Original Paper

Predicting Risk of Stroke From Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models

Eman M Alanazi[1,2], MSc; Aalaa Abdou[3], MD; Jake Luo[4], PhD

[1]Department of Health Informatics, College of Health Sciences, Saudi Electronic University, Riyadh, Saudi Arabia

[2]Department of Biomedical and Health Informatics, College of Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI, United States

[3]Department of Radiotherapy, Children's Cancer Hospital 57357, Cairo, Egypt

[4]Department of Health Informatics and Administration, College of Health Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI, United States

Corresponding Author:
Eman M Alanazi, MSc
Department of Health Informatics
College of Health Sciences
Saudi Electronic University
Abi Bakr As Siddiq Branch Rd
Riyadh, 13323
Saudi Arabia
Phone: 966 112613500
Email: e.alanazi@seu.edu.sa

Abstract

Background: Stroke, a cerebrovascular disease, is one of the major causes of death. It causes significant health and financial burdens for both patients and health care systems. One of the important risk factors for stroke is health-related behavior, which is becoming an increasingly important focus of prevention. Many machine learning models have been built to predict the risk of stroke or to automatically diagnose stroke, using predictors such as lifestyle factors or radiological imaging. However, there have been no models built using data from lab tests.

Objective: The aim of this study was to apply computational methods using machine learning techniques to predict stroke from lab test data.

Methods: We used the National Health and Nutrition Examination Survey data sets with three different data selection methods (ie, without data resampling, with data imputation, and with data resampling) to develop predictive models. We used four machine learning classifiers and six performance measures to evaluate the performance of the models.

Results: We found that accurate and sensitive machine learning models can be created to predict stroke from lab test data. Our results show that the data resampling approach performed the best compared to the other two data selection techniques. Prediction with the random forest algorithm, which was the best algorithm tested, achieved an accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and area under the curve of 0.96, 0.97, 0.96, 0.75, 0.99, and 0.97, respectively, when all of the attributes were used.

Conclusions: The predictive model, built using data from lab tests, was easy to use and had high accuracy. In future studies, we aim to use data that reflect different types of stroke and to explore the data to build a prediction model for each type.

(JMIR Form Res 2021;5(12):e23440) doi: 10.2196/23440

KEYWORDS

stroke; lab tests; machine learning technology; predictive analytics

Introduction

Stroke is a neurological deficit, primarily because of acute central nervous system focal injury caused by a vascular issue. It is a major cause of disability and death worldwide [1]. It estimated that the overall prevalence of stroke in the United States is 2.5%, and about 7 million Americans over the age of 20 years have experienced a stroke. The condition has a significant negative impact on patients' health and quality of life. It also has a negative impact on hospital services and the availability of beds and was estimated to cost the US economy about US $351.2 billion between 2014 and 2015 [2]. There are

two types of stroke: ischemic and hemorrhagic. Hemorrhagic stroke occurs because of a burst vessel that leads to bleeding in the brain, whereas ischemic stroke occurs because of a blockage of the arteries of the brain. Ischemic strokes are the most common, comprising 85% to 90% of all strokes [3]. This condition can be prevented by promoting health and increasing awareness of risk factors. There are many risk factors related to lifestyle, including obesity, diet, alcohol intake, and lack of physical activity [4]. Underlying conditions, such as diabetes, hypertension, and cardiovascular diseases, may also lead to stroke. Therefore, proper self-management of these diseases and the pursuit of a healthy lifestyle may prevent the occurrence of stroke.

In 2019, the American College of Cardiology/American Heart Association released the Guideline on the Primary Prevention of Cardiovascular Disease. The guideline recommends a complete assessment and examination of patients who are at risk of developing blockages in their arteries that may lead to a heart attack or stroke and might die as a result [5]. Now more than ever, physicians can access clinical evidence to identify high-risk patients using approaches such as acquiring a complete patient history and conducting thorough physical exams for risk assessment. Patient records contain many useful predictive factors, such as patient demographic (eg, age and gender), lifestyle (eg, diet and physical activity), and existing medical condition factors (eg, diabetes and hypertension), that might lead to stroke [5]. The growth of arterial blockages and decades of damage to blood vessels, which may lead to stroke, are often associated with these risk factors. If physicians can assess the risks of stroke easily and conveniently, strokes could be prevented at an earlier stage. This approach could save lives and reduce the economic burden of health care services. In the age of artificial intelligence and machine learning, a clinical decision support system has been developed to assist physicians to diagnose and identify individuals with a high risk of stroke. The potential of applying machine learning technologies in the cardiovascular domain is significant, from identifying individuals with a high risk of stroke [6,7] to predicting outcomes of patients following treatment [8,9]. Most of these studies use either health habits and lifestyle factors, such as smoking or alcohol consumption; conditions that predispose to strokes, such as hypertension and diabetes mellites; or neuroimaging, such as computed tomography and magnetic resonance imaging, to either classify or predict the disease.

