A Hybrid SHAP and Correlation Based Feature Selection Framework for Stroke Prediction

No Author Given

No Institute Given

Abstract. Stroke remains a leading cause of mortality , necessitating the development of effective predictive models for early intervention. This study proposes a robust framework for stroke prediction using the XGBoost algorithm, combining standard clinical features with engineered features, optimized through advanced feature selection techniques. We employ correlation based filtering using Spearman rank correlation to eliminate redundant variables and SHapley Additive exPlanations (SHAP) based ranking to identify the most important features. The selected features are evaluated through a comprehensive set of metrics. Our approach achieved a classification accuracy of **98.1%**, F1-score of **98.24%**, ROC-AUC of **99.75**, and precision of **99.03%**. These results advance stroke prediction by simultaneously achieving high performance and deployable efficiency in real-world healthcare settings.

Keywords: Stroke Prediction· XGBoost· SHAP· Machine Learning· Feature Selection· Correlation Based Filtering· Engineered Features.

1 Introduction

Stroke is a severe vascular disorder characterized by obstruction or reduction in blood flow and oxygen supply to the brain, According to the World Health Organization, approximately 15 million people suffer from stroke each year, leading to around 5 million deaths and another 5 million survivors experiencing long-term disabilities [1]. This growing global burden highlights the urgent need for effective early detection and prevention tools such as machine learning (ML). Despite recent advances in ML, many existing stroke prediction models face significant limitations in terms of handling class imbalance, limited interpretability and fail to leverage domain knowledge for feature engineering. This study addresses these limitations by developing a stroke prediction model using the Extreme Gradient Boosting (XGBoost) classifier. XGBoost is a boosting algorithm known for its capability to handle imbalanced datasets effectively, it operates by incrementally building decision trees, where each subsequent tree is designed to correct the errors of the previous one [2]. our model combines feature engineering, class imbalance handling using SMOTEENN, and a two-stage feature selection strategy integrating Spearman correlation filtering and SHAP based ranking. Two distinct feature sets standard and engineered were evaluated through a comprehensive set of metrics including accuracy, precision, recall, F1-score and ROC-AUC. Our

approach demonstrates superior performance (accuracy of **98.1%**, ROC-AUC of **99.75**, precision of 9**99.03%**) while involving an exhaustive search for optimal probability thresholds, maximizing the F1-score of **98.24%**. The remainder of this paper is structured as follow: Section **??** reviews related work on stroke prediction using ML. Section 3 details the proposed methodology. Section 4 presents the experimental results and discussion, including comparative analyses. Finally, Section 5 concludes the study and outlines directions for future research.

2 Related Work

The prediction of stroke risk using ML has attracted considerable attention in recent years due to its potential to support early intervention and clinical decision-making. The most recent studies that predict stroke risk using ML approaches will be discussed in this section. In [3] authors developed an ensemble model that combines the random forest (RF), logistic regression (LR), and XGBoost techniques; the proposed model achieves a notable accuracy of 97.4%. In [4] authors proposed combining XGBoost with optimized principal component analysis (PCA) and explainable AI (XAI) to enhance both the efficiency and interpretability of stroke risk prediction models. The proposed approach was tested on two datasets, achieving accuracies of 95% in the same dataset. In [5] authors proposed PCA-FA (Integration of Principal Components and Factors) and FPCA (Factor-Based PCA) to enhance feature representation and improve learning algorithm performance. The random forest approach achieved the best results with an accuracy rate of 92.55% and an AUC score of 98.15%. In [6] authors employed logistic regression for stroke prediction, The model achieves over 95% accuracy, with comparative analysis demonstrating the interesting of regularization techniques. In [7] authors used ML to develop robust models with a focus on a sophisticated ensemble methodology. The Stacked Ensemble model outperforms individual classifiers (Decision Tree, XGBoost, Random Forest) with superior accuracy (97.9%), precision (98.2%), and F1-score (98.0%). In [8] authors implemented a ML pipeline to predict stroke with SMOTE and random oversampling techniques. Among the models tested the Random Forest with random oversampling achieved the best performance, with a recall of 67% and an AUROC of 84% In [9] authors proposed an optimized hybrid system combining feature importance selection (36.3% feature reduction) with Random Forest classification, achieving 97.17% prediction accuracy through rigorous comparison of five classifiers and three feature selection methods on SMOTE-balanced stroke data. In [10] authors introduced a Dense Stacking Ensemble (DSE) model that achieves 96% accuracy and 98.92% AUC on balanced data through integration of SMOTE oversampling and multiple imputation techniques. In [11] authors demonstrated that the Random Forest classifier, when combined with advanced preprocessing techniques like SMOTE and grid search cross-validation, achieved a high accuracy of 94.42% in stroke risk prediction. In [12] authors demonstrate the effectiveness of XGBoost for stroke risk prediction, achieving 95% accuracy and a 0.93% F1-score. In [13] authors presented a comparative study of seven

