International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 9 Issue 12 December - 2025, Pages: 97-102

Soft Voting Ensemble Approach of Logistic Regression and Random Forest for Stroke Risk Prediction

Arinda Mahadesyawardani[1], Nur Chamidah[2], Marisa Rifada[3], Toha Saifudin[4]

[1,2,3,4] Department of Mathematics, Faculty of Science and Technology, Airlangga University
Surabaya, Indonesia
[1]arinda.mahadesyawardani-2022@fst.unair.ac.id
[2]nur-c@fst.unair.ac.id
[3]marisa.rifada@fst.unair.ac.id
[4]tohasaifudin@fst.unair.ac.id

Abstract: *Stroke is a major global health burden, making early risk prediction crucial for prevention and clinical decision-making. This study evaluates a Soft Voting Ensemble (SVE) that integrates Logistic Regression and Random Forest to enhance binary stroke classification. Using optimal parameters obtained through hyperparameter tuning with 5-fold cross-validation on the training set, the SVE consistently outperformed the individual models in both in-sample and out-of-sample evaluations. The ensemble achieved an in-sample F1-score of 0.80 and an AUC of 0.91, and an out-of-sample F1-score of 0.80 and an AUC of 0.89. Feature importance analysis identified age and lifestyle-related attributes as key contributors, aligning with established stroke risk factors. These findings highlight the capability of ensemble learning to support clinical assessment and risk stratification, offering a promising direction for developing more reliable stroke prediction systems.*

Keywords—soft voting ensemble; logistic regression; random forest; machine learning; stroke

1. INTRODUCTION

Stroke is a major non-communicable chronic disease with serious global implications and remains one of the leading causes of mortality and disability among adults. According to the Global Burden of Disease (GBD) 2021, stroke ranks as the third leading cause of death and the fourth leading cause of disability worldwide. The absolute number of stroke cases and deaths continues to rise significantly, driven by population aging, lifestyle transitions, and a range of complex contributing factors [1]. Despite substantial advances in stroke treatment, the global burden of the disease is expected to keep increasing, indicating that preventive efforts may face greater challenges than therapeutic approaches. This situation highlights the need for comprehensive research on stroke trends supported by technological advancements as part of early prevention strategies to mitigate future incidence rates and mortality risks.

Recent advances in technology have reshaped healthcare, particularly through the application of Machine Learning (ML). ML enables the development of accurate diagnostic and predictive models by identifying complex patterns within heterogeneous medical data, with its ability to analyze multivariate relationships and integrate diverse risk factors [2], ML offers considerable potential in building effective models for early stroke risk prediction. From a statistical perspective, risk factors such as advanced age, sex, hypertension, diabetes, metabolic syndrome, elevated BMI, blood glucose, and non-HDL cholesterol levels have been shown to be significantly associated with stroke incidence [3]. These biomarker-based indicators therefore hold strong potential to be incorporated into Machine Learning predictive models to enhance early identification of individuals at higher risk of stroke.

Previous studies have demonstrated the effectiveness of Machine Learning for stroke prediction. Patil et al. (2024) found that Random Forest (RF) achieved the highest accuracy of 94.85%, while Logistic Regression (LR) provided a strong balance between sensitivity and specificity (F1-score) of 90.84% [4]. These findings suggest that both classical linear models and ensemble-based algorithms can perform competitively. However, the previous study focused only on comparing individual models. Therefore, the present study proposes a Soft Voting Ensemble (SVE) that combines LR and RF, leveraging the complementary strengths of both. The linear decision of LR combined with the nonlinear modeling and robustness of RF are expected to improve predictive performance beyond single-model approaches.