Besides assessing known risk factors for stroke, scientists are trying to develop lab tests that can predict stroke. One of the major advantages of using lab test results for prediction is that lab tests are commonly collected in clinical settings, and the information is often well documented in patients' records. In this study, we explored data-driven approaches using supervised machine learning models to predict the risk of stroke from different lab tests.

Several studies have been able to identify independent laboratory tests that are correlated with stroke using descriptive statistical analysis. Sughrue et al [10] conducted a retrospective study in 2013 that identified 35 tests with a statistically significant correlation with a future stroke diagnosis. The most informative were for various types of cholesterol. Two of these 35 laboratory tests were urine tests, and 33 were blood, serum, or plasma tests. Some tests were positively associated with an outcome of stroke (ie, neutrophil count and percent; CD3+, CD8+, and T8 suppressor cells; monocytes; eosinophils; and CD3 cells), and others were negatively correlated (ie, hematocrit and lymphocytes). Their results show that it is possible to correlate future stroke with collected lab test data. Farah and Samra [11] conducted a retrospective study investigating the association between the neutrophil-to-lymphocyte ratio (NLR), mean platelet volume (MPV), and the risk of stroke. Two-tailed *t* tests showed no significant differences in the stroke group's MPV values compared with those in the control group. However, the NLRs of the stroke patients were significantly different compared with those of the control group. That study indicated the existence of a correlation between the level of NLR and stroke risk. NLR levels have been shown to be higher in stroke patients than in control groups. Feng et al [12] reviewed the scientific literature on the potential role and the possible epidemiological relationships between red cell distribution width (RDW) and ischemic stroke in a meta-analysis of 40 manuscripts from China National Knowledge Infrastructure and PubMed databases. They reported that patients with stroke had higher levels of RDW than those without strokes. Another study by Kaya et al [13] also investigated the association between baseline RDW level and stroke risk in patients with heart failure. These authors found that heart failure patients suffering from stroke had significantly increased basal RDW levels (mean 16.9, SD 1.14, vs mean 14.8, SD 1.6; $P<.001$) and serum uric acid levels (mean 8.8, SD 1.7, vs mean 7.5, SD 1.1; $P=.027$) compared with patients without stroke, according to the propensity score analysis. Giles et al [14] used data from a national cohort to investigate whether low folate levels were associated with ischemic stroke and found that folate concentrations of ≤9.2 nmol/L could be a risk factor for ischemic stroke (relative risk 1.37, 95% CI 0.82-2.29). Another study by Qin et al [15] concluded that there is a significant risk of first ischemic stroke in hypertensive patients with low levels of folate and vitamin B12.

These studies demonstrate the value of lab test results for predicting stroke. Our study aimed to leverage lab test results to build machine learning models for stroke prediction. We prepared the data sets using three data selection techniques for this study. After that, for each data selection technique, we applied four individual machine learning classifiers to prepare prediction models. We measured the performance of each prediction model using six different performance measures. Our results indicate that the data resampling technique outperformed the decision tree and random forest classifiers.

Methods

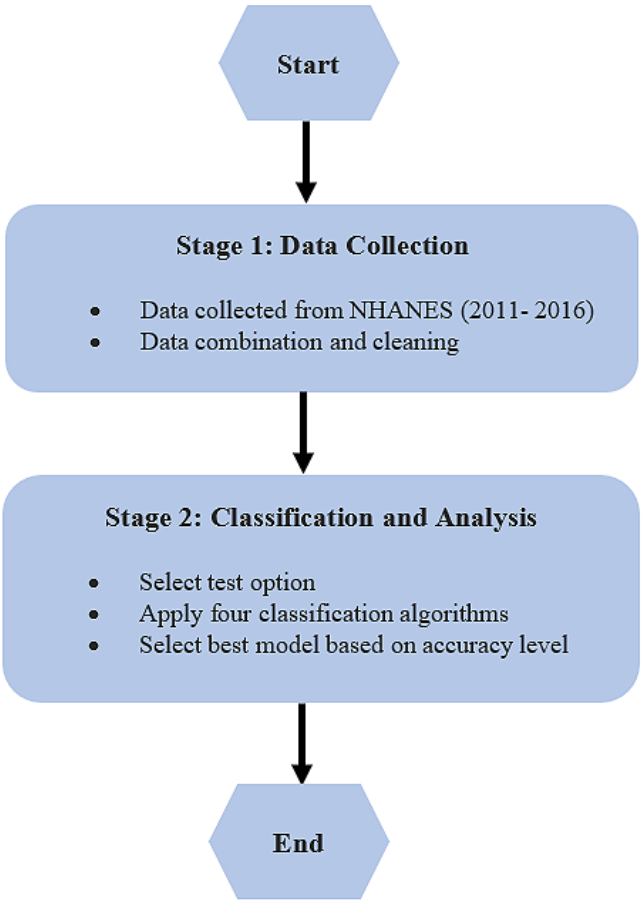
Overview

Figure 1 shows the outline of our investigation. In the first step, we collected data from the National Health and Nutrition Examination Survey (NHANES). In the second step, we selected the data using three data techniques for our prediction models. The first one was conducted without data resampling, the second

one included data imputation, and the third one was conducted with data resampling.

