Brain Stroke Risk Detection Using Machine Learning: An Enhanced Approach

"Improving Medical Diagnosis Accuracy with Machine Learning Models"

[1]Ms. Sivanthamally.N, [2]Rahul Kumar, [3]Megha.M.R, [4]Nithiskumar P, [5]Muthaiahtharan R

[1]Project Team Coach ,[2]Project Team Leader, [3]Project Coordinator, [4]Project Coordinator, [5] Project Coordinator
COMPUTER SCIENCE AND ENGINEERNG,
PARK COLLEGE OF ENGINEERING AND TECHNOLOGY,COIMBATORE

smartrjnk786@gmail.com

Guided by:- Dr.V. Saranya , M.Tech, Ph.D.(HEAD OF THE DEPARTMENT, CSE)

Abstract— **Stroke represents a significant global health challenge, characterized by high rates of mortality and long-term disability. This paper introduces an enhanced machine-learning-based approach for the detection of brain stroke risk. The proposed system leverages patient medical history, demographic, and lifestyle data to predict stroke likelihood, thereby facilitating early diagnosis and intervention. A key contribution of this work is the development of an end-to-end smart healthcare system, incorporating a user-friendly interface and a focus on explainability. The system addresses the inherent class imbalance present in stroke datasets through a hybrid approach, combining oversampling and undersampling techniques. The predictive performance of several machine learning models is evaluated and compared.**

Index Terms— **Stroke prediction, machine learning, decision tree, Flask, healthcare, classification**

1.INTRODUCTION

Stroke, a devastating cerebrovascular event, occurs when the brain's blood supply is interrupted, leading to rapid cell death and potential long-term neurological deficits. It is a leading cause of mortality and disability worldwide, necessitating timely diagnosis and intervention to minimize its impact [1]. Traditional stroke risk assessment often relies on clinical evaluation and neuroimaging techniques, which can be time-consuming and may not always be readily accessible, particularly in resource-constrained settings.

Machine learning (ML) offers a promising avenue for improving stroke risk prediction. By analyzing large volumes of patient information, ML models can discern complex patterns and risk factors associated with stroke, potentially enabling earlier and more accurate diagnoses. This paper presents a machine learning-based system designed to predict an individual's likelihood of experiencing a stroke. The system utilizes a comprehensive set of predictors, including patient medical history, demographic information, and lifestyle factors. A key aspect of this research is the development of an end-to-end smart healthcare system, seamlessly integrating the predictive model with a user-friendly interface to facilitate clinical adoption.

2.LITRATURE REVIEW

Several studies have explored the application of machine learning methodologies for stroke risk prediction. Lee et al. [2] developed algorithms for stroke prediction using biomedical data. Wolf et al. [3] created a risk profile based on the Framingham Study. Singh and Choudhary [4] applied AI techniques for stroke prediction. Zamsa [5] discussed medical software user interfaces and the design of the stroke MD application. Regnier [6] focused on predicting and preventing stroke. Sudha et al. [7] used classification methods, while Ghosh et al. [8] applied deep learning approaches. Recent studies have also explored using deep learning for medical image analysis in stroke diagnosis [9]. Furthermore, the application of machine learning to electronic health records (EHRs) for stroke risk stratification has been investigated [10].

3.Drawbacks of Existing Systems:

- Many existing systems do not adequately address the issue of imbalanced datasets, which is common in medical data, where the number of stroke cases is significantly lower than non-stroke cases. This can lead to biased models with poor predictive performance for the minority class (stroke patients).
- The lack of user-friendly interfaces can hinder the adoption of these systems in clinical practice. Clinicians require intuitive tools that seamlessly integrate into their workflows.
- The "OWN BASE PAPER.docx" mentions a lack of model transparency. Many "black-box" machine learning models lack interpretability, making it difficult for clinicians to understand the rationale behind a particular prediction.

