A Robust Methodology for Stroke Disease Prediction using a Hard Voting Classifier

Nouralhuda Ali Abdulsamad
Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah 61004, Iraq
pgs.nooralhuda.alii@uobasrah.edu.iq

Ali A.Yassin
Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah 61004, Iraq
ali.yassin@uobasrah.edu.iq (corresponding author)

Zaid Ameen Abduljabbar
Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah 61004, Iraq | Department of Business Management, Al-Imam University College, Balad 34011, Iraq | Shenzhen Institute, Huazhong University of Science and Technology, Shenzhen 518000, China
zaid.ameen@uobasrah.edu.iq

Mohammed S. Hashim
Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah 61004, Iraq
moh.salah@uobasrah.edu.iq

Vincent Omollo Nyangaresi
Department of Computer Science and Software Engineering, Jaramogi Oginga Odinga University of Science and Technology, Bondo 40601, Kenya | Department of Applied Electronics, Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu 602105, India
vnyangaresi@jooust.ac.ke

Received: 19 January 2025 | Revised: 20 February 2025 and 09 March 2025 | Accepted: 13 March 2025

 DOI: https://doi.org/10.48084/etasr.10292

ABSTRACT

The vast majority of strokes are caused by an unexpected occlusion of the blood vessels that supply the brain and the heart arteries. Early detection of the many warning symptoms of stroke can help reduce the severity of the stroke and save the patient's life. Although researchers have proposed a variety of diagnostic methods to detect this disease, the methods currently in use still need further improvement. In this paper, we propose an effective methodology that utilizes a hard voting classifier based on three Machine Learning (ML) models, namely, Random Forest (RF), K-Nearest Neighbors (KNN), and Extra Trees Classifier (ETC). First, a series of data quality improvement procedures were performed using the Synthetic Minority Oversampling Technique (SMOTE) approach for data balancing to ensure an unbiased training process without majority class dominance. Next, we divided the dataset into two parts, a training part and a testing part, and these data were fed to the models used. In the last phase, we implemented four ML algorithms to evaluate their effectiveness and then selected the three most effective models for integration into our proposed hard voting classifier. The hard voting outperformed the results of modern studies with an accuracy of 97.48%, a precision of 0.9802, a recall of 0.9691, and an F1 score of 0.9747. Furthermore, we applied K-fold cross-validation (K=10), which systematically partitions the dataset into multiple subsets, preventing overfitting and providing a robust estimate of model performance across different data splits, where a mean accuracy of 97.1% was achieved.

Keywords-stroke disease; machine learning; prediction; hard voting classifier; SMOTE; cross-validation

I. INTRODUCTION

A stroke, sometimes referred to as a brain attack, occurs when a cerebral blood vessel ruptures or when the blood supply to a particular region of the brain is interrupted [1]. When blood flow is interrupted, brain cells die rapidly due to lack of oxygen, resulting in a stroke. The consequences can include permanent disability, irreversible brain damage, or even death [2]. Stroke is the second leading cause of death and the leading cause of disability worldwide. According to the 2022 global stroke factsheet, the lifetime risk of stroke has increased by 50% in the last 17 years, and 1 in 4 people are expected to have a stroke in their lifetime [3]. Before the advent of modern medical technology, stroke detection relied primarily on clinical observation, patient history, and basic physical examination. These traditional methods were based on recognizing the outward symptoms of stroke, and diagnosis depended heavily on the skill and experience of healthcare providers. Today, information technology plays a critical role in the medical field, helping to improve diagnosis, patient care, and the overall efficiency of healthcare services. It is essential to improve the speed, accuracy, and efficiency of medical diagnoses, leading to satisfactory patient outcomes, improved healthcare delivery, and reduced costs [4].

Stroke can have a significant psychological impact on a patient, in addition to physical and cognitive impairments. Early prediction of the possible occurrence of a stroke is critical to prevent it through early medical intervention and appropriate treatment [5]. AI plays a large and important role in the early diagnosis of stroke [6]; it uses a set of different algorithms that learn about the patient's condition and train on it to predict future cases [7]. In medicine, Machine Learning (ML) methods for classification and prediction are most commonly used, especially on stroke-related datasets, to determine whether a case is positive or negative [8]. Extensive research has been conducted in the area of early stroke detection utilizing AI models. Most studies used ML models, whereas others implemented deep learning models. Researchers have also used an ensemble classifier that integrates multiple models. Our research will address the limitations in the accuracy of stroke diagnosis and prediction.

