Exploring Stroke Risk Identification by Machine Learning: A Systematic Review

Lelis Raquel Atencia Mondragon
20190175@aloe.ulima.edu.pe
https://orcid.org/0009-0002-7245-9577
Universidad de Lima, Perú

Melany Cristina Huarcaya Carbajal
20192902@aloe.ulima.edu.pe
https://orcid.org/0009-0005-1752-2950
Universidad de Lima, Perú

Rosario Guzmán Jiménez
rguzman@ulima.edu.pe
https://orcid.org/0000-0002-4550-7935
Universidad de Lima, Perú

Received: June 27, 2023 / Accepted: September 12, 2023
doi: https://doi.org/10.26439/ciis2023.7081

ABSTRACT. This work aims to systematize previous studies on stroke risk identification and its relationship with machine learning. A systematic review was conducted using the Web of Science and Scopus databases. The information was organized into three sections: stroke risk factors, data preprocessing techniques and techniques for identifying stroke risk with an emphasis on the most important features. The main results are as follows: risk factors are divided into modifiable (work environment and air pollution) and non-modifiable (sex, family history). The most commonly used data preprocessing techniques are SMOTE, standardization and value elimination/imputation. The most commonly used techniques for identifying stroke risk include support vector machine, random forest, logistic regression, naïve Bayes, k-nearest neighbors and decision tree.

KEYWORDS: stroke / models / machine learning / risk / identification

EXPLORACIÓN DE LA IDENTIFICACIÓN DEL RIESGO DE ACCIDENTE CEREBROVASCULAR MEDIANTE MACHINE LEARNING: UNA REVISIÓN SISTEMÁTICA

RESUMEN. Este trabajo busca sistematizar los estudios sobre la identificación del riesgo de sufrir un accidente cerebrovascular (ACV) en las bases de datos Web of Science y Scopus y su relación con el *machine learning*. La información se organizó en tres secciones: factores de riesgo del ACV, técnicas de preprocesamiento de datos y técnicas para identificar el riesgo de ACV haciendo énfasis en las características más relevantes. Los principales resultados son los siguientes: los factores de riesgo se dividen en modificables (ambiente de trabajo y contaminación del aire) y no modificables (sexo, hipertensión). Las técnicas de preprocesamiento más utilizadas son SMOTE, estandarización y eliminación/imputación de valores. Las técnicas más usadas para identificar el riesgo de sufrir ACV son *support vector machine*, r*andom forest*, *logistic regression*, *naive Bayes*, *k-nearest neighbors* y *decision tree*.

PALABRAS CLAVE: accidente cerebrovascular, modelos, aprendizaje automático, riesgo, identificación

1. INTRODUCTION

Cerebrovascular accident (CVA) or stroke is an acute phenomenon that occurs due to obstructions that prevent blood flow to the brain (World Health Organization, 2021). Factors such as smoking, overweight or obesity, as well as high cholesterol and glucose levels, influence this disease (Sarfo et al., 2022). It is estimated that approximately one out of four people will be prone to a stroke after the age of 25 (The GBD 2016 Lifetime Risk of Stroke Collaborators et al., 2018). Each year, stroke affects approximately 15 million people and 5 million of this group become disabled (World Health Organization Regional Office for the Eastern Mediterranean, n.d.).

In Peru, there has been an increase in stroke cases, and they are being considered one of the main causes of permanent disability in adults (Bernabé-Ortiz & Carrillo-Larco, 2021). The sequelae of this disease have an economic and health impact, affecting the quality of life of patients (Langhorne et al., 2000). According to studies conducted in the United States, the costs related to medical care, medications and loss of productivity reach significant figures (King et al., 2020).

Early detection and prevention of strokes are essential to reduce their impact. Identifying symptoms, such as vision problems as well as difficulty walking, reading or speaking, is crucial to seek help immediately (Centers for Disease Control and Prevention, 2022). It is estimated that about 80 % of strokes are preventable (Linn et al., 2014). In addition, researches that apply classification algorithms to identify stroke risk and related diseases have been conducted.