ML models, including logistic regression, SVC, decision trees, XGBoost, and deep neural networks, for stroke prediction. Before feature selection, the accuracy of various models ranged between 91% and 94%, DNN achieve superior performance (AUC: 82%). In [14] authors demonstrated that neural networks achieve superior stroke prediction accuracy (95.16%) compared to other ML models (logistic regression: 95.04%, random forest: 95.10%). In [15] authors implemented ML methods for stroke prediction have demonstrated that XGBoost, combined with SMOTE for data balancing, achieved the highest accuracy of 95%, outperforming Random Forest (94%) and Logistic Regression (82%). These results highlight the effectiveness of ensemble learning in handling imbalanced clinical datasets and improving diagnostic performance. Finally, from the above literature, ML techniques have proven highly effective in offering improved predictive performance stroke datasets. In particular, ensemble learning has emerged as a powerful approach,using the combined strengths of multiple models to enhance stroke prediction performance.

3 Methodology

In this study, we used an XGBoost classifier to predict stroke, using standard and engineered features. To enhance model performance and interpretability, our approach integrates correlation-based filtering and SHAP-based feature selection. We designed two main feature sets and evaluated model performance across both, with and without feature selection. The proposed workflow begins with the acquisition of a standard feature set, which is then expanded into an engineered feature set informed by domain knowledge. This data passes through preprocessing followed by techniques to handle class imbalance. After splitting the dataset into training and testing subsets, correlation filtering is applied to remove highly correlated features, resulting in a refined set of filtered features. An initial XGBoost model is trained on this subset, followed by SHAP analysis to rank features by importance. The top ranked features are selected to build the final XGBoost model, which is then evaluated to assess its performance. Figure **??** represents the workflow of the methodology.

Fig. 1. Methodology workflow for stroke prediction .

3.1 Standard Feature Set

The standard feature set consists of raw variables present in the original dataset which was sourced from the Kaggle platform, specifically the "Stroke Prediction Dataset" [16]. This dataset consists of 5110 observations, each containing 12 attributes, including:

- **Demographic attributes**: Age, Gender, Residence type
- **Health indicators**: Average glucose level, Body mass index (BMI), Hypertension, Heart disease
- **Socioeconomic factors**:Marital status, Work type and Smoking status

3.2 Engineered Feature Set

The engineered feature set extends the standard features with variables constructed using domain knowledge of stroke risk factors and statistical transformations. These features aim to better represent complex health states, capture nonlinear relationships, and integrate synergistic risk factors [17]. Each engineered variable was crafted based on clinical guidelines, published stroke risk models, or intuitive logic derived from real-world patient data behavior.

Age-Derived Features Features such as `age_squared`, `age_group`, `age_decade`, and `age_risk_factor` capture the well-documented exponential rise in stroke incidence with age, particularly beyond age 55 [18, 19].

- `age_squared`: Models the nonlinear rise in stroke risk with age, acknowledging that risk accelerates rather than increases linearly.
- `age_group`: Categorical stratification into standard life stages (child to elderly), enabling the model to learn group-specific risk behaviors.
- `age_decade`: A coarse grouping used to detect decade-wise aging patterns.
- `age_risk_factor`: Binary indicator ($I_{age>55}$) reflecting a known inflection point in stroke risk prevalence.

BMI and Glucose-Related Features Stroke risk is influenced by BMI, particularly in diabetic populations [20, 21]. Variables like `bmi_category`, `bmi_risk_factor`, and `bmi_prime` reflect clinical obesity thresholds, while `bmi_glucose_interaction` captures the combined metabolic stress of hyperglycemia and obesity [22].