A. Study Contributions

The main contributions of this study are:

- The analysis of the dataset provided the basis for performing preprocessing to improve the quality of the data for stroke prediction.
- The dataset was balanced using the Synthetic Minority Oversampling Technique (SMOTE) method. The use of this method plays a crucial role in balancing the dataset by generating synthetic samples, preventing bias towards the majority class, and ensuring an unbiased training process.
- In order to generate a single model that incorporates the strengths and qualities of these models and, as a result, helps to improve the prediction accuracy, we proposed the use of a hard voting classifier that incorporates the three models with the highest accuracy, namely the K-Nearest Neighbors (KNN), Random Forest (RF), and Extra Trees Classifier (ETC).
- We used additional measures, such as k-fold cross-validation, to evaluate the performance and robustness of a ML model, which were not addressed in previous studies. This process helps to assess how well the model generalizes to an independent dataset, which is crucial to ensure that the model performs well on unseen data. We also used the False Positive Rate (FPR) and the False Negative Rate (FNR) to provide insight into the types of errors our models make.

B. Related Works

In this section, we summarize relevant studies that have used ML and deep learning techniques for stroke prediction. Authors in [9] applied four ML classification techniques for early stroke prediction. They compared these algorithms on a reliable stroke health dataset containing factors and features that affect the accuracy of model performance. Hyperparameter tuning was applied to the ML algorithms to increase the accuracy of the results, and RF achieved the highest accuracy of 90%. Authors in [10] trained different AI classifiers on a trusted dataset of stroke sources, considering distinctive features (age, smoking status, body mass index, presence of hypertension, heart disease, previous stroke) for early stroke prediction. The results of the basic classifiers were combined using the weighted voting approach to achieve a high accuracy of 97%. Authors in [11] applied a set of ML algorithms including Decision Tree (DT), logistic regression, RF, and voting classification to the stroke dataset. The results showed that RF is the best algorithm in terms of model performance accuracy, which was estimated to be 96%. Authors in [12] trained five different ML models (KNN, DT, Support Vector Machine (SVM), RF, and Naïve Baye) for early stroke prediction and applied them to a reliable stroke prediction dataset from Kaggle. The results showed that Naïve Baye performed well and achieved the highest accuracy of 82%.

Authors in [13] proposed the use of four ML models, as well as hard and soft voting. The models were trained on a stroke dataset and the factors or features that affect the diagnosis of stroke were investigated. Several models were developed and the RF classifier achieved an accuracy of 94.6%. Authors in [14] proposed the use of one of the ensemble ML algorithms, the Xtreme Gradient Boosting algorithm, on a stroke dataset. The data were divided into 70% for training and 30% for testing, achieving an accuracy of 96%. Authors in [15] designed an AI model for early prediction on a stroke dataset. Several factors were examined and ML techniques were applied, with logistic regression achieving the highest accuracy estimated at 95.71%. To develop the best model for stroke prediction, authors in [16] suggested the use of seven ML algorithms: SVM, KNN, RF, Naive Bayes, stacked majority voting, and DT. They trained the model on a reliable dataset from Kaggle and several metrics were used to measure the accuracy of the model performance, namely, F1 score, recall, precision, and accuracy. After preprocessing and partitioning the data, the algorithms achieved an accuracy of about 96%. Authors in [17] used ML algorithms for an integrated approach to stroke prediction. They used data

cleaning, imbalance treatment, and evaluated the model with several metrics, including AUC, precision, recall, F1 score, and accuracy. RF achieved the highest accuracy of 94.85%.

Previous studies have struggled with how well the diagnosis was performed, possibly because they did not use appropriate parameters to balance the dataset, resulting in overlap with samples from the other category. No analytical method has been used to select the appropriate models for the dataset, due to limitations in classification accuracy.

