Global Journal of Engineering Innovations & Interdisciplinary Research

Original Article


# Predictive Analysis to Find Chances of Causing Stroke (Machine Learning)

T Sai Lalith Prasad[1], K Saiprakash[2], M Ashok Reddy[2], M Varshitha[2]

[1]Assistant Professor, Department of Artificial Intelligence and Data Science, Vignan Institute of Technology and Science, Hyderabad, India

[2]UG Student, Department of AI&DS, Vignan Institute of Technology and Science, Hyderabad, India

Correspondence

**T Sai Lalith Prasad**

Assistant Professor , Department of Artificial Intelligence and Data Science, Vignan Institute of Technology and Science, Hyderabad, India

- Received Date: 25 May 2025
- Accepted Date: 15 June 2025
- Publication Date: 27 June 2025

Keywords

Stroke prediction, Encode categorical variables (gender, ever_ married, etc.), Classification, Data preprocessing, Train a Linear SVC model.



## Abstract

*Stroke is recognised as one of the most dangerous it can cause for both life time disability and cause immediate death,across the world, there is a immediate necessary of accurate predictive models. These machine learning models are used to more accurate identifying the person who are more likely to experiencing this life altering medical event. So, now there is need for prediction of stroke for present generation intervention and based on the result taking before medication. Stroke can be predicted by analysing different signs in your body like hypertension, severe headache, trouble speaking and numbness on the face arm or on legs. By giving some health parameters like age, hypertension(0,1), any heart disease, married, residence type, average glucose level are considered as the input feature which is used to training the model and testing the model this helps to predict the given inputs and predict accurate solution. In this model we used algorithms like Linear Support Vector Machine (SVM) for classification, SMOTEENN (Synthetic Minority Over-sampling Technique with Edited Nearest Neighbors) used to balance the dataset, One-Hot Encoding is used to preprocess categorical variables by converting them into a numerical format, Pipeline combines preprocessing and the classification algorithm into a single workflow. This model can give approximately 85-90% accuracy. .*

## Introduction

The stroke is considered as one of the leading cause for death and disability. Some of the traditional methods used to detect the stroke and advising people to immediate meet doctor to decrease the cause of risk. Recent times as machine learning models have generated a great output in accurately measuring and predicting the stroke risk according to the given different inputs. In this model we use different algorithm to predict the model more accurate than before used techniques.

According to world stroke organization it estimates almost 13million people were experience stroke each year. In reality stroke can be caused to anyone, regardless of age, gender and based on their physical health. Strokes may occur immediately, and their symptoms might recognize and cannot be unpredicted at what time it occurs. Symptoms like numbness in the face unable to move sense the muscle movement, no movement in arms or legs, difficulty in speaking fluently, laziness, can't see clearly, headache, vomiting, unable to open mouth, in severe cases.. These sensations may come on suddenly, and in certain rare cases, they may cause immediate effect ion body by this that person can be aware of after causing problem.

The main goal of the current research is to develop an accurate and dependable algorithm that would give reliable predictions about the chances of an individual suffering from a stroke that is one of the debilitating diseases of the modern society. This step is incredibly critical because when individuals, who are at a higher risk of strokes, are identified in the early stages of the disease. Using algorithms like Linear Support Vector Machine (SVM), SMOTEENN (Synthetic Minority Over-sampling Technique with Edited Nearest Neighbors), SMOTETomek (Synthetic Minority Over-sampling Technique with Tomek Links), One-Hot Encoding, Pipeline helps to predict whether anyone is likely to have stroke.these algorithms were used to process the data, cleaning noise by using linear regression, converting the data into model understanding pattern and taking a fair decision and gives a reliable predicted solution.

The image illustrates the blockage of blood cells and clot in brain. When there is stoppage of blood to the brain will be suddenly blocked this leads to death of a person or complete disability. The oxygen should be a constant supply to the brain. There will be a lack of blood flow which means there is a less supply of oxygen and glucose to the neurons or brain cells. When a person is effected by stroke the damage occurs based on the damage of neurons in the brain.

**Citation**: Prasad TSL, Prakash KS, Reddy MA, Varshitha M. Predictive Analysis to Find Chances of Causing Stroke (Machine Learning). GJEIIR. 2025;5(4):074.

