AI-BASED APPROACH FOR EARLY DIAGNOSIS SUPPORT IN HEMORRHAGIC STROKE

Athanase SAWADOGO
Computer science department, UFR/SEA, Joseph KI-ZERBO University
athanasesaw@gmail.com

Lydie Simone TAPSOBA
Mathematics and Informatics laboratory (LAMI), Ouagadougou, Burkina Faso
lydie.tapsoba@ujkz.bf

Rodrique KAFANDO
Centre d'Excellence CITADEL, Université Virtuelle of Burkina Faso
rodrique.kafando@citadel.bf

Abdoul Kader KABORE
Centre d'Excellence CITADEL, Université Virtuelle of Burkina Faso
abdoulkader.kabore@citadel.bf

Aminata SABANE
Computer science department, UFR/SEA, Joseph KI-ZERBO University
aminata.sabane@ujkz.bf

Tegawende François d'Assise BISYANDE
Computer science department, UFR/SEA, Joseph KI-ZERBO University
tegawende.bissyande@citadel.bf

ABSTRACT

A hemorrhagic stroke is a life-threatening medical condition that happens when a blood vessel in your brain ruptures and bleeds. It constitutes a burden on health services and the victim's family. The current definitive diagnosis of stroke is based on brain scanning. However, the clinical diagnosis of hemorrhagic stroke is complex and depends on the skills and experience of the practitioner. Human diagnostic errors lead to delays in treatment and thus compromise clinical outcomes. Our vision is to propose an artificial intelligence approach for medical assistance in the early clinical diagnosis of hemorrhagic strokes. We studied and compared three machine learning models, namely logistic regression, Random Forest and artificial neural networks, to choose the best one after setting up a stroke dataset and identifying the most important characteristics. We can conclude that our system designed with artificial intelligence is important with satisfactory results to help health workers make the rapid diagnosis of hemorrhagic stroke and promote rapid treatment of suspected patients

KEYWORDS

hemorrhagic stroke, machine learning, clinical data.

1. INTRODUCTION

Stroke, a leading cause of global mortality, remains a significant public health concern (Incidence & Collaborators, 2018; Luft, Andrea & Katan, 2018). According to World Health Organization (WHO) reports cited by Mendis, Puska, & Norrving (2011), strokes account for 6.2 million deaths annually, with projections indicating sustained high mortality rates through 2030. Ischemic strokes, caused by blockages in cerebral blood vessels, and hemorrhagic strokes, resulting from ruptured arteries, are the two primary types (Types of stroke,

2023). Early diagnosis of hemorrhagic stroke is crucial for prompt medical intervention, as clinical symptoms often overlap with ischemic strokes, complicating accurate diagnosis (Shehab, et al., 2022). The complexity of hemorrhagic stroke diagnosis underscores the potential of artificial intelligence (AI) and data analysis techniques in medical applications (Shin & Lee, 2020; Shatte, Hutchinson, & Teague, 2019). Misdiagnosis can lead to treatment delays and adverse outcomes, emphasizing the need for reliable diagnostic tools. Our research aims to leverage AI to enhance clinical diagnosis by identifying pertinent risk factors and characteristics from stroke registries. This involves developing an AI model tailored to classify hemorrhagic strokes effectively, structured into literature review, methodology, and results sections to achieve our objectives.

