

Balancing and metaheuristic techniques for improving machine learning models in brain stroke prediction

Abd Allah Aouragh¹, Mohamed Bahaj¹, Fouad Toufik²

¹MIET Laboratory, Faculty of Sciences and Techniques, Hassan 1st University, Settat, Morocco

²Computer Sciences Laboratory, Higher School of Technology, Mohammed V University, Sale, Morocco

Article Info

Article history:

Received Mar 19, 2024

Revised Oct 19, 2024

Accepted Oct 23, 2024

Keywords:

Brain stroke

Genetic algorithm

Hyperparameter optimization

KMeansSMOTE

Machine learning

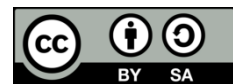
Oversampling

Random forest

ABSTRACT

A brain stroke, medically referred to as a stroke, represents a critical condition triggered by the disruption of blood flow to a region of the brain. Early detection of stroke is crucial to prevent fatal complications. In this study, we worked with an unbalanced dataset of 4981 entries on stroke, which we balanced using the K-means synthetic minority over-sampling technique (KMeansSMOTE) algorithm. We then employed five machine learning algorithms: decision tree, random forest, support vector machine, K-nearest neighbors, and gradient boosting. We compared the hyperparameter optimization of these algorithms using four metaheuristic techniques: gray wolf optimization, particle swarm optimization, genetic algorithm, and artificial bee colony. The models' effectiveness was evaluated using multiple metrics, such as accuracy, recall, precision, F1-score, and area under the receiver operating characteristic curve. Our findings indicate that the random forest optimized by the genetic algorithm achieved the best performance, with an accuracy of 97.39% and an F1-score of 97.35%. This study highlights the effectiveness of balancing and metaheuristics techniques in optimizing machine learning models for stroke forecasting.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Abd Allah Aouragh

MIET Laboratory, Faculty of Sciences and Techniques, Hassan 1st University

Settat, Morocco

Email: abdallahaouragh@gmail.com

1. INTRODUCTION

Brain stroke, also known as stroke, is a critical health problem worldwide, constituting one of the foremost causes of mortality and long-term handicap [1]. When a stroke occurs, the risk of death is high. If not fatal, a stroke can cause vision or speech impairment, paralysis, and confusion. Every year, 15 million people are reported to have a stroke: 5 million die, and 5 million are left permanently disabled, burdening families and societies [1], [2]. A stroke happens when blood flow to a section of the brain is disrupted, leading to oxygen deprivation, and subsequently, tissue damage. Prompt recognition and early intervention are paramount in mitigating the devastating consequences of this condition [3]. While there are many treatments available for stroke, including surgery, radiation therapy, chemotherapy, and targeted therapeutic approaches [4], these interventions can be costly, and their effectiveness often depends on how quickly the stroke is diagnosed and treated. Therefore, early diagnosis of stroke is of utmost importance, as it not only improves the health situation of patients but also reduces the costs associated with their rehabilitation [5].

Although potentially serious, stroke remains a challenging condition to predict and manage effectively [3], [4]. Machine learning techniques have become promising instruments in the healthcare field, facilitating the analysis of complex medical data and aiding clinical decision-making [6]. These advancements have also improved the prediction of various diseases, including brain tumors [7], liver disease [8], and others [9], notably

through the utilization of dataset balancing and hyperparameter optimization techniques [10], [11]. In the context of stroke, the utilization of machine learning algorithms and diverse optimization techniques for early prediction holds significant promise for improving patient care outcomes and alleviating the burden on healthcare systems. These innovative approaches enable predictive models to be tailored to the nuances of medical data, thereby enhancing their accuracy and reliability [12].

In this context, Daidone *et al.* [13] examined the use of machine learning techniques in stroke prediction. They emphasized the importance of developing sophisticated technologies to enhance diagnosis, treatment, and patient results. Their findings demonstrate that emerging machine learning techniques have shown exceptional precision in analyzing images, diagnosing subtypes of stroke, and predicting patient prognosis. However, several challenges persist, including data standardization and model validation. Sirsat *et al.* [14] conducted a study on the use of machine learning for accurate stroke prediction. They concluded that machine learning is a potent instrument in healthcare, providing tailored clinical care to stroke patients. They identified a need for research in certain underexplored areas, particularly in stroke treatment. The study highlighted the efficacy of random forest (RF) and support vector machine (SVM) in stroke forecasting. Emon *et al.* [15] built a model for the early forecasting of stroke-related diseases using various machine learning techniques. They identified several factors, including heart disease, body mass index, hypertension, stroke history, and age, as significant predictors. Their model involved training ten classifiers, the results of which were combined via weighted voting, resulting in 97% accuracy. The weighted voting classifier proved to be the best performer for predicting stroke, exhibiting the lowest false positive and false negative rates. Tazin *et al.* [16] developed stroke prediction models utilizing a variety of machine learning methods. Their research incorporated physiological data and algorithms, including logistic regression, RF classification, decision tree (DT) classification, and voting classifiers, to train four distinct models. Among these algorithms, RF emerged as the most effective, achieving an accuracy of approximately 96%. These results showcased notably higher accuracy compared to previous studies, affirming the reliability of the developed models.