We used 10-fold cross-validation to perform the train and test approach. To train models, we used four different machine learning classifiers, and six performance measures were used to assess the performance of the models. The elaborated descriptions of the data sets, classifiers, and performance metrics that were used are given below.

Figure 1. Flow diagram of the study methodology. NHANES: National Health and Nutrition Examination Survey.



Data Collection

The NHANES survey was conducted to examine the health and nutritional status of adults and children in the United States; "NHANES is a major program of the National Center for Health Statistics (NCHS). NCHS is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation" [16]. The data sets contain five domains: demographics, dietary data, examination data, laboratory data, and questioner data. Each domain contains several subdomains. Our focus was on data sets that contain information about laboratory tests. The data sets are available from 1999 to 2017, and we considered data from 2011 to 2015. The total number of participants was 15,714 during this period. To reduce the impact of imbalanced data, we noted that in the entire data set, there were about 17% of participants who had experienced a stroke. Therefore, we included total of 4186 participants, of whom 608 (14.5%) had experienced a stroke (Figure 2). The list of data attributes is shown in Table 1. The data sets contained 21 attributes, including each patient's age and gender as well as other lab test information for each respective patient. The data sets and their information are available online [16], where the data are presented from the year 2000 to the current year. For this study, the data were collected for each year and combined using the sequence number (SEQN). After combining and cleaning the data, we used the Waikato Environment for Knowledge Analysis (WEKA; version 3.8.0) system to build and test machine learning models.

Figure 2. Participant selection and prevalence of stroke in the National Health and Nutrition Examination Survey (NHANES).

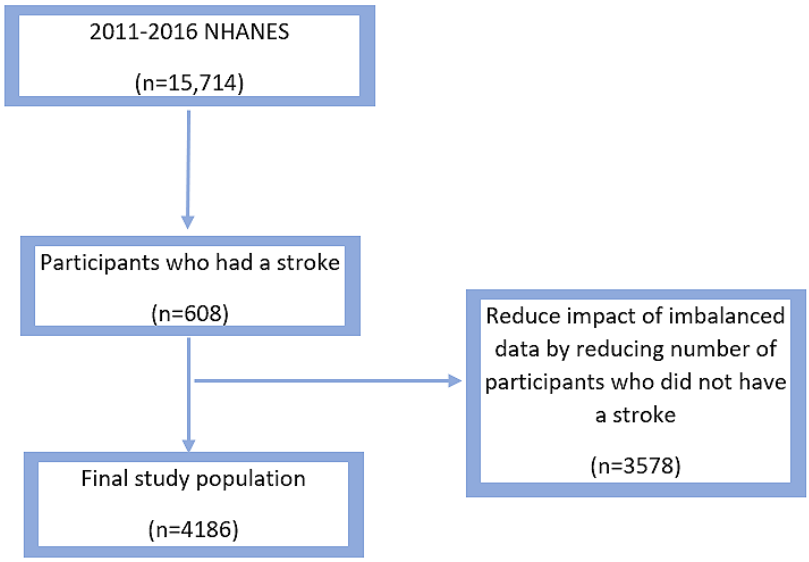


Table 1. List of the data attributes.

Feature[a]	Units
Age	Years
Gender	N/A[b]
Albumin, urine	ug/mL
Creatinine, urine	mg/dL
White blood cell count	1000 cells/µL
Lymphocytes	1000 cells/µL
Monocytes	1000 cells/µL
Segmented neutrophils	1000 cells/µL
Eosinophils	1000 cells/µL
Basophils	1000 cells/µL
Red blood cell count	Million cells/µL
Hemoglobin	g/dL
Hematocrit	%
Mean cell volume	fL
Mean cell hemoglobin	pg
Mean corpuscular hemoglobin concentration	g/dL
Red cell distribution width	%
Platelet count	1000 cells/µL
Mean platelet volume	fL
Cotinine, serum	ng/mL
Red blood cell folate	mg/dL

[a]All data types were numeric, except for “gender,” which was nominal.

[b]N/A: not applicable; this type of data did not have units.