2. STAT OF THE ART

The research we studied spans the period from 2018 to 2023. In most studies, the authors compared methods and algorithms to identify the best ones. Several studies focused on predicting pre-hospital outcomes (Zhao, et al., 2021; Qu, et al., 2022; Heo, JoonNyung, et al., 2019; Dr. V. Jyothsna & Dr. M. Rajkumar, 2023; Sailasya & Kumari, 2021; S K Uma & Rakshith S R, 2022; Bandi, Bhattacharyya, & Midhunchakkravarthy, 2020). They used some of machine learning algorithms such as Random Forest (RF), Logistic Regression(LR), Generalized Regression Neural Network (GRNN), Deep Neural Network (DNN), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees (DT), Naïve Bayes(NB), Multi-Layer Perceptron (MLP), Gaussian Naive Bayes (GNB), Bernoulli Naive Bayes (BNB), Radial Basis Function Kernel Support Vector Machine (RBG-SVM), Gradient Boosting Classifier (GB Classifier), Stochastic Gradient Descent (SGD), AdaBoost, AdaBoost with Stochastic Gradient Descent (SGD) to train models to classify and predict stroke cases in general. Their results were interesting, ranging from 80% to 98.54%. Unfortunately, they did not use the same data sizes, characteristics, algorithms, or evaluation metrics in their comparisons.

Classification models have been used to predict post-stroke functional outcomes (Chang et al., 2021; Kim, Choo, & Chang, 2021; Ashrafuzzaman, Saha & Nur, 2022). These authors compared various algorithms such as DT, NB, KNN, Linear Discriminant Analysis, AdaBoost (AB), SVM, LR, RF, and Deep Neural Networks (DNN). They focused on combining and comparing methods for predicting motor function outcomes after stroke (Chang et al., 2021; Kim, Choo, & Chang, 2021). Chang et al. (2021) used a stacking model that combines predictions from several ML models, demonstrating that this method does not improve outcomes. As their studies were based on determining functional outcomes, they used features that provided good results but are not accessible at the pre-hospital level.

Other authors specifically focused on hemorrhagic strokes, but rather on predicting the risk of intracranial hemorrhage (ICH) in patients on hemodialysis (Fengda Li, et al., 2023), as well as evaluating and comparing the performance of machine learning models to predict mortality 90 days post-discharge (Tang, et al., 2022). These last two studies also did not take into account patients in general clinical situations.

Studies show that the Logistic Regression algorithm is used in all stroke classification studies, followed by Random Forest. Random Forest achieves the highest precision in three studies (Zhao, et al., 2021; Dr. V. Jyothsna & Dr. M. Rajkumar, 2023; Bandi, Bhattacharyya, & Midhunchakkravarthy, 2020) with 83%, 95.56%, and 94.32% respectively. Neural networks are less frequently used but demonstrate greater efficiency in their applications compared to other methods (Qu et al., 2022; Tang et al., 2022; Ashrafuzzaman, Saha, & Nur, 2022; Kim, Choo, & Chang, 2021; Sailasya & Kumari, 2021; and S K Uma & Rakshith S R, 2022). From these studies, 13 critical clinical characteristics of hemorrhagic stroke have been identified (Qu et al., 2022; Tang et al., 2022; Li et al., 2023; Ashrafuzzaman, Saha, & Nur, 2022; Kim, Choo, & Chang, 2021; Bandi, Bhattacharyya, & Midhunchakkravarthy, 2020; Sailasya & Kumari, 2021; and Heo et al., 2019). In some studies, resampling methods such as SMOTE, RUS, ADASYN, Borderline, and SMOTEENN were applied to the study data, and the authors concluded that these methods do not improve model results (Tang et al., 2022; Sailasya & Kumari, 2021). Regarding evaluation metrics, ten studies used AUC to compare models (Qu, et al., 2022; Heo, JoonNyung, et al., 2019; Dr. V. Jyothsna & Dr. M. Rajkumar, 2023; Sailasya & Kumari, 2021; Bandi, Bhattacharyya, & Midhunchakkravarthy, 2020; Chang et al., 2021; Kim, Choo, & Chang, 2021; Ashrafuzzaman, Saha & Nur, 2022; Fengda Li, et al., 2023; Tang, et al., 2022). This literature review allows for a comparison of models such as Logistic Regression, Random Forest, and Artificial Neural Networks, guiding the selection of the best approach for an AI-based medical decision support system for the clinical diagnosis of hemorrhagic stroke, which will be presented in the following chapter.

