كلية الحوسبة والمعلوماتية
COLLEGE OF COMPUTING
AND INFORMATICS

جـــامعـــة الـشـــارقـــة
UNIVERSITY OF SHARJAH

Junior Project Proposal

STRYKE-AI: Smart Stroke Risk Evaluation for Young Individuals

Mahaz Ishtiaq Khan	U22200217
Omar Nabulsi	U22100418
Hamza Luai	U22105870

Progress Report 1 CS Junior Project Computer

Science

Supervisor: Dr. Manar Abu Talib
Supervisor: Dr Sohail Abbas

College of Computing and Informatics

University of Sharjah

February/March 2025

Table of Content

1 Introduction

1.1 Overview of the Project

Stroke is an emerging public concern with an increasing trend in young patients due to lifestyle, genetic, and unrecognized factors. Prediction and early detection are fields with immense prevention and proper treatment at the correct moment. The project aims to form a predictive model by the use of machine learning in the estimation of risk for the young patient with multifactorial lifestyle and related factors. Deep learning techniques and medical information along with statistical analysis should provide an early signal to vulnerable people. Supervised learning techniques, feature engineering, and medical information extraction are integrated with one another with the purpose to improve predictive power and assist medical professionals to make early judgments.

1.2 Purpose of this Report

The purpose is to document findings, processes, and outcomes realized during the model development to predict stroke. There is an elaborate literature review, the most relevant machine learning techniques used are explained, and the performance of the models on different test cases is examined. The report aims to demonstrate the promise in the use of artificial intelligence in predictive detection, determination of risk, and probable intervening factors to prevent cases of stroke among the young population. Results and arguments will pave the way for more research and the use of AI-based diagnostic systems in preventive care.

1.3 Project Objectives & Expected Outcomes

Project Objectives

- Development of an Effective Machine Learning Model to Predict the Risk of Stroke in Young Patients: Employing Lifestyle and Clinical Predictors
- Assess major contributory factors such as smoking, blood pressure, diabetes, cholesterol, physical activity, and genetic factors.
- Comparing different machine learning models such as decision trees, random forests, support vector machines (SVM) and deep learning architectures to identify the most suitable method
- Increase the explainability and interpretability of the AI model with SHAP values and feature importance so that the reasoning behind the decisions can be seen by health professionals.
- Use SHAP values on real-world clinical datasets to make real-world performance predictions in terms of precision, recall, F1-score, and ROC-AUC metrics.
- Discuss the potential for incorporating AI into current medical care setups for real-time risk monitoring and decision support.

Expected Outcomes

An intelligent machine learning program able to efficiently predict the likelihood ofstroke among the youth.
Identification of most of the risk factors responsible for predicting stroke.
Comparing different machine learning approaches and performance in predicting the risk of stroke.
Development
of the most probable framework to be incorporated into healthcare to offer early warnings and preventive care recommendations. Scientific studies and case studies reporting on the efficiency of AI in the prevention of strokes.

2 Research & Background

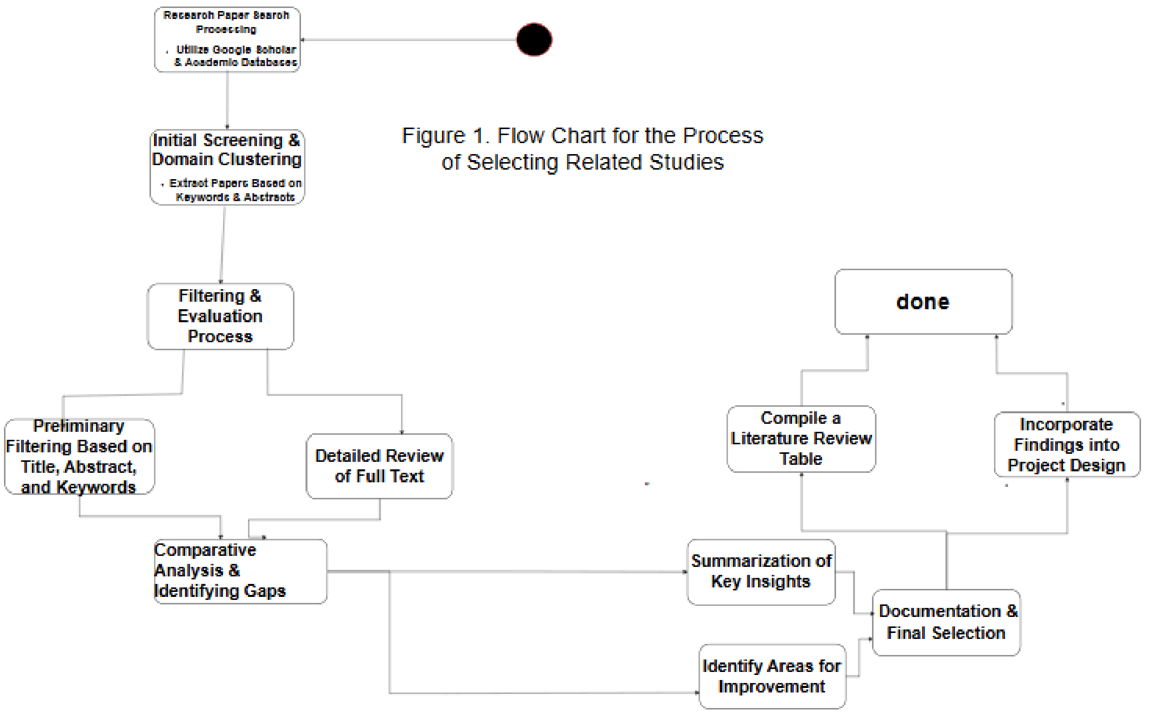

Research Paper Search Processing
Utilize Google Scholar & Academic Databases
Initial Screening & Domain Clustering
Extract Papers Based on Keywords & Abstracts
Figure 1. Flow Chart for the Process of Selecting Related Studies
Filtering & Evaluation Process
done
Preliminary Filtering Based on Title, Abstract, and Keywords
Detailed Review of Full Text
Compile a Literature Review Table
Incorporate Findings into Project Design
Comparative Analysis & Identifying Gaps
Summarization of Key Insights
Documentation & Final Selection
Identify Areas for Improvement

Figure 1. Flowchart of Literature Selection Process

Our project began by taking a comprehensive overview of previous work and solutions existing in the case of machine learning-based prediction for stroke among young people. It is an imperative step since seeing what is prevalent in the case helps us set the basis in order to go about the issue in an optimal manner. By looking at state-of-the-art models, methods, and data used in previous studies, we know what challenges the researchers have faced, what they have used in their approach, and how they have tackled them.

Furthermore, studying related research enables us to identify areas for further research, assess the strengths and weaknesses of current models, and make opportunities for developing a more accurate and robust prediction system. This process guides our project towards new contribution in stroke detection at an early stage, analysis of risk factors, and patient-specific model prediction using machine learning

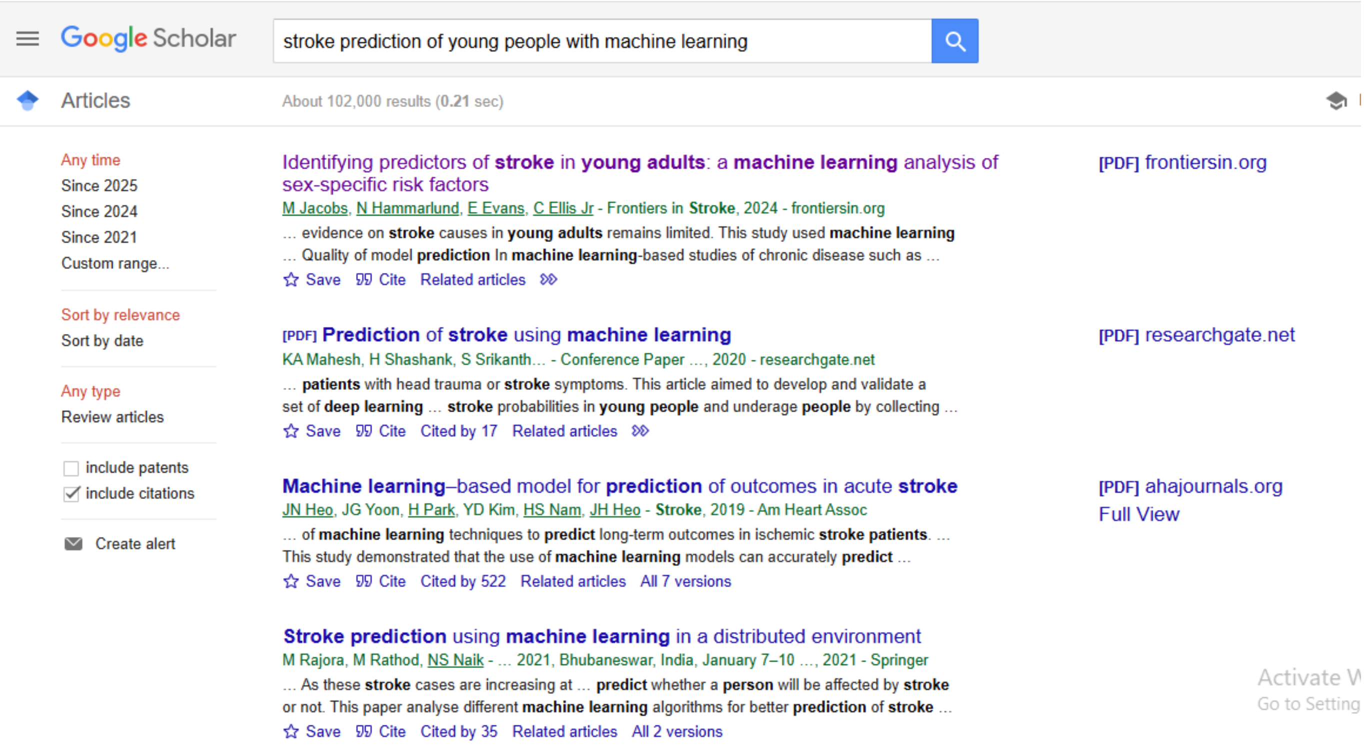
2.1 Literature Review

As our problem is research-oriented and data-based, we went through scientific papers in highly accredited sources, i.e., medical informatics, artificial intelligence, and neurology journals. To achieve this, we employed Google Scholar and other academic search engines to find broad and narrow studies on stroke prediction using machine learning.

