Southeast Europe Journal of Soft Computing ISSN 2233 – 1859
Available online: http://scjournal.ius.edu.ba

Vol 13, No. 1, March 2024, pp. 115-123

Brain Stroke Prediciton Model Based SMOTE and Machine Learning Algorithms

[1]Alhussain Waad Mohammed, [1]Hwraa Kareem Hmoud, [1]Ahmed Mohmed Abd Alzahra, [1]Ahmad Muneathir Sukar and [2*]Alaa Khalaf Hamoud

[1]Department of Computer Information Systems,
[2]Department of Cybersecurity,
College of Computer Science and Information Technology,
University of Basrah,
Basrah, Iraq.
**Corresponding Author*: alaa.hamoud@uobasrah.edu.iq

Article Info

Article history:

Article received on 10 01 2024

Received in revised form 10 02 2024

Keywords:

Brain Stroke; Feature Selection; SMOTE, Decision Tree; Logistic Regression; Random Forest; Gradient Boosted Trees; KNIME.

ABSTRACT: A brain stroke is a critical medical emergency that causes disability and death. The pre-diagnosis of this case can reduce the complications and problems that affect the brain as a result of being affected by the complications that occur during the injury. This study lists an analysis process on a brain stroke dataset using the KNIME tool, which provides a set of different machine learning components such as random forest, Decision Tree Learner, Gradient Boosted Trees Learner, and Logistic Regression algorithms. The problem of imbalanced data will be handled as part of data preprocessing. The factors that affect the brain stroke will be explored based on feature selection approaches such as forward feature selection, backward feature elimination, genetic algorithms, and random. The aim is to build a model that helps doctors diagnose the disease accurately based on the results we obtained from the study and analysis. The results showed that logistic regression outperformed the other algorithms after applying the algorithm with forward feature selection and backward feature elimination.

1. INTRODUCTION

According to the World Stroke Organization, 13 million people get a stroke each year, and approximately 5.5 million people will die as a result. So stroke is one of disorder leading causes of death and disability worldwide, and that is why its imprint is serious in all aspects of life [1]–[4]. Stroke not only affects the patient but also affects the patient's social environment, family, and workplace. In addition, contrary to popular belief, it can happen to anyone at any age, regardless of gender or physical condition. The functioning of various human body parts is essential to our existence. Strokes, a serious health condition, are a major cause of mortality, especially prevalent in individuals over 65 years [5], [6]. Similar to heart attacks impacting the heart, strokes impair brain function. They are caused either by a disruption in the brain's blood supply or by bleeding from ruptured brain vessels. This blockage or rupture prevents blood and oxygen from nourishing brain tissue. Strokes rank as the fifth leading cause of death globally in both developed and developing countries. Immediate medical attention significantly increases a stroke victim's recovery prospects. Delay in treatment can lead to death, irreversible damage, or severe brain impairment. Various factors contribute to stroke risk, including diet, sedentary lifestyle, alcohol consumption, tobacco use, personal and medical history, and other complications as outlined by the National Heart, Lung, and Blood Institutes. Recent advancements in clinical and medical services have been propelled by artificial intelligence, machine learning, and data science. Machine learning, a cornerstone of the current era, is employed for the early prediction of numerous health issues, including strokes. Early detection is vital for effective stroke treatment. In healthcare, machine learning plays an essential role in diagnosing and predicting diseases. Stroke prediction currently utilizes machine-learning algorithms. Managing large medical data sets requires robust data analysis tools, and AI's role in medicine is a key research area. AI systems can identify patients at higher risk of stroke based on their medical history. Factors like age, blood pressure, and sugar levels are analyzed to assess disease risk [7]–[11]. Classification algorithms are used for disease prediction in complex cases. Research has explored deep learning models like a multi-layer artificial neural network for stroke prediction. Other studies have focused on developing intelligent systems for this purpose [12]–[14].