3. METHODOLOGY

3.1. Proposal approach

The approach to our work is a process that extends from setting up the dataset to validating the solution. The main steps of this approach are: establishment of a dataset, exploration and preprocessing of data, selection of independent characteristics, division of preprocessed data into sub-bases, construction of models, evaluation and validation of models. The Figure 1 of the architecture clearly specifies this process.

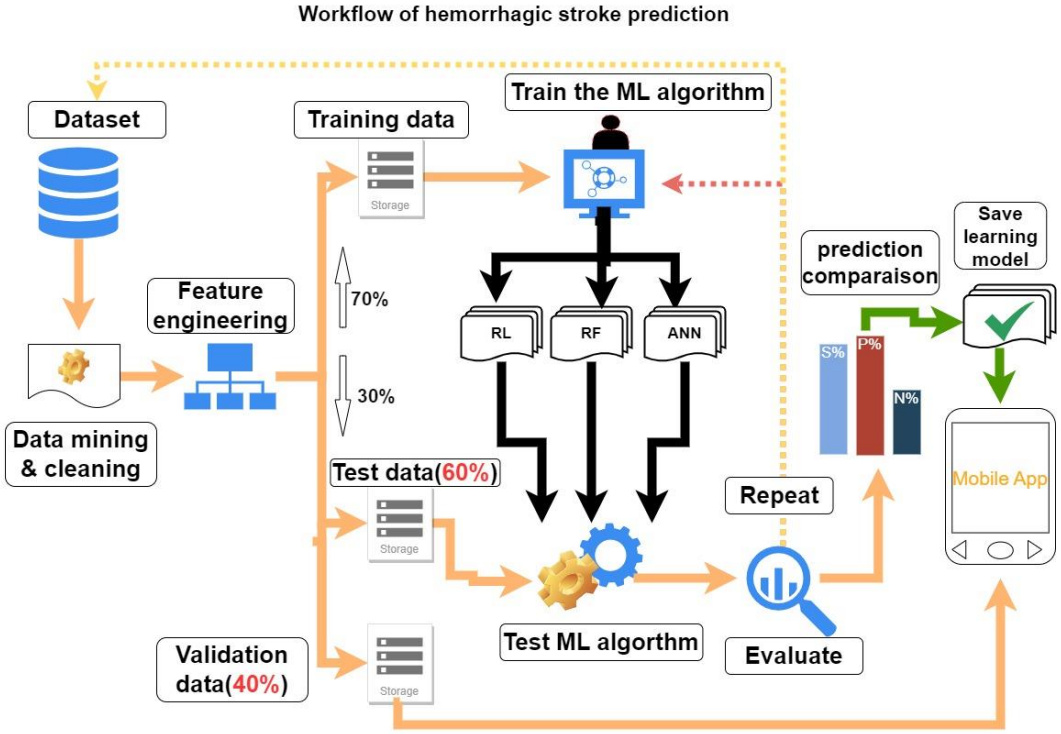

Figure 1. The architecture of the methodology. ML: Machine Learning, RL: Logistic Regression, RF: Random Forest, ANN: Artificial Neural Network

3.2. Implementation

3.2.1 Dataset

We used anonymized secondary data from a hospital in Burkina Faso, consisting of 148 records of patients who experienced stroke confirmed by CT and/or MRI, including only those with hemorrhagic or ischemic strokes. The dataset provided information on clinical symptoms, etiological assessment, treatment, and clinical course. Inclusion criteria focused on relevant clinical information for stroke patients with available radiographic results, while exclusion criteria omitted laboratory tests and scans except for stroke results used as the diagnostic target. Basic analysis and preprocessing involved extracting data from mdb and prj formats into csv and xlsx using Epi Info 7.2.5.0, followed by manual analysis in Excel to select clinical characteristics for AI tools. Initially, the dataset included 148 rows and 129 columns, covering sociodemographic, clinical, paraclinical, etiological, patient care, and pathology evolution data. Five additional columns were added to separate the patient's clinical history into individual characteristics, resulting in 28 selected variables for the study.

