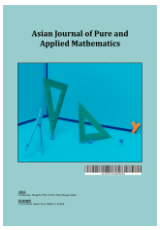

Asian Journal of Pure and Applied Mathematics

Volume 6, Issue 1, Page 174-181, 2024; Article no.AJPAM.1347

Application of Support Vector Machine Model for Prediction of Stroke Vulnerability Status

Okpe Anthony Okwori [a*], Moses Adah Agana [b] and Ofem Ajah Ofem [b]

[a] Department of Computer Science, Federal University Wukari, Nigeria.
[b] Department of Computer Science, University of Calabar, Nigeria.

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://prh.globalpresshub.com/review-history/1347

Original Research Article

Received: 24/06/2023
Accepted: 27/08/2023
Published: 20/06/2024

Abstract

Stroke is a chronic disease caused by lack of blood flow into some brain cells causing them to die due to oxygen deficiency. Cerebrovascular accidents (stroke) are the second leading cause of death and the third leading cause of disability and equally causes dementia and depression among the affected persons as well as their care takers. This disease affects people mostly at the peak of their life productive stage hence an urgent need for proactive measure through the prediction of stroke vulnerability using machine learning technique and subsequent stroke prevention. This paper aims at developing support vector machine model for the prediction of stroke vulnerability using healthcare_dataset_stroke_data obtained from Kaggle machine learning dataset repository after appropriate data preprocessing. It adequately employed the basic principles of machine learning to train the SVM model on the preprocessed dataset using python programming language. The SVM model was evaluated using python programming language sklearn evaluation metrics and the result obtained shows that support vector machine can adequately classified

**Corresponding author: E-mail: okwori@fuwukari.edu.ng;*

Cite as: *Okwori, Okpe Anthony, Moses Adah Agana, and Ofem Ajah Ofem. 2024. "Application of Support Vector Machine Model for Prediction of Stroke Vulnerability Status". Asian Journal of Pure and Applied Mathematics 6 (1):174-81. https://jofmath.com/index.php/AJPAM/article/view/163.*

patients as either vulnerable or not vulnerable to stroke using the stroke risk factors profile in the dataset as evident in its accuracy and area under the receiver operating characteristics curve (AUC) of 87% and 94% respectively.

Keywords: Stroke; machine_learning; support_vector_machine; python_programming; stroke_risk_ factors.

1 Introduction

Stroke is a medical codition caused by sudden death of some brain cells as a result of lost of blood flow to the brain due to blockage or rapture of brain ateries leading to oxegen deficiency. It is a multi-factorial medical condition that can lead to a pamanet physical disability among both elderly and young people. This illness has constituted a global nusants as it affect both high income, midum income and low income pupulation at the peak of their life time productivity. According to WHO [1], stroke occure as "rapidly developing clinical signs of focal (or global) disturbance of cerebral function, with symptoms lasting 24 hours or longer or leading to death, with no apparent cause other than vascular origin". Stroke is usually associated with some risk factors (behaviors or traits that makes one more vulnerable to a disease or medical condition). However, it is important to note that the presence of one or more risk factors in an individual does not guarantee that such person must surely develop the disease or the medical condition associated with the risk [2] but barely shows that the person have higher level of vulnerability to such disease or medical condition hence there is an urgent need to predict stroke vulnerability using a sophisticated machine learning technique such as support vector machine model in order to reduce stroke occurrence.

In general machine learning models plays a very relevant role in modern days medical practices as this area of artificial intelligence (AI) enable programs to analyze data, understand correlations between dependent and independent features in dataset in order to generate an insight for solving prediction problems [3]. Machine learning models provides a fast and efficient prediction outputs hence serves as a powerful tool in efficient system security which is needed by every user [4] healthcare services as it offers adequate personalized clinical care for stroke patients. Machine learning applications for disease prediction and diagnosis not only increases the processing speed but adequately reduces the processing cost [5].