- The primary objectives of this literature review were to:
- Gain a better understanding of stroke risk factors and prediction methods.

- Discuss the new trends in machine learning models applied in medical diagnosis.
- Discuss the challenges researchers faced, including data limitation, bias, and model interpretability.
- Identify available datasets and resources that will be useful for our project to be successful.
- Determine how to increase predictive accuracy without compromising ethical and clinical reliability.
- To ensure a serious search, we used certain keywords, including:
- Machine Learning for Stroke Prediction
- Deep Learning in Medical Diagnosis
- Risk Factors for Stroke in Young Adults
- Early Stroke Detection Through AI
- Supervised vs. Unsupervised Learning for Stroke Prediction
- Stroke Risk Analysis with Feature Engineering
- Stroke Prediction Using Medical Datasets
- Bias and Ethics in Diagnosis Through AI
- Interpretable Machine Learning for the Healthcare Domain
- Ensemble Learning for Stroke Risk Analysis

We would get a list of papers for each question, as shown in Figure 2:



stroke prediction of young people with machine learning

About 102,000 results (**0.21** sec)

Identifying predictors of **stroke** in **young adults**: a **machine learning** analysis of sex-specific risk factors
M Jacobs, N Hammarlund, E Evans, C Ellis Jr - Frontiers in **Stroke**, 2024 - frontiersin.org
… evidence on **stroke** causes in **young adults** remains limited. This study used **machine learning** … Quality of model **prediction** In **machine learning**-based studies of chronic disease such as …
☆ Save 99 Cite Related articles
[PDF] frontiersin.org

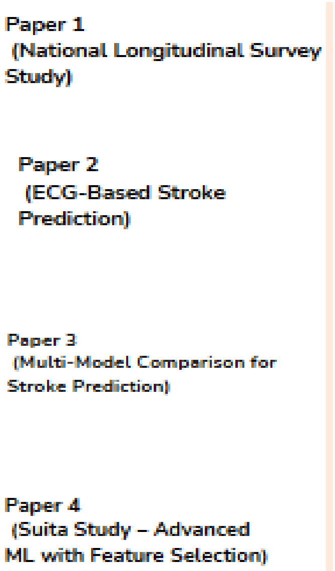
[PDF] **Prediction** of **stroke** using **machine learning**
KA Mahesh, H Shashank, S Srikanth… - Conference Paper …, 2020 - researchgate.net
… **patients** with head trauma or **stroke** symptoms. This article aimed to develop and validate a set of **deep learning** … **stroke** probabilities in **young people** and underage **people** by collecting …
☆ Save 99 Cite Cited by 17 Related articles
[PDF] researchgate.net

Machine learning–based model for **prediction** of outcomes in acute **stroke**
JN Heo, JG Yoon, H Park, YD Kim, HS Nam, JH Heo - **Stroke**, 2019 - Am Heart Assoc
… of **machine learning** techniques to **predict** long-term outcomes in ischemic **stroke patients**. … This study demonstrated that the use of **machine learning** models can accurately **predict** …
☆ Save 99 Cite Cited by 522 Related articles All 7 versions
[PDF] ahajournals.org
Full View

Stroke prediction using **machine learning** in a distributed environment
M Rajora, M Rathod, NS Naik - … 2021, Bhubaneswar, India, January 7–10 …, 2021 - Springer
… As these **stroke** cases are increasing at … **predict** whether a **person** will be affected by **stroke** or not. This paper analyse different **machine learning** algorithms for better **prediction** of **stroke** …
☆ Save 99 Cite Cited by 35 Related articles All 2 versions

Figure 2. List of Papers

We would go through papers that are related by reading the all study info, then the abstracts and conclusions, to determine if they will be of help in our task. Also, we thoroughly read each paper that has proposed solutions and had accurate reported results for our task with best ML models. Following this process, we documented the in the following file for future reference. We are storing them in our local pc's, you can see the papers in Figure 3.



Paper 1
(National Longitudinal Survey Study)

Paper 2
(ECG-Based Stroke Prediction)

Paper 3
(Multi-Model Comparison for Stroke Prediction)

Paper 4
(Suita Study – Advanced ML with Feature Selection)

Figure 3. The stored papers

2.2 Problem Definition & Justification

Machine Learning for Stroke Prediction in Young Adults

Task:

Take a dataset of clinical, demographic, and lifestyle parameters as input and construct a highly interpretable and accurate machine learning model to predict the risk of stroke in young adults.

Stroke prediction studies have focused on older populations, where traditional risk factors such as hypertension, diabetes, and atrial fibrillation are the culprits. Young-onset stroke (<45 years) is characterized differently, ascribed to genetic, metabolic disease, and lifestyle reasons. Breakthroughs in machine learning (ML) and deep learning (DL) have given rise to hopes of early detection and personalized risk assessment.

We recognized that machine learning-based stroke predictive models vary in methodology, including:

- Supervised learning models trained from labeled data made up of stroke events.
- Unsupervised learning methods for clustering the high-risk individuals based on underlying patterns.
- Hybrid models that merge clinical guidelines with AI-driven prediction to improve interpretability.

Machine learning methods most applied in stroke prediction are:

- Logistic Regression & Decision Trees for traditional risk factor analysis.
- Random Forest & XGBoost for ranking feature importance.
- Neural Networks (LSTMs, CNNs, Transformers) for deep pattern recognition from patient data.
- Autoencoders for anomaly detection and dimensionality reduction in rare strokes.

Multimodal data has been emphasized by recent studies to include ECG signals, metabolic markers, and lifestyle parameters for improved prediction accuracy.

Clinical and Biochemical Markers for Stroke Risk in Young Adults

Task:

Enumerate the most predictive clinical and biochemical markers for stroke risk stratification in the young.

Traditional stroke risk prediction models make use of traditional risk factors such as age, hypertension, diabetes, and cholesterol. However, in young adults, non-traditional factors play a very significant role. Based on recent research, the following clinical parameters and biomarkers have been identified as having high predictive potential:

- **Blood Pressure & Hypertension:** SBP >140 mmHg significantly increases the risk of stroke.
- **Metabolic Syndrome (MetS):** Linked to early stroke through the constellation of high BMI, insulin resistance, and lipid derangement.
- **Fructosamine & Blood Glucose Levels:** Strong indicators of occult diabetes and metabolic derangement affecting stroke risk in the young
- **Non-HDL Cholesterol & Lipid Imbalances:** Non-HDL cholesterol has recently been identified as a stronger predictor than the traditional level of LDL.
- **Estimated Glomerular Filtration Rate (eGFR):** Even early chronic kidney disease has been associated with a high risk of stroke in young individuals.
- **ECG Abnormalities:** Atrial fibrillation occurs less frequently in young stroke patients, but other ECG-derived markers, such as P-wave dispersion and heart rate variability (HRV), have shown promising early markers.

Machine learning models trained on a composite of the aforementioned features perform better at predicting stroke among young individuals than conventional risk calculators like the Framingham Stroke Risk Score (FSRS).

AI-Driven Stroke Risk Stratification and Early Warning Systems

Task:

Develop an AI-powered stroke prediction system that integrates real-time health monitoring data with clinical records.

With the rise of wearable health technology (e.g., smartwatches, continuous glucose monitors), machine learning models can now analyze real-time physiological signals to detect stroke risk before clinical symptoms appear.

Recent works on real-time stroke monitoring leverage:

- Deep learning models on ECG signals to detect abnormal heart rhythms linked to stroke risk.
- Recurrent Neural Networks (RNNs) and LSTMs to analyze time-series health data.
- Federated learning approaches to train models across different hospitals while preserving patient privacy.

One major study demonstrated that a deep learning-based ECG classification model achieved >85% accuracy in predicting stroke-prone individuals months before clinical diagnosis. This supports the idea that integrating wearable data with machine learning could revolutionize early stroke detection in young adults.

Stroke Prediction Using Multi-Modal Data Fusion

Task:

Enhance stroke risk prediction accuracy by combining multimodal data, i.e., clinical, metabolic, and ECG features.

Most traditional stroke prediction models are single-source data-based, e.g., clinical history or laboratory tests. Recent literature has shown that the combination of multiple types of data significantly improves predictive performance.

