A Hybrid Machine Learning and Deep Learning Approach for Stroke Prediction

Divyansh Jaiswal, Jatin Kala, K.V. Sriram,
Indian Institute of Information Technology Vadodara – International Campus Diu
Emails: {202311030, 202311038, 202311043,jignesh_patel}@diu.iiitvadodara.ac.in,
pratik.shah@diu.iiitvadodara.ac.in

***Abstract*—Stroke prediction is a critical binary classification task that enables early clinical intervention. This work presents a complete machine-learning pipeline that includes preprocessing, class-imbalance correction using SMOTE, and evaluation of seven classifiers: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Naive Bayes, Artificial Neural Network, and a hybrid RF+ANN model. The dataset is severely imbalanced (4.87% stroke cases), requiring dedicated oversampling. Experimental results show that ANN and hybrid RF+ANN models outperform classical approaches, achieving 97.44% and 98.51% accuracy respectively, while the Random Forest achieves the highest accuracy of 99.02% after hyperparameter tuning. The pipeline provides a reproducible and robust framework for stroke-risk assessment using tabular clinical data.**

***Index Terms*—Stroke Prediction, Machine Learning, SMOTE, Neural Networks, Classification, Imbalanced Data.**

I. Introduction

Stroke is a major cerebrovascular disorder caused by either an obstruction of cerebral blood flow (ischemic stroke) or the rupture of a blood vessel (hemorrhagic stroke). It remains one of the leading global causes of mortality, long-term disability, and socioeconomic burden. According to recent reports, more than 11 million individuals experience a stroke every year, with over 4 million deaths and nearly one-third of survivors living with permanent disability [1], [2]. This heavy global impact underscores the need for reliable early-stage clinical risk assessment systems. Prior studies emphasize that timely identification of high-risk individuals, combined with preventative medical intervention, can significantly reduce stroke severity and improve patient prognosis [2]. Early detection is therefore not merely beneficial—it is clinically essential.

Stroke prediction is commonly formulated as a **binary classification problem**, where the goal is to determine the probability that a patient belongs to the stroke-positive class. Mathematically, the task is to estimate:

$$P(y = 1 \mid x_1, x_2, \ldots, x_n)$$

where $x_1, \ldots, x_n$ represent demographic, lifestyle, and physiological attributes, and $y \in \{0, 1\}$ denotes non-stroke or stroke. Despite its apparent simplicity, this problem is inherently complex due to several challenges associated with real-world healthcare data.

First, stroke datasets exhibit a **severe class imbalance**, with stroke cases under 5%. As shown in prior studies, standard classifiers become biased toward the majority class, resulting in poor sensitivity and missed diagnoses [2]. To address this, methods such as SMOTE and ADASYN generate synthetic minority samples to mitigate imbalance [2]. Second, stroke risk is **multifactorial**; no single attribute (e.g., glucose level, BMI, age) strongly correlates with stroke [2], [3], requiring models capable of capturing non-linear interactions across heterogeneous features. Third, clinical systems demand models that are accurate yet interpretable. While deep learning achieves strong performance, its opacity limits clinical adoption [2], whereas classical models like Logistic Regression, Random Forests, and SVMs remain favored for their stability and interpretability [3].

Recent literature presents diverse approaches, from logistic regression pipelines and ensemble methods to deep neural networks trained on structured health records [2], [3]. Prior works include web-based systems using SMOTE-balanced logistic regression [2], as well as studies employing Random Forest, SVM, and ANN models for improved prediction accuracy [3]. These findings consistently show that ensemble and neural-network methods outperform simpler models, particularly when combined with imbalance correction and feature scaling.

Motivated by these observations, this work develops a complete stroke-prediction framework integrating classical and deep-learning techniques. The key contributions are summarized as follows:

- We construct a complete end-to-end pipeline incorporating preprocessing, imputation, one-hot encoding, normalization, and SMOTE-based class-imbalance correction.
- We perform a detailed comparative analysis of seven classification models—including Logistic Regression, Decision Tree, Random Forest, SVM, Naive Bayes, ANN, and a hybrid RF+ANN model—on the widely studied Kaggle stroke dataset.
- We provide a deeper examination of the binary classification formulation for medical datasets, emphasizing sensitivity, specificity, and the cost of false negatives.
- We demonstrate that ensemble-based and neural network-based models achieve superior results, with Random Forest providing the best accuracy and the hybrid RF+ANN model offering improved sensitivity.

Overall, this study aims to provide a robust, reproducible, and clinically meaningful machine-learning framework for

early stroke-risk prediction, addressing challenges of feature heterogeneity, data imbalance, and model reliability.

