Introduction

Medical literature and public health databases universally acknowledge that stroke risk factors fall into two categories:

- **Modifiable Risk Factors:** These include high blood pressure (the leading cause), heart disease (such as atrial fibrillation), diabetes, smoking, high cholesterol, obesity, physical inactivity, excessive alcohol consumption, drug use, and, in some cases, the use of oral contraceptives or a history of transient ischemic attacks (TIAs).
- **Non-Modifiable Risk Factors:** These consist of factors such as advanced age, genetic predisposition, gender (with strokes generally more frequent in men but often deadlier in women), and race/ethnicity.

This research set out to validate and refine these conventional risk factors through a comprehensive, data-driven approach. By employing exploratory data analysis (EDA), rigorous statistical testing, advanced machine learning (ML) and deep learning (DL) models, and interpretability techniques like SHAP analysis, we have uncovered nuanced insights that both confirm and extend existing knowledge.

Top Significant Risk Factors Derived from Data & Modelling

1. Age and Non-Linear Effects

- **Widespread Acceptance:** Age is widely recognized as a dominant non-modifiable risk factor for stroke, with risk increasing dramatically with age.
- **Data-Driven Findings**:
  - This study consistently revealed that age is the strongest predictor of stroke. In fact, individuals over 60 are at a 35.68× higher risk compared to those under 40.
  - The introduction of the *age_squared* feature in our models captured the non-linear escalation of risk, particularly showing that stroke risk accelerates sharply beginning around age 50–55.
  - SHAP analysis ranked both the linear (age) and non-linear (age_squared) components highly, confirming that a simple linear model might underestimate the true risk in older populations.

2. Cardiovascular Conditions: Hypertension and Heart Disease

- **Widespread Acceptance:** High blood pressure and heart disease are established as leading modifiable risk factors.
- **Data-Driven Findings**:
  - This analysis showed that heart disease increases stroke risk by approximately 4.08×, while hypertension contributes a 3.34× increase.

    - Notably, the effect of hypertension was even more pronounced in younger individuals (with an 8.24× risk increase for patients under 40), suggesting that early-onset hypertension may be a particularly potent signal of underlying vascular pathology.
    - SHAP interaction values further highlighted that when combined with age, these cardiovascular conditions have a **multiplicative** rather than additive effect on risk.

3. Metabolic Factors: Glucose Levels and BMI

- **Widespread Acceptance:** Diabetes and related metabolic abnormalities are well-known contributors to stroke risk.
- **Data-Driven Findings**:
    - Diabetic-range glucose levels (≥126 mg/dL) were associated with a nearly 3-fold (2.85×) increase in stroke risk. This analysis revealed a clear threshold effect at this clinical cutoff.
    - BMI showed a non-linear relationship: while risk increases significantly when moving from a normal BMI to an overweight category, the effect plateaus in the obese range. This suggests that metabolic dysregulation may be more nuanced than previously thought.
    - Both transformed (e.g., log-transformed glucose) and raw values of these metabolic indicators were among the top predictors in our models, as confirmed by SHAP analysis.

4. Sociodemographic and Lifestyle Factors

- **Widespread Acceptance:** Factors such as smoking, physical inactivity, and high cholesterol are typically acknowledged.
- **Data-Driven Findings**:
    - This study found that sociodemographic factors—such as work type and marital status—also play a significant role. For example, self-employed individuals showed a higher stroke risk, which could reflect work-related stress or differing access to healthcare.
    - The engineered variable risk_profile (a composite score derived from age, hypertension, and heart disease among others) ranked highly, reinforcing the importance of multi-factorial risk assessment.
    - Additionally, latent features extracted through PCA (e.g., PCA_3, which encapsulates lifestyle-demographic patterns) provided novel insights that extend beyond single risk factors.

Feature Interactions & Unexpected Findings

This data-driven approach not only confirmed the traditional risk factors but also revealed several unexpected patterns:

- **Nonlinear Relationships:** The introduction of the *age_squared* variable showcased the exponential risk increase with advancing age; a nuance often oversimplified in standard linear models.
- **Interaction Effects**:
  - A particularly counterintuitive finding was the strong interaction between age and hypertension. Although older patients inherently have a higher absolute risk, the relative impact of hypertension is more pronounced in younger patients.
  - Similarly, interactions between elevated glucose levels and age underscored that the detrimental impact of high glucose is magnified in older individuals.
- **Hidden Patterns:** Advanced anomaly detection methods (Isolation Forest and LOF) identified atypical risk profiles that may be underrepresented in traditional risk models. These profiles indicate that patients who deviate from typical patterns (for example, younger individuals with significant cardiovascular risk) may warrant additional clinical attention.

Conclusion & Public Health Implications

Summary of Findings

Our comprehensive study validates many of the widely accepted stroke risk factors while providing deeper, more nuanced insights:

- **Age** remains the most potent predictor, but its risk is best captured by considering non-linear effects.
- **Cardiovascular conditions** such as hypertension and heart disease are critical, with their interaction effects especially significant in younger demographics.
- **Metabolic abnormalities**, particularly elevated glucose levels and the nuanced behaviour of BMI, are robust predictors (risk factors) of stroke risk.
- **Sociodemographic factors** offer additional predictive power and highlight the importance of integrating lifestyle and environmental data into risk assessments.

Implications for Stroke Prevention

- **Enhanced Risk Stratification:** Data-driven risk scoring systems, like our 102-point scale, can more accurately stratify patients, enabling targeted screening and personalized intervention strategies.
- **Integration into Clinical Workflows:** Embedding these AI-powered tools into electronic health records (EHRs), mobile health applications, and public health dashboards can facilitate real-time risk assessments and improve patient outcomes.

- **Tailored Interventions:** Recognizing that young patients with early-onset hypertension or other atypical profiles are at disproportionately high risk, healthcare providers can design intervention strategies that are more finely tuned to individual risk profiles.

Future Directions: Next steps include leveraging natural language processing (NLP) to monitor psychosocial factors, integrating longitudinal data for dynamic risk predictions, and applying federated learning to improve model generalizability while preserving patient privacy.

In summary, our data-driven analysis not only reaffirms established stroke risk factors but also uncovers complex interactions and non-linear patterns that can enhance predictive modelling. These insights offer a roadmap for more effective, targeted public health strategies aimed at reducing the global burden of stroke.