Journal of Technology and Informatics (JoTI)
Vol. 4, No. 2, April 2023
P-ISSN 2721-4842
E-ISSN 2686-6102
DOI: 10.37802/joti.v4i2.278

Optimizing Accuracy of Stroke Prediction Using Logistic Regression

Mohammed Guhdar Mohammed[1], Amera Ismail melhum[2], Alaa Luqman Ibrahim[3]
[1]Department of Computer Science, Faculty of Science, University of Zakho, Zakho, Kurdistan region, Iraq,
[2]Department of Computer Science, Faculty of Science, University of Duhok, Duhok, Kurdistan region, Iraq,
[3]Department of Mathematics, Faculty of Science, University of Zakho, Zakho, Kurdistan region, Iraq,
e-mail: mohammed.guhdar@uoz.edu.krd[1], amera_melhum@uod.ac[2], alaa.ibrahim@uoz.edu.krd[3]
* Correspondence author: E-mail: mohammed.guhdar@uoz.edu.krd

Abstract: *An unexpected limitation of blood supply to the brain and heart causes the majority of strokes. Stroke severity can be reduced by being aware of the many stroke warning signs in advance. A stroke may result if the flow of blood to a portion of the brain stops suddenly. In this research, we present a strategy for predicting the early start of stroke disease by using Logistic Regression (LR) algorithms. To improve the performance of the model, preprocessing techniques including SMOTE, feature selection and outlier handling were applied to the dataset. This method helped in achieving a balance of class distribution, identifying and removing unimportant features and handling outliers. with the existence of increased blood pressure, body mass, heart conditions, average blood glucose levels, smoking status, prior stroke, and age. Impairment occurs as the brain's neurons gradually die, depending on which area of the brain is affected by the reduced blood supply. Early diagnosis of symptoms can be extremely helpful in predicting stroke and supporting a healthy lifestyle. Furthermore, we performed an experiment using logistic regression (LR) and compared it to a number of other studies that used the same machine learning model, which is logistic regression (LR), and the same dataset. The results showed that our method successfully achieved the highest F1 score and area under curve (AUC) score, which can be a successful tool for stroke disease prediction with an accuracy of 86% compared to the other five studies in the same field. The predictive model for stroke has prospective applications, and as a result, it is still significant for academics and practitioners in the fields of medicine and health sciences.*

Keywords: *Data Analysis Informatics, Logistic Regression (LR), Stroke Machine Learning, Stroke Prediction*