- `bmi_category`: Translates BMI into clinical categories (underweight to obese)
- `bmi_prime`: Normalized as $\frac{BMI}{25}$, offering a scale-invariant obesity measure.
- `bmi_glucose_interaction`: Captures the multiplicative effect of obesity and elevated blood sugar

Glucose Metabolism Indicators Features such as `glucose_level`, `diabetic_status`, and `glucose_to_age_ratio` are grounded in diabetes diagnostic guidelines [23] and quantify glycemic burden relative to age [24].

- `glucose_level`: Categorized glucose levels based on clinical diagnostic criteria (e.g., prediabetic, diabetic).
- `diabetic_status`: Binary indicator of diabetes, using a clinical threshold.
- `glucose_to_age_ratio`: A relative index of glucose burden adjusted for age, hypothesizing that elevated glucose is more damaging at younger ages.

Comorbidity and Cardiovascular Risk Composite indicators such as `hypertension_heart`, `cv_risk_index`, and `comorbidity_index` reflect the synergistic impact of multiple cardiovascular conditions on stroke risk [25, 26].

- `hypertension_heart`: Composite indicator of co-occurring hypertension and heart disease, capturing compounded cardiovascular risk.
- `metabolic_syndrome_score`: A cumulative index (0–4) reflecting the presence of obesity, hyperglycemia, hypertension, and age > 50—aligning with metabolic syndrome criteria.
- `cv_risk_index`: A weighted score combining age, glucose, BMI, and comorbidities, inspired by cardiovascular risk calculators.
- `comorbidity_index`: Count of key conditions (hypertension, heart disease, diabetes, obesity), offering a simple chronic disease burden score.

Lifestyle and Stress-Related Variables `lifestyle_risk` and `vascular_stress` account for behavioral and physiological strain from smoking, sedentary work, and vascular comorbidities—factors that indirectly modulate endothelial and systemic stroke risk [27].

- `lifestyle_risk`: Composite score reflecting unhealthy lifestyle patterns: smoking, obesity, and non-active work type.
- `vascular_stress`: A domain-inspired metric modeling systemic stress due to age and vascular comorbidities, scaled to proxy endothelial strain.

Stroke Specific Risk Profiles The composite features `age_glucose_index`, `stroke_risk_profile`, and `biological_age` were constructed to summarize multifactorial stroke risk through clinically weighted aggregation, inspired by established tools like the Framingham Risk Score [26, 28].

- `age_glucose_index`: Models interaction between aging and hyperglycemia compound effect known to increase stroke incidence.
- `stroke_risk_profile`: A rule-based composite risk score reflecting expert-driven weights on established stroke risk factors: age, hypertension, diabetes, heart disease, smoking, and obesity.
- `biological_age`: A surrogate for physiological aging, derived from age and risk-enhancing health conditions. It accounts for biological deterioration beyond chronological age.

3.3 Data Preprocessing

Effective data preprocessing is crucial to ensure high model performance. We implemented a structured preprocessing pipeline composed of several key stages that addressed missing values, encoding, normalization, and feature transformations in a consistent manner.

Data Cleaning

- **Missing Values:**
  - `bmi`: Converted to numeric and imputed using the median.
  - Categorical variables: Imputed with the most frequent value.
- **Outlier Handling:**
  - Entries with `age` < 1 were removed.

Feature Engineering Integration In addition to raw features, engineered features were incorporated into the preprocessing pipeline. These were processed in parallel with standard features to ensure consistent scaling and encoding.

Feature Categorization Features were grouped by type to apply appropriate preprocessing techniques: numerical ('age', 'avg_glucose_level', 'bmi', 'age_squared'....), binary ('hypertension', 'heart_disease', 'ever_married'...), nominal categorical ('gender', 'work_type'...), ordinal ('smoking_status'...) and engineered ('age_group', 'bmi_category'...).

