Ahliya Journal of Allied Medico-Technology Science
[Published by Palestine Ahliya University]

Vol. 01 Issue 01 (2024)
ISSN: 3007-9489

AI-Based Cerebral Vascular Accident (CVA) Analysis and Prediction

Hala Shaheen[1], Mutaz Rasmi Abu Sara[1], Khaled Sabarna[2], Labib Arafeh[1]
[1]IT Department, Faculty of Engineering and Information Technology, Palestine Ahliya University (Palestine)
✉ Halashahin98@gmail.com
✉ moutaz.a@paluniv.edu.ps
✉ l.arafeh@paluniv.edu.ps
[2]Nursing Department, Faculty of Allied Medical Sciences, Palestine Ahliya University (Palestine)
✉ k.sabarna@paluniv.edu.ps

Received:30/05/2024 Accepted:24/06/2024 Published:30/06/2024

Abstract: ***Early stroke detection significantly increases the prognosis for both survival and rehabilitation. Patients are more likely to receive appropriate therapy that minimizes brain damage and lowers the risk of consequences if a stroke is detected early on. Researchers are motivated to investigate the possibilities of artificial intelligence and machine learning technologies in creating new categorization systems that can identify and detect strokes more quickly and accurately due to their rapid development. This could potentially enhance the likelihood of surviving and recuperating. The support-vector machine (SVM), logistic regression, decision tree, random forest, Bayes nets, and K-nearest neighbor (KNN) algorithms are employed in this study's CRISP model technique. To enhance the final quality, the dataset was balanced using an oversampling technique, and the algorithms employed were subjected to principal components analysis (PCA). With an accuracy rate of 99%, the Random Forest algorithm is regarded as the optimum for prediction. Our study illustrates that the random forest classification model using the data balancing strategy outperforms the other strategies investigated, with a 99% classification accuracy and a 98% F1 score. The study also shows that the outcomes are unaffected by preprocessing with the PCA technique. The next objectives of the study are to use a larger dataset, various preprocessing methods, and machine learning models to enhance the framework models.***

Keywords: Cerebral Vascular Accident (CVA), Stroke Machine Learning (ML), Principal Component Analysis (PCA), CRISP-DM, Data Balancing, and Algorithms.

1. Introduction

A cerebral vascular accident (CVA) A stroke, also referred to as a brain attack, is an interruption in the flow of blood to cells in the brain. Stroke accounts for about 11% of all deaths worldwide, making it the second most common cause of death, according to the World Health Organization (WHO). Thus, efficient detection techniques are required to lower the chance of fatality sudden neurological loss that lasts longer than 24 hours is referred to as a stroke. Hemorrhagic strokes (15%) are brought on by bleeding inside the brain or surrounding tissues, while ischemic strokes (85%) are brought on by artery blockage. Atherosclerotic plaque accumulation, emboli, and localized blood clots can all cause ischemic strokes [1]. Paralysis, loss of vision or speech, and confusion can result from it. There are two primary forms of strokes: hemorrhagic stroke, which is brought on by a leaking vessel, and ischemic stroke, which is brought on by a blocked artery. The two main risk factors for stroke are tobacco use and high blood pressure, both of which can be considerably reduced by management [2]. Stroke risk can also be raised by other medical disorders such as heart disease and atrial fibrillation [2]. The most important thing to do when dealing with a stroke patient is to determine if the stroke is ischemic or hemorrhagic, as each form has different treatment guidelines. Time is also essential for maintaining neural function and halting more harm. Simultaneously, the general public has to be informed about stroke prevention strategies that involve positive lifestyle modifications. The time-consuming and error-prone nature of traditional stroke risk prediction approaches might postpone action and harm patient outcomes. Currently, the only way to determine a person’s risk for cerebral vascular diseases has been to use a combination of imaging such as an MRI scan, family history, demographic variables, and other risk factors evaluations such as inflammatory biomarker network with incident stroke risk, cognitive impairment, and imaging metrics [12].

