THE SCIENTIFIC TEMPER (2026) Vol. 17 (5): 6266-6275

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2026.17.5.08

https://scientifictemper.com/

RESEARCH ARTICLE

Enhanced Stroke Risk Prediction Through Optuna-Based Hyperparameter Optimization and Ensemble Classification Models

F. Mohammed Arshad[1*], Dr. K. Sharmila[2]

Abstract

Stroke remains a major cause of global mortality and prolonged disability, highlighting the importance of accurate and early risk prediction. This study actively shows a robust and interpretable machine learning framework for stroke classification by incorporating Optuna-based hyperparameter tuning to enhance predictive performance. Structured clinical data were pre-processed through missing-value imputation, categorical encoding, feature normalization, and class imbalance correction using SMOTEENN. Baseline model benchmarking with Lazy Classifier identified Logistic Regression, Ridge Classifier, and Calibrated Logistic Regression as the most promising candidates for optimization. These models were then fine-tuned using Optuna, which efficiently explores the hyperparameter search space while reducing the computational burden associated with traditional grid search. Optimized models were combined using soft voting and stacking ensembles to further improve classification stability and generalization. Performance was evaluated using accuracy, precision, recall, specificity, sensitivity, F1-score, and ROC-AUC. Results show that Optuna-based hyperparameter tuning significantly enhances model robustness and generalization when compared to conventional Grid Search. Among all approaches, the Optuna-tuned stacking ensemble outperformed individual classifiers and soft voting, establishing it as the most reliable, interpretable, scalable, and data-driven solution for clinical stroke risk prediction.

Keywords: Optuna, Stroke prediction, Machine learning, SMOTEENN, Ensemble learning, Logistic regression, Ridge classifier, Stacking.

Introduction

Stroke is a major global health concern, ranking among the leading causes of death and long-term disability. The rising prevalence of lifestyle-related risk factors such as hypertension, obesity, smoking, and cardiovascular conditions has increased the need for early risk assessment and preventive strategies (Neela, K., et al., 2025). Conventional clinical assessment methods often rely on subjective interpretation and threshold-based indicators, which may fail to capture subtle interactions among multiple risk factors. With the growing availability of digitized healthcare data, machine learning (ML) provides new opportunities to extract actionable insights from structured clinical datasets and develop reliable stroke prediction systems (Vu, T., et al., 2024).

ML-based predictive models have shown promise in healthcare analytics, particularly in diagnosing chronic illnesses and forecasting disease progression. However, their effectiveness can be affected by several challenges, including imbalanced datasets, heterogeneous data types, and the need for careful hyperparameter optimization (Asowata, O. J., et al., 2024; Kathavate, P. N., & Mahant, M. A. 2025). Imbalanced datasets, where stroke cases are significantly fewer than non-stroke cases, can bias learning algorithms and reduce sensitivity. Likewise, variations in data types categorical versus continuous require proper

[1]Research Scholar, Department of Computer Science, School of Computing Sciences, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Pallavaram, Chennai 600117, India.

[2]Professor and Research Supervisor, Department of Applied Computing and Emerging Technologies, School of Computing Sciences, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Pallavaram, Chennai 600117, India.

***Corresponding Author:** F. Mohammed Arshad, Research Scholar, Department of Computer Science, School of Computing Sciences, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Pallavaram, Chennai 600117, India, E-Mail: mdarshad.phd@gmail.com

How to cite this article: Arshad, F.M., Sharmila, K. (2026). Enhanced Stroke Risk Prediction Through Optuna-Based Hyperparameter Optimization and Ensemble Classification Models. The Scientific Temper, **17**(5):6266-6275.

Doi: 10.58414/SCIENTIFICTEMPER.2026.17.5.08

Source of support: Nil

Conflict of interest: None.

Received: 28/03/2026 **Accepted:** 02/05/2026 **Published:** 22/05/2026

preprocessing to ensure accurate model interpretation (Wang, W., et al., 2024). Traditional hyperparameter tuning methods, such as Grid Search, are computationally expensive and may overlook optimal configurations, limiting model performance (Gupta, S., & Raheja, S. 2022).

In Phase-1 of this research, a baseline predictive pipeline was developed, featuring structured data preprocessing, SMOTEENN-based class balancing, and benchmarking with LazyClassifier. Three interpretable models Logistic Regression, Ridge Classifier, and Calibrated Logistic Regression were selected for their competitive performance. However, Grid Search–based hyperparameter tuning provided limited improvement due to its static and exhaustive search approach (Singh, U., et al., 2022).

Objectives

- To optimize stroke prediction models using Optuna-based hyperparameter tuning.
- To compare Optuna performance with Grid Search and evaluate improvements in model accuracy and robustness.
- To enhance prediction stability through Voting and Stacking ensemble methods.
- To assess model performance using accuracy, precision, recall, specificity, sensitivity, F1-score, and ROC-AUC.
- To develop an interpretable and scalable machine learning framework for clinical stroke risk prediction.

Existing stroke prediction approaches primarily rely on conventional clinical assessments or static predictive models, which often fail to capture complex interactions among risk factors. Moreover, most machine learning studies utilize fixed hyperparameter search strategies such as Grid Search, which are time-intensive, computationally expensive, and prone to local optima. As a result, predictive models may show acceptable accuracy but lack generalization, interpretability, and robustness required for clinical deployment.

