




P-ISSN: 2788-9971 E-ISSN: 2788-998X

NTU Journal of Engineering and Technology

Available online at: <https://journals.ntu.edu.iq/index.php/NTU-JET/index>



Early Prediction of Stroke Risk Using Machine Learning Approaches and Imbalanced Data

Hassan M. Qassim 

Mosul Technical Institute, Northern Technical University, Iraq, hassanqassim@ntu.edu.iq

Article Informations

Received: 29-09-2024,
Revised: 15-11-2024,
Accepted: 15-12-2024,
Published online: 22-03-2025

Corresponding author:

Name: Hassan M. Qassim
Affiliation : Northern Technical
University
Email: hassanqassim@ntu.edu.iq

Key Words:

Decision Tree,
Imbalanced Data,
KNN,
LDA,
Naive Bayes,
Machine Learning Models.

ABSTRACT

Classifying medical datasets using machine learning algorithms could help physicians to provide accurate diagnosing and suitable treatment. For instance, stroke is one of the serious diseases that attacks many patients annually, and analyzing its symptoms in advance could save patients' lives. The warning signs of the stroke can be investigated to be used as attributes or predictors for machine learning models. This study evaluates the performance of four machine learning models to classify stroke datasets. Specifically, Decision Tree, Naïve Bayes, K- Nearest Neighbor (KNN) and Linear discriminant Analyses (LDA) models were trained on 11 attributes collected from 5110 patients to predict stroke risk. The findings showed that KNN outperformed the three other models with an achieved accuracy of 90%, while lowest accuracy was attained by LDA. Moreover, KNN model exhibited an acceptable prediction speed relative to the rest of models. The study also considered balancing the employed data prior validating the models to provide accurate classification. Cross-validation technique was used to avoid over-fitting and under-fitting during training phases.

@THIS IS AN OPEN ACCESS ARTICLE UNDER THE CC BY LICENSE:

<https://creativecommons.org/licenses/by/4.0/>



1. Introduction

Nowadays, machine learning models have attracted researchers' attention due its ability to analyse and classify various datasets [1]. These models manage complex pattern in the targeted dataset and address non-linear relationships to provide accurate and robust prediction. Moreover, they are capable to handle large imbalanced dataset through training in high-dimensional space. These characteristics make machine learning models critical and crucial in various domains and specifically in healthcare sector [2][3].

Namely, stroke is the second cause of death in the world and requires special attention to predict it is symptoms [4]. Hence, machine learning approaches are powerful tools that could be used to provide early detection of stroke symptoms. Moreover, stroke warning signs such as age, hypertension, heart diseases and person's weight are considered excellent attributes or features that make machine learning algorithms accurately detect the potential stroke [1]. However, there are only few researches on using machine learning algorithms to predict stroke.

For instance, Decision Tree and K-Nearest Neighbour (KNN) have been used to classify and diagnose stroke risk, where four values for the KNN ranging from 1 to 11 were used [5]. In this study 50 attributes, taken form potential stroke patients, were analysed. The attributes included various parameters that are directly related to stroke risk such as age, hypertension, sleep duration, smoking and alcohol consumption. Among employed machine learning algorithms, Decision Tree exhibited the highest accuracy followed by KNN of neighbours set to 1. However, the two models were trained on 807 samples and rises the uncertainty of models' efficiency on different datasets. Moreover, no information about the balanced data was explained in this study [6].

Furthermore, five machine learning models were trained on data collected from 425 patients to classify stroke disease [7]. In particular, Logistic Regression, Decision Tree, Support Vector Machine (SVM) and Random Forest were used for classification purposes by analysing 152 features. The findings showed that an accuracy of 90% can be achieved by Random Forest. Moreover, Naïve Bayes and Neural Network were also used to predict stroke risk by analysing imbalanced data [8]. Specifically, the data included 68,147 samples, where only around 0.4% of the participants have had the stroke and the rest are categorised as non-stroke participants. The study suggested reducing the number of non-stroke samples to 500 to eventually balancing the employed data. However, such balancing technique could compromise the models' accuracy, where the acquired accuracy was ranged from 72% - 75%. Furthermore, only six attributes

were analysed in this study including living province, marital status and education level [9].