In some cases, the data you use to train your data analysis model has an imbalanced distribution between classes. For example, there may be a small number of examples in the "infected" category and a large number of examples in the "not infected" category. The imbalanced dataset is the data that has an unequal representation of the final class. Some deep learning algorithms such as decision trees may need a balanced distribution between classes to achieve good classification performance. . Since the aim of most machine learning models is to utilize the final class as a base for the prediction, this will affect the reliability of model predictions If you are using unbalanced data, the synthesis minority oversampling technique (SMOTE) tool helps adjust the distribution of classes by adding artificial rows [15]–[17].

This study aims to demonstrate the effectiveness of machine learning algorithms in predicting brain strokes. It utilizes various algorithms on a data set to identify the best predictive approach for stroke onset. This study employs random forest (RF), decision tree (DT), and gradient boosted trees (GBT) and logistic regression (LR) as classification tools to predict stroke considering various related factors. The data preprocessing involved applying SMOTE to handle the problem of imbalanced data representation. The performance, error rate, incorrect classification, and correct classification are evaluated and compared through the scorer node to determine the highest accuracy among the results for the available data. The rest of the paper is organized as follows: Section 2 lists the related works and discusses the literature review; Section 3 explains the methodology framework; and Section 4 explains the concluded points and lists future works.

2. LITERATURE REVIEW

In [18], Padimi et al. proposed a model to predict brain strokes based on machine learning algorithms. provides a model for a detailed analysis and comparison of various machine learning algorithms for early stroke prediction. It discusses the importance of early detection of strokes and how machine learning algorithms can help minimize their disabling effects. The paper also covers the risk factors considered and the performance evaluation metrics used to compare the algorithms. The machine learning algorithms discussed in the paper include DT, RF, naive bayes, and self-training techniques. The results of the study show that the self-training technique outperformed the other algorithms in terms of accuracy and sensitivity.

In [19], Kokkotis et al. proposed model-based machine learning algorithms to predict stroke using imbalanced data. They explained how machine learning can be used to forecast strokes and offered an explanation for those predictions. An explainable machine learning pipeline with an emphasis on post-hoc explainability could pinpoint significant risk factors for stroke prediction and their effects on model output. They discussed the use of explainable artificial intelligence (AI) for DT as they proposed a method for generating global explanations of DT models by aggregating local explanations of individual predictions. The authors propose a causal approach to feature relevance quantification that takes into account the causal relationships between features and the target variable. The authors review the current understanding of the mechanisms underlying the increased risk of stroke in older adults and the potential interventions to prevent or treat stroke in this population.

In [20], Pitchai et al. proposed a model based an explainable machine learning research that utilized electromyography (EMG) bio-signals to predict stroke disorders. According to the study, stroke illnesses in patients who are considered to be in danger can be predicted and analyzed using deep learning and machine learning algorithms. Ensuring that signaling is the primary method used in extensive research into the early diagnosis and prognosis of strokes and other chronic diseases, like diabetes and heart disease, is the ultimate goal of the study.

In [21], Tazin et al. The goal of this research is to use machine learning models for the early detection and prediction of stroke disease. The authors compared the outcomes of four distinct machine learning algorithms: RF, DT, voting classifier, and logistic regression (LR), using a dataset that was made available to the public.

According to the study, these models are more reliable at predicting the chance of a stroke occurring because they have a significantly higher accuracy rate than those used in earlier studies. Together with a block diagram and flow diagram of the suggested system, the model offers a thorough explanation of the experimental methodology and procedures employed in the investigation. There is also discussion of the research's implications for stroke prevention and treatment worldwide.

In [22], Sailasya et al. investigated the application of machine learning algorithms to analyze different physiological factors and predict the likelihood of a brain stroke. The researchers compared the accuracy, precision score, and recall score of several machine learning algorithms, including K-Nearest Neighbors, LR, and Naïve Bayes. The results show that the Naïve Bayes Classification algorithm performed the best with an accuracy of approximately 82. The model concludes by discussing the potential applications in the fields of healthcare and medicine.