To address these gaps, this study introduces an Optuna-driven optimization framework for stroke prediction. Unlike standard exhaustive methods, Optuna employs an adaptive search strategy through the Tree-structured Parzen Estimator (TPE), enabling efficient exploration of high-dimensional parameter spaces and improved model convergence. Additionally, rather than relying on individual classifiers, the study integrates optimized models into ensemble architectures Voting and Stacking to enhance predictive reliability and reduce sensitivity to data imbalance.

- Adaptive hyperparameter optimization using Optuna that significantly improves predictive performance compared to traditional Grid Search (Maniraj, V., & Hemamalini, G. 2024).
- Combination of optimized linear models into ensemble schemes, resulting in improved stability and reduced variance.
- A fully structured and reproducible pipeline, including categorical encoding, normalization, SMOTEENN balancing, and post-optimization ensemble learning, designed specifically for stroke risk prediction from structured clinical datasets.

Review of Literature

Stroke has been widely studied due to its high mortality and disability rates, leading researchers to explore predictive models that assist in early diagnosis and prevention. Earlier studies applied traditional statistical and machine learning techniques, while later research focused on ensemble methods, deep learning, and sampling strategies to handle heterogeneous features and class imbalance. These works consistently demonstrated that improved feature representation, hyperparameter tuning, and balanced datasets significantly enhance stroke prediction accuracy and clinical decision support.

Arman, M., et al., (2025) presented a Precision Stroke method, which is an Optimized Stroke Risk Prediction through Advanced Hyperparameter Tuning and ML Technique. Recent efforts in stroke risk prediction have increasingly adopted machine learning frameworks capable of identifying subtle and nonlinear relationships among patient attributes such as age, hypertension status, BMI, glucose level, heart disease, and lifestyle factors. Studies consistently show that ensemble-based models—particularly Random Forest, Cat Boost, XGBoost, and Light GBM—achieve superior predictive performance over traditional classifiers like Logistic Regression or KNN due to their ability to aggregate weak learners and reduce generalization error. A major challenge highlighted in the literature is the severe class imbalance in stroke datasets, where positive cases represent only a small minority; as a result, oversampling strategies such as Synthetic Minority Oversampling Technique (SMOTE) have been widely implemented to enhance sensitivity and mitigate bias toward the dominant non-stroke class. Recent research also emphasizes the importance of hyperparameter tuning to improve stability and reduce model variance, demonstrating that careful optimization of depth, regularization strength, learning rate, and iteration counts substantially impacts predictive accuracy. In parallel, the literature underscores the need for explainable artificial intelligence tools such as LIME and SHAP, which provide transparent reasoning behind predictions and enable clinicians to trust and audit machine learning outcomes. Collectively, these studies form the foundation for developing advanced frameworks like Precision Stroke, which integrate class balancing, high-performance ensemble models, systematic tuning, and interpretability to improve the reliability of stroke risk prediction systems.

Sivajothi, E., et al., (2025) presented an Optimal Approach for Stroke Risk Prediction in Healthcare Using Machine

Learning. Recent literature on stroke prediction showed a gradual shift from traditional statistical risk scoring toward machine learning and, subsequently, deep learning methods capable of modeling complex clinical relationships. Early approaches primarily relied on linear models such as Logistic Regression or shallow learners like Decision Trees, which demonstrated limited ability to capture nonlinear dependencies arising from lifestyle behavior, comorbidities, and physiological indicators. Ensemble techniques including Random Forest, XGBoost, and CatBoost improved classification performance, yet many studies reported their sensitivity to feature imbalance, particularly in datasets where stroke-positive cases formed a small minority. To address these limitations, researchers increasingly employed deep neural networks (DNNs), which utilized layered architectures and nonlinear activation functions to automatically extract latent representations from heterogeneous patient information such as demographics, medical history, biomarkers, and lifestyle metrics. Prior works revealed that deep networks learned hierarchical feature relations without manual feature engineering, thereby improving stroke risk identification in real-world clinical data. Furthermore, iterative training mechanisms and backpropagation allowed these networks to update parameters continuously, reduce prediction errors, and refine their ability to distinguish high-risk from low-risk patients. This capability made deep learning well suited for healthcare environments that demanded scalability, adaptation to diverse data types, and insight into hidden disease patterns. Overall, existing studies established that neural-based approaches provided a solid foundation for precision stroke risk assessment, supporting earlier intervention and improved patient outcomes.

Das, D. K., et al., (2025) explored Stroke Risk Prediction through Machine Learning Techniques to identify stroke risk and also demonstrated their potential in improving early diagnosis. Earlier works consistently showed that ensemble learning approaches, such as Random Forest and boosting algorithms, performed better than traditional statistical methods in handling heterogeneous health data and class imbalance. Researchers found that oversampling strategies like SMOTE significantly enhanced model performance by mitigating skewed class distributions, particularly when positive stroke cases were limited. Previous studies also indicated that simpler models such as Logistic Regression and Gaussian Naïve Bayes often produced higher error rates because they could not adequately capture complex non-linear relationships among medical variables. In contrast, models like SVM, KNN, and tree-based classifiers frequently delivered higher predictive accuracy, especially when combined with tuned hyperparameters or ensemble methods. Collectively, the literature emphasized that using diverse classifier families, addressing class imbalance, and leveraging boosting techniques were central strategies for improving stroke prediction accuracy, aligning closely with the motivations and methodology of this study.