3.2.2 Exploratory analysis

We used Python and the Pandas library to visualize the dataset, which contains 28 columns and 148 rows. The columns include age, sex, profession, origin, sickle cell anemia, blood pressure, stroke history, diabetes, kidney failure, chronic brain disease, family medical history, swallowing disorder, vomiting, nausea, Glasgow Coma Scale, consciousness disorder, HIC syndrome, paralysis, paresis, CKD, motor deficiency, aphasia, commitment, alcohol consumption, stress, tobacco consumption, obesity, and stroke status (hemorrhage). Initially, the dataset columns contained strings, booleans, and integers, with no duplicated data. We identified zero values and calculated the number of null values, finding that some columns had missing values, but the present values exceeded the missing ones. Only the 'Age' attribute contained random numerical values. Using skewness-symmetry, we confirmed the data was not skewed. The 'AVCH' variable was the target for classification: 'False' indicated no risk of hemorrhagic stroke (indicating ischemic stroke), and 'True' indicated a positive case for hemorrhagic stroke. The dataset was almost balanced.

3.2.3 Data preprocessing

Data preprocessing is essential to remove noise and outliers, ensuring the model trains efficiently by addressing insignificant variables. The steps taken for data cleaning included balancing the data by aligning the number of observations for each class of the target variable. Initially, there were 78 positive AVCH cases and 70 negative cases; the majority class ("NOT AVCH") was subsampled to match the number of positive cases. Value conversion transformed column values into different types for consistency, and data scaling normalized the numeric variable "Age" to aid in studying variable correlations and preparing for independent variable selection. Correlation analysis determined the relationships between variables and assessed their importance concerning the target variable, aiding in selecting the most informative features. Data augmentation was then used to multiply data without duplicates based on specific criteria, resulting in a simulation dataset. After initial processing, 140 rows of data were obtained, and augmentation increased this to 50,000 samples, creating a large "simulation data" set for training effective models. Data augmentation is a strategy aimed at multiplying data without creating duplicates by using various criteria to conduct simulations. We used the “RandomOverSampler” class from the “imblearn.over_sampling” module of the “Imbalanced-Learn” (imblearn) library to balance our dataset. This class performs random oversampling of the minority classes to achieve a more balanced dataset. After instantiating “RandomOverSampler”, we applied the “fit_resample()” method from the same library. This method takes the imbalanced data as input and returns a balanced sample in terms of classes. However, this augmented data is far from real-world data, potentially leading to poorly performing models when deployed in real situations.

3.2.4 feature engineering

We used three feature selection methods to identify the most important features for training our models and built a new database with these influential features. These methods included the use of the ExtraTreeClassifier algorithm with feature_importances, SelectKBest from scikit-learn's sklearn.feature_selection module, and logistic regression employing the ExhaustiveFeatureSelector (EFS) from the mlxtend.feature_selection module. We summarized the results of these three methods, retaining variables that appeared twice in the results and those ranked in the top three that were not already selected. The final features selected for model training were: "Age", "Stroke", "ATH", "HIC Syndrome", "CKD", "Smoking", "Diabetes" and "Occupation".

3.2.5 Division of data

We considered both the original and augmented data. Each dataset was divided into sub-datasets with 70% for training and 30% for testing and validation, as shown in Table 1.

