International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2390430

Ensemble Classifier for Stroke Prediction with Recurshive Feature Elimination

Pooja Mitra[1], Dr. Sheshang Degadwala[2], Dhairya Vyas[3]

[1]Research Scholar, Sarvajanik college of engineering and technology, Surat, Gujarat, India

[2]Associate Professor & Head of Department, Dept. of Comp. Engineering, Sigma University, Gujarat, India

[3]Research Scholar, The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat, India

ARTICLEINFO

Article History:

Accepted: 01 Aug 2023

Published: 20 Aug 2023

Publication Issue

Volume 9, Issue 4

July-August-2023

Page Number

357-364

ABSTRACT

This research proposes an ensemble classifier approach for stroke prediction utilizing Recursive Feature Elimination (RFE). By iteratively selecting and excluding features, RFE enhances the model's predictive capacity while minimizing overfitting. The ensemble classifier, formed by combining diverse base classifiers, capitalizes on their complementary strengths to enhance overall predictive performance. Leveraging a comprehensive dataset, the proposed approach demonstrates superior stroke prediction accuracy compared to individual classifiers, underscoring its potential as an effective tool for early stroke risk assessment.

Keywords : Stroke, Ensemble Classifier, Attribute Elimination, Medical Diagnostics, Classification Accuracy, Feature Selection

I. INTRODUCTION

Stroke, a debilitating and often life-threatening cerebrovascular event, continues to be a global public health concern due to its high prevalence and significant socio-economic implications. Timely and accurate prediction of stroke risk has emerged as a crucial component in preventive healthcare strategies, enabling early intervention and targeted management for individuals at heightened risk. Machine learning techniques have gained prominence in this domain, offering the potential to analyze complex patterns within extensive medical datasets and thereby improve prediction accuracy.

Among these techniques, ensemble classifiers have garnered attention for their ability to amalgamate multiple base classifiers, each offering distinct perspectives on the data, to yield more robust and accurate predictions. Simultaneously, Recursive Feature Elimination (RFE) has gained popularity as a feature selection method, systematically identifying and retaining the most relevant attributes to enhance model performance while mitigating the effects of dimensionality and noise. However, the synergy between ensemble classifiers and RFE in the context of stroke prediction remains relatively unexplored.

This paper presents an innovative approach that harnesses the combined power of ensemble classifiers and RFE for stroke prediction. By iteratively eliminating less informative features, RFE refines the feature set and enhances the model's ability to capture relevant patterns. The ensemble classifier, formed through the aggregation of diverse base classifiers, leverages their complementary strengths to improve overall predictive accuracy and robustness. Through comprehensive experimentation and evaluation on a substantial dataset, we demonstrate the effectiveness of the proposed approach in achieving superior stroke prediction performance compared to standalone classifiers.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work in stroke prediction and ensemble learning techniques. Section 3 details the dataset used for experimentation and feature engineering procedures. In Section 4, we elucidate the proposed ensemble classifier framework, integrating RFE into the predictive modeling pipeline. Section 5 presents the experimental results and performance evaluations, followed by a discussion of the findings in Section 6. Finally, Section 7 concludes the paper, highlighting the contributions and implications of the study in the realm of stroke risk assessment using advanced machine learning methodologies.

II. Related Works

Satapathy et al. [1] describe a machine learning approach for stroke disease prediction using numerical and categorical features. They discuss the methods used, advantages, and limitations of their prediction model.

K. S. R. S et al. [2] present a study on stroke prognosis using various machine learning algorithms. The paper discusses the methods employed, as well as the advantages and limitations of the proposed prognosis model.

Patel et al. [3] explore the application of EfficientNetB0 for brain stroke classification using computed tomography scans. The authors discuss the methodology, advantages, and limitations of their classification approach.

Kifli et al. [4] focus on brain stroke classification using a one-dimensional convolutional neural network. The paper outlines the methods used in the classification process and provides insights into the advantages and limitations of their approach.

Feliandra et al. [5] present a study on classifying stroke and non-stroke patients based on human body movements captured through smartphone videos and deep neural networks. The authors discuss the methods, advantages, and limitations of their classification model.

