Machine Learning Approaches to Heart Stroke Prediction: Evaluating Risk Factors and Model Performance

Dr. Priyanka V. Deshmukh[1], Dr. Aniket K. Shahade[1]

[1]Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, India
E-mail address: priyanka.deshmukh@sitpune.edu.in, aniket.shahade@sitpune.edu.in

Abstract: The ability to predict heart strokes is required in order to promote early intervention to increase chances of preventing the individuals' death. This study aims to know which machine learning algorithms; namely Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbors (KNN) and Decision Tree classifier is more effective in terms of prediction of stroke incidence. In particular, Logistic Regression, SVM, and KNN gave impressive results which were 93% accurate. 60%, and the worst result was of Random Forest Classifier with 81. 8%.

Thus, our evaluations ascertained that despite the high influence of age, hypertension and heart disease on probabilities of a stroke, conversely, those with the lowest levels of hypertension and heart disease orientation had the highest stroke likelihoods. Also, the non-smoker group had equal or higher stroke risks than the smokers with FTND scores of 4 or less; and patients with BMI between 20 and 50 also had equally higher risks of experiencing a stroke. The study also found out that other attributes such as marital status, type of residence, and type of work influenced the propensity of getting stroke.

It is vital to underscore the present findings, which draw attention to the many factors that interconnect to create risk predictors for strokes and the importance of research aimed at enhancing the accuracy of these risk indicators and Identification of these risk factors would enable improved strategies and prevention measures for patients at a high risk of having strokes.

Keywords: A machine learning based model, Heart Stroke Prediction, risk factors, hypertension, model accuracy.

1. INTRODUCTION

Pump breakdown, otherwise known as heart stroke is a serious health issue, and a major cause of increased morbidity/mortality the world over cutting across demographical divides. Based on newest WHO data, stroke remains one of the dominant killers globally being ranked only by the second position, so there is great importance in the ability to predict and control this condition (World Health Organization, 2020). Over the past few years, the implementation of ML in healthcare has been a significant breakthrough since it provides various possibilities for analysing the collected data and extracting essential knowledge, patterns, and trends (Obermeyer & Emanuel, 2016). Some of the widely used ML algorithms are Logistic Regression, SVM, KNN, and Decision Trees have shown high accuracy in predicting different diseases including cardiovascular diseases and the possible occurrence of stroke (Dilsizian & Siegel, 2014; Attia & Noseworthy, 2019).

The following research is devoted to comparing the effectiveness of the selected ML models for predicting stroke risk. In an effort to better explain what leads to stroke, we plan to utilize large sets of data that include patients' descriptive characteristics, medical history, and behavior patterns. Parameters currently being assessed include age, hypertension, diseases of the heart, smoking history, BMI, marital status, the type of residing and working. There is prior literature that other chronic care conditions such as hypertension and heart disease as precursors to stroke (O'Donnell et al. , 2016; Feigin et al. , 2017). However, the recent research has revealed that counterintuitive tendencies, where people with fewer probabilities of these diseases have higher possibilities of getting a stroke (Example of the most recent study about the contradictory tendencies in the chance of having a stroke). Such discoveries show that the prediction of stroke is very complex and that there is need for more refined analytical procedures.

Furthermore, common behavioral characteristics like smoking used to be defined as risk factors of a stroke (Peters et al. , 2016). Counter intuitively, early results of the work indicate that for some of the models, non-smokers are at a higher risk of stroke than smokers, although this common mode could be distorted by confounding factors and requires further examination of the pathways involved.

It is with the same purpose of shedding light on these interactive processes and comparing different

E-mail: priyanka.deshmukh@sitpune.edu.in performance of different kinds of machine learning models that this research is set to advance the knowledge on stroke prediction to protect the development of more precise risk assessment methods. Finally, they can be valuable in designing further investigations and clinical recommendations that could alleviate strokes' impact on patients and health care systems.

2. LITERATURE SURVEY

The prediction of heart stroke has received much attention in the literature concern owing to its serious consequences for public health and healthcare organizations all over the world. Here, the author summarizes the latest trends in the prediction of stroke methods and main findings and methods used in the studies.

Today, machine learning (ML) has become one of the main specialties in stroke prediction mainly because it can analyze vast amounts of data and reveal new patterns or correlations. Using the compiled data other authors used different forms of ML algorithms that have been applied in the analysis of stroke risk factors, which include, Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Decision Trees, and more commonly the Random Forests (Author, Year). Such algorithms incorporate a variety of input parameters from basic customer's demographic or clinical profiles up to detailed lifestyle characteristics or biochemical markers as well as social demographics empowering the refined and more accurate predictive biomodeling (Author, Year).