Another study, done in Sudan, highlighted the importance of machine learning algorithms to classify stroke risk [10]. In this study, Decision tree and K-Nearest Neighbour (KNN) were used to classify 400 samples. The results showed that Decision Tree algorithm has outperformed the KNN algorithm and it is recommended to classify medical datasets. Moreover, a recent study showed that machine learning algorithms surpass deep neural network in terms of predicting stroke risk. Namely, the random forest algorithm showed the highest accuracy among various tested algorithms [11].

Although that the previous studies varied in their efficiency and accuracy in terms of predicting potential stroke, more information are required on imbalanced samples and the importance of involved attributes. This study analyses the efficacy of four popular machine learning algorithms in classifying stroke dataset, illustrating the capability of these models in diagnosing medical datasets and reducing false predictions.

The rest of the paper is arranged as follows: next section explains the materials and methods that were used in this study. Section 3 discusses the results and clarifies the differences between the models. Section 4 concludes the results and suggests a future work.

2. Materials and Methods

This section explains the steps that were followed to implement this study. Starting by collecting and pre-processing the data, then followed by explaining the employed machine learning models and validation process. Further details are clarified in the following paragraphs.

2.1 Data Collection and Pre-processing

The employed dataset involves information collected from 5110 patients, and lists the details of predictors that accompany the stroke. The dataset is characterised by providing 11 attributes or features and one outcome as follows [12]:

- 1- ID: Identification Number.
- 2- Gender: Male or Female, 59% of the participants are females, while 41% are males.
- 3- Age: Patients' Age that ranges from few months to 80 years.
- 4- Hypertension: "0" for having no hypertension, and "1" in case of hypertension existence.
- 5- Heart Disease: "0" for having no current heart diseases, and "1" in case of heart diseases existence.
- 6- Ever Married: refers to marriage status "No" or "Yes".
- 7- Work Type: that is either "government employing", "private employing", "never

- working”, or “child” for the young aged participant.
- 8- Residence Type: that is either “Rural” or “Urban”.
 - 9- Average Glucose Level: refers to the glucose level in participant’s blood.
 - 10- Body Mass Index (BMI): which is a value derived from the height and the weight of the participant.
 - 11- Smoking Status: refers to the participant’s status that is either “smokes”, “formerly smoked”, “never smoked”, and “unknown” in case of the unavailability information.
 - 12- Stroke: represents the outcome that is “0” for having no previous stroke, and “1” if the participant got stroke.

However, features of “ID”, “Work Type”, “Residence Type” and “Ever Married” were excluded from the training data as they have no effect on the likelihood of having the stroke or not [13]. Moreover, categorical features such as gender and smoking status were converted to numerical features to ensure it is suitability for most of machine learning models [14]. One-hot encoding labelling technique was used to convert gender and smoking status to numerical feature, such technique was chosen as it was previously proven to have promising results over other techniques [15].

Although, the information of 5110 patients have been collected for the purpose of stroke prediction, most of the participants have not suffered from stroke previously. Specifically, the dataset included the information of the 4861 participants of those who have not suffered from stroke, and 249 of those who had the stroke. Therefore, Synthetic Minority Over-sampling Technique (SMOTE) was used to address the unbalancing between the two classes. SMOTE create new samples based on the values of existing samples in the features space without duplicating the existed ones. Basically, the new sample is randomly taken from the line that connects two samples in the minority class as they are represented in the multidimensional space. SMOTE effectively increased the samples of the minority class (those who have obtained stroke in this study) to become similar to majority class (those who have not had stroke previously) [16].