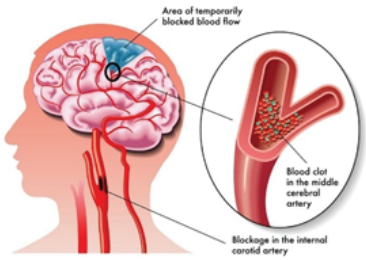


***Figure 1***. *Stroke Cause in Brain*

## Methodology

Preprocessing is the methodology for this stroke prediction system is centered as a well-structured machine learning that combines data preprocessing, class balancing, and predictive modeling. The process begins with cleaning and preparing the dataset. Categorical variables such as gender, marital status, work type, and residence type are encoded into numerical values using One-Hot-Encoder to make them understandable to the machine learning algorithm. Redundant or noisy features like BMI and smoking status are removed to enhance model focus and accuracy. Given the data instability in medical report datasets, where stroke cases are much fewer than non-stroke cases, the system employs advanced resampling techniques like SMOTEENN (Synthetic Minority Oversampling Technique with Edited Nearest Neighbors) and SMOTET-omek (Synthetical Minority Over sample Technique along Tomek Links).SMOTET-omek is a combination of both techniques one is oversampling and understanding the strategies. Tomek Links used to remove overlapping sample data which are near the decision boundary to improve performance. Pipeline used to combines preprocessing data and classification algorithm into a single flow for a proper execution.

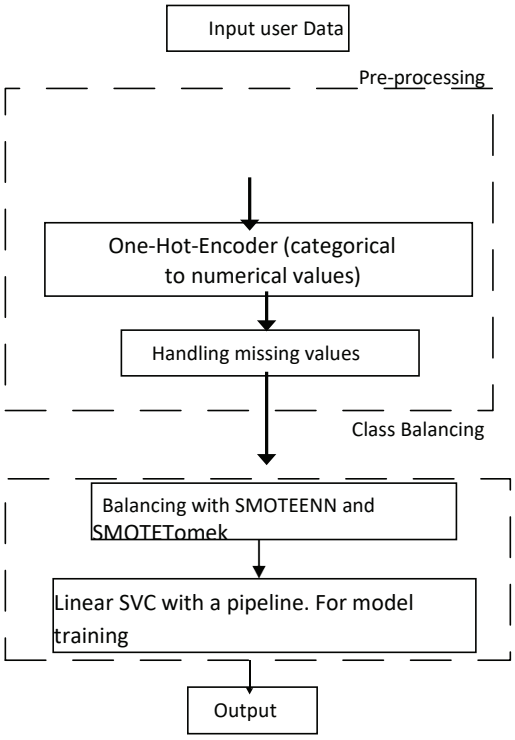


***Figure 2***. *Model of the proposed method*

## Modeling and Analysis

### Methodology: PREDICTIVE ANALYSIS TO FIND CHANCES OF STROKE (MACHINE LEARNING)

The modeling and analysis to implement the stroke predicting machine learning model involve creating a machine learning model which will accurately predict that someone have chances of getting stroke based on their personal and medical information. First, the dataset is prepared by cleaning it and organizing the features such as gender input as(male or female), age(in integer), marital status(married-1 or unmarried-0) and work type(working, children, employed), residence type, and glucose level given in only 1 or 0 only this helps model to understand easily this also handle text-based information like "gender" or "work type", it is converted into numbers using a technique called OneHotEncoding, so the model can understand the input data. Unnecessary features, like BMI and smoking status, are removed because they don't add much value to the predictions and also any records with unclear or unnecessary entries (like gender= "others" ) are removed to avoid confusion.