II. METHODOLOGY

In this section, the proposed methodology is presented in detail. Figure 1 illustrates the workflow of the proposed methodology.

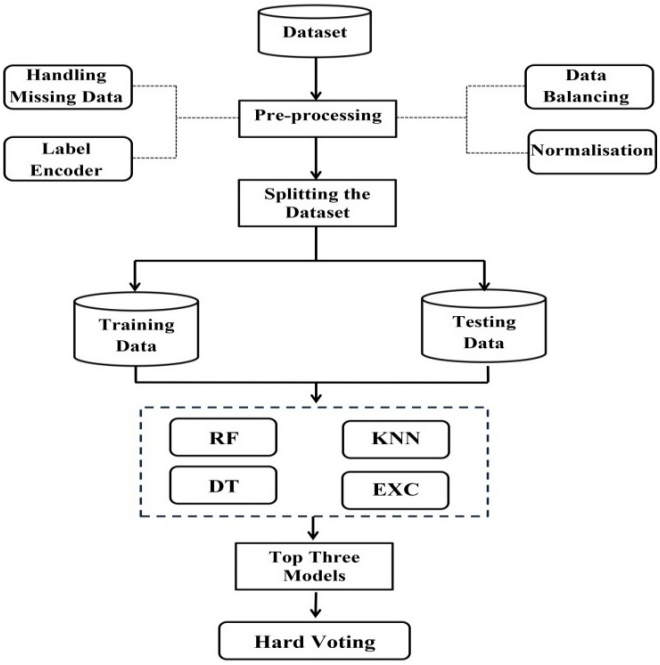

Fig. 1. The proposed methodology workflow.

A. Dataset Description

In this study, we selected the stroke dataset available on Kaggle [18]. It contains 5110 cases, 249 positive cases and 4861 negative cases, each with 11 features indicating information about the patients and their condition. In addition, it contains a final feature indicating whether the patient has a stroke or not [19]. There is a large imbalance between the number of positive and negative cases and other issues that need to be addressed. Therefore, a number of preprocessing steps are performed to improve the quality of the data.

B. Preprocessing

Before starting the classification and diagnosis process, it is necessary to examine the dataset to resolve possible problems and improve the quality of the data [20], which will help to provide the best training and achieve the highest classification accuracy [21]. Therefore, in this part, we perform several processes on the stroke dataset: missing data handling, label encoder, data balancing and normalization.

First, we removed the feature "id" because it is unimportant and does not affect the classification process. Then, we examined all the features to find the missing values and discovered that the feature "bmi" contains 201 missing values. Therefore, in this paper, we used the mean value to fill in the missing values. Second, the stroke dataset contains categorical features, namely, smoking status, ever married, gender, residence type, and work type. Considering that ML models only deal with numerical values [22], we converted the above features from categorical to numerical using the label encoder function. Third, the stroke dataset contains 5110 cases, of which 249 are positive and 4861 are negative. The number of negative cases is higher than the number of positive cases, so when ML models are trained on these data, they will have a large bias towards negative cases and learn more about them compared to the positive cases. Thus, the classification process will suffer from bias due to the large difference between the number of positive and negative cases [23]. Here, we used the SMOTE method to balance the dataset because it generates new samples of positive cases by using the KNN algorithm to find the nearest neighbors for a given case. It then draws a line between these cases and their neighbors, and then creates a new sample on the straight line connecting them [24].

After applying the SMOTE algorithm, the number of positive cases will be equal to the number of negative cases, which is 4861 for each. The main reasons for using the SMOTE method to balance the dataset are as follows: first, the SMOTE method balances the data so that the number of positive cases is exactly equal to the number of negative cases. Second, SMOTE generates synthetic examples rather than duplicating existing ones, which helps to reduce the risk of overfitting [25].

Finally, we used Standard Scaler to scale the features because there may be outliers in the dataset, and individual features may exhibit strange behavior if the dataset is not normally distributed [26]. The features have several dimensions and scales, which makes it difficult to model the dataset [27]. The relationship between misclassification errors and accuracy distorts the prediction results, so prior to modeling data scaling is required [28]. Standard Scaler is used because it ensures that features with different units have the same impact on the model.