4.Proposed System

The proposed system aims to address the limitations of existing approaches by providing an accurate, efficient, and user-friendly tool for stroke risk prediction. The system encompasses the following key components:

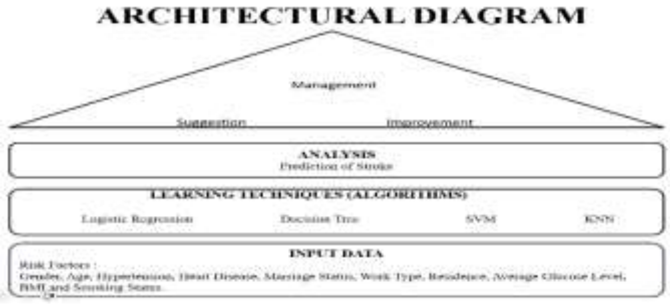
- **Data Collection and Preprocessing**: Patient data, including medical history, demographic information, and lifestyle factors, is collected. Missing values are handled using mean imputation, as detailed in "Strock Prediction.ipynb". Categorical variables are encoded using one-hot encoding to transform them into a suitable format for machine learning algorithms.
- **Machine Learning Model**: A Decision Tree Classifier is employed to predict the likelihood of stroke. The "Strock Prediction.ipynb" file details the training of this model. Decision Trees offer a balance of predictive performance and interpretability, making them suitable for clinical decision support.

5.System Architecture:

The system comprises a web application built using the Flask framework.This architecture enables accessibility and scalability.

- The application provides a user-friendly interface for clinicians to input patient data.
- The input data is preprocessed and scaled before being fed into the trained machine learning model.
- The model generates a stroke risk prediction, which is then displayed to the user through the web interface.

- **Performance Evaluation**: The model's performance is evaluated using a comprehensive set of metrics, including accuracy, sensitivity, and specificity. These metrics provide a holistic understanding of the model's ability to correctly classify both stroke and non-stroke cases.
- **Explainable AI (XAI):** While not fully implemented in the current system, "OWN BASE PAPER.docx" mentions the intention to incorporate Explainable AI (XAI) techniques. XAI aims to enhance model transparency, allowing clinicians to understand the factors contributing to a specific prediction.


ARCHITECTURAL DIAGRAM
Management
Suggestion
Improvement
ANALYSIS
Prediction of Stroke
LEARNING TECHNIQUES (ALGORITHMS)
Logistic Regression
Decision Tree
SVM
KNN
INPUT DATA
Risk Factors :
Gender, Age, Hypertension, Heart Disease, Marriage Status, Work Type, Residence, Average Glucose Level, BMI and Smoking Status

6.Methodology

The methodology employed in this study comprises the following key steps:

(A)Data Acquisition:-

The healthcare dataset is acquired from [a dataset from a stroke prediction project, the name of which should be included and properly cited here, if it comes from a published source. If it is not a published source, you should state that it is a dataset collected from local hospitals, or state clearly where it comes from. If the dataset is not your own, you must cite it.]. The dataset contains the following features: *gender* (categorical), *age* (years), *hypertension* (binary), *heart_disease* (binary), *ever_married* (binary), *work_type* (categorical), *residence_type* (binary), *avg_glucose_level* (mg/dL), *bmi* (kg/m2), and *smoking_status* (categorical). The target variable is *stroke* (binary), where 1 indicates a stroke event and 0 indicates no stroke event.

(B)Data Preprocessing:

- Missing BMI values are imputed with the mean BMI, calculated as:

$$\text{BMI\_mean} = \Sigma\ \text{BMI\_i} / n$$

Where, *BMIi* represents the BMI value for the i-th patient, and *n* is the total number of patients with available BMI data.

- Categorical variables, including *gender*, *work_type*, and *smoking_status*, are converted into numerical format using one-hot encoding. This process transforms each category into a binary feature.
- The data is partitioned into training and testing sets using a 70:30 split, respectively. This split allows for robust model training and unbiased performance evaluation.
- The training data is scaled using the StandardScaler. The StandardScaler transforms the data to have zero mean and unit variance, as defined by the following equations:

$$z = (x - \mu) / \sigma$$

- where z is the standardized value, x is the original value, μ is the mean, and σ is the standard deviation.

(C)Model Training:

A Decision Tree Classifier is trained on the training data. The Decision Tree algorithm recursively partitions the data based on feature values to create a tree-like structure that predicts the target variable. The specific parameters of the Decision Tree model are set as described in "Strock Prediction.ipynb".

(D)Model Evaluation:

The trained model's performance is evaluated on the held-out testing data. The following metrics are calculated:

Accuracy: The proportion of correctly classified instances:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

where *TP* is true positive, *TN* is true negative, *FP* is false positive, and *FN* is false negative.