Even though several efforts have been made using various machine learning techniques in predicting stroke occurrence with various prediction performance levels, the optimal prediction requirement has not been achieved [6] as those prediction models usually have large number of false positives and false negatives hence a need to use an SVM model for stroke prediction. The support vector machine algorithm is a derivative of statistical learning theory and is capable of compressing raw dataset into a support vector set and use it to learn and generate a classification decision function. It usually constructs an optimal hyperplane as the decision boundary [7] such that the distance between the positive and negative parts is maximum. The support vector machine algorithm was first implemented in 1963 by Vapnik and Alexey to draw hyperplane for linear classification which shows that the SVM was traditionally design with the basic principle of linear and non-probabilistic classification, however in recent time SVM has been developed to work on non-linear classification problems [8] by incorporating the kernel concept in a high-dimensional workspace and uses a constructor to estimate class membership probability [9] respectively. Support vector machine differentiates between two classes by generating a hyperplane that optimally separates the classes after the input data has been transformed mathematically into a high-dimensional space. The support vector machine technique is a data-driven technique with high discriminative power for classification when the sample size is small and a large number of variables are involved (high-dimensionality) as opined by Yu et al. [10].

2 Literature Review

In separate researches, [11,12] affirmed that support vector machine yields higher prediction performance than other related prediction algorithms. Jovel et al. [6] used stroke risk factor variables to develop a support vector machine learning based predictive model for stroke occurrence using the patient's medical records.

The researcher used cavite hospital dataset of 1500 patients on the SVM model and concluded that SVM model is an ideal machine learning algorithm for stroke prediction due to its high prediction performance. The performance of Stroke risk prediction model using several machine learning techniques such as support vector machines SVM, decision trees, nearest neighbors, and multilayer perception has been compared and was discovered that SVM is the most promising algorithm with high sensitivity and specificity, it reduces over-fitting and is capable of classifying non-linear data [13]. Kumari et al. [14] uses SVM Radial basis function (RBF) kernel to developed a support vector machine based diabetes classification model using the Pima Indian diabetic database at the UCI machine learning laboratory on Matlab. From the result obtained, the researcher concluded that SVM with RBF kernel can be successfully used for diagnosing diabetes and other common disease such as stroke with simple clinical measurements without explicit laboratory tests. [15] developed a heart disease prediction model using support vector machine algorithm. The researcher concluded that Support vector machine is highly efficient in heart disease prediction as evident in its high accuracy, specificity and sensitivity. Patil Patil et al. [16] developed a support vector machine learning based heart disease prediction model using datasets obtained from four deferent places: ClevelandClinic foundation (cleveland.data), Hungarian Institute of Cardiology (hungarian.data), Medical Center, Long Beach (long-beach-va.data) and University Hospital Zurich (switzerland.data) on Weka tools. The researchers found that the support vector machine heart disease classifier is highly accurate, sensitive, and specific and hence concludes that SVM is a very good algorithm for classifying binary events such as heart disease and stroke condition. Yu et al. [10] developed a web-based support vector machine model for diabetes classification using a dataset obtained from the 1999-2004 National Health and Nutrition Examination Survey (NHANES) of the U.S. population. This SVM web-based classifier was built using J2EE technology and other Java open-source frameworks such as Hibernate and Strut. The researchers found that, SVM models are very efficient in classifying people with common diseases such as diabetes and pre-diabetes and could be used in the classification of other complex diseases using the disease risk factors. Deepika et al. [5] developed a support vector machine based Breast Cancer prediction model using breast cancer dataset from UCI machine learning repository. The SVM breast cancer prediction model is capable of classifying the cancer tumor into the two basic stages of the breast cancer namely: benign or malignant. To achieve target classification performance, the dataset was preprocessed using the dimensionality reduction techniques and the clean data was then used for the training, validation and testing of the model. The researcher concluded that SVM could be used for efficient detection and prevention of breast cancer and other chronic diseases as the model classification accuracy is about 99%. The efficiency of support vector machine in binary classification was equally evident in the classification of cancer of the lung by Prajapati et al. [17]. The researchers were able to adequately classify cancerous and non-cancerous CT images using SVM model developed from the cancer imaging archive (TCIA) dataset on MATLAB image processing toolbox after the various dataset preprocessing were carried out. Emon et al. [18] developed a heterogenous ensemble machine learning models of ten base learners: Logistics Regression, Stochastic Gradient Descent, Decision Tree Classifier, AdaBoost Classifier, Gaussian Classifier, Quadratic Discriminant Analysis, Multi layer Perceptron Classifier, KNeighbors Classifier, Gradient Boosting Classifier and XGBoost using weighted voting approach for predicting stroke vulnerability status. Both the ten base learners and the voting ensemble model were evaluated using confusion matrix and its associated metrices and it was found that the area under the receiver operating characteristic curve (AUC-ROC) of all the models ranges from 73% to 93% which shows that all the models are good in classifying stroke vulnerability status.

