Publications

Available Online

J. Sci. Res. **16** (3), 637-646 (2024)

JOURNAL OF
SCIENTIFIC RESEARCH

www.banglajol.info/index.php/JSR

Leveraging AdaBoost and CatBoost to Classify the Likelihood of Brain Stroke

P. Nandal[1]*, S. Malik[2]

[1]Department of Computer Science and Engineering, Maharaja Surajmal Institute of Technology, New Delhi, Delhi 110058, India

[2]Department of Information and Technology, Maharaja Surajmal Institute of Technology, New Delhi, Delhi 110058, India

Received 28 July 2023, accepted in final revised form 26 May 2024

Abstract

Brain Stroke occurs when the blood flow to a portion of the brain is reduced or stopped, denying the brain's tissue nourishment and oxygen, which results in brain cell death. Many lives can be saved by early diagnosis, but the bulk of clinical datasets, including the stroke dataset, are unbalanced, which means that the majority of predictive algorithms are biased. By balancing the dataset, resampling methods improve machine learning algorithms' capacity for prediction. This study compares various algorithms on a stroke dataset to determine the likelihood of experiencing a stroke. In order to predict stroke, the authors of this work used two machine learning classifiers, AdaBoost and CatBoost, in conjunction with a well-known resampling technique called Synthetic Minority Oversampling Technique (SMOTE). A publicly available dataset was employed for the study. CatBoost outperformed AdaBoost and achieved an accuracy of 96 % when combined with SMOTE. The accuracy achieved using CatBoost was better than that of most previously developed models and is on par with other advanced models.

Keywords: Brainstroke; AdaBoost; CatBoost.

doi: https://dx.doi.org/10.3329/jsr.v16i3.67891 J. Sci. Res. **16** (3), 637-646 (2024)

1. Introduction

The two types of strokes are ischemic stroke and hemorrhagic stroke. The majority of strokes are ischemic in nature. When blood clots or other objects obstruct the blood vessels that supply the brain, an ischemic stroke occurs, according to the CDC (Centre for Disease Control and Prevention). When a blood artery ruptures or experiences pressure from bleeding, it suffers a hemorrhagic stroke. More than 12 million people are affected by stroke every year, which is a significant reason for disability and mortality around the globe [1]. The incident rate of brain stroke in India ranged from 105-152 per 100,000 individuals each year in the past decade [2]. Between 1990 and 2020, incidents of brain strokes in India have

[*] *Corresponding author*: priyankanandal@msit.in

increased by over a whopping 50 % [3]. The malady of stroke is getting worse in India, where, at present, it is the fourth leading cause of death [4]. Early detection and prevention of stroke are critical to reducing its impact because prompt treatment can greatly enhance results and lower the chance of long-term impairment [5]. One of the main risk factors for stroke is high blood pressure. A prior transient ischemic attack (TIA), high blood cholesterol, diabetes mellitus, smoking, obesity, and atrial fibrillation are additional risk factors [6]. Though there are other, less frequent causes of ischemic stroke, clogged blood arteries are typically the culprit. Stroke prediction can help to diagnose or anticipate a stroke early by identifying individuals who are at high risk for having a stroke. This can allow them to take preventive measures, such as lifestyle changes or medical treatment, to lessen their threat of bearing a stroke. Hours or days before the onset of a stroke, warning signals such as cognitive impairments, emotional control problems, and sadness may show up. We can identify the stroke using these symptoms and administer emergency care to prevent catastrophic brain damage. Therefore, by employing past knowledge of risk factors, we can forecast the onset of a stroke and lessen its effects. Early diagnosis and treatment of a stroke can also improve the chances of a full recovery and even prevent stroke in some cases. One side of the body may become weak or numb, speech or language difficulties may arise, sudden disorientation or trouble thinking, and sudden vision problems in one or both eyes are all signs of a stroke [7].

Machine learning (ML) applications are having a major impact on healthcare. ML is a type of artificial intelligence (AI) technology aimed at improving the speed and accuracy of doctors' work. AI shows promise in countries with overburdened health systems that currently lack qualified doctors. These technologies enhance quality of life by facilitating early diagnosis and enhancing care. In stroke care, a variety of AI applications are used, including decision support, early detection, and accurate diagnosis. In addition, compared to conventional statistical inference techniques, deep learning (DL) and ML can produce more effective and precise predictions.