INTRODUCTION

Thirteen million individuals get strokes annually, according to the World Stroke Organization. However, approximately 5.5 million patients die as an outcome. Stroke is the leading cause of disability and death worldwide, making its imprint critical in all aspects of life. Stroke impacts the client’s workplace, family, and social environment. In addition, centrally to the typical concept, a stroke can happen to any individual and at any age, regardless of physical condition or gender [1]. Stroke refers to an acute neural illness of the brain's blood vessels, which happens when the bloodstream to a sector of the brain halts. Therefore, the brain compartments get destitute of appropriate oxygen. Stroke comprises two sorts, which include hemorrhagic and ischemic. Additionally, strokes can be minor to very severe, with temporary or permanent damage. Hemorrhage involves the rupture of the blood vessels, which is rare and contributes to brain bleeding. The most frequent strokes, however, are ischemic strokes, in which an artery blockage or spasm stops blood flow to a particular region of the brain. Different aspects augment the probability of having a stroke. The factors include the presence of myocardial infarction, a transient stroke, a comparable stroke in the past, and other heart illnesses such as atrial fibrillation and heart failure. Other aspects that surge the chances of stroke include blood clotting disorders [2], alcohol consumption [3], [4], sedentary lifestyle, obesity, diabetes, high blood cholesterol, smoking, carotid stenosis from atherosclerosis, hypertension, and euphoric substances such as amphetamines and cocaine. In addition, stroke advances swiftly, which makes the symptoms different. Therefore, the symptoms can develop quickly or slowly. It is probable for an individual to wake up while sleeping with the signs. The core symptoms include immobility of the legs or arms [5]. In addition, the signs involve a drop in the mouth's angle, vomiting, headaches, decreased vision, dizziness, difficulty walking, challenges in speaking, and numbness on the face, legs, or arms. Ultimately, in acute strokes, the patient falls into a comma and loses consciousness. When a stroke is detected in an individual, a computerized tomography scan instantly offers an analysis. In an instance of ischemic stroke, the effective approach is magnetic resonance imaging (MRI) [6]. However, other diagnostic methods are used, including carotid ternary. In many instances, the initial twenty-four hours are significant. This is because the diagnostic highlights the treatment approach, typically pharmaceutical, although there are a few circumstances where the surgical approach gets incorporated. However, when the patient becomes unconscious, the ventilation system and induction in the ICU are necessary. Some clients heal after the disease, while others have challenges relying on the sternness of the stroke, such as difficulty swallowing food, inability to walk, emotional issues such as depression, difficulty comprehending speech and speaking, attention, concentration, and memory [7]. Recovery assists in regaining lost operation after a stroke. The necessary strategy gets established to enable the patient to instantly return socially and psychologically through the contributions of neurologists, speech therapy, and kinesis therapy [8]. To decrease the likelihood of a stroke, following a healthy diet without salt and fat is appropriate; quitting smoking, maintaining a normal weight, exercising regularly, and monitoring blood pressure is also appropriate. In addition, information and communication technologies play a significant role in the early forecast of illnesses such as hepatitis, sleep disorders, cholesterol, hypertension, and diabetes. In particular, a stroke will be a concern in the context of this scrutiny [8]. Much research scrutiny has been performed with the assistance of machine learning approaches. Early diagnosis of symptoms can be extremely helpful in predicting stroke and supporting a healthy lifestyle. This work uses machine learning (ML) to develop and evaluate a number of models in order to lay a great foundation for the long-term diagnosis and prediction of stroke incidence. Using machine learning and artificial intelligence (AI), in particular, is already having a substantial impact on the early diagnosis of some illnesses, including diabetes, hypertension, cholesterol, COVID-19, sleep disorders, hepatitis C, and others. We shall be especially concerned with strokes in the context of this research. Machine learning models such as Support Vector Machine (TSVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Bayesian Classifier (BC), and Random Forest (RF) have been used in numerous research studies [7], [9]–[13] for this particular condition.

RELATED WORK

The study community has demonstrated a substantial desire to establish methods and tools for predicting and monitoring different diseases that substantially affect individuals' health. This segment will provide modern work via machine learning to make predictions of stroke risk. The researchers tested different machine learning models such as random forest, K-nearest neighbor, logistic regression, and naïve Bayes to notice a stroke accurately [14]. In [15], the authors achieved a maximum reminiscence of 0.825% accuracy by performing a logistic regression algorithm for the stroke prediction task. This was used to analyze the obtained risk level of the strokes, such as high risk, moderate risk, and low risk, based on the primary attributes. As it is suggested by the authors, this model would perform better when compared to the existing models with the same method. This research is highly concentrated on improving random forest models. The author of [16] achieved 0.78% using logistic regression on the kaggle stroke database, and the authors of [17], [18] used the same model and the same kaggle database and achieved accuracy scores of 0.77% and 0.71%, respectively. A methodology was suggested to recognize different symptoms linked with stroke disease and the preventive gauges from social media resources. Logistic regression was applied to categorize stroke peril levels. The test outcomes of [19] with logistic regression demonstrated that the enhancing approach accomplished 0.79% Kaggle dataset was also used in the research [20]. The scrutiny work proposes the adoption of different machine learning algorithms. The algorithms included logistic regression, K-nearest neighbor, naïve Bayes, random forest, decision tree, and support vector machine. The naïve Bayes, like other machine learning algorithms, accomplished better correctness, with 0.82% for the estimate of stroke. In addition, the researchers targeted obtaining a stroke dataset from a health facility and categorizing the stroke through machine learning and mining algorithms. To show the influence of the risk elements on stroke projection, [21] also evaluated the client's computerized health records. The sorting exactness of the random forest, decision tree, and neural network over a thousand tests of the dataset of electronic health records were 0.7453%, 0.7431%, and 0.7502%, correspondingly. Moreover, the aptitude of machine learning models to evaluate fluid-attenuated inversion recovery and diffusion-weighted imaging of clients with stroke within twenty-four hours of stroke onset was assessed using automatic image processing techniques. Three machine learning approaches have been established to project the stroke onset for binary sorting, including support vector machine and random forest. The machine learning approaches were based on the specificity and sensitivity for recognizing clients in less than 4.5 hours and were related to individual readings of fluid-attenuated inversion recovery and diffusion-weighted imaging mismatch [22].