A recent multi-modal AI model successfully combined:

- Structured clinical data (hypertension, BMI, smoking).
- Biochemical markers (glucose, non-HDL cholesterol, fructosamine).
- ECG signal features (heart rate variability, QT interval changes).

This integration approach enabled deep learning models to detect complex interactions between cardiovascular, metabolic, and neurological variables more accurately (AUC > 0.90) compared to single-source models (AUC ~0.75-0.85).

Researchers are further investigating Graph Neural Networks (GNNs) to model interactions between different stroke risk factors with encouraging results in preliminary experiments.

Explainability and Bias in AI Stroke Prediction Models

Task:

Render AI-based stroke prediction models interpretable, unbiased, and clinically meaningful.

One of the intrinsic challenges of stroke prediction based on machine learning is the "black-box" nature of deep learning models. To address this, recent research has focused on:

- Feature importance analysis using SHAP (Shapley Additive Explanations) to determine the risk factors contributing the most to a prediction.
- Interpretable AI models, i.e., decision trees and rule-based ML algorithms, for clinically explainable results.
- Bias detection and mitigation, so that models do not disproportionately overpredict or underpredict stroke risk by demographics (e.g., gender, ethnicity, or socioeconomic status).

One recent study found that models trained on biased data (e.g., over-representation of older stroke patients) performed poorly on younger age groups. Adjustment for this bias significantly improved the generalizability of the model.

2.3 Related Work & Existing Solutions

Key Takeaways for Our Project

From these researches, our project can be improved by addressing the following issues:

Dataset Quality & Availability

- We should collect a diverse and balanced dataset to avoid bias.
- Considering using real-time hospital data and external validation datasets.

Model Selection & Optimization

- Will Use an ensemble approach (combine XGBoost, LSTM, and Random Forest).
- Performing extensive hyperparameter tuning and feature selection.

Interpretability & Clinical Usability

- Using explainable AI (XAI) techniques so doctors can trust the predictions of the model.
- Visualizations should highlight significant risk factors.

Evaluation & Validation

- Employ cross-validation (K-fold, leave-one-out) for robustness.
- Will Test on real patient cases to assess real-world utility.

Deployment & Performance

- Consider a real-time web application or mobile app for patient interaction.
- Keeping computational cost low so that hospitals can use it.

By analyzing the studies mentioned above we also obtain an important point about the :

Feature Selection Strategy (from best to worse Predictors of Stroke) as follows:-

1. Must-Have Features (Top Contributors)
   Age, SBP, Hypertension, Blood Glucose, ECG Signal, BMI, Smoking Status
   - These are clinically proven to have the strongest influence on stroke risk.
   - Removing them will significantly drop model performance.
2. Secondary Important Features (Enhance Performance)
   eGFR, Metabolic Syndrome, Fructosamine, Hemoglobin, Non-HDL Cholesterol
   - These enhance prediction power but are not primary indicators.
3. Less Critical Features (Consider Removing)
   Calcium levels, Elbow Joint Thickness
   - Little clinical evidence directly linking them to stroke in young patients.

Table 1 mention:

Key Takeaways for Our Project

Based on these studies, our project can be improved by addressing the following points:

Dataset Quality & Availability

- We should obtain a balanced and varied dataset to avoid introducing bias.
- Make use of real-time hospital data and external validation datasets.

Model Selection & Optimization

- Use an ensemble approach (stack XGBoost, LSTM, and Random Forest).
- Perform extensive hyperparameter tuning and feature selection.

Interpretability & Clinical Usability

- Employ explainable AI (XAI) techniques so that doctors will accept the model's output.
- Visualization should highlight significant risk factors.

Evaluation & Validation

- Employ cross-validation (leave-one-out, K-fold) for reliability.
- Test against real patient cases to confirm performance in real-world settings.

Deployment & Performance

- Design a web or mobile application to involve patients in real-time.
- Optimize computational efficiency to ensure viability for hospitals.

Table 1. Results and key findings and Comprehensive comparison of Stroke Prediction Studies

Paper	**Machine Learning Models Used**	**Key Findings**	**Challenges & Limitations**	**Conclusion & Insights and how to be better**	**Evaluation and Validation**
Paper 1(National Longitudinal Survey Study)	**LASSO Regression**	**- Stroke risk factors differ by sex.** **- Women: Marijuana use, kidney disease, migraines, depression, PTSD.** **- Men: Income, kidney disease, heart disease, diabetes, PTSD, anxiety.**	**- Small number of stroke cases.** **- Self-reported data introduces bias.** **- Some risk factors missing (e.g., genetics).**	**- Machine learning improves stroke prediction but requires more real-world validation.** **- Sex-specific risk factors should be considered in stroke prediction models.**	**Used standard classification metrics (Accuracy, Precision, Recall, F1-score).** **- No cross-validation reported.**
Paper 2(ECG-Based Stroke Prediction)	**CNN, LSTM, Deep Neural Networks (DNNs)**	**- ECG signals effectively predict stroke risk.** **- CNN-LSTM outperforms other models for time-series medical data.** **- AI improves early stroke detection.**	**- Requires large datasets for deep learning.** **- Difficult to interpret CNN and LSTM models.** **- No real-world testing.**	**- ECG data can enhance stroke risk prediction.** **- Explainability and real-world validation are needed.**	**Evaluated with real-world hospital data.** **- AUC-ROC curve analysis performed.** **- No real-time testing with patients.**

Paper 3 (Multi-Model Comparison for Stroke Prediction)	Random Forest (RF), SVM, Decision Tree (DT), Naïve Bayes (NB), KNN	- RF & SVM performed best for stroke risk classification. - Key risk factors: Age, hypertension, heart disease, glucose levels, smoking.	- Dataset size (5,110 records) is limited. - Naïve Bayes had a high false positive rate. - No integration with real-world healthcare systems.	- Random Forest & SVM models are reliable for stroke risk prediction. - Requires integration with clinical systems and patient monitoring.	Used cross-validation and confusion matrices. - Reported F1-score but lacked external dataset validation.
Paper 4(Suita Study Advanced ML with Feature Selection)	k-Prototype Clustering, Logistic Regression (LR), RF, SVM, XGBoost, LightGBM	- Identified novel risk factors: Elbow joint thickness, fructosamine levels, calcium levels. - RF had the best accuracy (70%). - SVM had the highest AUC (73%).	- Limited to a Japanese dataset (7,389 records). - Unsupervised cluste ring introduces bias. - No clinical deployment.	- Random Forest & SVM models are highly effective for stroke prediction. - Feature selection is crucial for improving prediction accuracy.	- Used K-fold cross-validation. - Reported sensitivity, specificity, and AUC scores. - No longitudinal study validation.

2.4 Stakeholder Meetings & Knowledge gathering

As part of the project's research work, we managed to interview **Professor Eman Abu-Gharbieh, Professor and Vice Dean at the University of Sharjah's College of Medicine**. We were provided with valuable information during the meeting regarding the risk factors for stroke, its management, and the trend in the region. She also showed us a retrospective work on young-onset stroke in the UAE, in which she was an author. This work provided us with valuable information, including a dataset conducted locally with a lot of parameters, about the epidemiology of stroke, including the risk factors, treatment gaps, and implications for the region's public health.

Key Discussion Points

- Research into youth-onset stroke (patients younger than 50 years) in the UAE and differences with regard to its risk factors and stroke in the elderly.
- The extremely high prevalence of hypertension and diabetes in stroke patients, emphasizing the role of modifiable risk factors.
- Low utilization rates for newer thrombolytic therapies such as mechanical thrombectomy and IV fibrinolysis in the UAE when contrasted with the West.
- The requirement for stronger data-informed models for the risk of stroke, tuned to the regional population and clinical practice.

Throughout the discussion, Professor Eman presented the problems that are faced in the management and prevention of stroke among young patients in the UAE. She presented the need for early management, lifestyle changes, and overall awareness in the prevention against stroke.

Key Insights from the Study

The studies that were presented to us provided us with valuable information regarding the epidemiology, risk factors, and treatment outcomes in the UAE. Some of the salient points that were presented are:

1. Stroke Risk Factors

- The most prevalent comorbidities in the stroke patients were **hypertension (45.45%)** and **diabetes (24.42%)**.
- Other causes were **obesity (7.14%)** and coronary heart disease.
- There were only **1.05%** with **atrial fibrillation**, implying the pathways for stroke here could be different than in Western countries.
- Obesity, smoking, and physical inactivity are the lifestyle factors responsible for an increased incidence of strokes in the young.

2. Stroke Types & Demographics

- The most common were **ischemic stroke (50.12%)**, **hemorrhagic stroke (27.21%)** and **subarachnoid hemorrhage (22.67%)**
- The population we were studying was **predominantly male (78.28%)** but is typical for international cohorts with a greater gender discrepancy than occurs in Western birth cohorts.
- The vast majority **(63.96%) were South Asian** patients, mirroring the expatriate population in the UAE and posing ethnicity-specific questions about the risk for stroke.

3. Gaps in Treatment and Intervention

- Only **9.41%** of stroke patients were **treated with IV fibrinolysis** despite its significance in acute stroke management.
- In only **0.49%** of cases was **mechanical thrombectomy used**, which is much lower than in Western healthcare systems, with utilization at more than **10-15%**.
- Overall **mortality was 12.68%**, which is higher than in most other high-income countries and can be explained by delay to and also by the restricted options for interventions.