Table 1. Dividing data based on original data and augmented data

	Total of data	**Training data**	**Test data**	**Validation data**
Pourcentages	100%	70%	60% from 30%	40% from 30%
Number of original data	140	98	29	13
Number of augmentation data	50000	35000	9000	6000

3.2.6 Models building

Based on the literature review, we opted to construct three classification models: logistic regression, Random Forest, and artificial neural networks, aiming to compare their efficiencies subsequently. **Logistic Regression (LR):** Utilizing LogisticRegressionCV from sklearn's model_selection module, which integrates cross-validation for automatic parameter selection. We employed GridSearchCV to explore hyperparameter combinations such as regularization parameter Cs= [2, 4, 5, 6], cross-validation settings cv= [4], penalty ['l1', 'l2'], and solver ['saga', 'liblinear']. Parameters were automatically chosen based on data size and "refit" was set to True for automatic model adjustment after parameter selection. **Random Forest (RF):** Constructed using RandomForestClassifier from sklearn.ensemble, with parameters set as oob_score=True, random_state=42, warm_start=True, and n_jobs=-1. Hyperparameter tuning via GridSearchCV determined the optimal number of estimators (n_estimators=15 with original data, n_estimators=30 with augmented data), considering data size variations and possible values like 15, 30, 40, and 50. **Artificial Neural Networks (ANN):** Developed using Keras from TensorFlow, configured with a sequential model. The network included an input layer (units=50, activation='relu', input_dim=8), three hidden layers (units=100, activation='relu'), and an output layer (unit=1, activation='sigmoid'). The model was compiled with optimizer='adam', loss='binary_crossentropy', and metrics=['accuracy']. Hyperparameters for layers and neurons were manually adjusted to optimize performance, bypassing formal hyperparameter selection techniques.

3.2.7 Evaluation and validation methods

In the first step, we assessed and compared the models' performance using evaluation metrics such as accuracy_score, f1_score, precision_score, recall_score, and roc_auc_score from the Sklearn.metrics module in Python. Additionally, we utilized confusion matrices. In the second step, we employed validation data to evaluate the best-performing model saved from the first step.

3.2.8 Development environment and tools

We developed our project using Windows 10 on a Lenovo PC with 500GO of disk and Google Colab for Python execution. Our toolkit included Epi Info 7 for data conversion, Excel for preprocessing, Draw.io for diagrams, overleaf for writing, PowerPoint for presentations.

4. OUTCOMING

4.1 results obtained

4.1.1 Comparison of the results obtained with the original data and simulation data

We compared the outcomes of the three trained models with both real data and simulated data, presenting the results in Table 2.

Table 2. Comparison of models built on the basis of original data et simulation data

Data Category	Models	Accuracy	F1-score	Precision	Recall	AUC	Confusion Matrix
Original data (140)	**Random Forest**	**73.07%**	**69.56%**	**66.66%**	**72.72%**	**73.03%**	**73.07%**
	ANN Model	50%	51.85%	43.75%	63.63%	51.81%	50%
	Logistic Regression	46.15%	22.22%	28.57%	18.18%	42.42%	53.84%
Simulation data (50000)	**Random Forest**	**99.33%**	**99.33%**	**98.32%**	**100%**	**99.15%**	**99.33%**
	ANN Model	98.61%	98.58%	100%	97.21%	98.60%	98.61%
	Logistic Regression	68.66%	66.61%	68.27%	68.95%	68.66%	99.15%

4.1.2 Saving and validating best model

When validating the saved model, the Random Forest model achieved a prediction accuracy of 99.05% using the validation data.

4.2. Discussion

4.2.1 Interpretation of results

We found that the Random Forest model outperforms other models with an accuracy of 99.33% on test results, compared to 98.61% for ANN and 68.66% for logistic regression. Due to its superior performance, we chose to save the Random Forest model and interpret its results. Considering accuracy, we notice that the correct predictions among all the predictions made were 99.33%. This shows that the model is effective in classifying hemorrhagic strokes using clinical data. This performance is crucial in the context of diagnosing hemorrhagic stroke, as it minimizes the risk of misclassifying healthy individuals as sick. Moreover, the model achieved a validation accuracy of 99.05% on new data, demonstrating its robustness and reliability. About the model that is strong classification performance, it can be attributed to several factors: algorithmically, the Random Forest model excels in mitigating overfitting by aggregating predictions from multiple decision trees, enhancing generalization to new data. Its capability to handle large and complex datasets, along with robustness in managing nonlinear characteristics, further improves its effectiveness. Automatic hyperparameter selection tailored to the dataset size optimized model performance. Additionally, the high data quality, ensured through rigorous preprocessing, and the positive impact of feature selection techniques collectively contributed to the model's success. In conclusion, we are pleased with the results, which align well with our expectations.

