DEEP LEARNING BASED BRAIN STROKE PREDICTION

[1] O Ramya Teja, [2]B. Rajya Laxmi, [3]P. Hema Sree, [4]J. Madhurika, [5]R. Bhavani
[1]Assistant Professor, [2,3,4,5]UG Students, Dept. of CSE (AI & ML), Malla Reddy Engineering College for Women (Autonomous), Hyderabad, India. E-Mail: ramyateja9@gmail.com

ABSTRACT
One of the main causes of adult humanity and disability is a stroke or a brain attack.It's a medical emergency, therefore obtaining aid as soon as you can is essential. Right quickly seeking medical attention can help avoid problems and brain damage. Predicting illness incidence, prognosis, and assisting physicians in prescribing disease therapy are just a few of the many predictive methodologies that have been widely employed in clinical decision-making. This approach of predicting analytical procedures for stroke was conducted out using a deep learning network on a brain illness dataset. The objective of this model is to build a deep learning application that uses a convolution neural network to recognize brain strokes. Three models have also been created to anticipate the results. The proposed study predicts and categories strokes using a dataset of CT scan (computed tomography) images.

INTRODUCTION
The fifth-leading cause of death worldwide is stroke. Stroke is a non-communicable infection that accounts for 11% of all mortality. It ranks as the fourth leading cause of death in India. The development of medical technology has made it possible to predict the onset of a stroke using machine learning. Machine learning algorithms are helpful in giving accurate analysis and making correct forecasts. The likelihood of a brain stroke is forecasted in this study using machine learning.
According to the key components of the techniques used and the outcomes obtained, Nave Bayes outperformed the other five classification algorithms and obtained a higher accuracy measure. The fact that this model was trained on text data rather than actual brain images is a disadvantage. The implementation of six machine learning classification methods is demonstrated in the study. This study can be expanded to incorporate all the most recent machine learning techniques. A dataset from Kaggle with a variety of physiological variables as its attributes is picked to continue with this task.
Based on an examination of these attributes, the final prognosis is made. The dataset is initially cleaned in order to make it easier for the machine learning model to grasp. At this point, the procedure involves data pre-processing.
The dataset is checked for null values and updated if necessary. Following label encoding, one hot encoding may then be used to convert string values into numbers, if necessary. The dataset is separated into train and test data after data pre-processing.
After then, a model is built utilising the new data and a number of categorization techniques. To find the most precise prediction model, accuracy for each of these approaches is calculated and compared. When the model has been trained and correctly decided, an HTML website and a Flask application are produced.
In the web application, the user enters the values for the forecast. The Flask application connects the web application with the trained model. The research concludes about which algorithm is most suitable for the prediction of stroke after thorough analysis.

LITERATURE SURVEY
The current approaches for predicting strokes based on medical history data primarily use traditional machine learning algorithms. These techniques could include decision trees, random forests, logistic

regression, and support vector machines (SVM).

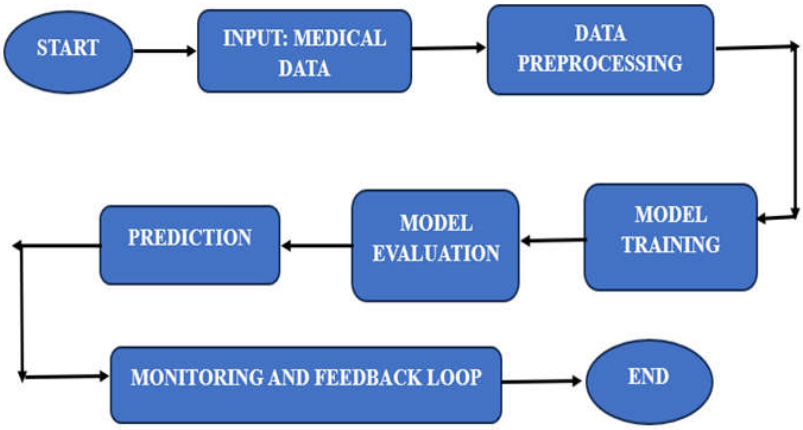
PROPOSED SYSTEM

The proposed system introduces a novel deep neural network model, specifically a convolutional neural network (CNN), for predicting the possibility of future strokes. The CNN model is designed to the most leverage the benefits of deep learning and excel in classification tasks.