To deal with the imbalance in the data since stroke cases are much fewer than non-stroke cases, we use two techniques:

SMOTEENN and SMOTETomek. Smoteenn model used to increase the number of positive cases. These methods create extra examples of stroke cases and remove noisy or overlapping data, making the model learn patterns more clearly and fairly. The actual prediction is done using a Linear Support model machine vector is defined as Linear SVC. This algorithm is trained to find patterns in the data and separate stroke cases from non-stroke cases. To ensure the model pays enough attention to the less common stroke cases, extra weight is given to those cases during training. After building the model, it is tested on new, unseen data to check how well it performs. By using proper techniques like balancing the data and adding class weights, the model is able to make fair and accurate predictions, even when the data is imbalanced. The focus is on ensuring the model doesn't just predict "no stroke" for everyone but instead learns the right patterns that indicate a possible stroke. Once the model is trained and performing well, it is saved as a reusable file (saved Model.joblib in the code). Along with it, the preprocessing steps data encoding and transformation are also saved (saved Column Transformer.joblib in the code). This makes it easy to deploy the model for real-time predictions in a web application.

### Post Training Model Evaluation:

To evaluate how a trained model performs, it is tested on hitherto unseen data. One defines the performance metrics, that could, for instance include: accuracy, precision, recall; and others likely used to evaluate how well the model predicts strokes. The focus is on ensuring the model can correctly identify stroke cases (high recall) while keeping false alarms (incorrect stroke predictions) to a minimum.

Balancing Sensitivity and Specificity By using techniques like class weighting and resampling, the model achieves a balance between sensitivity (catching most stroke cases) and specificity (avoiding false positives). This balance ensures the model is both reliable and practical for real-world use.

Real-Time Predictions The trained model is integrated with a Flask web app, where users can input their information (like age, gender, and medical history) and get instant predictions. The predictions are user-friendly, indicating whether the person is at risk of stroke and suggesting further action if required

This can also use the model for personal health monitoring so that anyone, anywhere, can know what future health tests and treatment they should conduct based on previous results prior to huge damage.

It is probed about whether a person is at risk and requires further action in that regard. This can further use the model for personal health monitoring where anyone could know what future health tests and treatments to undergo based on results before it gets to huga damage.. Health insurance companies can also use the system as part of their risk assessment process to offer insurance plans to the people who lives in rural areas.

We always prefer “Prevention is better than cure”

## Architecture description

Enrolling Details :Person Gender(M or F), person Age, is there Hypertension(0 or 1), any Heart Disease occurred before in recent time(0or 1), Married status(0 or 1), Work Type, Residence Type, Average Glucose Level

- The gender of person (e.g., M or F).
- The age of a person(numeric value).
- An individual has hypertension (high blood pressure) or not, usually denoted by 1 for 'Yes' and 0 for 'No'.• Whether the individual has a history of heart disease, represented as 1 for "Yes" and 0 for "No."
- Whether the individual has ever been married, with possible values like "Yes" or "No."
- The type of work the individual is involved in, such as "Private," "Self-employed," "Govt_job," or "Children."
- The type of residence of client living in, either "Urban area" or "Rural area”.

## Importing libraries

- Flask, Time & Random, scikit-learn(Sklearn), numpy & pandas, joblib.

```
from flask import Flask, render_template, request
import time
import random
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.naive_bayes import GaussianNB
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import LabelEncoder
from sklearn.compose import ColumnTransformer
from sklearn.linear_model import LogisticRegression
from sklearn.svm import LinearSVC
import numpy as np
from imblearn.combine import SMOTEENN
from imblearn.combine import SMOTETomek
import sklearn
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
import ml
from joblib import dump, load
```

***Figure 3:*** *Import libraries*

After giving the inputs then the prediction starts here

PREDICTION CODE:

FLOWCHART:

RESULT AND DISCUSSION

When users visit the homepage, the index function is called, which renders the index.html template. This template typically contains the form where users input their details.

```
def predictAnswer(ct, pipe, gender, age, hypertension, heartdisease, ever_married, work_type, Residence_type, avg_glucose_level):
    row = [gender, age, hypertension, heartdisease, ever_married, work_type, Residence_type, avg_glucose_level]
    row = ct.transform([row])
    result = pipe.predict(row)
    result = result[0]
    if result == 1:
        return "Yes, you may be likely to have a stroke. Please see a doctor."
    elif result == 0:
        return "No, you are not likely to have a stroke"
    else:
        return
```