4.2.2 Put into perspective with the state of the art

As we did not use the same dataset or data size as those in the state-of-the-art studies, direct comparison of our results is not feasible. However, Table 3 , we attempt to compare the numerical outcomes of our Random Forest models trained on the original reduced dataset and simulated data against those from the state-of-the-art studies we deemed superior.

Table 3 Put into perspective with the results of the state of the art

	High-performance model	**Data size**	**Accuracy**
(Zhao, et al., 2021)	Random Forest	4914	83%
(Dr. V. Jyothsna & Dr.M. Rajkumar, 2023)	Random Forest	-	95%
(Bandi, Bhattacharyya, & Midhunchakkravarthy, 2020)	Random Forest	4799	94%
Our study	Random Forest	140 for original data	73.07%
		50000 for simulation data	99.33%

The table illustrates that Random Forest accuracy rates reported in the state-of-the-art studies range from 83% to 95%. Despite having limited original data, we achieved 73.07% accuracy, which improved to 99.33% with simulation data. This suggests that these accuracy rates can improve significantly with sufficient data.

4.2.3 The difficulty encountered and the research limit

Our main challenge lies in the inadequate quantity of data available for model training. Despite initiating a data request procedure during the internship, delays in response and ethical considerations related to data usage hindered its success. Due to insufficient data, the mobile application we developed is not suitable for real-world deployment, as the integrated model was trained using simulation data.

5. CONCLUSION

Our research focused on developing an AI-based approach to aid in the early diagnosis of hemorrhagic strokes, recognizing their complexity and distinct clinical characteristics. We identified key stroke features using AI techniques and explored advanced algorithms suitable for diagnosis. Our work resulted in a high-performance Random Forest model achieving 99.33% accuracy, albeit trained on simulated data not directly applicable in real scenarios. Future steps involve adapting the model with real data to enhance its practical utility as a supportive tool for healthcare professionals, facilitating faster diagnosis and decision-making without replacing medical expertise.

ACKNOWLEDGEMENT

We extend our gratitude to Dr. R. Aristide YAMEOGO, cardiologist and faculty member at Joseph KI-ZERBO University, and the Director of the Interdisciplinary Center of Excellence, for generously providing us with anonymous data for our research.

REFERENCES

Ashrafuzzaman, M., Saha, S., & Nur, K. (2022). Prediction of Stroke Disease Using Deep CNN Based Approach. *Journal of Advances in Information Technology Vol*, *13*(6). From http://www.jait.us/issues/JAIT-V13N6-604.pdf

Bandi, V., Bhattacharyya, D., & Midhunchakkravarthy, D. (2020). Prediction of Brain Stroke Severity Using Machine Learning. *Revue d'Intelligence Artificielle*, *34*(6). From https://scholar.archive.org/work/e3ff3d4fwbdbtfwb2ntynco6pq/access/wayback_/http://www.iieta.org/download/file/fid/48077

Chang, S.-C., Chu, C.-L., Chen, C.-K., Chang, H.-N., Wong, A., Chen, Y.-P., & Pei, Y.-C. (2021). The comparison and interpretation of machine-learning models in post-stroke functional outcome prediction. *Diagnostics*, *11*(10), 1784. From https://www.mdpi.com/2075-4418/11/10/1784/pdf

Dr. V. Jyothsna, & Dr.M. Rajkumar. (2023). Application Of Supervised Machine Learning Techniques For Classification Of Brain Stroke. *Journal of Clinical Otorhinolaryngology, Head, and Neck Surgery, Associate Dean (Academic A fairs) and Associate Professor Department of CSIT, School of Computing, MOHAN BABU UNIVERSITYl vol*, *27*(1), 1001-1781. From https://www.lcebyhkzz.cn/article/view/2023/377.pdf