In light of this, this research aims to develop a machine learning-based model for assessing the risk of stroke in citizens. This will be accomplished by employing a variety of techniques to identify the contributing factors linked to stroke and then suggesting an integrated model to evaluate the risk of stroke.

The paper is structured as follows: Section 2 investigates the work done thus far on this issue and what has already been accomplished. Then, Section 3 explains the techniques and materials used by authors in the work. Section 4 then establishes the result findings. The conclusion is laid out in Section 5.

2. Literature Review

The use of machine learning algorithms to predict and categorize the likelihood of a brain stroke in humans has been extensively studied. Previous studies [8] used machine learning algorithms to classify stroke disease and predicted it with an accuracy of 96 % by focusing on Artificial Neural Networks. Singh and Chaudhary [9], based on a patient's risk characteristics, employed AI algorithms to forecast the possibility of a stroke happening to them. Principal Component analysis was used [10,11] to analyze the interdependence of the factors involved in the risk of stroke found in patient electronic health records. Nwosu *et al*. [10] concluded that all patient features may be used for stroke prediction because their analysis revealed that patient attributes were not strongly associated. The model performs better and takes less time to train when the feature subspace is reduced.

Work type, hypertension, average glucose level, heart disease, age, and ever-married were the only features left after utilizing statistical techniques like chi-squared, resulting in a performance accuracy of 96.8 % using a two-class boosted decision tree model [12]. Another study also used correlation analysis [11], and stepwise analysis was used to choose the best set of features. Yet another study [13] employed sampling techniques such as the Random sampling Technique (RUS), Random Oversampling Technique (ROS), and Synthetic Minority oversampling Technique (SMOTE) for stroke prediction in the development of machine learning models using unbalanced data, machine learning models for stroke prediction. In comparison, they found SMOTE successfully produced balanced results for Random Forest, which had an accuracy performance of 78 %.

Stroke prediction was created using the Naive Bayes, support vector machine, K-Nearest Neighbour, Decision Tree, Random Forest, Logistic Regression, neural network, XGBoost, SMOTE Technique, and as well as the cross-industry standard process for data mining (CRISP-DM) as a guide [14]. The records used for this purpose were 5110 in number with extremely unbalanced data problems.

The researchers found that Random Forest, with a 92 % accuracy rate, is the best model with fewer classification errors than contrasting algorithms and that there is a substantial likelihood that someone may experience a stroke if hypertension and heart disease are also present. Tazin *et al*. [15] used decision trees, voting classifiers, random forest, and logistic regression learning methods for brain stroke prediction, and they achieved 96 % accuracy with Random Forest classification. Chen *et al.* [16] recently used an ANN model to predict a stroke's readmission after 30 days. A number of studies have used gradient-boosting classifiers for brain stroke prediction. One example is a study by Zhang *et al*. [17], which used a gradient-boosting classifier to predict the likelihood of intracerebral hemorrhage (a type of stroke caused by bleeding in the brain) based on various clinical and imaging features. They found that the model was able to achieve an overall accuracy of 89 % in predicting stroke risk. The authors found that the classifier was able to achieve high levels of accuracy and outperformed other machine learning algorithms that were tested. Another example is a study by Li *et al*. [18], which determined the chance of stroke in people with atrial fibrillation (a type of irregular heartbeat) using a gradient-boosting classifier. According to the authors, the classifier could find high-risk individuals who would benefit from preventative care and achieve good performance. A study by Alanazi *et al*. using a neural network found that it could predict stroke with an AUC (area under the curve) of 0.84 in the primary care population from lab tests [19]. A study using support vector machines found that it could predict stroke with an AUC of 0.93 in the hospital population [20].

There are many studies using machine learning algorithms such as AdaBoost and CatBoost to predict stroke. These algorithms have proven effective in a variety of settings, including primary care clinics and hospitals. AdaBoost proved successful in reliably predicting stroke risk in a study of stroke prediction in the primary care population [21] with an AUC of 0.78. Another study found that AdaBoost could predict stroke with an AUC of 0.92 in a hospital population [22]. CatBoost has also been shown to be effective in predicting stroke. In a study of stroke prediction in a hospital population in which 4530 patients were involved, CatBoost was able to accurately predict stroke with an AUC of 0.833 [23]. These and other studies demonstrated the effectiveness of AdaBoost and CatBoost in building stroke prediction models. Based on the literature study of the most current techniques, we can conclude that none of them accurately classified stroke. Therefore, further improvement is required in the current techniques used for the prediction of stroke.