Different studies, including those carried out by Chantamit-o-pas and Goyal (2017), Alaa et al. (2019), Mohan et al. (2019), among others, have compared and evaluated the performance of various machine learning (ML) algorithms in predicting stroke risk. Some of these algorithms include deep learning (DL), naive Bayes (NB), support vector machine (SVM), random forest (RF), artificial neural network (ANN), adaptive boosting (AdaBoost) and gradient boosting (GB).

There is a gap in research conducted in Peru with respect to life cycle assessment (LCA) and ML. According to Chantamit-o-pas and Goyal (2017), future work using more risk factors is recommended. On the other hand, Dritsas and Trigka (2022) propose imaging-based studies and the use of DL techniques for stroke detection. However, so far, studies comparing and implementing ML algorithms to identify stroke risk have been mainly undertaken in countries such as the United States, England and China.

Unfortunately, no studies have been carried out for risk identification using ML algorithms based on Peruvian patient medical records, despite the fact that this population is also affected by strokes. Even the projects of the Peruvian Ministry of Health (MINSA) for the years 2020 to 2025 do not contemplate the use of ML for disease risk identification (Ministerio de Salud del Perú, 2020).

Castañeda-Guarderas et al. (2011) conducted a study that analyzed reports of stroke incidence in patients at the Cayetano Heredia Hospital between 2000 and 2009. In this study, arterial hypertension, atrial fibrillation and type 2 diabetes mellitus were identified as the most common associated conditions. One of the recommendations was to establish measures and interventions to reduce the mortality rate, prevent new events and improve patient outcomes.

Based on this context, the following research question arises: how has the application of machine learning algorithms in stroke risk detection been addressed in various studies?

In line with this, the overall objective of the present study is to perform a systematic review of studies related to the identification of stroke using ML. In addition, the following specific objectives are proposed: to identify the variables that significantly influence stroke risk, to identify the most appropriate preprocessing techniques, and to identify the characteristics of the most effective ML techniques. This compilation article is intended to provide an overview of the studies undertaken to date, providing a concise basis for achieving the abovementioned objectives.

2. METHODOLOGY

A systematic review of the literature was carried out with the purpose of identifying and analyzing relevant studies related to stroke risk identification using ML techniques. The main objective of this research was to gather detailed information on the most common risk factors, data preprocessing techniques and both risk identification algorithms and their major characteristics.

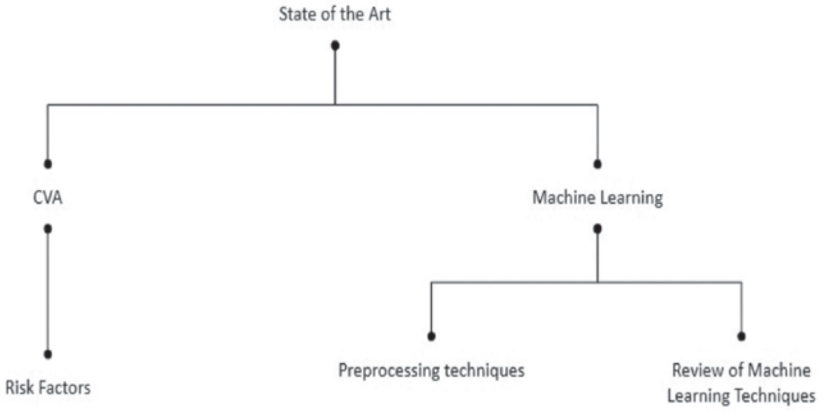
To perform the exhaustive search for studies, the renowned databases Web of Science and Scopus were accessed. The keywords "stroke," "risk" and "machine learning" in combination with Boolean operators were used to identify relevant studies. Inclusion criteria include studies related to the implementation of ML techniques for stroke risk identification and studies addressing risk factors with a more health-related approach.