3 Materials and Methods

In this paper, a Stroke Prediction model was developed for the prediction of likelihood of stroke using support vector machine algorithm. A supervised learning approach was used for the creation and training of the SVM model. Confusion matrix, using the True and False Positives and True and False Negatives was used to evaluate the performance of the SVM model. This system was implemented using python programming language as python is a general-purpose dynamic programming language that provides high-level readability, Simplicity, consistency, flexibility, platform independence and less Codes as it provides access to great libraries and frameworks. To develop the support vector machine model, relevant research information was obtained from related literatures. The system uses the healthcare_dataset_stroke_data obtained from Kaggle machine learning dataset repository to train, validate and test the support vector machine model. This dataset was downloaded and renamed as stroke and converted to comma separated

value (CSV) file "stroke.csv" using Microsoft excel for easy data preprocessing using Pandas machine learning library. It originally consists of 12 columns and 5110 rows. The dataset has 11 independent variables features: id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi and smoking_status with one dependent variable class label "stroke". The dataset is described in Table 1 as shown.

Table 1. Dataset description

Feature No.	Feature Name	Feature Description
1	Id	Unique identification number for each data point in the dataset
2	Gender	Male or Female
3	Age	Number of years of a patient
4	Hypertension	Presence or absence of hypertension
5	Heart_disease	Presence or absence of heart disease
6	ever_married	Married or not marred
7	work_type	Children, Private, Never worked, Govt. job or Self employed
8	Residence_type	Urban or rural residence
9	avg_glucose_level	Average quantity of glucose in the patient body
10	Bmi	The body mass index
11	smoking_status	Never smoked, formally smoked or smokes
12	Stroke	Presence or absence of stroke

To facilitate efficient performance of the SVM model, the dataset was preprocessed using various data preprocessing techniques such as feature selection, feature encoding, missing values detection and correction, class balancing, outliers detection and correction, feature scaling as well as proper turning of the various hyper parameters. After the entire data preprocessing exercise, the preprocessed stroke.csv has 9722 rows with 11 features that was split into training dataset, validation dataset and testing dataset respectively using sklearn python library. Out of the 9722 data records, 80% (7776) data items were used for training, 10% (973) data items were used for validation and 10% (973) data items were used for testing of the models respectively. The dataset was split into 80% training, 10% testing and 10% validation to prevent overfitting and to accurately evaluate the model performance. The training set is used to fit the model, the validation set is used to tune the hyperparameters, and the test set is used to measure the generalization error. To make binary classification with SVM, the decision boundary clearly separates the two classes and check if the new data instance belongs to either of the two classes. In this paper, a training dataset of the healthcare-dataset-stroke-data consisting of the two target classes "stroke" and "nostroke" denoted by small circles and small square shapes respectively were fitted into the support vector machine algorithm for learning. The SVM algorithm constructs hyperplanes H, H1 and H2 such that H1 is closest to the stroke target class and H2 is closest to the nostroke target class while H is positioned midway between H1 and H2. The line H that optimally separates the two target classes is referred to as the hyperplane while the distance between H1 and H2 is referred to as the classification interval as shown in Fig. 1.

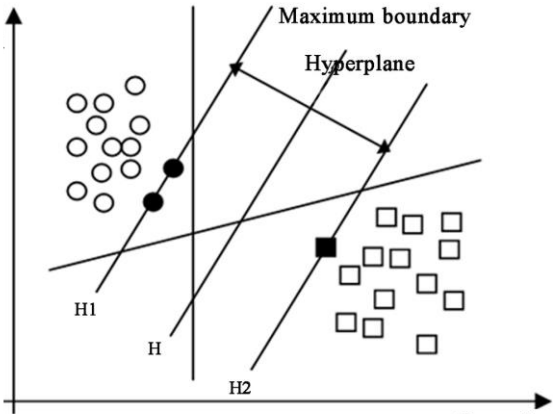

Fig. 1. SVM prediction model

The following algorithm describes the implementation of support vector machine model using python programming language.

Algorithm 1: support vector machine implementation algorithm

Step 1: Import the relevant libraries
Step 2: Import the dataset
Step 3: Read the dataset
Step 4: Pre-process the dataset
Step 5: split the dataset
Step 6: Create the Support Vector Machine model object
Step 7: Fit the model
Step 8: Predict result
Step 9: Evaluate the model

To generate the support vector machine model, a support vector Classifier (SVC) was imported from the python sklearn support vector machine library and the various hyperparameters were tuned. To train the model, the training dataset was fit into the SVC Classifier and evaluated using the stratified 10-fold cross-validation to obtain a suitable model performance and confusion matrix to compute the values of the various matrices.