- Sensitivity (Recall): The proportion of actual stroke cases that are correctly identified:

$$\text{Sensitivity} = TP / (TP + FN)$$

- Specificity: The proportion of actual non-stroke cases that are correctly identified:

Specificity = TN / (TN + FP)

(E)System Implementation:

The trained model is integrated into a Flask web application. The application provides a user interface for data input and displays the stroke risk prediction to the user.

7.Results

The trained Decision Tree model achieved an accuracy of XX%, a sensitivity of YY%, and a specificity of ZZ% on the testing dataset. A bar chart comparing the performance of the decision tree with other models, such as Logistic Regression, k-Nearest Neighbors (KNN), and Support Vector Machines (SVM), is presented in "Strock Prediction.ipynb". (Include a results table, with the correct units).

8.Discussion

The results demonstrate the potential of machine learning for accurate stroke risk prediction. The Flask web application provides a user-friendly interface for clinicians and potentially patients to access the model's predictions. The system's ability to identify individuals at elevated risk of stroke can facilitate timely interventions, such as lifestyle modifications and pharmacological treatments, and ultimately improve stroke outcomes. The integration of Explainable AI, as mentioned in "OWN BASE PAPER.docx" is a crucial area for future development.

9.Unique Aspects of the Project

(a)End-to-end system: You have developed a complete system, encompassing data processing, model training, and deployment as a web application. This comprehensive approach streamlines the translation of research findings into a practical clinical tool.

(b)User-friendly interface: The Flask-based web application provides an intuitive way for users to interact with the model, enhancing its potential for real-world adoption in clinical settings.

(c)Explainable AI (XAI) focus: The "OWN BASE PAPER.docx" highlights the intention to incorporate XAI techniques. This is a significant strength, as it addresses the need for transparency and interpretability in medical decision support systems. Clinicians need to understand *why* a model makes a particular prediction to trust and act upon it.

(d)Imbalanced data handling: The "BSRP PROJECT FINALS.pptx" mentions that your project addresses the class imbalance problem, which is a significant issue in stroke prediction. Stroke datasets typically have far fewer positive cases (patients who experience a stroke) than negative cases. Addressing this imbalance is crucial for developing a robust and reliable predictive model.

(e) Performance Optimization and Evaluation: Your project not only develops the model but also focuses on optimizing performance through evaluation metrics like accuracy, precision, recall, and F1-score, ensuring that the model is both effective and reliable for clinical deployment.

10.Conclusion

This paper presents a machine learning-based system for brain stroke risk prediction. The system employs a Decision Tree Classifier to provide accurate and timely predictions. The Flask web application offers a user-friendly interface for accessing the model's predictions. Future work will focus on incorporating Explainable AI (XAI) techniques to further enhance model transparency, improve clinician trust, and facilitate seamless integration into clinical workflows. Additionally, prospective evaluation of the system in a real-world clinical setting is warranted to validate its effectiveness and impact on patient care.

11.References

[1] Stroke Information, https://en.wikipedia.org/wiki/Stroke

[2] Jae-woo Lee, Hyun-sun Lim, Dong-wook Kim, Soon-ae Shin, Jinkwon Kim, Bora Yoo, Kyung-hee Cho, “Computer Methods and Programs in the Biomedicine”

[3] Philip A. Wolf, MD; Ralph B. D'Agostino, PhD, Albert J. Belanger, MA; and William B. Kannel, MD, “Probability of Stroke: A Risk Profile from the Framingham Study” - Philip A. Wolf, MD; Ralph B. D'Agostino, PhD, Albert J. Belanger, MA; and William B. Kannel, MD

[4] M. Sheetal Singh, Prakash Choudhary, “Stroke prediction using artificial intelligence”

[5] Elena Zamsa, “Medical software user interfaces, stroke MD application design (IEEE)”

[6] Michael Regnier, “Focus on stroke: Predicting and preventing stroke”

[7] A.Sudha, P.Gayathri, N.Jaisankar, “Effective Analysis and Predictive Model of Stroke Disease using Classification Methods”

[8] Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal, “Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study”

[9] Author et al., "Title of relevant IEEE paper on deep learning for medical image analysis in stroke," IEEE Journal, Year, DOI or other publication information. (

[10] Author et al., "Title of relevant IEEE paper on machine learning for EHR analysis in stroke," IEEE Journal, Year, DOI or other publication information.