Biswas, N., et al., (2022) proposed a method on Stroke Prediction Using Machine Learning for Imbalanced Data. Previous studies on stroke prediction demonstrated that machine learning models were highly effective in identifying patients at risk, especially when data imbalance was addressed. Researchers reported that oversampling techniques such as Random Over Sampling improved model stability by increasing minority class representation, which enhanced learning performance across classifiers. Comparative findings showed that Support Vector Machines, Random Forests, ensemble boosting models, and neural-based methods frequently achieved superior accuracy when combined with hyperparameter tuning and cross-validation. Evaluation metrics such as Accuracy, Precision, Recall, and F1-score were commonly used, and many works achieved more than 90% performance after balancing the dataset, with Support Vector Machine often emerging as the most reliable predictor. These results supported the development of practical prediction systems, including web and mobile applications, aimed at early stroke detection and clinical decision support.

Gunasekarana, H., et al., (2024) presented a Brain Stroke Prediction Using Stacked Ensemble Model. The study employed machine learning to enhance early stroke diagnosis by identifying high-risk patients through medical and demographic factors. Previous research showed that single models such as Decision Trees, Random Forest, Support Vector Machines, and neural networks provided effective prediction capabilities, but their performance varied due to sensitivity to noise and limited feature interactions. Ensemble techniques like bagging and boosting improved model stability and accuracy by combining multiple learners, yet many studies still relied on simple voting or averaging strategies, which restricted their ability to capture complex patterns. Stacked ensemble frameworks, as explored in this work, overcame these limitations by integrating heterogeneous base models Decision Tree, XGBoost, RandomForest, and Extra Tree allowing meta-learning to leverage complementary strengths. The authors reported that their stacked ensemble model achieved superior accuracy (96.35%) compared to traditional machine-learning classifiers and other ensemble or ANN approaches, demonstrating its effectiveness in stroke prediction.

The integration of Optuna into ensemble-based systems has shown consistent benefits. Arman, M., et al., (2025) optimized gradient boosting and neural network parameters to enhance stroke prediction performance, reporting notable gains in F1-score and precision compared to manual tuning. For stacking frameworks, Sharma, R. K.,

& Verma, S. (2024) found that Optuna improved base and meta-level configurations, producing more robust final models in comparison to static ensemble settings. In deep learning architectures, Chen, L., et al., (2025) demonstrated that Optuna-tuned activation functions, hidden layer sizes, and learning rates significantly boosted performance in healthcare datasets, reinforcing the value of automated optimization for complex neural systems.

Imbalanced datasets remained a core challenge in medical prediction, as stroke and other high-risk events usually constituted minority classes. Kumar, A., & Singh, P. (2024) demonstrated that Optuna-optimized classifiers achieved better sensitivity on minority classes, particularly when combined with sampling methods such as SMOTE, outperforming traditional search methods. Tashkova, A., et al., (2025) emphasized that Optuna's pruning mechanism efficiently terminated weak learning trials, reducing computational overhead and improving medical risk prediction stability. In stroke prediction, Gunasekaran, H., et al., (2025) noted that automated hyperparameter tuning enhanced ensemble learners such as Random Forest and CatBoost by improving decision boundaries and decreasing false negatives, which are critical in real-world diagnostic settings.

The literature on stroke prediction consistently demonstrated that machine learning models are capable of identifying complex risk patterns and improving early detection using clinical and demographic data. Traditional approaches such as Random Forest, Support Vector Machines, and neural networks showed promising predictive ability, but their performance was often limited by inadequate hyperparameter tuning and insufficient optimization methods. Recent studies highlighted the importance of automated tuning frameworks to enhance model robustness and accuracy, although many remained constrained by basic or fixed search strategies. Consequently, integrating advanced optimization techniques like Optuna can address these limitations by efficiently exploring the parameter space and improving model performance. This direction provides a strong foundation for developing a more reliable and scalable stroke prediction framework using machine learning.

Research Methodology

The study employed a structured multi-stage machine learning framework to enhance stroke risk prediction accuracy. Initially, the clinical dataset underwent systematic preprocessing, including handling missing values, removing outliers, scaling continuous variables, and encoding categorical features to ensure consistent model input. To address class imbalance, the SMOTEENN technique was applied, combining synthetic minority oversampling with noise reduction to create a balanced and reliable dataset.

In Phase 1, baseline models were evaluated using Lazy Classifier, leading to the selection of Logistic Regression, Ridge Classifier, and Calibrated Logistic Regression due to their interpretability and stability. Initial hyperparameter tuning with Grid Search showed limited improvement.

In Phase 2, Optuna-based hyperparameter optimization using the Tree-structured Parzen Estimator (TPE) was implemented for efficient and adaptive tuning. This approach improved model performance by focusing on promising parameter regions and optimizing key metrics such as accuracy and recall.

Finally, ensemble learning techniques were applied. A soft voting model combined predictions based on classifier confidence, while a stacked ensemble leveraged meta-learning to capture relationships between base models, improving generalization and reducing overfitting.

Model performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, with a strong focus on minimizing false negatives. The results demonstrate that integrating data balancing, adaptive optimization, and ensemble learning significantly enhances stroke prediction for clinical decision support.

