A Review On Stroke Risk Analysis

George Joseph, Bidusmita Das

Abstract: Stroke has become one of the most important causes of premature death and disability in low-income and medium- income countries. A possible reason could be the demographic changes and the increasing importance of modifiable risk factors. Poor people, however, are the most affected ones because of lack of funds. Recently however, countries like India, have observed a substantial rise in data related to stroke. Comprehensive study on the stroke risk can help in reducing the fatalities that occur every year due to stroke. This review represents in detail a literature survey on data collection, pre-processing and the various techniques that have been used for classification.

Index Terms: Analysis, Data Mining, Machine Learning, Risk, Stroke, Review, Survey, Lifestyle, Ischemic

1 INTRODUCTION

Learning about the factors that determine the risk of stroke requires a thorough understanding of the domain of environmental and physiological factors that affect stroke health. Substantial research has been performed to study the various features that affect the probability of stroke risk in an individual. Although these studies show strong results and has enabled the medical community in solving many real-world stroke related problems, the fact remains true that the poorer end of the worlds population are unable to receive these healthcare benefits because of being financially weak. In order to help many people have an easy access to healthcare, several Computer Aided Systems are used to help predict risk of stroke based on active factors. These systems collect information from the users that use these applications and then use the aforementioned data to extract meaning patterns and find solutions to a larger, more predominant, problem. There are namely three steps to follow in order to build a successful system for analysing stroke risk data. They are as follows :

- Data collection
- Data Pre-Processing
- Selecting a suitable algorithm

In the first stage, sufficient data is collected from various sources. The sources may include, but may not be limited to, personal data collected by frequent monitoring, data collected from a more credible source like hospitals or even data obtained from government approved agencies or websites that contain validated survey information.

The second stage, however, is the step that determines the effectiveness of the system being implemented. In involves removing all kinds of unnecessary information and only retain- ing the data that is required for the use case. It may include, but may not be limited to, removing outliers, handing imbalanced data, removing records with insufficient data and scaling data. In the final stage, the system is trained using advanced computing technologies such as a Machine Learning Model. Selecting an appropriate model that works best with the data in hand, providing the highest possible accuracy and precision, is important to build a system that has a low error rate.

- *SRM Institute of Science and Technology.*
- *Department of Computer Science and Engineering.*
- *thegeorgejoseph@gmail.com*
- *bidus21.bd@gmail.com*

A. Work Flow of a Typical Risk Analysis System

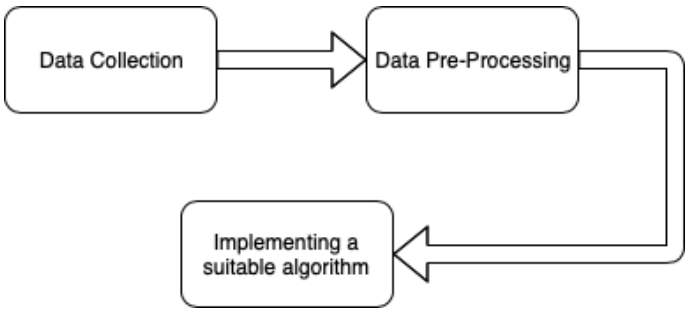

***Fig. 1**. Architecture Diagram*

2. DATA COLLECTION AND DATA PRE-PROCESSING

Collection of data that is required for the system is important when it comes to risk analysis. The sources these data come from are very important to determine the its quality and validity. Keeping the inclusion criteria of the quality of the data collected strict can result in insufficient data. A smarter approach to build the system would be to collect all kinds of related data and perform Pre-Processing tasks like data reduction, data imputation, replacing missing values, removing outliers, data transformation and cleaning.