Heo, JoonNyung, Yoon, Jihoon G, Park, H., Kim, Y., Nam, H., & Heo, J. (2019). Machine learning--based model for prediction of outcomes in acute stroke. *Stroke*, *50*(5), 1263-1265. From https://www.ahajournals.org/doi/full/10.1161/STROKEAHA.118.024293

Incidence, G. 2., & Collaborators, P. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet (London, England)*(392(10159)), 1789–1858. doi:10.1016/S0140-6736(18)32279-7

Kim, J., Choo, Y., & Chang, M. (2021). Prediction of motor function in stroke patients using machine learning algorithm: Development of practical models. *Journal of Stroke and Cerebrovascular Diseases*, *30*(8), 105856. From https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.105856

Types of stroke. (2023, 11 02). From coeuretavc: https://www.coeuretavc.ca/avc/questce-quun-avc/les-types-d-avc

Li, F., Chen, A., Li, Z., Gu, L., Pan, Q., Wang, P., . . . Feng, J. (2023). Machine learning-based prediction of cerebral hemorrhage in patients with hemodialysis: A multicenter, retrospective study. *Frontiers in Neurology*, *14*, 1139096. From https://www.frontiersin.org/articles/10.3389/fneur.2023.1139096/full

Luft, Andrea, & Katan, M. (2018, 05 23). Global Burden of Stroke. *Semin Neurol*, *38*, 208-211. doi:10.1055/s-0038-1649503

Mendis, S., Puska, P., & Norrving, B. (2011, 01 01). Global atlas on cardiovascular disease prevention and control. *World Heart Federation and World Stroke Organization*. From https://www.researchgate.net/publication/311885270_Global_atlas_on_cardiovascular_disease_prevention_and_control_WHO

Qu, S., Zhou, M., Jiao, S., Zhang, Z., Xue, K., Long, J., . . . others. (2022). Optimizing acute stroke outcome prediction models: Comparison of generalized regression neural networks and logistic regressions. *Plos one*, *17*(5), e0267747. From https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0267747

S K Uma, & Rakshith S R. (2022). Stroke Analysis Using 10 ML Comparison. *International Journal for Research in Applied Science & Engineering Technology (IJRASET) Vol, 10*(10), 2321-9653. From https://doi.org/10.22214/ijraset.2022.45895

Sailasya, G., & Kumari, G. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. *International Journal of Advanced Computer Science and Applications, 12*(6). From https://pdfs.semanticscholar.org/df5c/7d1bd7a59009dc51b9db903aa7f144241879.pd

Shatte, A., Hutchinson, D., & Teague, S. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological medicine, 49*(9), 1426-1448. From https://doi.org/10.1017/S0033291719000151

Shehab, Mohammad and Abualigah, Laith and Shambour, Qusai and Abu-Hashem, Muhannad A and Shambour, Mohd Khaled Yousef and Alsalibi, . . . Amir H. (2022). Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine, 145*, 105458. From https://doi.org/10.1016/j.compbiomed.2022.105458

Shin, Y., & Lee, I. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons, 63*(2), 157-170.

Tang, J., Wang, X., Wan, H., Lin, C., Shao, Z., Chang, Y., . . . Du, Y. (2022). Joint modeling strategy for using electronic medical records data to build machine learning models: an example of intracerebral hemorrhage. *BMC Medical Informatics and Decision Making, 22*(1), 278. From https://link.springer.com/article/10.1186/s12911-022-02018-x

Zhao, Y., Fu, S., Bielinski, S., Decker, P., Chamberlain, A., Roger, V., . . . Larson, N. (2021). Natural language processing and machine learning for identifying incident stroke from electronic health records: algorithm development and validation. *Journal of medical Internet research, 23*(3), e22951. From https://www.jmir.org/2021/3/e22951/