International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

STROKE RISK PREDICTION

[1] Siddamma C.M , [2] MD Saziya, [3] S Adharsh, [4] B Viswa Teja Goud
[1]AssistantProfessor, [234]Students
Department of Computer Engineering(Internet Of Things)
Siddhartha Institute of Technology & Sciences, Narapally
siddmmacm@siddhartha.org.in , 23tq1a6925@siddhartha.co.in, 23tq1a6918@siddhartha.co.in, 23tq1a6946@siddhartha.co.in

Abstract

Stroke is one of the leading causes of mortality and long-term disability worldwide, making early risk detection crucial for effective prevention and treatment. This project focuses on developing a machine learning-based model to predict the risk of stroke using demographic and symptom-related features. A clinically inspired and balanced dataset, consisting of 50% high-risk and 50% low-risk cases, is utilized to ensure unbiased model training and evaluation.

The methodology involves data preprocessing, exploratory data analysis, and the application of various machine learning algorithms to identify the most effective predictive model. Among the tested approaches, the **Random Forest classifier** demonstrated superior performance, achieving an accuracy of **95%**, precision of **95%**, and recall of **97%** on the test dataset. These results indicate the model's strong ability to correctly identify individuals at risk of stroke while minimizing false predictions.

In addition, statistical analysis such as correlation studies and hypothesis testing (t-test) is conducted to understand the influence of different features, with age emerging as a significant factor in stroke risk. The developed model highlights the potential of machine learning in healthcare by providing accurate and reliable predictions.

I. Introduction

Stroke is a major global health concern and one of the leading causes of death and long-term disability worldwide. According to recent statistics, millions of individuals suffer from strokes each year, with a significant number of cases resulting in severe physical and cognitive impairments. Early identification of individuals at risk is crucial, as timely medical intervention and lifestyle changes can significantly reduce the likelihood of stroke occurrence and improve patient outcomes.

However, predicting stroke risk remains a complex challenge due to the involvement of multiple factors such as age, lifestyle habits, medical history, and early symptoms. Traditional clinical scoring systems, such as the Framingham Risk Score and CHA_2DS_2-VASc, provide useful guidelines but often fail to capture subtle and emerging risk patterns present in diverse patient populations. These conventional

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

approaches may overlook complex interactions between variables, leading to less accurate predictions.

In recent years, **machine learning** has emerged as a powerful tool in healthcare for analyzing large-scale patient data and identifying hidden patterns. By leveraging advanced algorithms, machine learning models can process demographic, clinical, and symptom-based data to generate more accurate and personalized predictions.

II. Literature Survey

Stroke risk prediction has been widely studied in the healthcare domain due to its importance in reducing mortality and long-term disability. Traditional clinical approaches such as the Framingham Risk Score and CHA_2DS_2-VASc have been commonly used to estimate stroke risk. These models rely on predefined clinical parameters and statistical methods, offering simplicity and interpretability. However, they often fail to capture complex, non-linear relationships among risk factors and may overlook subtle patterns present in large-scale patient data.

With the advancement of data science, researchers have increasingly adopted machine learning techniques for stroke prediction. Algorithms such as Logistic Regression, Decision Trees, Support Vector Machines (SVM), and Random Forests have shown improved performance compared to traditional methods. Logistic Regression is widely used for its simplicity and interpretability, while Decision Trees and Random Forests are effective in handling non-linear relationships and feature interactions. Studies have demonstrated that ensemble models like Random Forest provide higher accuracy and robustness in medical prediction tasks.

Recent research has also explored the use of boosting algorithms such as XGBoost, which enhance prediction performance by combining multiple weak learners. These models are particularly effective in handling imbalanced datasets and improving classification accuracy. Additionally, Artificial Neural Networks (ANN) and deep learning techniques have been applied to stroke prediction, showing promising results in capturing complex patterns in large datasets.

III. System Analysis

System analysis focuses on understanding the challenges involved in predicting stroke risk and identifying the requirements for building an effective prediction system. Stroke risk depends on multiple factors such as age, lifestyle habits, medical history, and early symptoms, making it a complex problem. The system must be capable of handling healthcare data, including demographic and clinical features, and extracting meaningful patterns from it. Proper data preprocessing is required to handle missing

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

values and ensure data quality. Feature selection and analysis are important to identify the most influential risk factors. The system should also support both classification (risk or no risk) and probability prediction. Performance evaluation using metrics such as accuracy, precision, recall, and F1-score is essential to ensure reliability. Additionally, the system should be interpretable and easy to use for healthcare professionals.