Akter *et al.* [17] developed a precise model for predicting strokes utilizing machine learning algorithms. They evaluated DT, SVM, and RF models during training and testing. The efficiency of each classifier was assessed using various evaluation metrics such as accuracy, sensitivity, false negative rate, false positive rate, and error rate. The proposed model attained a maximum accuracy of 95.30% with the RF classifier, demonstrating its effectiveness in accurate stroke prediction. Paliwal *et al.* [18] investigated the substantial influence of early detection and swift intervention in mitigating stroke damage and enhancing survivors' quality of life. Employing a variety of machine learning methods, including DT, logistic regression, SVM, and RF, they devised a model targeting stroke prediction. The emphasis was on assessing the efficacy of oversampling techniques for managing unbalanced data. Their results revealed that the K-means synthetic minority over-sampling technique (KMeansSMOTE) technique yielded the highest accuracy of 96%, with minimal false positives and false negatives, showcasing its effectiveness in stroke prediction. Srivastav *et al.* [19] have created a model aimed at predicting stroke occurrence through different measures such as precision, recall, F1-score, and root mean square error (RMSE). An analytical comparison of prediction performance was conducted using several machine learning algorithms, and it was found that logistic regression yielded the best performance with an accuracy of 95.02%. Srinivas and Mosiganti [20] developed an ensemble learning model for stroke prediction, combining forecasts from extremely randomized trees, RF, and histogram-based gradient boosting (GB). This soft voting model improved precision and reliability, achieving 96.88% accuracy.

In our study, we extend previous research on machine learning applications for predicting brain strokes. Existing studies have explored various techniques, but no universally applicable method has emerged. Previous research primarily focused on developing and evaluating machine learning models using diverse datasets, often overlooking the critical importance of data balancing and hyperparameter optimization techniques. Our study addresses these gaps in the literature by introducing a comprehensive approach that evaluates the effectiveness of KMeansSMOTE, a powerful data balancing technique, in conjunction with advanced metaheuristic optimization techniques for hyperparameters, including gray wolf optimization (GWO), particle swarm optimization (PSO), genetic algorithm (GA), and artificial bee colony (ABC). We rigorously assess the combined impact of these techniques on model performance using five different machine learning algorithms: DT, SVM, K-nearest neighbors (KNN), GB, and RF. Importantly, our methodology demonstrates superior effectiveness compared to previous research endeavors, particularly in terms of achieving balanced datasets and optimized hyperparameters, which are crucial for enhancing model accuracy and robustness. This study not only fills a significant gap in the existing literature but also offers a novel contribution to the field of stroke prediction through a well-rounded and innovative methodological framework.

The following sections of this study are structured as follows: section 2 provides an in-depth explanation of the materials and methods utilized. Section 3 presents and discusses the results, including an

analysis of the effects of the techniques employed. Lastly, section 4 offers a summary of the key findings and proposes potential avenues for future investigations.

2. MATERIAL AND METHOD

2.1. Proposed methodology

In our study focused on predicting strokes using machine learning algorithms, we devised a methodology encompassing several crucial steps. Initially, we acquired the stroke dataset, which provides comprehensive information about patients afflicted by this medical condition. Subsequently, we conducted a data preprocessing phase, involving operations like encoding, data partitioning, and normalization, to ensure uniform data scaling. Furthermore, we addressed the challenge of data imbalance by assessing various balancing techniques, with KMeansSMOTE identified as the optimal solution. Subsequently, we assessed a variety of machine learning methods, such as DT, KNN, SVM, RF, and GB. To improve the efficacy of these models, we conducted hyperparameter optimization for each using metaheuristic optimization techniques such as GWO, PSO, GA, and ABC. Finally, Figure 1 illustrates the methodological approach adopted in this study.