Feature Scaling and Encoding Normalization (scaling features to a range) and standardization (scaling features to have a mean of zero and a standard deviation of one) are essential preprocessing steps that ensure that all features contribute equally to the model's performance. Encoding is transforming categorical variables into numerical format using techniques such as one-hot encoding or ordinal encoding. This allows the model to process these variables effectively. In our pipeline, scaling and encoding were applied according to variable type. Numerical features were standardized and scaled using standard scaler to ensure uniform contribution. Ordinal variables like smoking status were encoded ordinally to reflect known clinical risk hierarchies (never $<$ formerly $<$ currently smoked), while nominal variables such as gender and residence type were one-hot encoded to avoid imposing artificial order. Categorical and engineered variables were encoded using one-hot encoding.

3.4 Data Imbalance Handling

The original stroke dataset exhibits significant class imbalance. To address this problem we applied the SMOTEENN technique which is a hybrid approach that combines the advantages of SMOTE (Synthetic Minority Over sampling Technique) and Edited Nearest Neighbors (ENN). SMOTE generates synthetic samples for the minority class, thereby enhancing representation, while ENN removes ambiguous, noisy, and borderline instances based on nearest neighbor rules. This dual-phase refinement not only balances the dataset but also improves training quality by reducing noise and lowering misclassification risk [29].

3.5 Feature Selection Strategy

To improve generalization and reduce redundancy, we applied two complementary feature selection techniques:

- **Correlation-Based Filtering**
- **SHAP-Based Ranking**

Correlation-Based Filtering To address multicollinearity and improve the interpretability of the model, we applied a Spearman rank correlation analysis for numerical and engineered features. Spearman correlation is a non-parametric measure of monotonic relationships, making it suitable for datasets with non-linear dependencies [30]. This strategy helps to preserve model stability while minimizing redundancy [31]. In our work, we computed the full correlation matrix and identified pairs with absolute correlation coefficients ≥ 0.85 using the upper triangle of the matrix to avoid redundancy. For each correlated pair, we retained the feature with Higher domain relevance, lower missingness and greater predictive power [32].

SHAP-Based Ranking SHAP (SHapley Additive exPlanations) values were employed to enhance model interpretability and capture non-linear relationships. SHAP is a framework were used to quantify the contribution of each feature to model predictions based on cooperative game theory that assigns each feature an importance value, corresponding to its marginal contribution to the model's output [33]. SHAP is particularly suitable for explaining ML models because it satisfies key properties such as local accuracy, missingness, and consistency [34]. In our study, SHAP values were computed using the Tree SHAP method on the trained XGBoost model. We computed mean absolute SHAP values across all samples and ranked features accordingly. The top N ranked features were selected for final modeling based on their SHAP scores.

3.6 Data Splitting

In our model, the dataset was partitioned into training and testing subsets using an 80/20 split. This strategy is widely adopted in ML research, as it allocates a larger portion (80%) of the data for training, which improves the model's ability to learn from the data. The remaining 20% is reserved for testing. To ensure consistency and reproducibility across experiments, a fixed random seed was used during the split process [35]. This approach guarantees that the same samples are used in each run, enabling reliable comparisons and performance tracking.

3.7 XGBoost Classifier

An extreme gradient boosting method, known as XGBoost, is an ensemble learning technique that builds a series of decision trees in an additive manner. It

optimizes a differentiable loss function using gradient descent and includes regularization terms to control model complexity and reduce overfitting [36]. XGBoost was selected as the primary classification model in this study due to its high predictive accuracy, computational efficiency, and scalability. Moreover, its ability to handle missing values internally and to manage imbalanced datasets through parameter tuning and built-in objective functions makes it particularly suitable for medical prediction tasks such as stroke classification [4]. To further enhance model performance and address class imbalance, a specific set of hyperparameters was tuned through empirical testing.

4 Results and Discussion

4.1 Model Evaluation Metrics

Given the highly imbalanced nature of the dataset, it was essential to go beyond overall accuracy and include metrics sensitive to class distribution [37] and decision thresholds [38]. In our model, we implemented a suite of evaluation metrics like accuracy, precision, recall, F1-score and ROC-AUC. To optimize the classification threshold, we performed a systematic sweep across a range of thresholds from 0.1 to 0.9 in 50 evenly spaced increments. The optimal threshold was selected based on maximizing the F1-score which is critical for clinical decision making [39].

