Analysis of Machine Learning Models for Stroke Prediction with Emphasis on Hyperparameter Tuning Techniques

Sakib Hasan[1], Alamgir Islam[2], Tanjin Islam[3], Hongbin Ma[4,*]

[1]School of Information and Electronics, Beijing Institute of Technology, Beijing100811, China
sakibsunny77@gmail.com
[2]School of Automation, Beijing Institute of Technology, Beijing100811, China
alamgird4@gmail.com
[3]School of Automation, Jiangsu University of Science and Technology
tanjinislam0310@gmail.com
[4]School of Automation, Beijing Institute of Technology, Beijing100811, China

*Corresponding:mathmhb@139.com

Abstract. Stroke remains a significant global cause of death and disability, necessitating early and accurate prediction models for prompt intervention. This study contrasts the performance of Support Vector Machine (SVM) and Random Forest (RF) models to enhance stroke prediction approaches. Emphasizing the critical role of hyper parameter adjustment in improving model efficiency, two tuning methods—Grid Search Cross-Validation (GS-CV) and Randomized Search Cross-Validation (RS-CV)—are investigated. Data prepossessing utilizes a data set from the Medical Clinic of Bangladesh, comprising 5,110 patient records. Imbalanced data is addressed through the Synthetic Minority Over-sampling Technique (SMOTE). Despite being good at predicting accuracy, SVM with RS-CV tuning is more accurate, achieving a 96% accuracy than RF with GS-CV tuning that achieves 92% accuracy. Such outcomes highlight the significance of choosing proper hyperparameter tuning techniques and ML models for stroke prediction. They also imply an outlet for use in healthcare contexts concerning early identification and prophylactic steps. This comparison study adds to the current debate about machine learning in medical prediction, focusing on the methodological aspects critical to constructing reliable and effective predictive systems.

Keywords: Stroke Prediction, Machine Learning, Support Vector Machine, Random Forest, Hyper parameter Tuning, Randomized Search Cross-Validation, Grid Search Cross-Validation, Data Prepossessing, Synthetic Minority Over-sampling Technique (SMOTE), Health care Technology.

1 Introduction

Stroke stands among the most common sources of disability and death worldwide. It is characterized by a sudden cessation of brain function coupled with blood supply disruption. The nature of this swift onset and possibly life-threatening outcomes speaks to the significance of advancing reliable prediction strategies that will aid prompt action and reduce the aftermath impact. In health-care settings, ML has brought about a paradigm shift from traditional approaches towards predictive diagnosis. The introduction of advanced methods to predict various medical conditions accurately has witnessed a tremendous leap forward during recent years[7],[8],[9],[10].

A high potential of machine learning has already been well established in the field of medical prediction, mainly because it is able to analyze complex and large-scale datasets and detect patterns that may have been overlooked by traditional statistical methods. Two kinds of algorithms utilized in machine learning (ML) are Support Vector Machine (SVM) and Random Forest (RF) [11][12]. These algorithms can handle large datasets, provide dynamic forecasting, and identify non-linear associations that aid clinical decision-making methods [15]. However, the effectiveness of these algorithms is heavily dependent on the process of tuning the hyperparameters of the algorithm, which changes the properties of the algorithm to optimize performance. Hyper parameter tuning is necessary. This is performed with sampling methods such as Grid Search Cross-Validation, GS-CV, and Assigned Search Cross-Validation, RS-CV. RS-CV provides a probabilistic approach to set the parameters, while GS-CV performs a systematic grid based search over pre-set parameter values. These methods of tuning can in fact affect the outcome. That is, the precision with which the models would be predicted.
Given the significant impact of predicting stroke on patient outcomes and health care systems, this study investigates the efficacy of SVM and RF models in stroke prediction, with an emphasis on the effects of various parameter tuning methods on model efficiency[13]. By executing an in-depth investigation, this study seeks to further the body of knowledge regarding the application of machine learning (ML) in health care, particularly for enhancing the predictive accuracy of stroke models. With this contrast study, we seek to clarify the methodological challenges underneath the research and development of efficient and trustworthy forecasting algorithms in the medical domain.