SYSTEM ARCHITECTURE

The system architecture consists of the following components:

User Interface (Web Interface or API): Allows users to interact with the system and input their data. Data Collection Component: Gathers relevant data from external sources such as medical records and diagnostic tests. Data Storage Layer:



DATA FLOW DIAGRAM

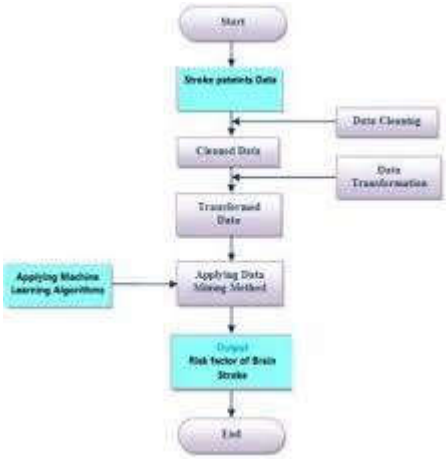


Fig.1. Dataflow Diagram

ADVANTAGES

Improved performance: The deep neural network CNN model has the potential to outperform classical machine learning algorithms in terms of accuracy and predictive capabilities for stroke prediction. Enhanced feature selection: Feature selection algorithms are applied to extract the most relevant features from the medical data, which improves themodel's efficiency and interpretability.

DEFINE THE PROBLEM

The current approaches for predicting strokes based on medical history data primarily use traditional machine learning algorithms. These techniques could include decision trees, random forests, logistic regression, and support vector machines (SVM).

Our project's goal is to effectively forecast strokes based on possibly modifiable risk factors by using machine learning techniques on large current data sets. Following that, it intended to develop an application that would offer each user a warning that was specifically tailored to their level of stroke risk as well as advice on how to change their lifestyle to lower that risk.

DEFINE THE MODULES AND THEIR FUNCTIONALITIES:Modules:

1. Data Collection
2. Data Preprocessing
3. Feature Extraction
4. Model Architecture Design
5. Model Training
6. Model Evaluation
7. Deployment and Integration
8. Monitoring and Update

Functionalities:

Data Collection: Gather relevant data related to strokes, including medical records, patient demographics, lifestyle factors, and diagnostic tests.

Data Preprocessing: Process and clean the collected data by removing inconsistencies, errors, or missing values. Perform tasks such as data normalization, feature scaling, handling missing data, and data augmentation. Feature Extraction: Extract relevant features from the pre- processed data. Use techniques like Principal Component Analysis (PCA) or convolutional layers to reduce dimensionality and capture essential information. Model Architecture Design: Design the deep learning architecture for stroke prediction. Define the type and number of layers (e.g., convolutional, recurrent, dense), activation functions, regularization techniques (e.g., dropout), and optimization algorithms (e.g., Adam, RMSprop).

Model Training: Train the designed model using the pre-processed data. Feed the input data to the model, compute the loss function, and update the model's weights through backpropagation. Repeat the training iteratively over multiple epochs.

Model Evaluation: Using a different test dataset, gauge the trained model's performance. Calculate metrics like F1 score, recall, accuracy, and precision. against determine the model's predictive power, compare it against known stroke cases.

Deployment and Integration: Deploy the trained model as an application or integrate it into existing healthcare systems. Create a user-friendly interface or API for inputting data and obtaining stroke prediction results in real time.

Monitoring and Update: Continuously monitor the performance of the deployed model, collect feedback, and update the model periodically. Ensure that the model remains accurate and effective as new data becomes available or as the population characteristics change.

These modules and functionalities work together to develop an intelligent system capable of predicting the likelihood of a stroke based on various input factors.

ALGORITHM

Convolutional Neural Network (CNN):

A CNN is a Deep Learning system that can take in an input image, give importance (learnable weights and biases) to various aspects/objects in the image, and be able to distinguish one from the other. In comparison to other classification methods, a Conv Net requires significantly less pre-processing. A Conv Net's architecture was influenced by the way the Visual Cortex is organised and is comparable to the connectivity network of neurons in the human brain. Only in a constrained area of the visual field known as the Receptive Field do individual neurons react to inputs. The entire visual field is covered by a group of similar fields that overlap.