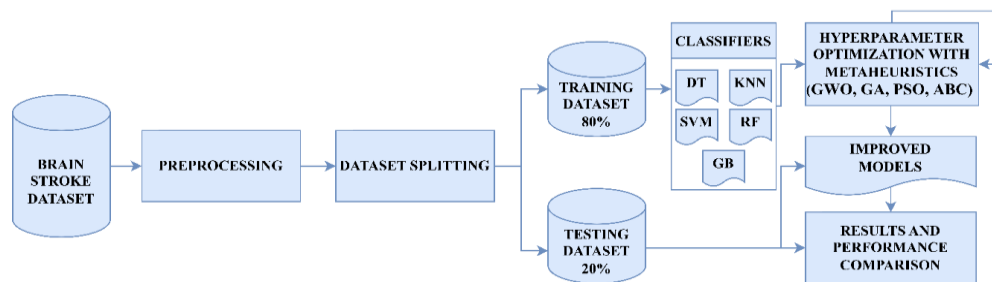


Figure 1. Suggested methodology

2.2. Dataset

In our brain stroke prediction project, we utilize the dataset sourced from Kaggle [21], which serves as a valuable asset in the medical domain. This dataset offers a robust foundation for researchers and healthcare practitioners interested in exploring the correlation between clinical attributes and brain strokes. Comprising 11 anthropometric and biological characteristics, the brain stroke dataset encompasses 4981 records, providing a comprehensive basis for brain stroke analysis and prediction. The brain stroke dataset is divided into two classes: stroke patients (4.97%) and non-stroke patients (95.03%). This significant class imbalance can present challenges during the training of machine learning models, leading to biased performance assessments. Therefore, addressing unbalanced classes is crucial to ensure accurate and balanced predictions. Further details regarding the dataset composition are outlined in Table 1.

Table 1. Dataset composition

Attribute	Description
Gender	Gender of patient
Age	Patient's age
Hypertension	0 for no hypertension, 1 for hypertension
Heart_disease	0 for no heart disease, 1 for heart disease
Ever_married	"No" or "Yes"
Work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
Residence_type	"Rural" or "Urban"
Avg_glucose_level	Average blood glucose level
BMI	Body Mass Index
Smoking_status	"formerly smoked", "never smoked" or "smokes"
Stroke	Target variable, 0: Healthy, 1: Affected

2.3. Balancing dataset

Dataset balancing is crucial in our study to ensure accurate and balanced predictions, particularly in the presence of unbalanced classes like stroke and non-stroke patients. In our research, we evaluated multiple data balancing techniques, including SMOTE, adaptive synthetic sampling (ADASYN), SVMSMOTE, and KMeansSMOTE [8], [22]–[24]. We selected these techniques due to their widespread adoption and

demonstrated effectiveness in medical research, where class imbalances are prevalent [8], [22]–[24]. Among these techniques, KMeansSMOTE emerged as the best-performing option. The KMeans algorithm is utilized to identify clusters in the original data, followed by a separate application of SMOTE to each group to generate synthetic data points that closely resemble examples from the minority class [25]. This approach enabled us to generate a synthetic dataset that better represents the minority class, enhancing our model's ability to generalize and produce exact predictions for stroke patients.

2.4. Machine learning algorithms

In our study, we evaluated several machine learning algorithms for stroke prediction. These algorithms were selected for their efficiency in modeling classification problems and have been widely used in similar medical fields to predict clinical outcomes [8], [23], [26]–[28]. The evaluated algorithms include as follows.

2.4.1. Decision tree

The DT is a simple and interpretable algorithm widely used for classification tasks. It can effectively capture non-trivial correlations between predictors and the outcome variable while also handling missing data and outliers efficiently. The decision rules it generates are explainable, making it particularly useful in fields such as medicine [26]–[28].

2.4.2. K-nearest neighbors

KNN is a basic classification technique that assigns data points to the most common class among their nearest neighbors. It's straightforward to implement and doesn't assume any specific data distribution, making it robust to noisy data. However, it may incur significant computational costs, particularly with high-dimensional datasets [23], [26]–[28].

2.4.3. Support vector machine

SVM is known for its effectiveness in handling high-dimensional data and finding non-linear decision boundaries. It works by finding the decision boundary that optimally segregates distinct classes while maximizing the margin between them. SVM is particularly useful in scenarios with complex data relationships but may require careful tuning for optimal performance [26]–[28].

2.4.4. Random forest

RF is a versatile ensemble learning algorithm known for its robustness and high accuracy. It builds numerous DT during the training phase and integrates their results to formulate predictions. RF is resilient to overfitting, performs well with high-dimensional data, and provides valuable insights into feature importance [7], [26]–[28].