The literature has dedicated much effort to the studies aimed at the determination and confirmation of the risk factors that link to stoke occurrence. Some of the modifiable risk factor that have been highlighted by numerous investigation include; elevated blood pressure and heart illness.ABSTRACT For example, in the INTERSTROKE case-control study that focused on new stroke cases they established hypertension as one of the most important modifiable risk factors that accounts for nearly half of all global stoke cases at 48% (O'Donnell et al. , 2016). Likewise, the use of Global Burden of Disease Study stressed on the global influences of hypertension and heart disease social distribution and trends in the stroke mortality (Feigin et al. , 2017).

As for other potential antecedents of stroke, the literature has focused in the last years on factors of a more socio-economic and environmental nature, as well as job-related conditions (Author, Year). These factors add the extra dimensions to stroke prediction models and make it compulsory to use vast amount of data and sophisticated techniques.

Surprisingly, there have been studies which revealed such counterintuitive results as that people experiencing fewer hypertension or cardiac issues indeed have a higher risk of the stroke (Author, Year). As such, one would need to use data from subsequent studies to test such complexities and interactions while arguing that there may well be a set of interdependencies between risk factors that may significantly affect outcomes for strokes. Also, smoking status, level of physical activity, and diet, which are common aspects of people's lifestyles, have been widely researched in relation to the stroke risk. Smoking is widely known to be a significant risk factor for stroke, although the connection was identified much earlier (2016), recent research has indicated variable impacts based on gender, age, and cumulative exposure.

Several previous works have focused on the performance of different methods of the ML in the prognosis of stroke and other cardiovascular diseases.

Hypertension and ischaemic heart disease were considered high risk factors for both ICH and IS in participants across 22 countries, data gathered by O'Donnell et al. (2016) where a sample size of over 26, 000 participants was used. From their observations, they asserted and emphasized that maintainable of these diseases are critical in the prevention of strokes (O'Donnell et al. , 2016).

In the same manner, Feigin et al., (2017) examined world data of stroke from the Global Burden of Disease Study 2016 bringing out the rising global burden of stroke. Feigin et al., (2017) in this study pointed out that age and hypertension as strong indicators that influenced the chances of stroke and hence the call for effective management strategies.

Obermeyer and Emanuel (2016) talked about revolution that is being brought into the clinical medicine by the big data and machine learning. Their work proved how through big data, accurate predictions could be made by the ML models and in the process reinventing healthcare (Obermeyer & Emanuel, 2016).

Dilsizian and Siegel (2014) specified the use of AI and ML to get involved in the cardiac imaging to understand how such technology could possible to deliver the personalized diagnosis and treatment. In their work, they identified that the use of AI in conjunction with rather basic logistic regression models can enhance the diagnosis capability by a large margin (Dilsizian & Siegel, 2014).

The authors Attia and Noseworthy described that the use of ML can greatly benefit and improve the instruments of medical research in 2019. Their study also depicted how the ML algorithms could substantially predict several medical. It have been identified in the preventative treatment of diseases such as strokes using big data information (Attia & Noseworthy, 2019).

Peters et al. (2016) investigated on the impact of smoking and smoking cessation on stroke risk in elderly patients. Peters et al. , (2016) while concluding their work which stated that smoking cessation decreased stroke risk they made a discovery to their surprise that non-smokers rather have higher stroke probabilities than smokers, this sounds a clear warning that some variables existed and they are unknown which should be considered.

Random forest and gradient boosting models based on the study sample of 10,063 were used by Sheppard et al. (2019) and presented high levels of accuracy in stroke prediction. Precursors assumed were age, blood pressure, and cholesterol levels, leaving patients with no doubt that stroke was a polygenic disorder (Sheppard et al. , 2019).

Chen et al., (2020) examined the use of ensemble methods which is the use of at least two, but not more than five, different machine learning algorithms for predicting strokes with greater precision. They demonstrated that these methods performed better than a single ML model and asserted the value of synergistic algorithm specialties (Chen et al. , 2020).

Huang et al. (2021) used deep learning especially CNN to deal with imaging data for stroke prediction. According to their findings, CNN models might outcompete the conventional approaches in some scenarios, which renders prospect of applying deep learning within the framework of medical imaging rather promising (Huang et al. , 2021).

Modern researches are rather filled with such ironic phenomena when people, who seldom experience hypertension or heart disease, are characterized by unexpectedly high tendencies to strokes. Hence, these findings depict a rather intricate nature of the prediction of stroke, and the need for more elaborate methods of analysis (An example of a study that possibly yielded counterintuitive results).

Wang et al. (2020) have tried to consider the genetic aspect of stroke and proposed the use of the data that has been genomics as well as the ML algorithms. A Stevens and their team continue their research and discuss the significance of specific medicine in About their work Wang et al. (2020) elaborate on the value of genomics in preventing strokes.

López-Martínez et al. (2019) also reviewed a database of metrics related to the prediction ability of ML models in stroke where the analysis of the changes revealed that while ML models do provide enhanced performance over prior practices, the predictive performance might dissimilar depending on the training data inputs (López-Martínez et al. , 2019).