3. BACKGROUND

In the field of machine learning, there are several tools available that aid in data mining and predictive modeling. One such tool is KNIME, shown in Figure 1, which stands out from the others due to its extensive collection of tools that can be used for machine learning. KNIME, an open-source data analytics platform, stands out for its user-friendly interface, enabling non-coders to build workflows easily. This GNU General Public License v3-licensed open-source predictive analytics platform works well with a variety of data formats. Everything from simple comma separated values (CSV) or microsoft excel (XLS) files to more complicated data structures like extensible markup language (XML), uniform resource locator (URLs), and relational databases including well-known ones like db2, Oracle, and MySQ. It empowers advanced analytics, machine learning, and workflow management. KNIME's efficiency, paired with its flexibility and extensive capabilities, positions it as a top choice for data analysis [23], [24].

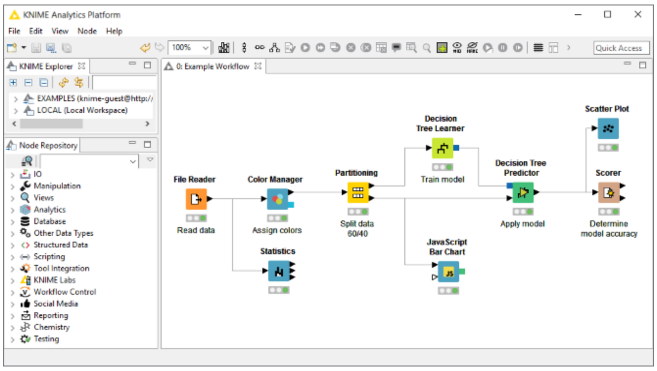

Figure 1:KNIME analytical tool

The problem of an imbalanced dataset is a major problem that may affect the model's accuracy. This problem reflects the data representation of the target class, where sometimes the data representation of a specific value exceeds the representation of other values. The synthetic minority oversampling technique (SMOTE) is one of the data preprocessing techniques that may be used to balance data representation and then make the model implementation more reliable. The technique works by creating synthetic data rows based on existing real data. You start by choosing a model from the underrepresented class (e.g., “active”) and choosing a nearest neighbor from the same class. We can then draw a line between these two models and choose a point along the line. We can create a new synthetic data row using this point and define the attribute values (cell values) of the new row based on this point. In this way, we were able to create synthetic data that supplemented the underrepresented class and made the distribution of classes more balanced. This can help us improve the performance of deep learning algorithms when training on this data [25]–[28].

There are many machine learning algorithms that can be implemented in the model. RF where the name implies, the RF algorithm is made up of many different DT working together as a group. The class that receives the most votes becomes the model prediction, and each individual tree in the RF arises from the class prediction. The RFalgorithm's fundamental idea is the majority opinion, which is a straightforward concept. A single DT model will not perform as well as a large number of relatively unconnected models (trees) acting as a committee, which is why the RF algorithm is so popular in data science [29], [30].

Next, DT Learner tool initiates a classification DT in the main memory. A nominal target attribute is required. Numerical or nominal attributes may be included in the other considerations when making decisions. At a given split point, numerical splits divide the domain into two partitions because they are always binary (having two outcomes). Nominal splits can have as many outcomes as nominal values, or they can be binary (having two outcomes). The nominal values in a binary split are separated into two subsets. The algorithm offers the gain ratio and the Gini index as two quality metrics for split computation. In addition, a post-pruning technique is available to decrease the size of the tree and improve prediction accuracy. The minimum description length principle forms the basis of the pruning technique. Gradient Boosted Trees Learner tool makes use of very shallow regression trees and a unique kind of boosting. As it is used in the RF, Simple Regression Tree, and Tree Ensemble nodes, the base learner for this ensemble method is a simple regression tree. By default, binary splits are used to build a tree for both nominal and numeric attributes (although multiway splits can be used for the latter). By evaluating every potential direction and choosing the one that produces the best outcome (i.e., the largest gain), the built-in missing value handling algorithm attempts to determine which direction missing values should go. The mean target value of the records within a leaf node in a regression tree is the predicted value for that leaf node. Therefore, if the variance of target values within a leaf is small, the predictions are optimal (relative to the training data). Splits that reduce the total squared errors in each of their offspring. Finnaly, Logistic Regression is a prediction node that must be connected to some data in order to test the required data. This is done only if the data contains the columns used by the learner model. Prediction is made by entering one or more inputs, through which the data gives groups of data in order to give outputs in the form of a linear slope. It is one of the most famous algorithms in statistics that always gives good results. The relationship between variables exists through certain linear equations, where the technique predicts the outputs through the specified inputs with a minimum of error [31]–[34].