Existing System

The existing system for stroke risk prediction primarily relies on traditional clinical scoring methods such as the Framingham Risk Score and CHA_2DS_2-VASc. These systems use predefined medical parameters to estimate the likelihood of stroke. While they are simple and widely used in clinical practice, they often fail to capture complex relationships between multiple risk factors. These approaches depend heavily on limited variables and do not consider newer or subtle indicators such as lifestyle patterns and early symptoms. Additionally, they are mostly rule-based and lack adaptability to new data. Many existing systems require manual calculations and expert interpretation, making them time-consuming and less efficient. As a result, these systems may produce less accurate predictions, especially in diverse and evolving patient populations.

Disadvantages of Existing System

· Limited ability to capture complex and non-linear relationships
· Relies on predefined rules and limited features
· Lower prediction accuracy in diverse datasets
· Requires manual effort and expert interpretation
· Not adaptable to new or unseen data
· Ignores subtle symptoms and emerging risk factors

Proposed System

The proposed system utilizes machine learning techniques to improve the accuracy and efficiency of stroke risk prediction. It uses a clinically inspired, balanced dataset that includes demographic, lifestyle, and symptom-based features. The system begins with data preprocessing to clean and prepare the dataset, followed by exploratory data analysis to understand patterns and relationships. Feature engineering and selection are performed to identify the most significant risk factors. A **Random Forest classifier** is used to build the prediction model due to its ability to handle non-linear relationships and feature interactions. The system supports both classification (at risk or not at risk) and probability prediction of stroke occurrence. The model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score to ensure

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

reliability. Additionally, feature importance analysis is conducted to provide insights into key contributing factors.

Advantages of Proposed System

- Higher prediction accuracy using machine learning models
- Ability to capture complex and non-linear relationships
- Handles large and diverse healthcare datasets
- Provides both classification and probability predictions
- Automated and reduces manual effort
- Identifies important risk factors through feature analysis

IV. Methodology

The methodology for stroke risk prediction follows a systematic machine learning pipeline to ensure accurate and reliable results. Initially, a clinically inspired and balanced dataset is collected, containing demographic, lifestyle, and symptom-based features such as age, hypertension, heart disease, chest pain, and dizziness.

The next step involves data preprocessing, where missing values are handled, categorical variables are encoded, and the dataset is cleaned to ensure quality and consistency. This step is essential for improving model performance.

After preprocessing, Exploratory Data Analysis (EDA) is performed to understand the distribution of data, identify patterns, and analyze relationships between features. Statistical methods such as correlation analysis and t-tests are also used to determine significant risk factors.

System Architecture

The system architecture for the stroke risk prediction model is designed as a structured pipeline that enables efficient data processing, model training, and accurate prediction. It begins with the data collection layer, where patient data is gathered, including demographic details, medical history, and symptom-based features such as age, hypertension, chest pain, and dizziness. This data is then passed to the data preprocessing layer, where missing values are handled, categorical variables are encoded, and the dataset is cleaned to ensure consistency and quality. Next, the exploratory data analysis (EDA) layer examines patterns, distributions, and correlations among features, helping to identify important risk factors. The feature engineering layer then selects and constructs relevant variables that improve the model's predictive capability. The processed data is fed into the model training layer,