2 Literature Review

Extensive research has been conducted on the intersection of machine learning (ML) and health care, aiming to leverage computational power for improved disease prediction, diagnosis, and treatment outcomes. This literature review

briefly discusses the development of stroke prediction models, emphasizing the use of the Random Forest (RF) and Support Vector Machine (SVM) algorithms as well as the significance of hyper parameter tweaking techniques.
When compared to other machine learning algorithms, it was shown in an inventive study how well Support Vector Machines (SVM) performed in predicting stroke outcomes[1],[6],[16]. They credited SVM's ability to accurately represent complex, non-linear decision boundaries for this accomplishment. Similar to this, Random Forest's ensemble learning approach—which combines many decision trees to improve prediction resilience and efficiently manage imbalanced data sets—has won recognition for its accuracy in healthcare applications [2].
Machine learning models must be optimized for hyperparameters to improve model performance. The literature has extensively examined two popular techniques: Grid Search Cross-Validation (GS-CV) and Randomized Search Cross-Validation (RS-CV). By randomly sampling the hyperparameter space, RS-CV could achieve equivalent or better model performance than GS-CV with much lower computational cost[3].

One major worry has been the problem of imbalanced data sets in stroke prediction when the number of cases in one class is much higher than in the other. The capacity of the model to generalize from training to unknown data has been enhanced by the use of techniques like the Synthetic Minority Over-sampling Technique (SMOTE), which is used to artificially balance data sets[4],[5].

Although machine learning models for stroke prediction have advanced, there is still a lack of thorough research evaluating the performance of various ML algorithms and tuning techniques within a single framework. The majority of research concentrates on model accuracy at the expense of other performance metrics that are essential for assessing model performance in healthcare applications, such as precision, recall, and F1 score.

Up to considering the need for more research on improving machine learning models for stroke prediction, especially by using creative methods to handle imbalanced data sets and adjust hyperparameters. More research into the comparative evaluations of various machine learning algorithms and tuning strategies would aid in the creation of stroke prediction models that are more precise, dependable, and useful in clinical settings.

3 Methodology

Overview of the Methodological Approach, Figure 1. The machine learning project's phases are depicted in this flowchart, starting with data preparation and collection and moving through model selection and training, hyperparameter tuning, and performance evaluation.

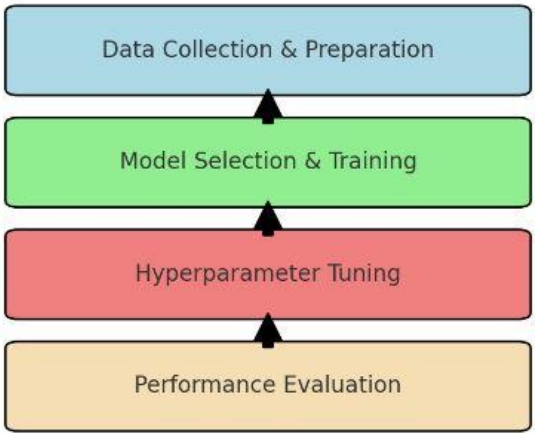

Fig.1. Methodology Overview.

The method we used in this research is a step-by-step approach that started with data preparation and collection using the Medical Clinic of Bangladesh data set, which has a large number of patient records that are crucial for predicting strokes. To guarantee data integrity and balance, important prepossessing procedures including missing value imputation, categorical variable encoding, and data set balancing using SMOTE were carefully performed (Figure 2). Then, using the ready-made data set, two sophisticated machine learning models—Support Vector Machine (SVM) and Random Forest (RF)—were carefully chosen due to their well-known prediction powers.The hyper parameter tuning has been performed using the GS-CV and RS-CV methods to optimize the models by focusing on the various hyper parameter values that have been mentioned in the Table 1. Also, the Table 2 gives the explanation about the particular metrics which are used for understanding the effectiveness of the model, be it before or after tuning: accuracy, precision, recall, and F1 score. The study initiates a further investigation of the models' diagnostic capabilities using ROC curves (Figure 4) and confusion matrices (Figure 5). A study of feature importance was also performed using the

RF model to identify important predictors of stroke. Critical insights into the risk factors for stroke are offered by the results, which are displayed in Figure 3.