Once the relevant studies were identified, relevant information was extracted using a data extraction matrix. The extracted data was analyzed using a descriptive approach, focusing especially on the comparison of the techniques used in stroke risk identification. The results were presented in a clear and concise manner using comparison tables that summarized the risk factors, preprocessing techniques and identification algorithms used in the reviewed studies.

3. RESULTS

This section presents a structured analysis of the literature reviewed and includes two subsections: one directly related to LCA, which contains the identified risk factors, and the other to ML. In turn, this subsection contains two subdivisions, i.e., one referred to data preprocessing techniques and the other to ML techniques for LCA identification. The sections are displayed in Figure 1.

Figure 1
Map of the structure of the state of the art



3.1 CVA

CVA is a complex disease and one of the leading causes of mortality in the adult population. Several factors are known to influence patient risk, including demographic variables, comorbidities or genetics (Torres-Aguila et al., 2019).

3.1.1 Risk factors

Before collecting patient medical records to identify stroke risk, it is necessary to know which variables influence the risk for this disease. This subsection shows the factors found in articles focused on predicting strokes with ML. Subsequently, another list of variables gathered from articles focused on the factors found in recent years is presented.

Table 1 shows the variables included in some of the datasets used to identify stroke risk with ML from the literature review. As shown, all the datasets include variables such as age, sex and comorbidities. However, some of these differ in taking into account variables concerning patient habits such as lifestyle, tobacco or alcohol consumption. Furthermore, according to the dataset analyzed in the study by Hippisley-Cox et al. (2017), family history also affects stroke risk.

In Table 2, information on risk factors with a more medical approach was systematized. Therefore, these factors have less presence in the collected datasets. However, the articles found highlight their importance.

Table 1

Datasets used for CVA risk identification

Authors	Source	Number	Data
Dritsas and Trigka (2022) Tazin et al. (2021) Ahammad (2022)	Stroke Prediction Dataset – Kaggle	5 110	Sex, age, hypertension, heart disease, married, occupational status, residence, glucose, BMI, smoking.
Chantamit-o-pas and Goyal (2017)	UCI Machine Learning Repository	899	Age, sex, blood pressure, chest pain, smoking, family history, hypertension, cholesterol, pulse, blood vessels.
Khdair and Dasari (2021)	South African Heart Disease – KEEL Repository	462	Systolic blood pressure, smoking, LDL cholesterol, adiposity, family history, type A behavior, obesity, alcohol, age, coronary artery disease.
Qu et al. (2022)	Shenzhen Hospital of Guangzhou University of Chinese Medicine	711	Age, sex, smoking, alcoholic beverage intake, medical history, weight, height, blood pressure, BMI, cholesterol, triglycerides, glucose, hemoglobin, among others.
Jabal et al. (2022)	Patients from two academic centers	443	Age, sex, accident scale comorbidities, diabetes, hypertension, hyperlipidemia, blood glucose levels, blood pressure.

Table 2

Risk factors according to studies with a medical approach

Authors	Factors	Study Subject	Findings
Torres-Aguila et al. (2019)	Leukoaraiosis, other clinical complications, blood components, acute treatments, neurological complications, genetic factors.	General population	Age and sex, which are used as covariates in association studies such as GWAS, had a very weak influence, at least in the short term.
Zhang et al. (2019)	Atrial fibrillation, heart failure, hypertension, genetic factors, chronic kidney disease, obstructive sleep apnea, migraine with aura, work environment, air pollution.	General population	Internal and external risk factors are important for stroke prevention. More importantly, a thorough understanding of each risk factor provides specific guidance in practice.

According to the comprehensive literature review, twelve of the articles showed these variables: age, sex and diseases. The disease most frequently repeated as a variable in the datasets was hypertension.

Although the three abovementioned variables are the most common ones, according to Torres-Aguila et al. (2019), age and sex had a very weak influence on stroke risk. The newly identified characteristics, as shown in Table 2, are work environment and air pollution. These, despite having significance, were not included in the found datasets.