Classification

Several different machine learning algorithms can handle a binary classification problem. In this study, we used four machine learning algorithms: naïve Bayes, BayesNet, J48 (Java implementation of C4.5 algorithm), and random forest. The performance of the algorithms was evaluated and compared for stroke prediction using lab test results as features. Details of the algorithms are as follows:

- The J48 algorithm creates a tree based on the C4.5 algorithm with pruning.
- The random forest algorithm creates a forest of random trees and outputs the mode of the classes created by individual trees.
- The naïve Bayes algorithm creates a classifier based on the naïve Bayes method, which assumes that all attributes are independent.
- The BayesNet algorithm creates a classifier based on non–naïve Bayes, which does not assume that all attributes are independent.

In the cross-validation approach, the data sets are divided into several equal portions; in general, 5-fold and 10-fold cross-validations are used when the data sets are equally divided into 5 and 10 portions [17]. With this approach, for each simulation, one portion of each data set is used to train the prediction model and the rest are used for validation. In this study, we used 10-fold cross-validation and, in this process, we divided the whole of each data set into 10 equal parts; each time, 10% of each data set was used to train the model and 90% was used for validation. In this task, three data analyses were conducted where the first data analysis applied each of the machine learning techniques on the data sets without data manipulation or resampling. The aim was to determine the baseline for the data sets among the various machine learning techniques. The imputation of missing data set entries was conducted in the second analysis. In statistics, imputation entails substituting missing data with values calculated using any of a number of techniques [18]. Imputation is a useful technique in remedying missing data, since missing data may lead to inaccurate predictions. We used the default ReplaceMissingValue filter in WEKA, which replaces all missing values for nominal and numeric attributes in a data set with the modes and means from the training data. Most of the features had 5% missing values, and one feature had 11% missing values. After the imputation of the missing data, data resampling was conducted in the third analysis. Data resampling is a commonly used technique, since training may result in nonuniformity of class labels. In this case, the resampling technique was applied to select a specific subset of data points for model training [19]. After resampling the data, the results of the first analysis should be improved because of the balancing of the data set distribution. A balanced distribution was achieved through the use of WEKA, which randomly resamples the data. Based on the available theoretical knowledge about resampling and imputation in statistics, the results after the third analysis should be improved.

Evaluation Metrics

Model accuracy was evaluated based on the following measures: recall or sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and area under the curve (AUC) (or area under the receiver operating characteristic [ROC] curve) to compare the four classifiers. Details of these measures are as follows:

- Sensitivity, also known as recall or true positive rate, is the number of true positives divided by the number of true positives plus the number of false negatives. It is the likelihood that the patient has a high risk of stroke [20].
- Specificity, also known as the true negative rate, is the proportion of individuals classified as nonstroke to the total number of actual nonstroke cases. It is the likelihood that a patient who does not have a risk of stroke will have a negative result [21].
- PPV, also known as precision, is the number of true positives divided by the number of true positives plus the number of false positives. It is the proportion of individuals who have suffered a stroke to the total number of participants classified as having a risk of stroke [22].
- NPV is the percentage of negative tests in patients who are free from the disease or the proportion of individuals who have not suffered a stroke to the total number of participants classified as not having a risk of stroke [22].
- Overall accuracy is the number of correctly classified instances over the total size of the data set [20].
- The AUC is the area under the ROC curve, which is constructed by plotting the true positive rate against the true negative rate [23].

We will also look at the Pearson correlation coefficient value of each independent predictor to investigate the relationship between each lab test and risk of stroke.

Results

In the NHANES data sets, 608 participants suffered from a stroke from 2011 to 2015. The median age of participants who had a stroke was 51 years for both men and women. The numbers of men and women who had a stroke were 220 (36.2%) and 190 (31.3%), respectively; 198 (32.6%) participants did not reveal their gender identity.

After the data collection process, the data were analyzed in three ways: without data resampling, with data imputation, and with data resampling. Data resampling techniques were used to tackle data imbalance problems in the data sets. These sampling techniques are widely used in machine learning–based prediction models in different areas [24]. Our first analysis was done without the data resampling technique, where the four machine learning algorithms were applied directly to the data sets. The first analysis produced poor results for all four classifiers. The best sensitivity rate among the classifiers in the first analysis was for the BayesNet model, followed by the naïve Bayes model. In the second analysis, we applied the data imputation technique to the data sets, which replaced missing values and deleted features that had more than 50% missing values; the prediction accuracy improved for all models, except for the

naïve Bayes model, whose performance decreased slightly after replacing the missing values.

In the third analysis, we resampled the data. After resampling, the prediction accuracy improved significantly for both the decision tree and random forest models, but only slightly for the naïve Bayes and BayesNet models. Table 2 shows the scores of accuracy, sensitivity, specificity, PPV, NPV, and AUC, according to the three data analysis techniques and four classifiers. The table shows that the random forest model was the best classifier with the data resampling technique. Figures 3 and 4 show the score comparisons among the three data selection techniques for the decision tree and random forest models, respectively. We considered the decision tree and random forest classifiers to compare the performance, as they significantly improved the performance in the third analysis. Both figures clearly show that the third analysis, the data resampling technique, outperformed the other two techniques for the decision tree and random forest classifiers.

