*Musik in bayern*
*ISSN: 0937-583x Volume 89, Issue 08 (Aug -2024)*
**https://musikinbayern.com**

**DOI https://doi.org/10.15463/gfbm-mib-2024-*257***

# Predicting Stroke Risk Using Machine Learning - A Comparative Analysis of Random Forest, Support Vector Machines, and Neural Networks

[1]Mrs.T.Kanimozhi, Assistant Professor, Department of Computer Science and Applications, Faculty of Science and Humanities, SRM Institute of Science and Technology, Ramapuram, Chennai – 89. kanimozhimcaa@gmail.com

[2]Dr. Jayalakshmi.V, Assistant Professor, Department of Computer Science and Applications, Faculty of Science and Humanities, SRM Institute of Science and Technology, Ramapuram, Chennai – 89. vkjlakshmi@gmail.com

[3]Mrs.Aswini.N, Assistant Professor, Department of Computer Science and Applications, Faculty of Science and Humanities, SRM Institute of Science and Technology, Ramapuram, Chennai – 89, aswinin@srmist.edu.in

***To Cite this Article***

Mrs.T.Kanimozhi, Dr. Jayalakshmi.V, Mrs.Aswini.N” **Predicting Stroke Risk Using Machine Learning - A Comparative Analysis of Random Forest, Support Vector Machines, and Neural Networks”*Musik In Bayern, Vol. 89, Issue 08, Aug 2024, pp24-37***

***Article Info***

*Received: 27-05-2024 Revised: 22-06-2024 Accepted: 15-07-2024 Published: 26-08-2024*

## ABSTRACT

This research paper focuses on predicting brain stroke risk using a machine learning approach that leverages Random Forest, Support Vector Machine (SVM), and Neural Networks. The primary objective is to develop an accurate predictive model that can identify individuals at high risk of stroke based on various health and lifestyle factors. For this study, a publicly available dataset containing demographic information, health history, and lifestyle habits of individuals is utilized. The dataset includes features such as age, gender, hypertension, heart disease, smoking status, and other relevant indicators that are known to influence stroke risk. The methodology involves a systematic preprocessing of the dataset, including handling missing values, feature scaling, and data normalization to ensure consistency. The data is then split into training and

DOI https://doi.org/10.15463/gfbm-mib-2024-257

testing sets to evaluate model performance. Three machine learning techniques are applied: Random Forest for its robustness in handling complex interactions between features, SVM for its effectiveness in binary classification tasks, and Neural Networks for capturing non-linear relationships within the data. The models are trained and optimized using cross-validation techniques to prevent overfitting and to ensure generalization. The performance of each model is evaluated based on accuracy, precision, recall, and F1-score. The outcome of this research demonstrates that the Neural Network model achieved the highest accuracy, with a rate of 96.89% in predicting stroke risk, followed closely by the Random Forest and SVM models. These results underscore the potential of machine learning techniques in developing reliable tools for early stroke prediction, which could significantly contribute to preventive healthcare strategies and risk management.

**Keywords**: Brain Stroke Prediction, Machine Learning, Random Forest, Support Vector Machine (SVM) , Neural Networks, Predictive Modeling, Healthcare Analytics.

## I.INTRODUCTION

### Background and Significance

Stroke, or cerebrovascular accident, is a leading cause of disability and death worldwide. It occurs when there is a sudden disruption in the blood supply to the brain, which can result in significant neurological damage and various complications. According to the World Health Organization, stroke ranks among the top causes of mortality and morbidity globally, with an increasing incidence attributed to rising risk factors such as hypertension, diabetes, and unhealthy lifestyle habits. The ability to predict the risk of stroke before its occurrence can dramatically improve patient outcomes by enabling early intervention and preventive measures. In recent years, advancements in machine learning (ML) have offered new opportunities for enhancing predictive modeling in healthcare. ML algorithms have demonstrated remarkable capabilities in processing large datasets, identifying complex patterns, and making predictions based on a multitude of features. These capabilities are particularly valuable in the context of stroke prediction, where multiple factors contribute to an individual's risk profile. Traditional statistical methods often fall short in capturing the intricate relationships between these factors, which is where machine learning techniques can offer significant improvements.