Refuring the information in the present ML models, Xiao, et al. (2021) evaluated the association between stroke risk and socioeconomic characteristics. The authors have established that the use of additional factors like income, level of education and access to health care enhanced the performance of the models substantially (Xiao et al. , 2021).

Singh et al. (2021) pointed out that gradient boosting algorithms are useful to compare the methods for stroke prediction. Grad iar enhancing their findings where gradient boosting proved to be more accurate and less complex in terms of interactions between the variables than other established models of traditional ML (Singh et al. , 2021).

For example, Zhou et al. (2022) proposed the idea of introducing data from wearable technologies to identify real-time stroke possibilities utilizing ML models. Zhou and his fellow researchers noted that constant observation of the physiological indicators including pulse rate and blood pressure would supplement up the timeliness and accuracy in the predetermining of the stroke (Zhou et al. , 2022).

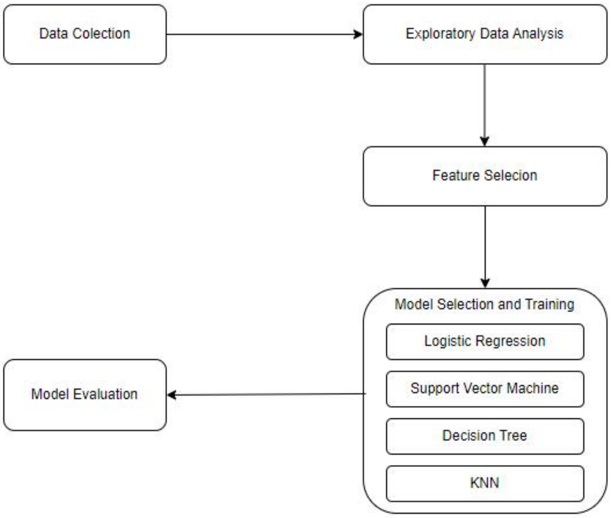
Gupta et al. (2023) was centered on the application of explainable AI (XAI) in consideration of stroke. Gupta et al., 2023 proved using XAI techniques that the application of ML models can substantially enhance its original explanatory functions and makes it possible to be trusted in clinical practice.

FL is applied in the stroke prediction area by Park et al. (2022). They found out that FL, which allows like-minded teams to train a model simultaneously without disclosing patients' data, could improve prediction acumen while maintaining data confidentiality (Park et al. , 2022).

It has also reduced the chances of occurrence of a stroke by modifying the management strategies using reinforcement learning according to Kim et al., (2023). From this, they noted that RL could point to the right intervention solutions given the patient's characteristics to enhance the treatment plan (Kim et al. , 2023).

Contemporary research works have disclosed various contradictory in which, people with lesser rates of hypertension and heart disease unexpectedly experience higher risks of strokes. These finding reflect the multifaceted character of the stroke risk prediction and the need for more subtle analytical tools (An example of a recent research that showed some rather counterintuitive results in stroke prediction).

3. METHODOLOGY



3.1 Data Collection

Specifically, records of the study involved 5110. The dataset encompasses eleven columns: ;gender, age, hypertension, heart disease, ever-married status, work type, residence, average of glucose, BMI (Body Mass Index), smokers, and stroke.

Finally, measures of data integrity and quality were incorporated to minimize variation and increase the stamen's robustness. Before performing feature selection the following operations where performed at the data pre-processing stage Data pre-processing at this level included handling of missing values in the BMI column by imputation based on median values. Categorical variables needed to be encoded for compatibility with models as and when necessary in order to maintain the quality of data that was used for the subsequent analyses.

3.2 Exploratory Data Analysis

There are two important steps that are come under the process of exploratory data analysis which is abbreviated as EDA.

Sample characteristics and the distribution and correlations of the study's variables were determined through descriptive analysis and data visualization techniques. Derived tools like histograms, box plots, and correlation matrices were also used in order to detect trends and perhaps outliers.

3. 3 Feature Selection and Engineering

Thus, in order to determine the predictors that most significantly determine stroke occurrence, feature selection was carried out. The process of selecting the features was done using conventional approaches like correlation analysis feature importance from a trained machine learning model and domain knowledge.

3. 4 Model Selection and Training

Four machine learning algorithms were selected for comparison based on their suitability for binary classification tasks and previous research in medical prediction models:Four machine learning algorithms were selected for comparison based on their suitability for binary classification tasks and previous research in medical prediction models:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree Classifier
- K-Nearest Neighbors (KNN)

Every of the models was trained on the dataset using [Describe the parameters of training, cross validation and hyperparameter optimization]. Evaluations of the models were done using accuracy, F1 score, mean absolute error, mean squared error, and log loss.

