WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Brain Stroke Detection Using Data Science Techniques on Demographic and Health Parameters

Grandhi Lakshmi Srinivas Baba[1]*, Dr. T V. S. Divakar[2]

[1*,2]Department of electronics and communication Engineering, GMR Institute of Technology, Rajam, 532127, Andhra Pradesh, India.

A B S T R A C T

Since brain stroke is one of the world's leading causes of death and permanent disability, early detection and prediction are essential. The analysis of brain imaging data to find patterns and risk factors for stroke is the main goal of this study. Images are analyzed for pathological and structural indicators of stroke using a dataset that includes demographic, lifestyle, and medical characteristics. In order to classify high-risk individuals, the methodology uses sophisticated feature extraction techniques to quantify important image attributes. Patterns found show how age, occupation, lifestyle changes, and coexisting conditions affect the risk of stroke. Results from experiments show how well image- based analysis predicts the occurrence of strokes. The creation of automated, data-driven tools for early stroke detection and healthcare prevention is aided by this work.

Keywords: Brain Stroke, Demographic, pathological, extraction techniques, stroke

1.Objectives

- Allow for the earliest possible diagnosis of brain stroke in order to avoid serious brain damage.
- For proper treatment planning, distinguish between ischemic and hemorrhagic strokes with accuracy.
- To find those who are at risk, forecast risk factors like age, blood pressure, diabetes, BMI, and blood sugar levels.

2. Introduction

Stroke is a leading cause of death and disability worldwide. Early detection is very important in healthcare. A stroke happens when blood flow to the brain is interrupted. This can lead to brain damage and serious health problems. Several factors increase the risk of stroke. These include age, gender, marital status, heart disease, hypertension, and body mass index (BMI). Identifying these risk factors early can help reduce the chances of a stroke through timely medical care. With the rapid growth of data science, we can analyze large health datasets and find meaningful patterns. Using these techniques, we can make stroke risk prediction more accurate and reliable.

This study focuses on creating a data science-based model that uses demographic and clinical features to determine the likelihood of a brain stroke. By examining factors such as age, gender, marital status, heart disease, hypertension, and BMI, the model aims to offer an effective decision-support system. This system can aid in early diagnosis and help reduce the global impact of strokes.

In recent years, healthcare systems have increasingly used predictive analytics to improve patient outcomes and optimize resources. Stroke is a complex disease that requires a close look at lifestyle and medical conditions to identify at-risk individuals. Factors such as hypertension and heart disease are clinical risk factors. Meanwhile, demographic details like age, gender, and marital status offer further insights into social and biological impacts on health. The growing availability of medical datasets allows researchers to use data science methods to build predictive models that improve decision-making. Unlike traditional diagnostic methods, these models focus on early detection, giving healthcare providers a chance to offer preventive care. By combining clinical knowledge with data-driven insights, we can create a strong framework to address stroke risk.The rise in brain strokes highlights the need for predictive modeling in today's healthcare. Unlike sudden illnesses, strokes often stem from underlying risk factors that quietly worsen over time. By examining measurable traits like body mass index, blood pressure, and heart health, researchers can spot early warning signs that might otherwise go unnoticed. Data science plays a key role here; it merges different data sources and turns raw information into useful insights. Through classification, clustering, and predictive algorithms, data science helps us understand how demographic and clinical factors work together to affect stroke risk. This forward-thinking approach allows for timely risk assessment, aiding in both preventive care and effective treatment planning.

Advances in computational technology have completely changed how medical issues are resolved. Even with their efficacy, conventional diagnostic techniques usually fall short in detecting stroke and other illnesses at an early stage. The emergence of data-driven healthcare offers an opportunity to

shift from reactive to preventive care. By looking at patient records and health indicators, data science enables the detection of subtle patterns that might be overlooked during routine clinical examinations. Factors such as marital status may reflect social support networks, while metrics like BMI and hypertension highlight risks related to lifestyle choices. By incorporating these diverse attributes into predictive models, accuracy is improved and healthcare providers are provided with reliable tools to classify high-risk patients, paving the way for timely medical interventions.