***Fig. 2**. TABLE I: Data Collection and Pre-Processing*

S. NO	STUDY OF	DATA	METHOD OF PREPROCESSING USED
1	Jae-woo Lee et al [1]	6,885,789 participants	Replacing missing values
2	Sun Ha Jee et al [2]	1,329,525 Koreans	Removing outliers
3	Miguel Monteiro et al [3]	541 patients	Data cleaning
4	Yonglai Zhang et al [4]	792 records	Data transformation into convenient structures
5	Yannan Yu et al [5]	Personal data,multiple scores	Data cleaning
6	Georgios Tsivgoulis et al [6]	2343 individuals	Remove noise and inconsistent data
7	Chen-Ying Hung et al [7]	EMC database	cleaning
8	X Zhang et al [8]	Caucasian population Individual participant data	Removing inconsistent data
9	King Chung Ho et al [9]	Personal dataset	Data reduction
10	Ahmet K. Arslan et al [10]	80 patients records	Cluster based outlier factor
11	Benjamin Letham et al [11]	Patients in MDCD database	Cleaning and reduction
12	Lisa Nobel et al [12]	Personal data	Data cleaning
13	Aditya Khosla et al [13]	Cardiovascular health stud(CHS) dataset	Data imputation,replace missing values
14	Cemil Colak et al [14]	297 Patient records	Removing outliers,extreme values, noise and inconsistent data
15	Xiao-Fei Zhang et al [15]	4400 male records	Data cleaning
16	Kristi Reynolds et al [16]	Medline records	Replace missing values, remove outliers