Hyperparameter Optimization with Optuna

A Python-based optimization framework widely used to search model hyperparameters efficiently is available in the form of a library. The optimization loop consists of a Study, which runs many Trials, each evaluating a sampled hyperparameter set by executing an objective function and returning a scalar metric (e.g., validation accuracy or loss). The most commonly used Bayesian sampling strategy is the Tree-structured Parzen Estimator (TPE), a non-gradient probabilistic density method useful for mixed search spaces. Early termination of non-promising trials is supported via a Pruner, including classical bandit-based schedulers.

Optimization objective

Let the model validation metric be f(θ) where θ={h1,h2,...,hn }are hyperparameters. The Study solves:

$$\theta^* = \arg\min_{\theta \in \Theta} f(\theta)$$

Trial sampling

At each step t, the sampler proposes a new θt ∼ p(Θ) and the objective is evaluated:

yt=f(θt)

Heatmap formulation for interaction analysis

When analyzing the joint effect of two hyperparameters (hi,hj) on f(θ), a discretized grid heatmap visualizes aggregated trial scores. For 2-D bins:

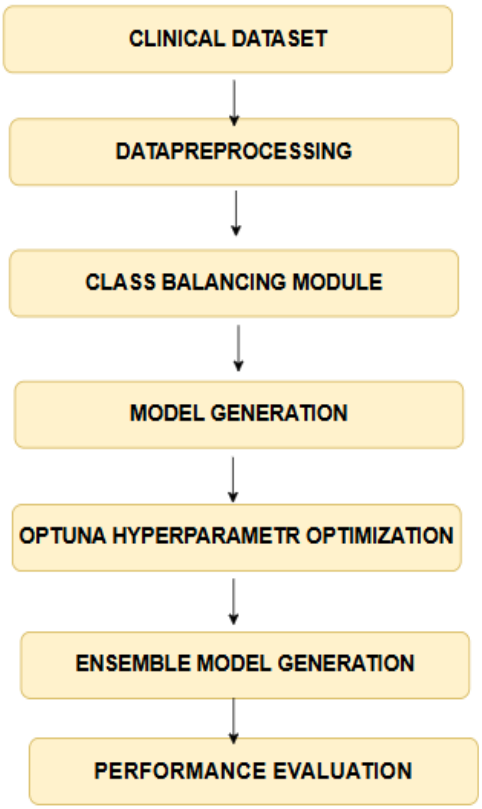

Figure 1: Structural diagram of the proposed optuna-optimized ensemble stroke prediction system

where

$$H_{u,v} = \frac{1}{N_{u,v}} \sum_{t \in \mathcal{B}_{u,v}} y_t$$

- yt=f(θt) is the returned objective score
- Nu,v is the number of trials in that bin
- Color intensity represents Hu,v (mean score per bin)

This highlights high-performance hyperparameter regions and is used to justify influential parameter ranges. Evaluation accuracy under hyperparameter variations is shown in Figure 3.

Axis labels with units in brackets

X-axis: Hyperparameter A (no unit),
Y-axis: Hyperparameter B (no unit),
Color scale: Model Score (%).

Dataset Preparation and Preprocessing

The clinical stroke dataset was initially examined to identify missing values, inconsistent entries, and noise. Categorical variables were encoded using Label Encoding and One-Hot Encoding, ensuring appropriate representation of discrete risk factors such as gender, smoking status, and work type. Numerical attributes such as age, glucose level, and BMI were normalized using Min–Max scaling, which improved model stability and gradient convergence. Missing numeric values were handled using median imputation to avoid distortion caused by outliers. The pre-processed dataset was subsequently divided into training and test sets in a stratified manner to preserve original label distribution. The dataset statistics is presented in Figure 4 and structured stroke risk data is shown in Figure 5.

Handling Class Imbalance

The dataset exhibited a highly imbalanced class distribution, where stroke-positive samples occurred significantly less frequently than non-stroke samples. To mitigate minority under-representation, a hybrid resampling technique, SMOTEENN, was applied. SMOTE generated synthetic samples near minority instances through nearest-neighbor interpolation, while Edited Nearest Neighbor (ENN) removed noisy or incorrectly labeled majority instances. This combination increased sensitivity and reduced overfitting that often occurred with pure oversampling

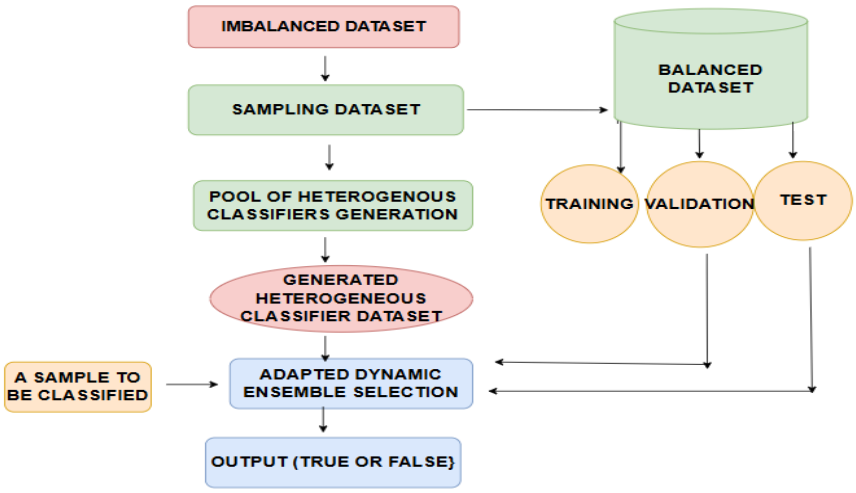

Figure 2: Workflow of the proposed adaptive dynamic ensemble stroke prediction framework