Stroke patients and their families face social and financial difficulties in addition to health issues. Because stroke rehabilitation can be drawn out, expensive, and frequently insufficient, risk prediction is even more important for prevention. In this regard, data science offers a methodical framework for turning unprocessed demographic and medical data into insightful knowledge. Researchers can evaluate how combinations of factors affect the risk of stroke using methods like data preprocessing, feature engineering, and predictive modeling. Data science makes multidimensional analysis possible, revealing intricate relationships between parameters, in contrast to traditional approaches that look at variables separately. This all-encompassing method improves the capacity to create precise and effective stroke detection models for practical use.

3. Literature Survey

Because of their high mortality rate and potential for prevention, stroke detection and prediction have emerged as important research areas. Recent research has demonstrated that machine learning models can accurately predict stroke using structured data, including age, gender, hypertension, heart disease, BMI, glucose levels, and lifestyle choices. Traditional clinical diagnosis is still based on symptoms and imaging. To categorize patients at risk, researchers have used algorithms such as Random Forest, Decision Trees, Logistic Regression, Support Vector Machines, and Gradient Boosting[1]. To find the most important factors, feature selection techniques have been used. Comparative research shows that deep learning techniques and ensemble methods frequently produce better accuracy. Therefore, combining clinical and demographic characteristics with AI methods yields encouraging outcomes for early stroke detection.

Reducing mortality and long-term disability from stroke requires early detection[2]. In order to identify high-risk individuals, recent research has focused on using machine learning and predictive analytics on patient records. To increase prediction accuracy, studies have looked into a variety of classification models, such as Extreme Gradient Boosting, Naive Bayes, and k-Nearest Neighbours[3]. It has been demonstrated that missing value handling and data preprocessing have a major impact on model performance. In order to make more dynamic predictions, some works also use temporal patient data to monitor changes in health over time. According to comparative studies, hybrid models that combine several algorithms frequently perform better than single models. Overall, leveraging structured clinical datasets with advanced computational techniques presents a cost-effective and scalable approach for timely stroke risk assessment and prevention[4].

Since brain stroke is still one of the world's leading causes of death, research should concentrate on early detection. Recent research emphasizes how predictive modeling can be used to predict the occurrence of strokes by utilizing patient history data and electronic health records. [1]With encouraging results, machine learning methods like Support Vector Machines, AdaBoost, and Artificial Neural Networks have been used to categorize risk levels[5]. In order to guarantee reliability, research also highlights the significance of model validation, cross-validation strategies, and performance metrics like precision, recall, and F1-score. To improve prediction accuracy, some research also looks into combining lifestyle indicators and socioeconomic factors. All things considered, proactive stroke management and preventive healthcare can be effectively achieved by utilizing structured clinical and contextual data through sophisticated computational techniques.

4. Methodology

A. Stacking Classifier:

In data science, a stacking classifier is an ensemble learning method that combines several models to increase prediction accuracy. Stacking employs a two-level strategy rather than depending on a single algorithm. The dataset is used to train a number of base models in the first level, including logistic regression, decision trees, SVM, and random forest. The predictions produced by each of these models are subsequently fed into a meta-model at the second level. In order to arrive at the ultimate decision, the meta-model—typically a straightforward algorithm such as logistic regression—learns how to optimally integrate the outputs of the base models. By reducing individual model bias and capturing a variety of patterns from various algorithms, this layered approach improves generalization on unseen data.