The main contribution of this study was to use logistic regression (LR) to implement the most accurate medical prediction system for the diagnosis of heart disease among many other studies that used the same dataset source and logistic regression by applying different preprocessing techniques such as SMOTE with different hyperparameters and correlation parameters to exclude less important features from the dataset.

METHOD

The research employed a machine learning technique called logistic regression to predict strokes by using data from the Kaggle competition, which included 5110 participants. After downloading the dataset, the next step is to prepare the dataset to handle missing values, data scaling, performing label encoding, and balancing data. These steps are called "data preprocessing." The model is built based on a machine learning algorithm. The research is highly dependent on the logistic regression (LR) algorithm. The random state parameter is set to 15, which results in the best performance and one of the difference points from previous researches, in addition, based on the correlation factor, the gender column was removed from the dataset for the training and testing stages, increasing the model's accuracy, and oversampling techniques were used to balance the data for the training stage. Here we chose to use SMOTE (synthetic minority oversampling technique) with a k-neighbor value of 50, which had a huge effect on testing and training, this was the key difference from previous research. After the model is built, five accuracy metrics are tested to evaluate the performance. The accuracy, precision, recall, F1 score, and area under curve (AUC) metrics are derived in this experiment to assess the effectiveness of the LR classifier. The 5110 datasets were divided into three groups: 70% for training, 30% for testing and validation. The data was split into two sections, one for learning or training a model and the other for model validation, and cross-validation was used to assess and contrast the results.

Dataset Description

The scrutiny (critical observation) was grounded on a Kaggle dataset [23] of participants above eighteen years old. They were 5110 and attributed to machine learning approaches as described below.

a. Age is the aspect where the participants were more than eighteen years old.
b. Gender refers to the sex of the participants where the number of women was 2994 and men was 2115.
c. The hypertension aspect referred to whether the participants had high blood pressure instances. The hypertensive participants were 9.75%.
d. The heart disease feature got linked to whether the participants had heart disease issues. However, the percentage of participants who had heart issues was 4.03%.
e. Ever married aspect represented the marital status of the participants, where 65.62% were married.
f. Work type indicated the four different types of employment status: never worked (0.43%), government job (12.86%), self-employed (16.03%), and privately employed (57.24%).
g. The residence type involved the living status of the participants, which was of two sorts, including 49.20% rural and 50.80% urban.
h. The stroke aspect represented if the participants previously had a stroke. The records showed that 5.53% of the participants had suffered from a stroke.
i. The smoking status aspect got categorized into three and captured the participants' smoking status. The categories included; formerly smoked at 17.32%, never smoked at 37.03%, and smoked at 15.44%.
j. BMI was measured in Kg/m^2 [23].
k. Age glucose level captured the average glucose level of the participants.

Analysis of long-term stroke risk

The preliminary dataset was detached into a test and training set, to evaluate the long-term peril of stroke. The risk aspects linked with stroke involve factors with which the machine learning approaches are filled to project the category of the new instance. Some factors leading to stroke include cardiac structural abnormalities, abnormal heart rhythm, illegal drugs, high blood lipids, and cholesterol. For instance, high cholesterol levels lead to hardening or thickening of the arteries due to plaque buildup. The Plaque is a build-up of substances from a fluid, such as cholesterol in the blood vessels which consists of calcium, cholesterol, and fatty substances. In addition, damaged heart valves damage the heart in the long run, which later increases the peril of stroke. Furthermore, an increase in the number of red blood cells causes clots to form faster and thickens the blood. Thus, it increases the peril of a stroke. The evaluation targets are to design machine learning approaches that reach the area under curve and sensitivity, enabling the appropriate projection of stroke occurrences [8].