2.2 Machine Learning Models

Four machine learning models were used to classify stroke datasets in terms of predicting whether the patient is likely to get stroke or not. Decision tree, Naïve Bayes, K-Nearest Neighbour (KNN), and Linear Discriminant Analysis (LDA) are the machine learning models that were trained on the above information to build a suitable mathematical model and ultimately predicting

potential stroke. Those algorithms were chosen due to their efficiency in classifying medical data [17][18]. The following paragraphs explains the four model in further details.

2.2.1 Decision Tree model

Decision Tree is a supervised machine learning model that separates the data into smaller subsets. This model is constructed from root node, branches and leaf nodes forming a hierarchical structure. The root node represents the tests on the features and branches represent the outcome of those tests, while the final classification is represented by the leaf node [19]. Decision Tree model selects features with highest information gain to ensuring maximised separation between different classes. Hence, the employed model has split the data four times using gini’s diversity index and was considered sufficient for the purpose of this study [20]. Moreover, the tree structure allows the model to capture various patterns makes it suitable for classification tasks.

2.2.2 Naïve Bayes model

Naïve Bayes model was chosen due to its efficiency and simplicity in various classification tasks. Furthermore, Naïve Bayes is well-known in handling high-dimensional data and effective in large features spaces. This model predicts the output by calculating the probability of each feature for a given class, then multiplies these individual probabilities and scales the output by overall class’s probability. Hence, the Naïve Bayes is known as a probabilistic classifier that relies on Bayes’ theorem. Kernel Density Estimation (KDE) was used in Naïve Bayes model to catch potential multiple peaks that could be existed in the employed data. Moreover, KDE allows the model to adapt to non-normal distributed data and eventually capture more complex relationships[17].

2.2.3 K- Nearest Neighbour (KNN) model

KNN is a simple and effective algorithm that is widely used for classification and regression tasks. Basically, KNN represents all data points as vectors in multidimensional space, and the classification is identified based on how far is the assigned point from the known points (figure 1). Euclidean technique was used to measure the distance between the points which was previously used and showed promised results. Furthermore, the value of K in the algorithm represents the number of points that need to be considered for classifying unseen data. In our model, one was chosen as the value of K to provide sensitive and accurate model that is able to classify new points in complex data. Moreover, the data was scaled to ensure accurate calculation for Euclidean distance for all contributed features [21].

2.2.4 Linear Discriminant Analysis (LDA) model

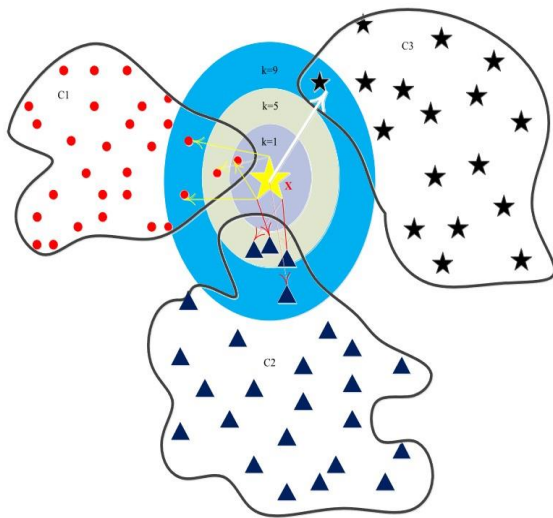


Figure 1. Concept of KNN Classification [22]

LDA is a popular linear classification technique that projects the data onto features space to separate between classes. LDA assumes that Gaussian distribution is applied to all classes in the dataset. It tries to find a linear relationship between the features that best separates the two appointed classes (stroke and non-stroke). Basically, the model calculates the probability of classifying assigned data to specific class [23].