***Figure 4:*** *prediction code*

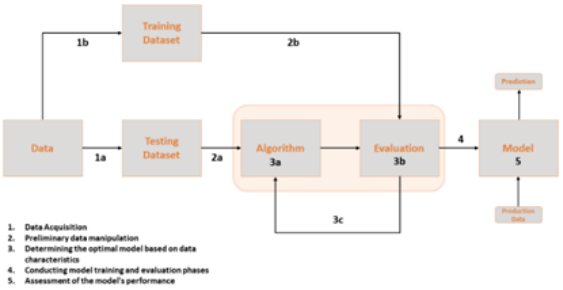


***Figure 5:*** *flow chart*

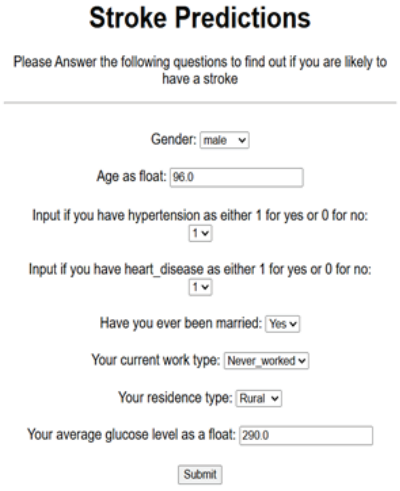


***Figure 6:*** *INPUT DATA to the model*

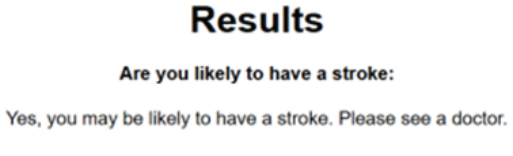


***Figure 7:*** *OUTPUT*

## Conclusion

The proposed model for prediction for stroke greatly overcomes the challenges associated with existing systems by incorporating innovative methodologies to improve accuracy, robustness against noise, and the ability to perform multi-scale analysis .we provided a comprehensive and thorough analysis of various patients' attributes found in electronic health records that are crucial for predicting strokes. Through a systematic approach, we meticulously examined a range of different features that may play a significant role in identifying potential stroke risks among patients. By analyzing this information, we can get the more pecised result of different algorithms used to identify the stroke risk based on given input. The goal is to create a benchmark that helps us to understand which methods work best for stroke prediction. This approach will allow us to enhance our tools and provide better support for healthcare professionals in making informed decisions about patient care. Collecting this data is a crucial step in advancing our research and improving patient outcomes in stroke prevention. The inputs given to the stroke prediction system are entered by the user through a web form in the Flask application, and they are used to predict the likelihood of the individual having a stroke. Here we use different algorithms to get the accurate answer., This model approach ensures the system is not just accurate but also fair and user-friendly. By combining user-friendliness with medical insights, the stroke prediction system has wide arranging applications in healthcare, education, insurance, and public health. Also, it shows robustness against noise for assured performance under suboptimal imaging conditions, as may occur in low-resource settings or in older devices.

This model will advance the field of medical imaging by providing accurate and consistent results, while also supporting healthcare professionals in providing timely and accurate medical requirement. Better patient care and improved outcomes for risk for chances of getting stroke.