Machine learning algorithms have demonstrated significant potential in precisely estimating the risk of stroke based on many risk variables. This can facilitate the early identification of individuals at high risk and prompt management [3]. To predict the presence of stroke disease with a range of associated factors, machine learning methods were used in this study as classification algorithms. Several techniques were employed to reduce dimensionality and balance the data; principal component analysis (PCA) is one of the approaches used. following the reduction of the dimensions. To find out which classification model predicts the dataset more accurately, many performance indicators are found and evaluated, including precision, recall, f-1 score, and precision. Improving stroke prediction will ultimately lower stroke-related mortality, which is the driving force behind this endeavor. In this study, machine learning algorithms were applied as classification algorithms to predict the presence of stroke disease with a variety of associated characteristics. Different methods were used to balance the data and to reduce dimensionality, a principal component analysis (PCA) approach is used. After reducing the dimensions. Different performance metrics such as precision, precision, recall, and f-1 score of classification models are identified and compared with each other to determine which one predicts more accurately in the dataset. The motivation for this work is to enhance stroke prediction, ultimately reducing stroke-related mortality.

2. Methodology

Data science projects follow diverse paths in different fields. Despite the variation, they share six basic pillars, forming the project life cycle [7]. CRISP-DM, developed by IBM in the late 1990s, provides a generalizable framework for data modeling [8]. This methodology has been widely adopted across various fields and tasks, proving particularly valuable for data science projects [8]. In this study, the CRISP-DM workflow system was relied upon for analysis, figure 1 shows the CRISP-DM steps.

Understand the work: The Business Understanding phase is the first phase in the CRISP-DM model of data analysis, focusing on the precise needs and requirements of stakeholders. Based on this stage, several important issues were identified to study stroke risk prediction to reduce deaths:

- Stakeholders: Hospitals, healthcare providers, and public health organizations.
- The primary goal: To reduce the risk of death due to stroke by predicting the probability of its occurrence in individuals.
- The plan: Develop a machine learning model to assess stroke risk based on patient data.
- Accurate predictions: The model must accurately predict the patient's likelihood of having a stroke.
- Reduced mortality: Early detection and intervention based on predictions should lead to fewer deaths from stroke.

Understand the data: Hence, the quest is to build a smarter stroke prediction system to reduce deaths, it is necessary to focus on the data available to the project at a high level by answering several questions, including:

- Data file: The dataset contains information about the patient (gender, age, various diseases, smoking status, etc.). Each row in the dataset provides information relevant to the patient, as well as whether or not the person has had a stroke.
- Data dictionary file: A data dictionary is a document that lists the name and explanation for each feature in a data set.
- Information included in the data: The file contains 5110 rows and 12 columns.; There is a mixture of data types: (3 floats, 4 integers, 5 objects).
- Missing values: The dataset contains a small number of missing values. The missing values will be assigned to the BMI column and the missing values will be removed from the smoking status column.
- Features exploration: Distribution of data for each column. exploring features is important to solve subsequent data problems, such as outliers, and knowing how important the features are for building the model, as well as how balanced the target data is.

1.1 Visualize some features:

Through the graph shown in Figure 2, it was noted that the target “stroke” is classified into two categories: Stroke-afflicted and non-stroke-related. These two classifications are not parallel and need to process the imbalanced data so that machine learning models can interpret and avoid bias in learning for the higher class at the expense of the lower class.

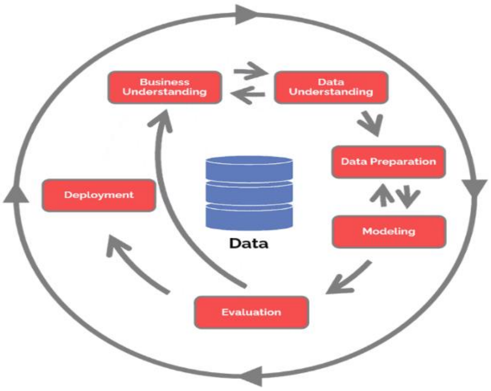

Figure 1: The 6 Stages of the CRISP-DM Process [9]

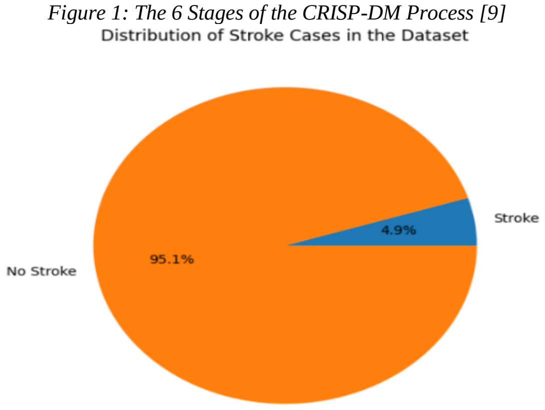

Figure 2: Distribution of Stroke Cases in the Dataset

Figures 3 and 4 show that there are outlier values for some columns, namely average glucose level and body mass index (BMI), and these values were removed from them as shown in Figures 5 and 6.