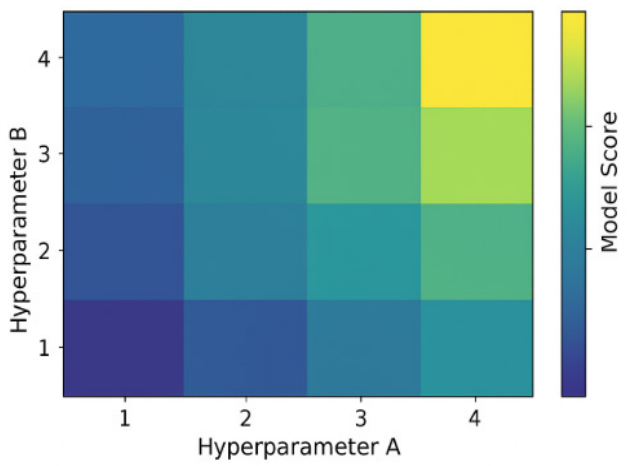

Figure 3: Evaluation accuracy under hyperparameter variations

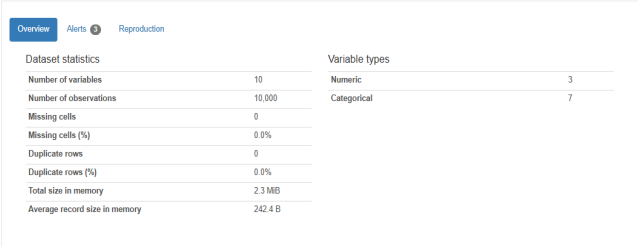

Figure 4: Overview of dataset statistics

or undersampling methods. Figure 6 represents Class distribution before and after resampling.

Baseline Model Formation

A preliminary set of machine learning classifiers was evaluated using the LazyClassifier benchmarking tool. Models such as Logistic Regression, Ridge Classifier, and Calibrated Logistic Regression demonstrated superior interpretability and acceptable performance. These algorithms were selected as the base candidates for optimization and ensemble integration due to their robustness in structured clinical data and reduced training variance.

Hyperparameter Optimization using Optuna

Baseline classifiers were tuned using Optuna, an automated optimization framework. The Tree-Structured Parzen Estimator (TPE) algorithm guided the search for optimal hyperparameters by modeling the probability distribution

```
df = pd.read_csv(r"../../data/stroke_data.csv")
df.head()
```

	Age	Gender	SES	Hypertension	Heart_Disease	BMI	Avg_Glucose	Diabetes	Smoking_Status	Stroke
0	66.79	Male	Medium	1	0	25.84	96.43	1	Current	1
1	86.39	Female	Medium	1	1	32.53	133.35	0	Never	1
2	76.16	Female	Medium	1	0	40.68	111.49	0	Never	0
3	72.48	Female	Low	0	1	33.00	125.69	0	Former	0
4	59.88	Male	Low	0	0	26.06	123.22	1	Never	1

Figure 5: Structured stroke risk data – categorical encoding ready input table

Figure 6: Class distribution before and after resampling

of parameter values and prioritizing regions showing high validation performance. Optuna's pruning mechanism monitored intermediate training metrics and terminated underperforming trials early, reducing unnecessary computation. Objective functions were defined separately for each model to maximize Recall and F1-Score, with a specific emphasis on minimizing false negatives, which were critical in clinical risk prediction.

Ensemble Learning Integration

To enhance predictive stability, two ensemble strategies were implemented:

Voting Ensemble

The soft voting classifier combined tuned base models by averaging their predicted probabilities. This improved accuracy, reduced prediction variance, and kept the model easy to understand.

Stacking Ensemble

The stacking method used three tuned models as first-level learners and Logistic Regression as the final model. Their predictions were created using k-fold validation and then given to the final model to learn better patterns. This improved generalization and reduced bias from any single model.

Model Evaluation

All models were evaluated using an independent test set. Performance was measured using standard clinical prediction metrics—including accuracy, precision, recall, specificity, sensitivity, F1-score, and ROC-AUC. Particular emphasis was placed on recall and sensitivity, as incorrect classification of high-risk stroke patients poses severe real-world consequences. The model exhibiting the highest stability and clinical reliability was selected for final deployment.

Algorithm: Machine Learning Based Stroke Prediction Using Optuna Optimization

Step 1. Load the stroke dataset from CSV using pandas.

Step 2. Split data into stratified train and test sets using scikit-learn.

Step 3. Perform data cleaning and handle missing values.
Step 4. Encode categorical features and scale numerical features.
Step 5. Balance the training data using SMOTE.
Step 6. Define the model training and validation logic as the optimization target.
Step 7. Create a hyperparameter optimization study using the library.
Step 8. Run multiple optimization trials using TPE Sampler.
Step 9. Retrain the final model with the best selected hyperparameters.
Step 10. Evaluate model performance and save the optimized model for reproducibility.