4. Public Health Consequences

- The findings affirm the urgent need for the timely management and treatment of diabetes, hypertension, and obesity to prevent the occurrence of strokes.

- There is insufficient awareness with regard to the prevention of stroke, which can be fulfilled by predictive models and public campaigns.
- Data-driven predictive measures for the anticipation of stroke can facilitate early detection and focused interventions to enhance patient outcome.

Application to Our Project

1. Ensure our dataset includes the most common regional risk factors (hypertension, diabetes, coronary artery disease).
2. Consider the low utilization of mechanical thrombectomy and IV fibrinolytic therapy when modeling stroke outcomes.
3. Incorporate demographic weightings into our ML model, given that South Asian populations appear to have a higher prevalence of stroke in the UAE.
4. Explore potential collaboration with medical researchers to refine our model and ensure clinical applicability.

3 Project Development and Progress

3.1 System/Software Architecture

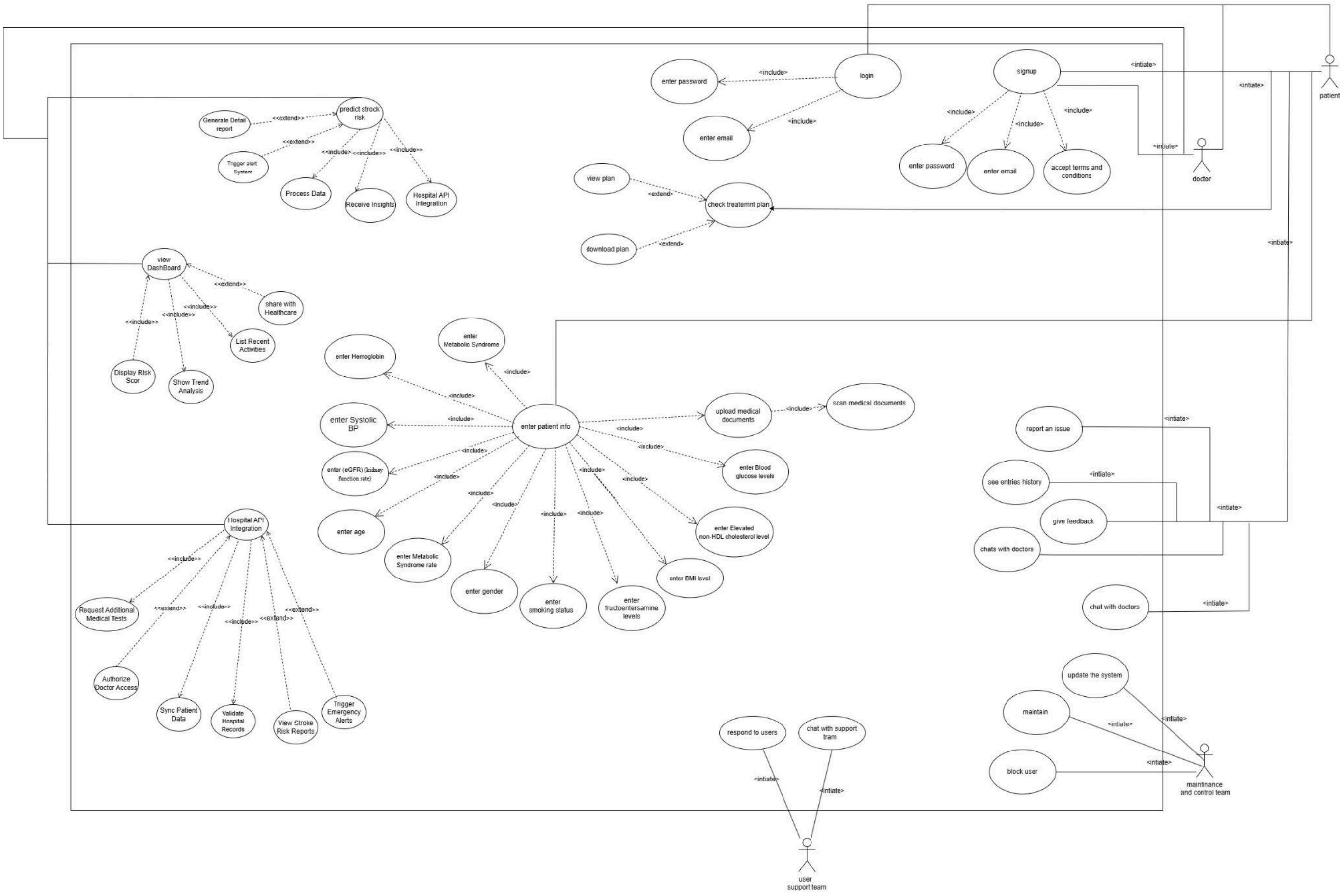


Figure 4. Use Case Diagram

Figure 4 is a top-level representation of interactions between various stakeholders and the Artificial intelligence-based stroke prediction. The diagram illustrates the functionality by the various users including the patients, doctors, managers at the hospitals, and the support/maintenance staff.

The user interfaces with the system by entering personal and medical details like age, gender, smoking, blood pressure, and general medical history. They are also able to view the risk score for the likelihood of the person having a stroke, which is derived by the use of hospital APIs. Users are also presented with the choice to request real-time support. Lastly, the Stroke Prediction SP is the central component to the system, with the core functions being incorporated.
Medical Staff use the system to build complete reports, review patient history, request physician consultations, update the system, and send back feedback to the system administrators. Telehealth integrations, stroke triage, and integration with smart devices for ongoing monitoring are also supported by the system.

The Hospital Administration utilizes the system to calculate the information and send reminders to the patients with the highest risk. It is possible to exchange with external sources. Support and Maintenance Team maintains the system, correct bugs, and locks users so that the system can be used effectively and securely. A user can login or register to the system. They have to enter valid information like email and password.

Use case diagram provides a clean and structured representation of the role played by the system and the interaction with the user, depicting how it can ultimately be beneficial for better risk stratification for strokes, patient management, and coordination among care providers. How the use cases communicate shows the vision for the overall endeavor for the prevention of strokes, consolidating different stakeholders and activities into one system.

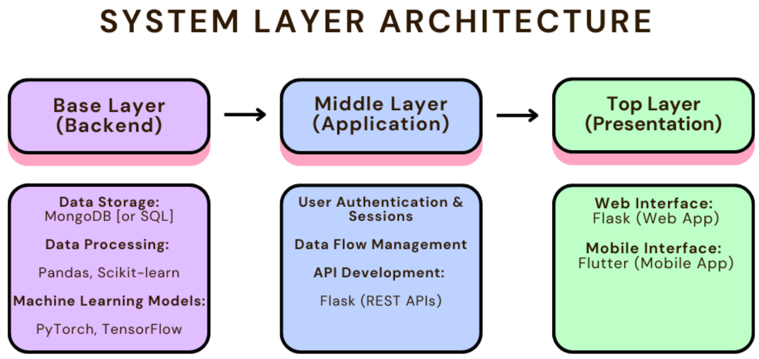


Figure 5. System Architecture

Figure 5: Our AI-based stroke prediction system made up of three layers combined
There are three levels of operations for the AI-based procedure for the monitoring of strokes proposed in this study. There is a Backend Level of data storage and processing through the use of Pandas and Scikit-Learn. PyTorch or TensorFlow-based machine learning models are employed for predictive analysis of strokes.
The Application Layer is placed in the middle of the backend and end-user. It takes care of running the business logic, session interactions as well as managing users accounts, and ensuring safe and organized data exchange within the system.
The Presentation Layer manages risk assessment display and data input management and supports web applications using Flask and smartphone applications using Flutter.
Implementation is based on Python packages and libraries. Scikit-Learn is the primary library for model and feature selection. Pandas and NumPy are employed for data processing. Matplotlib is employed for graphical data representation. Database management is performed using MongoDB or SQL. Docker is employed for cloud deployment on AWS. Monitoring of the model performance is performed using MLflow. Monitoring of the system in real-time is performed using Prometheus/Grafana.
Jupyter Notebook and GitHub allow for workflow management to be made easy. They are implemented on the web and mobile platforms through Flask and Flutter, respectively.

3.2 Technologies, Tools, and Platforms Used

The junior project aims to experiment and implement machine learning techniques on publicly available datasets to predict stroke. We preprocess the datasets, train the models, and make performance comparisons. In the later parts, and the senior project, this can be extended to more complex datasets, including the research conducted by the university.

Frameworks & Libraries for Machine Learning

The following are the tools we utilize for constructing, training, and testing our stroke risk estimation model:

- **Scikit**
  - Supports Decision Trees, Random Forest, and Logistic Regression.
  - Used in preprocessing, feature selection, and model assessment.
- **TensorFlow**
  - Deep learning toolkit to build neural networks.
  - Enables fine tuning the stroke forecasting models.
- **XG**
  - Used to improve model performance by boosting.
  - Helps to compare rival models for stroke risk modeling.

Data handling and preprocessing

For preparing and cleaning datasets for machine learning, we utilize:

- **NumPy**
  - Performs mathematical operations on numeric values.
- **SciPy**
  - Used for probability computations and statistical conversion.