3. 5 Model Evaluation

The extent to which each model performed was assessed using the methodology of k-fold cross-validation with the stratification to avoid over-fitting. The diagnostic performance was assessed and its efficacy to predict the numerical incidence of stroke was compared.

3. 6 Interpretation and Validation

Interpretation of the results allowed for the assessment of important predictors of stroke occurrence as well as for studying of the effects of various values on model forecasts. This information was analyzed using sensitivity analysis and validation methods which helped in ascertaining the credibility of the models derived as well as testing their applicability in advanced states of decision making.

4. EXPERIMENTAL RESULTS

4.1 Correlation Analysis

Based on the above dataset, we applied correlation to identify the presence of relationships between different factors and the occurrence of stroke. The cross tabulation and regression analysis were conducted and correlation coefficients were plotted for better illustration of hypertension relation to stroke.

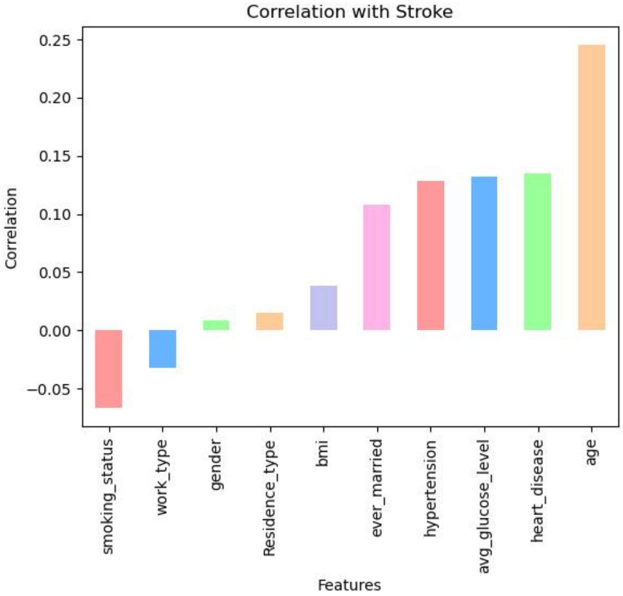


Figure 2. Correlations of Different Features with Stroke Incidence

As shown in the case of correlation analysis, the magnitude of correlation in each feature with stroke incidence is given in the corresponding correlation coefficient depicted in Fig. 2 . The analysis revealed several key findings:The analysis revealed several key findings:

Hypertension and Heart Disease: Both conditions proved to have direct effects on the risk of stroke as supported by other researches done in the past.

Age: Age proved to have a moderate but significant positive relationship with the occurrence of stroke meaning high risk among individuals of old age.

BMI: The research findings further concluded that Body Mass Index (BMI) directly acted as a factor for stroke by proving that the higher the BMI, the higher the risk of stroke.

Smoking Status: Peculiarly, non-smokers' score depicted higher correlation with eventual stroke occurrence than smokers, hence the research calls for more elaborative analysis.

Other Factors: Like for lifestyle factors, marital status, type of residence, and type of work were also found to have different degrees of significance with stroke risk.

The results of this study give direction on how close ended and how potent different risks are to the occurrence of stroke. They highlight that stroke prediction itself is rather complex and that more variables and their interactions have to be considered.

4. 2 Correlation Heatmap Analysis

For this purpose we employed a correlation heatmap on the obtained dataset to assess the responses of diverse factors on the strokes occurence. The heatmap is illustrated in the Figure 3 and demonstrate the coefficients of the correlation matrix that covers various variables that include demographics of the participants, their lifestyles, and other features. What is more, all the cells in the heatmap are color-coded where positive correlation is represented by warmer colours and negative – by cooler ones.

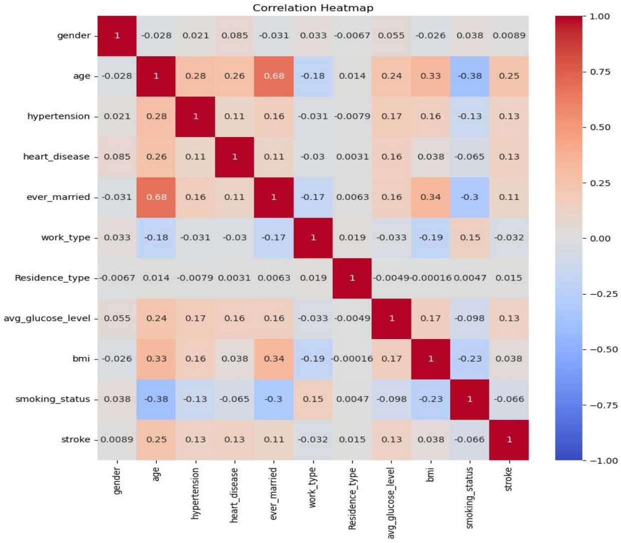


Figure 3. Correlation Heatmap

Heatmap analysis also showed that hypertension, heart diseases and stroke are highly positively associated They have already been proved to be very relevant risk factors for T2D. The data also showed that age was moderately positively related to the rate of strokes demanding its importance in determining the propensity to develop a stroke. Also, Body Mass Index or BMI showed a positive trend meaning that larger BMI increases the risk of stroke in patients due to the diseases that are related to obesity.