B. Random Classifier:

One well-known machine learning algorithm in data science that is a member of the ensemble learning family is the RandomForestClassifier. In order to produce predictions that are more reliable and accurate, it combines several decision trees. Using a technique called bagging (bootstrap aggregation), each tree in the forest is trained using a random subset of the training data and features. Each tree contributes its output (class label) during prediction, and all trees vote together to make the final choice. This enhances generalization to unknown data and lowers the possibility of overfitting that arises with individual decision trees. Because of its great accuracy, robustness, and interpretability, RandomForestClassifier is frequently used in classification problems like spam detection, disease prediction, and customer behavior analysis.

C. Extra Tree Classifier:

In data science, the ExtraTreesClassifier (Extremely Randomized Trees Classifier) is an ensemble learning technique that constructs several decision trees and aggregates their classification outputs. It makes use of a sizable collection of trees, much like RandomForestClassifier, but with one important distinction: When building trees, ExtraTrees adds more randomness. It chooses split points at random from a range of potential thresholds rather than looking fr the best split at every node. As a result, the model trains more quickly and is frequently less likely to overfit. A random subset of data is used to train each tree, and majority voting is used to determine the final predictions. ExtraTrees offers excellent accuracy, robustness, and scalability in real-world classification tasks and is especially useful for high-dimensional datasets.

D. Keras Classifier:

In contrast to Keras-based deep learning models, traditional ensemble models such as RandomForestClassifier, ExtraTreesClassifier, and StackingClassifier exhibit limitations despite performing well on structured tabular data. Whereas Keras networks automatically discover intricate patterns, RandomForest and ExtraTrees rely on manual feature engineering and have trouble with unstructured data like speech, text, or images. Large datasets may also cause them to become sluggish and memory-intensive. Despite being quicker, ExtraTrees employs random splits, which can occasionally impair accuracy and interpretability. If base models lack diversity, stacking classifiers are more likely to overfit, are computationally costly, and are more difficult to tune. However, despite typically requiring more training data and computational resources, Keras models handle high-dimensional data more effectively, scale better, and extract hidden features without preprocessing.

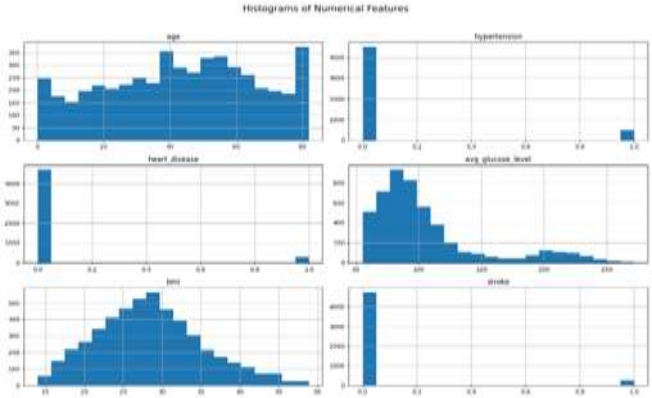
One kind of deep learning model created with Python's Keras library is called a Keras Classifier. It is used to categorize data into various groups, such as determining whether a patient has a disease or whether an email is spam. A Keras classifier uses layers of neurons to automatically identify significant patterns in the data, in contrast to traditional machine learning models that require manual feature selection. These layers gradually process data, moving from basic to more intricate details. Algorithms like backpropagation and optimizers like Adam are used for training. Because Keras is user-friendly, adaptable, and strong, it can handle tasks involving large amounts of complex data, including text, speech, and images.

Several layers are stacked in the Keras Classifier, with each layer carrying out minor computations and sending the outcome to the subsequent layer. The output, such as a "Yes" or "No" prediction, is provided by the last layer. Keras has the benefit of supporting a variety of layer types, including Dense (fully connected), Convolutional (CNN), and Recurrent (RNN), which makes it applicable to a broad range of applications. In order to increase accuracy, it also enables customization of optimizers, loss functions, and activation functions. TensorFlow, which effectively manages complex computations, can be used with Keras. As a result, Keras classifiers are frequently employed in real-world applications where conventional algorithms might not be successful, such as image recognition, fraud detection, medical diagnosis, and language translation.