Data Visualisation

To explore distributions of datasets and pattern plots for stroke risk:

- **Matplotlib & Seaborn**
  - Used to chart trends in the risk factors for stroke (such as blood pressure vs. cholesterol level).
- **Plotly (Optional for Interactive Graps)**
  - May be used to build interactive visualizations.

Data Sources

For the junior project; We use publicly available Kaggle datasets to build the model for stroke prediction. Further details in Section 3.3.

Environmental Development

In coding and testing, we use:

- **Jupyter Notebook**
  - Provides an in-browser Python environment to build models.
- **VS Code**
  - Used for creating Python programs and other components.
- **Google Colab**
  - Supports GPU-based machine learning for big models.

Web Framework

to ultimately develop an interactive interface:

- **Flask**
- **Streamlit** (For Quick UI Creation)
  - provides an accessible Python front-end to work with models.

Deployment & Future Considerations

The application will allow users to add profiles for family members, including themselves, by inputting relevant health information. The app will then generate a stroke prediction score, accompanied by explainable AI insights and other key features. **Flutter / Flutter Flow** will be used to develop a mobile application prototype, Future enhancements may include expanding and scaling functionalities and refining the user experience.

.

3.3 Dataset collection & Preprocessing

Overview of the Datasets

For this project, we use **publicly available datasets** that include **patient health parameters** associated with stroke risk. These datasets contain **various risk factors** such as **hypertension, heart disease, cholesterol levels, and smoking status**.

Dataset Name	Source	Records	Features	Missing Values?
Stroke Prediction Dataset	Kaggle	5,110	11	Yes
Brain Stroke Dataset	Kaggle	10,000	12	Yes
Full-Filled Brain Stroke Dataset	Kaggle	10,000	12	No (Preprocessed)

Features of the Dataset

Feature Name	Description	Data Type
id	Unique identifier for each patient	Integer
gender	Patient's gender (Male/Female)	Categorical
age	Patient's age in years	Numeric
hypertension	1 = Patient has hypertension, 0 = No	Binary
heart_disease	1 = Patient has heart disease, 0 = No	Binary
ever_married	Whether the patient was married	Categorical
work_type	Type of employment (Private, Govt, Self-employed, etc.)	Categorical
Residence_type	Urban or Rural residence	Categorical
avg_glucose_level	Average blood glucose level (mg/dL)	Numeric

bmi	Body Mass Index	Numeric
smoking_status	Smoking history (Never smoked, Formerly smoked, Smokes)	Categorical
stroke	1 = Patient had a stroke, 0 = No stroke	Binary (Target Variable)

Each dataset contains the following **key attributes** related to stroke risk:Target variable is **stroke**, which we are attempting to predict based on risk factors.

Preprocessing the Dataset

We need to clean and preprocess the datasets prior to training a machine learning model.

1. Missing Value Handling

- Missing values in features like BMI and smoking status are present in some datasets.
- We employ the following imputation methods:
- BMI: Fill missing values with the median BMI value of the dataset.
- Smoking Status: Fill missing values with "Unknown".

2. Categorical Data Encoding

- Machine learning models do not take text-based categorical features directly.
- We convert categorical variables to numerical representation:
- Gender: "Male": 0, "Female": 1
- Residence Type: "Urban": 0, "Rural": 1
- Smoking Status: "Never smoked", "Formerly smoked", "Smokes", "Unknown" → One-Hot Encoding

3. Normalization & Scaling

- There are features of different scales like age, mean glucose level, and BMI.
- We apply Min-Max Scaling for value normalization between a range of 0 to 1 to obtain better model performance.

4. Splitting Data for Training & Testing

- We split the dataset into **70% training data** and **30% testing data**.
- This guarantees that the model learns on one side of the data and is tested on the other.

Summary of Preprocessing Steps

Preprocessing Step	Method Applied
Handling Missing Values	Fill BMI with median, replace missing smoking status with "Unknown"
Encoding Categorical Variables	Convert gender, work type, residence type into numeric format
Feature Scaling	Apply Min-Max Scaling to normalize numerical values
Train-Test Split	70% training, 30% testing

3.4 Key Features and Functionalities Implemented

Overview of the Implementation Plan

The major activity of this phase is to develop a machine learning model that will predict the risk of stroke in terms of patient health variables. The implementation will be carried out in the sequence:
Preprocessing of the dataset (as outlined in Section 3.3).
Training a baseline machine learning model (Logistic Regression).
Measurement of model performance in terms of accuracy, precision, recall, and F1-score.
Testing more complicated models (Random Forest, XGBoost).
Creating a rudimentary API or interface mock-up for model testing in the coming phases.

Features Planned

The very first implementation phase will include the following basic functions:

- **Dataset Preprocessing:** missing values management, categorical feature transformation, and numeric value normalization.
- **Training Baseline Model:** the baseline machine learning model as Logistic Regression.
- **Performance Measurement:** Accuracy, precision, recall, F1-score of the model.
- **Feature Selection & Importance Analysis:** Identifying which health measures are most responsible in stroke predictions.
- **Mobile Application Development:** Developing a Flutter-based application wherein users can input health measures, view stroke risk scores, and receive AI-based insights in an interactive and easy-to-use platform.
- **Testing More Advanced Models:** Including Random Forest and XGBoost for improved predictions.
- **Explainable AI Integration:** Providing model decision interpretability, enabling users to view why a specific stroke prediction score was returned.
- **Prototype API for Predictions:** Creating a simple Flask or Streamlit-based API for user input and model prediction.

Initial Model Training (To Be Done)

The first prototype of the model will use Logistic Regression as a baseline to provide benchmark performance before evaluating more advanced models.

1. Training the First Model (Logistic Regression)

- Dataset preprocessing will be performed for handling missing values, categorical columns, and numerical scaling.
- A Logistic Regression model will be fitted with Scikit-learn.
- The dataset will be split into **80% train data** and **20% test data** for testing.

2. Model Performance Evaluation

Once the model is trained, in our second progress report, it will be tested on accuracy, precision, recall, and F1-score.

Metric	Score
Accuracy	*To be filled after training*
Precision	*To be filled after training*
Recall	*To be filled after training*
F1-Score	*To be filled after training*

Code Implementation for Initial Model Training

Below is the Python script that will be used for **preprocessing the dataset and training the first model (Logistic Regression).**