2.4.5. Gradient boosting

GB is a robust ensemble learning method that constructs a potent predictive model by iteratively incorporating weak learners. Often DTs, to minimize the loss function. It effectively handles complex datasets, reduces bias and variance, and can model complex interactions between predictors and the response variable [23], [26]–[28].

2.5. Hyperparameter optimization with metaheuristics

Optimizing hyperparameters is crucial for fine-tuning the effectiveness of machine learning models. Traditionally, this optimization is performed using methods such as grid search, which evaluates every potential combination of parameters to detect the best configuration [8]. However, this approach may be inefficient for large or complex parameter spaces. Metaheuristic techniques provide a more effective alternative, efficiently exploring the solution space to discover optimal configurations within reasonable time frames. Unlike grid search, they offer greater flexibility, reduced susceptibility to local minima, and enhanced suitability for addressing the complexities of hyperparameter optimization [29], [30]. In our study, we opted to utilize the following four techniques, widely recognized and extensively employed in the research community:

2.5.1. Gray wolf optimization

GWO is a metaheuristic approach influenced by the social organization and hunting behavior of gray wolves. The algorithm operates by emulating the hunting strategies of gray wolves, in which the pack members collaborate to track and capture prey. Initially, the locations of the alpha, beta, delta, and omega wolves are randomly initialized within the exploration space. Then, the algorithm continually modifies the locations of the wolves depending on predefined equations, incorporating the alpha, beta, delta, and omega wolves' positions

to guide the search toward optimal solutions. GWO is advantageous because it requires fewer control parameters, is easy to implement, and exhibits fast convergence [29], [30].

2.5.2. Genetic algorithm

GA is a metaheuristic algorithm inspired by natural selection and evolution, employing principles such as selection, crossover, and mutation to seek optimal solutions within a defined problem space. Starting with an initial population of potential solutions, represented as chromosomes, the algorithm selects individuals with higher fitness scores for reproduction. Through crossover, selected pairs exchange genetic information to generate offspring, while mutation introduces small changes to maintain population diversity. This iterative process continues until convergence to an ideal or quasi-optimal solution [29], [30].

2.5.3. Particle swarm optimization

PSO is a metaheuristic algorithm guided by social bird flocking or fish schooling behavior. In PSO, a population of possible solutions, represented as particles, moves across the search space to find optimal solutions. Each particle adapts its location using its individual history and the collective actions of nearby particles. The algorithm iteratively adjusts the velocity and position of particles based on the most optimal positions found by individual particles and the entire swarm. This cooperative behavior allows PSO to efficiently navigate through the search space and converge towards the best solutions [29], [30].

2.5.4. Artificial bee colony

ABC is a metaheuristic optimization algorithm inspired by the hunting activity of honeybees. It involves three primary elements: working bees, observer bees, and scout bees. Working bees seek out food sources and communicate their findings to observer bees, who then select food sources based on this information, while scout bees explore new food sources. The algorithm emulates the cooperative attitude of bees in locating and exploiting food sources, facilitating efficient exploration and exploitation of the solution space. ABC is known for its simplicity, few control parameters, and effectiveness in exploring search spaces, making it a commonly chosen option in various optimization problems [29], [30].

3. RESULTS AND DISCUSSION

In this section, we delve into the outcomes and discussions stemming from our investigation into brain stroke prediction. The experiments were conducted on a computing system powered by an AMD Ryzen 7 5700G processor with Radeon graphics. Our software environment was configured with the Python language operating within Jupyter Notebook, supplemented by essential libraries like Pandas, Scikit-learn, and HypONIC.

To evaluate the efficiency of our machine learning models, we utilized a variety of assessment metrics, including accuracy, recall, precision, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). These measures serve as essential indicators of our models' predictive prowess and their capacity to generalize to new data. We selected these evaluation metrics due to their ability to offer a thorough insight into our models' performance [31], [32]. The results across different stages of the study are outlined in Tables 2–7. These results offer valuable insights into the comparative effectiveness of the algorithms in predicting brain strokes, considering both the data balancing and the hyperparameter optimization strategies employed.