3. Materials and Methods

This section describes the dataset, methodology employed, pre-processing, and proposed algorithms. Fig. 1 gives a description of the process.

3.1. *Methodology*

In this study, using an open-access dataset for stroke prediction, the authors suggested AdaBoost and CatBoost for predicting cerebral stroke. The pre-processing of the dataset included handling missing values, removing outliers, utilizing the one-hot encoding technique, and normalizing the features using various value ranges. To assess the AdaBoost and CatBoost models, the authors employed the ten-fold cross-validation technique, which utilized more than one train-test split of the data. One-fold serves as the test set in the end, and nine-folds serve as the train sets in the ten-fold cross-validation. The authors used the SMOTE on the training set to balance the samples of the two classes after dividing the dataset. Finally, the authors used the test set to calculate the evaluation metrics and assess the performance of the suggested tuning ensemble.

3.2. *Dataset*

The investigation was conducted using the stroke prediction dataset. There are 5110 rows and 12 columns in this dataset. The output column stroke is represented by one of two numbers, 1 or 0 respectively. When the value was 1, stroke risk was recognized; when it was 0, no stroke risk was noticed. The probability of the output column (stroke) in this dataset being 0 is higher than the probability of the same column being 1. Only 249 rows have the value 1 in the stroke column, whereas 4861 rows have 0. Data preparation was employed to balance the data and boost accuracy.

Three of the 11 columns are numeric, while the remaining eight are categorical, making one hot encoding or label encoding an effective pre-processing technique to prepare the model for categorizing future input.

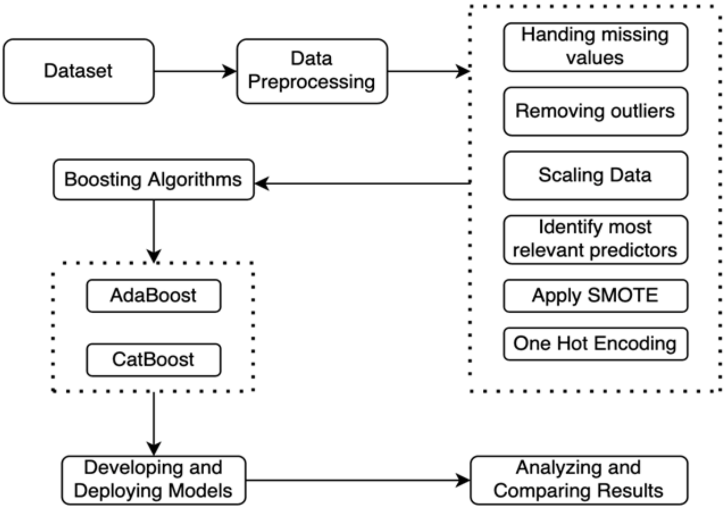


Fig. 1. Workflow of the process.

Table 1. The variables and their type.

Variable Name	Variable Type
Gender	Categorial
Age	Numeric
Hypertension	Categorial
Heart_disease	Categorial
Work_type	Categorial
Avg_glocose_level	Numeric
BMI	Numeric
Smoking_status	Categorial
Ever_married	Categorial
Residence_type	Categorial
Stroke	Categorial

3.2. *Data pre-processing*

Data pre-processing is necessary before training and evaluating the models since machine learning techniques are data-based, and it helps the models perform better. The subsequent actions are taken into account during the pre-processing stage. The missing values in categorial variables were filled using the mode, while continuous variables were filled using linear imputation. Label Encoder was used for ordinal features like ever_married and residence type to transform categorical features into numerical values, and One Hot Encoder was used for nominal categories like work_type and smoking_status to do the same for ordinal features like ever_married and residence_type. Feature scaling was implemented using Z-score normalization, which enables features with extremely disparate ranges of values to have comparable ranges of values. SMOTE (Synthetic Minority Oversampling Technique) is a technique utilized to address the problem of imbalanced datasets in machine learning. In unbalanced datasets, one type of data is underrepresented compared to another. Machine learning models may be more likely to anticipate the majority class and less likely to correctly forecast the minority class as a result, which might be an issue when training them. SMOTE creates new minority-class data samples based on existing minority-class data samples. It does this by choosing a sample of data from a minority class and locating its close neighbours. Then, a fresh data sample is generated by randomly choosing one of the closest neighbours and slightly altering the feature values. The minority class and the dominant class are then balanced by repeating this procedure.