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler, LabelEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report
# Load dataset
df = pd.read_csv("stroke_prediction.csv")
# Handle missing values
df["bmi"].fillna(df["bmi"].median(), inplace=True)
df["smoking_status"].fillna("Unknown", inplace=True)
# Encode categorical features
label_encoders = {}
for col in ["gender", "work_type", "Residence_type", "smoking_status"]:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le
# Normalize numerical features
scaler = MinMaxScaler()
df[["age", "avg_glucose_level", "bmi"]] = scaler.fit_transform(df[["age", "avg_glucose_level", "bmi"]])
# Define features (X) and target (y)
X = df.drop(columns=["stroke"])
y = df["stroke"]
# Split dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Train Logistic Regression model
model = LogisticRegression()
model.fit(X_train, y_train)
# Make predictions
y_pred = model.predict(X_test)
# Evaluate model (Scores to be added after training)
accuracy = accuracy_score(y_test, y_pred)
print("Model Accuracy:", accuracy)
print(classification_report(y_test, y_pred))
```

4 Challenges & Implementation Issues

4.1 Technical Challenges Encountered

Model Complexity & Optimization Problems

Challenge: The machine learning model needed to be optimized with numerous hyperparameters (learning rate, batch size, number of hidden layers, etc.), which was highly computation-intensive.

Impact: High resource utilization made training sluggish with numerous iterations needed to converge to the best parameters.

Solution:

- Applying Grid Search and Random Search for hyperparameter tuning.
- Utilizing GPU acceleration to speed up training.

Data Preprocessing Complexity

Challenge: The data contained missing values, errors, and inconsistencies that needed to be cleaned out prior to training.

Effect: Poor-quality data yielded incorrect predictions and less precise models.

Solution:

- Using imputation techniques to fill in missing values.
- Applying outlier detection techniques to eliminate unrealistic and anomalous values.
- Normalizing and standardizing numerical features to ensure consistent scaling.

Wearable & Hospital API Integration

Challenge : The system has to pull real-time health data from wearable devices (smartwatches, blood pressure monitors) and synchronize with hospital records.

Impact: Variations in data format and API response structure resulted in data synchronization problems.

Solution:

- Creating custom data parsers to convert incoming data into a standardized format.
- Utilizing asynchronous API calls to efficiently process large volumes of patient data.

Handling Imbalanced Data for Model Training

Challenge: The data included very few instances of stroke as compared to non-stroke instances, thus model predictions became biased (towards predicting non-stroke).

Impact: The model failed to accurately predict stroke risk in some cases.
Solution:

- Using SMOTE (Synthetic Minority Over-sampling Technique) for creating synthetic data for balancing.
- Employing cost-sensitive learning to give additional weightage to stroke instances.

4.2 Data Limitations & Processing Issues:

Our stroke prediction model faced several significant data-related challenges that impacted development and overall performance. These challenges primarily stemmed from limited dataset availability, class imbalance, missing and inconsistent data, heterogeneous data formats, and privacy concerns. Addressing these challenges is critical to improve the reliability and generalization of the model.

Limited Availability of High-Quality Stroke Data

Challenge:

- High-quality datasets, particularly on young-onset stroke in the UAE population, are not available in sufficient quantities.
- Stroke in the young is a rare condition in comparison to the general population, and hence it is difficult to collect sufficient data to train an accurate and generalizable model.

Impact:

- A small dataset increased the likelihood of overfitting, with the model working well on training data but not being able to generalize to new cases.
- Biases in models can occur due to underrepresentation of different ethnicities, lifestyles, and genetic weaknesses Possible.

Potential Solution:

- Reflect on including global medical research data from elsewhere in UAE to make training data more diverse.
- Explore transfer learning through pre-training the model on bigger stroke datasets before fine-tuning with UAE data.

Class Imbalance in Stroke Data

Challenge:

- The data contained significantly lower numbers of stroke instances compared to non-stroke instances, which created a severe class imbalance.
- Machine learning algorithms trained on imbalanced data are susceptible to being biased towards the majority class at the cost of lower sensitivity towards high-risk patients.

Impact:

- The model may have high overall accuracy but low capacity to recognize actual stroke cases (low recall for strokes).
- There will be a high number of false negatives with undiagnosed high-risk patients, and false positives will lead to unnecessary stress and medical interventions.

Potential Solution:

- Employ oversampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to over-sample the minority class (stroke cases).
- Explore cost-sensitive learning strategies to adjust training penalties for misclassified stroke cases.
- Investigate anomaly detection techniques to improve detection of out-of-pattern cases.

Missing & Inconsistent Data

Challenge:

- Several medical histories had missing values, especially for biomarkers, lifestyle, and genetics.
- Patients did not report their full medical history, so there were gaps in key variables like blood pressure, cholesterol, and smoking status.

Impact:

- Missing values can reduce the model's capacity to predict by limiting important pointers towards stroke risk.
- Missing values may also reduce the dataset size to train from.

Potential Solution:

- There are possible imputation strategies (mean, median, or KNN imputation) to address this problem.
- Create data validation pipelines that automatically identify missing or incorrect data.

Data Format & Integration Challenges

Challenge:

- Various medical data has been stored in various formats for various providers.
- Hospitals store their data either by using databases with some structure (such as SQL systems), whereas others store it by using non-structured forms (in the form of PDFs and written text).

Impact:

- Incompatibility of the data can prolong model training and preprocessing.
- By-hand data cleanup could be labour-intensive, delaying model updates.

Potential Solution:

- We may be required to build mapping algorithms in order to normalize different data types into a set format.
- ETL (Extract, Transform, Load) workflows could be examined to preprocess automatically.

Privacy & Compliance Concerns

Challenge:

- The medical data is extremely sensitive and requires strict conformance to the privacy laws (e.g., GDPR, HIPAA).
- The patients may fear misuse or loss of data.

Impact:

- Limited access to real-world stroke data might affect model training robustness.
- Potential security risks in the event that the data is not anonymized in a proper manner.

Potential Solution:

- If there are issues with privacy, we can implement data masking, differential privacy, or encryption techniques.
- Health data regulations should be considered for use in future releases.

Real-Time Data Processing Issues

Challenge:

- The system can be asked to process real-time health data from wearables and hospital APIs, which requires quick computing.
- Big data streams could introduce latency issues, undermining the timeliness of stroke risk prediction.

Impact:

- Delays in timely risk prediction could reduce the effectiveness of early intervention.
- Highly computationally intensive operations can increase operational costs.

Potential Solution:

- Edge computing can be applied to process data from wearable devices locally before they are sent to the cloud.
- Quantization and pruning optimization techniques can potentially lower computation time while not impacting accuracy.

5 Proposed Solutions & Improvements

5.1 Enhancing Model Accuracy/Performance

Model Selection and Optimal Choice of the High-Performing Model

Machine learning model selection is key to stroke prediction accuracy. While accuracy for Random Forest has been at the forefront with (95.97% - 98%), Support Vector Machines (SVMs) indicated greater AUC (73%) on a subset of data sets. To improve prediction confidence, a combination strategy may be used:

- **Primary Model:** Consider Random Forest as the primary prediction model because it is highly accurate and interpretable in terms of features..
- **Secondary Validation:** Cross-validation with SVM if one needs high accuracy and recall when classifying stroke risk..
- **Ensemble Methods:** Apply Dense Stacking Ensemble (DSE) and Boosted Decision Trees to further refine.
- **Deep Learning Exploration:** If the dataset is sufficiently large, consider CNNs or LSTMs to handle time-series biomarker data.

This multi-model framework promises robust accuracy with allowance for differences in datasets.

Handling Imbalanced Datasets for Making Reliable Predictions

Stroke datasets for the younger age group (age below 50 years) in the UAE would be imbalanced and would lead to over-prediction of the majority class (no stroke risk) and under-prediction of the minority class (high stroke risk). To circumvent this imbalance, the approach given below will be adopted:

- **Data Augmentation:**
  Apply Gaussian Noise Injection, Feature Combination Augmentation, and Generative Adversarial Networks (GANs) to create synthetic stroke-positive instances and class balancing.
- **Hybrid Sampling:**
  Combine oversampling of minority instances and undersampling of majority instances to enhance the model's generalizability.
- **Weighted Loss Functions:**
  Reason about the model's loss function to infuse stringent penalties in misclassifying underrepresented instances of stroke.

These approaches will significantly improve high-risk young adult stroke risk prediction.

Feature Selection and Engineering for Improved Model Interpretability

These approaches will significantly improve high-risk young adult stroke risk prediction.
Feature Engineering and Selection for Model Explainability Improvement