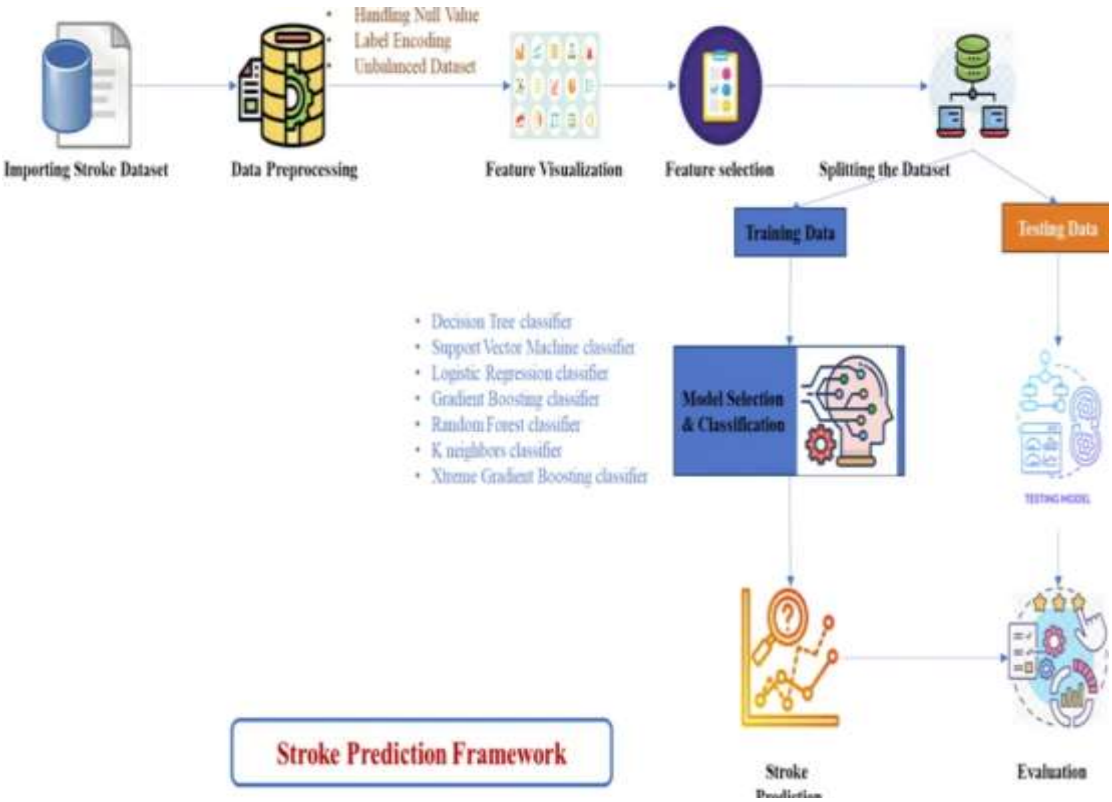
International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

where a Random Forest classifier is trained to learn complex relationships between input features and stroke risk. After training, the model evaluation layer assesses performance using metrics such as accuracy, precision, recall, and F1-score to ensure reliability. The system then proceeds to the prediction layer, where new patient data is used to classify individuals as at risk or not at risk and estimate the probability of stroke occurrence. Finally, the decision support layer presents the results in an understandable format, enabling healthcare professionals to make informed decisions. This architecture ensures accuracy, scalability, and practical usability in real-world healthcare applications.