2.3 Models' Validation

Cross-validation technique was used to evaluate the performance of the four machine learning models. Essentially, cross-validation divides the dataset into several equal subset that also known as folds, then the training and validation phases are applied to those folds. The training phase is repeated according to the number of folds while keeping one different fold for validation at each sequential training phase. Such technique reduces the chance of overfitting and generalizes the model to unseen, new data [24]. The employed data was divided to 5 folds, hence the four models were trained and validated 5 times using different fold for validation. Specifically, each training phase used 4 folds as training set and one fold as validation set. Following steps clarifies the process of 5-Fold Cross-Validation

- A- Dividing the Data: the data was divided into five equal or nearly equal subsets
- B- Training and Validation:

- First Iteration: the assigned model used folds 2, 3, 4, and 5 as training data while used fold 1, for validation.
 - Second Iteration: the assigned model used folds 1, 3, 4, and 5 as training data while used fold 2, for validation.
 - Third Iteration: the assigned model used folds 1, 2, 4, and 5 as training data while used fold 3, for validation.
 - Fourth Iteration: the assigned model used folds 1, 2, 3, and 5 as training data while used fold 4, for validation.
 - Third Iteration: the assigned model used folds 1, 2, 3, and 4 as training data while used fold 5, for validation.
- C- Metrics Calculation: the performance metrics are calculated for all iteration and averaged to evaluate model's performance.

Consequently, cross-validation use all data for training and validation phases reducing the risk of under-fitting and ensures the model's robustness.

2.4 Performance's Metrics

Accuracy, True Positive Rate (TPR), False Negative Rate (FNR), prediction speed, and training time were used to evaluate the four models. Firstly, accuracy represents the ratio of correct predictions to the total observations and can be expressed as follows:

$$\text{Accuracy} = \frac{\text{No.of Correct Predictions}}{\text{Total Observations}} \tag{1}$$

Secondly, TPR, also known as recall and sensitivity, represents the ability of the model to predict true positives to the sum of true positives and false negatives instances, and it is mathematically represented as follows

$$\text{True Positive Rate(TPR)} = \frac{\text{True Positive(TP)}}{\text{True positive(TP)+False Negative (FN)}} \tag{2}$$

Thirdly, FNR represents how often is the model incorrectly classify the positive instances as negative. Hence FNR is considered an important key metric when dealing with medical datasets, where failing to detect positive instances could lead to further serious consequences. FNR is mathematically represented as follows

$$\text{False Negative Rate(FNR)} = \frac{\text{False Negative (FN)}}{\text{True positive(TP)+False Negative (FN)}} \tag{3}$$

Prediction speed, the fourth metric, reflects how quickly the model predicts a new dataset, which is

important when dealing with large dataset. Finally, training time represents the duration in seconds that was needed by the model to train on the assigned data, this includes processing the data and building the model [25]. Next section explains the findings that were achieved using the four models and evaluates their performance.

3. Results and Discussion

In this study, the performance of four machine learning algorithms, Decision Tree, Naïve Bayes, KNN and LDA, was evaluated in terms of their ability to classify stroke patients. Accuracy, True Positive rate (TPR), False Negative Rate (FNR), prediction speed and training time were the key metrics that were used to assess the four models. Moreover, the 5-fold cross-validation technique was used to ensure that all data have been taken into account during training phase, and the obtained results are not dependent on a specific part of data.

Herein and after, first class refers to those who have not had stroke previously and is represented by “0”, while second class refers to those who have had stroke previously and is represented by “1”. Following paragraphs explain the findings of each model and their implications.

Decision tree achieved a validation accuracy of 79.8%, where the model poorly classified the two classes as shown in confusion matrix in figure 2. Moreover, a percentage of 67.9% and 91.8% were achieved as TPR's values for the two classes respectively. Lastly, high FNR was obtained for the first class with a value of 32.1%, which should be avoided when dealing with medical data. Similar Accuracy of around 79% was achieved by Naïve Bayes model, however this model showed slightly better performance in terms of TPR and FNR values for the first class. On other hand, LDA model exhibited the lowest accuracy, where a percentage of 72.3% was achieved. Furthermore, undesired values for the TPR and FNR were obtained using LDA, which proves that LDA is not recommended for classifying such data. Finally, an accuracy of 90.5% was achieved using KNN model to classify stroke patients' data. Moreover, KNN performed relatively better in terms of TPR and FNR for the both classes when compared to the three models as shown in figure 2.