3.4. *Algorithms used*

3.4.1. *AdaBoost*

AdaBoost is used to classify problems like prognosis and diagnosis in medicine. It works by combining a collection of weak classifiers, or models, into a powerful classifier that can accurately anticipate desired outcomes. AdaBoost has shown effective results in a variety of medical applications, including the prediction of heart disease and breast cancer recurrence.

3.4.2. *CatBoost*

A machine learning technique called CatBoost was created especially for categorical data. It is a variation of the popular machine learning algorithm AdaBoost, which is used in classification tasks, including medical diagnosis and prognosis. CatBoost creates a powerful classifier that can precisely predict desired outcomes by merging a number of weak classifiers, or models, that perform somewhat better than chance. The algorithm is shown successful in a variety of medical applications, such as the prediction of heart disease.

3.5. *Evaluation metrics*

In order to determine the efficacy of the machine learning algorithms implemented in this study, we used the assessment metrics of precision, recall, f1- score, and support. "Support" in a classification report refers to the number of samples or instances for each class. It provides information on how often or widely data points in a classification task are distributed among several classes.
Accuracy is the fraction of prediction that is predicted correctly.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ positive + True\ Negative + False\ Negative} \tag{1}$$

Precision is the percentage that demonstrates the model's capacity to avoid classifying negative information as good.

$$Precision = \frac{True\ Positive}{True\ Positive+False\ positive} \quad (2)$$

The recall is the percentage that indicates the model's capacity to categorize all of the positive samples.

$$Recall = \frac{True\ Positive}{True\ Positive+False\ Negative} \quad (3)$$

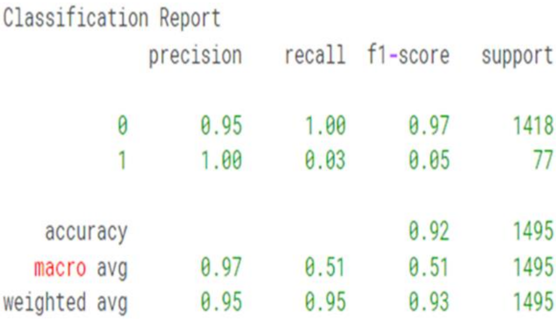
The weighted harmonic mean of recall and precision is known as the F1-score.

$$F1\_score = 2\frac{precision*recall}{precision+recall} \quad (4)$$

4. Results Discussion

4.1. *Results*

The open-source Kaggle dataset is used to predict strokes using the CatBoost and AdaBoost classifiers. The performance outcomes of the models are shown in the table. According to Table II, the model created with CatBoost had the highest accuracy, f1-score, precision, and recall, with a score of 98.23 %. AdaBoost has an accuracy score of 98.07 %.

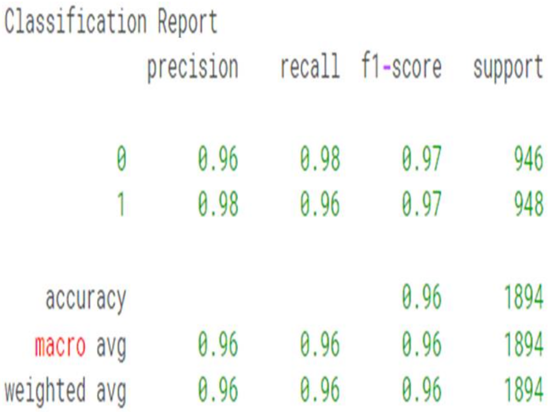


```
Classification Report
              precision    recall  f1-score   support

           0       0.95      1.00      0.97      1418
           1       1.00      0.03      0.05        77

    accuracy                           0.92      1495
   macro avg       0.97      0.51      0.51      1495
weighted avg       0.95      0.95      0.93      1495
```

Fig. 2. Classification report of CatBoost without SMOTE.