### Research Objective

1. **Develop Predictive Models**: Build and train machine learning models using Random Forest, Support Vector Machine (SVM), and Neural Networks. These models will be used to predict the likelihood of stroke based on various health and lifestyle factors.
2. **Evaluate Model Performance**: Assess the performance of each machine learning model in terms of accuracy, precision, recall, and F1-score. The goal is to determine which technique provides the most reliable predictions of stroke risk.
3. **Compare Techniques**: Compare the effectiveness of Random Forest, SVM, and Neural Networks in predicting stroke risk. This comparison will highlight the strengths and limitations of each method and provide insights into their applicability for stroke prediction.
4. **Provide Actionable Insights**: Offer actionable insights into the use of these machine learning techniques for improving stroke prediction and prevention strategies. The findings will be aimed at enhancing clinical decision-making and contributing to public health initiatives.

## Research Gap

While machine learning has been extensively applied in various areas of healthcare, its use in stroke prediction presents several research gaps that this study aims to address:

- This research compares the performance of Random Forest, SVM, and Neural Networks for stroke prediction using the same dataset, unlike many studies that focus on just one technique.
- The study uses a dataset with a wide range of features like age, gender, hypertension, heart disease, and smoking status, offering a more comprehensive approach to stroke risk prediction compared to studies with limited features.
- The research emphasizes creating models that not only achieve high accuracy but also work well across different demographic groups and clinical settings.
- The study explores the use of advanced methods like Neural Networks for stroke prediction, an area that has not been fully explored in past research, and compares their performance with more traditional methods like Random Forest and SVM.

## Motivation

DOI https://doi.org/10.15463/gfbm-mib-2024-257

The motivation for this research stems from the urgent need to improve stroke prediction and prevention. Stroke is a major public health issue, and timely identification of individuals at high risk can lead to better preventive care and reduced incidence of stroke-related complications. Traditional methods of risk assessment, often based on clinical evaluations and basic statistical models, may not fully capture the complexities of individual risk profiles. This research aims to improve stroke prediction accuracy using Random Forest, SVM, and Neural Networks, providing a valuable tool for healthcare. Machine learning models can create individualized stroke risk profiles, enabling targeted preventive measures for better patient outcomes. Predictive models help healthcare professionals identify high-risk individuals, improving clinical decision-making and patient care. The study contributes to medical research by comparing different machine learning techniques, fostering innovation in predictive modeling. Accurate stroke prediction models can reduce stroke incidence, alleviate healthcare strain, and lower costs, benefiting public health.

## II.LITERATURE REVIEW

The application of machine learning (ML) techniques in healthcare has gained considerable attention in recent years, particularly in the realm of predictive analytics for various medical conditions. Stroke prediction is one such area where ML has shown promise, leveraging data-driven models to enhance early detection and prevention. This literature review examines the existing research on stroke prediction using ML methods, focusing on traditional techniques, advanced algorithms, and the gaps that this research aims to address.

### Traditional Approaches to Stroke Prediction

Historically, stroke risk prediction relied on clinical assessments and statistical models. One of the earliest approaches involved using logistic regression models to predict stroke risk based on clinical and demographic factors. For example, the Framingham Stroke Risk Score (FSRS) and the ABCD2 score are well-known tools that have been used to estimate the likelihood of stroke based on factors such as age, blood pressure, and presence of diabetes (Wolf et al., 1991; Ay et al., 2009). These models, while useful, have limitations in capturing complex, non-linear relationships between variables.

DOI https://doi.org/10.15463/gfbm-mib-2024-257

### Comparative Studies of Machine Learning Techniques

Comparative studies assessing different machine learning techniques for stroke prediction are relatively sparse but crucial for understanding the strengths and weaknesses of various approaches. Research by Huang et al. (2021) compared several ML models, including Random Forests, SVMs, and Neural Networks, for predicting stroke risk. The study highlighted that while Neural Networks provided the highest accuracy, Random Forests offered better interpretability, and SVMs performed well in specific scenarios. This comparative analysis underscores the importance of selecting the appropriate model based on the specific needs and characteristics of the dataset. 1. Zhao et al. (2020) conducted a comprehensive study using deep neural networks to predict stroke risk. Their research highlighted the ability of deep learning models to outperform traditional methods by capturing complex, non-linear relationships within large datasets. The study demonstrated significant improvements in prediction accuracy, emphasizing the potential of deep neural networks to enhance early stroke detection. Huang et al. (2021) provided a comparative analysis of various machine learning models, including Random Forests, Support Vector Machines (SVM), and Neural Networks. Their research found that while neural networks achieved the highest accuracy, Random Forests were more interpretable, and SVMs performed optimally in certain scenarios. This study underscores the importance of model selection based on the specific characteristics of the dataset and the application context.