4. METHODOLOGY

The methodology framework is depicted in Figure 2. The dataset is read in the first step, while the required data preprocessing in performed in step two. The feature selection algorithm in performed in step three, while the model is trained using machine learning algorithm in the next steps. The final step involves model evaluation.

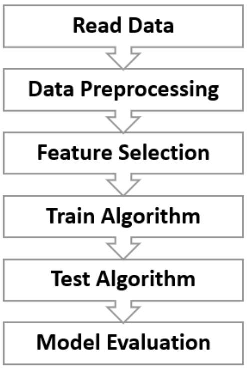

Figure 2: Methodology framework

The dataset [35]consists of several rows and columns, where the number of rows is about 4981 and the number of columns is 12. The first column represents the gender, either male (1966) or female (2767), as the number of females is greater than the number of males, while the second column represents the patient's age, where the age range is between 3 months and 82 years. The third column represents hypertension, which represents the percentage of high blood pressure in the patient, as it consists of two numbers, zero and one, meaning the pressure is high (413), or not (4320). The fourth column represents heart disease and also consists of two numbers: zero, which represents not infected (4505) and one, which represents infected (228). The fifth column represents the marriage status of the person, whether married or not. The sixth column represents the patient's type of work, whether it is private (2712), self-employed (739), government job (611), etc. (671). The seventh column represents the type of residence, whether urban (2397) or rural (2336). The eighth column represents the average clucose level in the patient while the ninth column represents the average blood sugar level (bmi). The tenth column represents the patient's body mass index, and the eleventh column represents the smoking status column, which represents whether the person is a smoker (336), never (893), unknown (724), or formerly smoked (1107). The last column represents the infection column, and the entry is either zero, which represents the uninfected (4733) or one, which represents the infected person (248).

Table 1: Dataset metadata

Seq	**Coulmn**	**Unique values**	**Missing Values**
1	Gender	2	None
2	Age	104	None
3	Hypertension	2	None
4	heart_disease	2	None

5	ever_married	2	None
6	work_type	4	None
7	Residence_type	2	None
8	avg_glucose_level	3895	None
9	Bmi	342	None
10	smoking_status	4	None
11	Stroke	2	None

Due to the data representation in the final class (infected or not), the dataset is considered imbalanced which may affect the final model prediction and results reliability. The number of infected person is lower than uninfected, where the role of SMOTE algorithm to make a balance between these two categories. The data distribution will be balanced in order to reduce the bais toward the uninfected class of the dataset, improve the result accuracy, and make the model more reliable [28], [36], [37]. The model implementation flow is depicted in Figure 3.

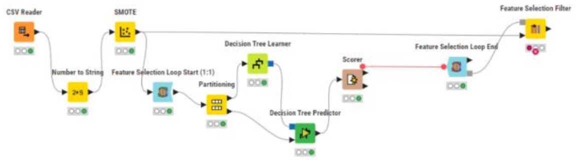

Figure 3: Model implementation in KNIME

The CSV reader tool is utilized to import data from CSV files. This tool efficiently reads and loads data into the KNIME environment. Different tools had been employed within KNIME to analyze and preprocess the data. The tool (to String) is utilized to transform the final class into String. Next, the SMOTE tool is utilized to balance the dataset. As mentioned before, the dataset is imbalanced

For algorithm selection, different machine learning algorithms were utilized, including DT, RF, GBT, and LR in our workflow. These algorithms are suitable for various tasks, such as classification or regression. The scorer tool had been utilized in the model to measure the accuracy of the model for prediction. The first result showed that the accuracy was low, and due to that, the feature selection algorithms were utilized. The feature selection is implemented by adding a feature selection loop to identify the factors influencing the brain stroke, where the aim is to improve the prediction accuracy by keeping the relevant features. There are four strategies for feature selection:

1. *Forward Feature Selection (FFS)*: this machine learning technique begins with no features and progressively adds ones that notably enhance model performance. The impact of each addition is assessed through metrics such as accuracy and mean squared error. This method helps in refining the model by selecting only impactful features, thereby streamlining the model's complexity, reducing noise in the data, and lessening the risk of overfitting. It's widely applicable in diverse sectors like finance and healthcare and is utilized until further additions don't markedly improve performance.