The final prediction of the SVE is obtained by averaging the class probabilities generated by the base models and selecting the class with the highest probability, this allows the ensemble to mitigate the limitations of individual models and achieve more optimal predictive performance [5]. Recent work by Samuel & Pandi (2025) demonstrated that a weighted Soft Voting Ensemble can achieve strong predictive performance for stroke classification, reporting an accuracy of 92.31% using tree-based and gradient boosting models [6]. This evidence supports the feasibility of applying SVE for binary stroke prediction with different base models such as LR and RF. Furthermore, This study also aligns with global public health initiatives, particularly SDG 3.4, which emphasizes reducing non-communicable disease mortality through early prevention.

2. CLASSIFICATION METHODS

2.1 Logistic Regression

Binary Logistic Regression models the probability of a binary outcome (0 or 1) under the assumption that each observation follows a Bernoulli distribution. The probability of the positive class ($y = 1$) given a feature vector x is modeled using the logistic (sigmoid) function, as in

$$P(y = 1|x) = \sigma(x\beta) = \frac{1}{1+e^{-x\beta}} \quad (1)$$

The model parameters are estimated by maximizing the Bernoulli log-likelihood. Equivalently, this is formulated as minimizing the binary cross-entropy (log loss), which is differentiable and therefore suitable for optimization via gradient-based methods [7]. To prevent overfitting and to promote coefficient sparsity, an L1 regularization term is incorporated into the objective function as in

$$L_{L1}(\beta) = -\sum_{i=1}^{m}\left[y^{(i)}\log\left(\hat{P}^{(i)}\right) + \left(1 - y^{(i)}\right)\log(1 - \hat{P}^{(i)})\right] + \lambda\sum_{j=1}^{n}|\beta_j| \quad (2)$$

where $\hat{P}^{(i)}$ denotes the predicted probability of sample i, parameter λ controls the strength of the penalty, and β_j are the feature coefficients [8]. The L1 penalty encourages sparsity by driving less informative coefficients toward zero, effectively performing embedded feature selection. In this study, the *liblinear* solver is employed because it efficiently optimizes logistic regression with L1 regularization.

2.2 Random Forest

Random Forest is an ensemble classification algorithm that combines multiple independent decision trees to improve prediction accuracy. The model generates different training sets using bootstrap sampling and selects a random subset of features at each node to determine the best split. Each tree is grown fully without pruning, and the final prediction is obtained by aggregating the outputs of all trees, enabling the model to classify new data more reliably [9].

A common splitting criterion used in Random Forest is the Gini impurity. For a binary classification problem with class probabilities $p_{n,A}$ and $p_{n,B}$, the Gini impurity at node n is defined as

$$G = 1 - (p_{n,A}^2 + p_{n,B}^2) \quad (3)$$

During tree construction, the algorithm evaluates all candidate features and thresholds by partitioning samples into left and right subsets, and selecting the split that minimizes the weighted impurity

$$G_{split} = f_{left}\, G_{left} + f_{right}\, G_{right} \quad (4)$$

Where G_{left} and G_{right} denote the impurities of the resulting subsets. This process identifies the optimal feature and threshold that yield the lowest impurity, ensuring more homogeneous nodes and improving classification performance [10].

2.3 Soft Voting Ensemble

Rather than depending on a single model, ensemble learning integrates multiple classifiers to enhance generalization and improve predictive accuracy [11]. A widely applied ensemble strategy is the voting method, which aggregates outputs from individual models. In soft voting, the predicted class probabilities from each classifier are summed (optionally with weights), and the class with the highest average predicted probability across the base learners, expressed as

$$y = arg\,\max_j \frac{1}{N}\cdot\sum_{i=1}^{N} P_i^j(x) \quad (5)$$

where y denotes the selected class, N is the number of base models, and $P_i^j(x)$ represents the predicted probability for class j from model i.

2.4 Evaluation Metrics

Several standard evaluation metrics: accuracy, precision, recall, F1-score, and AUC-ROC are employed to assess the performance of the SVE model and its base classifiers. These metrics are derived from the confusion matrix components: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The calculation formulas for these performance measures are given below.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

$$F1\ Score = \frac{2\times precision\times recall}{precision+recall} \quad (9)$$

While the AUC-ROC evaluates how well a model distinguishes between classes. The ROC curve reflects the trade-off between true and false positives, while the AUC provides a single measure of overall class separability [2].