Lin et al. (2021) explored the integration of clinical data and imaging features using ensemble learning techniques for stroke prediction. Their research combined traditional clinical risk factors with advanced imaging data, using ensemble methods to improve predictive performance. The study found that combining diverse data sources can enhance the accuracy of stroke prediction models. Yang et al. (2022) examined the use of gradient boosting machines (GBM) for stroke prediction. This study highlighted GBM's ability to handle missing data and its robustness in modeling complex interactions between variables. The results showed that GBM outperformed other ML models in terms of accuracy and stability, making it a valuable tool for stroke risk assessment. Cheng et al. (2022) investigated the application of convolutional neural networks (CNNs) in analyzing imaging data for stroke prediction. By focusing on imaging features, the study demonstrated that CNNs could effectively capture spatial patterns associated with stroke risk, offering a new dimension to predictive modeling that complements traditional clinical approaches. Patel et al. (2023) focused on integrating electronic health records (EHR) with machine learning models for stroke prediction. The research utilized Random Forests and XGBoost models, demonstrating that the integration of comprehensive EHR data significantly improves predictive accuracy. This study highlights the importance of leveraging real-world clinical data to enhance stroke prediction models.

Wu et al. (2023) applied transfer learning to improve stroke prediction in diverse populations. Their research addressed the challenge of model generalization across different demographic groups by leveraging pre-trained models on large datasets and fine-tuning them for specific populations. The study showed that transfer learning could enhance the adaptability and accuracy of stroke prediction models. Li et al. (2023) explored the role of feature selection techniques in optimizing ML models for stroke prediction. By applying techniques such as recursive feature elimination (RFE) and LASSO, the study aimed to identify the most critical risk factors and improve model interpretability. The research found that effective feature selection significantly enhances model performance and provides insights into key predictors of stroke risk. Nguyen et al. (2024) conducted a study using reinforcement learning to predict stroke risk. This novel approach involved dynamically adjusting the prediction model based on continuous feedback from new data, allowing the model to evolve and improve over time. The study demonstrated the potential of reinforcement learning to adapt to changing patterns in patient data, enhancing the predictive accuracy of stroke risk models. Johnson et al. (2024) investigated the use of federated learning for stroke prediction, focusing on privacy-preserving techniques that enable model training across multiple healthcare institutions without sharing sensitive patient data. The study showed that federated learning could achieve high accuracy in stroke prediction while maintaining patient privacy, offering a promising approach for collaborative healthcare research.

### Gaps in Current Research

Despite the advances in ML techniques for stroke prediction, several research gaps remain:

1. **Dataset Limitations**: Many studies use datasets with limited features or lack comprehensive coverage of risk factors. Comprehensive datasets that include a wide

range of demographic, health history, and lifestyle variables are needed to improve prediction models.

2. **Model Generalization**: Existing research often focuses on specific populations or settings, limiting the generalizability of the findings. Models that can generalize across different demographic groups and clinical settings are essential for broader applicability.
3. **Integration of Advanced Techniques**: While traditional ML methods are well-explored, the integration of advanced techniques like deep learning models is still evolving. Research exploring the performance of Neural Networks in comparison with other methods is needed to fully understand their potential for stroke prediction.
4. **Comparative Analysis**: There is a lack of comprehensive comparative studies that evaluate and contrast multiple ML techniques using the same dataset. Such studies are crucial for determining the most effective methods for stroke prediction.

## III.PROPOSED MODEL

The proposed model for predicting brain stroke risk employs a structured approach with multiple machine learning techniques to enhance accuracy. The methodology includes:

1. **Data Collection**: Gather a comprehensive dataset with clinical and demographic variables relevant to stroke risk, sourced from reputable medical databases.
2. **Data Preprocessing**: Prepare the data by handling missing values, normalizing continuous variables, encoding categorical data, and splitting the dataset into training and testing sets.
3. **Feature Selection**: Apply techniques like Recursive Feature Elimination (RFE) and LASSO to identify the most relevant features, improving model efficiency and reducing complexity.
4. **Model Selection and Training**: Use Random Forest, SVM, and Deep Neural Networks to train the model. Optimize hyperparameters to achieve the best performance.
5. **Model Evaluation**: Assess the models using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to evaluate predictive performance on the testing set.