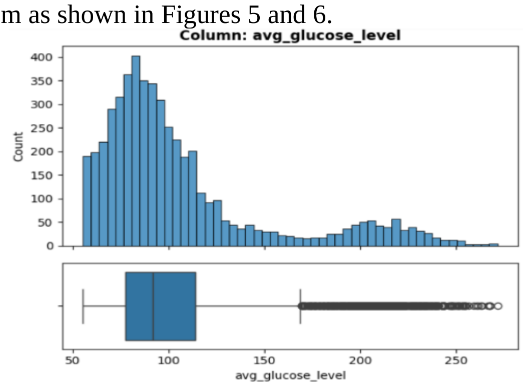

Figure 3: Visualize and explore a column of average glucose

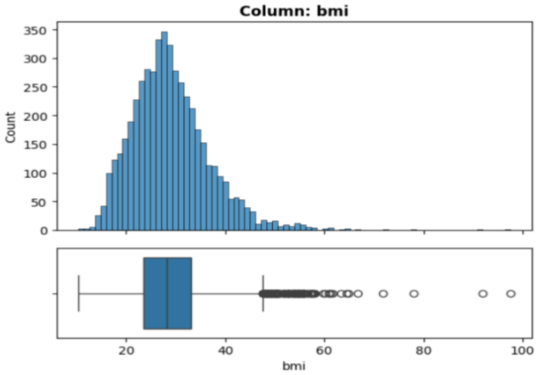

Figure 4: Visualize and explore a column (BMI)

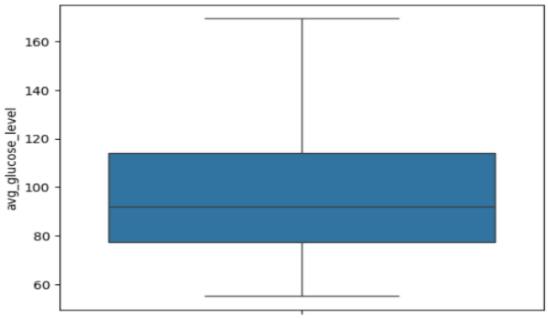

Figure 5: Visualize the average glucose column after

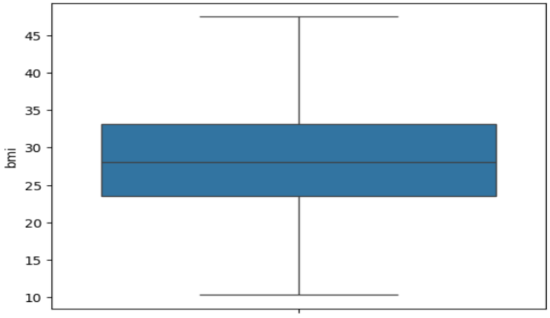

Figure 6: Visualize the BMI column after removing outliers
removing the outlier

3. Data Preparation:
After the data has been comprehended and suppressed during the understanding phase, it needs to be ready for modeling. By this point, it was determined which characteristics should be included in the model and how to translate the category and string features into numerical features for machine learning modeling.

4. Modeling: Support vector machine (SVM), logistic regression, decision trees, random forests, Bayes nets, and K-nearest neighbor (KNN) are the models that will be defined in this step. Principal component analysis (PCA) was added to the techniques utilized, along with oversampling to balance the data and enhance the quality of the outcome.

5. Evaluation: Every model will be assessed at this point, and the best model will be showcased. Table 1 shows the results.