3. DATA AND PROCEDURE

This study uses secondary data from the Stroke Prediction Dataset sourced from Kaggle. In this study, only 510 patient observations with 9 clinical features were utilized to predict stroke events. These features include demographic, medical, and lifestyle-related variables, which are described in detail in Table 1. This dataset was selected because it is publicly accessible, widely used in health prediction research, and has a data structure well-suited for binary classification in Machine Learning.

3.1 Research Variable

The research variables are described in the following table.

Table 1: Research Variables

Variable	Description	Category	Scale
Y	Stroke	0=No stroke	Nominal

		1=Stroke	
X_1	Gender	0=Female 1=Male	Nominal
X_2	Age	-	Ratio
X_3	Hypertension	0=No hypertension 1=Hypertension	Nominal
X_4	Heart disease	0=No heart disease 1=Heart disease	Nominal
X_5	Ever married	0=No 1=Yes	Nominal
X_6	Residence type	0=Rural 1=Urban	Nominal
X_7	Average glucose level	-	Ratio
X_8	BMI	-	Ratio
X_9	Smoking status	0=Never smoke 1=Formerly smoked 2=Smokes	Ordinal

3.2 Research Procedure

The data analysis in this study was carried out through the following steps:

1. Exploratory Data Analysis.
   Conducting descriptive and graphical analysis to understand data distribution.
2. Predicting the risk of stroke using a binary classification approach using the SVE.
   a. Splitting the dataset into in-sample and out-sample partitions with a ratio of 90:10.
   b. Performing hyperparameter tuning on the in-sample data for LR and RF using Grid Search with Stratified 5-Fold Cross-Validation.
   c. Building classification models on the in-sample data using the best estimator obtained from tuning, selected based on the highest Macro F1 score.
   d. Combining the two best estimators through a SVE, where the final prediction is determined by averaging the predicted class probabilities.
   e. Predicting stroke events on the out-sample data using the base models and the ensemble model to generate final classification results.
   f. Showing feature importance.
3. Model Performance Evaluation.
   a. Comparing the performance of LR, RF, and SVE on both in-sample and out-sample data using evaluation metrics including accuracy, precision, recall, F1-score, and AUC.
   b. Selecting the optimal model based on evaluation outcomes and drawing conclusions regarding the most effective method for predicting stroke risk.

4. Result and Discussion

After following the research procedure, the results were obtained from the analysis.

4.1 Exploratory Data Analysis

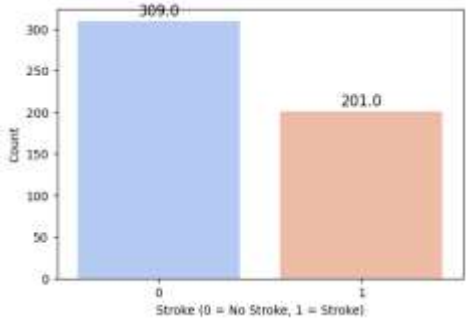

Fig. 1. *Distribution of Target Variable*

Based on Fig. 1, the distribution of the target variable shows an imbalance, with 309 patients categorized as non-stroke and 201 as stroke.

Table 2: Descriptive Statistics for Numerical Features

Feature	Mean	Std	Min	Max
Age	58.10	16.79	15	82
Average glucose level	117.58	55.5	55.78	271.74
BMI	29.53	5.49	14.10	48.90

For numerical featires based on Table 2 shows that Age ranges from 15 to 82 years with an average of about 58 years, indicating that most patients are older adults who are more vulnerable to stroke. The average glucose level is 117.58 with a relatively large standard deviation of 55.50, showing substantial variability in blood glucose among patients. Meanwhile, the mean BMI of 29.53 falls within the overweight category, suggesting that weight-related factors may also influence stroke outcomes.