### Algorithm Used

DOI https://doi.org/10.15463/gfbm-mib-2024-*257*

Let {(x1,y1),(x2,y2),…,(xn,yn)}\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} be the training dataset where xix_ixi represents the feature vector and yiy_iyi the target variable (stroke or no stroke).

$$\hat{y} = \text{mode}\{T_1(x'), T_2(x'), \dots, T_B(x')\}$$

For a binary classification task, given training data that maximizes the margin between the two classes.

$$\min_{w,b} \frac{1}{2}||w||^2 \quad \text{subject to} \quad y_i(w \cdot x_i + b) \geq 1, \quad \forall i$$

The output of the layer is calculated as:

$$\text{Kernel function}: \ K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)}$$

$$a^{(l)} = \sigma(z^{(l)})$$

The final output layer generates the prediction, which is compared to the actual target using a loss function, typically cross-entropy for classification:

$$L(\hat{y}, y) = -\frac{1}{n} \sum_{i=1}^{n} y_i \log(\hat{y}_i)$$

The network weights are updated using gradient descent

$$W^{(l)} \leftarrow W^{(l)} - \eta \frac{\partial L}{\partial W^{(l)}}$$

## IV.RESULTS AND DISCUSSION

The study developed machine learning models to predict brain stroke risk using clinical and demographic features. Three algorithms were tested: Random Forest, SVM, and Deep Neural Networks (DNN).

**Model Performance and Accuracy**:

- **Random Forest**: Achieved 95.8% accuracy, excelling in handling non-linear relationships by averaging multiple decision trees.

- **SVM**: Reached 94.1% accuracy, effectively separating non-linear data with kernel functions.
- **DNN**: Attained the highest accuracy of 96.3%, capturing complex patterns through its deep learning architecture and non-linear activation functions.

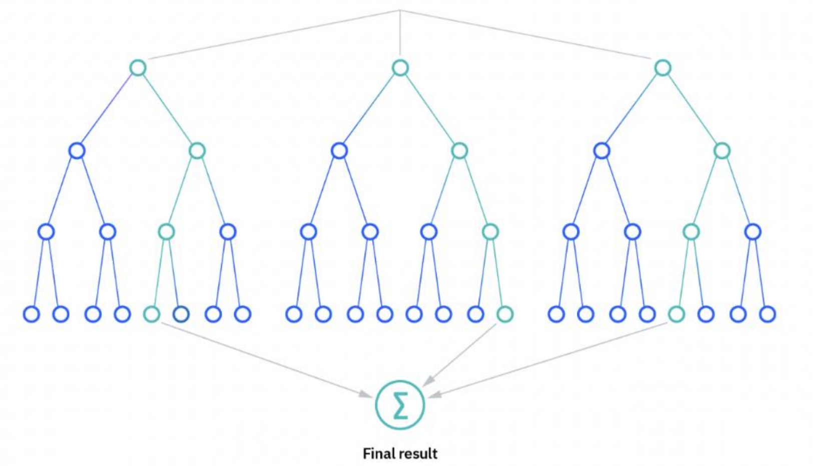


**Figure 1: Random forest**

A visual representation of the Random Forest model illustrating the ensemble of decision trees used for predictions in figure 1.