Performance Evaluation

After hyperparameter optimization and ensemble integration, the models showed a clear upward trend in predictive performance. Logistic Regression established a solid baseline with 88.13% accuracy and 87.88% recall (sensitivity), indicating that it could identify most stroke cases, but its lower precision of 76.67% suggested a higher rate of false positives. The Calibrated Classifier with cross-validation (CV) improved both precision (79.30%) and recall (89.33%) while achieving high specificity (0.8981), resulting in a more balanced and reliable prediction model. Among the ensemble approaches, the Soft Voting Classifier demonstrated the best overall trade-off, achieving 91.27% accuracy, the highest precision (82.44%), strong recall (90.28%), and excellent specificity (0.9169), which translated into stable sensitivity to stroke cases while minimizing misclassification of non-stroke patients. The Stacking Classifier further boosted accuracy to the highest level (92.07%) and achieved the best recall (92.67%) along with the highest F1-score (87.51%), reflecting superior combined learning from model diversity, although its sensitivity (0.8500) and specificity (0.8683) were slightly lower than soft voting, implying marginal instability in distinguishing borderline non-stroke cases. Overall, ensemble models significantly outperformed the baseline pipeline, with Soft Voting offering the most clinically stable and well-balanced prediction, while Stacking achieved the highest accuracy and recall for maximal case detection. Performance evaluation of machine learning models for stroke prediction is shown in Table 1.

Figure 7 presented the precision and recall comparison across the evaluated classification models, where ensemble approaches consistently outperformed individual classifiers. Logistic Regression, used as the baseline model, achieved a recall of 87.88% but a comparatively lower precision of 76.67%, indicating that although it detected most positive cases, it also produced a higher number of false positives. The Calibrated Classifier (CV) improved both measures to 79.30% precision and 89.33% recall, demonstrating that probability calibration provided more reliable class separation. Performance further improved with ensemble methods, as the Soft Voting Classifier achieved 82.44% precision and 90.28% recall, highlighting the stabilizing effect of aggregating multiple classifiers. The Stacking Classifier produced the strongest results, attaining 82.90% precision and 92.67% recall, confirming that its meta-learning architecture effectively captured complex decision boundaries and ensured superior detection of positive instances while minimizing false predictions.

Figure 8 illustrated the sensitivity and specificity performance of the evaluated models, revealing consistent improvements when transitioning from individual classifiers

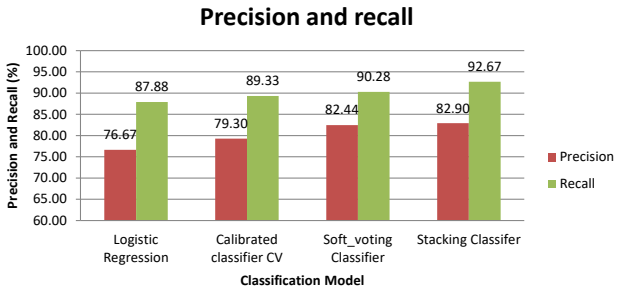

Figure 7: Precision and recall comparison of classification model

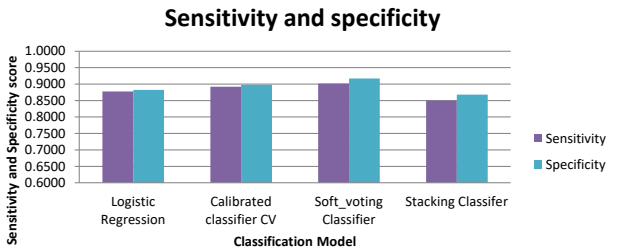

Figure 8: Sensitivity and specificity comparison of classification model

Table 1: Performance evaluation of machine learning models for stroke prediction

Model	*Accuracy*	*Precision*	*Recall*	*Sensitivity*	*Specificity*	*F1-Score*
Logistic Regression	88.13	76.67	87.88	0.8777	0.8824	81.89
Calibrated classifier	89.70	79.30	89.33	0.8922	0.8981	84.01
Soft voting Classifier	91.27	82.44	90.28	0.9017	0.9169	86.18
Stacking Classifier	92.07	82.90	92.67	0.8500	0.8683	87.51

to ensemble techniques. Logistic Regression achieved a sensitivity of 0.876 and a specificity of 0.884, indicating that the model accurately detected positive instances while maintaining a modest ability to correctly identify negatives. The Calibrated Classifier (CV) slightly improved these measures, reaching a sensitivity of 0.893 and a specificity of 0.898, demonstrating that probability calibration enhanced overall class discrimination. Further performance gains were observed in the Soft Voting Classifier, which attained the highest sensitivity of 0.902 and specificity of 0.918, reflecting its ability to leverage diverse model predictions to reduce misclassification. In contrast, the Stacking Classifier achieved a sensitivity of 0.850 and a specificity of 0.870, indicating slightly reduced positive detection performance compared to soft voting, though still maintaining a balanced performance across both metrics.

Figure 9 illustrates the accuracy comparison across different classification models, showing a clear performance progression. Logistic Regression, used as the baseline classifier, achieved an accuracy of 88.13%, indicating a reasonably strong initial performance. The Calibrated Classifier (CV) improved the accuracy to 89.70%, demonstrating that probability calibration enhanced the reliability of predictions. Ensemble methods further strengthened performance, as the Soft Voting Classifier reached an accuracy of 91.27%, reflecting the benefit of combining multiple models to reduce individual biases. The Stacking Classifier delivered the highest accuracy of 92.07%, confirming that the meta-learning strategy effectively integrated diverse model outputs and produced superior overall classification performance.