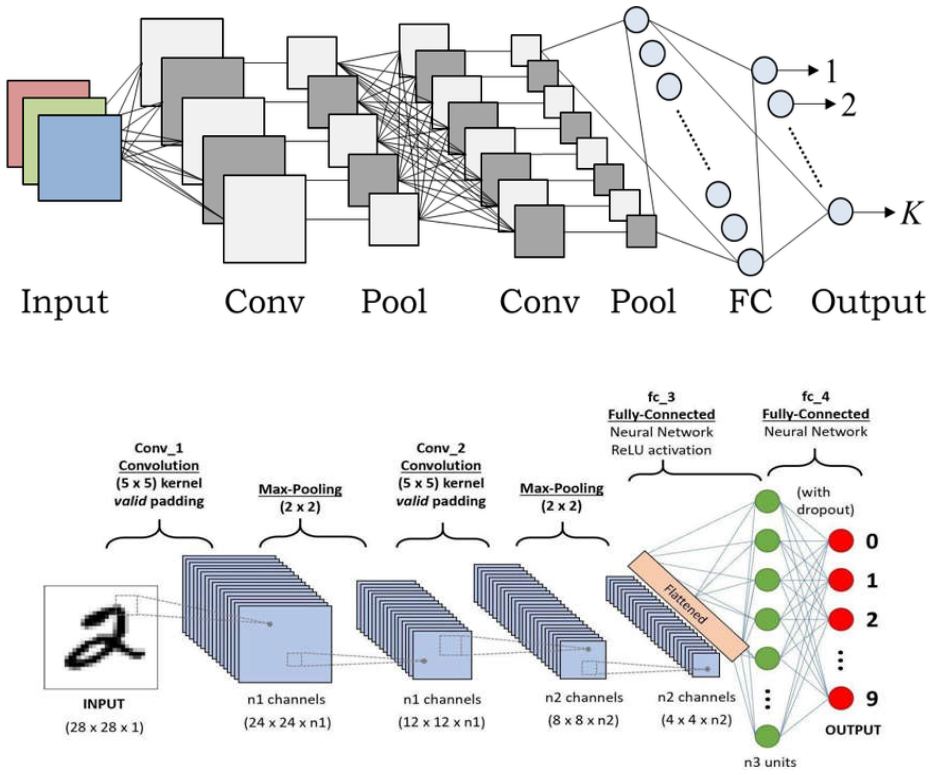


Fig.2. Architecture of CNN

Feature Selection Algorithm:

The number of input variables is reduced using feature selection methods to the ones that are thought to help a model predict the target variable the best. In this post, we've used the three feature selection techniques shown below to choose the optimal characteristics for our model.

Single-factor selection:

We started off by using the univariate feature selection method. To ascertain the extent to which each feature is connected to the outcome variable, the univariate feature selection technique investigates each feature separately. Univariate statistical methods are used to choose the top attributes. For univariate selection, there are several options. To eliminate everything except K's top-scoring characteristics from them, we utilized Select KBest. The default selection functionwas utilized, and each feature's score is given in Table I.

Sl	Specs	Score
1	age	3635.226911
7	avg_glucose_level	1718.285446
3	heart_disease	87.987436
2	hypertension	75.449498
4	ever_married	20.622787
8	bmi	15.894122
9	smoking_sratus	3.369423
5	work_type	2.925901
6	residence_type	0.600717
0	gender	0.239001

Fig.3. Table

The value of a feature

Methods that assign a score (or value) to the input characteristics of the prediction model are referred to as feature importance methods. Each feature's "significance" during prediction is indicated by the score. A feature with a high score value has a greater influence on the model'sability to predict a given variable. With fewer input characteristics, this method is used to grasp the data and model better. Using the predictive model's feature significance characteristic, thisapproach calculates the feature score of each dataset feature and eliminates the features with low scores to obtain the highest score, which simplifies the model and improves performance. Additionally, a built-in class that goes along with Tree-Based Classifiers is called feature significance. Here, as shown in Fig. 2, we used an additional tree classifier to separate the key 10 items from the data set.