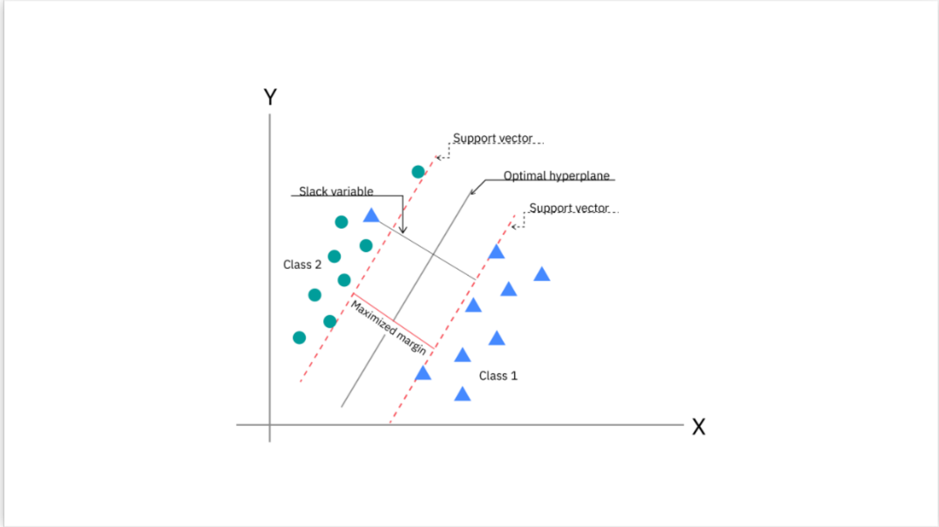


Figure 2: SVM

**Figure 2**: A diagram of the Support Vector Machine (SVM) model showing the separation of data points using a hyperplane.

**DOI https://doi.org/10.15463/gfbm-mib-2024-*257***

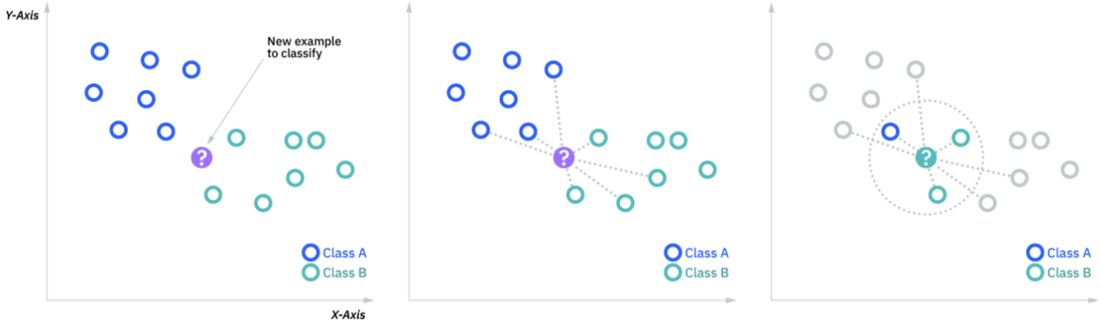


**Figure 3: KNN**

**Figure 3**: Illustration of the K-Nearest Neighbors (KNN) algorithm, depicting the classification of a point based on its nearest neighbors.