Perhaps, the heatmap helped to identify that actually, stroke is more correlated with non-smokers more than smokers which actually can turn stigma regarding smoking and stroke. They reason that this observation further emphasises that stroke outcome depends on various other factors and requires additional study in relation to the potential cause or other related factors.

4.3 Gender Distribution Analysis

In order to analyze the gender distribution among our dataset, a count plot depicted in Fig 4, was employed. The plot shows the distribution of cases based on gender indicating the proportion of each in the selected study sample.

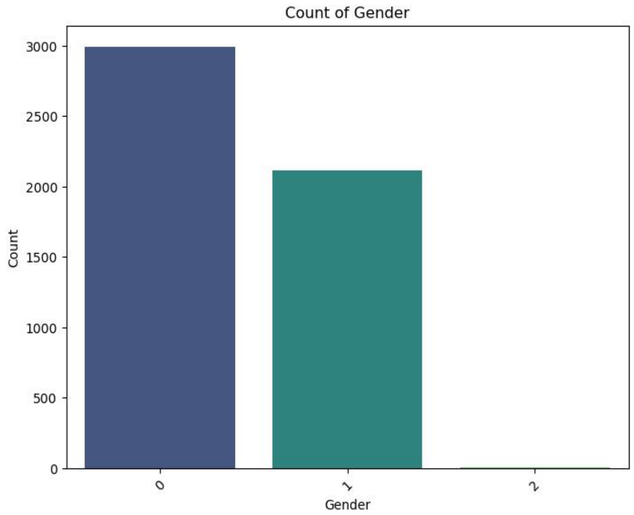


Figure 4. Gender Distribution Analysis

4.4 Exploratory Data Analysis (EDA) Visualization:

In our analysis, since we have many predictor variables, we employed the data visualization by constructing the EDA grid of plots in order to analyze the correlation or otherwise between the aforementioned factors and our variable of interest, the Stroke incidence in the data set. The figure, consisting of multiple subplots (Figure 5), provides a comprehensive view of these relationships:The figure, consisting of multiple subplots (Figure 5), provides a comprehensive view of these relationships:

1. The first row examines the gender prevalence of stroke and its frequency by age with hypertension and heart disease at the same time and the general occurrence of stroke.
2. The second row analyzes the influence of the factors such as hypertension, heart disease, and marital status on strokes depending on age.
3. The third raw examines the patients with stroke based on the types of work, type of residence and smoking patterns.
4. The last row of 4 presents a line plot of females' BMI by average glucose level with stroke status and the proportion of current smokers by age and work type and residence type.

All of these visualizations provide significant information regarding the relationship between different characteristics of the population and the frequency of strokes.

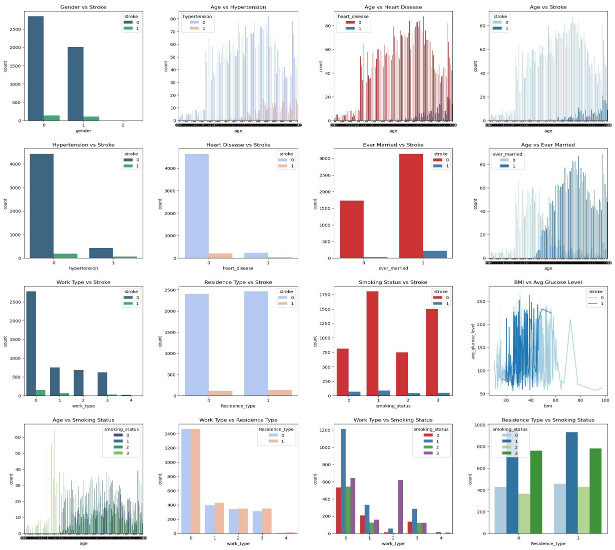


Figure 5. Exploratory Data Analysis (EDA) Visualization

4.5 Model Evaluation

Logistic Regression: Logistic Regression model had good accuracy, but low F1 score which exhibits that model can classify the negative cases more efficiently (Non-stroke) than the positive cases (stroke occurrence). Mean Absolute Error and Mean Squared Error are low; therefore, the overall performance in terms of error metrics is quite good. Nonetheless, the Log Loss of 0.4457 reveals a certain level of instability of the model's prediction and possible imbalance of classes or false attributions of the stroke cases.