In Table 2, the performance of the models on the original dataset, which is highly imbalanced, is presented. Although some models achieved high overall accuracy, with the best value reaching up to 95.79% for the SVM model, it is crucial to mention that these models are not considered effective due to their poor performance on other metrics such as precision, recall, and AUC-ROC, which do not exceed 28%, 15%, and 55%, respectively. These results indicate that these models are not suitable for the task of stroke prediction due to their inability to generalize well and effectively capture positive instances. After balancing with KMeansSMOTE, the model performances are significantly improved, as shown in Table 3. There is a significant increase in all evaluation metrics compared to the highly imbalanced original dataset. Particularly, GB stands out by achieving the highest values of accuracy (95.56%), precision (94.30%), F1-score (95.56%), and AUC-ROC (95.58%). Additionally, KNN and RF perform well, with the highest recall of 97.29%. After applying GWO to the dataset balanced by KMeansSMOTE, Table 4 highlights a clear performance improvement. Once again, GB stands out by achieving the highest values of accuracy (97.32%), precision (97.49%), F1-score (97.28%), and AUC-ROC (97.32%). Meanwhile, SVM achieves the highest recall of 97.29%.

After applying GA to the dataset balanced by KMeansSMOTE, Table 5 demonstrates a notable performance improvement. This time, RF stands out by achieving the highest values of accuracy (97.39%), precision (97.63%), F1-score (97.35%), and AUC-ROC (97.39%). Additionally, SVM achieves the highest

recall of 97.21%. After applying PSO to the dataset balanced by KMeansSMOTE, Table 6 reveals a significant performance improvement. In this case, GB stands out by achieving the highest values of accuracy (97.22%), precision (97.35%), F1-score (97.17%), and AUC-ROC (97.22%). Additionally, SVM achieves the highest recall of 97.29%. After applying ABC to the dataset balanced by KMeansSMOTE, Table 7 illustrates a significant performance improvement. This time, GB stands out by achieving the highest values of accuracy (97.32%), precision (97.29%), recall (97.29%), F1-score (97.29%), and AUC-ROC (97.32%). Additionally, SVM also achieves the highest recall of 97.29%.

Table 2. Original dataset

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
DT	91.04	10.99	15.87	12.99	55.11
KNN	95.38	25.00	4.76	8.00	52.07
SVM	95.79	0.00	0.00	0.00	50.00
RF	95.59	20.00	1.59	2.94	50.65
GB	95.59	28.57	3.17	5.71	51.41

Table 3. Balanced KMeansSMOTE dataset

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
DT	94.15	92.79	95.57	94.16	94.17
KNN	94.15	91.41	97.29	94.26	94.20
SVM	95.04	92.97	97.29	95.08	95.07
RF	94.82	92.82	97.00	94.87	94.85
GB	95.56	94.30	96.86	95.56	95.58

Table 4. Balanced dataset + GWO Optimization

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
DT	94.40	93.67	95.07	94.36	94.41
KNN	94.26	91.65	97.21	94.35	94.30
SVM	97.11	96.87	97.29	97.08	97.12
RF	95.04	93.56	96.57	95.04	95.06
GB	97.32	97.49	97.07	97.28	97.32

Table 5. Balanced dataset + GA Optimization

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
DT	95.67	94.81	96.50	95.65	95.68
KNN	94.44	91.96	97.21	94.51	94.48
SVM	97.25	97.28	97.14	97.21	97.25
RF	97.39	97.63	97.07	97.35	97.39
GB	96.69	96.31	97.00	96.65	96.69

Table 6. Balanced dataset + PSO Optimization

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
DT	94.72	94.01	95.36	94.68	94.73
KNN	94.72	92.46	97.21	94.78	94.75
SVM	97.11	96.87	97.29	97.08	97.12
RF	95.21	93.83	96.64	95.21	95.23
GB	97.22	97.35	97.00	97.17	97.22

Table 7. Balanced dataset + ABC Optimization

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
DT	94.44	93.79	95.00	94.39	94.44
KNN	94.65	92.33	97.21	94.71	94.68
SVM	97.11	96.87	97.29	97.08	97.12
RF	94.96	93.37	96.64	94.98	94.99
GB	97.32	97.29	97.29	97.29	97.32

The results highlight the benefits of data balancing and hyperparameter optimization in stroke prediction. By combining the KMeansSMOTE method with the GA and the RF model, we reached the highest

performance, with an accuracy rate of 97.39% and an F1-score of 97.35%. The measurements for the different models of the best combination are illustrated in Figure 2. This approach proved to be effective in enhancing the overall predictive capability of the models. The GB and RF models exhibited optimal performance in accuracy, precision, F1-score, and AUC-ROC due to their intrinsic ability to capture complex relationships between variables and provide accurate predictions. Their capability to train weak models and aggregate them significantly improved the overall model performance. Conversely, the SVM stood out for recall due to its utilization of decision hyperplanes that maximize the margin between different classes, enabling it to more easily identify positive examples. The GA outperformed other metaheuristics because of its ability to efficiently explore solution spaces and identify sets of optimal parameters for the models. By employing concepts of natural selection and crossover, the GA could avoid local minima and converge towards more performant solutions within reasonable timeframes.