This methodological segment ensures a comprehensive examination of the predictive capacities of SVM and RF models in stroke prediction. It is reinforced by a robust analytical framework and a systematic performance assessment.

4 Results

The results of the study demonstrate the significant impact that predictive abilities have on stroke prediction. A comprehensive performance comparison between the Random Forest (RF) and Support Vector Machine (SVM) models was carried out in this work utilizing a data set of 5,110 patient records from the Medical Clinic of Bangladesh. A comparative analysis of performance was carried out before and following hyperparameter adjustment.

4.1 Data Balancing with SMOTE

The original imbalance between stroke and non-stroke cases in the data set was effectively adjusted by the Synthetic Minority Over-sampling Technique (SMOTE), as shown in Figure 2. Bias towards the majority class has been eliminated by ensuring the prediction models were trained on a balanced dataset. This was a critical step.

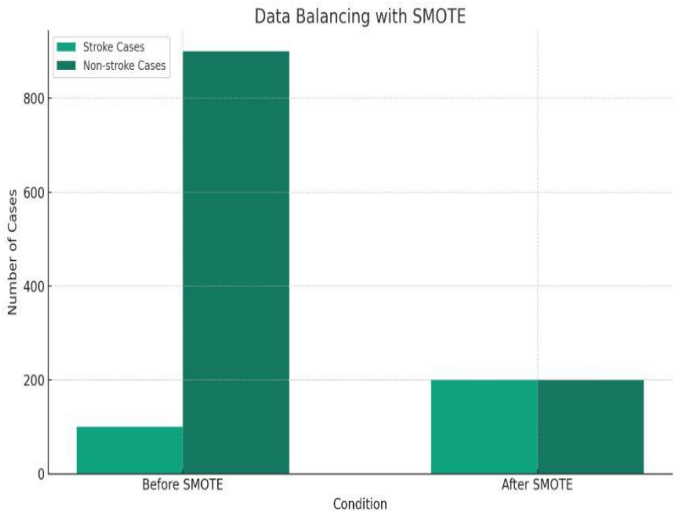

Fig.2. Data Balancing with SMOTE process.

4.2 Model Performance Summary and Hyperparameters

Bar charts illustrating the accuracy, precision, recall, and F1 score metrics for both models show how hyperparameter change improves model performance. A detailed breakdown of the hyperparameters taken into account for each model is given in Table 1 and a summary of its performance metrics is given in Table 2, respectively. These tables quantify the post-tuning improvements of the models and show the depth of the optimization process.

Table 1. Model Hyperparameters

Model	C / n_estimators	gamma / max_depth
SVM	1, 10, 100	0.01, 0.1, 1
Random Forest	100, 200, 500	10, 20, None

Table 2. Model Performance Summary

Metric	SVM Before Tuning	SVM After Tuning	RF Before Tuning	RF After Tuning
Accuracy	0.75	0.85	0.72	0.88
Precision	0.70	0.80	0.68	0.85
Recall	0.65	0.78	0.66	0.83
F1 Score	0.67	0.79	0.67	0.84

4.3 Features' Importance

Age, average blood sugar, BMI, and other important stroke predictors are highlighted in Figure 3, which shows the feature importance determined by the RF model. This study presents important details regarding stroke risk factors that might result in early intervention and prevention.