Logistic regression (LR)

Logistic regression (LR) is a different technique that will be included in the suggested framework. It is a statistics binary classifier that was first created for binary issues and then expanded to deal with multi-class problems [12]. The output of the model is a binary variable, which indicates the likelihood that a certain incident would fall under the "Stroke" class and 1 captures the likelihood that an occurrence will belong to the "Non-Stroke" class of sensors. Following is the linear connection between model parameters and log-odds with base b [24]:

$$log_b \frac{p}{1-p} = \beta_0 + \beta_i f_{i1} + \cdots + \beta_n f \quad (1)$$

where (P) Probability of Success and (1-P) is Probability of Failure, while Estimating the values of betas involves the probability concept, odds and log odds.

The use of logistic regression simplifies the mathematics for determining how many factors or status, such as age, gender, height and the placement of advertisements, affect a particular outcome. The generated models can be used to evaluate the relative efficacy of various interventions for various classes, such as young versus old or male against female.

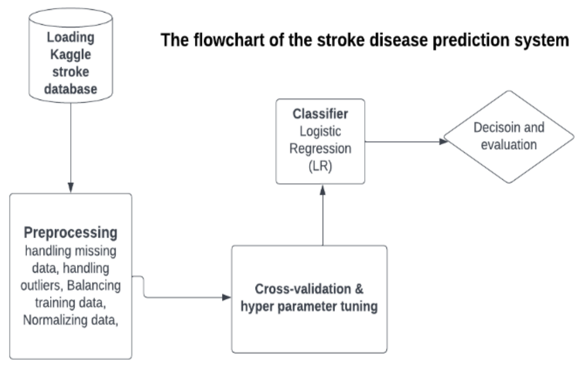
The flowchart of the stroke disease prediction system
Loading Kaggle stroke database
Preprocessing handling missing data, handling outliers, Balancing training data, Normalizing data,
Cross-validation & hyper parameter tuning
Classifier Logistic Regression (LR)
Decisoin and evaluation

Figure 1. Framework of Proposed System

Preprocessing

After inspecting dataset there was 201 null values for BMI column The null values are filled by using the data column's mean. Then ID column is excluded from the dataset as it has no impact on data manipulation. Next correlation is performed on the whole dataset as shown in Figure 2 below.

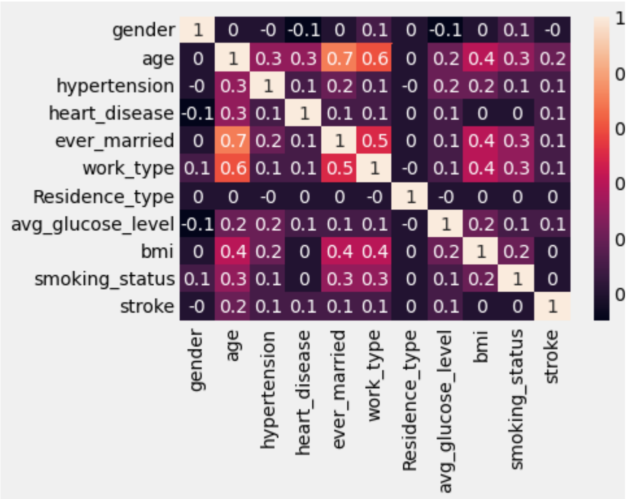

Figure 2. Finding Correlation of Dataset Columns

Figure 2 indicates that ever married, avg_glucose_level, heart disease, hypertension, age, and BMI are correlated positively with the target feature which is stroke. However, gender is negatively correlated with stroke. As a part of per-processing, from the above available dataset, the gender attribute has been excluded from this research.

Note: in Figure 2. 0 value means the correlation is below 0.0X but still there is a correlation. Unlike -0 which is less or equal to actual 0 value.

Checking outliers in dataset

By checking the abnormal distance of some data values from other values, it is obvious that there are such data points that need to be handled, and to handle this issue we used graph and computed the nth percentile of the given data (array elements) along the specified axis to exploit outliers then statistically solved by iterating on all outlier data. Figure 3 exploits the outliers in the dataset, while Figure 4 shows the same dataset after handling the problem through statistical methods.