Table 2. Results of three data analysis techniques.

Technique and classifier	Accuracy	Sensitivity	Specificity	PPV[a]	NPV[b]	AUC[c]
Without data resampling						
Naïve Bayes	0.82	0.34	0.88	0.27	0.91	0.76
BayesNet	0.82	0.38	0.89	0.37	0.90	0.88
Decision tree	0.83	0.33	0.87	0.14	0.95	0.73
Random forest	0.86	0.55	0.86	0.01	0.99	0.87
Data imputation						
Naïve Bayes	0.81	0.32	0.88	0.25	0.91	0.74
BayesNet	0.86	0.53	0.92	0.54	0.92	0.85
Decision tree	0.88	0.61	0.91	0.46	0.95	0.74
Random forest	0.90	0.89	0.90	0.33	0.99	0.85
Data resampling						
Naïve Bayes	0.82	0.33	0.88	0.29	0.90	0.74
BayesNet	0.87	0.53	0.93	0.57	0.92	0.85
Decision tree	0.93	0.76	0.95	0.72	0.96	0.86
Random forest	0.96	0.97	0.96	0.75	0.99	0.97

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

[c]AUC: area under the curve.

Figure 3. Performance comparison among three data selection techniques for the decision tree model. AUC: area under the curve; NPV: negative predictive value; PPV: positive predictive value.

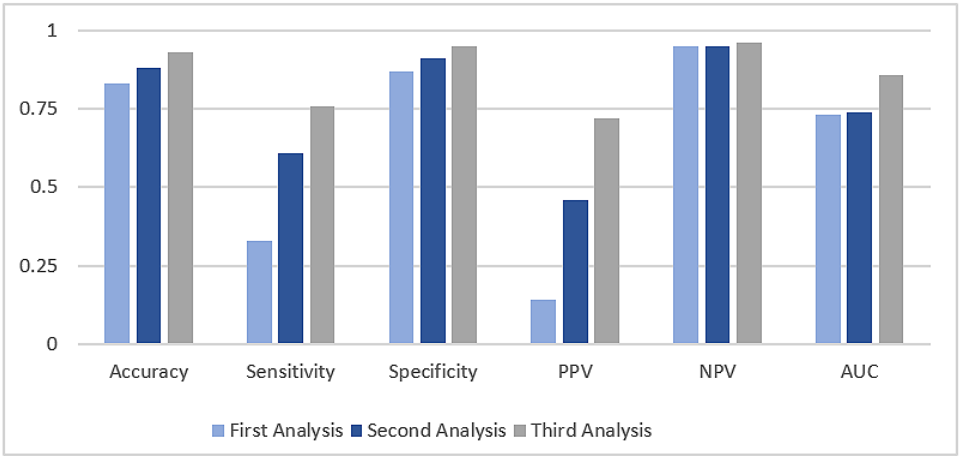


Figure 4. Performance comparison among three data selection techniques for the random forest model. AUC: area under the curve; NPV: negative predictive value; PPV: positive predictive value.

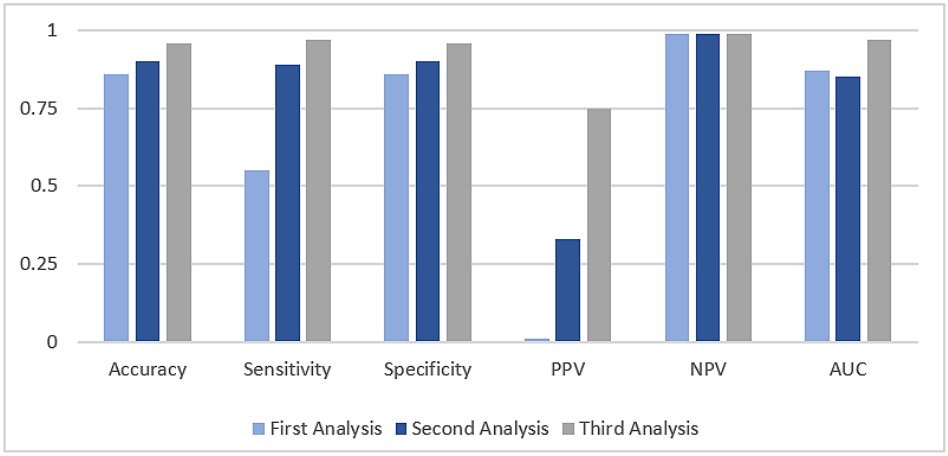


Table 3 shows the results from Pearson correlation analysis of the independent predictors.