3.EVALUATING THE ALGORITHM USED

The data for the analysis of stroke risk may be of any form. This may include, but may not be limited to, numeric data, categorical data and/or image data. Therefore, finding an algorithm that maximises the accuracy is a challenging task. Commonly used algorithms according to research are SVM, ANN, DNN, Logistic Regression, Cox's Regression Model, Random Forest, Stochastic Gradient Boosting, MLP etc. Jae-woo Lee et al[1] developed a system that classifies10 year probability of stroke based on certain risk factors such as age, record of hypertension, reported heart disease, physical activity, total cholesterol, smoking habits,presence of diabetes,BMI.The probability of stroke was determined using the cox model by studying the information of the user and uploading a lifestyle correction message to reduce stroke risk. However, there was a limitation in accurately identifying the stroke type that was caused due to the factors, the model produced an AUC of 80percent. Sun Ha Jee et al [2] proposed a system based on Framingham heart study for a stroke risk prediction model. This model was performed on patients with risk factors of blood pressure, cholesterol, fasting blood sugar and a 13-year research was performed to predict probability using the cox model, it performed estimation from models with actual stroke cases. Limitations were that there was lack of relevant risk factors, hospital diagnosis could not be verified. Advantage was the sample size. The area under ROC curve was 82percent for men and 81percent for women. Miguel Monteiro et al [3] developed a system that makes use of machine learning techniques to predict the individuals recovery three months after the initial stroke. Classifiers are trained and the results of various machine learning methods were compared. Different experiments have been conducted on L1 regularised LR, decision trees, SVM, Random forest, XGBoost and the results are analysed. AUC has been used to measure performance. It has been inferred that training the classifiers with more data results in increase in average AUC. The resulting AUC can range up to 0.936. Yonglai Zhang et al [4] in their system used feature selection and classification method to find the risk of stroke. Feature selection determines distinguishing characters from existing features, this includes patients basic clinical and physiological characteristics and symptoms that can identify potential risk. Uses SVM,ANN and glow-worm swarm optimisation techniques. The classification accuracy obtained was 82.58percent and AUC is 0.8948. Yannan Yu et al [5] developed a system that collects patients details for 24 hours to collect the factors that helps source perfusion MRIs predict the severity of hemorrhagic transformation in stroke. This detects and predicts the severity in hemorrhagic transformation.Bayesian modelling and Gaussian processes have been employed to review perfusion. Machine Learning algorithms such as SVM,Multiple Linear Regression are used in the prediction model. Models reached an accuracy of above 80percent on this dataset. Georgios Tsivgoulis et al [6] proposed a system which monitors a patients blood pressure and tells the risk profile based on the Framingham stroke data. Considers clinical evidence of hypertension complications, previous stroke history, status of dipping and recordings for a periods of four years. The results provide evidence regarding the predictive utility of ABPM in terms of individualised stroke risk quantification. Chen-Ying Hung et al [7] in their system compares deep neural nets and other ML algorithms for predicting stroke in a large amount of data. Utilises algorithms such as deep neural network(DNN),GBDT,LR, SVM on the data to achieve an effective prediction model. DNN and GBDT achieve AUC 0.915 and 0.918,LR and SVM have less effective performance. X Zhang et al [8]in their system describes the relation between cholesterol, stroke and coronary disease. Factors in the likelihood of age, sex, BP, BMI, record of smoking were considered and log linear or linear regression model has been implemented to find relation among cholesterol levels and stroke risk and other heart diseases. The data suggests a relation that is neither negative nor null between cholesterol and ischemic stroke. King Chung Ho et al [9] proposed a system which employs machine learning techniques for classifying strokes that are ischemic by image studying with onset stroke time that could provide information in deciding treatment for stroke patients.Algorithms such as LR, random forest,SVM, stepwise multilinear regression(SMR) are constructed and the performance is compared. This proposed model provides information to clinicians to formulate an intervention treatment for clinicians of acute stroke. Ahmet K. Arslan et al [10] in their system uses a lot of mining techniques to predict strokes that are ischemic. SVMs, SGBs and PLRs have been employed in the prediction model, this study makes computer aided medical approaches effective in prediction of ischemic stroke and explores the hidden features of the dataset.SVM has the highest performance in prediction according to various metrics. Benjamin Letham et al [11] developed a highly accurate predictive model which is interpretable. Bayesian rules that are listed have been employed to build the prediction model, factors such as age, cerebrovascular disorder, altered state of consciousness are considered. Different experiments have been conducted for female only and male only to obtain BRL point estimates. In the predictive medicinal domain, interpretable models are very useful. Lisa Nobel et al [12] in their system proposed a personalised tool that can be used by health professionals and people alike to assess stroke risk. It also assesses lifestyle indicators like such as obesity, unhealthy diet, physical inactivity, alcohol consumption, elevated blood glucose level. Uses Cox's model to predict stroke risk of the individual. A very easy method to spread information regarding risk and prevention. Aditya Khosla et al [13] developed a system which compares Cox's model with other ML techniques to predict stroke. The feature selection method is based conservative- mean and SVM. A larger AUC is obtained in this model as compared to Cox's. This combination had a 15percent lesser error rate. Cemil Colak et al [14] proposed a system to predict stroke outcome utilising KDP, ANN and SVM. 81.3percent specificity was obtained by the MLP trained model that used a quick propagation algorithm. It also produced 78.4percent sensitivity, 80.7percent model accuracy, and 0.869 AUC. This proposed neural net model may be helpful in making clinical decisions regarding stroke with AUC value of 0.905 in training set. The testing set gave a value of 0.928. Xiao-Fei Zhang et al [15] developed a system that predicts a score for the risk in heart disease among the Chinese population. Varieties of stroke like hemorrhagic and ischemic were scored separately. Risk estimates are based on variables such as age, smoking habits,BP and cholesterol. Multivariate cox's model is used. The average age for occurrence of

stroke was found to be 45 years. AUC was 0.76 for heart disease,0.82 for hemorrhagic and 0.72 for ischemic stroke respectfully. Kristi Reynolds et al [16] proposed that alcohol consumption and stroke risk could be related. Based on the different level of alcohol consumption relative risk of stroke was measured using random-effects model and meta regression analysis. A possible error in the data could be due to the alcohol consumption data that was self-assessed by the individuals as heavy alcohol consumers can falsify their consumption record. This model suggests that reducing the alcohol intake can reduce the chances of stroke.