2. *Backward Feature Elimination (BFE):* this is a feature selection method used in machine learning, particularly for models like linear regression. It starts with all available features and removes the least significant ones iteratively. This process involves fitting a model with all features, evaluating each feature's significance (usually via p-value), and eliminating the least significant ones. The cycle continues until all remaining features are sufficiently significant, typically below a p-value threshold like 0.05. This method simplifies the model by removing non-essential features, enhancing model accuracy, and preventing overfitting. It's a practical choice due to its simplicity and computational efficiency.

3. *Genetic Algorithm (GA):* in machine learning, feature selection mimics natural evolutionary processes. They generate a group of potential solutions, where each represents a different feature combination. The performance of these combinations in a predictive model determines their 'fitness'. Through natural selection-inspired processes like crossover and mutation, the algorithm refines these combinations over generations, aiming to discover the most effective feature set. This approach is particularly beneficial for large feature spaces where traditional selection methods are less feasible, offering a robust solution for complex gubernatorial optimization problems in feature selection.

4. *Random:* In machine learning, random feature selection is a straightforward method where features are selected randomly from the full set without any specific performance-based criteria. This approach differs from systematic methods like forward feature selection or backward feature elimination. For initial, quick selection or unbiased exploration of data, random feature selection offers a baseline but may not always provide the most effective feature combination for optimal model performance.

In this training, we applied four distinct algorithms to our data, achieving the previously mentioned accuracy levels. Each algorithm will be thoroughly explained to understand their unique contributions and methodologies in data training:

A. *Random Forests (RF):* which is a machine learning algorithm recognized for its high accuracy. It operates by creating numerous DT during the training phase and then making predictions based on the majority vote or average outcome of these trees. This approach is effective for both classification and regression tasks. Its strength lies in its ability to manage overfitting and handle complex datasets with many variables. Nonetheless, the actual accuracy of RF can vary based on the specific characteristics of the dataset and the way it's implemented.

B. *Decision Tree (DT):* the accuracy of a DT algorithm measures its effectiveness in correctly predicting or classifying data. In this structure, internal nodes represent attribute tests, branches indicate test outcomes, and leaf nodes denote class labels. The tree's rules are defined by paths from the root to each leaf. Accuracy hinges on the tree's ability to distinctly separate data, influenced by the tree's depth and data complexity. While DTs are easy to understand, they may overfit in complex or noisy datasets.

C. *Gradient-Boost Trees (GBT)*: are evaluated on their accuracy in predicting or classifying correctly. This technique involves building trees in a sequence, each correcting the previous tree's errors. The final decision is derived by integrating the output of all trees. Noted for their high accuracy in handling complex data, gradient-boosted trees demand precise tuning to avoid overfitting and are widely utilized for their adaptability and effectiveness in diverse scenarios.

D. *Logistic Regression (LR)*: the accuracy in logisti regression gauges how effectively it classifies or predicts binary outcomes. It employs a logistic function for modeling binary targets, useful in scenarios like determining win/lose or healthy/sick. Its performance, however, can vary based on dataset properties and model specifications. Known for its straightforwardness and clarity, logistic regression is widely used in binary classification tasks.