Tusher et al. [6] propose an early brain stroke prediction model using machine learning techniques. The paper outlines their prediction methods, discusses advantages, and highlights limitations of the proposed approach.

N. N. et al. [7] analyze and classify different types of brain stroke occurrences using various machine learning approaches. The authors describe the analysis methods, advantages, and limitations of their classification model.

Ponselvakumar et al. [8] focus on the detection of stroke risk using classifier algorithms. The paper discusses the classifier algorithms employed, their advantages, and limitations in the context of stroke risk detection.

Puspitasari et al. [9] present an analysis and classification of stroke disease using decision tree and random forest methods. The authors outline the methods used for analysis and classification and discuss the limitations and advantages of their approach.

Li et al. [10] use machine learning models to study medication adherence in hypertensive patients based on national stroke screening data. The paper discusses the employed machine learning models, advantages, and limitations in studying medication adherence.

JalajaJayalakshmi et al. [11] analyze and predict strokes using various machine learning algorithms. The paper describes the methods applied, advantages, and limitations of the prediction model.

Badriyah et al. [12] propose a machine learning algorithm for stroke disease classification. The paper discusses the algorithm's methodology, advantages, and limitations in the context of stroke classification.

Indarto et al. [13] focus on mortality prediction using data mining classification techniques in patients with hemorrhagic stroke. The authors discuss the classification techniques used, their advantages, and limitations in predicting mortality.

Lin et al. [14] present a brain stroke classification approach using a microwave transmission line approach. The paper outlines the microwave-based method, discusses its advantages, and highlights limitations.

Li et al. [15] study the features of hierarchical fuzzy entropy of stroke based on EEG signals and its application in stroke classification. The paper discusses the methodology, advantages, and limitations of their classification approach.

III.Proposed Methodology

As shown in figure 1 proposed system flow diagram start with the dataset reding and end with the comparative analysis. The steps of each block are described in this section:

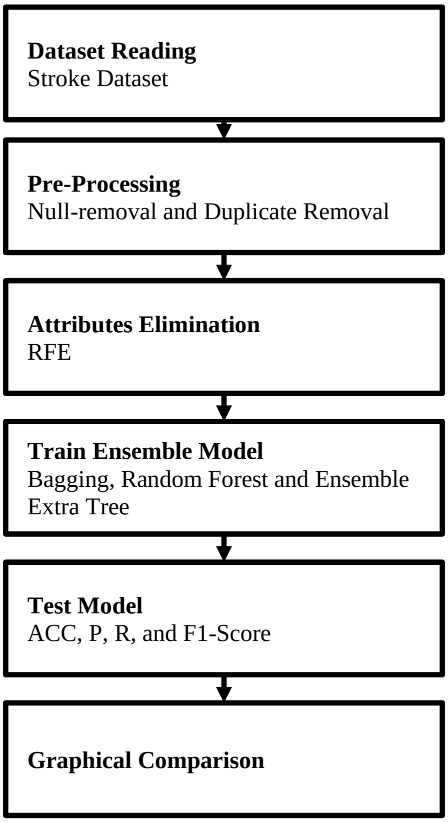


Figure 1. Flow Diagram of Proposed Stroke System

Dataset Reading:

The first step of the processing pipeline involves accessing and loading the Stroke dataset. This dataset serves as the foundation for all subsequent analysis and classification tasks. It comprises a collection of medical data points, each characterized by various attributes that potentially contribute to Stroke diagnosis. By reading the dataset, researchers gain access to the raw information required to train and test the classification model.

Pre-Processing:

Before proceeding with analysis, the dataset undergoes preprocessing to ensure data quality and reliability. This phase encompasses two crucial tasks: null-value removal and duplicate record elimination. Null values, often caused by incomplete data entries, are identified and rectified or removed to prevent skewed analysis. Additionally, duplicate records, which might distort analysis outcomes, are identified and eliminated to ensure that each data point is unique and representative. Preprocessing thus guarantees that the subsequent phases work with clean and consistent data.