Table 1: Models' performance comparison

No Data Balancing, No PCA				
Algorithm	Accuracy	Precision	Recall	F1
SVM	93%	0%	0%	0%
LR	93%	0%	0%	0%
decision tree	89%	9%	7%	8%
random forest	92%	0%	0%	0%
Bayes nets	8%	7%	100%	13%
KNN	93%	20%	2%	3%
Using Data Balancing and PCA				
Algorithm	Accuracy	Precision	Recall	F1
SVM	80%	75%	86%	80%
LR	77%	73%	82%	77%
decision tree	97%	94%	100%	97%
random forest	**98%**	96%	100%	**98%**
Bayes nets	75%	76%	70%	73%
KNN	92%	86%	100%	92%
Using Data Balancing				
Algorithm	Accuracy	Precision	Recall	F1
SVM	84%	79%	90%	84%
LR	77%	73%	80%	76%
decision tree	97%	93%	100%	96%
random forest	**99%**	97%	100%	**98%**
Bayes nets	48%	48%	100%	65%
KNN	92%	85%	100%	92%

The Random Forest algorithm achieved the best results with data balancing with an accuracy of ninety-nine percent. The Bayes nets model achieved the lowest Accuracy at 75%.

6. Deployment: Depending on project requirements, the dissemination phase can be as simple as creating a final report/presentation summarizing findings and recommendations.

3. Results
The findings were examined using several important criteria to thoroughly assess how well various classification algorithms performed. These requirements consist of:

– Accuracy: The most logical metric is accuracy. It provides an answer to the query, "What proportion or percentage of all the model's predictions was correct?"; Stated differently, accuracy is the percentage of accurate predictions the model produced out of all the predictions.
– Recall: Recall provides an answer to the following query: "What proportion/percentage of all the samples that belong to the positive class did the model classify correctly?" The recall will increase when the number of false negatives decreases.
– Precision: Precision provides an answer to the following query: "How often is our model correct when it predicts the class to be the positive class?" It asks, "Out of all the samples that the model predicted to be positive, what proportion/percentage were true positives?" It focuses on how reliable the model is at predicting the positive class.
– F1 Score: This score indicates how effectively the model balances recall and precision in the confusion matrix [10]. Figure 7 presents the formulas for the calculation of the accuracy, recall, and precision from the confusion matrix.

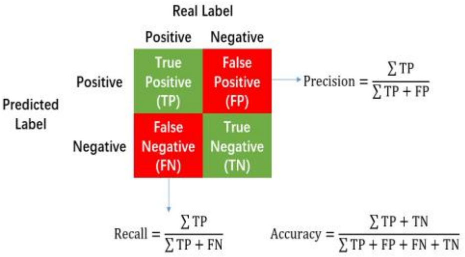

***Figure 7**: Calculation of Accuracy Recall and Precision*

Accuracy: The algorithms that classified patients as either unwell or healthy were the Balanced Random Forest and Balanced Decision Tree algorithms, which had the best accuracy (99% and 97%, respectively). In terms of accuracy, the Balanced SVM and Balanced LR algorithms come in second and third, respectively, at 84% and 77%. Overall, most algorithms performed well in terms of accuracy, and after Principal Component Analysis (PCA) and balancing, there was a noticeable increase.

Recall: The Bayes Nets method fared better than other algorithms in terms of correctly recalling information, attaining 100% of the time, indicating its capacity to identify all patients who are infected. In terms of recall, the Balanced LR and Balanced Decision Tree algorithms come in second and third, respectively, with 100% and 80%. Notably, high recall was attained by the Balanced Random Forest and Balanced SVM algorithms (97% and 90%, respectively).

Precision: All patients recognized as sick were correctly classified using the Balanced KNN and Balanced Bayes Nets algorithms, which had the highest accuracy (100% and 76%, respectively). In terms of accuracy, the Balanced Random Forest and Balanced Decision Tree algorithms come in second and third, respectively, at 97% and 93%. Overall, most algorithms performed well in terms of accuracy, and after Principal Component Analysis (PCA) and balancing, there was a noticeable increase.

F1 Score: The algorithms for balanced random forest and balanced decision tree had the greatest F1 measurements (98% and 96%, respectively), showing that recall and precision were well-balanced. In terms of F1 measurement, the Balanced SVM and Balanced LR algorithms rank second (84% and 76%, respectively). When it came to F1 measurement, most algorithms performed well overall, and their performance significantly improved following Principal Component Analysis (PCA) and balancing.