5. RESULTS DISCUSSION

This section lists the results obtained from implementing the machine learning algorithms with feature selection algorithms after implementing the SMOTE filter to balance data representation. The data had been divided into two sets (train with 80% of data) and (20% of data). The feature selection algorithms (FFS, BFE, GA, and random) were utilized to determine the feature set that may affect the final class. The model implementation in Figure 4 shows that the best algorithm with prediction is LR with FFS and BFE with 100% accuracy, followed by LR with GA and random feature selection with 98.3% and 95.9%, respectively. Next, the DT algorithm is the next algorithm in prediction accuracy, with 93.2% and 92.9% for the GA and FFS, respectively, while the accuracies are 91.7% and 90% for the DT with BFE and random feature selection. The gradient-boosted trees prediction accuracies with feature selection algorithms are (90.4%, 90.3%, 88.9%, and 88.8%) with BFE, GA, random, and FFS, respectively. While the algorithm with the lowest accuracies compared with the previous algorithms is R (87%, 87%, 81.4%, and 80.3%) with BFE, FFS, random, and GA feature selection.

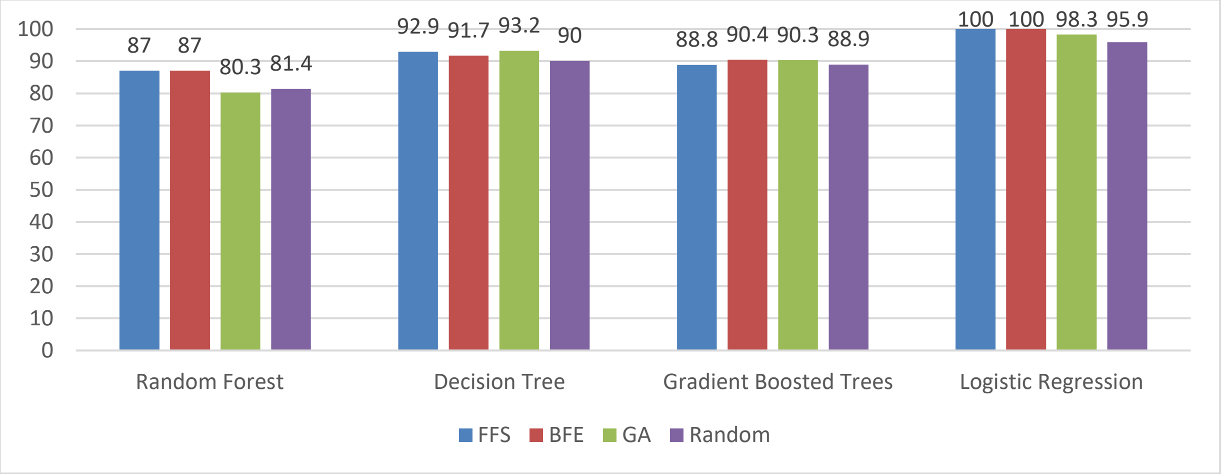

Figure 4: Comparison of algorithms' accuracy

Table 2: Candidate features according to Logistic Regression

Accuracy	Algorithm	Features
100%	FFS	4,8,9
100%	BFE	4,7,8,10
98.3%	GA	1,2,8,9
95.9%	Random	2,6

According to that, the LR algorithm has the optimal prediction accuracy compared with other algorithms. In Table 2, the best accuracies of the algorithm are with the FFS and BFE feature selection algorithms. The feature sets (4, 8, and 9) and (4, 7, 8, and 10) have the most correlations with the final class (stroke). The features (heart disease, average clocuse level, bmi, residence type, and smoking type) had the most correlations with the brain stroke.

6. CONCLUSION AND FUTURE WORKS

A brain stroke is a medical case that affects human lives and is considered a medical concern. This case may cause different critical situations, such as disability and death. Different factors may affect human health and cause a stroke, such as smoking status, blood pressure, cholesterol level, and age. Finding and determining the features that may affect the brain stroke may help in preventing this condition and helping medical institutions. The model implementation showed that LR had the optimal accuracy of 100% in prediction compared with RF, DT, and GBT algorithms. The feature selection step is crucial since it can explore the most correlated features for the final class. The model showed that the feature set (heart disease, average clocuse level, bmi, residence type, and smoking type) had the most correlations with the brain stroke. These results can encourage clinical staff and individuals to pay more attention to these features. In the future, this model can be tested against new real-time data and give recommendations. The model can be integrated with different data of patients (such as image data of diagnosis) to explore the other factors that may affect the brain stroke.