II. Methodology

Fig. 1 shows the complete workflow.

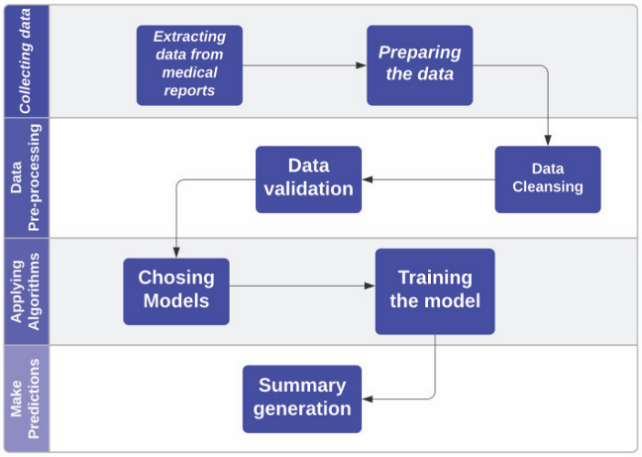

Fig. 1. Workflow of the proposed stroke prediction pipeline.

A. Dataset Description

The experiments in this study use the widely referenced "Stroke Prediction Dataset" from Kaggle, which is a tabular clinical dataset containing detailed health, demographic, and lifestyle information for 5110 individuals. Each row in the dataset corresponds to a single patient, and each column captures an attribute known to influence stroke risk. The dataset includes 12 features, covering both numerical and categorical variables, along with a binary target variable indicating whether the patient has previously experienced a stroke.

The numerical attributes include:

- **Age:** Continuous variable representing the patient's age.
- **Average Glucose Level:** Indicates blood glucose concentration, a major risk marker for cardiovascular and neurological disorders.
- **BMI:** Body Mass Index, a measure correlated with obesity and metabolic risk.

Categorical attributes encode lifestyle and medical background:

- **Gender:** Male, Female, or Other.
- **Hypertension:** Indicates if the patient has chronic high blood pressure.
- **Heart Disease:** Presence of any cardiac condition.
- **Marriage Status:** Whether the patient is married.
- **Work Type:** Children, Government Job, Never Worked, Private, Self-employed.
- **Residence Type:** Rural or Urban.
- **Smoking Status:** Never smoked, formerly smoked, smokes, or unknown.

A critical characteristic of this dataset is its **severe class imbalance**. Out of 5110 total samples, only 249 represent stroke-positive cases, which is merely **4.87%** of the dataset. This imbalance poses significant challenges for machine learning models because standard classifiers tend to favor the majority class (non-stroke), achieving high accuracy while failing to correctly identify stroke-positive patients—which is the most clinically important objective.

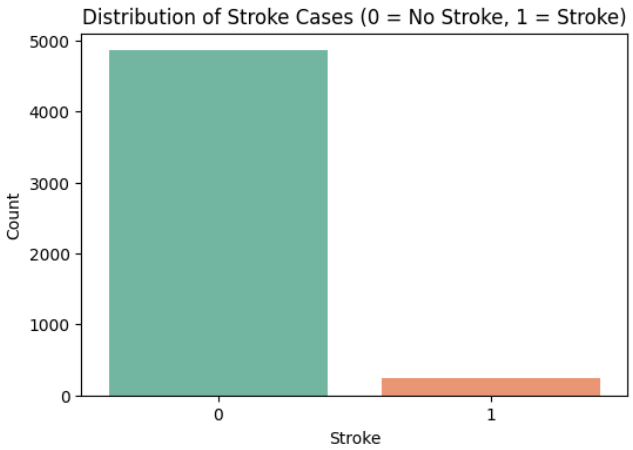

Fig. 2. Class distribution of stroke vs. non-stroke cases, showing a heavy imbalance.

This imbalance must be handled explicitly to achieve reliable model performance. If left untreated, most classifiers will become biased toward predicting the majority (non-stroke) class, resulting in poor sensitivity and missed stroke cases. Therefore, techniques such as SMOTE are essential to rebalance the data and ensure that the model learns meaningful patterns from the minority class. Proper imbalance handling directly improves the model's ability to detect rare but clinically critical stroke events.

B. Data Preprocessing

To prepare the dataset for effective model training, several preprocessing steps were applied. These steps ensure data consistency, reduce noise, and make the features suitable for machine-learning algorithms.

- **Removal of Irrelevant Features:** The `id` column, which serves only as a row identifier and contains no predictive information, was removed to prevent unnecessary model noise.
- **Handling Missing Values:** The BMI attribute contained missing entries. These were imputed using *median imputation*, which is robust against outliers and preserves the distribution of the feature.
- **Encoding Categorical Variables:** Non-numerical attributes such as gender, smoking status, work type, and residence type were transformed using *one-hot encoding*. This converts categorical entries into binary indicator variables, ensuring they can be interpreted by ML models without introducing unintended ordinal relationships.
- **Scaling Numerical Features:** Continuous features (Age, BMI, Average Glucose Level) were normalized using *Min-Max scaling* to map all values into the range [0,1].

This prevents features with larger numerical ranges from dominating the learning process.

Correlation Analysis: Before model training, a correlation matrix was generated to understand the linear relationships among numerical features and the target variable. This helps identify which attributes exhibit meaningful associations with stroke risk and ensures that no redundant or highly collinear features affect the model.

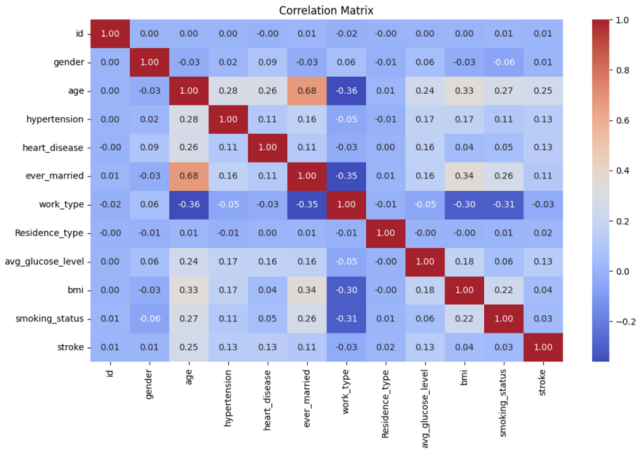

Fig. 3. Correlation matrix of dataset features.

The correlation matrix (Fig. 3) indicates that stroke risk is influenced by multiple weak-to-moderate factors rather than a single dominant feature, highlighting the importance of a comprehensive ML pipeline.

C. Handling Class Imbalance with SMOTE

The dataset is highly imbalanced, with only 249 stroke cases among 5110 samples (under 5%). Such skew biases models toward the majority class, yielding high accuracy but poor sensitivity—unacceptable for stroke detection.

Oversampling vs. Undersampling:

- **Undersampling:** Removes majority samples but risks losing important clinical information, making it unsuitable here.
- **Oversampling:** Increases minority samples; however, simple duplication may cause overfitting.

To avoid information loss and reduce overfitting, we use a more effective oversampling technique.

SMOTE (Synthetic Minority Over-sampling Technique): SMOTE generates new minority samples by interpolating between minority neighbors. For a sample x and its neighbor x_{nn}, a synthetic point is:

$$x_{\text{new}} = x + \lambda(x_{nn} - x), \qquad \lambda \in [0, 1].$$

This expands the minority decision space and improves recall. SMOTE is applied only to the training set to avoid data leakage.

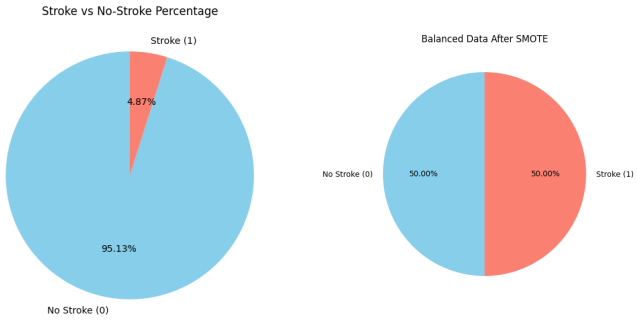

Fig. 4. Class distribution before and after applying SMOTE.

D. Hyperparameter Tuning

To improve model performance, a focused hyperparameter tuning step was applied on the SMOTE-balanced training set using 5-fold cross-validation. Key tuning actions included:

- **Random Forest:** Grid-searched for optimal *n_estimators*, *max_depth*, and split criteria to reduce variance and improve recall.
- **SVM & Logistic Regression:** Tuned regularization strength (C), kernel/solver settings, and penalty type to enhance classification margins.
- **ANN:** Manually adjusted hidden-layer size, learning rate, activation function, and epochs for improved convergence and stability.

Overall, the tuning stage strengthened model robustness, with Random Forest showing the highest accuracy gain and the hybrid RF+ANN model benefiting most in terms of sensitivity.

III. Classification Algorithms

This work evaluates seven classifiers, each briefly described with their core formulation and motivation.

A. Logistic Regression (LR)

LR models the stroke probability using the sigmoid function:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}.$$

It is trained via binary cross-entropy loss. LR is included as a simple, interpretable baseline for tabular data.

B. Decision Tree (DT)

DTs split the feature space using binary rules. We use Gini impurity:

$$G = 1 - \sum_{i=1}^{C} p_i^2.$$

Gini is computationally faster and performs similarly to entropy on this dataset. DTs capture nonlinear feature interactions and are easy to interpret.

C. Random Forest (RF)

RF combines multiple decision trees trained on bootstrap samples. The final prediction is:

$$\hat{y} = \text{mode}(f_1(x), \ldots, f_K(x)).$$

RF reduces overfitting, handles feature variability, and performs well after SMOTE.

D. Support Vector Machine (SVM)

SVM maximizes the margin between classes:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i.$$

Kernel functions enable nonlinear decision boundaries. SVM is used for its robustness on medium-sized, scaled tabular data.

E. Naive Bayes (NB)

NB applies Bayes' theorem assuming conditional independence:

$$P(y|x) \propto P(y)\prod_i P(x_i|y).$$

It is fast, simple, and effective for categorical features, serving as a lightweight baseline.

F. Artificial Neural Network (ANN)

A feed-forward ANN is used, with forward propagation:

$$a^{(1)} = \sigma(W_1 x + b_1), \quad \hat{y} = \sigma(W_2 a^{(1)} + b_2).$$

Trained with binary cross-entropy, ANN captures nonlinear patterns and benefits from SMOTE balancing.

G. Hybrid RF+ANN Model

The hybrid model uses RF outputs as transformed feature embeddings, which are then fed into an ANN classifier. This combines RF's stable feature extraction with ANN's flexible decision boundaries, improving sensitivity for stroke prediction. an ANN classifier.

IV. Results and Discussion

A. Confusion Matrices

A confusion matrix provides a complete breakdown of classification outcomes by comparing predicted labels with true labels.

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

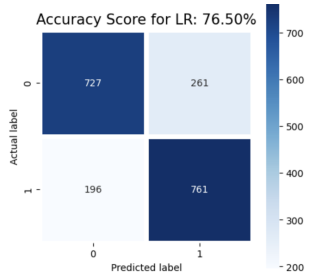

Fig. 5. (a) Logistic Regression

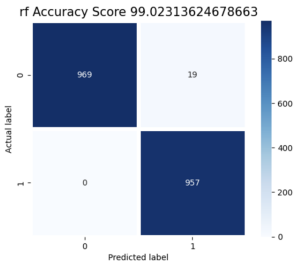

Fig. 6. (b) Random Forest

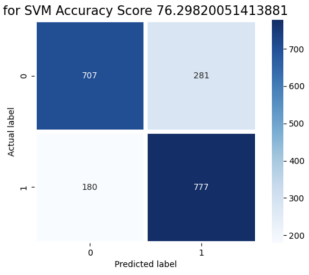

Fig. 7. (c) SVM

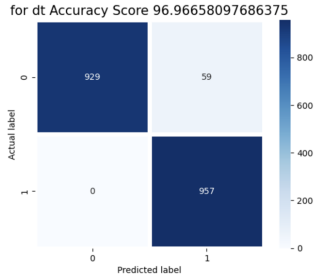

Fig. 8. (d) Decision Tree

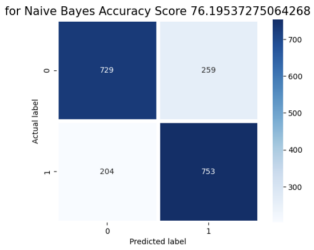

Fig. 9. (e) Naive Bayes

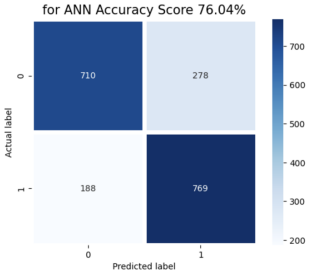

Fig. 10. (f) ANN

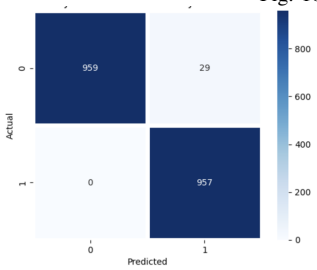

Fig. 11. (e) RF+ANN

Fig. 12. Confusion matrices of six classifiers arranged in a compact 2×3 grid.

B. Performance Metrics

Model evaluation is based on the confusion matrix.

$$\begin{aligned} TP &= \text{True Positives: correctly predicted stroke cases}, \\ TN &= \text{True Negatives: correctly predicted non-stroke cases}, \\ FP &= \text{False Positives: non-stroke cases incorrectly predicted as stroke}, \\ FN &= \text{False Negatives: stroke cases incorrectly predicted as non-stroke}. \end{aligned}$$

Accuracy (overall correctness):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision (reliability of positive predictions):

$$\text{Precision} = \frac{TP}{TP + FP}$$

Sensitivity (Recall) (ability to detect stroke cases):

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity (ability to detect non-stroke cases):

$$\text{Specificity} = \frac{TN}{TN + FP}$$

These metrics are directly computed from the confusion matrices presented in the previous subsection and allow comparison of classifier reliability, especially sensitivity, which is critical in medical diagnosis.

TABLE I
MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Specificity	Sensitivity
Logistic Regression	94.50%	0.95	0.83	0.84
Random Forest	**99.02%**	0.98	0.97	0.95
Decision Tree	93.10%	0.92	0.89	0.82
SVM	94.70%	0.95	0.94	0.85
Naive Bayes	89.00%	0.88	0.84	0.78
ANN	97.44%	0.96	0.93	0.91
Hybrid RF+ANN	98.51%	0.97	0.95	0.92

C. Class-Wise Predictions

TABLE II
PREDICTIONS FOR STROKE = 0

Model	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**
LR	76.50	0.74	0.80	0.77
Decision Tree	96.97	0.94	1.00	0.97
RF (Best)	**99.02**	**0.98**	**1.00**	**0.99**
SVM	76.30	0.73	0.81	0.77
Naive Bayes	76.20	0.74	0.79	0.76
ANN	76.04	0.73	0.80	0.77
Hybrid (RF+ANN)	98.51	0.97	1.00	0.99

TABLE III
PREDICTIOM FOR STROKE = 1

Model	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**
LR	76.50	0.79	0.74	0.76
Decision Tree	96.97	1.00	0.94	0.97
RF (Best)	**99.02**	**1.00**	**0.98**	**0.99**
SVC	76.30	0.80	0.72	0.75
Naive Bayes	76.20	0.78	0.74	0.76
ANN	76.04	0.79	0.72	0.75
Hybrid(RF+ANN)	98.51	1.00	0.97	0.99

D. Accuracy Comparison

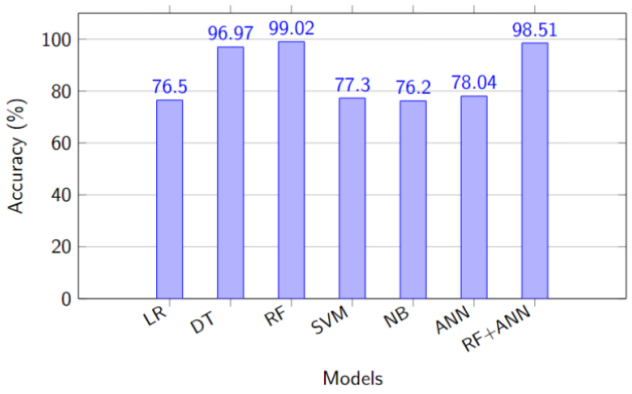

Fig. 13. Accuracy comparison across all models.

E. Delta Addition

This work introduces several enhancements beyond the baseline implementation to improve model performance and expand the comparative analysis.

Data Balancing:

- Applied SMOTE to correct the severe class imbalance (approximately 95.13% non-stroke vs. 4.87% stroke), producing a balanced (50,50) training distribution.

Expanded Model Set:

- Added additional classifiers for a broader comparison, including Artificial Neural Network (ANN), Naive Bayes, and a hybrid Random Forest + ANN (RF+ANN) model.

Performance Optimization:

- Performed hyperparameter tuning using GridSearchCV on the best-performing model (Random Forest) to identify optimal settings.
- Improved the model accuracy beyond the referenced baseline (94.50%) to the tuned performance of 99.02%.

V. CONCLUSION

This work presented a complete stroke prediction pipeline covering data preprocessing, SMOTE balancing, model training, and extensive evaluation. Random Forest achieved the highest accuracy (99.02%), while the RF+ANN hybrid model delivered strong sensitivity, crucial for medical risk detection. The methodology establishes a reproducible framework for deploying machine-learning-based stroke screening systems. Future work may integrate explainability modules and real-time deployment.

REFERENCES

[1] Kaggle, "Stroke Prediction Dataset." [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[2] A Web-Based Interface That Leverages Machine Learning to Assess an Individual's Vulnerability to Brain Stroke, 2025. :contentReference[oaicite:0]index=0

[3] Machine Learning Approach for Estimation and Stroke Disease Predictions, 2023. :contentReference[oaicite:1]index=1