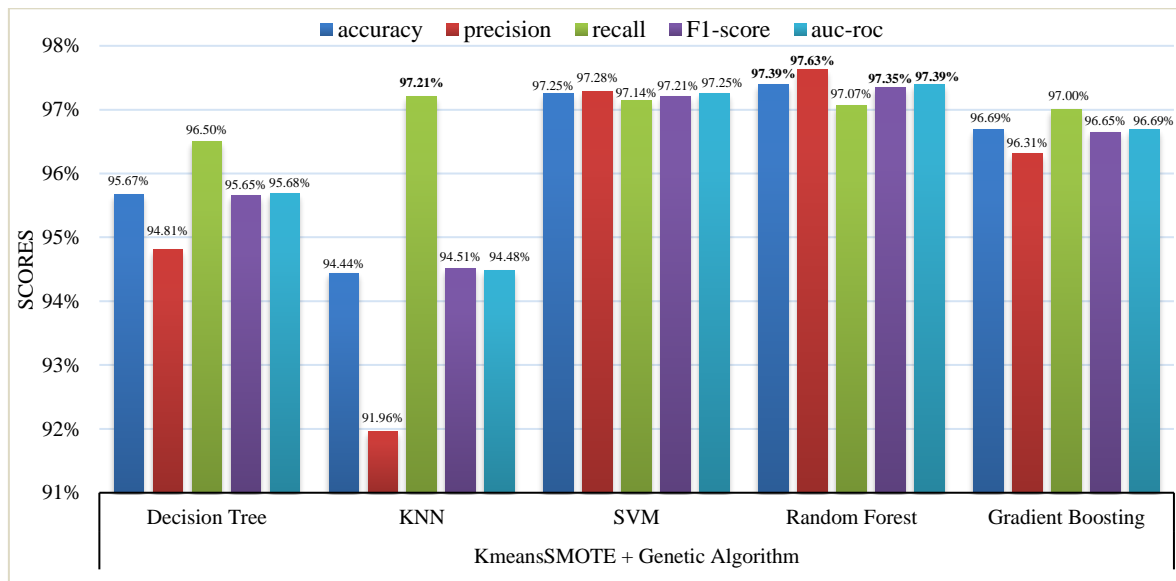


Figure 2. Metrics of the best combination

Comparing our results to previous studies, we observe that methods such as GB and RF consistently perform well, aligning with findings from other research. Our study extends these findings by specifically highlighting the impact of different hyperparameter optimization techniques on model performance. Unlike previous work, we employed KMeansSMOTE for data balancing in conjunction with various optimization algorithms, including GWO, PSO, and GA. Notably, the combination of KMeansSMOTE with the GA and RF yielded the highest accuracy and F1-score, significantly outperforming previous models. This underscores the critical role of data balancing and hyperparameter optimization in enhancing prediction accuracy and robustness. Our findings demonstrate that these techniques can significantly improve model performance compared to prior studies that did not utilize such comprehensive optimization strategies. By systematically addressing the challenges of data imbalance and optimizing hyperparameters, our approach provides a more robust framework for stroke prediction, offering improved accuracy and reliability. Despite these encouraging findings, it is important to acknowledge certain restrictions of our study. Our dataset may be limited in terms of representativeness, potentially restricting the generalizability of our findings to other populations. Additionally, some important variables may not have been accounted for in our analysis, which could affect the validity of our results. For future research, it would be advantageous to further examine the use of these techniques in other areas of predictive medicine, as well as their integration into real clinical settings. Additional studies could also delve into the impact of these techniques on treatment decisions and patient outcomes to better understand their potential in a clinical context.

4. CONCLUSION

In conclusion, this study sheds light on the significance of data balancing and hyperparameter optimization techniques in stroke prediction. The findings demonstrate that the integration of KMeansSMOTE with the GA and the RF model yielded the highest predictive performance, showcasing an accuracy rate of

97.39% and an F1-score of 97.35%. These results underscore the capability of advanced machine learning methods to improve stroke prediction accuracy. Moving forward, the successful application of these techniques opens avenues for enhanced stroke risk assessment and early intervention strategies in clinical settings. Moreover, the robust performance of the GB and RF models suggests their suitability for real-world deployment in healthcare systems, where accurate stroke prediction can significantly impact patient outcomes. Furthermore, this study prompts further exploration into the integration of advanced machine learning algorithms with domain-specific knowledge and additional patient data sources, such as genetic markers or lifestyle factors. Such interdisciplinary approaches hold promise for refining stroke prediction models and tailoring interventions to individual patient profiles. Ultimately, the findings presented in this study contribute to the ongoing research efforts aimed at leveraging machine learning for proactive healthcare management. By elucidating the effectiveness of data balancing and hyperparameter optimization techniques, this research advances our understanding of stroke prediction and paves the way for future innovations in clinical practice.