REFERENCES

[1] X. W. Gao, R. Hui, and Z. Tian, “Classification of CT brain images based on deep learning networks,” *Comput Methods Programs Biomed*, vol. 138, 2017, doi: 10.1016/j.cmpb.2016.10.007.

[2] K. Overgaard, “The effects of citicoline on acute ischemic stroke: A review,” *Journal of Stroke and Cerebrovascular Diseases*, vol. 23, no. 7. 2014. doi: 10.1016/j.jstrokecerebrovasdis.2014.01.020.

[3] J. G. Merino, “Clinical stroke challenges: A practical approach,” *Neurology: Clinical Practice*, vol. 4, no. 5. 2014. doi: 10.1212/CPJ.0000000000000082.

[4] O. Ozaltin, O. Coskun, O. Yeniay, and A. Subasi, “A Deep Learning Approach for Detecting Stroke from Brain CT Images Using OzNet,” *Bioengineering*, vol. 9, no. 12, 2022, doi: 10.3390/bioengineering9120783.

[5] S. Yalçın and H. Vural, “Brain stroke classification and segmentation using encoder-decoder based deep convolutional neural networks,” *Comput Biol Med*, vol. 149, 2022, doi: 10.1016/j.compbiomed.2022.105941.

[6] M. S. Sirsat, E. Fermé, and J. Câmara, “Machine Learning for Brain Stroke: A Review,” *Journal of Stroke and Cerebrovascular Diseases*, vol. 29, no. 10. 2020. doi: 10.1016/j.jstrokecerebrovasdis.2020.105162.

[7] A. Väänänen, K. Haataja, K. Vehviläinen-Julkunen, and P. Toivanen, “AI in healthcare: A narrative review,” *F1000Res*, vol. 10, 2021, doi: 10.12688/f1000research.26997.2.

[8] P. Apell and H. Eriksson, “Artificial intelligence (AI) healthcare technology innovations: the current state and challenges from a life science industry perspective,” *Technol Anal Strateg Manag*, vol. 35, no. 2, 2023, doi: 10.1080/09537325.2021.1971188.

[9] J. Yu, S. Park, C. M. B. Ho, S. H. Kwon, K. H. cho, and Y. S. Lee, “AI-based stroke prediction system using body motion biosignals during walking,” *Journal of Supercomputing*, vol. 78, no. 6, 2022, doi: 10.1007/s11227-021-04209-1.

[10] J. Yu, S. Park, S. H. Kwon, K. H. Cho, and H. Lee, “AI-Based Stroke Disease Prediction System Using ECG and PPG Bio-Signals,” *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3169284.

[11] J. Yu, S. Park, S. H. Kwon, C. M. B. Ho, C. S. Pyo, and H. Lee, “AI-based stroke disease prediction system using real-time electromyography signals,” *Applied Sciences (Switzerland)*, vol. 10, no. 19, 2020, doi: 10.3390/app10196791.

[12] F. Farrokhi *et al.*, “Investigating Risk Factors and Predicting Complications in Deep Brain Stimulation Surgery with Machine Learning Algorithms,” *World Neurosurg*, vol. 134, 2020, doi: 10.1016/j.wneu.2019.10.063.

[13] S. Mushtaq and K. S. Saini, “A Review on Predicting Brain Stroke using Machine Learning,” in *Proceedings of the 17th INDIACom; 2023 10th International Conference on Computing for Sustainable Global Development, INDIACom 2023*, 2023.

[14] A. Cerasa *et al.*, “Predicting Outcome in Patients with Brain Injury: Differences between Machine Learning versus Conventional Statistics,” *Biomedicines*, vol. 10, no. 9. 2022. doi: 10.3390/biomedicines10092267.

[15] Q. Hang, J. Yang, and L. Xing, “Diagnosis of rolling bearing based on classification for high dimensional unbalanced data,” *IEEE Access*, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2919406.