In summary, stroke risk factors can be classified into modifiable and non-modifiable. The main non-modifiable factors are usually sex and family history. On the other hand, the main modifiable factors are associated with the patients' health status, such as arterial hypertension, diabetes and hypercholesterolemia, and to habits such as alcohol and tobacco consumption.

3.2 Machine learning

Data processing is crucial for any application process with ML algorithms. This section gathers information on processing techniques and algorithms used for LCA risk identification with ML.

3.2.1 Preprocessing Techniques

In order to implement the techniques for the identification of stroke or similar diseases, it is necessary to preprocess the data. These techniques are useful to perform a correct and adequate training of algorithms since they can present null values, outliers, values with a format that is not in accordance with the type of variables used and can cause the results to be erroneous. Table 3 shows the main preprocessing techniques in the literature reviewed.

Table 3

Data preprocessing techniques used for CVA risk identification

Authors	Technique	Purpose
Mohan et al. (2019)	Removal of data with missing values, multiclass and binary classification.	Edit missing values.
Dritsas and Trigka (2022)	Redundant data reduction, variable selection, data discretization, resampling method.	Avoid degradation of prediction accuracy.
Chen et al.(2022)	Missing data processing: continuous variables were filled with linear imputation and categorical variables with mode.	Edit missing values.
Shoily et al. (2019)	Data standardization.	Work with quantitative data.
Tazin et al. (2021)	Missing data analysis, data balancing, one-hot encoding. Use of SMOTE.	Remove unnecessary data for better efficiency, convert qualitative variables into numbers for training, eliminate null values.

continues

continued

Authors	Technique	Purpose
Nusinovici et al. (2020)	Missing data imputation, multicollinearity, dummy variables for categorical factors.	Ensure better accuracy.
Lin et al. (2020)	LOWESS and standard deviation.	Eliminate illogical evaluations and adjust quantities to the same scale.
Qin et al. (2021)	Data standardization, SMOTE, hyperparameter optimization.	Normalize, balance and produce the best performance.
Zhang et al. (2022)	Elimination of variables with more than 30% missing data and R method.	Eliminate inconsistencies and identify correlated variables.

The findings in the table reveal a number of common techniques used in machine learning data preprocessing. Among the most recurrent techniques are missing data imputation, data standardization, redundant data removal, data balancing and variable selection. Mohan et al. (2019) focus on removing data with missing values to improve data integrity, while Shoily et al. (2019) advocate data standardization, which facilitates quantitative data analysis. Dritsas and Trigka (2022) prioritize redundant data reduction and variable selection as key approaches to maintain prediction accuracy. Tazin et al. (2021) emphasize the importance of data balancing, while Qin et al. (2021) focus on hyperparameter optimization. These techniques are applied in different contexts, contributing significantly to the quality of the results obtained in the process of data preparation for machine learning models. In summary, these findings underline the critical importance of data preprocessing in improving data quality, which is supported by the research of the aforementioned authors.

3.2.2 Techniques for CVA identification

In the literature reviewed, a number of different algorithms and techniques have been implemented for stroke risk identification. Depending on the authors and the proposed objectives, these techniques can be evaluated individually to test their performance. This individual evaluation occurs mainly when the authors develop a hybrid algorithm, such as that of Zhang et al. (2022), who implemented a model based on logistic regression (LR), ANN, RF and GB.

Another measure of evaluation occurs when the authors compare techniques to demonstrate which one is the best performing. Comparative articles are presented in most of the literature reviewed. Among them is Dritsas and Trigka (2022), who used a dataset with 5110 records and 11 features from the Kaggle platform. The authors compared more than five ML models, including NB, LR, RF and Stacking, in order to determine which one was the best performing in identifying stroke risk. As a result, the stacking method performed better on the basis of the area under the curve (AUC).