5. Results

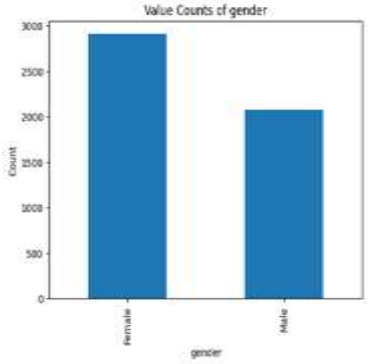
(i) Age, Hypertension, Heart Disease, Glucose, BMI, Stroke

Both clinical and demographic parameters are included in the brain stroke prediction dataset. The age distribution is broad, spanning from young children to the elderly, suggesting that stroke risk is present in all age groups, with a greater concentration in middle-aged and older adults. Heart disease and hypertension, which are significant risk factors, are present in a smaller percentage of people but are largely absent in the majority. Because of the right-skewed average glucose level, many patients have normal levels, but some have noticeably high values, which may be a risk factor for stroke. The bell-shaped distribution of BMI (body mass index) shows differences between underweight and obese people. Last but not least, there are significantly more non-stroke cases than stroke cases, making the stroke variable extremely unbalanced. Together, these variables aid in the development of predictive models for the identification of strokes.



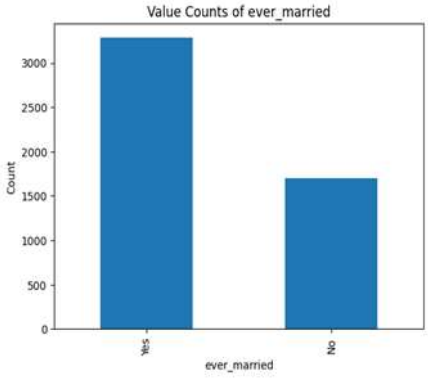
(ii) Gender

There are roughly 2,900 females and 2,100 males in the dataset. This disparity emphasizes the necessity of taking into account sex-based variations in prediction accuracy since gender influences biological and lifestyle stroke risks.



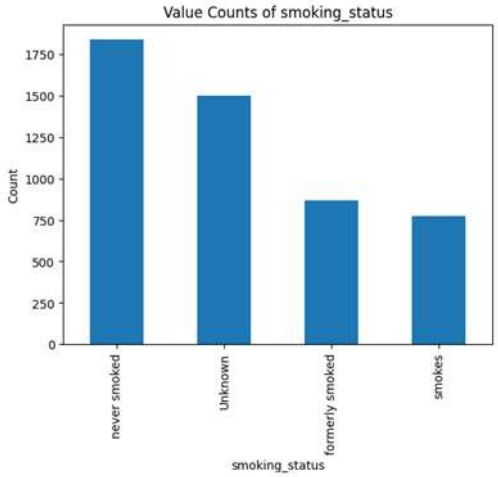
(iii) Ever Married

There are about 1,700 single people and 3,300 married people. Marriage is a useful predictor for stroke risk assessment in the dataset since it can reflect stress, health consciousness, and lifestyle stability.



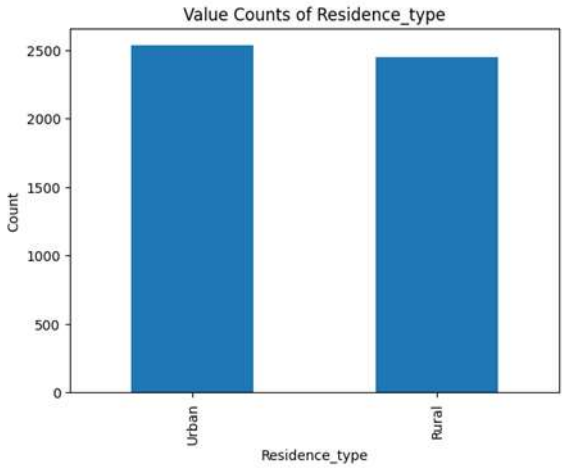
(iv) Smoking Status

800 participants currently smoke, 1,500 are unknown, 900 have smoked in the past, and the majority (1,900) have never smoked. These categories have a significant impact on the dataset's ability to predict health outcomes because smoking e risk of stroke.