C. Splitting the Dataset

After performing the initial data processing, which consisted of four stages to improve the data quality and classification accuracy, we divided the stroke dataset into two parts, namely training and testing. In this study, we used 80% for training and 20% for testing, so that the comparison of our results with previous works would be fair and flexible. Also, we used the K-fold cross-validation metrics, where the selected value of is 10.

D. Classification Models

The ML models used are considered to be among the best for accurate classification according to the type and distribution

of the stroke dataset. The proposed methodology is mainly based on four ML models, specifically RF, ETC, DT, and KNN. These four models were nominated as the best performing models based on a series of experiments we conducted as well as the analysis of previous studies. Stroke detection is performed in two stages according to the proposed methodology; in the first stage, we test the performance of our proposed models separately based on a set of metrics. In the second stage, we drop the worst performing model and pass the remaining three models to the hard voting classifier. The hard voting provides the final prediction, as shown in Figure 2.

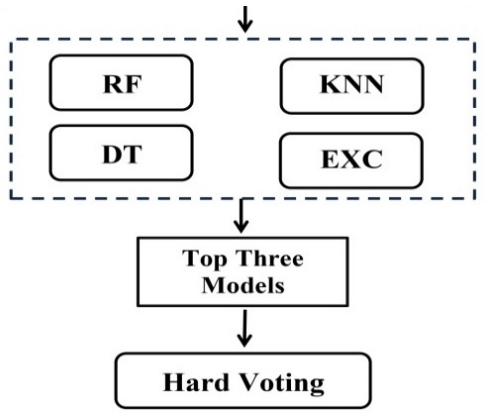

Fig. 2. The process of selecting the best model for stroke prediction.

To analyze the contribution of each model within the hard voting classifier, an ablation study was conducted. This involved evaluating the individual performance of each model and systematically removing them from the voting ensemble to observe their impact on the final classification accuracy. Therefore, we performed model-only performance, two-model combinations, three-model combinations, and four-model combinations to find the best combination of models that play critical roles in the hard voting classifier. The three model combinations are the best, with ETC contributing the most to accuracy, followed by RF. KNN contributed to robustness but had a relatively small impact. For the models, the default hyperparameters of Scikit-learn were utilized. This was done to ensure a fair comparison of models without hyperparameter tuning bias and to increase repeatability, enabling future researchers to reproduce findings and show baseline model performance for fine-tuning.

III. RESULTS AND DISCUSSION

We used several metrics to show how well ML models perform in these experiments. These include accuracy, K-fold cross-validation, F1 score, precision, recall, AUC and ROC curves, FPR, and FNR. First, we show the performance of the models before and after dataset balancing and discuss its importance in improving the performance of the models. Then, we compare the performance of our proposed model with previous studies in the same area using the same data set.

A. First Experiment

In this experiment, we applied the proposed methodology to the stroke dataset, where a number of preliminary procedures were performed, the most important of which was the dataset balancing process. In this experiment, a balanced dataset was used using SMOTE. The benefit of this process is that it eliminates the bias during the training process due to the large difference between the number of positive and negative cases. As a result, the number of negative cases became equal to the number of the positive cases, resulting in a total of 9722 cases. In addition, the dataset was divided into 80% training and 20% testing and fed to the models used. Table I shows the performance results of the models after balancing the dataset, and Table II explains the confusion matrix for each model. A significant improvement in the performance of the models is observed due to the balancing of the dataset, which eliminated the bias during the training process, leading to proper and correct training of the models. In addition, ETC achieved the highest classification accuracy of 97.12%. Figure 3 shows the AUC and ROC curves for each model after balancing the dataset. It can be seen that ETC is clearly superior in terms of AUC, which reached 0.9686, the highest value compared to the rest of the models. This good performance is due to the balancing of the dataset, which eliminates bias during the training process.