Table 2 illustrates the comparison of the results for this study with other related works.

Table 2.: Comparison of the results with other studies

		Proposed models		[4]	[6]		[11]	
Algorithm	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
SVM	84%	84%	-	-	80	81.1	-	-
LR	77%	76%	79	-	78	77.6	70	67
decision tree	97%	96%	94	95	66	77.6	69	71
random forest	**99%**	**98%**	96	96	73	80.4	73	71
Bayes nets	48%	65%	-	-	-	-	-	-
KNN	92%	92%	-	-	80	80.4	-	-

4. Conclusion

A stroke is considered a potentially fatal medical disease that has to be treated as soon as possible to prevent further worsening. Using machine learning models and techniques may help in the early detection of the stroke and reduce its serious effects. In this research, several machine learning algorithms have been examined to accurately predict stroke based on various features that were used in the dataset. The study shows that with a classification accuracy of 99% and an F1-score of 98%, the random forest classification model with the use of the data balancing technique performs better than the other techniques examined. The study also finds that the use of the PCA algorithm as a preprocessing step does not affect the results. The study's next goals are to improve the framework models using a larger dataset, using different preprocessing techniques and machine learning models.

References

[1] Pikula A, Howard BV, Seshadri S. Stroke and Diabetes. In: Cowie CC, Casagrande SS, Menke A, Cissell MA, Eberhardt MS, Meigs JB, Gregg EW, Knowler WC, Barrett-Connor E, Becker DJ, Brancati FL, Boyko EJ, Herman WH, Howard BV, Narayan KMV, Rewers M, Fradkin JE, editors. Diabetes in America. 3rd ed. Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases (US); 2018 Aug. CHAPTER 19. PMID: 33651535. Available: https://pubmed.ncbi.nlm.nih.gov/33651535/

[2] World Health Organization (WHO), " Stroke, Cerebrovascular accident". Available: https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html

[3] K. Mridha, S. Ghimire, J. Shin, A. Aran, M. M. Uddin and M. F. Mridha, "Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study With a Web Application for Early Intervention," in IEEE Access, vol. 11, pp. 52288-52308, 2023, doi: 10.1109/ACCESS.2023.3278273.

[4] 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, 7633381. https://doi.org/10.1155/2021/7633381

[5] Ivanov, I. G., Kumchev, Y., & Hooper, V. J., An optimization precise model of stroke data to improve stroke prediction. Algorithms, 16(9), 417, 2023. https://doi.org/10.3390/a16090417

[6] Gangavarapu Sailasya and Gorli L Aruna Kumari, “Analyzing the Performance of Stroke Prediction using ML Classification Algorithms” International Journal of Advanced Computer Science and Applications(IJACSA), 12(6), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120662.

[7] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999).

CRISP-DM 1.0: Step-by-step data mining guide. Available: https://api.semanticscholar.org/CorpusID:59777418

[8] Quantum, Data Science project management methodologies, Medium, 2019, August 20. Available: https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb

[9] Hotz, N., What is CRISPR-DM? Data Science Process Alliance, 10 September 2018. Available: https://www.datascience-pm.com/crisp-dm-2/

[10] J. Ma, Y. Ding, J. C. P. Cheng, Y. Tan, V. J. L. Gan and J. Zhang, "Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective," in IEEE Access, vol. 7, pp. 148059-148072, 2019, doi: 10.1109/ACCESS.2019.2946401.

[11] Elangovan, Viswa Priya Subramaniyam; Devarajan, Rajeswari; Khalaf, Osamah I.; Sharif, Mhd Saeed; and Elmedany, Wael (2024) "Analysing an imbalanced stroke prediction dataset using machine learning techniques," Karbala International Journal of Modern Science: Vol. 10: Iss. 2, Article 8. doi: https://doi.org/10.33640/2405-609X.3355

[12] Association of Incident Stroke Risk With an IL-18-Centered Inflammatory Network Biomarker Composite Richard A. Martirosian, Crystal D. Wiedner, Jasmin Sanchez, Katherine T. Mun, Kiran Marla, Cristina Teran, Marissa Thirion, David S. Liebeskind, Emer R. McGrath, originally published May 2024. doi: https://doi.org/10.1161/STROKEAHA.123.044719Stroke. 2024; 55:1601–1608