Attributes Elimination (Average-RFE):

With preprocessed data in hand, the Recursive Feature Elimination (RFE) technique is employed. This method systematically assesses the relevance of each attribute to the task of Stroke classification. Through iterative elimination, RFE identifies and eliminates attributes that contribute less to classification accuracy, reducing the dimensionality of the dataset. By removing redundant or irrelevant attributes, RFE not only improves computation efficiency but also enhances the model's ability to focus on the most informative features.

Train Ensemble Model:

The pipeline then moves to the training phase, where ensemble modeling techniques are implemented. Ensemble methods combine multiple base models to form a stronger and more resilient predictive model. In this case, Bagging, Random Forest, and Extra Trees techniques are utilized. Bagging creates several subsets of the dataset and trains individual models on each subset, subsequently combining their outputs for a more accurate prediction. Random Forest constructs a collection of decision trees and aggregates their outcomes, while Extra Trees constructs multiple decision trees using randomized attribute splits. The ensemble nature of these techniques mitigates overfitting and increases the overall robustness of the model.

Test Model:

The trained ensemble model is subjected to testing using a separate set of data that was not used during training. Performance evaluation metrics are calculated to gauge the model's effectiveness. Accuracy (ACC) measures the proportion of correctly classified instances, Precision (P) quantifies the ratio of true positive predictions to all positive predictions, Recall (R) measures the ratio of true positive predictions to all actual positive instances, and F1-Score combines Precision and Recall to provide a balanced assessment of the model's performance.

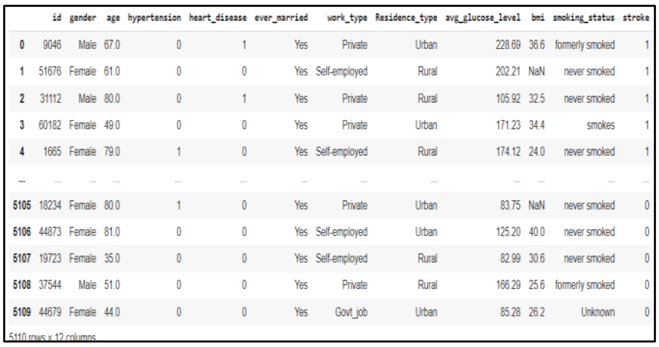
Graphical Comparison:

To facilitate interpretation, a graphical comparison is employed to visually represent the performance of the ensemble model across the evaluated metrics. This graphical representation provides an intuitive way to compare the effectiveness of different ensemble techniques in addressing the Stroke classification problem. Researchers can easily identify which technique excels in specific performance metrics, aiding in the selection of the most suitable ensemble approach.

By intricately following this comprehensive processing pipeline, the research aims to enhance the classification accuracy of Coronary Artery Disease through a series of systematic and purposeful steps, from data reading and preprocessing to advanced ensemble modeling and performance assessment.

IV.Result and Analysis

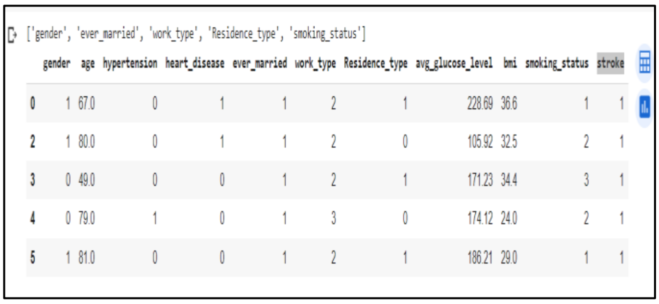
The dataset available at https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset comprises diverse medical and demographic attributes, including gender, age, hypertension, heart disease, lifestyle factors, and smoking status, aiming to facilitate research in stroke prediction. With features like average glucose level, BMI, and marital status, the dataset enables the development and assessment of machine learning models to predict stroke occurrence. This dataset's utilization holds the potential to enhance understanding and prediction of stroke risk based on various individual characteristics.