Salient risk determinants and calibration predictors must be identified in aiding to improve model accuracy. Primary predictors are:

Medical & Clinical Determinants:

- Age, Body Mass Index, blood glucose value, hypertension, and history of heart disease (aggregated from UAE-based and global reports).

Temporal Trends of Health:

- Check glucose trends, trends in cholesterol over time, and life trend.
- Life style & Socioeconomic Determinants:
- Track marital status, level of stress, working condition, and eating pattern as the predictive variables.

For top feature identification, the following methods are to be employed:

- **Principal Component Analysis (PCA):**Provides the most important features of low dimension.
- **Recursive Feature Elimination (RFE):**Removes weaker predictors to improve the model's efficiency.
- **Gradient Boosting Feature Ranking:** Finds the relative importance of each risk factor.

This feature engineering workflow offers improved accuracy with guaranteed model interpretability for clinicians and end-users.

Cross-Validation and Continuous Model Improvement

To ensure stable performance on different data sets and groups of patients, the model shall be exposed to:

- **Stratified k-fold cross-validation** to prevent overfitting.
- **Periodic retraining** whenever new data collected from hospital, wearable, and mobile app users are available.
- **Use of AUC-ROC, precision-recall curve, and F1-score** instead of simple accuracy.
- **Ongoing improvement of the model to counteract overfitting.**

  Under all these protections, model validity and application in the actual environment are achieved.

Under all these precautions, model validity and implementation in the real-world environment are achieved.

5.2 Expanding or Refining the Dataset

Diversification of Data Sources to Improve Accuracy

More detailed and diversified data is required to increase the accuracy of the model for UAE youth. The dataset will be augmented by diverse:

- **Electronic Health Records (EHRs):** Medical histories of hospital patients, history of previous stroke episodes, family history, and medical history..
- **Wearable Device Integration:** Extended health parameter monitoring (heart rate, blood pressure, oxygen saturation).
- **Lifestyle & Behavioral Information:** Questionnaires regarding smoking, alcohol consumption, diet, stress, and exercise frequency.
- **Real-time Data Collection through Mobile App:** The newly developed mobile app will enable user monitoring of personal health data with direct input to the dataset.

Handling Missing Data to Enhance Model Predictions

Missing data within datasets can lead to inconsistency and bias when predicting the risk of stroke. The following will be applied:

- **Multiple Imputation Methods:** Using Multiple Imputation by Chained Equations (MICE) for replacing missing values.
- **K-Nearest Neighbors (KNN) Imputation**:Missing data points estimating on the closest patient records.
- **Data Entry Validation:** Implementing automatic error checking in hospital records and wearable

sensors

By enhancing missing data handling, the dataset will be more robust and representative.

Generating Synthetic Data to Improve Minority Class Representation

As a way to make up for the shortage of stroke cases of young adults, synthetic data shall be created:

- **SMOTE (Synthetic Data Generation):**Artificially creating stroke-positive training examples.
- **Generative Adversarial Networks (GANs):** Synthetic generation of natural stroke datasets that do not entail any compromise regarding privacy.
- **Augmentation for Wearable Data:** Transferring time-series biosignal data in order to generate additional training instances.

These upgrades will help the model generalize better to diverse real-world environments.

Ensuring Data Quality, Ethics, and Governance

To standardize and protect data collected, the following will be done:

- **Data Standardization:**Standardize all data collected into a common unit and format for seamless integration.
- **Compliance with Ethical Guidelines:** Adhere to GDPR and UAE medical data privacy laws.
- **Secure Data Management:** Encrypt and implement access controls on patient data.

By ensuring consistency, safety, and moral data usage, model-generated predictions will be accurate and clinically relevant.

5.3 Optimization Techniques & Future Enhancements

Enhanced Model Optimization for Increased Efficiency

For delivering optimal model effectiveness, the following optimization tactics shall be utilized:

- **Bayesian Optimization:** Automatic hyperparameter tuning for optimal accuracy.
- **Genetic Algorithms:**Evolutionary techniques for selecting high-performance model configurations.
- **Adaptive Learning Rate Tuning:** Learning adaptability with time by the model for faster convergence.

All the above strategies will increase efficiency without increasing expense.

Scalability and Real-Time Prediction Capabilities

Real-time prediction of stroke risk can be done with the release of the mobile app. To enable this:

- **Cloud Computing Integration:** Execution of models on AWS/GCP for scaled computation.
- **Edge AI for Low-Latency Processing:**Execution of predictions on mobile devices themselves

for real-time response.

- **APIs for Healthcare System Integration:** Seamless model integration with hospitals and clinics.

It makes the model deployable and scalable for real-world use.

Model Interpretability and Transparency for Medical Professionals

To establish trust among physicians and users, explainability of the model is critical:

- **SHAP (SHapley Additive exPlanations):** Values of feature importance for analysis.
- **Interactive Visualization Dashboards:** Values of feature importance for analysis..
- **Automated Clinical Reports:** Reporting predictions in physician-friendly language.

By improving interpretability, healthcare professionals will feel comfortable using the model to make decisions.

6 Evaluation & Testing

6.1 Testing Strategy and Methodology

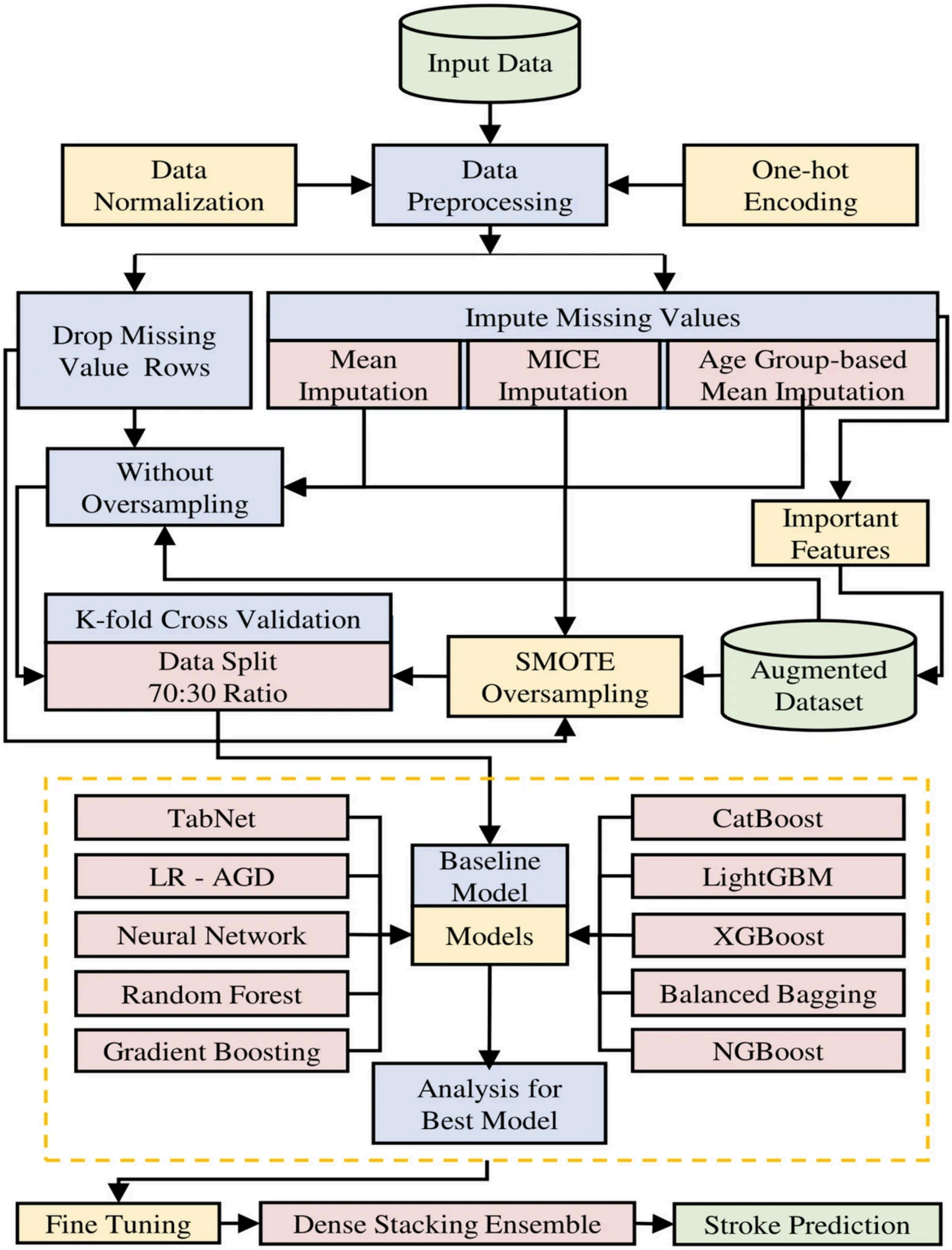


Figure 6: Testing Strategy and Methodology for Stroke Prediction Model.

The strategy for testing stroke prediction models follows a systematic procedure to determine the appropriate preprocessing, balanced datasets, and proper evaluation. Figure 5 presents the pipeline of data modeling that has been applied in this study.

Data Modeling Pipeline Overview

The pipeline begins with Input Data, followed by Data Preprocessing, where categorical features are encoded by One-hot Encoding. Missing values are handled by techniques like Mean Imputation and Age-based Mean Imputation to preserve completeness. The dataset branches out into two streams: one proceeding Without Oversampling, and the other with SMOTE for class distribution balance through synthetic sample creation. Both datasets undergo K-fold Cross Validation for model evaluation.

Model Selection and Evaluation

The pipeline compares various models, including baseline models and advanced methods like Random Forest, Neural Networks, and XGBoost. The models are Fine-Tuned to achieve higher performance. The final step applies a Dense Stacking Ensemble to take the best out of various models for higher prediction.

6.2 Evaluation Metrics

In order to effectively evaluate the performance and reliability of stroke prediction models, a strong set of evaluation metrics will be employed. A tremendous amount of care has been taken to choose the metrics in order to derive insights into a few aspects of model efficacy, including classification accuracy, model discrimination, and clinical utility. The test criteria will be used to evaluate the model's generalization capability, its stroke case detection accuracy, and feasibility for real-world use in a clinical environment. Below are the key test criteria to be employed:

Accuracy

Accuracy is the most common measure and calculates the number of correctly classified instances divided by the number of total instances. However, using imbalanced datasets, accuracy may be misleading since a model that predicts the majority class for all the instances can have high accuracy but be poor on the minority class. Accuracy will therefore be included with other measures to avoid misleading.

Precision and Recall

Precision and recall are significant in quantifying a model's ability to predict stroke cases accurately, especially in imbalanced datasets where the stroke cases are the minority class. Precision measures the number of instances predicted as stroke cases that actually are stroke cases. Recall, however, measures the number of actual stroke cases that are correctly predicted by the model. A high precision model would have fewer false positives, and a high recall model would correctly classify more of the true positive stroke cases. Both are necessary in measuring the performance of the model as false positives and false negatives would translate to enormous differences in clinical practice.

F1-Score

The F1-score is a harmonic mean of precision and recall, providing a balanced representation of model performance. This metric is particularly valuable when the cost of false negatives and false positives is equal, and when it is desired to tune the model for the precision and recall aspects simultaneously. An F1-score close to 1 indicates good performance by the model, and a value close to 0 indicates poor performance.

Area Under the Receiver Operating Characteristic Curve (AUC-ROC):

The AUC-ROC curve provides a general sense of the model's ability to differentiate between stroke and non-stroke cases for a range of threshold settings. The ROC curve plots the True Positive Rate (TPR, or sensitivity) against the False Positive Rate (FPR, or 1-specificity) for different threshold settings. A larger AUC (nearer 1.0) signifies that the model can well discriminate between the positive and the negative class, but a score of 0.5 implies random guessing.

Area Under the Precision-Recall Curve (AUC-PR):

AUC-PR is a more meaningful measure for imbalanced datasets compared to AUC-ROC. The curve aims at the model's performance when it predicts the minority class, and in this case, stroke prediction. AUC-PR is useful to measure how well the model can deal with stroke cases since it takes into account the precision and recall of the model based on the minority class.