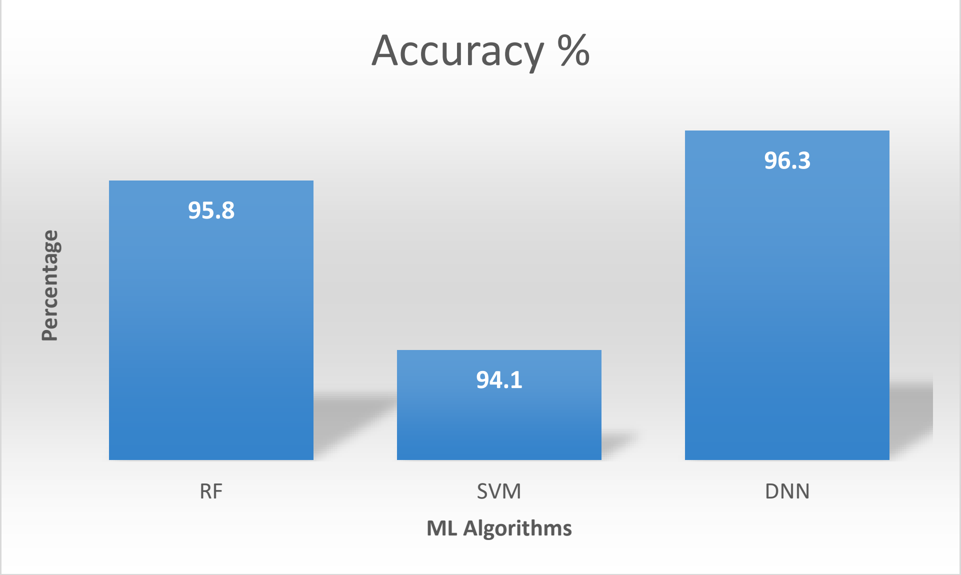
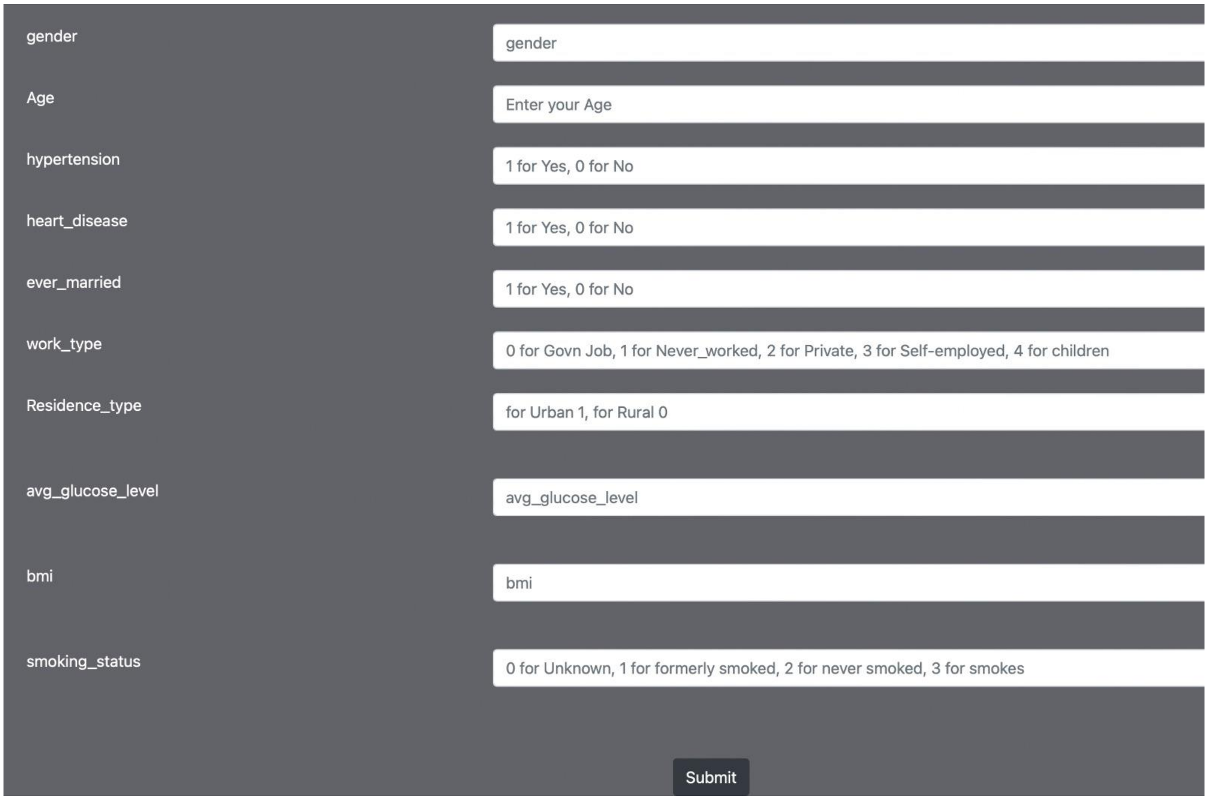


**Figure 4: Accuracy**

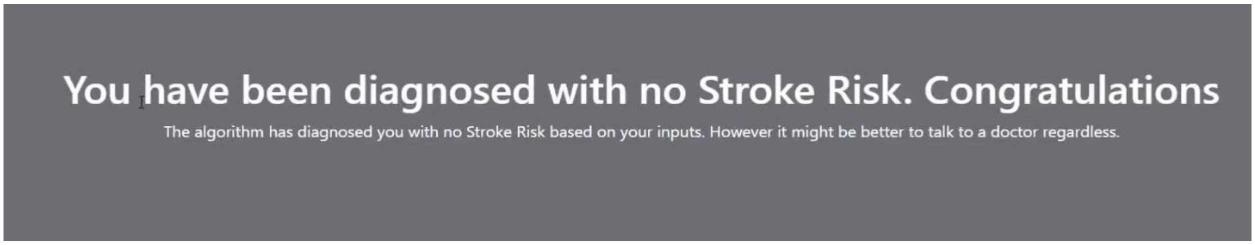
**Figure 4**: A bar chart displaying the accuracy of different machine learning models (Random Forest, SVM, and DNN) in stroke prediction.