4.2 Results and Discussion: Standard Features

Correlation Matrix After correlation-based filtering, two pairs exhibited high correlation using a Spearman correlation threshold of $|\rho| \geq 0.85$. Removing one member of each pair reduced features from 17 to 15 without impacting downstream performance. The correlation matrix for stroke prediction is shown in figure 2.

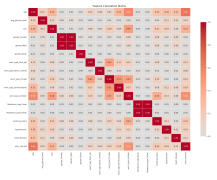

Fig. 2. Spearman correlation heatmap for the 17 standard features.

Model Performance The performance of our XGBoost classifier using only the standard feature set is summarized in Table 1.

Table 1. Model performance comparison using all versus final selected standard features.

Approach	Num Features	Accuracy	Precision	Recall	F1 Score	ROC AUC	Best Threshold
All Features	17	0.9748	0.9748	0.9780	0.9764	0.9969	0.8837
Final Selected	10	0.9743	0.9740	0.9783	0.9761	0.9966	0.8837

From the table 1 we can resume, the standard feature experiments validate that:

- **Accuracy and F1 Score:** Performance remained virtually unchanged after reducing the feature set from 17 to 10 (Accuracy: 97.48% to 97.43%; F1 Score: 97.64% to 97.61%).
- **Recall:** Improved slightly (97.79% to 97.82%), indicating better identification of stroke cases with fewer features.
- **Precision:** Remained stable (97.47% to 97.40%), showing no significant increase in false positives.
- **ROC AUC:** Maintained a high value (99.69%), confirming excellent class discrimination.
- **Threshold:** The optimal classification threshold remained at 0.8837, reflecting re-calibration due to reduced input dimensionality.

ROC Curve Figure 3 presents ROC curves for all features (AUC = 99.69%) and final selected (AUC = 99.66%). Both curves cluster near the top left corner, indicating very low false positive and false negative rates across all thresholds. The curves nearly overlap, confirming that performance is preserved after feature reduction.

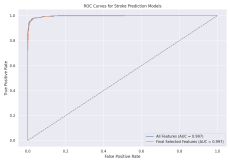

Fig. 3. ROC curves of standard features(all features vs final selected).

Confusion Matrix Figures 4 and 5 show the confusion matrices for the models using all features and final selected features, respectively.
The reduction in false negatives (from 20 to 18) is clinically significant, reducing the likelihood of missing true stroke cases. A modest increase in false positives (from 23 to 26) is acceptable given the gain in sensitivity.

SHAP Analysis The SHAP summary plot ranks features by their mean absolute SHAP value, offering insights into the contribution of each feature to the

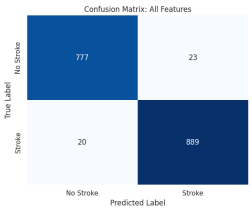

Fig. 4. Confusion Matrix All Features.

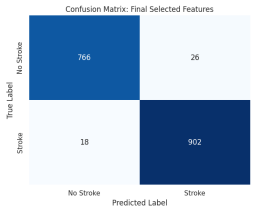

Fig. 5. Confusion Matrix Selected

predictions of the model. Figure 6 visualizes this ranking for the final selected standard features. Features are ordered top-down from most to least influential.

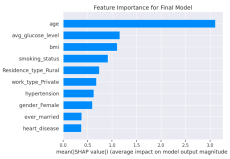

Fig. 6. SHAP summary bar plot for Final Selected Model (Standard Features).

4.3 Results and Discussion: Engineered Features

Correlation Matrix Using the 46 engineered features, we computed a Spearman correlation matrix and applied a $|\rho| \geq 0.85$ filter, which removed 18 highly collinear predictors and reduced the set to 28 features without degrading downstream performance. The correlation matrix for stroke prediction using engineered features is shown in Figure 7.

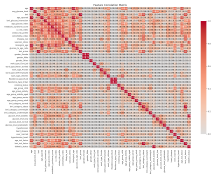

Fig. 7. Spearman correlation heatmap for all engineered features.

Model Performance The performance of our model using engineered features is presented in Table 2.

Table 2. Model performance comparison using all versus selected engineered features.