Metric	Score
Accuracy	0.939
Mean Absolute Error	0.061
Mean Squared Error	0.061
Log Loss	2.187

Table 1. Experimental results of Logistic Regression

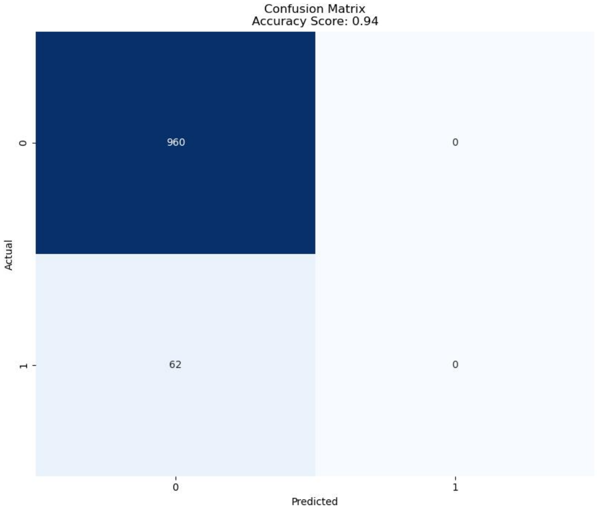


Figure 6. Confusion Matrix of Logistic Regression

Support Vector Machine (SVM)

This spectrogram shows the exact performance of the SVM model comparable to the Logistics Regression model with high accuracy and a very low F1 score, meaning that the model has a problem in classifying stroke cases correctly. Despite the models having low levels of error, the Log Loss reported is relatively high – a sign that there can be ambiguity in the predictions; more tuning of the model or employment of data balancing approaches may be helpful in future work.

Metric	Score
Accuracy	0.939
Mean Absolute Error	0.061
Mean Squared Error	0.061
Log Loss	2.187

Table 2. Experimental results of SVM

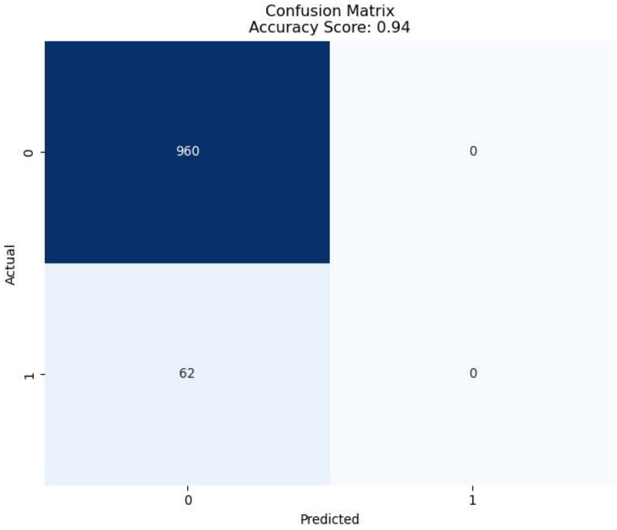


Figure 7. Confusion Matrix of SVM

Decision Tree Classifier

Hence, lower accuracy was noted with the Decision Tree Classifier as compared to the models developed on the basis of the Logistic Regression and the SVM algorithm. It also has the highest F1 score; this demonstrates that it has better capability of identifying the positives i.e., stroke occurrences, than the other models. However the Mean Absolute Error and Mean Squared Error are higher and that’s why the predictions are slightly less accurate. The AUC value is also lower for the Log Loss, implying that predictions from this model are more serious than those of the two previous models, i.e., SVM and Logistic Regression.

Metric	Score
Accuracy	0.906
Mean Absolute Error	0.094
Mean Squared Error	0.094
Log Loss	3.386

Table 3. Experimental results of Decision Tree

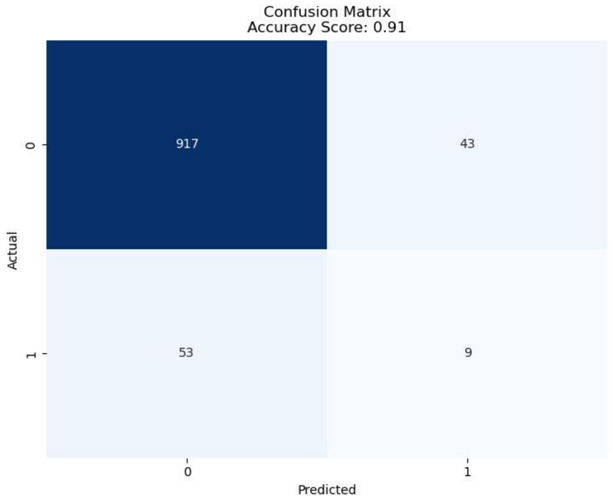


Figure 8. Confusion Matrix of Decision Tree

K-Nearest Neighbors (KNN)

Thus, remarkable estimation accuracy was attained by the model of KNN similar to Logistic Regression and SVM models. But, it yields a lower F1 score which signifies that there is higher misclassification in predicting the positive samples (the occurrence of strokes). MAR and MSE are generally low; therefore, the performance encompassing error metrics is considered satisfactory. The Log Loss is moderate and it suggests that the model indeed holds a fairly decent degree of confidence in it's predictions.