Another application of comparative articles can be seen in the work of Liu et al. (2021), whose main objective was to identify the main risk factors for stroke in a province in China. For this purpose, they had two datasets. One contained 2000 records of stroke patients and the other one more than 27 000 records from a stroke prevention project in China. They were categorized into low risk, medium risk and high risk. Both datasets had more than 100 features. In order to meet their goal, they implemented decision tree (DT) and RF models. The DT algorithm showed that the main features were hypertension, physical inactivity and diabetes mellitus. While RF showed that the main features were hypertension, hyperlipidemia and physical inactivity.

As can be seen, the applications of the techniques are diverse. In the literature reviewed, some preference was observed for SVM, LR, NB, K-Nearest Neighbor (KNN), RF and DT.

As for SVM, Sailasya and Kumari (2021), Ahammad (2022) and Khdair and Dasari (2021) chose this technique because it is useful for classification and regression of variables. In addition, Dinesh et al. (2018) claimed that this model is quite well known and used due to its efficiency. However, Ahammad (2022) acknowledged that its performance may decrease in datasets with a large number of variables. In turn, it can be observed that in the article by Nusinovici et al. (2020) the SVM showed the lowest AUC among the other models. The implemented dataset contains more than 10 000 records. Similarly, in the paper by Alaa et al. (2019), where the implemented dataset consisted of more than 400 000 records, the SVM accounted for the lowest AUC. On the other hand, the work of Mohan et al. (2019), where the dataset consisted of 303 records, the SVM algorithm obtained an average value among the other records. One contrast to mention is the paper by Lin et al. (2020), whose dataset consisted of more than 58 000 records and the SVM algorithm showed the best result. This may be due to the variables implemented.

Regarding the LR model, Dritsas and Trigka (2022), Khdair and Dasari (2021) and Sailasya and Kumari (2021) highlighted its usefulness for binary classifications; Dinesh et al. (2018) agreed with the authors and added the efficiency of the model. However, they mentioned that, in case more variables were present, multinomial logistic regression was applied. Furthermore, Tazin et al. (2021) added that this model was quite widely used for the prediction of dichotomous variables.

Giving a focus on the NB model, Dinesh et al. (2018) mentioned that it is one of the best ML classification options due to its scalability. Furthermore, Dritsas and Trigka (2022) observed that it ensures the maximization probability according to the independence of the variables. Likewise, Shoily et al. (2019) claimed about its ease of use, its simplicity of handling when multiple variables are present, as well as its scalability and efficiency with discrete and continuous data. However, they remarked that it performs better on small datasets. Regarding the last point, no cases can be provided to support Shoily et al. (2019) except that the model has had average results in the different metrics.

The KNN model, according to Shoily et al. (2019), is the simplest algorithm because it does not require training. Moreover, Ahammad (2022) claimed that, by requiring a *k* value in order to calculate the nearest neighbors, it could be complicated in large datasets.

Dritsas and Trigka (2022) stated that RF creates a subset of instances for classification and regression tasks. For that reason, it is optimal in jobs. Tazin et al. (2021) performed a correlation between that mentioned by Dinesh et al. (2018) and the algorithm logic and explained that RF creates different decision trees per attribute. In addition, it can take care of data preprocessing. Dritsas and Trigka (2022) added that these trees are created during training. Moreover, they highlighted the flexibility of this model. Likewise, Ahammad (2022) highlighted the efficiency of RF with numerical and categorical variables; however, he pointed out that predictions may take time at the training stage.

Concerning the DT algorithm, Tazin et al. (2021) emphasized its usefulness for regression and classification. In addition, they alluded to its simplicity of understanding and replication and commented that one of its characteristics is its support for decision-making and the low need for data cleaning. Dritsas and Trigka (2022) complemented this point of view by saying that DT gives support in the reduction of errors as the model is built. Nikam et al. (2020) added that DT organizes features into different targets. Mohan et al. (2019) also highlighted the speed and simplicity of its implementation. On the other hand, Ahammad (2022) stated that the algorithm can obtain a good accuracy depending on the dataset. However, he detailed that it may present delays if the dataset is large.

Table 4 presents the main characteristics of the ML techniques found in the literature review.