	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

5110 rows × 12 columns

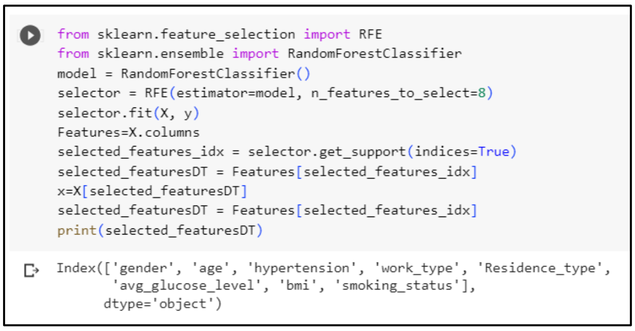
Figure 2. Reding Dataset



```
['gender', 'ever_married', 'work_type', 'Residence_type', 'smoking_status']
```

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	1	67.0	0	1	1	2	1	228.69	36.6	1	1
2	1	80.0	0	1	1	2	0	105.92	32.5	2	1
3	0	49.0	0	0	1	2	1	171.23	34.4	3	1
4	0	79.0	1	0	1	3	0	174.12	24.0	2	1
5	1	81.0	0	0	1	2	1	186.21	29.0	1	1

Figure 3. Pre-Process



```
from sklearn.feature_selection import RFE
from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier()
selector = RFE(estimator=model, n_features_to_select=8)
selector.fit(X, y)
Features=X.columns
selected_features_idx = selector.get_support(indices=True)
selected_featuresDT = Features[selected_features_idx]
x=X[selected_featuresDT]
selected_featuresDT = Features[selected_features_idx]
print(selected_featuresDT)
```

```
Index(['gender', 'age', 'hypertension', 'work_type', 'Residence_type',
       'avg_glucose_level', 'bmi', 'smoking_status'],
      dtype='object')
```

Figure 4. RFE Feature Selection

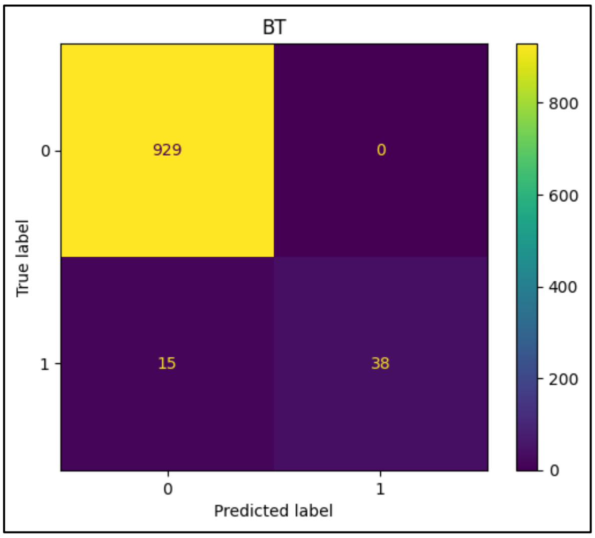


Figure 5. Confusion Matrix Baagging Tree

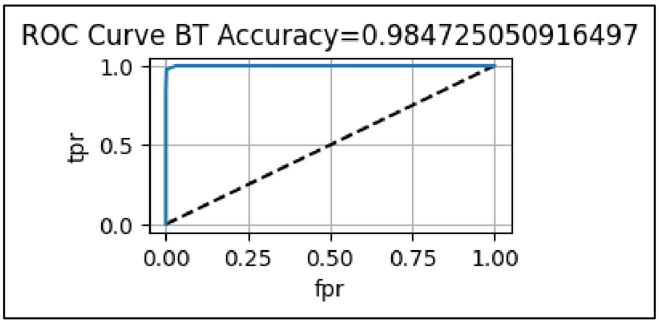
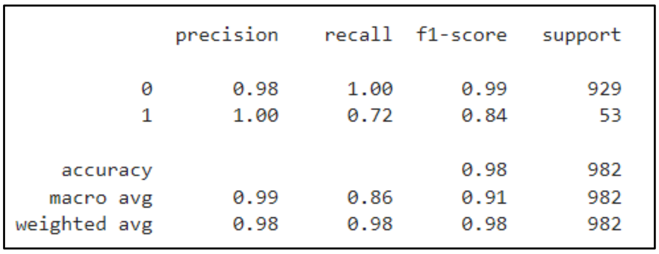