4 Evaluation result of the SVM and Discussion

The support vector machine was evaluated using confusion matrix and its related matrices such as Accuracy score, Precision score, Recall score, F1 score, Sensitivity score, Specificity score, and AUC score respectively. Table 2 below shows the confusion matrix generated by support vector machine which is used in the computation of the values of all its related metrics.

Table 2. Confusion matrix for support vector machine

N = 973		**Actual values**	
		Positive (yes)	**Negative (no)**
Predicted values	Positive (yes)	Tp = 411	Fp = 76
	Negative (no)	Fn = 46	Tn = 440

From Table 2 above 411 data instances were predicted positive and are actually positive (true positive), 440 data instances were predicted negative and are actually negative (true negative), 76 data instances were predicted positive but were actually negatives (false positives) while 46 data instances were predicted negatives and there are actually positive (false negatives). The various values of the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) in the confusion matrix above were used to compute the values of the various evaluation metrics as shown in Table 3.

Table 3. Results of support vector machine evaluation

Matrix	**Accuracy**	**Precision**	**Recall**	**F1 score**	**Sensitivity**	**Specificity**	**AUC**
Value	0.87	0.85	0.91	0.88	0.91	0.84	0.94

From Table 3 above it can be seen that support vector machine have predicted accuracy of 87%, precision of 85%, recall of 91%, F1 score of 88%, sensitivity of 91%, specificity of 84% and AUC of 94%. This shows that the algorithm performs very well in stroke disease prediction using the healthcare-dataset-stroke-data.

Its eminent performance is evident in the value of its area under receiver operating characteristic AUC- ROC curve as shown in Fig. 2 below.

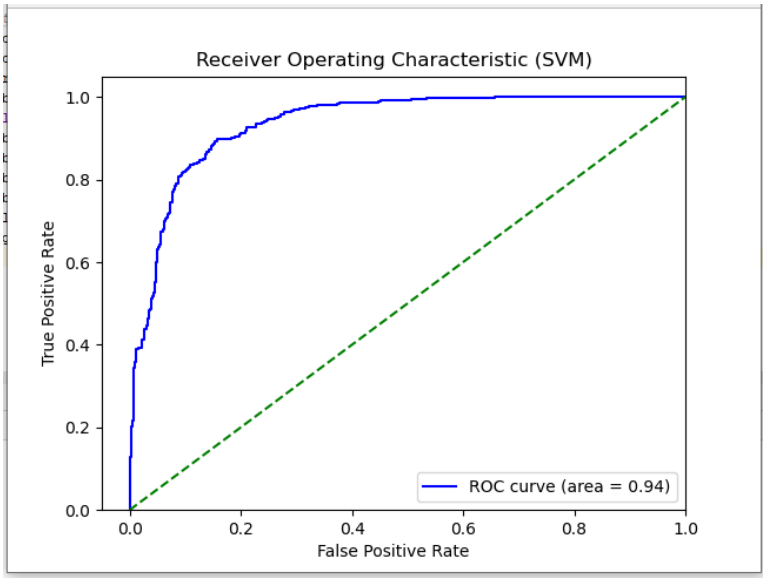

Fig. 2. AUC-ROC curve for support vector machine

The AUC – ROC curve of Fig. 2 above shows the performance of SVM at various thresholds settings. ROC is a probability curve and AUC represents the degree or measure of separability between the class stroke and the class no stroke. It shows how much the support vector machine model is capable of distinguishing between classes that are vulnerable to stroke and classes that are not vulnerable to stroke. the higher value of 94% for the AUC demonstrated how good the developed model is at predicting patients that are vulnerable to stroke as been vulnerable and patient that are not vulnerable to stroke as not been vulnerable.

5 Conclusion

In this paper, an efficient support vector machine based stroke prediction model was developed using the healthcare_dataset_stroke_data obtained from Kaggle machine learning repository to assist medical practitioners predict a very highly accurate and dependable stroke vulnerability status of an individual. Timely prediction of stroke vulnerability status of a patient enhances proactive major for stroke prevention or proper management of stroke condition that significantly increase the chances of long-term longevity of life and overall survival. This paper demonstrated the efficacy of machine learning models in disease prediction and diagnoses as the SVM model was able to classify individuals as vulnerable or not vulnerable to stroke with 87% accuracy 94% AUC. Hence it can be used as an efficient machine learning application for predicting and preventing stroke occurrence. Although this result is efficient and within the acceptable range when compared to current literatures, it can be improved upon by hybridizing good multiple binary classifiers to achieve a better prediction performance.