Table 3. Pearson correlation coefficient values of independent predictors.

Independent predictor of stroke	Pearson correlation coefficient (*r*)
Age	0.26
Gender	0.13
Red cell distribution width (%)	0.18
Lymphocytes (%)	0.15
Red blood cell folate (ng/mL)	0.13
Segmented neutrophils (%)	0.12
Hemoglobin (g/dL)	0.11
Red blood cell count (million cells/μL)	0.11
Hematocrit (%)	0.09
Lymphocytes (1000 cells/μL)	0.08
Segmented neutrophils (1000 cell/μL)	0.07

Discussion

Principal Findings

From the previous section, we noticed that our models had the potential to perform stroke prediction using lab test data. Our results show that the random forest model was the best classifier after conducting the data resampling technique.

Also, several observations can be made from the results in Table 3. We identified nine lab tests, in addition to age and gender, that effectively correlated with stroke occurrence. These correlations were calculated using the Pearson correlation coefficient. These results align with other research that showed a linear relationship between some of these variables and stroke. Several studies have shown that age is correlated with the risk of stroke. According to Muntner et al [2], stroke incidence doubles after the age of 45 years, and 70% of all strokes occur over the age of 65 years. Many studies have investigated the relationship between baseline RDW and stroke. They found that elevated RDW is a risk factor in ischemic stroke [12,13,25].

One of the novel correlations that were found in this study is the lymphocyte percentage. Lymphocytes are white blood cells, including B cells, T cells, and natural killer cells. Lymphocyte percentage is positively associated with stroke occurrence. There have been no studies suggesting that lymphocyte percentage can be a predictor of stroke, but different studies have examined the use of immune cells as biomarkers to predict stroke outcome [26,27]. There is one study that showed a negative correlation between hematocrit and stroke occurrence [10]. Folate deficiency has various clinical manifestations. Our finding that serum folate level was correlated with the risk of stroke is in line with the finding of Giles et al [14], who found that a serum folate concentration of ≤9.2 nmol/L may slightly increase the risk for ischemic stroke. Other studies have shown that folic acid therapy involving folic acid, vitamin B12, and vitamin B6 reduced the risk of ischemic stroke [15,28]. Neutrophils, which are normally the most abundant circulating white blood cells and respond quickly to infection, also contribute to the main processes causing an ischemic stroke, as they facilitate the development of blood clots. Neutrophils are, therefore, also of

considerable importance as targets for treating and preventing ischemic stroke [29]. A study by Sughrue et al [10] produced results similar to ours regarding the positive association between neutrophils and stroke occurrence. Hemoglobin levels can predict the risk of stroke. Observational studies have reported an independent association between red blood cell count, hematocrit, and hemoglobin concentration and the risk of developing stroke [30,31].

The correlations between these different lab tests and stroke were found in several studies. However, this is the first study that used all of these different attributes to build a prediction model using machine learning algorithms. Our results showed that a prediction model can be created using the random forest algorithm and could achieve an accuracy of 0.96.

Conclusions

Machine learning applications are becoming more widely used in the health care sector. The prediction of stroke using machine learning algorithms has been studied extensively. However, no previous work has explored the prediction of stroke using lab tests. The results of several laboratory tests are correlated with stroke. Building a prediction model that can predict the risk of stroke from lab test data could save lives. In this study, we created a prediction model using the random forest algorithm and achieved a 96% accuracy rate. The model can be integrated with electronic health records to provide a real-time prediction of stroke from lab tests. Because of the nature of the data, we could not predict the type of stroke: hemorrhagic or ischemic. In future studies, we aim to use data that provide information about different types of stroke to build prediction models for each type.

Acknowledgments

EMA conducted the research design, data collection, and data analysis and wrote the original draft. AA assisted with the literature review of the lab tests. JL revised and edited the original draft and provided guidance throughout the whole research process. This study received no external funding.

Conflicts of Interest

None declared.