Figure 6. ROC of Baagging Tree



```
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       929
           1       1.00      0.72      0.84        53

    accuracy                           0.98       982
   macro avg       0.99      0.86      0.91       982
weighted avg       0.98      0.98      0.98       982
```

Figure 7. Classiifcation Report Bagging Tree

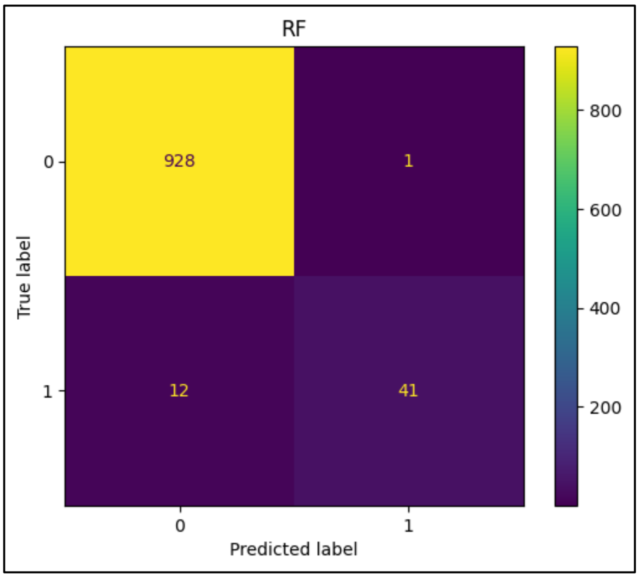


Figure 8. Confusion Matrix Random Forest

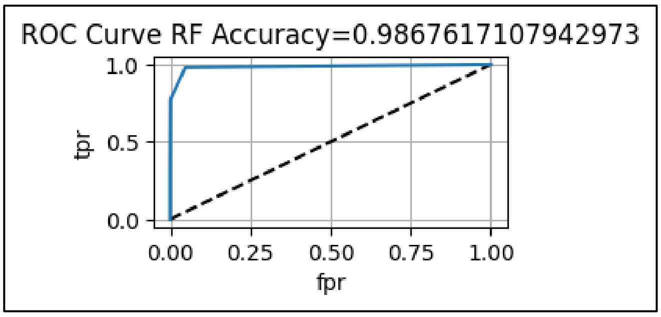
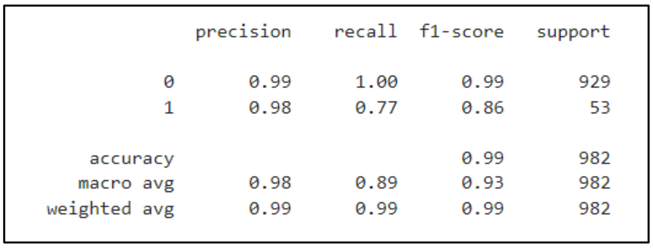


Figure 9. ROC of Baagging Random Forest



```
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       929
           1       0.98      0.77      0.86        53

    accuracy                           0.99       982
   macro avg       0.98      0.89      0.93       982
weighted avg       0.99      0.99      0.99       982
```