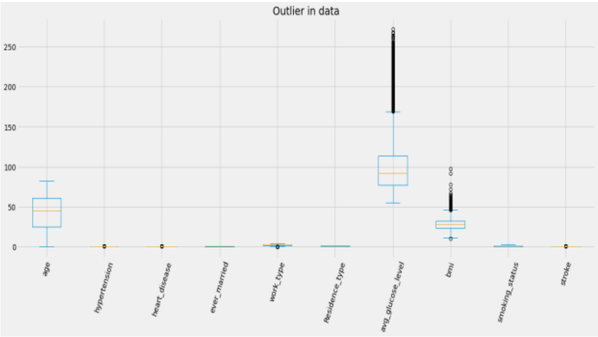

Figure 3. Detecting Outliers in Dataset

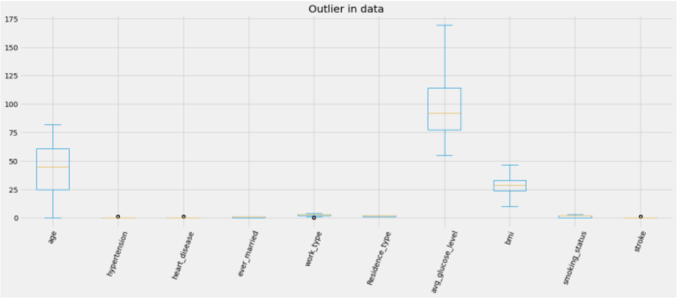

Figure 4. Handling Outliers in Dataset

Checking Balance in Dataset

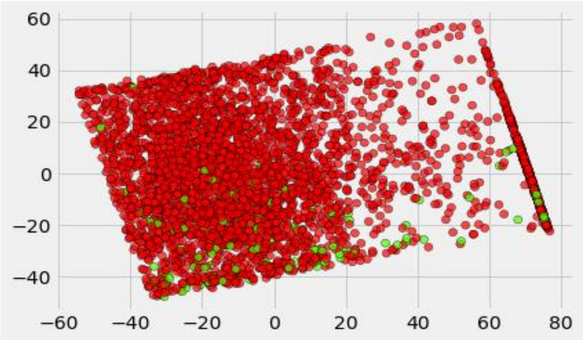

Figure 5. Dataset Imbalance Form (Before Oversampling)

Figure 5 indicates the number of samples from class 0 and class 1. It is obvious that the training data is imbalanced. To solve this problem, we have applied some oversampling techniques such as the Synthetic Minority Oversampling Technique (SMOTE) which is an algorithm that augments data by generating synthetic data points depending on the original ones [25]. The k_neighbor parameter was set to 50, which resulted in the best performance. Figure 6 shows the same dataset after applying the SMOTE algorithm.

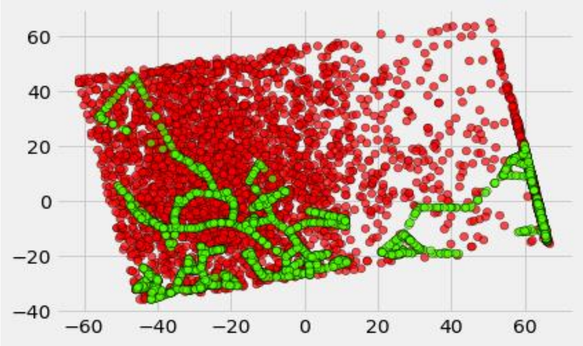

Figure 6. Dataset Balanced Form (After Applying Oversampling)

RESULT AND DISCUSSION

As show in Table 1., The accuracy of the logistic regression model in the test data was discovered to be 0.86 percent correct. Precision was 0.875 percent, and recall, which showed that the LR model correctly identified the percentage of actual positive stroke cases, was 0.865 percent. The F1 score and AUC of our model are 0.87 percent and 0.93 percent, respectively. The results show that the classifier can correctly predict occurrences based on the patterns used in the training activity as well as in Table 1. Our method achieve the best result compared to all other researchers who used the same database and LR method. The model can therefore be used to precisely predict potential strokes.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (2)$$