Table 4

Advantages and disadvantages of machine learning techniques for CVA risk identification

ML Algorithm	Advantages	Disadvantages
SVM	Useful for classification and regression of variables.	Performance may decrease in datasets with large number of variables.
LR	Useful for binary classifications.	Need for multinomial logistic regression for multiple variables.
NB	Ease of use, scalability, efficiency with discrete and continuous data.	Not specified.
KNN	Simplicity, as it does not require a prior training process.	In large datasets, the choice of a K value in KNN can be complicated.
RF	Optimal performance in creating subsets of instances, flexibility in model building, efficiency with numerical and categorical variables	At the training stage, predictions may experience delays.
DT	Useful in regression and classification, easy to understand and replicate, low need data cleaning.	May experience delays when the dataset is large.

4. DISCUSSION

With respect to the section on risk factors, it should be noted that the authors agree on various characteristics, including age, sex, hypertension, among others. In addition, there are non-modifiable factors such as age, sex and family history, and modifiable factors such as arterial hypertension, diabetes, hypercholesterolemia and habits, including alcohol and tobacco consumption. It is worth mentioning that the need to explore the patient's medical history should be evaluated, since some diseases or comorbidities that are not so frequently observed should be considered in order to better identify the risk of suffering a CVA.

In the section on preprocessing techniques, which are fundamental for obtaining satisfactory results in identifying CVA risk, the discussion highlights the importance of similarities among the implemented techniques, as well as the need to evaluate and perform an exploratory analysis on the dataset. This ensures the implementation of appropriate preprocessing techniques aligned with the research objectives. For example, at the moment of carrying out a research work related to the exposed topic, it is recommended to analyze in the dataset the null values, distribution of variables, standardization and the need to perform data balancing, which are the characteristics of the datasets most frequently addressed in the present research. As a solution, techniques such as data elimination, standardization and data balancing are used.

In the section on implementation techniques, the SVM, LR, RF and NB algorithms are the ones that showed the best performance. It should be noted that these algorithms displayed a lower performance than usual in few works due to dataset characteristics such as its dimensions and the preprocessing techniques used. Even so, these techniques can be a useful tool in future research works considering the size of the dataset and the objective of the work to be developed.

5. CONCLUSIONS

This study aimed to carry out a systematic compilation of articles related to stroke and related diseases. It was found that the main risk factors for stroke are age, sex, hypertension, blood pressure, smoking, alcohol consumption, family history and hypercholesterolemia. In addition, the research showed that the main preprocessing techniques are SMOTE, standardization and value elimination/imputation. These techniques can be developed with different tools. Also, it introduced the most commonly used identification techniques: SVM, RF, LR, NB, KNN and DT. These findings can serve as a guide for future empirical research to identify the risk of CVA.

REFERENCES

Ahammad, T. (2022). Risk factors identification for stroke prognosis using machine-learning algorithms. *Jordanian Journal of Computers and Information Technology*, *8*(3), 282-296. https://doi.org/10.5455/jjcit.71-1652725746

Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H., & Van Der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423 604 UK Biobank participants. *PLOS ONE*, *14*(5), e0213653. https://doi.org/10.1371/journal.pone.0213653

Bernabé-Ortiz, A., & Carrillo-Larco, R. M. (2021). Tasa de incidencia del accidente cerebrovascular en el Perú. *Revista Peruana de Medicina Experimental y Salud Pública*, *38*(3), 399-405. https://dx.doi.org/10.17843/rpmesp.2021.383.7804

Castañeda-Guarderas, A., Beltrán-Ale, G., Casma-Bustamante, R., Ruiz-Grosso, P., & Málaga, G. (2011). Registro de pacientes con accidente cerebro vascular en un hospital público del Perú, 2000-2009. *Revista Peruana de Medicina Experimental y Salud Pública*, *28*(4), 623-627. http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1726-46342011000400008

Centers for Disease Control and Prevention (2022). Stroke signs and symptoms. https://www.cdc.gov/stroke/signs_symptoms.htm.