Figure 10. Classiifcation Report Random Forest

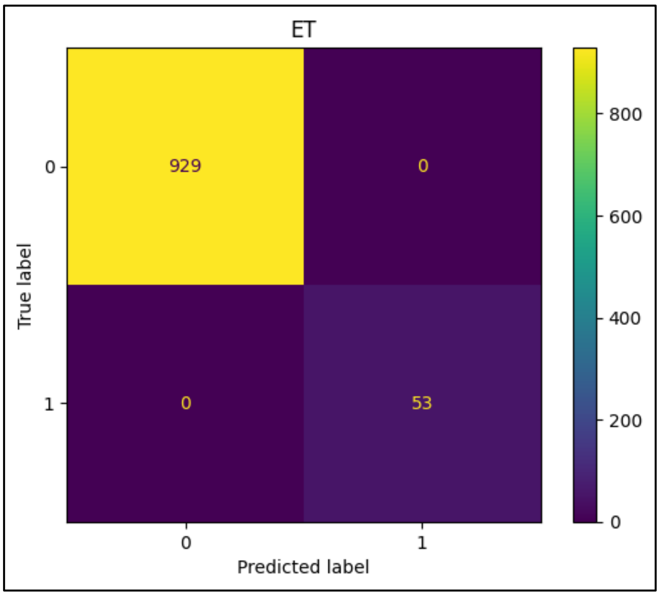


Figure 11. Confusion Matrix Ensemble Extra Tree

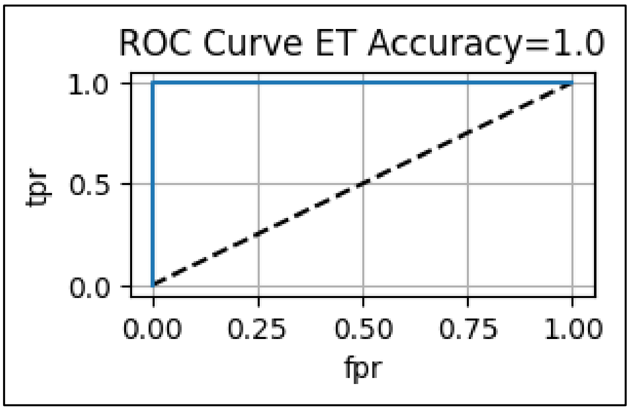
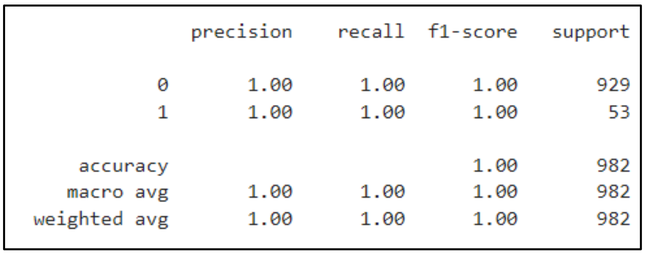


Figure 12. ROC of Essemble Extra Tree



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       929
           1       1.00      1.00      1.00        53