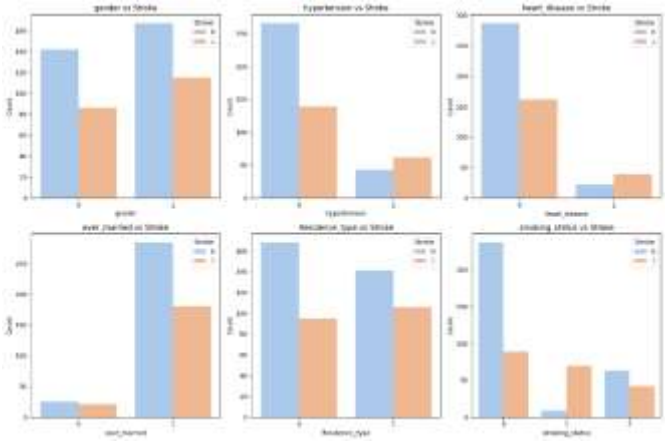

Fig. 2. *Descriptive Statistics for Categorical Features*

Based on Fig. 2, the categorical feature comparisons indicate that stroke cases are more prevalent among individuals with hypertension and heart disease. Patients with hypertension show a noticeably higher frequency of stroke compared to those without hypertension. A similar trend is observed for heart disease, where individuals with heart disease tend to have more stroke occurrences. Additionally, most stroke patients appear to belong to the "ever married" category, suggesting a correlation with age, since older individuals are more likely to be married.

Overall, the EDA results highlight that age, hypertension, heart disease, smoking status, and BMI may play important roles in predicting stroke events in this dataset.

4.2 Model Classification Performance

Numerically, the performance evaluation results demonstrate notable differences among the three classification models used in this study: Logistic Regression, Random Forest, and the Soft Voting Ensemble.

Table 3: ML Models Performance on Insample Data

Model	Accuracy	Precision	Recall	F1	AUC
LR	0.7909	0.7432	0.7183	0.7271	0.8412
RF	0.7995	0.7370	0.7682	0.7513	0.8663
SVE	0.8388	0.7801	0.8232	0.8011	0.9147

Based on Table 3, RF achieved higher performance than LR with an AUC of 0.8663, as well as improvements in recall and F1-score. This indicates that RF is more capable of capturing complex patterns within the training data. However, the SVE outperformed both base models. The SVE achieved an in-sample AUC of 0.9147, along with increased accuracy (0.8388) and F1-score (0.8011). These improvements suggest that combining probabilistic outputs from LR and RF enhances the model's capacity to generalize while leveraging the complementary strengths of both algorithms.

Table 4: ML Models Performance on Outsample Data

Model	Accuracy	Precision	Recall	F1	AUC
LR	0.8039	0.7083	0.8500	0.7727	0.8919
RF	0.7843	0.7143	0.7500	0.7317	0.8742
SVE	0.8235	0.7200	0.9000	0.8000	0.8935

Table 4 shows the out-of-sample evaluation reinforces these findings. While LR and RF achieved similar predictive capability with AUC scores of 0.8919 and 0.8742 respectively, the SVE slightly surpassed them with an AUC of 0.8935. The ensemble also obtained the highest recall (0.9000) and F1-score (0.8000), meaning it better identifies stroke-positive cases while maintaining a balanced precision–recall tradeoff. This is particularly valuable in medical prediction tasks where minimizing false negatives is critical.