References

1. Sacco RL, Kasner SE, Broderick JP, Caplan LR, Connors JJB, Culebras A, American Heart Association Stroke Council, Council on Cardiovascular Surgery and Anesthesia, Council on Cardiovascular Radiology and Intervention, Council on Cardiovascular and Stroke Nursing, Council on Epidemiology and Prevention, Council on Peripheral Vascular Disease, Council on Nutrition, Physical Activity and Metabolism. An updated definition of stroke for the 21st century: A statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2013 Jul;44(7):2064-2089. [doi: 10.1161/STR.0b013e318296aeca] [Medline: 23652265]
2. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics-2019 Update: A report from the American Heart Association. Circulation 2019 Mar 05;139(10):e56-e528 [FREE Full text] [doi: 10.1161/CIR.0000000000000659] [Medline: 30700139]
3. European Stroke Initiative Executive Committee, EUSI Writing Committee, Olsen TS, Langhorne P, Diener HC, Hennerici M, et al. European Stroke Initiative Recommendations for Stroke Management – Update 2003. Cerebrovasc Dis 2003;16(4):311-337 [FREE Full text] [doi: 10.1159/000072554] [Medline: 14584488]
4. Boden-Albala B, Sacco RL. Lifestyle factors and stroke risk: Exercise, alcohol, diet, obesity, smoking, drug use, and stress. Curr Atheroscler Rep 2000 Mar;2(2):160-166. [doi: 10.1007/s11883-000-0111-3] [Medline: 11122740]
5. Arnett D, Blumenthal R, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 2019 Sep 10;74(10):e177-e232. [doi: 10.1016/j.jacc.2019.03.010] [Medline: 30894318]
6. Manuel DG, Tuna M, Perez R, Tanuseputro P, Hennessy D, Bennett C, et al. Predicting stroke risk based on health behaviours: Development of the Stroke Population Risk Tool (SPoRT). PLoS One 2015;10(12):e0143342 [FREE Full text] [doi: 10.1371/journal.pone.0143342] [Medline: 26637172]
7. Lee J, Lim H, Kim D, Shin S, Kim J, Yoo B, et al. The development and implementation of stroke risk prediction model in National Health Insurance Service's personal health record. Comput Methods Programs Biomed 2018 Jan;153:253-257 [FREE Full text] [doi: 10.1016/j.cmpb.2017.10.007] [Medline: 29157457]
8. Nielsen A, Hansen MB, Tietze A, Mouridsen K. Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning. Stroke 2018 Jun;49(6):1394-1401. [doi: 10.1161/strokeaha.117.019740]
9. Rondina JM, Filippone M, Girolami M, Ward NS. Decoding post-stroke motor function from structural brain imaging. Neuroimage Clin 2016;12:372-380 [FREE Full text] [doi: 10.1016/j.nicl.2016.07.014] [Medline: 27595065]
10. Sughrue T, Swiernik MA, Huang Y, Brody JP. Laboratory tests as short-term correlates of stroke. BMC Neurol 2016 Jul 21;16:112 [FREE Full text] [doi: 10.1186/s12883-016-0619-y] [Medline: 27439507]