Chantamit-o-pas, P., & Goyal, M. (2017). Prediction of stroke using deep learning model. In D. Liu, S. Xie, Y. Li, D. Zhao, & E. S. El-Alfy (Eds), *Neural Information Processing* (pp. 774-781). Springer. https://doi.org/10.1007/978-3-319-70139-4_78

Chen, S. D., You, J., Yang, X. M., Gu, H. Q., Huang, X. Y., Liu, H., ... & Wang, Y. J. (2022). Machine learning is an effective method to predict the 90-day prognosis of patients with transient ischemic attack and minor stroke. *BMC Medical Research Methodology*, *22*(1), 1-11. DOI: 10.1186/s12874-022-01672-z

Dinesh, K., Arumugaraj, K., Santhosh, K. D., & Mareeswari, V. (2018). Prediction of cardiovascular disease using machine learning algorithms. *2018 International Conference on Current Trends Towards Converging Technologies (ICCTCT)*. https://doi.org/10.1109/icctct.2018.8550857

Dritsas E, Trigka M. (2022) Stroke risk prediction with machine learning techniques. *Sensors*, *22*(13), 4670. doi: 10.3390/s22134670

Hippisley-Cox, J., Coupland, C., & Brindle, P. (2017). Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ*, 357: j2099. https://doi.org/10.1136/bmj.j2099

Jabal, M. S., Joly, O., Kallmes, D., Harston, G., Rabinstein, A., Huynh, T., & Brinjikji, W. (2022). Interpretable machine learning modeling for ischemic stroke outcome prediction. *Frontiers in Neurology*, *13*, 884693. https://doi.org/10.3389/fneur.2022.884693

Khdair, H., & Dasari, N. M. (2021). Exploring machine learning techniques for coronary heart disease prediction. *International Journal of Advanced Computer Science and Applications*, *12*(5), 28-36. http://dx.doi.org/10.14569/IJACSA.2021.0120505

King, D., Wittenberg, R., Patel, A., Quayyum, Z., Berdunov, V., & Knapp, M. (2020). The future incidence, prevalence and costs of stroke in the UK. *Age and ageing*, *49*(2), 277-282. https://doi.org/10.1093/ageing/afz163

Langhorne, P., Stott, D. J., Robertson, L., MacDonald, J., Jones, L., McAlpine, C., Dick, F., Taylor, G. S., & Murray, G. (2000). Medical complications after stroke: A multicenter study. *Stroke*, *31*(6), 1223-1229. https://doi.org/10.1161/01.str.31.6.1223

Lin, C. H., Hsu, K. C., Johnson, K. R., Fann, Y. C., Tsai, C. H., Sun, Y., Lien, L. M., Chang, W. l., Chen, P. L., Lin, C. L., Hsu, C. Y., & Taiwan Stroke Registry Investigators (2020). Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. *Computer Methods and Programs in Biomedicine*, *190*, 105381. https://doi.org/10.1016/j.cmpb.2020.105381

Linn, L., Eberwine, D., & Oliel, S. (2014, May 15). La OPS/OMS insta a las personas en las Américas a chequear su presión arterial para prevenir infartos y accidentes cerebrovasculares. Organización Panamericana de la Salud. https://www.paho.org/es/enlace/hipertension

Liu, J., Sun, Y., Ma, J., Tu, J., Deng, Y., He, P., Li, R., Hu, F., Huang, H., Zhou, X., & Xu, S. (2021). Analysis of main risk factors causing stroke in Shanxi province based on machine learning models. *Informatics in Medicine Unlocked*, *26*, 100712 https://doi.org/10.1016/j.imu.2021.100712

Ministerio de Salud del Perú (2020). Agenda digital del sector salud 2020-2025. http://bvs.minsa.gob.pe/local/MINSA/5165.pdf

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542-81554. https://doi.org/10.1109/access.2019.2923707

Nikam, A., Bhandari, S., Mhaske, A., & Mantri, S. (2020). Cardiovascular disease prediction using machine learning models. *2020 IEEE Pune Section International Conference (PuneCon),* 22-27. https://doi.org/10.1109/punecon50868.2020.9362367