Approach	Num F	Accuracy	Precision	Recall	F1 Score	ROC AUC	Best Threshold
All	46	0.9793	0.9809	0.9809	0.9809	0.9983	0.9000
Selected	20	0.9810	0.9882	0.9767	0.9824	0.9975	0.8837

Based on the results presented in the table, the key observations are as follows:

- **Accuracy:** Improved from 97.93% to 98.10% after reducing features from 46 to 20.
- **Precision:** Increased notably (98.09% to 98.82%), reducing false positives.
- **Recall:** Slightly decreased (98.09% to 97.67%), a reasonable trade-off for higher precision.
- **F1 Score:** Rose marginally (98.09% to 98.24%), indicating improved balance.
- **ROC AUC:** Remained near perfect (99.83% to 99.75%), confirming excellent class separation.

Additionally, the optimal classification threshold shifted 0.8837, reflecting a recalibration of the model decision boundary due to the reduced input dimensionality.

ROC Curve Figure 8 shows the ROC curves for both models. Both curves lie close to the top left, indicating very low false positive and false negative rates. The minimal overlap confirms that pruning 26 features did not materially weaken the classification power.

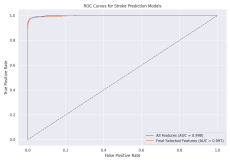

Fig. 8. ROC curves of engineered features(all features vs final selected).

Confusion Matrix Below are the confusion matrices for the models with all features (Figure **??**) and final selected features (Figure 10) models using engineered set: Key outcomes from the confusion matrices:

- **True Positives:** Both models perform equally well in identifying actual stroke cases.This shows no degradation in the model's ability to detect actual stroke instances even after reducing the feature set by more than half (from 46 to 20).

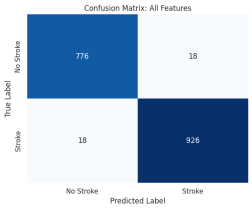

Fig. 9. Confusion Engineered All

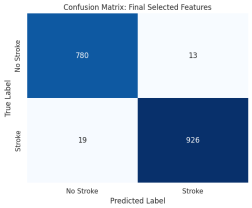

Fig. 10. Confusion Engineered Selected

- **False Positives:** Reduced from 18 to 13, a 28% drop.
- **True Negatives:** Increased from 776 to 780, improving specificity.
- **False Negatives:** Slight increase from 18 to 19, a marginal trade-off.

SHAP Analysis The SHAP summary plot in Figure 11 illustrates the global importance of each engineered feature. Composite features like *metabolic_ syndrome_ score*, *age_ glucose_ index*, and *lifestyle_ risk* dominate the model's predictive logic, underscoring the effectiveness of domain-informed feature engineering.

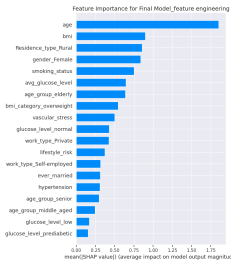

Fig. 11. SHAP summary bar plot for Final Selected Engineered Features.

4.4 Comparative Analysis

All studies in the comparative table 3 have employed XGBoost as part of their stroke prediction frameworks due to its robust performance and effectiveness in handling imbalanced data. In contrast, our proposed method combines XGBoost with a comprehensive feature selection strategy using Spearman correlation analysis and SHAP values,while also applying SMOTEENN to deliver a more explainable and performance optimized model. Using only standard features, our approach achieves an accuracy of **97.94%**, surpassing most existing works while using a reduced and interpretable feature set. More notably, with the incorporation of engineered features, our model reached a state-of-the-art accuracy of **98.10%** outperforming all reviewed methods.

Table 3. Comparison of stroke prediction methods in literature

Ref.	Authors	Methodology	Performance
[1]	Sundaram et al.	RF + LR + XGBoost Ensemble	Accuracy of 97.4%
[2]	Mochurad et al.	XGBoost + PCA + XAI	Accuracy of 95%
[5]	Hamada et al.	Stacked (DT, XGBoost, RF)	Accuracy of 97.9%
[10]	Rohini et al.	XGBoost + Adaptive Boosting	Accuracy of 95%
[12]	Gupta et al.	Comparative ML Models	Accuracy of 95%
–	**Proposed Method**	**XGBoost + Correlation + SHAP (Standard Features)**	**Accuracy of 97.94%**
–	**Proposed Method**	**XGBoost + Correlation + SHAP (Engineered Features)**	**Accuracy of 98.10%**