TABLE I. MODEL PERFORMANCE AFTER DATASET BALANCING

Model	Accuracy (%)	Precision	F1 score	Recall	10-fold (%)
RF	96.45	0.9669	0.9644	0.9619	96.56
DT	87.56	0.8747	0.8756	0.8765	93.68
ETC	97.12	0.9771	0.9710	0.9650	96.33
KNN	92.24	0.8854	0.9259	0.9702	92.02

TABLE II. THE CONFUSION MATRICES OF THE MODELS BEFORE BALANCING THE DATASET

Model	Confusion matrix	
RF	966	2
	54	0
DT	941	27
	49	5
ETC	965	3
	54	0
KNN	967	1
	53	1

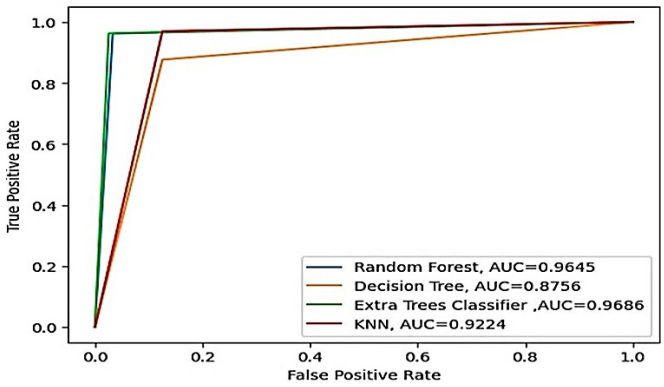

Fig. 3. AUC and ROC curves for the models after dataset balancing.

B. Second Experiment

In the first experiment, we found that DT is the worst performing model. According to our proposed methodology, we drop this model and pass the remaining three models (RF, ETC, and KNN) to the hard voting classifier to get the final stroke prediction. We conducted this experiment using the balanced dataset and Table III shows the results of the hard voting performance. As can be seen from this table, hard voting performed much better than the models that were tested separately in the first experiment. It achieved an accuracy of 97.48%, a precision of 0.9802, an F1 score of 0.9747, a recall of 0.9691, and a 10-fold accuracy of 97.1%. Here we see the importance of using voting classifiers in the prediction process by merging more than one model to produce a single strong model that carries the characteristics and strengths of the models it is fed. Figure 4 shows the AUC and ROC curves for the hard voting classifier and Figure 5 shows the accuracy and K-fold cross-validation values for all the models, indicating that the hard voting classifier is superior to the other models. Clearly, the hard voting classifier outperformed the other models with an accuracy of 97.48% and a K-fold of 97.1%.

To provide insight into the types of errors our models make, we used the FPR and FNR, presented in Table IV, to evaluate their reliability in different scenarios. Also, to demonstrate the strength of our results and the contributions of our work, we compare the performance results of our proposed methodology with previous studies in Table V.

TABLE III. PERFORMANCE RESULTS OF THE HARD VOTING CLASSIFIER

Metrics	Hard voting classifier (RF, ETC, KNN)	
Accuracy (%)	97.48	
Precision	0.9802	
F1 score	0.9747	
Recall	0.9691	
10-Fold (%)	97.1	
Confusion matrix	954	19
	30	942

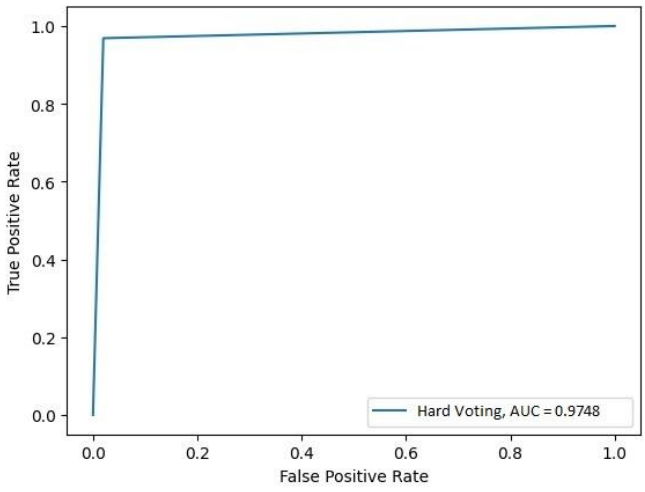

Fig. 4. The AUC and ROC curves for hard voting classifier.