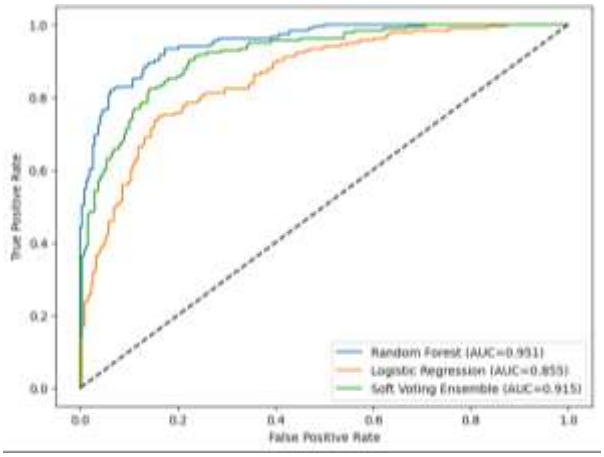

Fig. 3. *Comparison of AUC-ROC Performance on Insample*

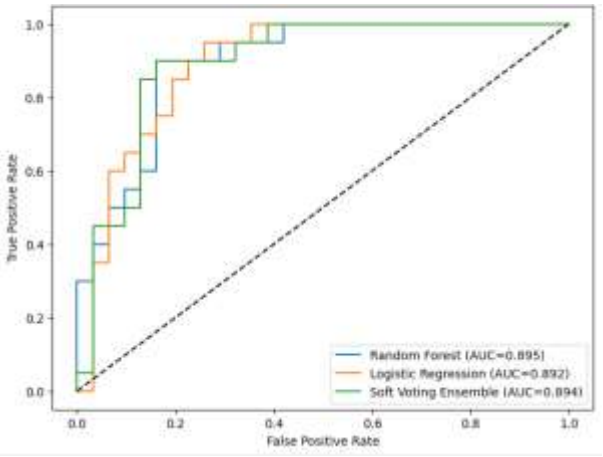

Fig. 4. *Comparison of AUC-ROC Performance on Outsample*

The ROC curve comparison on Fig. 3 and Fig. 4 visualizes these differences. In the in-sample plot, the RF curve lies closest to the top-left corner, with the SVE curve following closely, and LR positioned further below. The out-of-sample ROC curves show a reduced but consistent pattern: the SVE and LR curves are nearly overlapping, slightly above the RF curve. These curves confirm that the ensemble model maintains competitive separability across unseen data.

Overall, the SVE offers the best balance between discrimination ability and classification tradeoffs in both evaluation stages. This finding highlights the advantage of combining heterogeneous models to achieve more reliable

stroke prediction performance than either LR or RF individually.

4.3 Feature Importance

The Soft Voting Ensemble shows that age, smoking status, and marital status are the most influential predictors of stroke, with age contributing the highest importance. These importance scores were obtained by averaging the Random Forest importance and the normalized Logistic Regression coefficients.

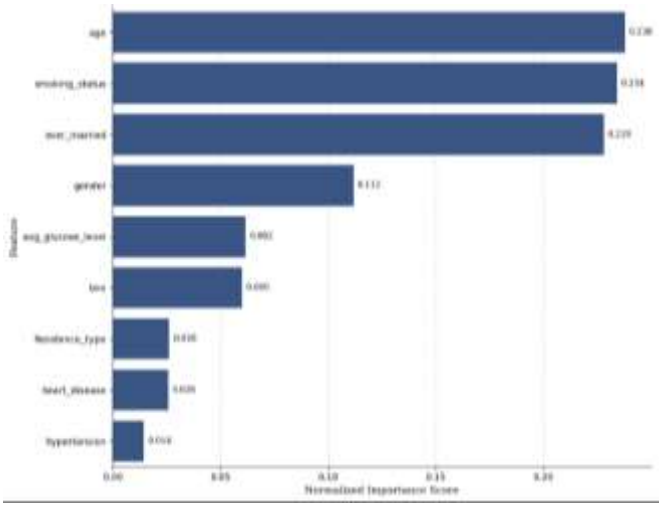

Fig. 5. *Feature Importance of SVE*

Based on Fig. 5, This indicates that both base models consistently emphasize demographic and lifestyle factors as key drivers of stroke risk. Gender provides a moderate contribution, while glucose level and BMI show weaker influence. Meanwhile, residence type, heart disease, and hypertension contribute the least. Overall, demographic and lifestyle characteristics dominate stroke prediction in this dataset compared with clinical indicators.