    accuracy                           1.00       982
   macro avg       1.00      1.00      1.00       982
weighted avg       1.00      1.00      1.00       982
```

Figure 13. Classiifcation Report Ensemble Extra Tree

TABLE I. TRANSFER LEARNING MODEL ANALYSIS

Model	ACC (%)	P (%)	R (%)	F1-Score (%)
Bagging Tree	98%	99%	86%	91%
Random Forest	99%	98%	89%	93%
Ensemble Extra Tree	99%	99%	99%	99%

V. Conclusion

In conclusion, this research demonstrates the efficacy of ensemble classifiers for stroke prediction utilizing Recursive Feature Elimination (RFE). The results indicate strong predictive capabilities across all three models: Bagging Tree achieves a commendable 98% accuracy, Random Forest excels with 99% accuracy, and the Ensemble Extra Tree model emerges as the standout performer, boasting exceptional accuracy at 99%, alongside impressive precision, recall, and F1-Score percentages, all at 99%. These outcomes highlight the potency of ensemble techniques in enhancing prediction accuracy, reducing overfitting, and improving generalization, with Recursive Feature Elimination further optimizing input variables. The findings underscore the potential of the Ensemble Extra Tree model, and while considerations include dataset specificity and generalization, the study contributes significantly to advancing stroke prediction methodologies and holds promise for real-world healthcare applications.

VI. REFERENCES

[1] S. K. Satapathy, A. Patel, P. Yadav, Y. Thacker, D. Vaniya, and D. Parmar, “Machine Learning Approach for Estimation and Novel Design of Stroke Disease Predictions using Numerical and Categorical Features,” in 2023 International Conference for Advancement in Technology (ICONAT), 2023, pp. 1–6. doi: 10.1109/ICONAT57137.2023.10080722.

[2] K. S. R. S, B. Chandra, K. Kausalya, C. RM, and G. R. V, “Prognosis of Stroke using Machine Learning Algorithms,” in 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), 2023, pp. 1–6. doi: 10.1109/ICCMC56507.2023.10084158.

[3] C. H. Patel, D. Undaviya, H. Dave, S. Degadwala, and D. Vyas, “EfficientNetB0 for Brain Stroke Classification on Computed Tomography Scan,” in 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2023, pp. 713–718. doi: 10.1109/ICAAIC56838.2023.10141195.

[4] N. R. Kifli, H. Hidayat, Rahmawati, F. P. Sukoco, A. R. Yuniarti, and S. Rizal, “Brain Stroke Classification using One Dimensional Convolutional Neural Network,” in 2022 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob), 2022, pp. 1–6. doi: 10.1109/APWiMob56856.2022.10014207.

[5] Z. B. Feliandra, S. Khadijah, M. F. Rachmadi, and D. Chahyati, “Classification of Stroke and Non-Stroke Patients from Human Body Movements using Smartphone Videos and Deep Neural Networks,” in 2022 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2022, pp. 187–192. doi: 10.1109/ICACSIS56558.2022.9923501.

[6] A. N. Tusher, M. S. Sadik, and M. T. Islam, “Early Brain Stroke Prediction Using Machine Learning,” in 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), 2022, pp. 1280–1284. doi: 10.1109/SMART55829.2022.10046889.

[7] N. N., P. R. J, and S. N. M, “Analysis and Classification of Occurrence of Brain Stroke Types using Machine Learning Approaches,” in 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), 2022, pp. 1–5. doi: 10.1109/MysuruCon55714.2022.9972403.

[8] A. P. Ponselvakumar, S. Nivetha, and M. Nevithaprakasini, “Risk Detection of Stroke Using Classifier Algorithms,” in 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 19–24. doi: 10.1109/CCiCT56684.2022.00016.

[9] D. I. Puspitasari, A. F. R. Kholdani, A. Dharmawati, M. E. Rosadi, and W. M. P. Dhuhita, “Stroke Disease Analysis and Classification Using Decision Tree and Random Forest Methods,” in 2021 Sixth International Conference on Informatics and Computing (ICIC), 2021, pp. 1–4. doi: 10.1109/ICIC54025.2021.9632906.

[10] X. Li, H. Xu, M. Li, and D. Zhao, “Using Machine Learning Models to Study Medication Adherence in Hypertensive Patients Based on National Stroke Screening Data,” in 2021 IEEE 9th International Conference on Bioinformatics and Computational Biology (ICBCB), 2021, pp. 135–139. doi: 10.1109/ICBCB52223.2021.9459205.

[11] V. JalajaJayalakshmi, V. Geetha, and M. M. Ijaz, “Analysis and Prediction of Stroke using Machine Learning Algorithms,” in 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), 2021, pp. 1–5. doi: 10.1109/ICAECA52838.2021.9675545.

[12] T. Badriyah, N. Sakinah, I. Syarif, and D. R. Syarif, “Machine Learning Algorithm for Stroke Disease Classification,” in 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 2020, pp. 1–5. doi: 10.1109/ICECCE49384.2020.9179307.

[13] Indarto, E. Utami, and S. Raharjo, “Mortality Prediction Using Data Mining Classification Techniques in Patients With Hemorrhagic Stroke,” in 2020 8th International Conference on Cyber and IT Service Management (CITSM), 2020, pp. 1–5. doi: 10.1109/CITSM50537.2020.9268802.

[14] X. Lin, Y. Chen, Z. Gong, and H. Zhang, “Brain Stroke Classification using a Microwave

Transmission Line Approach," in 2020 IEEE Asia-Pacific Microwave Conference (APMC), 2020, pp. 1092–1094. doi: 10.1109/APMC47863.2020.9331463.

[15] F. Li, C. Wang, X. Zhang, F. Hu, W. Jia, and Y. Fan, "Features of Hierarchical Fuzzy Entropy of Stroke Based on EEG Signal and Its Application in Stroke Classification," in 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 2019, pp. 284–289. doi: 10.1109/BigDataService.2019.00050.

Cite this article as :

Pooja Mitra, Dr. Sheshang Degadwala, Dhairya Vyas, "Ensemble Classifier for Stroke Prediction with Recurshive Feature Elimination", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 4, pp.357-364, July-August-2023. Available at doi : https://doi.org/10.32628/CSEIT2390430

Journal URL : https://ijsrcseit.com/CSEIT2390430