Additionally, table 1 was suggested to clarify the differences between the four models using the key metrics elaborated in previous section. It is clearly shown that KNN model performed well over the rest of models, where an accuracy of 90.5 was achieved. Decision tree and Naïve Bayes models came in the second place with an accuracy of 79.8% and 79.4% respectively, while the lowest accuracy was achieved by LDA model. In respect to prediction speed, LDA model outperformed the rest

with speed of over 400,000 predictions per second, and the lowest speed was achieved by Naïve Bayes model. Moreover, KNN model showed an acceptable prediction speed of 46,000 observations per second. Training time is the third metric that was calculated to evaluate the models' performance. Again, KNN and LDA models exhibited the lowest training time of around 1 – 2 seconds. However, the highest training time was taken by Naïve Bayes model and this was expected as the classification process relies on the probabilities. True positive rate (TPR) varied dramatically for the four models and for the two classes. For instance, TPR is high for the second class when using Decision Tree model and significantly low for the first class using the same model. Moreover, Low TPR was achieved for the first and second class using Naïve Bayes and LDA. The last metric, False Negative Rate (FNR), is considered a substantial metric especially when classifying medical data. Similar to TPR, FNR varied between the classes and the models, where lowest FNR was achieved for the first class using KNN. The highest FNR, was obtained using LDA when classifying second class, and such rate should be avoided in medical aspects. In general, KNN showed relatively lowest FNR when compared to the rest of models.

The superior performance of the KNN model over the employed models belongs to its sensitivity and simplicity to structured data. Moreover, the findings suggest that KNN is recommended for medical diagnoses, where the failed detection of serious diseases leads to further severe consequences. To conclude, KNN model was identified as the best model to predict stroke risk due to its outperformance over the Decision Tree, Naïve Bayes and LDA models.

4. Conclusion and Future Works

Four machine learning models that are Decision Tree, Naïve Bayes, LDA and KNN, were used to classify stroke patients' data. Five metrics, accuracy, prediction speed, training time, TPR and FNR, were used to assess the models' performance. The findings showed that KNN achieved an accuracy of 90% making it outperformed the rest of the models, while lowest accuracy was obtained using LDA. Moreover, KNN model showed high prediction speed and low training time, making it suitable for medical data classification.

Future works could involve applying the four machine learning models to different datasets and explore the differences.