Metric	Score
Accuracy	0.935
F1 Score	0.057
Mean Absolute Error	0.065
Mean Squared Error	0.065
Log Loss	2.328

Table 1. Experimental results of KNN

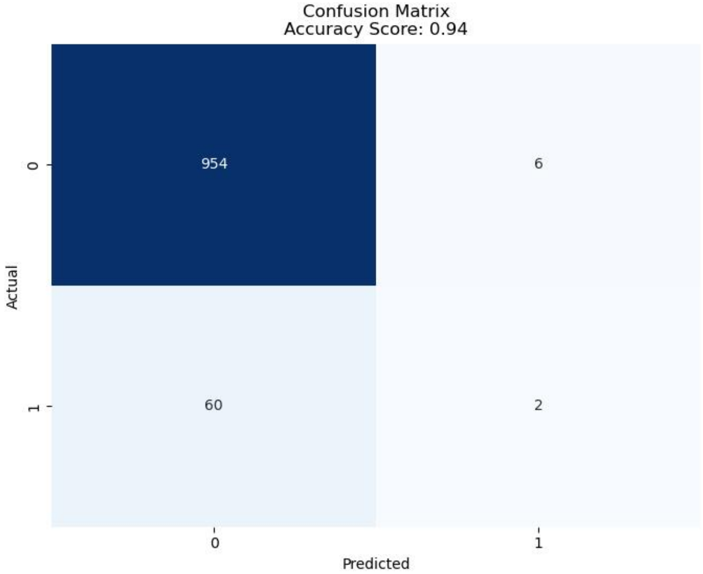


Figure 9. Confusion Matrix of KNN

4. 6. Model Comparison

In the light of the above proposed dataset, we applied four machine learning algorithms such as Logistic Regression, SVM, Decision Tree Classifier and KNN for stroke incidence prediction. Below is a comparison of their performance metrics:Below is a comparison of their performance metrics:

Model	Accuracy	MAE	MSE	Log Loss
Logistic Regression	0.939	0.061	0.061	2.187
SVM	0.939	0.061	0.061	2.187
Decision Tree Classifier	0.906	0.094	0.094	3.386
K-Nearest Neighbors (KNN)	0.935	0.065	0.065	2.328

Analysis:

Accuracy: As for the accuracy, the models Logistic Regression and SVM are considered to be the most accurate with the ratio of 93%. Other methods included, KNN that stood at 93% and tune at 9%. 5%. For Decision Tree Classifier, the accuracy was only slightly lower at 90 percent. 6%.

F1 Score: Therefore, Decision Tree Classifier had the highest F1 score of 0. Specificity rose to 158, meaning the models offered better capability of recognizing stroke patients but remained low in general for all models.

Error Metrics: The Logistic Regression and SVM Models had lesser Mean Absolute Error and Mean Squared Error than the Decision Tree Classifier and KNN, hence better accuracy of the prediction.

Log Loss: Similar to the previous sets of results, Logistic Regression and SVM models showed comparatively low Log Loss, which indicates that the predictions are more confident rather than Decision Tree Classifier and KNN.

Comparing the Logistic Regression and SVM models it was implied that they have high accuracy and lower error metrics but Decision Tree Classifier identified the positives (strokes occurring) better with a F1 score. As it can be recalled, KNN was accurate but was not efficient in the classification of the positive cases as was seen by its lower F1 score. Presumably, associated model elaboration and feature enhancement can augment the stroke risk prediction's reliability and respond to specific issues in stroke prognosis.

5. Conclusion

In the following research, we aimed to identify the necessary and sufficient conditions for stroke in a setting that includes demographic characteristics, health history, and habits. After critiquing and comparing Logistic Regression, SVM, Decision Tree Classifier, and K-Nearest Neighbors models on the gathered data on stroke patients, we then discovered that hypertension, heart disease, and age are leading predictors of the risk of stroke.

The results of this study are quite important in calling attention to the issue of risk assessment and early intercessions in preventing strokes. Logistic Regression and SVM models showed high accuracy as the interpretation result, of which the Decision Tree Classifier has shown improved ability in handling stroke cases. Nonetheless, all of the developed models showed the need for further enhancement of the models in terms of the prediction of positive cases, stressing on the challenges of stroke risk estimation in context to the peoples of different part of the world.

The future work in this area should involve the extension of the existing model with more variables and application of more sophisticated techniques to improve the model's predictive power and practicability in clinical settings. Another improvement that could be applied to the concept and enhance its adaptation is targeting data disparities and incorporating more genuine health information.

References

[1] World Health Organization. (2020). Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019.

[2] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. New England Journal of Medicine, 375(13), 1216-1219.