5. Conclusion

This study shows that the Soft Voting Ensemble combining Logistic Regression and Random Forest offers improved stroke prediction performance compared with the individual base models. The ensemble achieved higher F1-scores and competitive AUC values in both in-sample and out-of-sample evaluations, indicating better accuracy and generalization for binary classification.

Feature importance analysis highlights age, smoking status, and marital status as key predictors of stroke risk, while glucose level and BMI contribute moderately. These results confirm that both demographic and lifestyle attributes play significant roles in stroke prediction. Overall, the study demonstrates the potential of ensemble learning for medical risk prediction and provides a foundation for future research using larger datasets and more advanced ensemble approaches to further strengthen predictive capability.

6. Acknowledgment

The author would like to express sincere gratitude to the academic advisor of the Statistics Study Program at Airlangga University, as well as to all parties who contributed to this research. Appreciation is also extended to the dataset provider, whose open-access resource served as the foundation for this study. The author further acknowledges the support and constructive feedback from colleagues throughout the development of this work.

7. References

[1] L. Yang *et al.*, "Trends in stroke incidence and mortality in China, Japan, and South Korea (1990–2021) with projections to 2035," *Sci. Rep.*, vol. 15, no. 1, p. 25370, 2025, doi: 10.1038/s41598-025-10840-2.

[2] F. Asadi, M. Rahimi, A. H. Daeechini, and A. Paghe, "The most efficient machine learning algorithms in stroke prediction: A systematic review," *Heal. Sci. Reports*, vol. 7, no. 10, p. e70062, Oct. 2024, doi: https://doi.org/10.1002/hsr2.70062.

[3] T. Vu *et al.*, "Machine Learning Approaches for Stroke Risk Prediction: Findings from the Suita Study," *Journal of Cardiovascular Development and Disease*, vol. 11, no. 7. p. 207, 2024, doi: 10.3390/jcdd11070207.

[4] N. Patil and A. Sumarsono, "Stroke Prediction Using Machine Learning," *J. Res. Eng. Comput. Sci.*, vol. 2, no. 1, pp. 61–72, 2024.

[5] A. Srinivas and J. P. Mosiganti, "A brain stroke detection model using soft voting based ensemble machine learning classifier," *Meas. Sensors*, vol. 29, p. 100871, 2023, doi: https://doi.org/10.1016/j.measen.2023.100871.

[6] R. Samuel and T. Pandi, "Optimizing brain stroke detection with a weighted voting ensemble machine learning model," *Sci. Rep.*, vol. 15, no. 1, p. 31215, 2025, doi: 10.1038/s41598-025-14358-5.

[7] Y. S. Suh, S. K. Shin, D. Baang, and S. M. Seo, "A Brief Review of a Machine Learning Programming of Simple Logistic Regression," Oct. 2018.

[8] N. Alamsyah, B. Budiman, E. Setiana, and V. Claudia Jennifer, "The Role of L1 Regularization in Enhancing Logistic Regression for Egg Production Prediction ," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 10, no. 4 SE-Articles, pp. 821–832, May 2025, doi: 10.33480/jitk.v10i4.6409.

[9] Q. Ren, H. Cheng, and H. Han, "Research on machine learning framework based on random forest algorithm," *AIP Conf. Proc.*, vol. 1820, no. 1, p. 80020, Mar. 2017, doi: 10.1063/1.4977376.

[10] I. Reis, D. Baron, and S. Shahaf, "Probabilistic Random Forest : A Machine Learning Algorithm for Noisy Data Sets," *Astron. J.*, vol. 157, no. 1, p. 16, 2019, doi: 10.3847/1538-3881/aaf101.

[11] K. Akyol, E. Uçar, Ü. Atila, and M. Uçar, "An ensemble approach for classification of tympanic membrane conditions using soft voting classifier," *Multimed. Tools*

Appl., vol. 83, no. 32, pp. 77809–77830, 2024, doi: 10.1007/s11042-024-18631-z.