Disclaimer (Artificial Intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] WHO. World Health Organization (WHO). Definition of Stroke; 202.
Available:https://www.World Health Organization (WHO) Definition of Stroke (ngontinh24.com)

[2] Cleveland C. Know Your Risk Factors for Stroke.
Available: https://my.clevelandclinic.org/health/articles/13398-know-your-risk-factors-for-stroke, accessed on 25th February, 2022

[3] Harini D, Akash S, Durai RSV, Archana J. Heart Disease Prediction and Medicine Prescription using SVM. International Journal of Research in Engineering. Science and Management. 2019;2(4): 41-44.

[4] Ogbu HN, Agana MA. Intranet security using a LAN packet sniffer to monitor traffic. In Natarajan M.(Eds). 2019;9(8):57-68: CCSIT, NCWMC, DaKM

[5] Deepika S, Kapilaa Ramanathan K, Devi N. Prediction of Breast Cancer Using SVM Algorithm, International Journal of Applied Engineering Research. 2021;16(4):316-320.

[6] Jovel TR, Alexander AH. Developing a Predictive Model of Stroke using Support Vector Machine. IEEE. 2019;32:35-40.

[7] Jianfang C, Min W, Yanfei L, Qi Z. Improved support vector machine classification algorithm based on adaptive feature weight updating in the Hadoop cluster environment. Plos One. 2019;14(4). Available: https://doi.org/10.1371/journal.pone.0215136

[8] Lumbanraja FR, Fitri E, Ardiansyah Junaidi A, Rizky P. Abstract classification using support vector machine algorithm (Case Study: Abstract in a Computer Science Journal). Journal of Physics: Conference Series. 2021;1-13.

[9] Pedregosa p. Support vector machines, scikit-learn: Machine learning in python; 2011.
Available:https://scikit-learn.org/stable/modules/svm.html#scores-probabilities,
Accessed on 17th February, 2022

[10] Yu W, Tiebin L, Rodolfo V, Marta G, Muin JK. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes, Medical Informatics and Decision Making; 2010.
Available:https://bmcmedinformdecismak.biomedcentral.com/track/pdf/10.1186/1472-6947-10-16

[11] Ahmet KA, Cemil C, Ediz S. Different medical data mining approaches based prediction of ischemic stroke. Computer methods and programs in biomedicine.
Available:https://www.researchgate.net/publication/299375650_Different_Medical_Data_Mining_Approaches_Based_Prediction_of_Ischemic_Stroke

[12] Youn-Jung S, Hong-Gee K, Eung-Hee K, Sangsup C, Soo-Kyoung L. Application of support vector machine for prediction of medication adherence in heart failure patients, Healthc Inform pres. 2010;253-259.
Available:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3092139/pdf/hir-16-253.pdf

[13] Haewon B. Predicting the swallow-related quality of life of the elderly living in a local community using support vector machine. International Journal of Environmental Research and Public Health. Available:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6862249/pdf/ijerph-16-04269.pdf

[14] Kumari VA, Chitra R. Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications (IJERA). 2013;3(2):1797-1801.

[15] Sandhya Y. Prediction of Heart Diseases using Support Vector Machine, International Journal for Research in Applied Science & Engineering Technology (IJRASET). 2020;8(II):2321-9653.

[16] Patil M, Jadhav R, Patil V, Bhawar A, Chillarge G. Prediction and analysis of heart disease using SVM algorithm. International Research Journal of Engineering and Technology (IRJET). 2019;6(3): 872-875.

[17] Prajapati P, Hande V, Ingale A, Dwivedi S. Lung cancer detection and classification using SVM. Journal of Emerging Technologies and Innovative Research (JETIR). 2019;6(5):60-64.

[18] Emon MU, Keya MS, Meghla TI, Rahman M, Mamun SA, Kaiserk MS. Performance Analysis of Machine learning approaches in stroke prediction. International Conference on Machine Intelligence and Emerging Technologies. 2022;1-6.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
https://prh.globalpresshub.com/review-history/1347