REFERENCES

- [1] S. J. Murphy and D. J. Werring, "Stroke: causes and clinical features," *Medicine*, vol. 48, no. 9, pp. 561–566, 2020, doi: 10.1016/j.mpmed.2020.06.002.
- [2] WHO, "Stroke, cerebrovascular accident," *Health Topic*, World Health Organization, 2021. Accessed: Jun. 25, 2024. [Online]. Available: <https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html>
- [3] D. Kuriakose and Z. Xiao, "Pathophysiology and treatment of stroke: Present status and future perspectives," *International Journal of Molecular Sciences*, vol. 21, no. 20, pp. 1–24, 2020, doi: 10.3390/ijms21207609.
- [4] M. O. Owolabi *et al.*, "Primary stroke prevention worldwide: translating evidence into action," *The Lancet Public Health*, vol. 7, no. 1, pp. e74–e85, 2022, doi: 10.1016/S2468-2667(21)00230-9.
- [5] T. N. Rochmah, I. T. Rahmawati, M. Dahlui, W. Budiarto, and N. Bilqis, "Economic burden of stroke disease: A systematic review," *International Journal of Environmental Research and Public Health*, vol. 18, no. 14, 2021, doi: 10.3390/ijerph18147552.
- [6] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," *International Journal of Intelligent Networks*, vol. 3, pp. 58–73, 2022, doi: 10.1016/j.ijin.2022.05.002.
- [7] A. A. Aouragh and M. Bahaj, "Combining transfer learning with CNNs and machine learning algorithms for improved brain tumor classification from MRI," in *Artificial Intelligence, Data Science and Applications*, 2024, pp. 391–397, doi: 10.1007/978-3-031-48573-2_56.
- [8] A. A. Aouragh and M. Bahaj, "Feature selection and dimensionality reduction for unbalanced liver disease classification with optimized machine learning algorithms," in *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, 2023, pp. 547–552, doi: 10.1109/CiSt56084.2023.10409967.
- [9] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: a comprehensive review," *Healthcare*, vol. 10, no. 3, 2022, doi: 10.3390/healthcare10030541.
- [10] K. M. Hasib *et al.*, "A survey of methods for managing the classification and solution of data imbalance problem," *Journal of Computer Science*, vol. 16, no. 11, pp. 1546–1557, 2020, doi: 10.3844/JCSSP.2020.1546.1557.
- [11] J. Kong, W. Kowalczyk, D. A. Nguyen, T. Back, and S. Menzel, "Hyperparameter optimisation for improving classification under class imbalance," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2019, pp. 3072–3078, doi: 10.1109/SSCI44817.2019.9002679.
- [12] C. H. Lin *et al.*, "Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry," *Computer Methods and Programs in Biomedicine*, vol. 190, 2020, doi: 10.1016/j.cmpb.2020.105381.
- [13] M. Daidone, S. Ferrantelli, and A. Tuttolomondo, "Machine learning applications in stroke medicine: Advancements, challenges, and future prospective," *Neural Regeneration Research*, vol. 19, no. 4, pp. 769–773, 2024, doi: 10.4103/1673-5374.382228.
- [14] 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.
- [15] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. Al 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 (ICECA)*, 2020, pp. 1464–1469, doi: 10.1109/ICECA49313.2020.9297525.
- [16] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. M. Khan, "Stroke disease detection and prediction using robust learning approaches," *Journal of Healthcare Engineering*, vol. 2021, 2021, doi: 10.1155/2021/7633381.
- [17] B. Akter, A. Rajbongshi, S. Sazzad, R. Shakil, J. Biswas, and U. Sara, "A machine learning approach to detect the brain stroke disease," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2022, pp. 897–901, doi: 10.1109/ICSSIT53264.2022.9716345.
- [18] S. Paliwal, S. Parveen, M. A. Alam, and J. Ahmed, "Improving brain stroke prediction through oversampling techniques: a comparative evaluation of machine learning algorithms," *Preprints*, pp. 1–18, 2023, doi: 10.20944/preprints202306.1444.v1.
- [19] S. Srivastav, K. Guleria, and S. Sharma, "Machine learning models for early brain stroke prediction: a performance analogy," in *2023 World Conference on Communication & Computing (WCONF)*, 2023, pp. 1–6, doi: 10.1109/WCONF58270.2023.10235070.
- [20] A. Srinivas and J. P. Mosiganti, "A brain stroke detection model using soft voting based ensemble machine learning classifier," *Measurement: Sensors*, vol. 29, 2023, doi: 10.1016/j.measen.2023.100871.
- [21] F. Soriano, "Stroke prediction dataset," *Kaggle*, 2020. Accessed: Jun. 25, 2024. [Online]. Available: <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>
- [22] T. D. Piyadasa and K. Gunawardana, "A review on oversampling techniques for solving the data imbalance problem in classification," *International Journal on Advances in ICT for Emerging Regions (ICTer)*, vol. 16, no. 1, pp. 22–31, 2023, doi: 10.4038/icterv.16i1.7260.
- [23] A. A. Aouragh and M. Bahaj, "Advancing breast cancer diagnosis with machine learning: exploring data balancing, feature selection, and bayesian optimization," in *2023 IEEE 6th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech)*, 2023, pp. 1–6, doi: 10.1109/CloudTech58737.2023.10366058.
- [24] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of classification methods on unbalanced data sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/ACCESS.2021.3074243.