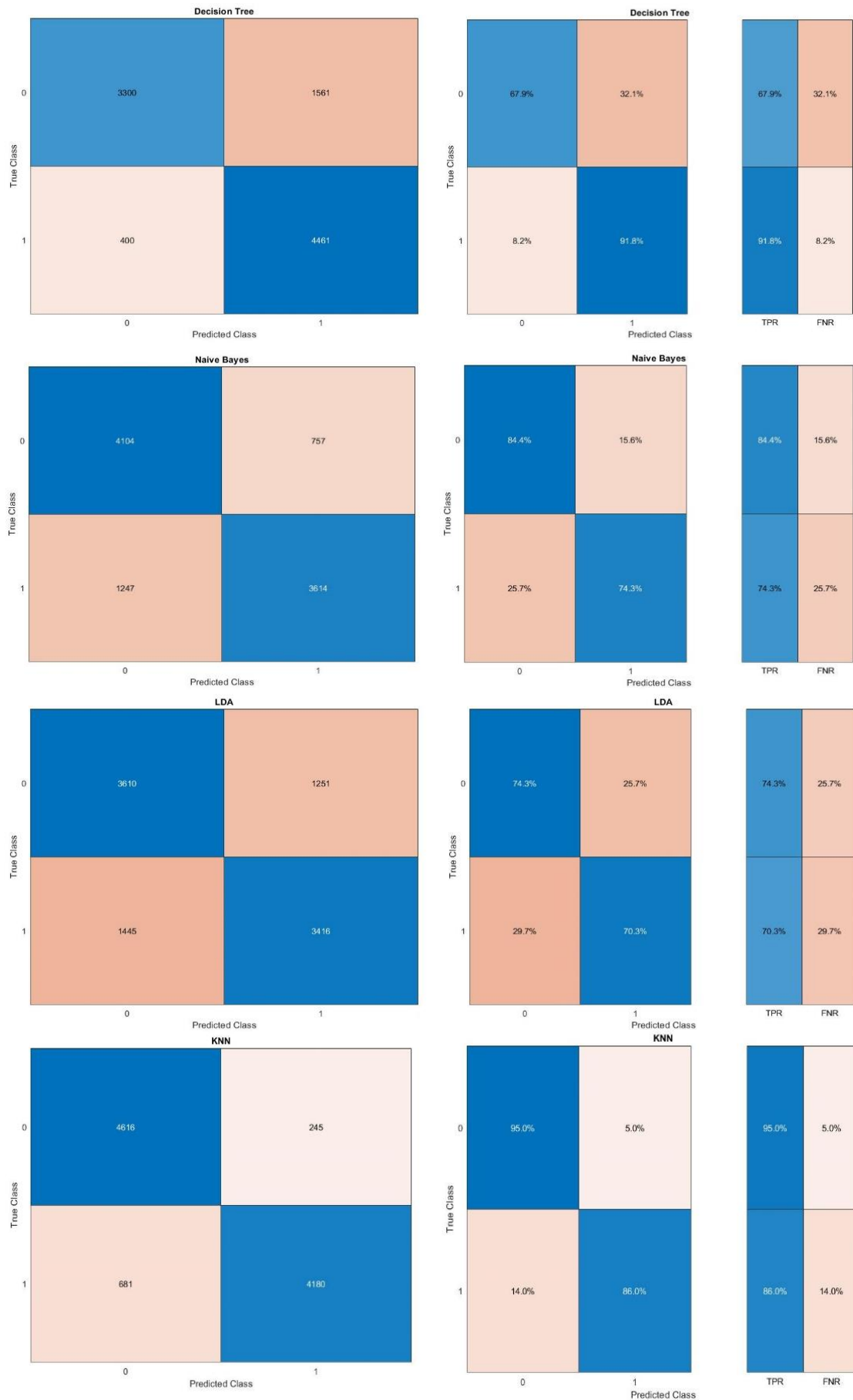


Figure 2. Shows the Confusion Matrices, TPR and FNR of the Four Models.

Table 1. Shows the Five Performance Metrics that were used to evaluate the Four Models.

Model	Accuracy (%)	Prediction Speed (Observations/Second)	Training Time (Seconds)	TPR (%)		FNR (%)	
				1 st Class	2 nd Class	1 st Class	2 nd Class
Decision Tree	79.8	240,000	4.23	67.9	91.8	32.1	8.2
Naïve Bayes	79.4	820	53.5	84.4	74.3	15.6	25.7
LDA	72.3	410,000	0.96	74.3	70.3	25.7	29.7
KNN	90.5	46,000	2.0	95.0	86.0	5.0	14.0