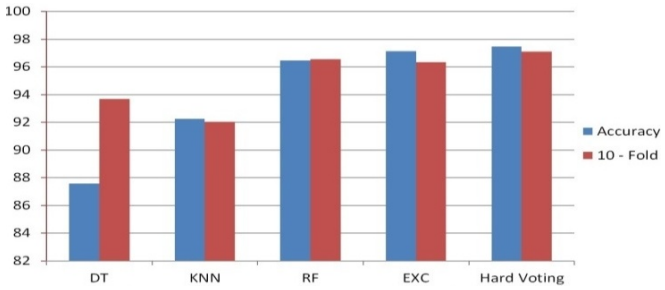

Fig. 5. Accuracy and K-fold for all models.

TABLE IV. FALSE POSITIVE AND FALSE NEGATIVE RATES FOR ALL MODELS

Model	FPR (%)	FNR (%)
DT	12.54	12.35
KNN	12.54	2.98
RF	3.29	3.81
ETC	2.26	3.50
Hard voting	1.95	3.09

TABLE V. PERFORMANCE COMPARISON OF THE PROPOSED METHODOLOGY WITH RELATED WORKS

Reference	Dataset	Best model	Accuracy (%)
[11]	Stroke prediction dataset	RF	96
[5]	Stroke prediction dataset	Naive bayes	82
[13]	Stroke prediction dataset	RF	94.6
[14]	Stroke prediction dataset	Extreme gradient, Boosting algorithm	96
[15]	Stroke prediction dataset	Logistic regression	95.71
[16]	Stroke prediction dataset	RF	96
[17]	Stroke prediction dataset	RF	94.85
Proposed methodology	Stroke prediction dataset	Hard voting	97.48

Table V presents a comparative analysis of the performance of our proposed methodology against related studies in stroke prediction. Overall, this study proposes a novel integration of ML models in a hard voting classifier for stroke prediction by combining three high-performing models (RF, ETC, and KNN) to improve prediction accuracy. In addition, the work uses K-fold cross-validation, AUC-ROC analysis, and FPR/FNR to ensure robustness. Our method outperforms existing stroke prediction models, achieving a higher accuracy of 97.48%, improved generalization through 10-fold cross-validation (97.1% mean accuracy), and more reliable predictions by reducing the FPR (1.95%) and the FNR (3.09%). In addition, the hard voting classifier outperforms the individual models, improving accuracy by +1.48% over RF and +0.36% over ETC, thus increasing the F1 score by +1.03% over RF and +0.37% over ETC. These enhancements make our model more suitable for real-world stroke prediction applications, providing a reliable decision support tool for early diagnosis. In conclusion, this study highlights the potential of ML-based stroke prediction to reduce the burden of stroke, particularly in

resource-limited healthcare environments. However, reliance on synthetic data, dataset bias, and feature limitations require further research before clinical deployment.

IV. CONCLUSION

Stroke is a serious condition that should be prevented before it occurs and is the second leading cause of death. Before the advent of modern medical technology, stroke detection relied primarily on the skill and experience of healthcare providers. Today, information technology plays a critical role in the medical field, helping to improve diagnosis, patient care, and the overall efficiency of healthcare services. Machine Learning (ML) models are very helpful in the early prediction of cases that are likely to have a stroke, so that the necessary measures can be taken to preserve the patient's life. In this paper, we propose an effective methodology that utilizes a hard voting classifier based on three ML models, namely, Random Forest (RF), K-Nearest Neighbors (KNN), and Extra Trees Classifier (ETC). First, we performed a series of initial treatments to improve the data quality and balance the dataset by using the Synthetic Minority Oversampling Technique (SMOTE) method, which helps to eliminate bias during the training process. Next, we divided the dataset into two parts, a training part and a testing part, and these data were fed to the models used. In the last phase, we implemented four ML algorithms to evaluate their effectiveness, and then selected the three most effective models for integration into our proposed hard voting classifier. The proposed hard voting outperforms the results of modern research, achieving an accuracy of 97.48%, a precision of 0.9802, a recall of 0.9691, and an F1 score of 0.9747. It also achieved a mean accuracy of 97.1% in a 10-fold cross-validation trial. This research lays the groundwork for AI-driven stroke prediction by demonstrating the potential of ensemble learning. By advancing scalable, interpretable, and data-driven stroke prediction models, this research paves the way for early intervention strategies, ultimately reducing stroke mortality and improving global healthcare outcomes. In future work, the integration of real-world clinical data, real-time edge AI applications, and explainable decision-making frameworks will empower healthcare providers to make proactive, data-driven interventions.