- [25] W. Li, "Imbalanced data optimization combining K-means and SMOTE," *International Journal of Performability Engineering*, vol. 15, no. 8, pp. 2173–2181, 2019, doi: 10.23940/ijpe.19.08.p17.21732181.
- [26] A. A. Aouragh, M. Bahaj, and N. Gherabi, "Comparative study of dimensionality reduction techniques and machine learning algorithms for alzheimer's disease classification and prediction," in *2022 IEEE 3rd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, 2022, pp. 1–6, doi: 10.1109/ICECOCS55148.2022.9983211.
- [27] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, 2019, doi: 10.1186/s12911-019-1004-8.
- [28] D. Paul, G. Gain, and S. Orang, "Advanced random forest ensemble for stroke prediction," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 11, no. 3, 2022, doi: 10.17148/ijarccce.2022.11343.
- [29] S. Yarat, S. Senan, and Z. Orman, "A comparative study on PSO with other metaheuristic methods," *International Series in Operations Research and Management Science*, vol. 306, pp. 49–72, 2021, doi: 10.1007/978-3-030-70281-6_4.
- [30] T. Dokeroglu, E. Sevinc, T. Kucukyilmaz, and A. Cosar, "A survey on new generation metaheuristic algorithms," *Computers and Industrial Engineering*, vol. 137, 2019, doi: 10.1016/j.cie.2019.106040.
- [31] S. A. Hicks *et al.*, "On evaluation metrics for medical applications of artificial intelligence," *Scientific Reports*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-09954-8.
- [32] J. Lever, M. Krzywinski, and N. Altman, "Classification evaluation," *Nature Methods*, vol. 13, no. 8, pp. 603–604, 2016.

BIOGRAPHIES OF AUTHORS






Abd Allah Aouragh    is currently pursuing his Ph.D. at the MIET Laboratory, Faculty of Sciences and Techniques, Hassan First University, Settat, Morocco. His research is focused on advancing medical diagnosis support systems, with a particular emphasis on leveraging machine learning, deep learning, and computer vision techniques to develop innovative solutions. He can be contacted at email: abdallahaouragh@gmail.com.



Mohamed Bahaj    received his Ph.D. in Mathematics and Computer Science from the University Hassan 1st, Morocco, and currently serves as a Full Professor in the Department of Mathematics and Computer Sciences at the University Hassan 1st, Faculty of Sciences & Technology, Settat, Morocco. With a robust academic background, he has contributed over 60 peer-reviewed papers, spanning areas such as intelligent systems, ontologies engineering, partial and differential equations, numerical analysis, and scientific computing. He has provided valuable peer review services for various journals and mentored several Ph.D. students in computer sciences and mathematics. He actively engages in workshops, seminars, and academic forums to enhance teaching methodologies and research practices. He can be contacted at email: mohamedbahaj@gmail.com.



Fouad Toufik    received his Ph.D. in Computer Science from the University Hassan 1st, Settat, Morocco. He currently holds the position of Professor of Computer Sciences at the Higher School of Technology SALE, Mohammed V University, Morocco. With expertise in artificial intelligence, big data, and database architectures, his research interests lie at the intersection of these fields. His academic pursuits aim to advance knowledge and innovation in these areas, contributing to the development of cutting-edge technologies. He can be contacted at email: toufik.fouad@gmail.com.