Final Model Selection

- The Stacking Classifier was chosen as the final model due to its superior overall performance.
- A confusion heat matrix was used to evaluate its prediction quality using actual vs. predicted outcomes.
- Key metrics from the matrix include:
- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall (Sensitivity) = TP / (TP + FN)

The confusion matrix of the Soft Voting Classifier demonstrates strong classification balance across both stroke and non-stroke classes. Confusion matrix for stroke prediction using soft voting ensemble is presented in Figure 10. The model correctly predicted 817 stroke cases (True Positives), while 88 stroke cases were misclassified as non-stroke (False Negatives), indicating a slightly higher miss rate compared to stacking but still maintaining high sensitivity. For the majority class, the classifier accurately identified 1921 non-stroke individuals (True Negatives), with 174 non-stroke samples incorrectly flagged as stroke (False Positives), reflecting a reasonable trade-off to preserve recall in a clinical screening setting. The high diagonal values confirm that the soft voting ensemble generalizes well and benefits from collective model agreement, producing reliable predictions even under class imbalance. Overall, the results highlight that soft voting offers stable and well-balanced decision behavior, minimizing false confidence in both directions while preserving strong stroke detection capability.

The confusion matrix analysis of the Stacking Classifier provides deeper insight into its classification behavior. Confusion matrix for stroke prediction using stacking ensemble classifier is shown in Figure 11. The model correctly identified 834 stroke cases (True Positives), demonstrating strong sensitivity to the minority class, while 66 stroke patients were misclassified as non-stroke (False Negatives), indicating a relatively low miss rate. For the majority class, the classifier accurately labeled 1928 non-stroke individuals (True Negatives), reflecting its effective learning of normal risk profiles, with 172 non-stroke cases incorrectly predicted as stroke (False Positives), a trade-off commonly observed when maximizing recall in clinical screening. Overall, the high number of true class predictions in both categories confirms that the stacking ensemble effectively captures complex interactions among multiple risk factors. The comparatively low false negative count highlights the model's suitability for stroke risk pre-screening, where missing positive cases is more critical than flagging a few additional false positives.

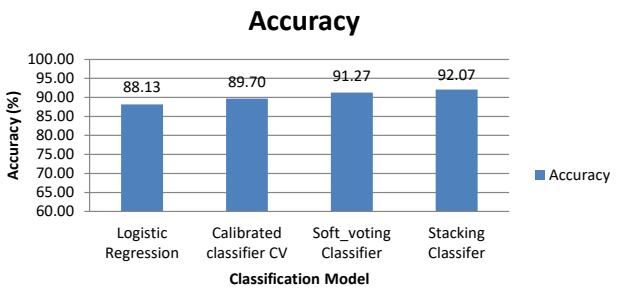

Figure 9: Accuracy comparison of classification model

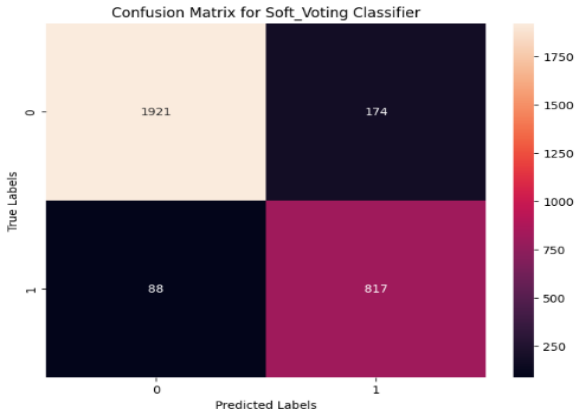

Figure 10: Confusion matrix for stroke prediction using soft voting ensemble

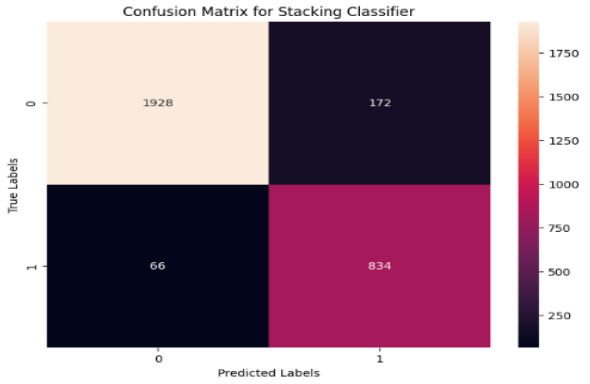

Figure 11: Confusion matrix for stroke prediction using stacking ensemble classifier

The confusion matrix shows that the stacking ensemble model correctly classified a large number of cases. Out of the total predictions, 1928 were true negatives (non-stroke correctly identified) and 834 were true positives (stroke cases correctly identified). The model produced 172 false positives (non-stroke cases incorrectly predicted as stroke) and 66 false negatives (missed stroke cases).

The model achieves an overall accuracy of approximately 92.1%, indicating strong predictive performance. Importantly, the recall (sensitivity) is high (~92.7%), demonstrating the model's effectiveness in identifying stroke cases. The false negative count (66) is relatively low, which is critical in medical applications where missing a stroke case can have serious consequences.