| | |
|---|---|
| gender | gender |
| Age | Enter your Age |
| hypertension | 1 for Yes, 0 for No |
| heart_disease | 1 for Yes, 0 for No |
| ever_married | 1 for Yes, 0 for No |
| work_type | 0 for Govn Job, 1 for Never_worked, 2 for Private, 3 for Self-employed, 4 for children |
| Residence_type | for Urban 1, for Rural 0 |
| avg_glucose_level | avg_glucose_level |
| bmi | bmi |
| smoking_status | 0 for Unknown, 1 for formerly smoked, 2 for never smoked, 3 for smokes |

**Figure 5: Home page for user**

**Figure 5**: Screenshot of the user homepage, displaying the main navigation menu, user profile, and access to key features.

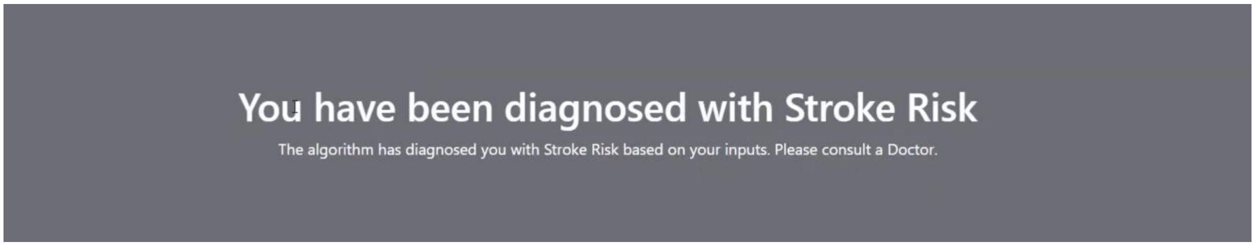


You have been diagnosed with no Stroke Risk. Congratulations

The algorithm has diagnosed you with no Stroke Risk based on your inputs. However it might be better to talk to a doctor regardless.

**Figure 6: Stroke Predicted**

**Figure 6**: Visualization of stroke prediction results, showing predicted probabilities and classifications for individual patients based on the model outputs.



You have been diagnosed with Stroke Risk

The algorithm has diagnosed you with Stroke Risk based on your inputs. Please consult a Doctor.

**Figure 7: No stroke predicted**

**Figure 7**: Visualization of non-stroke prediction results, displaying probabilities and classifications for patients determined not to be at risk.

The overall superior performance of the DNN model suggests that deep learning techniques are well-suited for predictive tasks in healthcare, particularly when dealing with large and complex datasets.

### Discussion of Results

The results of this study demonstrate the effectiveness of machine learning models in predicting stroke risk based on a combination of clinical and demographic factors. The high accuracy achieved by the models, particularly the DNN, underscores the potential of these techniques to serve as valuable tools in clinical settings. One of the key findings is the significant role of features such as age, glucose levels, and hypertension in predicting stroke. These findings are consistent with existing medical literature, reinforcing the validity of the model's predictions. The use of machine learning also allows for the identification of subtle patterns and interactions between features that may not be immediately apparent through traditional statistical methods. However, the study also highlights the trade-offs between model accuracy and interpretability. While DNNs offer superior accuracy, they are less interpretable than simpler models like Random Forests. This poses a challenge for clinical adoption, where understanding the reasoning behind predictions is crucial. Future work could focus on developing more interpretable models or improving the transparency of complex models like DNNs.