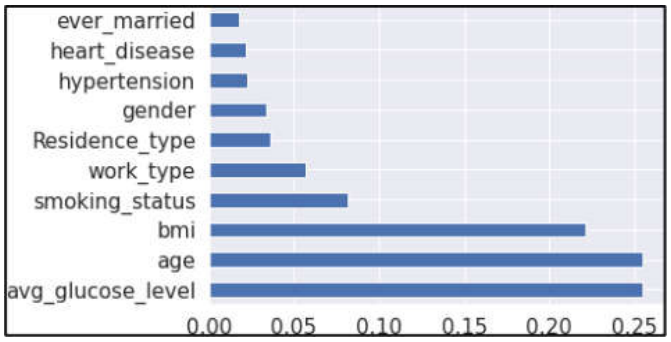


Fig.4. The value of the feature

Heatmap and correlation matrix:

A 2D matrix that shows the correlation coefficients between the characteristics is called a correlation matrix with a heatmap. In order to determine which character in the data set is mostconnected to the target variable, it knows how the features are related to one another and the strength of the relationships. A correlation plot often has several numerical variables, each of which is represented by a columnpositive values indicate a positive association, while negativevalues indicate a negative association. Figure 3 displays the heat map and correlation matrix for the dataset utilized in our investigation. The correlation value in this case is open to any number between -1 and 1. The relationships between variables may be plainly seen through the color coding of the cells.

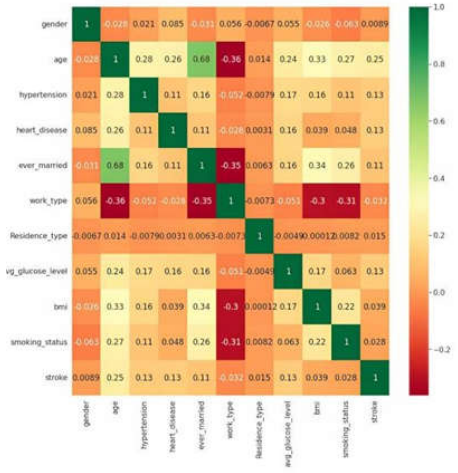


Fig.5. Heat map and correlation

Sample for Training and Validation:

Our collected dataset has been split into two groups: a training group and a testing group. One is utilised to evaluate the model's fit, whereas the other is utilised to fit the learning model. As opposed to this, we used 80% of the dataset's data (i.e., 4088 samples) for training and 20% of the dataset (i.e., 1022 samples) for testing. Before we separate the dataset, we shuffle it for more extensive validation and training.

Traditional Machine Learning Methods:

Logistic regression

A supervised learning approach called logistic regression [21] is used to predict the likelihoodof the outcome variable. When the output of the result variable is either 0 or 1, this approach fits the data the best. When there are only two possible output values for the dataset, logistic regression is chosen.

Classification using decision trees:

Regression and classification issues are handled using the Decision Tree classification method. The information components are comparing yield variables in this approach, which is also a supervised learning strategy. It is made to look like a tree. The two components of a decision tree algorithm are the Decision Node and the Leaf Node. The information is divided up by the hub before it, and the final option is the node that produces the result.

Classification using random forests:

Multiple autonomous decision trees that were freely created on any arbitrary subset of the dataare combined to form Random Forests. These trees are created throughout the training process, and each decision tree produces outcomes. Voting is a mechanism that is used for this classification algorithm's final prediction. According to this method, each decision tree supports a certain outcome class. The class with the highest number of votes is chosen by the random forest as the final expectation.

Support vector device:

The Support Vector Machine (SVM), one of the most well-known supervised learning algorithms, is used to address classification and regression problems. But fundamentally, it is employed in machine learning to address Classification issues. The SVM computation's objective is to establish the best line or limit that can categorise the n-layered space, enabling us to quickly classify the subsequently discovered new informative item. The name for this best-choice limitation is a hyper-plane.