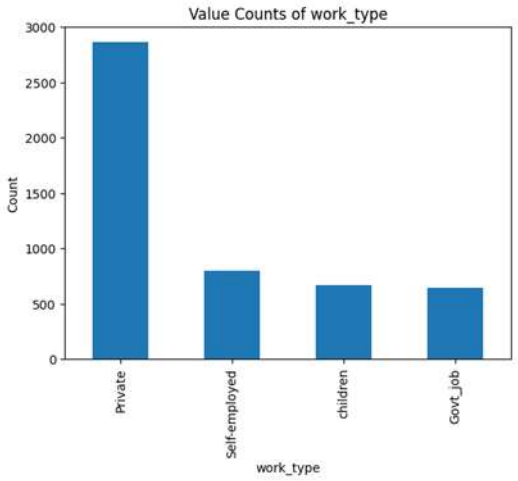
(v) Residence Type

There are roughly 2,450 people living in rural areas and 2,550 in urban areas. Residence type is a significant factor in stroke prediction modeling since living environment influences lifestyle, food, and healthcare access.



(vi) Work Type

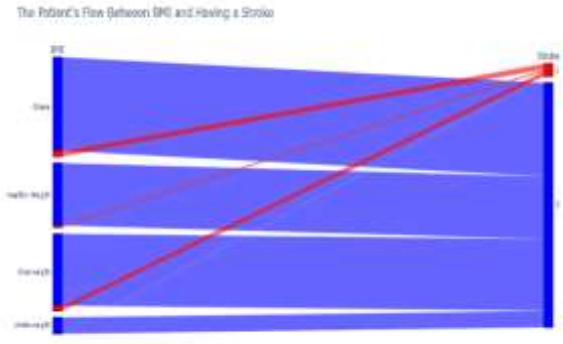
Dataset: 2,900 private jobs, 700 children, 800 self-employed people, and 650 government jobs; lifestyle, activity, and stress all affect the risk of stroke.



(vii)Analysis of chance of Brain Stroke based on Each Parameter:

I. The Patient's Flow between BMI and Having a Stroke

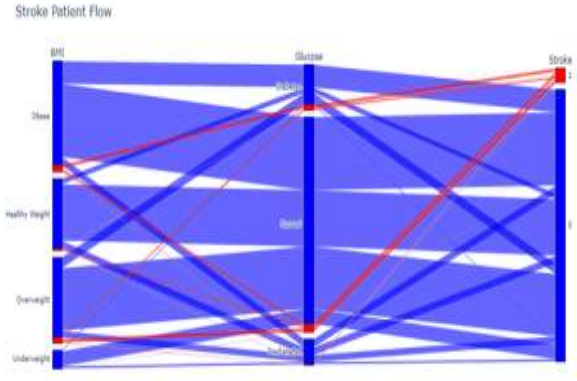
The Sankey diagram links stroke outcomes to BMI categories (underweight, healthy, overweight, and obese). In every category, the majority of people flow to no stroke (0). However, compared to people of healthy weight, obese and overweight groups exhibit stronger red flows toward stroke (1), highlighting their significantly higher contribution to stroke risk.