[16] X. Li and L. Zhang, “Unbalanced data processing using deep sparse learning technique,” *Future Generation Computer Systems*, vol. 125, 2021, doi: 10.1016/j.future.2021.05.034.

[17] A. S. Hussein, T. Li, C. W. Yohannese, and K. Bashir, “A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE,” *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, 2019, doi: 10.2991/ijcis.d.191114.002.

[18] V. Padimi, V. S. Telu, and D. D. Ningombam, “Performance analysis and comparison of various machine learning algorithms for early stroke prediction,” *ETRI Journal*, 2022, doi: 10.4218/etrij.2022-0271.

[19] C. Kokkotis *et al.*, “An Explainable Machine Learning Pipeline for Stroke Prediction on Imbalanced Data,” *Diagnostics*, vol. 12, no. 10, 2022, doi: 10.3390/diagnostics12102392.

[20] R. Pitchai *et al.*, “An Artificial Intelligence-Based Bio-Medical Stroke Prediction and Analytical System Using a Machine Learning Approach,” *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/5489084.

[21] 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,” *J Healthc Eng*, vol. 2021, 2021, doi: 10.1155/2021/7633381.

[22] 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, doi: 10.14569/IJACSA.2021.0120662.

[23] A. Fillbrunn, C. Dietz, J. Pfeuffer, R. Rahn, G. A. Landrum, and M. R. Berthold, “KNIME for reproducible cross-domain analysis of life science data,” *Journal of*

Biotechnology, vol. 261. 2017. doi: 10.1016/j.jbiotec.2017.07.028.

[24] KNIME AG, “KNIME Analytics Platform | KNIME,” *Knime*. 2019.

[25] S. Alija, E. Beqiri, A. S. Gaafar, and A. K. Hamoud, “Predicting Students Performance Using Supervised Machine Learning Based on Imbalanced Dataset and Wrapper Feature Selection,” *Informatica*, vol. 47, no. 1, 2023.

[26] M. B. M. Kamel *et al.*, “A Comparative Study of Supervised/Unsupervised Machine Learning Algorithms with Feature Selection Approaches to Predict Student Performance,” *International Journal of Data Mining, Modelling and Management*, vol. 15, no. 4, 2023, doi: 10.1504/ijdmmm.2023.10055032.

[27] 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.

[28] R. K. Tripathi, L. Raja, A. Kumar, P. Dadheech, A. Kumar, and M. N. Nachappa, “A Cluster Based Classification for Imbalanced Data Using SMOTE,” in *IOP Conference Series: Materials Science and Engineering*, 2021, p. 12080.

[29] L. Breiman, “Random Forests,” *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[30] J. Bai, Y. Li, J. Li, X. Yang, Y. Jiang, and S. T. Xia, “Multinomial random forest,” *Pattern Recognit*, vol. 122, 2022, doi: 10.1016/j.patcog.2021.108331.

[31] “Pattern Recognition and Machine Learning,” *J Electron Imaging*, vol. 16, no. 4, 2007, doi: 10.1117/1.2819119.

[32] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, vol. 398. John Wiley & Sons, 2013.

[33] N. Landwehr, M. Hall, and E. Frank, “Logistic Model Trees,” *Mach Learn*, vol. 59, no. 1, pp. 161–205, 2005, doi: 10.1007/s10994-005-0466-3.

[34] S. Sperandei, “Understanding logistic regression analysis,” *Biochem Med (Zagreb)*, vol. 24, no. 1, 2014, doi: 10.11613/BM.2014.003.

[35] Jillani Soft Tech, “Brain Stroke Dataset,” https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset.

[36] A. Desiani, S. Yahdin, A. Kartikasari, and I. Irmeilyana, “Handling the imbalanced data with missing value elimination SMOTE in the classification of the relevance education background with graduates employment,” *IAES International Journal of Artificial Intelligence*, vol. 10, no. 2, p. 346, 2021.

[37] A. M. Sowjanya and O. Mrudula, “Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms,” *Applied Nanoscience (Switzerland)*, vol. 13, no. 3, 2023, doi: 10.1007/s13204-021-02063-4.