Fig. 3. *TABLE II: Algorithm Evaluation*

S. NO	STUDY BY	ALGORITHM USED	PERFORMANCE METRICS
1	Jae-woo Lee,Hyun-sun Lim,Dong-wook Kim,Soon-ae Shin,Jinkwon Kim,Bora Yoo,Kyung-hee Cho	Cox's Proportional Hazard Regression Model	B coefficient,P value,Hazard ratio
2	Sun Ha Jee,Ji Wan Park,Sang-Yi Lee,Byung-Ho Nam,Hwang Gun Ryu,Su Young Kim,Youn Nam Kim,Ja Kyoung Lee,Sun Mi Choi,Ji Eun Yun	Cox Model	Coefficient,Hazard ratio,95%CI
3	Miguel Monteiro,Ana Catarina Fonseca,Ana Teresa Freitas,Teresa Pinho e Melo,Alexandre P. Francisco,Jose M. Ferro,Arlindo L, Oliveira	L1 Regularized Logistic Regression,decision tree, SVM,random forest,xgboost	AUC
4	Yonglai Zhang,Wenai Song,Shuai Li,Lizhen Fu, and Shixin Li	SVM, glow-worm swarm optimization algorithm,ANN	AUC,Accuracy
5	Yannan Yu,Danfeng Guo,Min Lou,David Liebesking,Fabien Scalzo	SVM,Linear regression,decision tree, neural network,kernal spectral regression	AUC-ROC,AUC-PR, DICE's coefficient
6	Georgios Tsivgoulis,Maria Pikilidou,Aristeidis H.Katsanos,Kimon Stamatelopoulos,Fotios Michas,Aikaterini Lykka,Cristina Zompola,Angeliki Filippatou,Efstthios Boviatsis,Konstantinos Voumvorakis,Nikolaos Zakopoulos,Efstathios Manios	Multiple linear regression analysis	Pearson's correlation coefficient
7	Chen-Ying Hung,Wei-Chen Chen,Po-Tsun Lai,Ching-Heng Lin, Chi-Chun Lee	Deep neural network(DNN),GBDT,LR, SVM	Sensitivity,specificity, accuracy,UAR
8	X Zhang,A Patel,H Horibe,Z Wu,F Barzi,A Rodgers,S MacMahon,M Woodward	Log linear or linear regression	Mean,sd,p-value
9	King Chung Ho,William Speier,Haoyue Zhang,Fabien Scalzo,Suzie El-Saden,Corey W. Arnold	Logistic regression,random forest,SVM,stepwise multilinear regression(SMR)	Relative minimun,relative max,variance,relative variance
8	X Zhang,A Patel,H Horibe,Z Wu,F Barzi,A Rodgers,S MacMahon,M Woodward	Log linear or linear regression	Mean,sd,p-value
9	King Chung Ho,William Speier,Haoyue Zhang,Fabien Scalzo,Suzie El-Saden,Corey W. Arnold	Logistic regression,random forest,SVM,stepwise multilinear regression(SMR)	Relative minimun,relative max,variance,relative variance
10	Ahmet K. Arslan Cemil Colak Ediz Sarihan	SVM,stochastic gradient boosting(SGB),penalized logistic regression(PLR)	AUC,sensitivity,positive and negative predictive value,specificity
11	Benjamin Letham, Cynthia Rudin,Tyler H. McCormick,David Madigan	Bayesian Rule Lists(BRL), Apriori	Mean,accuracy, standard deviation,AUC ROC
12	Lisa Nobel,Nancy E. Mayo,James Hanley, Lyne Nadeau,Stella S. Daskalopoulou	Cox proportional hazards regression model	Prevalence%, HR, 95% confidence interval,B-estimate
13	Aditya Khosla,Yu Cao,Cliff Chiung-Yu Lin, Hsu-Kuang Chiu,Junling Hu,Honglak Lee	SVM,margin based censored regression(MCR), cox regression, regularized logistic regression	Concordance index, Area under ROC curve
14	Cemil Colak, Ersa Karaman , M. Gokhan Turtay	KDP,ANN,SVM and MLP	Accuracy and AUC
15	Xiao-Fei Zhang,John Attia,Catherine D'Este,Xue-Hai Yu,Xi-Gui Wu	Cox model, multivariate cox model	AUC,ROC, Regression coefficient
16	Kristi Reynolds,L. Brian Lewis, John David L. Nolen, Gregory L. Kinney, Bhavani Sathya, Jiang He	Random-effects model and meta regression analysis	G/d,p-value

4. CONCLUSION

With sufficient study on the various techniques used to assess the risk of stroke, important features that affect the probability of risk of stroke are, but not limited to, alcohol consumption, smoking status, hypertension, heart disease, blood sugar level, sex, age etc. Algorithms like Cox's regression is widely used to analyse risk, and is highly accurate as well. Other algorithms that provide high accuracy in classification include SVM, SGB and ANNs. It is also inferred that performance metrics like AUC can be marginally improved by implementing the aforementioned algorithms on large data.

