1. INTRODUCTION

In India, brain stroke is the country's fourth-leading cause of death. According to a survey of World Health Organization (WHO), accounting for 11% of all fatalities. The brain is the most crucial part of our human body. The cerebrum, brainstem, and cerebellum make up the brain. The brain's weight is approximately three pounds, which is the main organ for the human senses, thinking ability, creativity, and balance. The brain is around 60% fat and combines water, protein, carbohydrates, and salts. The skull bones of the head enclose and shield the brain from external harm.

If the person is older than 55, the risk of a brain stroke is higher than that of an average person. People of all other races and ethnicities have a lower risk of stroke than African Americans and Hispanics. By 2050, low and moderate-income nations will account for more than 80% of the estimated 15 million additional strokes worldwide .Additionally, the expense of stroke hospitalization is rising, necessitating the development of new technology to aid in clinical diagnosis, treatment, clinical event prediction, referral of viable therapeutic approaches, and rehabilitation programs.

2. LITERATURE REVIEW

Md. Shafiul Azam et al. [1] proposed various pre-processing techniques to balance the dataset. By using three machine learning algorithms, Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF), they predict whether a person has a chance of getting a brain stroke or not. The main aim of the thesis is to predict the risk of brain stroke and to analyze the performance of these algorithms.

This paper identifies the symptoms of brain stroke and predicts whether the person is suffering from ischemic or hemorrhage stroke. Also, a comparison between men and women to know who has a high chance of risk. Nur Sakinah et al. [4] considered different CT scan images of both ischemic stroke and hemorrhage stroke images as a dataset performed pre-processing techniques to improve image quality. Also, by applying further deep learning and machine learning algorithms like Random Forest (RT), Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbor, and Multi-Layer Perceptron (MLP-NN) to detect whether the patient is suffering from ischemic stroke or hemorrhage stroke.

3 PROBLEM DEFINITION

Using Machine learning algorithms be used to accurately predict the likelihood of an individual experiencing a stroke based on their demographic information, lifestyle habits, medical history, and other risk factors.

The existing system follows different classification algorithms like Logistic Regression, Decision Tree Classification, Random Forest classification, K-Nearest Neighbors, Support Vector Machine and Naïve Bayes. By using these algorithms, it is possible to determine whether a person is at risk for a brain stroke. A decision tree is a graphical representation of a decision-making process that uses a tree-like model of decisions and their possible consequences.

It is a type of supervised learning algorithm used in machine learning and data mining. The decision tree starts with a single node, called the root, which represents the entire dataset ,by using this algorithm we predict the brain stroke. Due to lack of awareness of brain stroke, there may be a chance of risks having it, maybe they will consult the doctor for betterment and may use mandatory precautions that should be taken to avoid such strokes and reduce the overall rate of death caused by brain stroke.

4. DATASET

4.1 DATA SET

gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status
Male	67	0	1	Yes	Private	Urban	228	36	formerly smoked
Female	61	0	0	Yes	Self-employed	Rural	202	37	never smoked
Male	80	0	1	Yes	Private	Rural	105	32	never smoked
Female	49	0	0	Yes	Private	Urban	171	34	smokes
Female	79	1	0	Yes	Self-employed	Rural	174	24	never smoked
Male	81	0	0	Yes	Private	Urban	186	29	formerly smoked
Male	74	1	1	Yes	Private	Rural	70	27	never smoked
Female	69	0	0	No	Private	Urban	94	22	never smoked
Female	59	0	0	Yes	Private	Rural	76	30	Unknown
Female	78	0	0	Yes	Private	Urban	58	24	Unknown
Female	81	1	0	Yes	Private	Rural	80	29	never smoked
Female	61	0	1	Yes	Govt_job	Rural	120	36	smokes
Female	54	0	0	Yes	Private	Urban	104	27	smokes
Male	78	0	1	Yes	Private	Urban	219	30	Unknown
Female	79	0	1	Yes	Private	Urban	214	28	never smoked
Female	50	1	0	Yes	Self-employed	Rural	167	30	never smoked
Male	64	0	1	Yes	Private	Urban	191	37	smokes
Male	75	1	0	Yes	Private	Urban	221	25	smokes
Female	60	0	0	No	Private	Urban	89	37	never smoked
Male	57	0	1	No	Govt_job	Urban	217	26	Unknown
Female	71	0	0	Yes	Govt_job	Rural	193	22	smokes
Female	52	1	0	Yes	Self-employed	Urban	233	48	never smoked
Female	79	0	0	Yes	Self-employed	Urban	228	26	never smoked
Male	82	0	1	Yes	Private	Rural	208	32	Unknown
Male	71	0	0	Yes	Private	Urban	102	27	formerly smoked
Male	80	0	0	Yes	Self-employed	Rural	104	23	never smoked
Female	65	0	0	Yes	Private	Rural	100	28	formerly smoked
Male	58	0	0	Yes	Private	Rural	189	26	Unknown
Male	69	0	1	Yes	Self-employed	Urban	195	28	smokes

Dataset Description:

- The Txt file dataset is used for predicting brain stroke.
- Size of the dataset: 316.97 KB
- The dataset contains a total of 5110 entries.
- 5110 entries are used.
- The dataset contains 10 fields.
- Training data

: 4088 (80%)

- Testing data

: 1022 (20%)

- Dataset is collected from the reference of:

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

- govt hospital srikaikulam. (20% of data).