Where; TP=True Value
FP – False Positive
TN – True Negative
FN – False Negative

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

$$F1Score = \frac{2xrecallxprecision}{recall+precision} (5)$$

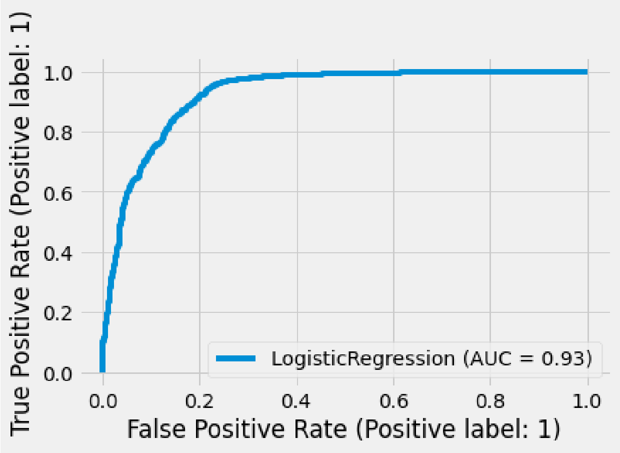

Figure 7. Area Under the Curve (AUC) Performance Metric for the Proposed System

Table 1. Result Comparison

ML Models	Precision	Recall	F-measure	AUC	Accuracy	References
LR	0.78%	0.78%	0.78%	0.78%	0.78%	[16]
LR	0.78%	0.78%	0.78%	Non	0.77%	[17]
LR	0.7%	0.76%	0.73%	Non	0.71%	[18]
LR	0.79%	0.79%	0.79%	87.7%	0.79%	[19]
LR	Non.	Non.	Non.	Non	0.83%	[15]
LR	**0.88%**	**0.87%**	**0.87%**	**0.93%**	**0.86%**	**Ours**

In all criteria taken into account, our LR based model underneath the chosen base models was the most effective. Focusing on the AUC metric demonstrates that our model achieved the best results among all other LR based models, demonstrating that the model can successfully differentiate stroke from non-stroke cases with a high likelihood of 0.93 percent. The F-measure is a suitable ratio that can reveal the effectiveness (i.e., accuracy) of the machine learning techniques on the data. The F-measure shows that our model outperforms all other LR based models with the same dataset with a score of 0.87 percent. This study's use of a publicly accessible dataset poses a restriction. Unlike data from a hospital or institute, this data has a specific size and set of characteristics. Although the latter could provide more thorough health profiles of the participants and richer information, accessing such data is typically tedious and challenging due to privacy concerns.

CONCLUSION AND SUGGESTION

A stroke threatens an individual's life and should be treated or prevented to dodge unprotected complications. Currently, with the swift evolution of machine learning, decision-makers, medical experts, and clinical providers can execute the developed approaches to discover relevant aspects of stroke instances and evaluate the respective risk or probability [26]. Therefore, machine learning can assist in mitigating the severe consequences and early stroke prediction. This scrutiny explores the impact of different machine learning algorithms to recognize the most appropriate feature for projecting stroke grounded on various aspects that seizures the participant's contours. The healthcare sector generates enormous amounts of complicated data using data mining techniques about patients, healthcare resources, diagnosis of diseases, electronic records of patients, hospital instruments, etc. [22]. For the diagnosis of stroke patients, many research used logistic regression, Results indicate that, in comparison to the other five LR based model, our model can be a useful tool for predicting the occurrence of stroke with the accuracy of 0.86 percent, an AUC of 0.93 percent and an F-measure of 0.87 percent, the model can distinguish itself from other models and outperforms the other methods.

REFERENCES

[1] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, “Stroke Disease Detection and Prediction Using Robust Learning Approaches,” *Journal of Healthcare Engineering*, vol. 2021, 2021.

[2] V. Tutwiler, A. D. Peshkova, I. A. Andrianova, D. R. Khasanova, J. W. Weisel, and R. I. Litvinov, “Contraction of blood clots is impaired in acute ischemic stroke,” *Arteriosclerosis, thrombosis, and vascular biology*, vol. 37, no. 2, pp. 271–279, 2017.

[3] P. B. Gorelick, “Alcohol and stroke.,” *Stroke*, vol. 18, no. 1, pp. 268–271, 1987.