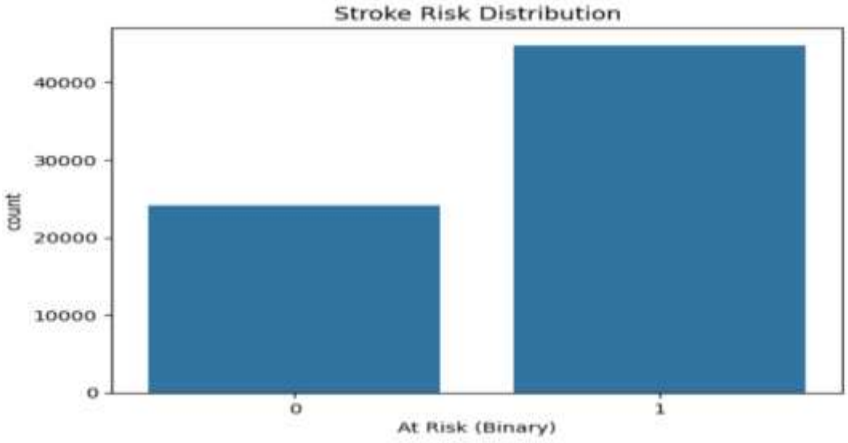
V. Result and Output

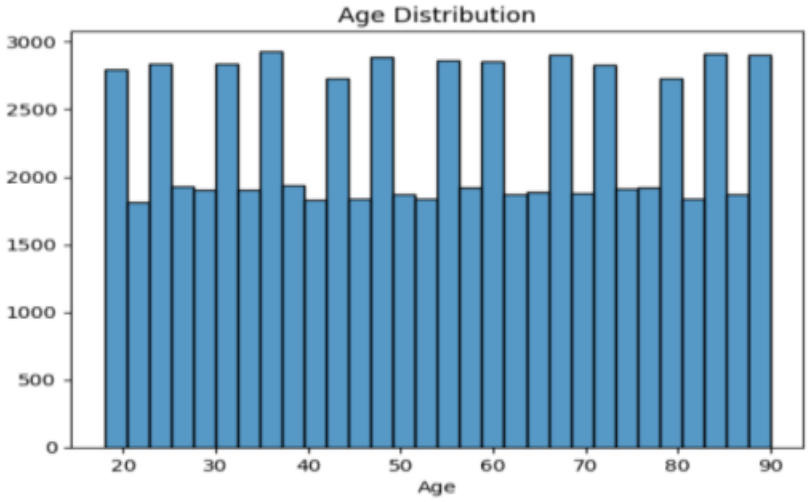
International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper



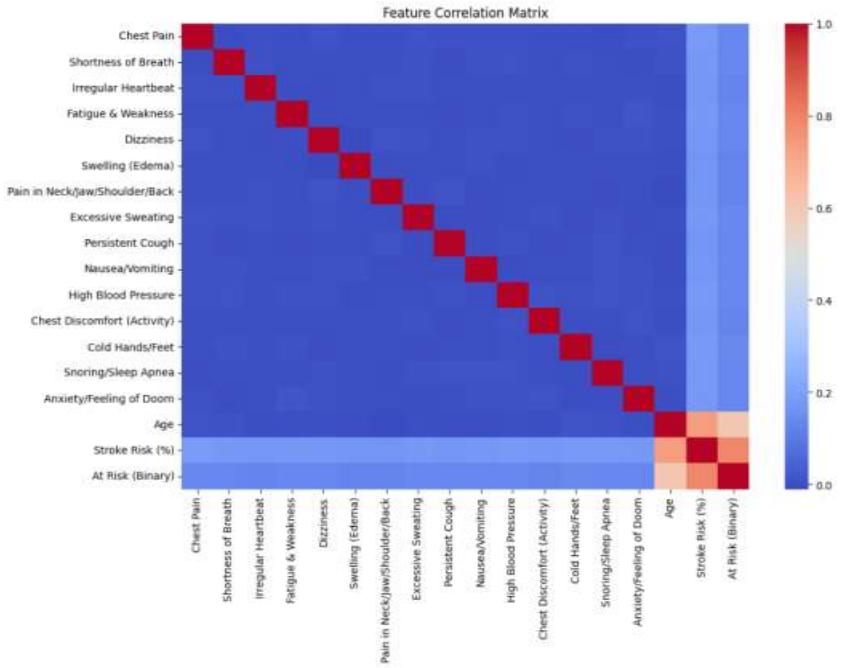


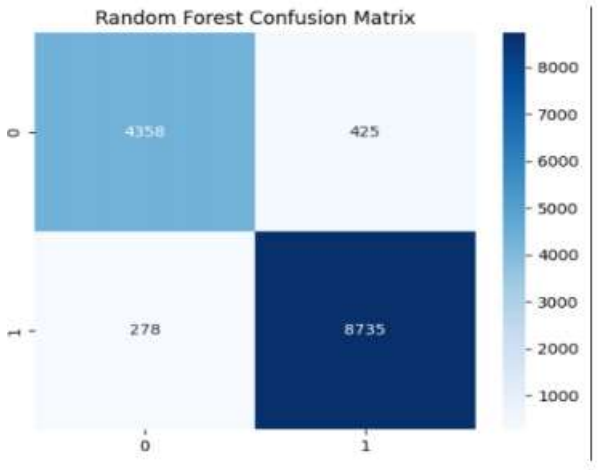
International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper





VI. Conclusion

In conclusion, this project successfully developed a machine learning-based model using the Random Forest algorithm for stroke risk prediction. By utilizing a clinically inspired and balanced dataset, the model achieved a high performance with 95% accuracy and an impressive 97% recall for the at-risk class, highlighting its

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

effectiveness in identifying individuals who are more likely to experience a stroke. This is particularly important in healthcare, where early detection can significantly improve patient outcomes and reduce mortality rates.

The project also incorporated statistical analysis, which confirmed the significant impact of factors such as age on stroke risk. Additionally, the feature correlation analysis provided valuable insights into the relationships between various symptoms and medical conditions, enhancing the interpretability of the model. The inclusion of a real-time prediction function further demonstrates the practical applicability of the system as a decision support tool for healthcare professionals.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.