REFERENCES

[1] Lee, J. W., Lim, H. S., Kim, D. W., Shin, S. A., Kim, J., Yoo, B., Cho, K. H. (2018). The development and implementation of stroke risk prediction model in National Health Insurance Service's personal health record. Computer methods and programs in biomedicine, 153, 253-257.

[2] Jee, S. H., Park, J. W., Lee, S. Y., Nam, B. H., Ryu, H. G., Kim, S. Y.,

[3] ... Yun, J. E. (2008). Stroke risk prediction model: a risk profile from the Korean study. Atherosclerosis, 197(1), 318-325.

[4] Monteiro, M., Fonseca, A. C., Freitas, A. T., Pinho e Melo, T., Francisco,

a. P., Ferro, J. M., Oliveira, A. L. (2018). Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 15(6), 1953-1959.

[5] Zhang, Y., Song, W., Li, S., Fu, L., Li, S. (2018). Risk detection of stroke using a feature selection and classification method. IEEE Access, 6, 31899-31907.

[6] Yu, Y., Guo, D., Lou, M., Liebeskind, D., Scalzo, F. (2017). Prediction of hemorrhagic transformation severity in acute stroke from source perfusion MRI. IEEE Transactions on Biomedical Engineering, 65(9), 2058-2065.

[7] Sivgoulis, G., Pikilidou, M., Katsanos, A. H., Stamatelopoulos, K., Michas, F., Lykka, A., ... Zakopoulos, N. (2017). Association of ambulatory blood pressure monitoring parameters with the Framingham Stroke Risk Profile. Journal of the neurological sciences, 380, 106-111.

[8] Hung, C. Y., Chen, W. C., Lai, P. T., Lin, C. H., Lee, C. C. (2017, July). Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 3110-3113). IEEE.

[9] Asia Pacific Cohort Studies Collaboration. (2003). Cholesterol, coronary heart disease, and stroke in the Asia Pacific region. International journal of epidemiology, 32(4), 563-572.

[10] Ho, K. C., Speier, W., Zhang, H., Scalzo, F., El-Saden, S., Arnold,

[11] C. W. (2019). A Machine Learning Approach for Classi-

fying Ischemic Stroke Onset Time from Imaging. IEEE transactions on medical imaging.

[12] Arslan, A. K., Colak, C., Sarihan, M. E. (2016). Different medical data mining approaches based prediction of ischemic stroke. Computer methods and programs in biomedicine, 130, 87-92.

[13] Letham, B., Rudin, C., McCormick, T. H., Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics, 9(3), 1350-1371.

[14] Nobel, L., Mayo, N. E., Hanley, J., Nadeau, L., Daskalopoulou, S. S. (2014). MyRisk Stroke Calculator: a personalized stroke risk assessment tool for the general population. Journal of Clinical Neurology, 10(1), 1-9.

[15] Khosla, A., Cao, Y., Lin, C. C. Y., Chiu, H. K., Hu, J., Lee, H. (2010, July). 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). ACM.

[16] Colak, C., Karaman, E., Turtay, M. G. (2015). Application of knowledge discovery process on the prediction of stroke. Computer methods and programs in biomedicine, 119(3), 181-185.

[17] hang, X. F., Attia, J., D'Este, C., Yu, X. H., Wu, X. G. (2005). A risk score predicted coronary heart disease and stroke in a Chinese cohort. Journal of clinical epidemiology, 58(9), 951-958.

[18] Reynolds, K., Lewis, B., Nolen, J. D. L., Kinney, G. L., Sathya, B., He, J. (2003). Alcohol consumption and risk of stroke: a meta-analysis. Jama, 289(5), 579-588.