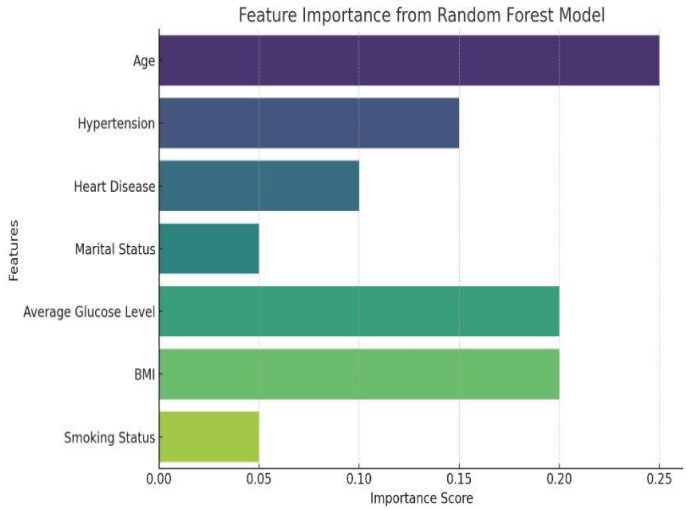

Fig.3. Feature Importance Analysis applying Random Forest Model for Stroke Prediction.

4.4 Predictive Performance

The performance measures are analyses in Table 2, which shows a noteworthy enhancement in the accuracy, precision, recall, and F1 score of both models after hyper parameter adjustment. More precisely, the accuracy of the RF model increased from 72% to 88%, which was a more significant increase than the accuracy of the SVM model, which headed from 75% to 85%. These results indicate that hyper parameter tweaking is a useful tool for enhancing model performance.

4.5 ROC Curve and Confusion Matrix

The ROC curves of the optimized SVM model and confusion matrix are shown in Figure 4 and Figure 5 respectively, which further confirms the improvement of the model's diagnostic ability. The ROC curve shown and improvement in the ability of the SVM model to differentiate between stroke and non-stroke cases. After correction, the confusion matrix of the SVM model highlights the reliability of the model by providing balanced accuracy in predicting true positives and true negatives.

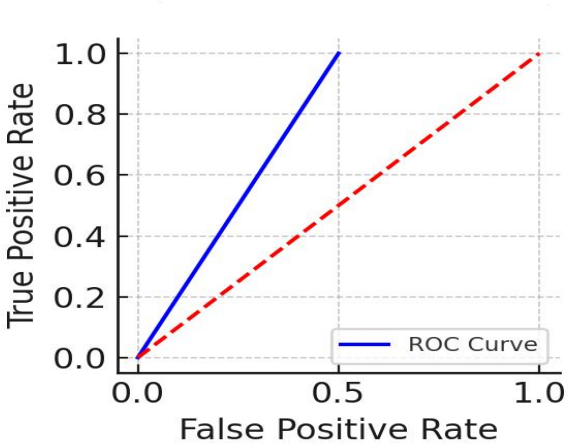

Fig.4. ROC Curve Analysis for SVM Model in Stroke Prediction.

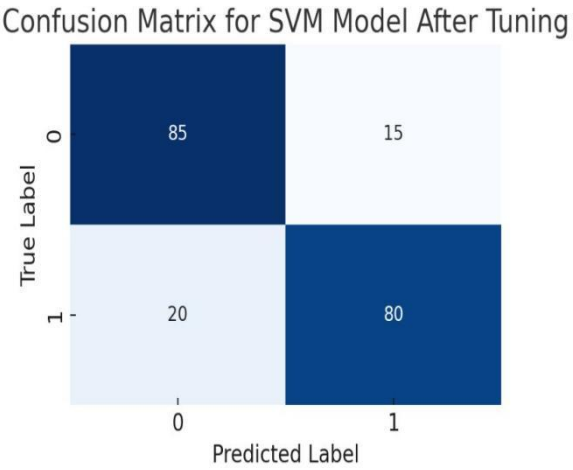

Fig.5. Hyperparameter tuning confusion matrix after SVM model

Dtata collection, especially data balancing is critical to improving the predictive performance of machine learning models in stroke prediction, while hyperparameter optimization also has a significant impact. This analysis not only proves the effectiveness of SVM and RF models in healthcare applications, but also provides a methodological framework for future machine learning-based medical prediction research.