5. SYSTEM REQUIREMENTS AND SPECIFICATIONS

5.1 HARDWARE REQUIREMENTS

- **Processor**

: Intel i5 or higher

- **RAM**

: 8 GB or higher

- **Hard Disk**

: 128 GB

5.2 SOFTWARE REQUIREMENTS

- **Operating System**

: Windows 7/8/10/11

- **IDE**

: Jupyter Notebook

- **Technology Used**

: Python 3.9

5.3 TECHNOLOGIES USED:

NumPy:

Numerous people have contributed to the open-source program NumPy. Large, multi dimensional arrays and matrices are supported by the Python programming language's Numpy library, which also provides a vast variety of high-level mathematical operations for use on these arrays. The ancestor of Numpy, Numeric, was created by Jim Holguin with contributions from several other developers. In 2005, Travis Oliphant created Numpy by incorporating features of the competing Num array into Numeric, with extensive modifications.

Pandas:

A software package called panda was created for the Python programming language to manipulate and analyze data. It includes the specific data structures and procedures for working with time series and mathematical tables. Numpy is a free software distributed under the BSD license's three clauses.

Scikit-Learn:

Scikit-learn is perhaps the foremost helpful library for machine learning in python. Scikit-learn contains a lot of efficient tools for machine learning and statistics including classification, regression, clustering, and dimensionality reduction. The components used are Supervised learning algorithms, cross-validation, Unsupervised learning algorithms, various toy datasets, and feature extraction.

OS:

The OS module in Python provides functions for interacting with the operating system. The module provides a portable way of using operating system - dependent functionality. OS comes under Python's standard utility modules. The *os* and *os. path* modules include many functions to interact with the file system.

Seaborn:

Seaborn is a library for visualization of graphical statistical plotting in Python. Seaborn provides many color palettes and defaults beautiful styles to create many statistical plots in Python more attractive.

Matplotlib:

Matplotlib is a cross-platform for data visualization and graphical charting package in Python and its numerical extension NumPy. As a result, it acts as an open-source replacement for MATLAB. The APIs (Application Programming Interfaces) of Matplotlib can also be used to include charts in GUI programmers.