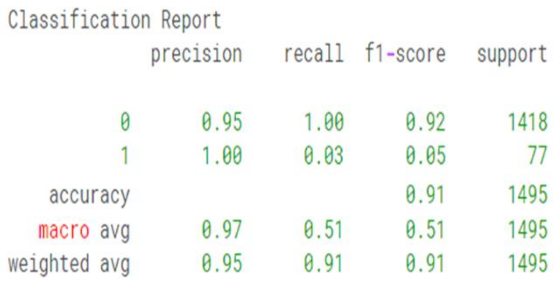


```
Classification Report
              precision    recall  f1-score   support

           0       0.96      0.98      0.97       946
           1       0.98      0.96      0.97       948

    accuracy                           0.96      1894
   macro avg       0.96      0.96      0.96      1894
weighted avg       0.96      0.96      0.96      1894
```

Fig. 3. Classification report of CatBoost with SMOTE.

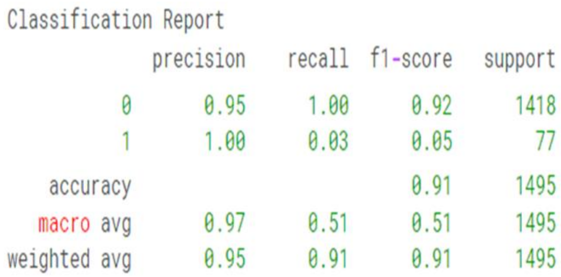


```
Classification Report
                  precision    recall  f1-score   support

             0         0.95      1.00      0.92      1418
             1         1.00      0.03      0.05        77

      accuracy                             0.91      1495
     macro avg         0.97      0.51      0.51      1495
  weighted avg         0.95      0.91      0.91      1495
```

Fig. 4. Classification report of AdaBoost without SMOTE.



```
Classification Report
                  precision    recall  f1-score   support

             0         0.95      1.00      0.92      1418
             1         1.00      0.03      0.05        77

      accuracy                             0.91      1495
     macro avg         0.97      0.51      0.51      1495
  weighted avg         0.95      0.91      0.91      1495
```

Fig. 5. Classification report of AdaBoost with SMOTE.

4.2. *Comparison with other algorithms*

Table 2 shows the comparison with previous algorithms. The performance of the training models is judged by accuracy, f1-score, precision, and recall.

Table 2. Comparison with previous algorithms.

Model	Acccuracy	F1-score	Precision	Recall
Logistic Regression	0.72	0.81	0.96	0.72
KNN Classifier	0.72	0.81	0.94	0.72
Naïve Nayes	0.58	0.70	0.96	0.58
SVM	0.71	0.80	0.95	0.71
AdaBoost	0.85	0.83	0.82	0.85
CatBoost	0.92	0.93	0.95	0.95
AdaBoost (SMOTE)	0.91	0.91	0.95	0.91
CatBoost (SMOTE)	0.96	0.96	0.96	0.96

5. Conclusion

Using AdaBoost and CatBoost algorithms; the authors suggested a stroke prediction method for this study. A number of ML techniques, including logistic regression, KNN classifier, naive Bayes, and SVM, were compared to the suggested method. The contrasted models were assessed using different performance indicators, including accuracy, precision, recall, and f1-score. It was discovered that the suggested method could handle a sizable dataset and offer a sizable improvement in accuracy over that of existing methods.

The proposed method's susceptibility to noise is its only drawback. Thus, in the future, we intend to improve the data analysis process by introducing an effective noise removal method. One of the main advantages of AdaBoost and CatBoost is that they are relatively easy to implement and require less hyperparameter tuning. These algorithms can also process large amounts of data and handle high-dimensional input spaces well. Using AdaBoost, we were able to achieve an accuracy of 85 % without SMOTE, and using SMOTE, we achieved an accuracy of 91 %. On the other hand, employing CatBoost resulted in an accuracy of 92 % without SMOTE and 96 % with SMOTE. This implies that CatBoost is the better choice of algorithm for stroke prediction. This accuracy achieved is better than many other algorithms and is on par with other top algorithms employed for classifying the likelihood of brain stroke prediction. The model's performance can be improved in the future by incorporating additional data sources such as imaging or genetic data. It would also be useful to evaluate the generalizability of the AdaBoost and CatBoost models to different populations and settings in order to determine their potential for use in a wider range of contexts.