REFERENCES

[1] J. T. Marbun, Seniman, and U. Andayani, "Classification of stroke disease using convolutional neural network," *Journal of Physics: Conference Series*, vol. 978, no. 1, Mar. 2018, Art. no. 012092, https://doi.org/10.1088/1742-6596/978/1/012092.

[2] O. Almadani and R. Alshammari, "Prediction of Stroke using Data Mining Classification Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 1, pp. 457–460, Jan. 2018, https://doi.org/10.14569/IJACSA.2018.090163.

[3] V. L. Feigin *et al.*, "World Stroke Organization (WSO): Global Stroke Fact Sheet 2022," *International Journal of Stroke*, vol. 17, no. 1, pp. 18–29, Jan. 2022, https://doi.org/10.1177/17474930211065917.

[4] S. Negash, P. Musa, D. Vogel, and S. Sahay, "Healthcare information technology for development: improvements in people's lives through innovations in the uses of technologies," *Information Technology for Development*, vol. 24, no. 2, pp. 189–197, Apr. 2018, https://doi.org/10.1080/02681102.2018.1422477.

[5] R. S. Jeena and S. Kumar, "Machine intelligence in stroke prediction," *International Journal of Bioinformatics Research and Applications*, vol. 14, no. 1–2, pp. 29–48, Jan. 2018, https://doi.org/10.1504/IJBRA.2018.089192.

[6] R. J. Mohammed *et al.*, "A Robust Hybrid Machine and Deep Learning-based Model for Classification and Identification of Chest X-ray Images," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 16212–16220, Oct. 2024, https://doi.org/10.48084/etasr.7828.

[7] M. Kong, Q. He, and L. Li, "AI Assisted Clinical Diagnosis & Treatment, and Development Strategy," *Strategic Study of Chinese Academy of Engineering*, vol. 20, no. 2, pp. 86–91, Apr. 2018, https://doi.org/10.15302/J-SSCAE-2018.02.013.

[8] S. Campagnini, C. Arienti, M. Patrini, P. Liuzzi, A. Mannini, and M. C. Carrozza, "Machine learning methods for functional recovery prediction and prognosis in post-stroke rehabilitation: a systematic review," *Journal of NeuroEngineering and Rehabilitation*, vol. 19, no. 1, Jun. 2022, Art. no. 54, https://doi.org/10.1186/s12984-022-01032-4.

[9] H. Ahmed, S. F. Abd-el Ghany, E. M. G. Youn, N. F. Omran, and A. A. Ali, "Stroke Prediction using Distributed Machine Learning Based on Apache Spark," *International Journal of Advanced Science and Technology*, vol. 28, no. 15, pp. 89–97, Nov. 2019.

[10] M. U. Emon, M. S. Keya, T. I. Meghla, Md. M. Rahman, M. S. A. Mamun, and M. S. Kaiser, "Performance Analysis of Machine Learning Approaches in Stroke Prediction," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology*, Coimbatore, India, 2020, pp. 1464–1469, https://doi.org/10.1109/ICECA49313.2020.9297525.

[11] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, "Stroke Disease Detection and Prediction Using Robust Learning Approaches," *Journal of Healthcare Engineering*, vol. 2021, no. 1, Nov. 2021, Art. no. 7633381, https://doi.org/10.1155/2021/7633381.

[12] 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, pp. 539–545, Jun. 2021, https://doi.org/10.14569/IJACSA.2021.0120662.

[13] H. Al-Zubaidi, M. Dweik, and A. Al-Mousa, "Stroke Prediction Using Machine Learning Classification Methods," in *2022 International Arab Conference on Information Technology*, Abu Dhabi, United Arab Emirates, 2022, pp. 1–8, https://doi.org/10.1109/ACIT57182.2022.10022050.