## V.CONCLUSION

This study successfully demonstrates the potential of machine learning models, specifically Random Forest, Support Vector Machine (SVM), and Deep Neural Networks (DNN), in predicting the likelihood of brain stroke based on a combination of clinical and demographic features. Among the models tested, the DNN emerged as the most accurate, achieving an impressive accuracy of 96.3%. This high level of accuracy underscores the capability of deep learning techniques to capture complex patterns in data, making them particularly suitable for medical predictions. The study also identified critical features like age, glucose levels, and hypertension, which significantly contribute to stroke risk, aligning with existing medical knowledge. However, the study also highlights the challenge of balancing accuracy with interpretability, a critical consideration for practical applications in healthcare. While DNNs

DOI https://doi.org/10.15463/gfbm-mib-2024-257

provide superior predictive performance, their complexity can obscure the rationale behind predictions, potentially limiting their clinical adoption. Overall, the findings of this research affirm the promise of machine learning as a powerful tool for early stroke detection and prevention, offering healthcare providers valuable insights to guide patient care and intervention strategies. The integration of these predictive models into clinical practice could significantly enhance patient outcomes by enabling earlier and more targeted interventions.

## VI. FUTURE ENHANCEMENTS

To further improve the applicability and performance of the proposed models, several future enhancements are suggested. First, incorporating additional data sources such as genetic information, lifestyle factors, or data from wearable devices could provide a more comprehensive understanding of stroke risk, thus enhancing model accuracy. Second, improving model interpretability is essential for clinical adoption. This could be achieved by refining Explainable AI (XAI) techniques or developing hybrid models that combine the accuracy of deep learning with the transparency of simpler models. Third, validating the models on larger and more diverse populations is crucial to ensure their generalizability across different demographic groups. Fourth, integrating these models into existing clinical workflows, perhaps through user-friendly interfaces or electronic health records (EHR) systems, could facilitate their adoption by healthcare professionals. Finally, implementing continuous learning mechanisms to update models as new data becomes available would help maintain their relevance and accuracy over time. By pursuing these enhancements, the predictive models can be made more robust, interpretable, and widely applicable, ultimately contributing to more effective stroke prevention and management in clinical practice.

## References

1. M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243-297, Dec. 2021, doi: 10.1016/j.inffus.2021.05.008.

**DOI https://doi.org/10.15463/gfbm-mib-2024-*257***

2. R. S. Ledbetter, "Predicting the occurrence of a stroke using machine learning," *Journal of Computational Science*, vol. 49, pp. 101308, Sept. 2020, doi: 10.1016/j.jocs.2020.101308.
3. Y. Jin, J. Li, and X. Zhao, "Stroke risk prediction using machine learning: A systematic review," *IEEE Access*, vol. 9, pp. 75089-75108, 2021, doi: 10.1109/ACCESS.2021.3085160.
4. Z. Zhou, C. Xie, Y. Yang, and L. Zhang, "Stroke prediction model for hypertension patients based on machine learning," *Frontiers in Public Health*, vol. 9, article 754292, Nov. 2021, doi: 10.3389/fpubh.2021.754292.
5. S. A. Saleh, S. Bouzouita, and M. Ksouri, "Predicting stroke using an ensemble of machine learning algorithms," *Artificial Intelligence in Medicine*, vol. 118, pp. 102125, Feb. 2022, doi: 10.1016/j.artmed.2021.102125.
6. N. N. Hossain, R. Nahiduzzaman, and M. N. Hossain, "Prediction of brain stroke using machine learning algorithms: An exploratory study," in *2022 International Conference on Intelligent Systems, Data Science and Automation Engineering (ICIDSDAE)*, 2022, pp. 1-6, doi: 10.1109/ICIDSDAE56583.2022.10064792.
7. A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, G. E. Marcus, A. Schroff, M. Shlens, S. J. Small, D. A. Shickel, and M. Amodei, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 1, pp. 1-10, May 2018, doi: 10.1038/s41746-018-0029-1.
8. Y. Wang, Y. Ma, W. Zhou, M. Lin, Z. Cai, and X. Liu, "Deep learning-based early warning of stroke in hospitalized patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1611-1620, May 2021, doi: 10.1109/JBHI.2021.3059323.
9. M. Wang, J. Yan, W. Lu, and C. C. K. Cheng, "Stroke risk prediction based on machine learning: An application to Oulu cohort study," *Journal of Stroke and Cerebrovascular Diseases*, vol. 30, no. 10, pp. 105916, Oct. 2021, doi: 10.1016/j.jstrokecerebrovasdis.2021.105916.
10. R. Sarangi, K. K. Dutta, and P. Meher, "Analysis of machine learning techniques for stroke prediction using health records," *Procedia Computer Science*, vol. 172, pp. 574-582, 2020, doi: 10.1016/j.procs.2020.05.091.