References

1. A. C. Fonseca and S. I. Savitz, Organizational Update - *World Stroke Conf. 2021* (2022) https://doi.org/10.1161/STROKEAHA.122.038782
2. S. Kamalakannan, A. S. Gudlavalleti, V. S. M. Gudlavalleti, S. Goenka and H. Kuper, Ind. J. Med. Res. **146**, 2 (2017) https://doi.org/10.4103/ijmr.IJMR_516_15
3. S. P. Jones, K. Baqai, A. Clegg, R. Georgiou, C. Harris et al., Int. J. Stroke. **17**, 2 (2022). https://doi.org/10.1177/17474930211027834
4. J. Kim, T. Thayabaranathan, G. A. Donnan, G. Howard, V. J. Howard et al., Int. J. Stroke. **15**, 8 (2020). https://doi.org/10.1177/1747493020909545
5. L. L. Yan, C. Li, J. Chen, J. J. Miranda, R. Luo et al., Eneurologicalsci. **2**, 21 (2016). https://doi.org/10.1016/j.ensci.2016.02.011
6. A. K. Boehme, C. Esenwa, and M. S. Elkind, Circulation Res. **120**, 3 (2017) https://doi.org/10.1161/CIRCRESAHA.116.308398
7. W. Syed, O. A. Qadhi, A. Barasheed, E. AlZahrani, and M. B. A. Al-Rawi, Front. Public Health **11** (2023). https://doi.org/10.3389/fpubh.2023.1131110
8. R. Jothiramalingam, A. Jude, R. Patan, M. Ramachandran, J. H. Duraisamy, and A. H. Gandomi, Neural Comput. Appl. **33**, 4445 (2021). https://doi.org/10.1007/s00521-020-05238-2
9. M. S. Singh and P. Choudhary - *Annual Industrial Automation Electromechanical Eng. Conf.* (2017). https://doi.org/10.1109/IEMECON.2017.8079581
10. C. S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli, and D. John - *Annual Int. Conf. of the IEEE Eng. in Medicine and Biology Society (EMBC)* (Berlin, Germany, 2019). https://doi.org/10.1109/EMBC.2019.8857234
11. S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, Healthcare Analytics **2**, ID 100032 (2022). https://doi.org/10.1016/j.health.2022.100032
12. S. Ray, K. Alshouiliy, A. Roy, A. AlGhamdi, and D. P. Agrawal, *Intermountain Engineering, Technology and Computing (IETC)* (Orem, UT, USA, 2020), pp. 1-6. https://doi.org/10.1109/IETC47856.2020.9249117
13. Y. Wu and Y. Fang, Int. J. Environ. Res. Public Health **17**, 6 (2020). https://doi.org/10.3390/ijerph17061828

14. S. Gupta and S. Raheja – *Int. Conf. on Cloud Computing, Data Science & Eng.* (Noida, India, 2022). https://doi.org/10.1109/Confluence52989.2022.9734197
15. T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. M. Khan, J. Healthcare Eng. **2021**, ID 7633381 (2021). https://doi.org/10.1155/2021/7633381
16. Y. C. Chen, J. H. Chung, Y. J. Yeh, S. J. Lou, H. F. Lin et al., Front. Neurol. **13**, ID 875491 (2022). https://doi.org/10.3389/fneur.2022.875491
17. X. Xu, J. Zhang, K. Yang, Q. Wang, X. Chen, and B. Xu, Brain Behavior **11**, 5 (2021). https://doi.org/10.1002/brb3.2085
18. C. Jiang, T. -G. Chen, X. Du, X. Li, L. He et al., Chinese Med. J. **134**, 19 (2021). https://doi.org/10.1097/cm9.0000000000001515
19. G. C. O'Connell, K. B. Walsh, C. G. Smothers, S. Ruksakulpiwat, B. L. Armentrout et al., BMC Neurol. **22**, 1 (2022). https://doi.org/10.1186/s12883-022-02726-x
20. D. Park, E. Jeong, H. Kim, H. W. Pyun, H. Kim et al., Diagnostics **11**, 10 (2021). https://doi.org/10.3390/diagnostics11101909
21. E. M. Alanazi, A. Abdou, and J. Luo, JMIR Formative Res. **5**, 12 (2021). https://doi.org/10.2196/23440
22. C. Rao, M. Li, T. Huang, and F. Li, Comput. Model. Eng. Sci. **139**, 699 (2024). https://doi.org/10.32604/cmes.2023.044898
23. Q. Wang, J. Yin, L. Xu, J. Lu, J. Chen et al., Neurol. Sci. (2024). https://doi.org/10.1007/s10072-024-07329-7