Naive Bayes classifier:

Based on the Bayes hypothesis, the supervised learning technique known as the Naive Bayes algorithm handles classification problems. Most often, it is used for text characterization tasks that demand a sizable training set. The naive Bayes method, one of the simplest and most useful classification algorithms, supports the creation of quick AI models that are capable of making quick predictions.

CNN Proposed Approach:

Modern AI technology known as Convolution Neural Networks (CNNs) has significantly advanced the field of medical applications. Day by day, CNN gained popularity as it outperformed all other models with great accuracy and few errors. A CNN's essential building blocks are convolution layers, pooling layers, and fully associated layers. We created a 1D-CNN to take use of the dataset's structured 1D features, and the model also produces 1D outputs using the CNN method we recommend. To train the model, ten attributes from our dataset are used.

By sending engendering on a preparation dataset, a model's performance under particular sections and loads is assessed. Then, with a propensity to increase estimation, learnable bounds, parts, and loads are updated in accordance with the disaster regard by backpropagation. Highlight extraction produced by hand is not necessary for CNN. Due to the large number of learnable boundaries to gauge, it is unquestionably more information-hungry and hence more computationally expensive, necessitating the use of graphics processing units (GPUs) for modelpreparation.

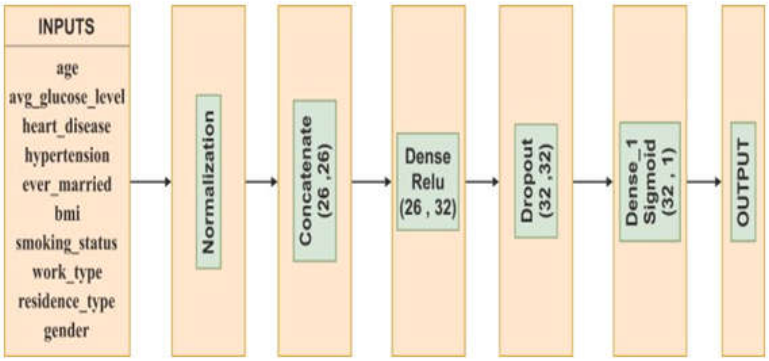


Fig.6. CNN Model

Workflow of the Proposed Approach:

The "Health care dataset for stork prediction" data collection, which comprises 10 characteristics and 1 target column, was first acquired for the project's construction. Then, some data preparation is done, such as eliminating null values and unnecessary data from the dataset.To remove the null values and provide a more uniform performance, we compute the mean value of that column and substitute it for the null value. The category and String data are then normalized using two normalization methods. The model is then trained using all 10 characteristics as parameters. Next, we create our model. In our model, the Dense layer serves as the input layer, and the RELU activation function is used. After normalisation and string lookup, all the features in our model are concatenated, and the input shape is when it reaches the first dense layer. However, when it enters the final dense layer following the dropout layer, it changes, and the output form is as expected.

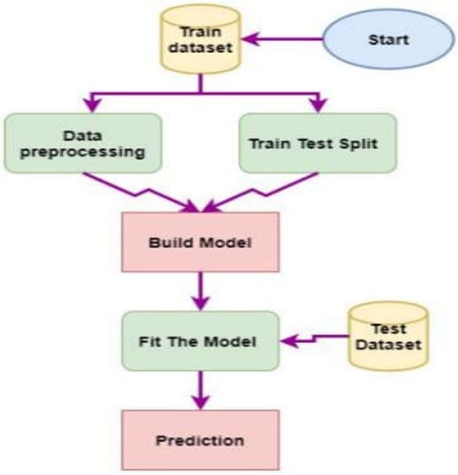


Fig.7. work flow.

LIBRARIES:

TensorFlow:

TensorFlow, developed by Google, is an open-source deep learning library with high- level APIs like Keras. It supports building and training machine learning models, including CNNs, and provides efficient execution on CPUs and GPUs.

Kera's:

Kera's is a user-friendly, high-level neural networks API that runs on top of TensorFlow and other backend libraries. It provides an intuitive interface for building deep learningmodels, including CNNs, with minimal code complexity.