References

- [1] Sirsat MS, Fermé E, Câmara J. Machine Learning for Brain Stroke: A Review. *J. Stroke Cerebrovasc. Dis.* 2020.
- [2] Shehab M, Abualigah L, Shambour Q, et al. Machine learning in medical applications: A review of state-of-the-art methods. *Comput. Biol. Med.* 2022.
- [3] Yousef Shaheen M. Adoption of machine learning for medical diagnosis Adoption of machine learning for medical diagnosis [Internet]. *Sci. Prepr.* 2021. p. 0–2. Available from: <https://www.scienceopen.com/>.
- [4] Al-Mekhlafi ZG, Senan EM, Rassem TH, et al. Deep Learning and Machine Learning for Early Detection of Stroke and Haemorrhage. *Comput. Mater. Contin.* 2022. p. 775–796.
- [5] Amini L, Azarpazhouh R, Farzadfar MT, et al. Prediction and control of stroke by data mining. *Int. J. Prev. Med.* 2013. p. S245–S249.
- [6] Tyagi S, Mittal S. Sampling approaches for imbalanced data classification problem in machine learning. *Lect. Notes Electr. Eng.* 2020. p. 209–221.
- [7] Monteiro M, Fonseca AC, Freitas AT, et al. Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 2018. p. 1953–1959.
- [8] Kansadub T, Thammaboosadee S, Kiattisin S, et al. Stroke risk prediction model based on demographic data. *BMEiCON 2015 - 8th Biomed. Eng. Int. Conf.* 2016.
- [9] Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Comput. Surv.* 2019.
- [10] Yahiya S, Yousif A, Bakri M. Classification of Ischemic Stroke using Machine Learning Algorithms. *Int. J. Comput. Appl.* 2016. p. 26–31.
- [11] Rahman S, Hasan M, Sarkar AK. Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques. *Eur. J. Electr. Eng. Comput. Sci.* 2023. p. 23–30.
- [12] Fedesoriano. Stroke Prediction Dataset [Internet]. 2020. Available from: <https://www.kaggle.com/datasets/fedesoria/stroke-prediction-dataset/data>.
- [13] Zhu C, Tran PM, Leifheit EC, et al. The association of marital/partner status with patient-reported health outcomes following acute myocardial infarction or stroke: Protocol for a systematic review and meta-analysis. *PLoS One.* 2022.
- [14] Liudmila P, Gleb G, Aleksandr V, et al. Catboost: unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* 2018. p. 6638–6648.
- [15] John T. Hancock* and Taghi M. Khoshgoftaar. Survey on categorical data for neural networks | Enhanced Reader. *J Big Data* [Internet]. 2020;7. Available from: moz-extension://170f95a9-cae2-4bce-87af-8558f3cef2be/enhanced-reader.html?openApp&pdf=https%3A%2F%2Fjournalofbigdata.springeropen.com%2Fcounter%2Fpdf%2F10.1186%2F40537-020-00305-w.pdf.
- [16] Joloudari JH, Marefat A, Nematollahi MA, et al. Effective Class-Imbalance Learning Based on SMOTE and Convolutional

- Neural Networks. Appl. Sci. 2023.
- [17] Uddin S, Khan A, Hossain ME, et al. Comparing different supervised machine learning algorithms for disease prediction. BMC Med. Inform. Decis. Mak. 2019.
- [18] Jawalkar AP, Swetcha P, Manasvi N, et al. Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting. J. Eng. Appl. Sci. 2023.
- [19] Quinlan JR. Induction of Decision Trees. 2007;81–106.
- [20] Louppe G. Understanding Random Forests: From Theory to Practice [Internet]. 2014. Available from: <http://arxiv.org/abs/1407.7502>.
- [21] Ehsani R, Drabløs F. Robust Distance Measures for kNN Classification of Cancer Data. Cancer Inform. 2020.
- [22] Xing W, Bei Y. Medical Health Big Data Classification Based on KNN Classification Algorithm. IEEE Access. 2020. p. 28808–28819.
- [23] Adebisi MO, Arowolo MO, Mshelia MD, et al. A Linear Discriminant Analysis and Classification Model for Breast Cancer Diagnosis. Appl. Sci. 2022.
- [24] Cherradi B, Terrada O, Ouhmida A, et al. Computer-Aided Diagnosis System for Early Prediction of Atherosclerosis using Machine Learning and K-fold cross-validation. 2021 Int. Congr. Adv. Technol. Eng. ICOTEN 2021. 2021.
- [25] Huang C, Li SX, Caraballo C, et al. Performance Metrics for the Comparative Analysis of Clinical Risk Prediction Models Employing Machine Learning. Circ. Cardiovasc. Qual. Outcomes. 2021. p. E007526.