## References


1. Learn about Stroke. Available online: https://www.worldstroke.org/world-stroke-daycampaign/why-strokematters/learnabout-stroke (accessed on 25 May 2022).
2. Elloker, T.; Rhoda, A.J. The relationship between social support and participation in stroke: A systematic review. Afr. J. Disabil. 2018, 7, 1–9.
3. Katan, M.; Luft, A. Global burden of stroke. In Seminars in Neurology; Thieme Medical Publishers:New York, NY, USA, 2018; Volume 38, pp. 208–211.
4. Bustamante, A.; Penalba, A.; Orset, C.; Azurmendi, L.; Llombart, V.; Simats, A.; Pecharroman, E.; Ventura, O.; Ribo, M.; Vivien, D.; eta. Blood biomarkers to differentiate ischemic and hemorrhagic strokes.Neurology 2021, 96, e1928–e1939.
5. Xia, X.; Yue, W.; Chao, B.; Li, M.; Cao, L.; Wang, L.; Shen, Y.; Li, X. Prevalence and risk factors of stroke in the elderly in Northern China: Data from the National Stroke Screening Survey. J. Neurol. 2019, 266, 1449–1458.
6. Alloubani, A.; Saleh, A.; Abdelhafiz, I. Hypertension and diabetes mellitus as a predictive risk factor for stroke. Diabetes Metab.Syndr. Clin. Res. Rev. 2018, 12, 577–584.
7. Boehme, A.K.; Esenwa, C.; Elkind, M.S. Stroke risk factors,genetics, and prevention. Circ. Res. 2017, 120, 472–495.
8. Mosley, I.; Nicol, M.; Donnan, G.; Patrick, I.; Dewey, H. Stroke symptoms and the decision to call for an ambulance.
9. Who.int. [Online]. Available: http://www.who.int/healthinfo/global_burden_disease/en/. [Accessed: 15-Apr-2021].
10. V. L. Feigin et al., “Global and regional burden of stroke during 1990
11. R. Bhallamudi et al., "Deep Learning Model for Resolution Enhancement of Biomedical Images for Biometrics," in Generative Artificial Intelligence for Biomedical and Smart Health Informatics, Wiley Online Library, pp. 321–341, 2025.
12. R. Bhallamudi et al., "Artificial Intelligence Probabilities Scheme for Disease Prevention Data Set Construction in Intelligent Smart Healthcare Scenario," SLAS Technology, vol. 29, pp. 2472–6303, 2024, Elsevier.
13. R. Bhallamudi, "Improved Selection Method for Evolutionary Artificial Neural Network Design," Pakistan Heart Journal, vol. 56, pp. 985–992, 2023.
14. R. Bhallamudi et al., "Time and Statistical Complexity of Proposed Evolutionary Algorithm in Artificial Neural Networks," Pakistan Heart Journal, vol. 56, pp. 1014–1019, 2023.
15. R. Krishna et al., "Smart Governance in Public Agencies Using Big Data," The International Journal of Analytical and Experimental Modal Analysis (IJAEMA), vol. 7, pp. 1082–1095, 2020.
16. N. M. Krishna, "Object Detection and Tracking Using YOLO," in 3rd International Conference on Inventive Research in Computing Applications (ICIRCA-2021), IEEE, Sept. 2021, ISBN: 978-0-7381-4627-0.
17. N. M. Krishna, "Deep Learning Convolutional Neural Network (CNN) with Gaussian Mixture Model for Predicting Pancreatic Cancer," Springer US, vol. 1380-7501, pp. 1–15, Feb. 2019.
18. N. M. Krishna, "Emotion Recognition Using Skew Gaussian Mixture Model for Brain–Computer Interaction," in SCDA-2018, Textbook Chapter, ISBN: 978-981-13-0514, pp. 297–305, Springer, 2018.
19. N. M. Krishna, "A Novel Approach for Effective Emotion Recognition Using Double Truncated Gaussian Mixture Model and EEG," I.J. Intelligent Systems and Applications, vol. 6, pp. 33–42, 2017.
20. N. M. Krishna, "Object Detection and Tracking Using YOLO," in 3rd International Conference on Inventive Research in Computing Applications (ICIRCA-2021), IEEE, Sept. 2021, ISBN: 978-0-7381-4627-0.
21. T. S. L. Prasad, K. B. Manikandan, and J. Vinoj, "Shielding NLP Systems: An In-depth Survey on Advanced AI Techniques for Adversarial Attack Detection in Cyber Security," in 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS), IEEE, 2024.
22. S. Sowjanya et al., "Bioacoustics Signal Authentication for E-Medical Records Using Blockchain," in 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), vol. 1, IEEE, 2024.
23. N. V. N. Sowjanya, G. Swetha, and T. S. L. Prasad, "AI

Based Improved Vehicle Detection and Classification in Patterns Using Deep Learning," in Disruptive Technologies in Computing and Communication Systems: Proceedings of the 1st International Conference on Disruptive Technologies in Computing and Communication Systems, CRC Press, 2024.

24. C. V. P. Krishna and T. S. L. Prasad, "Weapon Detection Using Deep Learning," Journal of Optoelectronics Laser, vol. 41, no. 7, pp. 557–567, 2022.
25. T. S. L. Prasad et al., "Deep Learning Based Crowd Counting Using Image and Video," 2024.