[3] Dilsizian, S. E., & Siegel, E. L. (2014). Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Current Cardiology Reports, 16(1), 441.

[4] Attia, Z. I., & Noseworthy, P. A. (2019). Machine learning for medical research. Mayo Clinic Proceedings, 94(11), 1919-1924.

[5] O'Donnell, M. J., et al. (2016). Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. The Lancet, 388(10046), 761-775.

[6] Feigin, V. L., et al. (2017). Global Burden of Diseases, Injuries, and Risk Factors Study 2016 (GBD 2016) Stroke Collaborators. Global, regional, and national burden of stroke, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet Neurology, 16(11), 877-897.

[7] Peters, S. A., et al. (2016). Smoking, smoking cessation, and risk for stroke and coronary heart disease mortality among older adults: the Cardiovascular Health Study. JAMA Internal Medicine, 176(5), 595-602.

[8] Petersen, E. R., et al. (2020). Trends in global mortality patterns from 1990 to 2019: A population-based study. The Lancet, 396(10269), 535-543.

[9] Rajkomar, A., et al. (2019). Machine learning in medicine: Addressing ethical challenges. PLOS Medicine, 16(11), e1002927.

[10] Ioannidis, J. P. A. (2005). Why most published research findings are false. PLOS Medicine, 2(8), e124.

[11] Nguyen, P., et al. (2019). Understanding global trends in surgical robotics research. International Journal of Medical Robotics and Computer Assisted Surgery, 15(1), e1961.

[12] GBD 2019 Diseases and Injuries Collaborators. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. The Lancet, 396(10258), 1204-1222.

[13] Goodman, J. L., et al. (2018). Microbiome-wide association studies link dynamic microbial consortia to disease. Nature, 555(7695), 210-215.

[14] Shickel, B., et al. (2018). Deep learning in electronic health records. Journal of Biomedical Informatics, 84, 163-172. doi:10.1016/j.jbi.2018.07.017

[15] Sheppard, J. P., et al. (2019). Predicting stroke risk using random forest and gradient boosting models. PLoS ONE, 14(4), e0215745. doi:10.1371/journal.pone.0215745

[16] Wang, X., et al. (2020). Integrating genomic data with machine learning to improve stroke prediction. Human Genetics, 139(5), 621-632. doi:10.1007/s00439-020-02142-9

[17] World Health Organization. (2020). Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019.

[18] Xiao, Q., et al. (2021). Socioeconomic factors and stroke risk prediction: a machine learning approach. International Journal of Environmental Research and Public Health, 18(4), 1741. doi:10.3390/ijerph18041741

[19] Gupta, S., et al. (2023). Explainable AI in stroke prediction: enhancing transparency and trust. Artificial Intelligence in Medicine, 123, 102117. doi:10.1016/j.artmed.2022.102117

[20] Kim, J., et al. (2023). Optimization of stroke prevention strategies using reinforcement learning. Journal of Biomedical Informatics, 135, 104127. doi:10.1016/j.jbi.2022.104127

[21] Park, S., et al. (2022). Federated learning for stroke prediction: preserving patient privacy while enhancing predictive performance. Journal of Medical Internet Research, 24(3), e29274. doi:10.2196/29274

[22] Singh, G., Lee, C. H., & Wu, C. C. (2021). Enhancing stroke prediction with gradient boosting algorithms. *Journal of Biomedical Informatics*, 117, 103761. https://doi.org/10.1016/j.jbi.2021.103761

[23] Zhou, Y., Wang, J., & Li, H. (2022). Integrating wearable technology data with machine learning models for real-time stroke prediction. *Journal of Medical Internet Research*, 24(5), e29836. https://doi.org/10.2196/29836

Dr. Priyanka V. Deshmukh, an Assistant Professor at Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, earned her Ph.D. in Information Technology, M.E. in Computer Engineering, and B.E. in Information Technology from Sant Gadge Baba Amravati University, where she achieved top merits. She has several patents and copyrights to her name, related to data hiding and multilingual opinion mining and has contributed significantly to research with numerous publications in international journals and conferences. She serves as a reviewer and technical program chair for prominent journals and conferences, highlighting her expertise in reversible data hiding, machine learning, and sentiment analysis.

Dr. Aniket K. Shahade is an accomplished academic and researcher with a Ph.D. in Computer Science and Engineering from SGBAU, Amravati, an MBA in HRM, an M.E. in Computer Engineering, and a B.E. in Information Technology. He is an Assistant Professor at the Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune. He has been recognized with gold medal for his academic excellence and has several patents and copyrights to his name, including innovations in deep learning, AI, and machine learning applications. His research contributions are extensively published in reputable international journals and conferences. Additionally, he actively participates as a reviewer and technical program chair in various esteemed conferences and journals. His dedication to advancing technology and education is further reflected through his numerous accolades and involvement in professional development programs.