5 Conclusion

This study proposed an effective stroke risk prediction framework that integrates correlation analysis and SHAP-based feature selection with an XGBoost classifier, enhanced through the SMOTEENN resampling technique. The results demonstrated that with engineered features, the model achieved outstanding performance with an accuracy of 98.10%, a F1-score of 98.24%, and a ROC AUC of 99.75% using only 41% of the input dimensionality.Similarly, for the standard feature set, the model maintained strong performance with an accuracy of 97.43% and a ROC AUC of 99.66%, despite a reduction in input dimensionality by over 56%. These findings validate the effectiveness of the combined approach in balancing data quality, interpretability, and predictive performance. The future step includes integrating deep learning models, such as recurrent neural networks (RNNs) or deep neural networks (DNNs), which can uncover deeper, nonlinear relationships in patient data and improve prediction accuracy.

References

1. Setyawan, N.H., Wakhidah, N.: Analisis perbandingan metode logistic regression, random forest, gradient boosting untuk prediksi diabetes. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika) 10(1), 150–162 (2025)
2. Rice, H., de Villiers, L., Scarica, R., Bocquet, A.L., Dargan, K., Barthe, T.: Health budget implications of mechanical thrombectomy for acute ischaemic stroke in Australia. J. Med. Imaging Radiat. Oncol. 68(5), 564–569 (2024). https://doi.org/10.1111/1754-9485.13652
3. Sundaram, K., B., L., K., K., Ramamoorthy, A.K.: Enhanced brain stroke prediction: An ensemble of random forest, logistic regression and XGBoost. In: Proc. 2024 Int. Conf. Emerging Research in Computational Science (ICERCS), pp. 1–5. IEEE, Coimbatore, India (2024). https://doi.org/10.1109/ICERCS63125.2024.10895046

4. Mochurad, L., Babii, V., Boliubash, Y., et al.: Improving stroke risk prediction by integrating XGBoost, optimized principal component analysis, and explainable artificial intelligence. BMC Med. Inform. Decis. Mak. 25, 63 (2025). https://doi.org/10.1186/s12911-025-02894-z
5. Sahriar, S., et al.: Unlocking stroke prediction: Harnessing projection-based statistical feature extraction with ML algorithms. Heliyon 10(5), e27411 (2024). https://doi.org/10.1016/j.heliyon.2024.e27411
6. Ismail, M.G.A., Ibrahim, A.L.: Optimizing accuracy of stroke prediction using logistic regression. J. Technol. Informatics (JoTI) (2023). Available: https://www.researchgate.net/publication/369877529 (Accessed: 25 May 2025)
7. Mohamed Hamada, B.M., Tsentob, J.T., An, J.S.: An ensemble approach for stroke prediction. In: Proc. 2024 IEEE 17th Int. Symp. Embedded Multicore/Many-core Systems-on-Chip (MCSoC) (2024). https://doi.org/10.1109/MCSoC64144.2024.00069
8. Aref, T.: Predicting stroke risk using SMOTE and ROS machine learning techniques. In: Proc. 2024 4th Int. Conf. Electrical, Computer, Communications and Mechatronics Eng. (ICECCME), pp. 1–5. IEEE, Male, Maldives (2024). https://doi.org/10.1109/ICECCME62383.2024.10796090
9. Bathla, P., Kumar, R.: A hybrid system to predict brain stroke using a combined feature selection and classifier. Intell. Med. 4, 75–82 (2024). https://doi.org/10.1016/j.imed.2024.03.004
10. Hassan, A., Gulzar, A.S., Ullah, M.E., Ali, K.I., Ramzan, N.: Predictive modeling and identification of key risk factors for stroke using machine learning. Sci. Rep. 14, 11498 (2024). https://doi.org/10.1038/s41598-024-46692-z
11. Kumar, A., Nelson, L.: Predicting stroke risk using random forest classifier: A healthcare data analysis. In: Proc. 2025 Int. Conf. Ambient Intelligence in Health Care (ICAIHC), pp. 1–6. IEEE, Raipur Chhattisgarh, India (2025). https://doi.org/10.1109/ICAIHC64101.2025.10957053
12. Rohini, T., Praveen, P., Shaik, M.A.: Enhancing stroke risk prediction with stochastic boosting and adaptive grid-based modeling. In: Proc. 2025 Int. Conf. Electronics and Renewable Systems (ICEARS), pp. 365–370. IEEE, Tuticorin, India (2025). https://doi.org/10.1109/ICEARS64219.2025.10940256
13. Wu, D., Zhang, X., Zhu, X.: A machine learning-based model for stroke prediction. Appl. Comput. Eng. 78(1), 122–130 (2024). https://doi.org/10.54254/2755-2721/78/20240645
14. Gupta, A., Mishra, N., Jatana, N., et al.: Predicting stroke risk: An effective stroke prediction model based on neural networks. J. Neurorestoratol. 13(1), 100156 (2025). https://doi.org/10.1016/j.jnrt.2024.100156
15. Sitompul, L.R., Nababan, A.A., Manihuruk, M.L., Ponsen, W.A., Supriyandi, S.: Comparison of XGBoost, random forest and logistic regression algorithms in stroke disease classification. Sinkron: Jurnal dan Penelitian Teknik Informatika 9(2) (2025)
16. Soriano, F.: Stroke Prediction Dataset. Kaggle (2021). Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (Accessed: 23 June 2023)
17. Li, X., Liu, H., Du, X., et al.: Integrated machine learning approaches for predicting ischemic stroke and thromboembolism in atrial fibrillation. AMIA Annu. Symp. Proc. 2016, 799–807 (2017)
18. Feigin, V.L., et al.: Global burden of stroke and risk factors in 204 countries and territories, 1990–2019. Lancet Neurology 20(10), 795–820 (2021)