11. Farah R, Samra N. Mean platelets volume and neutrophil to lymphocyte ratio as predictors of stroke. J Clin Lab Anal 2018 Jan;32(1):1-4 [FREE Full text] [doi: 10.1002/jcla.22189] [Medline: 28303662]
12. Feng G, Li H, Li Q, Fu Y, Huang R. Red blood cell distribution width and ischaemic stroke. Stroke Vasc Neurol 2017 Sep;2(3):172-175 [FREE Full text] [doi: 10.1136/svn-2017-000071] [Medline: 28989807]
13. Kaya A, Isik T, Kaya Y, Enginyurt O, Gunaydin ZY, Iscanli MD, et al. Relationship between red cell distribution width and stroke in patients with stable chronic heart failure: A propensity score matching analysis. Clin Appl Thromb Hemost 2015 Mar;21(2):160-165 [FREE Full text] [doi: 10.1177/1076029613493658] [Medline: 23804231]
14. Giles WH, Kittner SJ, Anda RF, Croft JB, Casper ML. Serum folate and risk for ischemic stroke. First National Health and Nutrition Examination Survey epidemiologic follow-up study. Stroke 1995 Jul;26(7):1166-1170. [doi: 10.1161/01.str.26.7.1166] [Medline: 7604408]
15. Qin X, Li J, Spence JD, Zhang Y, Li Y, Wang X, et al. Folic acid therapy reduces the first stroke risk associated with hypercholesterolemia among hypertensive patients. Stroke 2016 Nov;47(11):2805-2812. [doi: 10.1161/strokeaha.116.014578]
16. National Health and Nutrition Examination Survey. Centers for Disease Control and Prevention. URL: https://wwwn.cdc.gov/nchs/nhanes/Default.aspx [accessed 2020-04-10]
17. Sohan F, Rahman SSMM, Munna TA, Allayear SM, Rahman H, Rahman M. NStackSenti: Evaluation of a multi-level approach for detecting the sentiment of users. In: Proceedings of the 4th International Conference on Next Generation Computing Technologies. 2018 Presented at: 4th International Conference on Next Generation Computing Technologies; November 21-22, 2018; Dehradun, India p. 38-48. [doi: 10.1007/978-981-15-1718-1_4]
18. Marasinghe MG, Koehler KJ. Statistical Data Analysis Using SAS: Intermediate Statistical Methods. 2nd edition. Cham, Switzerland: Springer International Publishing AG; 2018.
19. Dugan TM, Mukhopadhyay S, Carroll A, Downs S. Machine learning techniques for prediction of early childhood obesity. Appl Clin Inform 2017 Dec 19;06(03):506-520. [doi: 10.4338/aci-2015-03-ra-0036]
20. Li X, Tao S, Jamal-Omidi S, Huang Y, Lhatoo SD, Zhang G, et al. Detection of postictal generalized electroencephalogram suppression: Random forest approach. JMIR Med Inform 2020 Feb 14;8(2):e17061 [FREE Full text] [doi: 10.2196/17061] [Medline: 32130173]
21. Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. 4th edition. London, UK: Springer-Verlag; 2014.
22. Hsu C, Liu C, Tain Y, Kuo C, Lin Y. Machine learning model for risk prediction of community-acquired acute kidney injury hospitalization from electronic health records: Development and validation study. J Med Internet Res 2020 Aug 04;22(8):e16903 [FREE Full text] [doi: 10.2196/16903] [Medline: 32749223]
23. Du Z, Yang Y, Zheng J, Li Q, Lin D, Li Y, et al. Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: Model development and performance evaluation. JMIR Med Inform 2020 Jul 06;8(7):e17257 [FREE Full text] [doi: 10.2196/17257] [Medline: 32628616]
24. Sohan M, Kabir M, Jabiullah M, Rahman SSMM. Revisiting the class imbalance issue in software defect prediction. In: Proceedings of the 2nd International Conference on Electrical, Computer and Communication Engineering. 2019 Presented at: International Conference on Electrical, Computer and Communication Engineering; February 7-9, 2019; Cox's Bazar, Bangladesh p. 1-6. [doi: 10.1109/ecace.2019.8679382]
25. Song S, Hua C, Dornbors D, Kang R, Zhao X, Du X, et al. Baseline red blood cell distribution width as a predictor of stroke occurrence and outcome: A comprehensive meta-analysis of 31 studies. Front Neurol 2019;10:1237 [FREE Full text] [doi: 10.3389/fneur.2019.01237] [Medline: 31849813]
26. Wang Y, Liu J, Wang X, Liu Z, Li F, Chen F, et al. Frequencies of circulating B- and T-lymphocytes as indicators for stroke outcomes. Neuropsychiatr Dis Treat 2017 Oct;13:2509-2518. [doi: 10.2147/ndt.s148073]
27. Liesz A, Hu X, Kleinschnitz C, Offner H. Functional role of regulatory lymphocytes in stroke. Stroke 2015 May;46(5):1422-1430. [doi: 10.1161/strokeaha.114.008608]
28. Spence JD, Yi Q, Hankey GJ. B vitamins in stroke prevention: Time to reconsider. Lancet Neurol 2017 Sep;16(9):750-760. [doi: 10.1016/S1474-4422(17)30180-1] [Medline: 28816120]
29. Jickling GC, Liu D, Ander BP, Stamova B, Zhan X, Sharp FR. Targeting neutrophils in ischemic stroke: Translational insights from experimental studies. J Cereb Blood Flow Metab 2015 Jun;35(6):888-901 [FREE Full text] [doi: 10.1038/jcbfm.2015.45] [Medline: 25806703]
30. Kim M, Jee SH, Yun JE, Baek SJ, Lee D. Hemoglobin concentration and risk of cardiovascular disease in Korean men and women - The Korean Heart Study. J Korean Med Sci 2013 Sep;28(9):1316-1322 [FREE Full text] [doi: 10.3346/jkms.2013.28.9.1316] [Medline: 24015036]
31. Chang Y, Hung S, Ling W, Lin H, Li H, Chung S. Association between ischemic stroke and iron-deficiency anemia: A population-based study. PLoS ONE 2013 Dec 9;8(12):e82952. [doi: 10.1371/journal.pone.0082952]

Abbreviations

AUC: area under the curve
CDC: Centers for Disease Control and Prevention

MPV: mean platelet volume
NCHS: National Center for Health Statistics
NHANES: National Health and Nutrition Examination Survey
NLR: neutrophil-to-lymphocyte ratio
NPV: negative predictive value
PPV: positive predictive value
RDW: red cell distribution width
ROC: receiver operating characteristic
SEQN: sequence number
WEKA: Waikato Environment for Knowledge Analysis

Edited by G Eysenbach; submitted 12.08.20; peer-reviewed by T Taveira-Gomes, K Kades, A Sule; comments to author 15.09.20; revised version received 14.11.20; accepted 15.10.21; published 02.12.21

Please cite as:
Alanazi EM, Abdou A, Luo J
Predicting Risk of Stroke From Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models
JMIR Form Res 2021;5(12):e23440
URL: https://formative.jmir.org/2021/12/e23440
doi: 10.2196/23440
PMID: