Stroke Prediction
A Data-Driven Framework for Clinical and Public Health Applications

Nikhil Chikmagalur Prasanna Kumar

Abstract

This paper presents a comprehensive study on stroke prediction using machine learning and deep learning approaches, integrating rigorous analytical techniques with practical deployment strategies. We develop predictive models, derive a risk scoring system, and outline a framework for integrating these tools into clinical and public health workflows.

1 Introduction

1.1 The Public Health Impact of Stroke

Stroke is a leading cause of mortality and long-term disability worldwide, placing a significant burden on healthcare systems and diminishing survivors' quality of life. Despite advances in acute treatment, early risk identification remains the most effective strategy for prevention. With approximately 13.7 million new strokes occurring annually—and projections indicating an increase due to aging populations—the economic impact is substantial. This impact includes direct healthcare costs, lost productivity, and extensive informal caregiving expenses.

1.2 The Role of Predictive Modeling in Stroke Prevention

Conventional clinical risk assessments typically use fixed scores and a limited set of risk factors, often missing complex interactions and atypical risk patterns. In contrast, predictive modeling using machine learning and deep learning can:

- Identify complex, non-linear relationships among risk factors,
- Capture interaction effects that traditional methods may overlook,
- Recognize unusual risk profiles,
- Provide personalized risk assessments,
- Enable more efficient resource allocation by focusing on high-risk populations.

This shift from reactive treatment to proactive prevention allows for targeted interventions—such as lifestyle modifications, optimized medication, and closer monitoring—that can potentially avert strokes.

1.3 Research Objectives

This study aims to develop a comprehensive stroke prediction framework with the following objectives:

Objective	Description
1. Risk Factor Identification	Quantify key predictors of stroke risk through rigorous statistical analysis and feature importance assessment.
2. Predictive Model Development	Design and compare traditional machine learning and deep learning approaches to optimize accuracy and clinical relevance.
3. Model Interpretability	Use SHAP analysis to elucidate model predictions and enhance transparency for healthcare providers.
4. Clinical Translation	Develop an interpretable risk scoring system that converts complex model outputs into a practical clinical tool.
5. Implementation Framework	Create a robust deployment infrastructure, including an API for seamless integration into healthcare workflows.
6. Public Health Impact	Generate actionable insights for targeted screening and prevention strategies to inform public health policies.

Table 1: Research Objectives

By integrating advanced analytical techniques with clinical practice, this research seeks to provide a scalable and interpretable framework for stroke risk assessment, ultimately aiming to reduce the burden of this debilitating condition.

2 Data Analysis & Exploratory Data Analysis (EDA)

2.1 Dataset Description and Characteristics

The analysis employs a healthcare dataset of 5,110 patient records, capturing demographic, lifestyle, and clinical variables. It includes 11 predictors and a binary target variable indicating stroke occurrence. The dataset exhibits a marked class imbalance with 249 stroke cases (4.9%) versus 4,861 non-stroke cases (95.1%).

2.1.1 Dataset Overview:

Attribute	Details
Total Records	5,110
Target Variable	Stroke occurrence (0 = No, 1 = Yes)
Stroke Cases	249 (4.9%)
Non-Stroke Cases	4,861 (95.1%)

Table 2: Dataset Overview

Variables are classified as follows:

2.1.2 Numerical Variables:

- **Age (years):** Range: 0.08–82.00; Mean: 43.23; Median: 45.00
- **Average Glucose Level (mg/dL):** Range: 55.12–271.74; Mean: 106.15; Median: 91.88
- **BMI:** Range: 10.30–97.60; Mean: 28.89; Median: 28.10

2.1.3 Categorical Variables:

- **Gender:** Female (58.6%), Male (41.4%), Other (<0.1%)
- **Hypertension:** 0 (90.3%), 1 (9.7%)
- **Heart Disease:** 0 (94.6%), 1 (5.4%)
- **Ever Married:** Yes (65.6%), No (34.4%)
- **Work Type:** Private (57.2%), Self-employed (16.0%), Government job (12.9%), Children (13.4%), Never worked (0.4%)
- **Residence Type:** Urban (50.8%), Rural (49.2%)
- **Smoking Status:** Never smoked (37.0%), Unknown (30.2%), Formerly smoked (17.3%), Currently smokes (15.4%)

2.2 Data Quality Assessment and Distribution Analysis

2.2.1 Missing Value Analysis:

The dataset contained missing BMI values in 201 records (3.93%). Analysis showed these records had a higher average age (52.05 vs. 42.87 years) and a substantially higher stroke rate (19.90% vs. 4.26%), suggesting that the missingness itself is informative.

2.2.2 Numerical Variable Distributions:

- **Age:** Slight left skew (-0.14) with a uniform distribution across adult age groups.
- **Average Glucose Level:** Strong positive skew (1.57); majority below 100 mg/dL with a long tail.
- **BMI:** Moderate positive skew (1.06) with most values concentrated in the overweight range (25–30).

2.2.3 Categorical Variable Analysis:

- **Medical Conditions:** Hypertension (9.7%) and heart disease (5.4%) align with population statistics.
- **Work Type:** Dominated by private sector (57.2%); "Children" represents a distinct group.
- **Smoking Status:** "Unknown" category (30.2%) poses analysis challenges.

2.3 Key Patterns and Relationships with Stroke

2.3.1 Age and Stroke Risk:

Age is a powerful predictor with a marked non-linear relationship. Stroke risk increases dramatically with age as detailed in Table 3:

Age Group	Stroke Risk	Risk Ratio
<40	0.37%	1.00 (reference)
40–60	3.84%	10.41
>60	13.15%	35.68

Table 3: Age and Stroke Risk

Condition	Risk with Condition	Risk without Condition	Risk Ratio / OR
Hypertension	13.25%	3.97%	3.34 (risk ratio), 3.70 (OR)
Heart Disease	17.03%	4.18%	4.08 (risk ratio), 4.71 (OR)

Table 4: Medical Conditions and Stroke Risk

2.3.2 Medical Conditions:

Hypertension and heart disease are strongly associated with stroke as detailed in Table 4:

2.3.3 Glucose Level and Stroke Risk:

A clear threshold effect is observed around diabetic levels as detailed in Table 5:

Glucose Category	Stroke Risk	Risk Ratio
Normal (<100 mg/dL)	3.58%	1.00 (reference)
Pre-diabetic (100–125 mg/dL)	3.71%	1.04
Diabetic (>=126 mg/dL)	10.19%	2.85

Table 5: Glucose Levels and Stroke Risk

2.3.4 BMI and Stroke Risk:

The relationship between BMI and stroke is non-linear as detailed in Table 6:

2.4 Insights from EDA:

- **Non-linear Relationships:** Key variables like age and glucose levels exhibit non-linear effects, necessitating careful transformation or use of non-linear modeling techniques.
- **Threshold Effects:** Particularly with glucose levels, a distinct threshold at 126 mg/dL suggests value in categorical transformation.
- **Interaction Potential:** The stark risk differentials, especially in age groups, indicate likely interactions between age and other risk factors.
- **Class Imbalance:** The severe imbalance (4.9% positive cases) requires specialized modeling approaches.
- **Missing Data Signal:** The non-random missingness in BMI suggests that missing values carry predictive information.
- **Categorical Complexity:** Variables such as work type and smoking status, notably with ambiguous categories like "Children" and "Unknown," need tailored handling.

BMI Category	Stroke Risk
Underweight (<18.5)	0.30%
Normal (18.5–24.9)	2.82%
Overweight (25–29.9)	5.32%
Obese (>=30)	5.10%

Table 6: BMI Categories and Stroke Risk

These insights form the basis for subsequent statistical analyses, feature engineering, and model development strategies aimed at capturing the complex risk patterns inherent in stroke prediction.

3 Statistical Analysis & Unsupervised Learning

3.1 Statistical Methods for Identifying Significant Risk Factors

To rigorously assess the relationships between potential risk factors and stroke occurrence, we applied several statistical techniques appropriate for both numerical and categorical variables.

3.1.1 Correlation Analysis:

For numerical variables, Pearson correlation coefficients with the stroke outcome were computed as detailed in Table 7:

Variable	Correlation Coefficient	Significance
Age	0.245	$p < 0.001$
Average Glucose Level	0.132	$p < 0.001$
BMI	0.042	$p < 0.05$

Table 7: Correlation Analysis for Numerical Variables

Age exhibits the strongest correlation, suggesting a moderate linear relationship. However, its nonlinear impact on stroke risk—evident from earlier EDA—indicates that linear measures may understate its true effect.

3.1.2 Chi-Square Tests for Categorical Variables:

Associations between categorical predictors and stroke were evaluated using chi-square tests along with Cramer's V as detailed in Table 8:

Notably, heart disease and hypertension display the strongest associations with stroke, while residence type and gender are not statistically significant.

3.1.3 Mann-Whitney U Tests:

Due to non-normal distributions in key numerical variables, Mann-Whitney U tests were conducted to compare stroke and non-stroke groups, detailed in Table 9:

Variable	Cramer's V	Significance
Heart Disease	0.135	$p < 0.001$
Hypertension	0.128	$p < 0.001$
Ever Married	0.108	$p < 0.001$
Work Type	0.098	$p < 0.001$
Smoking Status	0.076	$p < 0.001$
Residence Type	0.015	$p > 0.05$
Gender	0.010	$p > 0.05$

Table 8: Chi-Square Tests for Categorical Variables

Variable	Z-Score	Significance
Age	-17.83	$p < 0.001$
Average Glucose Level	-5.90	$p < 0.001$
BMI	-3.88	$p < 0.001$

Table 9: Mann-Whitney U Tests

The exceptionally high Z-score for age confirms its role as the strongest differentiator between groups.

3.1.4 Interaction Effects:

Stratified analyses explored how risk factors interact across age groups. For instance:

Age and Hypertension:

Age Group	Risk with Hypertension	Risk without Hypertension	Risk Ratio
<40	2.70%	0.33%	8.24
40–60	7.27%	3.43%	2.12
>60	17.91%	11.85%	1.51

Table 10: Age and Hypertension Interaction

Age and Heart Disease:

These analyses reveal that while absolute risks are higher in older patients, the relative risk increase due to conditions like hypertension is most pronounced in younger individuals.

3.2 Unsupervised Learning for Pattern Discovery

To uncover latent structures and complex patterns in the data, several unsupervised techniques were employed.

3.2.1 Dimensionality Reduction:

- **Principal Component Analysis (PCA):** The first principal component explained 28.6% of variance (strongly associated with age, heart disease, hypertension, and average glucose level), while the second explained 16.4% (capturing metabolic factors like glucose, BMI). The top four components explained approximately 72% of the total variance, highlighting an "age-comorbidity spectrum" and a secondary axis of metabolic health.

Age Group	Risk with Heart Disease	Risk without Heart Disease	Risk Ratio
40–60	6.78%	3.72%	1.82
>60	20.09%	11.88%	1.69

Table 11: Age and Heart Disease Interaction

- **Non-linear Techniques (t-SNE & UMAP):** Revealed a clear separation of age groups, with stroke cases predominantly clustering among older patients. Also revealed distinct clustering patterns and boundary regions where stroke and non-stroke cases intermingle, underscoring complex, non-linear relationships.

3.2.2 Clustering Analysis:

- **K-Means Clustering (k=4):** Optimal clustering (k=4, as determined by silhouette analysis) yielded the following patient subgroups, detailed in Table 12:

Cluster	Size (%)	Stroke Rate	Key Characteristics
Cluster 0	33.2	1.47%	Young (avg. age 29.8), very low hypertension (2.1%) and heart disease (0.8%), normal glucose (88.6 mg/dL)
Cluster 1	28.6	4.32%	Middle-aged (avg. age 44.3), moderate hypertension (5.8%), low heart disease (2.9%), slightly elevated glucose (95.2 mg/dL)
Cluster 2	24.1	6.18%	Older (avg. age 54.7), moderate hypertension (11.2%), moderate heart disease (7.6%), elevated glucose (108.9 mg/dL)
Cluster 3	14.1	15.82%	Elderly (avg. age 68.5), high hypertension (29.3%), high heart disease (18.8%), high glucose (142.5 mg/dL)

Table 12: K-Means Clustering Results

- **Hierarchical Clustering:** Achieved 83.7% agreement with K-Means, showing similar risk stratification.
- **DBSCAN Clustering:** Identified 3 natural clusters and 257 noise points (5.0%), Notably, the high-risk cluster had a 14.9% stroke rate, while noise points exhibited an even higher stroke rate (18.3%), suggesting that atypical patient profiles may be critical risk indicators.

3.2.3 Anomaly Detection:

Two methods were applied to flag atypical cases:

- **Isolation Forest:** Detected 5% of records (255 patients) as anomalies, with a stroke rate of 16.47% (versus 4.13% in the rest). An average age of 63.8 years (compared to 42.0 years). Elevated hypertension (26.3% vs. 8.8%) and heart disease (14.5% vs. 4.9%).
- **Local Outlier Factor (LOF):** Also identified 5% anomalies, with a stroke rate of 13.33% (versus 4.30% in normal cases), though with only 61.2% overlap with Isolation Forest results.

3.2.4 Association Rule Mining:

Using the Apriori algorithm, key association rules among categorical variables were extracted. Top rules: These rules detailed in Table 13, reinforce the frequent co-occurrence of advanced age, cardiovascular conditions, and metabolic abnormalities in stroke patients.

Rule	Confidence	Lift
{Age > 60, Heart Disease} → {Hypertension}	0.87	3.21
{Heart Disease, Hypertension} → {Age > 60}	0.79	2.86
{Age > 60, Glucose Diabetic} → {Hypertension}	0.72	2.67
{Hypertension, Formerly Smoked} → {Age > 60}	0.81	2.93
{Ever Married, Heart Disease} → {Age > 60}	0.85	3.08

Table 13: Association Rules for Stroke Patients

3.2.5 Insights from Statistical Analysis & Unsupervised Learning:

- **Risk Factor Hierarchy:** Both statistical tests and unsupervised techniques consistently highlight age as the dominant predictor, followed by heart disease, hypertension, and elevated glucose levels.
- **Non-linear Relationships:** Evidence from both PCA and non-linear dimensionality reduction confirms that the relationships between risk factors and stroke are non-linear, necessitating advanced modeling strategies.
- **Patient Subgroup Stratification:** Clustering analyses identified distinct subgroups with up to a ten-fold difference in stroke risk, validating the utility of data-driven stratification.
- **Anomalous Profiles:** Anomaly detection methods revealed that patients deviating from typical patterns bear significantly higher stroke risks, suggesting that outlier information can be an important predictive signal.
- **Interaction Effects:** The interaction between age and cardiovascular conditions, particularly in younger cohorts, indicates that early-onset hypertension and heart disease may signal an elevated relative risk.
- **Latent Dimensions:** PCA uncovered latent factors corresponding to an "age-comorbidity spectrum" and metabolic health, which may further enhance predictive modeling.

4 Data Preprocessing & Feature Engineering

A comprehensive preprocessing pipeline was developed to address data quality issues and to extract informative features, based on insights from our exploratory and statistical analyses.

4.1 Data Cleaning and Transformation Strategy

4.1.1 Missing Value Handling:

The dataset contained missing BMI values in 201 records (3.93%). Analysis showed that these records had a higher average age (52.05 vs. 42.87 years) and a substantially higher stroke rate (19.90% vs. 4.26%). To preserve this predictive signal, we adopted a dual strategy:

- **Missing Indicator:** A binary feature (missing_bmi_indicator) was created to flag missing BMI values.
- **KNN-based Imputation:** A K-Nearest Neighbors imputation (with k=5) was used to estimate BMI values based on related features (e.g., age, gender).

4.1.2 Feature Scaling and Transformations:

To optimize model performance and address non-normal distributions:

- **Standardization:** All numerical features were standardized using z-score normalization:

$$z = \frac{(x - \mu)}{\sigma}$$

- **Logarithmic Transformation:** The average glucose level, originally exhibiting a right skew of 1.57, was log-transformed, reducing skewness to 0.34 and better capturing physiologically relevant percent changes.
- **Categorical Encoding:** One-hot encoding was applied to gender, ever_married, residence_type, and smoking_status (with drop-first to avoid multicollinearity). Special handling was applied for `work_type` to account for the distinct "children" category.

4.2 Feature Engineering: Based on Domain Knowledge

Drawing from clinical insights and statistical findings, several new features were engineered to capture complex relationships:

4.2.1 Age-Related Features:

Given age's dominant, non-linear influence on stroke risk, we generated:

- **Age Categories:** Binary indicators for key thresholds:
  - `age_under_40`: Age <40 (stroke risk: 0.37%)
  - `age_40_to_60`: Age 40–60 (stroke risk: 3.84%)
  - `age_over_60`: Age >60 (stroke risk: 13.15%)
- **Age Squared:** A quadratic term computed as:

$$\texttt{age\_squared} = (\texttt{standardized\_age})^2$$

  This transformation better models the exponential risk increase, evidenced by a 35.68× risk ratio between patients over 60 and those under 40.

Medical Category Features: Clinical thresholds were embedded into the feature space for Glucose and BMI:

- **Glucose Categories:**

Feature	Threshold	Stroke Risk
`glucose_normal`	<100 mg/dL	3.58%
`glucose_prediabetic`	100–125 mg/dL	3.71%
`glucose_diabetic`	>=126 mg/dL	10.19%

Table 14: Glucose Categories

- **BMI Categories:**

Feature	Range	Stroke Risk
`bmi_underweight`	<18.5	0.30%
`bmi_normal`	18.5–24.9	2.82%
`bmi_overweight`	25–29.9	5.32%
`bmi_obese`	>=30	5.10%

Table 15: BMI Categories

Interaction Features: To capture the varying influence of risk factors across age groups, interaction terms were engineered:

- **Age-Condition Interactions:**
  - `age_hypertension`: Standardized age multiplied by the hypertension indicator.
  - `age_heart_disease`: Standardized age multiplied by the heart disease indicator.
- **Comorbidity Count:** Sum of hypertension and heart disease indicators (values 0, 1, or 2) to reflect cumulative cardiovascular risk.

4.2.2 Risk Profile Feature:

A simplified risk stratification variable was constructed based on age and key clinical conditions:

- **Low Risk:** Age <50, no hypertension, no heart disease (stroke risk: 0.68%).
- **Moderate Risk:** Age 50–70 without conditions OR Age <50 with conditions (stroke risk: 4.67%).
- **High Risk:** Age >70 OR Age 50–70 with conditions (stroke risk: 16.09%).

This feature yields a 24-fold risk difference between high- and low-risk categories.

4.2.3 Incorporating Unsupervised Learning Features:

Unsupervised learning techniques provided additional dimensions:

- **Principal Component Features:**
  - **PCA_1 (Age-Comorbidity Component):** Explains 18.2% of variance; highly correlated with age (0.82), hypertension (0.65), and heart disease (0.59).
  - **PCA_2 (Metabolic Component):** Explains 12.4% of variance; primarily correlated with BMI (0.72) and glucose level (0.68).
  - **PCA_3 (Lifestyle-Demographic Component):** Explains 10.1% of variance; associated with smoking history, work type, and marital status.
- **Cluster Membership Feature:** A categorical variable denoting membership in one of four clusters identified via K-means (ranging from Cluster 0 with a 1.47% stroke rate to Cluster 3 with a 15.82% stroke rate).

4.2.4 Handling Class Imbalance:

With only 4.9% of cases being stroke-positive, we employed several techniques to mitigate class imbalance:

- **SMOTE:** Synthetic Minority Over-sampling Technique was applied to the training data to increase the proportion of stroke cases.
- **Stratified Sampling:** Both train-test splitting and cross-validation were performed using stratified methods to maintain class proportions.
- **Class Weights:** Applied in model training to ensure that minority class examples received appropriate emphasis.

4.2.5 Insights from Feature Engineering:

The rigorous preprocessing and feature engineering process led to several key insights:

- **Domain Knowledge Integration:** Embedding clinical thresholds (e.g., diabetic glucose levels) produced interpretable and clinically relevant features.
- **Non-linear Transformations:** The quadratic term for age and log transformation for glucose effectively captured the observed non-linear risk patterns to account for complex risk relationships.
- **Predictive Signal in Missing Data:** The creation of a missing BMI indicator preserved an important risk signal that would otherwise be lost with basic imputation.
- **Interaction Effects:** Age-condition interactions and the comorbidity count allowed the model to account for differential effects of risk factors across age groups.
- **Multidimensional Representations:** PCA components and cluster membership features provided valuable abstractions, capturing latent relationships not apparent in raw features.
- **Balanced Learning Despite Imbalance:** Techniques such as SMOTE, stratified sampling, and class weighting ensured that the models could effectively learn patterns despite the inherent class imbalance.

Overall, these preprocessing and feature engineering steps transformed the raw dataset into a structured, enriched feature space, laying a robust foundation for subsequent modeling and predictive analysis.

5 Machine Learning & Deep Learning Models

5.1 Model Selection and Implementation Strategy

To capture the complex, non-linear relationships in stroke prediction while ensuring interpretability, we implemented a diverse set of models spanning traditional machine learning and deep learning techniques.

5.2 Traditional Machine Learning Models:

- **Logistic Regression:** Utilized as a baseline model with balanced class weights. Its interpretable coefficients relate directly to stroke risk despite limitations in modeling non-linear effects.
- **Random Forest:** Ensemble of 100 decision trees (max depth=10, min samples split=10, min samples leaf=4) with balanced class weights, automatically modeling non-linear interactions and providing feature importance metrics.
- **Gradient Boosting (XGBoost):** Configured with 200 trees (max depth=6), learning rate 0.1, subsample ratio 0.8, and scale_pos_weight of 19.5; early stopping based on validation performance.

5.3 Deep Learning Architectures:

- **Feedforward Neural Network (FNN):** A three-layer deep network (256-128-64 neurons) employing ReLU activations, batch normalization, dropout (0.3), and L2 regularization ($\lambda = 0.001$), optimized using Adam (learning rate=0.001) with early stopping on validation AUC.
- **TabTransformer:** Leverages transformer encoders for tabular data by embedding categorical features (embedding dimension=32) and employing 8 attention heads across 4 encoder layers, followed by an MLP head (128 neurons, dropout rate=0.2).

- **Autoencoder-based Model:** A two-phase model with unsupervised pretraining of an encoder (64-32-16 neurons) and a mirrored decoder, followed by a 2-layer classifier (32-16 neurons) on the compressed 16-dimensional latent space. This hybrid approach facilitates robust feature learning through noise reduction and dimensionality compression.

5.4 Model Training and Evaluation Methodology

Data was split using an 80% training and 20% test split with stratification (fixed seed 42). Five-fold stratified cross-validation with SMOTE (applied only to training folds) and class weighting was used.

Primary optimization focused on AUC-ROC, with secondary emphasis on recall given the clinical imperative to minimize false negatives. Additional metrics included PR-AUC, F1-score, and overall accuracy.

5.5 Model Performance Results

The following table summarizes key performance metrics across models:

Model	AUC-ROC	**Precision** (Stroke)	**Recall** (Stroke)	**F1-Score** (Stroke)	PR-AUC
Logistic Regression	0.8335	0.14	0.78	0.24	0.2314
Random Forest	0.7879	0.16	0.06	0.09	0.1503
Gradient Boosting	0.8134	0.17	0.00	0.00	0.1721
FNN	0.8253	0.1376	0.7895	0.2344	0.1893
TabTransformer	0.7717	0.1295	0.7632	0.2214	0.1409
Autoencoder	**0.8382**	**0.1012**	**0.8684**	**0.1813**	**0.1789**

Table 16: Model Performance Comparison

The Autoencoder achieved the highest AUC-ROC (0.8382) and recall (0.8684), correctly identifying 33 of 38 stroke cases on the test set. Its confusion matrix is:

True Negatives: 436 False Positives: 293
False Negatives: 5 True Positives: 33

Thresholds were optimized (e.g., 0.12 for the Autoencoder) to enhance sensitivity.

5.6 Model Comparison and Selection Rationale

While traditional models like logistic regression provided interpretability and tree-based models offered strong overall accuracy, they fell short in recall—an essential metric in clinical settings for stroke prediction. The Autoencoder's superior discriminative power and high recall (0.8684),aligning with the clinical priority of minimizing missed stroke cases made it the preferred model for clinical use.

5.6.1 Key Considerations:

Key considerations in model selection include:

- **Superior Discriminative Ability:** Highest AUC-ROC indicates robust overall performance.
- **Sensitivity / Enhanced Recall:** Highest recall ensures minimal missed stroke cases. Missing only 5 stroke cases, a critical factor given the high cost of false negatives.
- **Representation Learning:** The unsupervised pretraining phase facilitates learning of robust, compressed features, improving generalizability.
- **Threshold Optimization:** Adjusting decision thresholds maximizes clinical utility.

5.7 Insights from Model Development

- **Precision vs. Recall Trade-Off:** The models highlighted the challenge of balancing precision with recall. Given that false negatives pose a greater clinical risk, the emphasis was on maximizing recall.
- **Deep Learning Superiority:** The Autoencoder-based approach benefited from unsupervised pretraining, which captured latent patterns that traditional models and even standard neural networks (FNN) could not fully exploit.
- **Threshold Optimization:** The unsupervised pretraining phase facilitates learning of robust, compressed features, improving generalizability.
- **Threshold Optimization:** : Adjusting the decision threshold was pivotal in achieving high sensitivity, demonstrating that performance improvements can be realized through post-training calibration.

In summary, the Autoencoder model was selected based on its superior ability to identify at-risk patients, providing a strong foundation for subsequent development of a risk scoring system and integration into clinical workflows.

6 SHAP Analysis & Feature Importance

6.1 Understanding Model Decisions Through SHAP Analysis

Model interpretability is paramount in healthcare to validate clinical decisions and build trust in AI-driven tools. Utilized SHAP (SHapley Additive exPlanations) to quantify the contribution of each feature to stroke risk predictions, offering insights at both global and individual levels. SHAP analysis involved:

6.1.1 Methodology for SHAP Analysis

For each model:

- **Global Feature Importance:** Calculated average absolute SHAP values to determine overall impact and to rank features.

- **Direction of Effect:** Assessing whether higher feature values increase or decrease stroke risk.
- **Feature Interactions:** Evaluating SHAP interaction values to understand how features modify each other's impact.
- **Individual Case Explanations:** Using force plots to illustrate how features combine to yield specific risk predictions.

6.2 Key Findings from the Autoencoder Model

The Autoencoder—the selected model for stroke prediction—demonstrated the following global feature ranking:

Feature	Average SHAP Value	Rank
PCA_3 (Lifestyle-Demographic)	0.06814	1
PCA_1 (Age-Comorbidity)	0.05529	2
age_squared	0.05219	3
age	0.03074	4
risk_profile	0.02786	5
age_under_40	0.01858	6
glucose_log	0.01853	7
age_hypertension	0.01460	8
avg_glucose_level	0.01402	9
glucose_normal	0.01291	10

Table 17: Feature Importance from SHAP (Autoencoder Model)

6.2.1 Key Observations and insights:

- **Complex Patterns:** The top two features are PCA components, indicating that multidimensional, unsupervised representations capture critical risk patterns.
- **Age Effects:** The quadratic term (`age_squared`) ranks higher than the linear age term, confirming non-linear risk dynamics with stroke risk.
- **Glucose and Engineered Features:** Both transformed (`glucose_log`) and raw glucose measures appear, while the engineered `risk_profile` ranks highly—validating our domain-informed feature engineering.
- **Interactions:** The `age_hypertension` term is significant, emphasizing the interaction between age and cardiovascular risk.

6.3 Direction of Feature Effects

The SHAP summary plot revealed the direction of feature impacts:

- **Features Increasing Stroke Risk:** Advanced age and `age_squared` (its squared transformation), higher glucose levels, PCA_1 (Age-Comorbidity Component), presence of hypertension and heart disease, and high `risk_profile` scores.
- **Features Decreasing Stroke Risk:** Younger age (`age_under_40`), normal glucose levels (`glucose_normal`), lower BMI values, and absence of comorbidities.

6.4 Feature Impact Visualization and Interpretation

SHAP summary plots illustrate:

- **Feature Clustering:** Grouping of related features such as age-related, glucose-related, PCA components, and medical indicators.
- **Distribution of Impact:** Wide variation in SHAP values for some features (e.g., age) versus binary impacts for others (e.g., heart disease).
- **Feature Value Correlation:** For most features, higher values correlate with higher SHAP values (positive impact), while some (e.g., `glucose_normal`) exhibit inverse relationships.

6.5 SHAP Dependence Plots

Dependence plots revealed:

- **Age Dependence:** Minimal impact until approximately age 50, followed by an exponential increase post-60, confirming the non-linear relationship.
- **Glucose Dependence:** Little impact until around 125 mg/dL, then a sharp increase at diabetic thresholds.
- **BMI Dependence:** A non-linear pattern with highest impact in the overweight range (25–29.9), with a plateau in the obese category.

6.6 Individual Case Explanations

SHAP force plots provided detailed explanations for individual patients:

- **High-Risk Patient Example:** A patient with advanced age (+0.213), high glucose (+0.185), and hypertension (+0.112) had a predicted stroke probability of 0.837 (base value: 0.049).
- **Low-Risk Patient Example:** A patient with, Base value 0.049, with contributions from young age (-0.036), normal glucose (-0.022), and absence of hypertension (-0.009); resulting in a predicted probability of 0.003.
- **Anomalous Case Example:** A young patient with hypertension (+0.139) and very high glucose (+0.127), leading to a predicted probability of 0.289 despite young age, highlighting the nuance captured by interaction terms.

6.7 Interaction Effects Between Features

SHAP interaction analysis revealed critical interactions:

- **Age and Hypertension:** The interaction shows the highest impact in younger patients, moderate in middle-aged, and the lowest relative impact in older patients, highlighting the strong signal of early-onset hypertension.
  - **Younger patients:** Hypertension multiplies stroke risk dramatically (risk ratio 8.24).
  - **Older patients:** Although absolute risk is higher, the relative impact is lower.
- **Age and Glucose Levels:** Elevated glucose has negligible impact in young patients, becomes moderate in middle-aged individuals, and is strongly predictive in older patients.
- **BMI and Other Risk Factors:** BMI's effect is amplified in the presence of hypertension and elevated glucose, reflecting metabolic syndrome patterns.

6.8 Clinical Implications of SHAP Analysis

The SHAP analysis yields several practical insights for stroke risk assessment:

- **Age-Stratified Risk Assessment:** Different weighting for risk factors based on age is crucial and essential.
- **Focus on Early-Onset Conditions:** The pronounced impact of hypertension and heart disease in younger patients calls for early intervention.
- **Metabolic Health Monitoring:** Glucose management is critical, particularly at and above the diabetic thresholds, warrant aggressive management.
- **Personalized Risk Interpretation:** Individual risk should be assessed in the context of the complete profile rather than isolated metrics.
- **Identification of Atypical Cases:** SHAP analysis helps identify patients with non-traditional risk profiles. Recognizing anomalous profiles can aid in targeting preventive measures for patients who do not fit standard risk models.

6.9 Comparative Insights Across Models

While all models consistently identified age, glucose levels, and cardiovascular conditions as risk factors, However few differences were noted:

- Logistic Regression emphasized direct medical indicators, such as comorbidity count.
- Tree-based models distributed importance across a broader set of features but sometimes missed critical interactions.
- The Autoencoder's balanced weighting of both engineered and unsupervised features, combined with its superior recall, validates its selection for clinical deployment. This analysis not only reinforces our model selection but also informs the development of a nuanced risk scoring system that integrates complex interactions and non-linear effects.

These insights validate our model selection and inform the development of a risk scoring system.

7 Risk Scoring System for Stroke Prediction

To bridge the gap between complex machine learning insights and clinical applicability, developed a point-based risk scoring system. This system translates the SHAP-derived feature importance from our best-performing Autoencoder model into an intuitive, clinically usable tool.

7.1 Development of a Risk Scoring Framework

Systematic approach to risk score development involved:

7.1.1 Methodology for Developing the Risk Score

The process involved:

- **Feature Selection:**
  - Starting with the top 15 features by SHAP importance from the Autoencoder model, ensuring representation across key clinical domains (demographics, medical conditions, laboratory values) with clear directional effects.
- **Point Assignment Process:**
  - A maximum score of 102 points was set to provide sufficient granularity.
  - SHAP values were normalized to allocate points proportionally.
  - **Points were assigned in accordance with each feature's importance and the direction of its effect (positive for risk-increasing, negative for risk-reducing).**
- **Scale Development:**
  - A 0–102 point scale was constructed, where higher scores correspond to greater stroke risk.

- Clinically meaningful thresholds were established through statistical validation on the test set.
- The resulting score ranges were mapped to intuitive risk categories for ease of interpretation.

7.2 Components of the Risk Scoring System

The full scoring system assigns points based on patient characteristics, as summarized in the table below:

Feature	Points	Direction
PCA_3 (Lifestyle-Demographic)	+19	Variable
PCA_1 (Age-Comorbidity)	+15	Positive
age_squared	+15	Positive
age	+9	Positive
risk_profile	+8	Positive
age_under_40	-5	Negative
glucose_log	+5	Positive
age_hypertension	+4	Positive
avg_glucose_level	+4	Positive
glucose_normal	-4	Negative
age_over_60	+3	Positive
bmi_obese	+3	Positive
PCA_2 (Metabolic)	+3	Positive
ever_married_Yes	+3	Positive
ever_married_No	-2	Negative

Table 18: Full Risk Scoring System (102-point Scale)

Note: "Variable" indicates that the impact may vary based on the patient's overall profile.

7.3 Clinical Version Without PCA Components

Recognizing that settings without advanced computational resources might not calculate PCA components, also developed a simplified version that retains approximately 92% of the full model's performance. The alternative point assignments are detailed in Table 19:

This version is agreeable to calculation using basic arithmetic, making it suitable for point-of-care decision support.

7.4 Interpretation of Risk Scores for Clinical Decision-Making

Based on analysis of 767 patients in our Test set, three risk categories (Low, Moderate & High) were defined, detailed in Table 20:

7.5 Clinical Decision Support Applications:

- **Screening Prioritization:** Identifies patients for targeted intensive vascular screening based on risk category. Also facilitates efficient allocation of diagnostic resources
- **Intervention Selection:** Guides clinicians in determining appropriate preventive interventions like:
  - Low-risk patients may be advised on lifestyle modifications.
  - Moderate-risk individuals might require combined lifestyle and pharmacological interventions.
  - High-risk patients warrant comprehensive management and frequent monitoring.
- **Monitoring Frequency:** Provides guidance on follow-up intervals like
  - Higher scores suggest the need for more frequent follow-up and reassessment.
  - Changes in risk score over time can signal improvements or deterioration in health status.
- **Patient Communication:** Offers a simple metric for communicating risk to patients.
  - A numerical score (0–102) simplifies the explanation of risk to patients. A concrete score simplifies understanding compared to abstract probabilities
  - Clear risk categories enable shared decision-making and goal setting for risk reduction. By clearly delineating modifiable factors, patients can see the impact of interventions.

In summary, the risk scoring system effectively translates complex machine learning outputs into a practical

Feature	Points	Direction
age_squared	+22	Positive
age	+18	Positive
risk_profile	+12	Positive
glucose_diabetic	+10	Positive
age_hypertension	+9	Positive
heart_disease	+8	Positive
hypertension	+8	Positive
age_over_60	+7	Positive
bmi_obese	+6	Positive

Table 19: Simplified Risk Scoring System (Without PCA Components)

Risk Category	Score Range	Patients	Stroke Rate	Key Characteristics
Low Risk	0–31	363 (47%)	1.38%	Young to middle-aged, few risk factors
Moderate Risk	31–71	397 (52%)	8.06%	Mixed age, some risk factors
High Risk	71–102	7 (1%)	14.29%	Elderly with multiple risk factors

Table 20: Risk Categories and Clinical Interpretation

tool for clinical practice. It not only supports screening and intervention decisions but also facilitates clear communication with patients, thereby enhancing overall stroke prevention efforts.

8 Deployment: FLASK API & Integration in EHR, Website, Mobile App

8.1 Overview of the Flask API

To make our stroke prediction models accessible in real-world healthcare settings, developed a robust Flask API that operationalizes the entire prediction workflow by handling data preprocessing, feature engineering, model inference, and explainability via SHAP analysis. By providing real-time stroke risk predictions, the API serves as the backbone for various deployment channels, including Electronic Health Records (EHR), web portals, and mobile applications.

8.2 API Design and Architecture

The Flask API was engineered with several key principles:

- **Comprehensive Prediction Pipeline:** Includes data preprocessing, feature engineering, model inference, and explainability (via SHAP analysis).
  - **Data Preprocessing**: Implements the same transformations used during model training (e.g., standardization, categorical encoding).
  - **Feature Engineering**: Automatically derives features such as quadratic age (age_squared), glucose and BMI categories.
  - **Model Inference**: Produces stroke probabilities and assigns risk categories.
  - **Explainability**: Utilizes SHAP analysis to return the top contributing features for each prediction.
- **Model Flexibility:** Defaults to the Autoencoder model but supports FNN and TabTransformer models; compatible with both TensorFlow and PyTorch.
- **RESTful Design:** Clearly defined endpoints, standard HTTP methods and status codes, structured JSON requests/responses, and Swagger documentation.
- **Production Readiness:** Supports containerization via Docker, robust error handling, input validation, and horizontal scalability.

8.3 Key API Endpoints

The API exposes endpoints for various functionalities:

- **/prediction/predict (POST):** Accepts patient data in JSON format and returns stroke probability, risk category, and contributing factors.
- **/prediction/select-model (POST):** Allows switching between different model architectures.
- **/preprocessing/feature-importance (GET):** Returns global feature importance rankings.
- **/risk-score/calculate (POST):** Calculates the 102-point risk score from patient data.

8.4 Example API Interaction:

8.4.1 Request:

```
{
  "age": 67.0,
  "gender": "Male",
  "hypertension": 1,
  "heart_disease": 1,
  "avg_glucose_level": 228.69,
  "bmi": 36.6,
  "smoking_status": "formerly smoked",
  "ever_married": "Yes",
  "work_type": "Private",
  "Residence_type": "Urban"
}
```

8.4.2 Response:

```
{
  "stroke_probability": 0.7823,
  "stroke_prediction": 1,
  "risk_category": "High Risk",
  "risk_score": 78,
  "top_factors": [
    {"feature": "age_squared", "contribution": 0.1843,
    "direction": "increases"},
    {"feature": "glucose_diabetic", "contribution": 0.1205,
    "direction": "increases"},
    {"feature": "heart_disease", "contribution": 0.0982,
    "direction": "increases"},
    {"feature": "hypertension", "contribution": 0.0875,
    "direction": "increases"},
    {"feature": "bmi_obese", "contribution": 0.0524,
    "direction": "increases"}
  ],
  "recommendations": [
    "Schedule comprehensive vascular assessment",
    "Regular blood pressure monitoring",
    "Glucose management program",
    "Cardiac follow-up",
    "Weight management program"
  ],
  "model_used": "Autoencoder"
}
```

8.5 Integration into Healthcare Systems

The API engineered can be integrated into various technological platforms like:

8.5.1 Electronic Health Record (EHR) Integration:

- **FHIR Integration:** Accepts and returns FHIR (Fast Healthcare Interoperability Resource)-formatted data, enabling integration with systems such as Epic, Cerner, and Allscripts.
- **HL7 Support:** Supports HL7 message parsing for legacy systems.
- **EHR-Specific Workflows:** Facilitates point-of-care risk assessment, population health screening, and decision support alerts.
  - **Point-of-Care Risk Assessment:** Provides real-time risk scores during patient encounters.
  - **Population Health Screening:** Enables batch processing to identify high-risk patients.
  - **Decision Support Alerts:** Generates real-time alerts for clinicians with actionable recommendations.

8.5.2 Web Platform Integration:

- **Public Health Portal:** Provides a user-friendly interface for population-level risk assessment and geographic risk mapping. This helps supports resource allocation planning based on risk distribution
- **Provider Dashboard:** Secure web application for monitoring risk trends, patient stratification, and outcome tracking.
- **Patient Education Website:** Patient-facing tool for self-assessment, educational content, and goal tracking.

8.5.3 Mobile Application Integration:

- **Sahatna App Integration:** API can be easily integrated with Sahatna Mobile Application. The Lightweight SDK for iOS and Android, provides offline risk calculation and push notification support.
- **Patient-Facing Mobile App:** Simplified risk assessment, longitudinal tracking, and integration with wearable health devices.
- **Provider Mobile Tools:** Bedside risk assessment tools optimized for clinical workflows.

8.6 Use Cases and Implementation Examples

- **Primary Care:** Integration into routine preventive health workflows for automated risk assessment during annual physicals.
- **Hospital Systems:** System-wide stroke risk screening, integration with discharge planning, specialist referrals, and population health analytics.
- **Public Health Departments:** Community-level risk assessment, targeted prevention programs, resource allocation, and outcome tracking.
- **Patient-Centered Applications:** Personalized risk assessments and self-management tools for continuous monitoring and behavior change.

In Summary, the Flask API serves as a critical bridge between advanced stroke prediction models and their practical application in diverse healthcare environments. Its flexible, scalable, and standards-compliant design enables integration across EHR systems, web platforms, and mobile applications, thereby enhancing clinical decision-making and facilitating targeted preventive strategies. This deployment framework not only maximizes the potential impact of our predictive models but also ensures that the insights are actionable at every level of patient care.

9 Key Insights & Public Health Strategies

9.1 Summary of Significant Stroke Risk Factors

Our comprehensive analysis, integrating statistical tests, machine learning models, and SHAP interpretability, consistently identified several key risk factors for stroke. These findings provide a robust foundation for targeted intervention strategies.

- **Age-Related Risk:**
  - Individuals over 60 have a 35.68× higher risk compared to those under 40.
  - SHAP analysis revealed a non-linear risk increase beginning around age 50–55, effectively captured by the quadratic feature `age_squared`.
  - The impact of age is further amplified when combined with other risk factors, supporting both age-based screening and tailored approaches for younger at-risk individuals.
- **Cardiovascular Conditions:**
  - Heart disease is associated with a 4.08× risk ratio (odds ratio: 4.71).
  - Hypertension shows a 3.34× risk ratio (odds ratio: 3.70), with young-onset hypertension demonstrating an even higher relative risk (8.24× for patients under 40).
  - The coexistence of these conditions has a multiplicative rather than additive effect on stroke risk, with SHAP interactions underscoring their joint influence.

- **Metabolic Factors:**
  - Diabetic-range glucose (greater than or equal 126 mg/dL) is associated with a 2.85x increased risk, with a clear threshold effect at the clinical diabetic cutoff was observed, with the impact intensifying with age.
  - BMI exhibits a non-linear relationship, with risk increasing from normal to overweight but plateauing in the obese category.
  - Elevated glucose and high BMI together show synergistic impacts on stroke risk.
- **Sociodemographic Factors:**
  - Work type analysis revealed that self-employed individuals have the highest stroke risk (7.94%).
  - Work type, Marital status and smoking history also play roles, with "ever_married" and "formerly smoked" being notable predictors, while gender and rural/urban residence show minimal direct effects.
  - The PCA-derived lifestyle-demographic component (PCA_3) encapsulates complex sociodemographic patterns beyond individual variables.
- **Novel Patterns and Interactions:**
  - **Age–Hypertension Interaction:** Younger patients with hypertension exhibit a dramatically higher relative risk compared to older patients.
  - **Plateau Effect in BMI:** There is a significant risk increase from normal to overweight BMI, which then stabilizes in the obese range.
  - **Detection of Anomalous Profiles:** Anomaly detection techniques (e.g., Isolation Forest, LOF) revealed atypical risk profiles that merit clinical attention.
  - **Synergistic Effects and Clustering:** Dimensionality reduction and association rule mining highlighted how risk factors cluster together, offering a nuanced understanding of stroke risk.
  - Interaction effects, such as between age and hypertension, and the plateau effect in BMI, provide deeper insights into risk.

9.2 Data-Driven Recommendations for Public Health Policies

Based on our integrated findings, propose the following evidence-based strategies:

- **Targeted Screening Programs:**
  - **Age-Stratified Screening:** Implement comprehensive vascular assessments for those over 60, targeted screening for ages 40–60 with risk factors, and specialized protocols for younger patients with early-onset hypertension. Rationale behind this recommendation is the Clear 35-fold risk difference across age groups; strong age interactions with other factors.
  - **Cluster-Based Stratification:** Utilize risk clusters to intensify monitoring for high-risk groups and anomalous profiles. Rationale behind this recommendation is because Unsupervised analysis revealed distinct subgroups with 10-fold differences in risk.
  - **Occupation-Specific Screening:** Develop programs targeting high-risk occupational groups, especially self-employed individuals. Rationale behind this recommendation is the significant variation in risk by work type suggests targeted occupational interventions.
- **Preventive Interventions:**
  - **Hypertension Control:** Prioritize aggressive management, particularly for younger patients. Set stringent blood pressure targets (less than 130/80 mmHg). Rationale behind this recommendation is because Young-onset hypertension shows a marked 8.24× risk increase, highlighting preventive value.
  - **Glucose Management:** Focus interventions at the diabetic threshold and prevent progression in pre-diabetics. Rationale behind this recommendation is because a distinct threshold effect was observed; highest metabolic impact in older patients.
  - **Integrated Risk Management:** Combine clinical and lifestyle interventions for patients with overlapping risk factors. Rationale behind this recommendation is because significant interactions indicate that multifactorial interventions yield the greatest benefit.
- **Technology Implementation:**
  - **Risk Score Integration:** Embed the 102-point risk scoring system in primary care settings and develop mobile self-assessment tools via the API developed. Rationale behind this recommendation : a validated scoring system offers an accessible tool for real-time risk assessment and tracking.
  - **API-Driven Solutions:** Deploy the Flask API for real-time predictions, population health dashboards, and continuous monitoring. Rationale behind this recommendation : seamless integration across platforms enhances screening, intervention, and follow-up care.
  - **Specialized Technology:** Use remote monitoring, medication adherence tools, and telehealth for high-risk groups.

- **Public Education and Awareness:**
  - **Risk-Stratified Education:** Develop tailored educational materials and awareness campaigns for high-risk demographics. Rationale behind this recommendation is risk stratification supports personalized education, driving behavioral change and early intervention.
  - **Early Warning Signs:** Emphasize recognition of stroke symptoms in high-risk groups through simplified, accessible materials. Rationale behind this recommendation is early recognition significantly improves outcomes; high-risk groups benefit most from targeted education.

9.3 AI-Driven Tools for Preventive Healthcare

Our AI-driven tools contribute to enhanced preventive strategies by:

- **Enabling Real-Time Risk Assessment:** Enables point-of-care decision support and automated screening. Integration into EHRs and mobile applications provides immediate, point-of-care risk evaluation.
- **Delivering Personalized Risk Profiling:** Models capture complex interactions and identify atypical risk patterns, supporting individualized care and assessments.
- **Optimizing Resource Optimization:** Focuses interventions on high-risk individuals to maximize resource efficiency. High-risk patients receive focused interventions, while low-risk individuals avoid unnecessary procedures.
- **Supporting Continuous Learning:** Models evolve with new data, maintaining accuracy over time. The system adapts with new data, continuously refining predictions and improving preventive strategies.
- **Integration Across the Care Continuum:** Links primary care, specialty care, public health, and patient self-management by deploying predictive stroke AI solution.

9.4 Socioeconomic Considerations and Broader Implications

- **Occupational Patterns:** Address work-related stress and healthcare access disparities.
- **Access to Care:** Ensure risk assessment tools are available in underserved communities.
- **Social Determinants:** Incorporate economic, environmental, and educational factors in prevention strategies.
- **Cost-Effectiveness:** Emphasize potential savings from early interventions.

By combining advanced predictive analytics with actionable public health strategies, our approach transforms stroke prevention from a reactive to a proactive endeavour. This integrated model not only reduces the overall burden of stroke but also ensures equitable access to preventive care through tailored, data-driven interventions.

10 Conclusion & Future Directions

10.1 Summary of Key Findings and Contributions

This study has yielded significant contributions:

- **Model Development and Performance:**
  - The Autoencoder model achieved an AUC-ROC of 0.8382 and a recall of 0.8684.
  - Feature engineering captured non-linear relationships and interactions, with PCA components providing valuable abstractions.
  - The 102-point risk scoring system translates complex model outputs into a clinically applicable tool.
- **Risk Factor Identification:**
  - Age is the dominant risk factor, with a 35.68× higher risk for individuals over 60.
  - Heart disease and hypertension are key predictors, with early-onset hypertension being particularly significant.
  - Diabetic-range glucose levels and non-linear BMI effects further influence risk.
- **Clinical Translation:**
  - The risk scoring system stratifies patients effectively, and the Flask API enables real-time integration in clinical settings.
- **Public Health Implications:**
  - Age-stratified screening, targeted interventions, and data-driven resource allocation are recommended.

10.2 Integration of AI and Healthcare: Lessons Learned

Key lessons include:

- Balancing model complexity with interpretability is critical.
- Clinical context must guide model evaluation and threshold adjustments.

- Robust data preprocessing and attention to data quality are essential.
- Implementation requires customization, standards compatibility, and consideration of regulatory and privacy issues.

10.3 Future Directions

Promising directions for future research include:

- **NLP-Powered Monitoring:** Integrate NLP to assess psychosocial factors, enable continuous monitoring via chatbots, and deliver personalized interventions.
- **Longitudinal Data Analysis:** Incorporate time-series data for dynamic risk modeling.
- **Multimodal Data Integration:** Combine clinical, imaging, genomic, wearable, and environmental data.
- **Federated Learning:** Implement privacy-preserving, distributed learning across institutions.
- **Intervention Effectiveness Research:** Evaluate the impact of risk-stratified interventions through randomized controlled trials.
- **Stroke Subtype Prediction:** Develop specialized models for different stroke subtypes.

10.4 Broader Implications for Data-Driven Public Health

This work demonstrates a paradigm shift from reactive to proactive healthcare:

- Personalized risk assessment enables earlier intervention.
- Efficient resource allocation is achieved through targeted prevention strategies.
- Ethical considerations and transparent communication of model limitations are essential.
- The framework provides a template for applying AI to other public health challenges.

10.5 Final Remarks

This comprehensive study illustrates the potential of advanced machine learning and deep learning techniques to enhance stroke prediction and prevention. By integrating rigorous analytical methods with practical deployment strategies, we offer a framework that identifies high-risk individuals, informs targeted interventions, and has the potential to reduce the global burden of stroke. Continued innovation and collaboration among data scientists, clinicians, and public health experts will be critical to translating these findings into meaningful improvements in healthcare outcomes.

11 Reference

11.1 Stroke Prediction Data:

The Stroke Prediction Data utilized in this study can be found here : Data

11.2 Github Code Repository:

Complete Code implementation done for this study can be found here : Code