[14] A. M. A. Rahim, A. Sunyoto, and M. R. Arief, "Stroke Prediction Using Machine Learning Method with Extreme Gradient Boosting Algorithm," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 595–606, Aug. 2022, https://doi.org/10.30812/matrik.v21i3.1666.

[15] P. Agarwal, M. Khandelwal, Nishtha, and D. A. K. Kadam, "Brain Stroke Prediction using Machine Learning Approach," *Iconic Research And Engineering Journals*, vol. 6, no. 1, pp. 273–277, Jul. 2022.

[16] N. Alageel, R. Alharbi, R. Alharbi, M. Alsayil, and L. A. Alharbi, "Using Machine Learning Algorithm as a Method for Improving Stroke Prediction," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 4, pp. 738–744, Apr. 2023, https://doi.org/10.14569/IJACSA.2023.0140481.

[17] R. Kuksal, G. A. Reddy, M. Vaqur, A. Bhatt, and K. Joshi, "Brain stroke prediction using machine learning," *AIP Conference Proceedings*, vol. 2919, no. 1, Mar. 2024, Art. no. 090013, https://doi.org/10.1063/5.0185040.

[18] "Stroke Prediction Dataset." Kaggle. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

[19] S. Rahman, M. Hasan, and A. K. Sarkar, "Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques," *European Journal of Electrical Engineering and Computer Science*, vol. 7, no. 1, pp. 23–30, Jan. 2023, https://doi.org/10.24018/ejece.2023.7.1.483.

[20] M. J. J. Ghrabat, G. Ma, I. Y. Maolood, S. S. Alresheedi, and Z. A. Abduljabbar, "An effective image retrieval based on optimized genetic algorithm utilized a novel SVM-based convolutional neural network classifier," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, Aug. 2019, Art. no. 31, https://doi.org/10.1186/s13673-019-0191-8.

[21] W. A. Awadh, M. S. Hashim, and A. S. Alasady, "Implementing the Triple-Data Encryption Standard for Secure and Efficient Healthcare Data Storage in Cloud Computing Environments," *Informatica*, vol. 48, no. 6, pp. 173–184, Apr. 2024, https://doi.org/10.31449/inf.v48i6.5641.

[22] K. Potdar, T. S. Pardawala, and C. D. Pai, "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," *International Journal of Computer Applications*, vol. 175, no. 4, pp. 7–9, Oct. 2017.

[23] I. Letteri, A. D. Cecco, A. Dyoub, and G. D. Penna, "A Novel Resampling Technique for Imbalanced Dataset Optimization." arXiv, Dec. 30, 2020, https://doi.org/10.48550/arXiv.2012.15231.

[24] M. Hashim and A. Yassin, "Using Pearson Correlation and Mutual Information (PC-MI) to Select Features for Accurate Breast Cancer Diagnosis Based on a Soft Voting Classifier," *Iraqi Journal for Electrical and Electronic Engineering*, vol. 19, no. 2, pp. 43–53, Dec. 2023, https://doi.org/10.37917/ijeee.19.2.6.

[25] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 6390–6404, Sep. 2023, https://doi.org/10.1109/TNNLS.2021.3136503.

[26] J. Sola and J. Sevilla, "Importance of input data normalization for the application of neural networks to complex industrial problems," *IEEE Transactions on Nuclear Science*, vol. 44, no. 3, pp. 1464–1468, Jun. 1997, https://doi.org/10.1109/23.589532.

[27] M. J. J. Ghrabat, G. Ma, Z. A. Abduljabbar, M. A. Al Sibahee, and S. J. Jassim, "Greedy Learning of Deep Boltzmann Machine (GDBM)'s Variance and Search Algorithm for Efficient Image Retrieval," *IEEE Access*, vol. 7, pp. 169142–169159, 2019, https://doi.org/10.1109/ACCESS.2019.2948266.

[28] P. Ferreira, D. C. Le, and N. Zincir-Heywood, "Exploring Feature Normalization and Temporal Information for Machine Learning Based Insider Threat Detection," in *2019 15th International Conference on Network and Service Management*, Halifax, Canada, 2019, pp. 1–7, https://doi.org/10.23919/CNSM46954.2019.9012708.