PyTorchs:

PyTorch is an open-source deep-learning library known for its dynamic computationalgraphs and Pythonic programming interface. It offers a flexible and efficient platform for building neural networks, including

CNN.

NumPy:

The foundational Python library for numerical computing is called NumPy. It offers effective numerical operations on arrays, which are multidimensional data structures. In deep learning, NumPy is frequently used for preprocessing and data manipulation.

Pandas:

Data structures and operations for effective data analysis are provided by the potent data manipulation package known as Pandas. It introduces the Data Frame, a two-dimensional structure resembling a table that makes processing structured data simple.

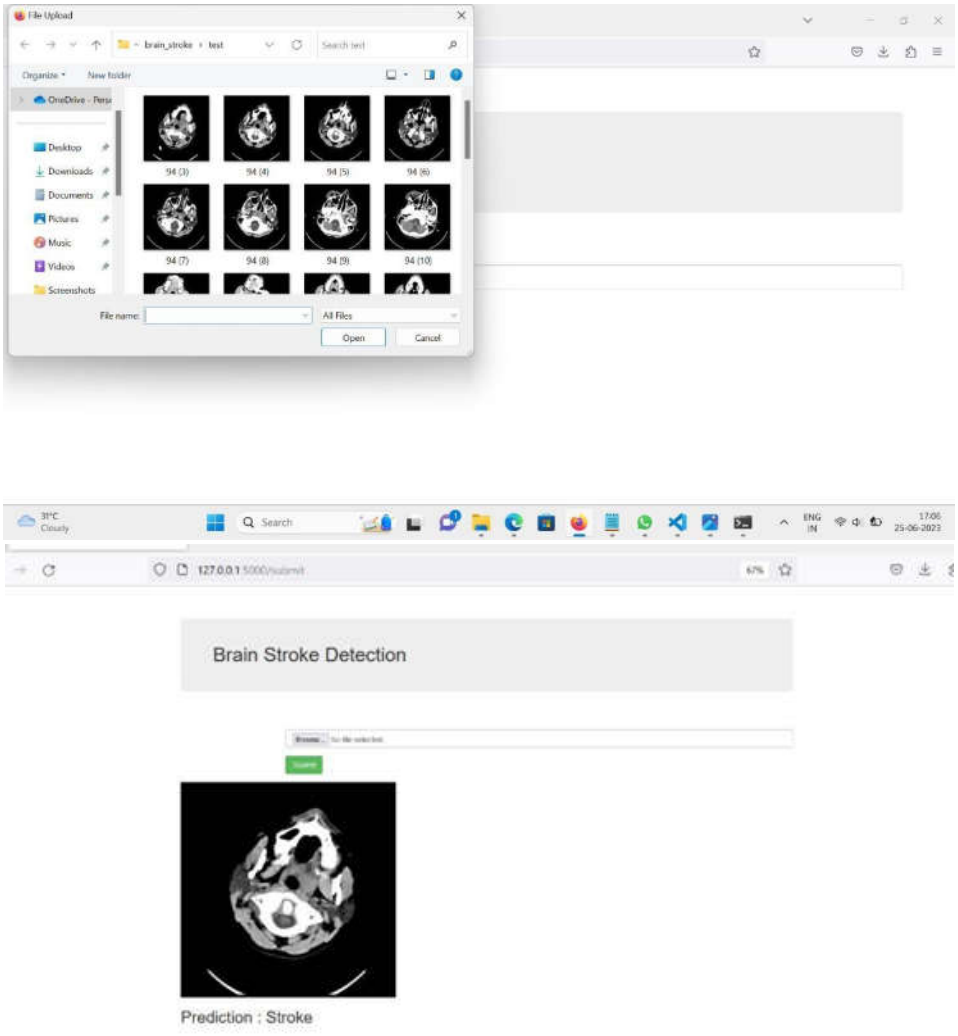
Matplotlib:

