidden groupings within the heart disease patients.

The learning curve of XGBoost and Random Forest on the test and training data set can be seen in figure 6 below. However, this time the Random Forest converged quicker and had a lower MSE compared to the XGBoost. Nonetheless, the XGBoost's learning curve on the test set still didn't overlap on the training set, which again displays its capability to avoid over-fitting.

XGBoost was able to achieve a stroke onset rate of 16.3% and a F1 score of 0.84 for the heart disease patients from the training data. This high F1 score demonstrated the model's high sensitivity for stroke prediction. The final stroke onset rate predicted for the test set within the heart disease patients was 14%.

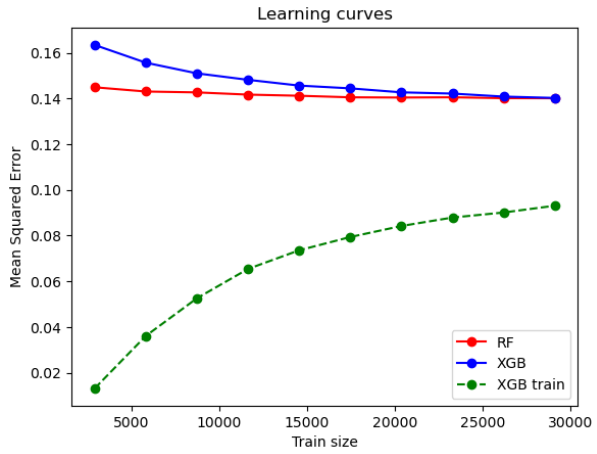

Fig. 6: Learning curve of XGBoost & Random Forest on the training and test set of the heart disease patients.

IV. Conclusion

In summary, an integrated machine learning approach to forecast the development of stroke was developed which utilized both supervised as well as unsupervised machine learning. Stroke prediction was performed on the original by first performing feature selection utilizing the importance obtained from the Random Forest model. The top 10 features with the highest importance values were verified with prerequisite medical knowledge, such as how features related to weight were all related to stroke. After performing normalization and winsorization, the performances of all the learning algorithms on the training set were compared with LightGBM achieving the lowest MSE but due to its instability, XGBoost was chosen for the stroke prediction instead. XGBoost was able to retrieve a stroke onset rate of 4.1% on the test set which was reasonable as the real onset rate was around 3%, validating the reliability of the model.

The same process was repeated with a new data set where only samples with heart disease were included. The new data set was reduced to 32,353 samples and new number of total features was 132. During feature selection, the model outputted the same top 10 features as the previous task but contained a single different feature. This feature was "POTATOE1" and was also validated with medical knowledge as an important stroke risk factor. Afterwards, the learning algorithms were compared on the training set of heart disease patients and Random Forest achieved the best result, with the lowest MSE of 0.136. The learning curve of XGBoost and Random Forest was also visualized, showing that even though Random Forest had converged quicker and had a lower MSE, the learning curve of XGBoost on the training and test data set did not overlap, proving its robustness to over-fitting. Then, XGBoost was utilized for stroke prediction on the heart disease patients and a stroke onset rate of 16.3% as well as an F1 score of 0.84 was obtained. Furthermore, its stroke onset rate predicted for the heart disease patients from the test set was 14%.

V. Future Prospects

More supervised as well as unsupervised learning algorithms will be performed, such as FLD and SVM utilizing various kernels. Moreover, deep learning will be experimented for further dimensionality reduction. Furthermore, unsupervised learning will be repeated on the subset of the original data set but instead of heart disease patients, it may be a subset based on other features such as a certain age group or frequent smokers, to identify possible hidden groupings.

References

[1] W. Longstreth, C. Bernick, A. Fitzpatrick, M. Cushman, L. Knepper, J. Lima, and C. Furberg, "Frequency and predictors of stroke death in 5,888 participants in the cardiovascular health study," *Neurology*, vol. 56, no. 3, pp. 368–375, 2001.

[2] D. R. Cox, "Regression models and life-tables (with discussion)," *J. Roy. Statist. Soc. Ser. B*, vol. 34, pp. 187–220, 1972.

[3] T. A. Manolio, R. A. Kronmal, G. L. Burke, D. H. O'Leary, and T. R. Price, "Short-term predictors of incident stroke in older adults: the cardiovascular health study," *Stroke*, vol. 27, no. 9, pp. 1479–1486, 1996.

[4] A. Spooner, E. Chen, A. Sowmya, P. Sachdev, N. A. Kochan, J. Trollor, and H. Brodaty, "A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction," *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.

[5] M. W. Kattan, "Comparison of cox regression with other methods for determining prediction models and nomograms," *The Journal of urology*, vol. 170, no. 6S, pp. S6–S10, 2003.

[6] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 183–192.

[7] M. Marino, J. Lucas, E. Latour, and J. D. Heintzman, "Missing data in primary care research: importance, implications and approaches," *Family Practice*, vol. 38, no. 2, pp. 199–202, 2021.

[8] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[9] World Stroke Organization. [Online]. Available: https://www.world-stroke.org/assets/downloads/WSO_DontBeTheOne_PI_Leaflets_-_WEIGHT.pdf

[10] M. Yousufuddin and N. Young, "Aging and ischemic stroke," *Aging (Albany NY)*, vol. 11, no. 9, p. 2542, 2019.

[11] K. J. Joshipura, A. Ascherio, J. E. Manson, M. J. Stampfer, E. B. Rimm, F. E. Speizer, C. H. Hennekens, D. Spiegelman, and W. C. Willett, "Fruit and vegetable intake in relation to risk of ischemic stroke," *Jama*, vol. 282, no. 13, pp. 1233–1239, 1999.