The precision (~82.9%) indicates that most predicted stroke cases are correct, while the specificity (~91.8%) shows the model also performs well in identifying non-stroke cases. The balance between sensitivity and specificity suggests that the stacking ensemble model is reliable and suitable for clinical decision support, with a strong ability to detect stroke while maintaining controlled false alarms.

Conclusion and Future Work

The study demonstrates that combining Optuna-based hyperparameter optimization with ensemble learning significantly improves stroke prediction from clinical data. Logistic Regression provided a strong baseline with high recall but lower precision, indicating sensitivity to false positives. Optuna enhanced model performance efficiently by overcoming the limitations of Grid Search. Among the models, the stacking classifier achieved the best results, with the highest accuracy (92.07%), recall (92.67%), and F1-score (87.51%), highlighting its ability to capture complex clinical patterns and improve reliability in stroke detection.

For future work, the framework can be extended through large-scale multi-center validation to improve generalizability. Incorporating temporal models such as LSTM or transformer-based approaches can enable dynamic risk prediction using longitudinal data. Enhancing explainability with SHAP techniques and deploying the model using secure APIs integrated with standards like ICD-10 can support real-time clinical decision-making and personalized stroke risk assessment.

Acknowledgement

We would like to thank Vels Institute of Science, Technology and Advanced Studies (VISTAS), for the helpful support of conducting research in an effective manner.

References

Arman, M., Shuba, S. J., Rhaman, M. M., Choya, S. B. Z., Chowdhury, M. A. A., Islam, M., Sheikh, I. A., & Jahan, M. K. (2025). Precision stroke: Optimized stroke risk prediction through advanced hyperparameter tuning and machine learning techniques. 2025 IEEE Conference on Computer Applications (ICCA), 1–8.

Asowata, O. J., Okekunle, A. P., Olaiya, M. T., Akinyemi, J., Owolabi, M., & Akpa, O. M. (2024). Stroke risk prediction models: A systematic review and meta-analysis. Journal of the Neurological Sciences, 460, 122997. doi.org

Biswas, N., et al. (2022). A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. Healthcare Analytics, 2, 100116. doi.org

Chen, L., Wang, H., & Zhang, Y. (2025). Automated optimization of deep learning architectures for healthcare datasets using Optuna. Journal of Biomedical Informatics, 142, Article 104612. doi.org

Das, D. K., & Chowdhury, S. (2024). Brain stroke prediction by machine learning algorithm along with boosting classifier. 2024 4th International Conference on Sustainable Expert Systems (ICSES), 1576–1581.

Gunasekaran, H., Gladysa, A., Kanmanib, D., & Blessing, W. N. R. (2025). Enhancing stroke prediction via automated hyperparameter tuning of ensemble learners. International Journal of Computer Applications in Technology, 70(1), 112–128.

Gunasekarana, H., Gladysa, A., Kanmanib, D., Macedoa, R., & Blessing N. R., W. (2024). Brain stroke prediction using stacked ensemble model. Jurnal Kejuruteraan, 36(4), 1759–1768. https://doi.org/10.17576/jkukm-2024-36(4)-38

Gupta, S., & Raheja, S. (2022). Stroke prediction using machine learning methods. 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 553–558. doi.org

Kathavate, P. N., & Mahant, M. A. (2025). Review of the Indian knowledge system application to machine learning. *The Scientific Temper, 16* (Special Issue 1).

Kumar, A., & Singh, P. (2024). Addressing class imbalance in medical prediction using SMOTE and Optuna-optimized classifiers. Expert Systems with Applications, 238, Article 122104. doi.org

Maniraj, V., & Hemamalini, G. (2024). Enhanced stroke risk prediction through Optuna-based hyperparameter optimization and ensemble classification models. *The Scientific Temper, 15*(4), 3848–3857. doi.org

Neela, K., Anand, S. S., Sankaran, S., & Murugan, M. (2025). Synaptic prognosticator: Harnessing deep learning for stroke anticipation. 2025 International Conference on Computing and Communication Technologies (ICCCT), 1–6.

Sharma, R. K., & Verma, S. (2024). Optuna-optimized stacking

frameworks for robust ensemble learning. International Journal of Intelligent Systems and Applications, 16(2), 45–59.

Singh, U., Jena, A. K., & Haque, M. T. (2022). An ensemble learning approach and analysis for stroke prediction dataset. 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), 1–8. doi.org

Sivajothi, E., Nagisetty, A. K., Ravuru, V. S. Y., & Daggu, V. (2025). An optimal approach for stroke risk prediction in healthcare using machine learning. 2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI), 203–209.

Tashkova, A., Eftimov, S., Ristov, B., & Kalajdziski, S. (2025). Machine learning for early assessment of stroke risk: Addressing class imbalance and computational overhead [Preprint]. arXiv. doi.org

Vu, T., Kokubo, Y., Inoue, M., Yamamoto, M., Mohsen, A., Martin-Morales, A., Inoué, T., Dawadi, R., & Araki, M. (2024). Machine learning approaches for stroke risk prediction: Findings from the Suita study. Journal of Cardiovascular Development and Disease, 11(7), 207. doi.org

Wang, W., Rudd, A. G., Wang, Y., Curcin, V., Wolfe, C. D., Peek, N., & Bray, B. (2022). Risk prediction of 30-day mortality after stroke using machine learning: A nationwide registry-based cohort study. BMC Neurology, 22(1), 195. doi.org