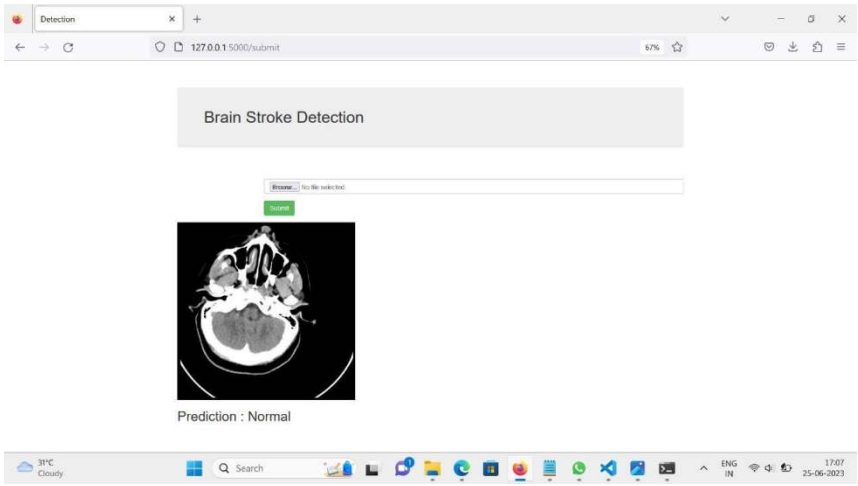
Matplotlib is a widely used plotting library in Python. It provides a flexible and comprehensive set of functions for creating various types of plots and visualizations.

Scikit-learn:

Popular Python machine learning package Scikit-learn offers a variety of tools and algorithms for applications including classification, regression, and clustering.

RESULTS





CONCLUSION

The project's major goal is to detect a stroke in the brain in advance with the highest level of accuracy. This will aid in lowering the death rate from brain stroke. The most traumatic one is a stroke of the brain. For classification, the suggested system employs a Convolutional NeuralNetwork.This project can be carried out by developing a website where anyone can submit a CT brain image for classification. For the same dataset classification, different Machine Learning Algorithms can be employed. The project utilises a Kaggle data set of healthcare information. We have utilised various categorization models to analyse the dataset. Additionally, we created a CNN model to assess the model's performance and forecast whether a person will experience a stroke or not. The experimental finding demonstrates that the proposed model is more efficient than certain other models already in use, with a promising accuracy of 95,5 percent. This model can be used to diagnose a patient and predict their likelihood of experiencing a stroke in the near future. Finally, the analysis of stroke in relation to various traits revealed a broad pattern among the characteristics that are associated with a higher risk of having stroke disease.

FUTURE SCOPE

This project aids in predicting the risk of stroke in older individuals and for those who are dependent on the risk factors indicated in the project. The same project may be expanded in the future to provide the stroke percentage using project output. By gathering data on the relevant risk factors and contacting doctors, this study can also be used to determine the likelihood of stroke in young people and children. Thus, this study aids in predicting the risk of stroke using a prediction model and offers customised warnings and lifestyle corrective messages via a web application. By doing this, it exhorts medical consumers to increase their motivation for managing their health and bring about changes in their health-related behaviours.

REFERENCES

[1] D. Pastore, F. Pacifici, B. Capuani, et al., “Sex-Genetic interaction in the risk for cerebrovascular disease,” International Journal of Environmental Research and Public Health, vol. 24, no. 24, pp. 2687-2699, 2017.

[2] The top 10 causes of death. (2020). [Online]. Available: https://www.who.int/news- room/fact-sheets/detail/the-top-10- causes-of-death

[3] U. R. Acharya, S. L. Oh, Y. Hagiwara, et al., “Automated EEGbased screening of depression using deep convolutional neural network,” Computer Methods and Programs in Biomedicine, vol. 161, pp. 103-113, 2018.

[4] Stroke facts. (2021). [Online]. Available: https://www.cdc.gov/stroke/facts.htm

[5] R. Ramyea, S. Preethi, K. Keerthana, et al., “An intellectual supervised machine learning algorithm for the early prediction of hyperglycemia,” in Proc. Innovations in Power and Advanced Computing

Technologies, 2021, pp. 1-7.
[6] N. S. Adi, R. Farhany, R. Ghina, et al., “Stroke risk prediction model using machine learning,” in Proc. International Conference on Artificial Intelligence and Big Data Analytics, 2021, pp. 56-60.