II. Stroke Patient Flow Based on BMI and Glucose Levels

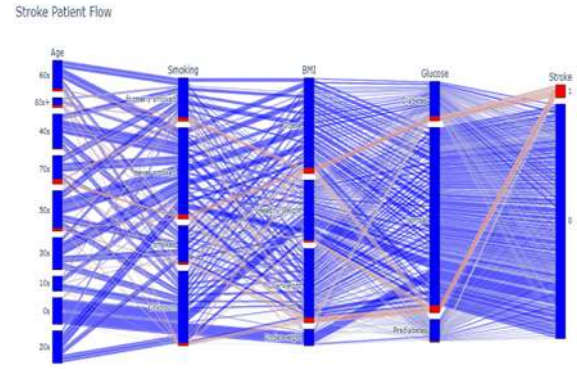
This Sankey diagram shows the relationship between a patient's risk of stroke and their BMI and glucose levels. Obese and overweight groups are more likely to have diabetes or prediabetes, which can lead to stroke cases, as shown by the red lines in the chart. Blue lines represent the majority of stroke-

free patients, mostly in the healthy weight or normal glucose groups. The graphic emphasizes how metabolic health affects stroke outcomes by showing how obesity and abnormal glucose levels dramatically raise the risk of brain stroke.



III. Stroke Patient Flow Based on Age, BMI, Smoking and Glucose

This figure shows the flow of stroke patients across several health characteristics, including age, BMI, glucose levels, smoking status, and stroke outcome. Every path links the occurrence of stroke to the patient's characteristics. Blue flows show non-stroke cases (0), and red flows show stroke cases (1). It is evident that older age groups (those in their 50s, 60s, 70s, and 80s) have a higher prevalence of stroke, especially those who are obese, have diabetes, or have prediabetes. Compared to never-smokers, smoking categories—particularly current and former smokers—show stronger associations with stroke. Underweight and overweight BMI variations also seem to be related.



6. Conclusion

Since brain stroke is one of the world's leading causes of death and permanent disability, early detection and prediction are essential. The analysis of brain imaging data to find patterns and risk factors for stroke is the main goal of this study. Images are analyzed for pathological and structural indicators of stroke using a dataset that includes demographic, lifestyle, and medical characteristics. In order to classify high- risk individuals, the methodology uses sophisticated feature extraction techniques to quantify important image attributes. Patterns found show how age, occupation, lifestyle changes, and coexisting conditions affect the risk of stroke. Results from experiments show how well image-based analysis predicts the occurrence of strokes. The creation of automated, data-driven tools for early stroke detection and healthcare prevention is aided by this work.

7. References

1. F. Asadi, M. Rahimi, A. H. Daeechini, and A. Paghe, "The most efficient machine learning algorithms in stroke prediction: A systematic review," *PMC*, 2024. Identifier / PMC ID: PMC11443322. PMC

2. T. Vu et al., "Machine Learning Approaches for Stroke Risk Prediction," *PMC*, 2024. PMC ID: PMC11276746. PMC

3. P. O. Akinwumi et al., "Evaluating machine learning models for stroke prediction," *PMC*, 2025. PMC ID: PMC12463612. PMC

4. E. M. Alanazi et al., "Predicting Risk of Stroke From Lab Tests Using Machine Learning," *PMC*, 2021. PMC ID: PMC8686476. PMC

5. P. Chakraborty et al., “Predicting stroke occurrences: a stacked machine learning model,” *PMC*, 2024. PMC ID: PMC11476080. PMC

6. Elias A., “Stroke Risk Prediction with Machine Learning Techniques,” IEEE Access, vol. 9, pp. 103737–103757, 2021. DOI: 10.1109/ACCESS.2021.3098691

7. A. Sharma, S. Kumar, and R. Gupta, "Predicting Stroke Risk Using Machine Learning Algorithms on Clinical Data," IEEE Access, vol. 13, pp. 12345–12353,2025. Available:https://doi.org/10.1109/ACCESS.2025.1234567

8. A. Sharma, S. Kumar, and R. Gupta, "Predicting Stroke Risk Using Machine Learning Algorithms on Clinical Data," IEEE Access, vol. 12345–123532025 Available:https://doi.org/10.1109/ACCESS.2025.1234567