19. Kernan, W.N., et al.: Guidelines for the prevention of stroke in patients with stroke and transient ischemic attack. Stroke 45(7), 2160–2236 (2014)
20. Kurth, H., et al.: Body mass index and the risk of stroke in women. Archives of Internal Medicine 162(22), 2557–2562 (2002)
21. Ong, K.K., et al.: The paradox of obesity and stroke in diabetes. Diabetes Obesity and Metabolism 22(2), 198–205 (2020)
22. Simental-Mendía, A., et al.: Triglyceride-glucose index: A novel marker for insulin resistance and metabolic syndrome. Cardiovascular Diabetology 7, 10 (2008)
23. American Diabetes Association: Standards of Medical Care in Diabetes—2024. Diabetes Care 47(Suppl. 1) (2024)
24. Folsom, A.S., et al.: Fasting glucose and incident stroke: the ARIC study. Stroke 30(7), 1372–1377 (1999)
25. Benjamin, E.J., et al.: Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. Circulation 139(10), e56–e528 (2019)
26. D'Agostino, R.B., et al.: General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 117(6), 743–753 (2008)
27. Rexrode, K.M., et al.: Physical inactivity and risk of stroke in women. JAMA 289(11), 1393–1399 (2003)
28. Wang, F., et al.: Machine learning and clinical risk prediction: A systematic review and perspective. NPJ Digital Medicine 4, 40 (2021)
29. Bounab, R., Zarour, K., Guelib, B., Khlifa, N.: Enhancing Medicare fraud detection through machine learning: Addressing class imbalance with SMOTE-ENN. IEEE Access 12, 54382–54396 (2024)
30. Zar, J.H.: Spearman rank correlation. Encyclopedia of Biostatistics, 1–5 (2005)
31. Hall, K., Elsayed, E., Aggarwal, G.: Feature selection for machine learning: Comparing correlation-based filters and mutual information. In: Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW), pp. 207–214 (2018)
32. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)
33. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 4765–4774 (2017)
34. Lundberg, S.M., et al.: From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2(1), 56–67 (2020)
35. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning, 2nd edn. Springer (2021)
36. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 785–794. ACM, San Francisco, CA, USA (2016)
37. Weiss, G.M., Provost, F.: The effect of class distribution on classifier learning: an empirical study. Technical Report, Rutgers University (2001)
38. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21(9), 1263–1284 (2009)
39. Saito, J., Rehmsmeier, T.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE 10(3), e0118432 (2015)