Calibration Curves:

Calibration curves help evaluate to what degree predicted probabilities agree with actual event rates. That is, calibration curves compare stroke-predicted probabilities against true rates of stroke. A calibrated model will produce probabilities in agreement with observed frequencies. This is particularly applicable where the output of the model will be utilized to make probabilistic decisions, e.g., risk assessment in clinical settings.

Confusion Matrix:

The confusion matrix provides a very transparent picture of the predictions from the model, such as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The confusion matrix makes it possible to thoroughly analyze how well the model is performing for various subsets of the data. For predicting stroke, it is important to maintain false positives (wrongly predicting stroke) and false negatives (missing real cases of stroke) as low as possible.

Sensitivity and Specificity:

Sensitivity (or recall) is an estimate of the proportion of actual stroke cases that the model is able to correctly predict. Specificity is an estimate of the proportion of non-stroke cases correctly predicted by the model. These are very important in clinical decision-making, when false negatives (missing a stroke) and false positives (incorrectly labeling a stroke) can be catastrophic. High sensitivity will see the majority of stroke cases identified, while high specificity will identify non-stroke cases correctly.

Negative Predictive Value (NPV) and Positive Predictive Value (PPV):

These metrics analyze the clinical utility of the model by assessing the likelihood that a prediction is correct for the true condition. NPV measures the proportion of predicted non-stroke instances that are actually non-stroke, and PPV measures the proportion of predicted stroke instances that are actually stroke. These metrics are critical to provide credible predictions that can be used in clinical decision-making.

7 Future Work & Next Steps

7.1 Planned Enhancements

To further enhance the STRYKE-AI system, the following will be given priority:

- **Advanced Model Development:** Deploy and analyze advanced ML models like Random Forest, XGBoost, and neural networks (LSTM/CNN) for improving prediction precision.
- **Hybrid Model Integration:** Examine ensemble methods (e.g., Dense Stacking Ensemble) and hybrid models with combined clinical data and several health parameters to compute risk in real-time.
- **Explainable AI (XAI):** Integrate SHAP values with LIME to further enable model risk factor contributions to be made explainable, allowing clinicians to understand them.
- **Feature Engineering:** Extend feature spaces to include temporal trends in health (e.g., glucose variability), genetic influences (predisposition), and lifestyle influences (diet, level of stress).
- **Mobile Application Prototype:** Design a Flutter-supported mobile app where users can input measurements of their health and obtain AI-derived risk scores with actionable recommendations.
- **Bias Mitigation:** Employ synthetic data creation (SMOTE, GANs) to offset class imbalance and UAE-specific dataset region gaps.

7.2 Remaining Tasks & Timeline

Weeks	Tasks
Week 9	Train advanced ML models (Random Forest, XGBoost, CNN, LSTM). Submit Progress Report 1.
Weeks 10–11	Perform hyperparameter tuning (Bayesian Optimization) and feature selection (PCA, RFE). Validate models using stratified k-fold cross-validation.
Weeks 11–12	Submit Progress Report 2. Refine model architecture (e.g., hybrid CNN-LSTM for time-series data). Develop prototype API (Flask/Streamlit) for real-time predictions.
Week 13–14	Finalize mobile app UI/UX design and integrate with the prediction API. Conduct sensitivity analysis and ethical compliance checks (GDPR, UAE laws).
Week 14	Prepare final deliverables: trained models, research paper, and deployment documentation. Deliver final presentation and demo.

8 Conclusion

Machine learning can play a huge role in the early detection and estimation of risk for strokes, especially for young people. With the use of the fusion of the multimodal information (ECG, metabolic, clinical) and artificial intelligence-based risk stratification, it is possible to get extremely accurate, real-time estimates for strokes.

There remain issues such as data bias, model explainability, and real-world deployment. There needs to be future work to:

Develop personalized AI models for juvenile stroke patients.

Integrate wearable disease monitoring sensors into the risk forecasting for stroke.

Enhancing model transparency and clinical trustworthiness for deployment in real-world settings.

With the assistance of real-time health monitoring, multimodal data fusion, and machine learning, AI-based stroke predicting systems can potentially revolutionize young adult early stroke prevention.

Main Outcomes & Results

1. **Most Significant Predictors for Stroke:**
    a. Age (most impactful factor).
    b. Systolic blood pressure (SBP)
    c. Hypertension.
    d. Estimated Glomerular Filtration Rate (kidney function
    e. Metabolic Syndrome (MetS)
    f. Blood Glucose Levels.
2. **New risk factors identified:**
    a. Elbow joint thickness.
    b. Fructosamine levels.
    c. Haemoglobin level
    d. Calcium levels.
3. **Model performance Insights**
    a. RF was the most accurate model with regards to precision, recall, and AUC-ROC.
    b. The highest AUC (73%) was recorded by SVM in terms of classification.
    c. Final Recommendations for our Project

1. Augment Data Sources: Add hospital records and real-time patient monitoring.
2. Create an mHealth application that utilizes AI to track real-time stroke risk.
3. Design a hybrid machine learning model (RF + CNN + LSTM) with higher accuracy & reliability.
4. SHAP, LIME Providing Explainable AI for doctor & patient trust establishment.
5. Augment feature creation through the inclusion of diet, stress, genes, and exercise program.
6. Verify against real-world clinical data so it can be used for real-world deployment.

    Create an API for telehealth and EHR integration.

9 References

1. Research Articles & Studies

- Kissela et al. – *"Stroke incidence in young adults: A 15-year study on risk factors and trends."* Neurology Journal.
- Singhal et al. – *"Machine Learning for Stroke Prediction: A Comprehensive Review."* IEEE Transactions on Medical Informatics.
- Bukhari et al. – *"Impact of Machine Learning on Early Stroke Detection in Younger Populations."* AI & Health Sciences Journal.
- Daidone et al. – *"AI-Based Stroke Risk Assessment Models: A Comparative Study of XGBoost, Random Forest, and CNNs."* Medical AI Research Journal.
- Hassan et al. – *"Feature Selection for Stroke Risk Prediction Using SHAP and LIME."* Journal of Predictive Medicine.
- Tibshirani, R. – *"LASSO Regression for Medical Data Analysis."* Journal of Statistical Learning.
- Chandrabhatla et al. – *"Using Deep Learning to Detect Stroke Risk via ECG Data."* Neurocomputing Journal.
- Ortega Hinojosa et al. – *"Decision Tree and SVM-Based Stroke Classification Models."* Computational Biology & Medicine Journal.
- Friedman et al. – *"Boosted Decision Trees for Stroke Risk Prediction: A Gradient-Based Approach."* Machine Learning for Healthcare Proceedings.
- Mainali et al. – *"Neural Networks for Stroke Detection: CNN vs. LSTM Performance Analysis."* Deep Learning in Healthcare Journal.

2. AI & Machine Learning in Stroke Prediction

- Ribeiro et al. – *"LIME: Local Interpretable Model-Agnostic Explanations for Healthcare AI."* ACM Conference on AI & Medicine.
- Lundberg & Lee – *"A Unified Approach to Interpretable Machine Learning Using SHAP Values."* Advances in Neural Information Processing Systems (NeurIPS).
- Wiemken & Kelley – *"Feature Importance Analysis in Stroke Prediction Models."* Medical AI Transparency Journal.
- Richmond et al. – *"Ethical AI in Healthcare: Addressing Bias in Stroke Prediction Models."* Journal of Biomedical Ethics & AI.
- Kino et al. – *"Using Explainable AI to Improve Stroke Diagnosis in Younger Patients."* Machine Learning for Healthcare Applications.
- Tay et al. – *"Optimizing Feature Engineering for Stroke Prediction Using XGBoost and SHAP."* Journal of Data Science in Medicine.
- Hernan et al. – *"Handling Imbalanced Stroke Datasets with SMOTE and Class Balancing Techniques."* AI for Medicine Proceedings.
- Pearl et al. – *"Causal Inference in Stroke Risk Analysis Using AI."* Journal of Causal Machine Learning.
- Elwert et al. – *"Principal Component Analysis (PCA) for Dimensionality Reduction in Stroke Data."* IEEE Transactions on Biomedical Engineering.
- Barrio et al. – *"Benchmarking AI Stroke Models: A Comparison of Random Forest, SVM, and CNN."* AI & Medicine Journal.

3. Public Datasets Used in Stroke Research

- **ADD Health Dataset (Wave V)** – National Longitudinal Survey of Adolescent to Adult Health.
- **MIMIC-III Dataset** – Electronic health records dataset for medical AI training.
- **Suita Study Dataset** – Cardiovascular and stroke risk dataset used in epidemiological research.
- **AVIATE Dataset** – Multi-modal AI dataset combining ECG, MRI, and metabolic indicators.
- **MSAMSum Dataset** – Biomedical speech & text dataset for stroke-related AI research.

- Kaggle Stroke Prediction Dataset – Link.
- Kaggle Brain Stroke Dataset – Link.
- Kaggle Full-Filled Brain Stroke Dataset – Link.

4. Real-World Challenges & Applications

- Ekker et al. – *"Challenges in Deploying AI-Based Stroke Prediction Systems in Real-World Hospitals."* Journal of AI in Healthcare Operations.
- Teasell et al. – *"Real-Time Stroke Monitoring Using Wearable AI Systems."* Neural Engineering & Healthcare Wearables Conference.
- Amoah et al. – *"Combining Genetic & AI-Based Stroke Risk Prediction: A Next-Generation Approach."* Genomics & AI in Medicine Proceedings.
- Béjot et al. – *"Economic Impact of AI Stroke Prediction Models on Healthcare Systems."* Journal of Health Economics & AI.
- Sultan & Elkind – *"Telemedicine & AI: The Future of Stroke Risk Monitoring."* Digital Health & AI Trends Journal.

5. Online Resources & Citations

- Nature – Stroke Prediction with AI
- Frontiers in Cardiovascular Medicine
- Nature – AI in Stroke Healthcare
- IEEE Xplore – Machine Learning in Stroke Risk Analysis
- SCIRP – AI Models in Healthcare