[4] K. Reynolds, B. Lewis, J. D. L. Nolen, G. L. Kinney, B. Sathya, and J. He, “Alcohol consumption and risk of stroke: a meta-analysis,” *Jama*, vol. 289, no. 5, pp. 579–588, 2003.

[5] J. Alberto and T. Rodríguez, “Stroke prediction through Data Science and Machine Learning Algorithms,” *no. Ml*, 2021.

[6] M. Fatahi and O. Speck, “Magnetic resonance imaging (MRI): A review of genetic damage investigations,” *Mutation Research/Reviews in Mutation Research*, vol. 764, pp. 51–63, 2015.

[7] C. Sharma, S. Sharma, M. Kumar, and A. Sodhi, “Early Stroke Prediction Using Machine Learning,” in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, 2022, pp. 890–894.

[8] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, “Prediction of cardiovascular disease using machine learning algorithms,” in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, 2018, pp. 1–7.

[9] J. F. Medina-Mendieta, M. Cortés-Cortés, and M. Cortés-Iglesias, “COVID-19 forecasts for Cuba using logistic regression and gompertz curves,” *MEDICC review*, vol. 22, pp. 32–39, 2022.

[10] R. Xiao, X. Cui, H. Qiao, X. Zheng, and Y. Zhang, “Early diagnosis model of Alzheimer’s Disease based on sparse logistic regression,” *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 3969–3980, 2021.

[11] P. Johnson *et al.*, “Genetic algorithm with logistic regression for prediction of progression to Alzheimer’s disease,” *BMC bioinformatics*, vol. 15, no. 16, pp. 1–14, 2014.

[12] S. Nusinovici *et al.*, “Logistic regression was as good as machine learning for predicting major chronic diseases,” *Journal of clinical epidemiology*, vol. 122, pp. 56–69, 2020.

[13] A. S. Abdalrada, O. H. Yahya, A. H. M. Alaidi, N. A. Hussein, H. T. Alrikabi, and T. A.-Q. Al-Quraishi, “A predictive model for liver disease progression based on logistic regression algorithm,” *Periodicals of Engineering and Natural Sciences (PEN)*, vol. 7, no. 3, pp. 1255–1264, 2019.

[14] W.-W. Chang, S.-Z. Fei, N. Pan, Y.-S. Yao, and Y.-L. Jin, “Incident Stroke and Its Influencing Factors in Patients With Type 2 Diabetes Mellitus and/or Hypertension: A Prospective Cohort Study,” *Frontiers in Cardiovascular Medicine*, vol. 9, 2022.

[15] V. Bandi, D. Bhattacharyya, and D. Midhunchakkravarthy, "Prediction of Brain Stroke Severity Using Machine Learning.," *Rev. d'Intelligence Artif.*, vol. 34, no. 6, pp. 753–761, 2020.

[16] G. Sailasya and G. L. A. Kumari, "Analyzing the performance of stroke prediction using ML classification algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021.

[17] A. A. Ali, "Stroke prediction using distributed machine learning based on Apache spark," *Stroke*, vol. 28, no. 15, pp. 89–97, 2019.

[18] M. S. Azam, M. Habibullah, and H. K. Rana, "Performance Analysis of Various Machine Learning Approaches in Stroke Prediction," *International Journal of Computer Applications*, vol. 175, no. 21, pp. 11–15, 2020.

[19] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," *Sensors*, vol. 22, no. 13, p. 4670, 2022.

[20] Z. Liu, C. Yang, X. Wang, and Y. Xiang, "Blood-based biomarkers: a forgotten friend of hyperacute ischemic stroke," *Frontiers in neurology*, p. 797, 2021.

[21] C. S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli, and D. John, "Predicting stroke from electronic health records," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 5704–5707.

[22] H. Lee *et al.*, "Machine learning approach to identify stroke within 4.5 hours," *Stroke*, vol. 51, no. 3, pp. 860–866, 2020.

[23] Kaggle, "Stroke Prediction Dataset," *Kaggle*. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (accessed Sep. 11, 2022).

[24] R. E. Wright, "Logistic regression.," 1995.

[25] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[26] M. J. O'Donnell *et al.*, "Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study," *The lancet*, vol. 388, no. 10046, pp. 761–775, 2016.