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., Wong, T. Y., & Cheng, C. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, *122*, 56-69. https://doi.org/10.1016/j.jclinepi.2020.03.002

Qin, Q., Zhou, X., & Jiang, Y. (2021). Prognosis prediction of stroke based on machine learning and explanation model. *International Journal of Computers, Communications & Control*, *16*(2), artículo 4108. https://doi.org/10.15837/ijccc.2021.2.4108

Qu, Y., Zhuo, Y., Lee, J., Huang, X., Yang, Z., Yu, H., Zhang, J., Yuan, W., Wu, J., Owens, D., & Zee, B. (2022). Ischemic and haemorrhagic stroke risk estimation using a machine-learning-based retinal image analysis. *Frontiers in Neurology*, *13*: 916966. https://doi.org/10.3389/fneur.2022.916966

Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. *International Journal of Advanced Computer Science and Applications*, *12*(6), 539-545. https://doi.org/10.14569/ijacsa.2021.0120662

Sarfo, F. S., Ovbiagele, B., Akpa, O., Akpalu, A., Wahab, K., Obiako, R., Komolafe, M., Owolabi, L., Ogbole, G., Calys-Tagoe, B., Fakunle, A., Sanni, T., Mulugeta, G., Abdul, S., Akintunde, A. A., Olowookere, S., Uvere, E. O., Ibinaiye, P., Akinyemi, J., ..., & SIREN. (2022). Risk factor characterization of ischemic stroke subtypes among West Africans. *Stroke*, *53*(1), 134-144. https://doi.org/10.1161/STROKEAHA.120.032072

Shoily, T. I., Islam, T., Jannat, S., Tanna, S. A., Alif, T. M., & Ema, R. R. (2019, July). Detection of stroke disease using machine learning algorithms. *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT*), 1-6. https://doi.org/10.1109/icccnt45670.2019.8944689

Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S., & Khan, M. M. (2021). Stroke disease detection and prediction using robust learning approaches. *Journal of Healthcare Engineering*, *2021*, 1-12. https://doi.org/10.1155/2021/7633381

The GBD 2016 Lifetime Risk of Stroke Collaborators (2018, December 19). Global, regional, and country-specific lifetime risks of stroke, 1990 and 2016. *The New England Journal of Medicine*, *379*(25), 2429-2437. https://doi.org/10.1056/nejmoa1804492

Torres-Aguila, N. P., Carrera, C., Muiño, E., Cullell, N., Cárcel-Márquez, J., Gallego-Fabrega, C., González-Sánchez, J., Bustamante, A., Delgado, P., Ibanez, L., Heitsch, L., Krupinski, J., Montaner, J., Martí-Fàbregas, J., Cruchaga, C., Lee, J-M., Fernández-Cadenas, I., & Acute Endophenotypes Group of the International Stroke Genetics Consortium (ISGC) (2019). Clinical variables and genetic risk factors associated with the acute outcome of ischemic stroke: A systematic review. *Journal of Stroke*, *21*(3), 276-289. https://doi.org/10.5853/jos.2019.01522

World Health Organization (2021, June 11). *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

World Health Organization Regional Office for the Eastern Mediterranean (n.d.). *Stroke, cerebrovascular accident*. https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html

Zhang, L., Niu, M., Zhang, H., Wang, Y., Zhang, H., Mao, Z., Zhang, X., He, M., Wu, T., Wang, Z., & Wang, C. (2022). Nonlaboratory-based risk assessment model for coronary heart disease screening: Model development and validation. *International Journal of Medical Informatics*, *162*, 104746. https://doi.org/10.1016/j.ijmedinf.2022.104746

Zhang, S., Zhang, W., & Zhou, G. (2019). Extended risk factors for stroke prevention. *Journal of the National Medical Association*, *111*(4), 447-456. https://doi.org/10.1016/j.jnma.2019.02.004