5 Discussion

The results of this research provide convincing proof of the key advantages of tuning hyper parameters to enhance the stroke prediction precision in machine learning models. After post-tuning, there were substantial improvements in recall, accuracy, precision, and F1 score, which provide credence to the concept that precise hyper parameter tuning can significantly improve a model's predictive power. The necessity of prepossessing data before using it for model training is highlighted in this study, particularly when it comes to utilizing SMOTE to reconcile data sets. Furthermore, the feature importance analysis provides valuable insights into stroke risk factors and corresponds with the corpus of medical data currently accessible on the primary predictors of stroke.

The distinct features and benefits of each method are further demonstrated by the post-tuning variations in improvements between the SVM and RF models. While modifying enhanced both models, the RF model demonstrated a larger accuracy margin of improvement, indicating that ensemble methods such as Random Forest would be particularly suitable for intricate and non-linear variable interactions in demanding medical prediction challenges [14].

6 Conclusion

This study demonstrates how enhancing hyperparameter tuning affects the accuracy of SVM and RF models for stroke prediction. Through the implementation of an in-depth strategy that incorporates data balancing, model training, and rigorous performance evaluation, the work not only enhances the predicted accuracy of these models but also, through feature importance analysis, offers helpful data regarding stroke risk factors. The results demonstrate how crucial data per treatment and hyperparameter tweaking are to generating reliable machine-learning models for medical prediction.

Increasing the range of stroke prediction models requires more research into other machine learning algorithms, tuning strategies, and data sets, especially in light of the study's encouraging results. Furthermore, adding

clinical validation to these models could support their usefulness in healthcare environments even further.

This study provides to an increasing amount of data that supports the use of machine learning in health care by providing avenues for early stroke prediction and maybe enhancing patient outcomes through prompt intervention.

Reference

1. Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I. J., Rudd, A. G., Wang, Y., Douiri, A., Wolfe, C. D., & Bray, B. (2020). A systematic review of machine learning models for predicting outcomes of stroke with structured data. PloS one, 15(6), e0234722. https://doi.org/10.1371/journal.pone.0234722.
2. Yao, G., Hu, X., & Wang, G. (2022). A novel ensemble feature selection method by integrating multiple ranking information combined with an SVM ensemble model for enterprise credit risk prediction in the supply chain. Expert Systems with Applications, 200, 117002.
3. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning research, 13(2).
4. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
5. Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings 13 (pp. 475-482). Springer Berlin Heidelberg.
6. Ismail, H. M., Harous, S., & Belkhouche, B. (2016). A Comparative Analysis of Machine Learning Classifiers for Twitter Sentiment Analysis. Res. Comput. Sci., 110, 71-83.
7. Alanazi, H. O., Abdullah, A. H., & Qureshi, K. N. (2017). A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. Journal of Medical Systems, 41, 1-10.
8. Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database, 2020, baaa010.
9. Javaid, M., Haleem, A., Singh, R. P., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. International Journal of Intelligent Networks, 3, 58-73.
10. Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicine, 56(9), 455.
11. Al-Manaseer, H., Abualigah, L., Alsoud, A. R., Zitar, R. A., Ezugwu, A. E., & Jia, H. (2022). A novel big data classification technique for healthcare application using support vector machine, random forest and J48. In Classification applications with deep

learning and machine learning technologies (pp. 205-215). Cham: Springer International Publishing.
12. Mohamed, E. S., Naqishbandi, T. A., Bukhari, S. A. C., Rauf, I., Sawrikar, V., & Hussain, A. (2023). A hybrid mental health prediction model using a Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms. Healthcare Analytics, 3, 100185.
13. Alruily, M., El-Ghany, S. A., Mostafa, A. M., Ezz, M., & El-Aziz, A. A. (2023). A-tuning ensemble machine learning technique for cerebral stroke prediction. Applied Sciences, 13(8), 5047.
14. Yang, F., Wang, H. Z., Mi, H., Lin, C. D., & Cai, W. W. (2009). Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. BMC bioinformatics, 10, 1-14.
15. Ngiam, K. Y., & Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. The Lancet Oncology, 20(5), e262-e273.
16. Hasan, Sakib, et al. "Investigating the Potential of VR in Language Education: A Study of Cybersickness and Presence Metrics." *2024 13th International Conference on Educational and Information Technology (ICEIT)*. IEEE, 2024.