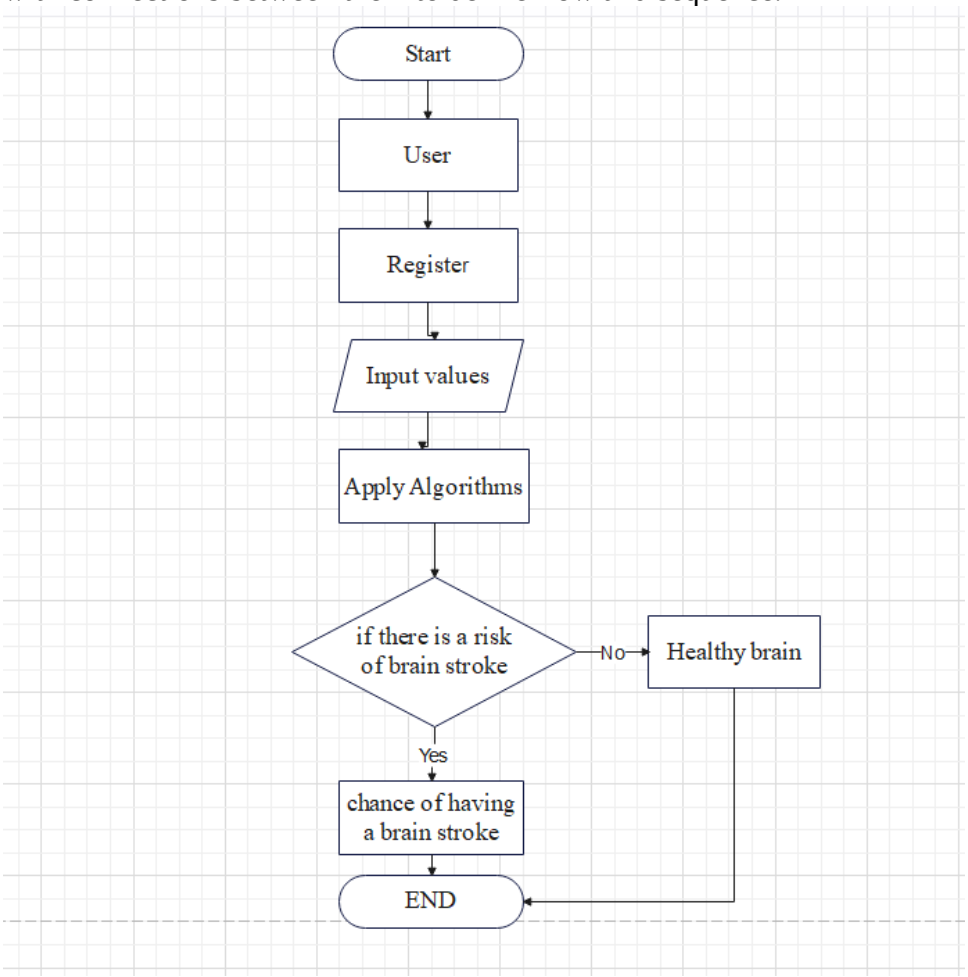
6. SYSTEM DESIGN

6.1 UML DIAGRAMS:

The term “UML” stands for "Unified Modeling Language.”A general-purpose modeling language with standards, UML is used in object-oriented software engineering. The Object Management Group is in charge of creating and managing standards. The objective of UML is to establish itself as a standard language for modeling object-oriented computer programs consisting primarily of a meta-model and the notation in its current version.

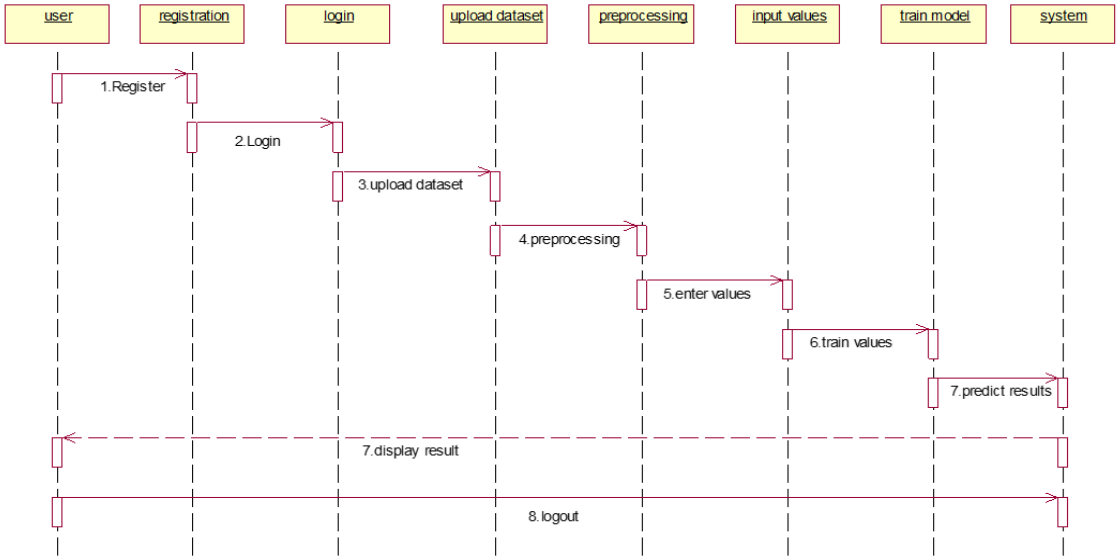
6.1.1 FLOWCHART DIAGRAM:

A flowchart is a diagram that demonstrates how a system or process works. They are widely used in many different fields to examine, organize, enhance and communicate often a complex process in simple and understandable diagrams. Rectangles, ovals, diamonds and many other shapes are used in flowcharts to define the type of step, along with the arrows with connections between them to define flow and sequence.



6.1.2 SEQUENCE DIAGRAM:

A series diagram is a shape of an interplay diagram that suggests how strategies interact with one to any other and in what order. It's referred to as a Message Sequence Chart. Sequence diagrams are also known as occasion diagrams, situations, and timing diagrams.



7. CONCLUSION

Diagnostic applications of machine learning are growing in the medical industry. This thesis aims to predict stroke whether a person has a risk of getting a stroke or not by utilizing a pre-processed dataset. This can be achieved by employing three machine learning techniques, including Decision Tree (DT), Naïve Bayes (NB), and Artificial Neural Networks (ANN). By utilizing a few user-provided inputs and forecasting accuracy, the system assists in the cost effective and efficient prediction of brain stroke. A comparison of each approach is also included. Naïve Bayes Classification, which was selected, performs the best overall, with a 95% accuracy rate. The suggested method is the backbone of the stroke patient healthcare system. This study may be expanded to identify the probability of stroke by gathering a dataset made up of images, such as brain CT scans.

8. REFRENCE

1. Kaggle. “Healthcare stroke Patients in Python” kaggle.com/surajdidwania/healthcare-stroke-patientsinpython/data?fbclid=IwAR21bmwdw1jItWPIMRMEB_CehjYuh5wH6IQOlVvOvOyBp0umH4X1d9Zk7g

2. T. Lumley, R. A. Kronmal, M. Cushman, T. A. Manolio, and S. Goldstein. A stroke prediction score in the elderly: Validation and web-based application. J. Clin. Epidemiol., 55(2):129–136, February 2002.

3. Mujtaba, M. A., Azam, M. S., &Rana, H. K., "Performance evaluation of various data mining classification techniques that correctly classify banking transaction as fraudulent", GUB Journal of Science and Engineering, vol. 4, no. 1, pp. 59-63, 2017.