ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 6 June 2024

An Effective Machine Learning Model for Stroke Prediction System

[1] Anoop Yadav, [2] Puneet, [3] Avirat Anant, [4] Pranav Kumar, [5] Aman Singh, [6] Priyanshu Kumar Yadav

[1] [2] [3] [4] [5] [6] Lovely Professional University, India

Corresponding Author Email: [1] vyadav99x1@gmail.com, [2] puneet.30430@lpu.co.in, [3] aviratanant595@gmail.com, [4] pranavkumar12384@gmail.com, [5] amanrathode123@gmail.com, [6] priyanshuyadav39976@gmail.com

***Abstract*— The brain stroke prediction project endeavours to construct an effective machine learning model for anticipating the likelihood of stroke occurrence based on diverse demographic, lifestyle, and health-related attributes. Utilizing a dataset encompassing features such as age, gender, hypertension, heart disease, average glucose level, body mass index, smoking status, among others, seven distinct classification algorithms were trained and assessed. Notably, both Logistic Regression and Support Vector Classifier achieved an accuracy score of approximately 97.16 percentage. The Random Forest Classifier emerged as the most promising model with the highest accuracy of around 97.2 percentage. This project underscores the potential of machine learning in facilitating early detection and intervention strategies to mitigate stroke risks effectively within clinical settings.**

***Index Terms*— Brain stroke prediction, machine learning, logistic regression, support vector classifier, random forest classifier, accuracy assessment, early detection.**

I. INTRODUCTION

Stroke is a major worldwide health issue which causes serious challenges with regard to the rates of death and disability. Stroke is the most common cause of long-term disability, which involves a sudden cessation of blood flow to the brain that causes serious neurological deficits. Even with advances in medical research, minimizing the negative effects of stroke still depends on early detection [15] and treatment. Stroke risk assessment has always been based on known variables including age, diabetes, and hypertension. To improve the precision and accuracy of stroke prediction models, however, new developments in machine learning and healthcare analytics provide encouraging ways forward. Machine learning algorithms are able to reveal complex patterns and associations by analysing large datasets that include a variety of demographic, lifestyle, and clinical information. This allows us for more accurate prediction of a person's risk of brain stroke so that people can take care of their self's in earlier stage.

Healthcare analytics can be revolutionized by machine learning algorithms, a branch of artificial intelligence that can identify intricate relationships across vast and diverse datasets. When compared to traditional risk assessment methods alone, these algorithms can predict an individual's likelihood of having a stroke with superior accuracy by analysing a wide range of factors, including age, gender, heart disease, hypertension, average glucose level, body mass index (BMI), and smoking status, among others. In this study, we use a large and diverse dataset that includes a range of demographic, lifestyle, and health-related factors to conduct a thorough investigation of machine learning-based stroke prediction.

Our approach entails the utilization of multiple classification algorithms, including Logistic Regression, Support Vector Classifier (SVC), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC), AdaBoost Classifier (ABC), and K-Nearest Neighbours Classifier (KNC). Each of these algorithms offers unique advantages and considerations in terms of model complexity, interpretability, and predictive performance. Through rigorous evaluation and comparison of these models' performance in predicting stroke occurrence, our goal is to identify the most effective approach for early detection and intervention strategies.

Furthermore, our project delves into the assessment of traditional stroke risk factors vis-à-vis machine learning-based predictions, providing insights into the added value of incorporating a broader range of variables in stroke prediction. Through interdisciplinary collaboration between machine learning and healthcare domains, we aim to harness the power of data-driven insights to revolutionize stroke prevention and management strategies. By developing reliable and clinically relevant predictive models, we aspire to assist healthcare professionals in identifying individuals at heightened risk of stroke, thereby enabling timely interventions and personalized care plans to reduce the burden of stroke-related morbidity and mortality on a population scale.

In summary, this project embodies a concerted effort to leverage advanced machine learning techniques and comprehensive datasets to enhance our understanding of stroke risk factors and improve patient outcomes through early detection and intervention strategies. Through interdisciplinary collaboration and rigorous evaluation, we

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 6 June 2024

strive to pave the way for more effective stroke prevention and management approaches, ultimately contributing to improved public health outcomes and enhanced quality of life for individuals at risk of stroke.

II. LITERATURE REVIEW

Sirsat et al. [1] categorize machine learning approaches for stroke prediction, highlighting Support Vector Machine (SVM) as optimal in several instances and emphasizing the need for further investigation, particularly in stroke treatment prediction. Emon et al. [2] develop a weighted voting classifier for stroke prediction, demonstrating superior accuracy compared to individual classifiers. Sailasya and Kumari. [3] analyse the performance of machine learning algorithms in stroke prediction, with Naïve Bayes exhibiting the highest accuracy. Singh and Choudhary. [4] compare different methods for stroke prediction, achieving remarkable accuracy using decision tree algorithms and backpropagation neural networks. Dritsas and Trigka. [5] introduce a stacking method for stroke risk prediction, outperforming other methods with high accuracy and AUC values. Tusher et al. [6] propose a system for early brain stroke [13] prediction, demonstrating significant reliability and accuracy using various classification algorithms. Heo et al. [7] explore stroke outcome prediction using natural language processing-based machine learning of radiology reports, showcasing the superiority of deep learning algorithms. Puri et al. [8] focus on heart stroke prediction, identifying SVM as effective, especially when considering linear and quadratic decision boundaries. Arslan et al. [9] compare medical data mining approaches for predicting ischemic stroke, with SVM showing the best predictive performance. Srinivas et al. [10] discuss the potential of data mining techniques in healthcare, particularly in predicting heart attacks [11] using classification-based approaches, emphasizing the importance of discovering hidden information from healthcare data for effective decision-making.

In conclusion, the literature review provides insights into the diverse array of machine learning and data mining techniques employed in stroke and heart attack prediction within healthcare. Through the examination of ten research papers, it becomes evident that these methodologies hold significant promise in enhancing patient care by enabling early detection and accurate prediction [12] of cardiovascular events. Various algorithms, including Support Vector Machine (SVM), decision trees, Naïve Bayes, and deep learning models, have been explored across different studies, each showcasing their effectiveness in different aspects of prediction. SVM emerges as a consistent performer in stroke and heart attack prediction, while deep learning models, particularly convolutional neural networks (CNN), demonstrate notable success in outcome prediction based on radiology reports. The literature also underscores the importance of feature selection, data pre-processing, and model evaluation techniques in optimizing prediction accuracy and reliability. Overall, these findings underscore the transformative potential of machine learning and data mining techniques in revolutionizing healthcare by facilitating personalized treatment, early intervention, and improved patient outcomes in the realm of cardiovascular disease prediction [14]. However, further research is warranted to address challenges such as data heterogeneity and model interpretability, thereby advancing the practical application of these methodologies in clinical settings.

III. METHODOLOGY

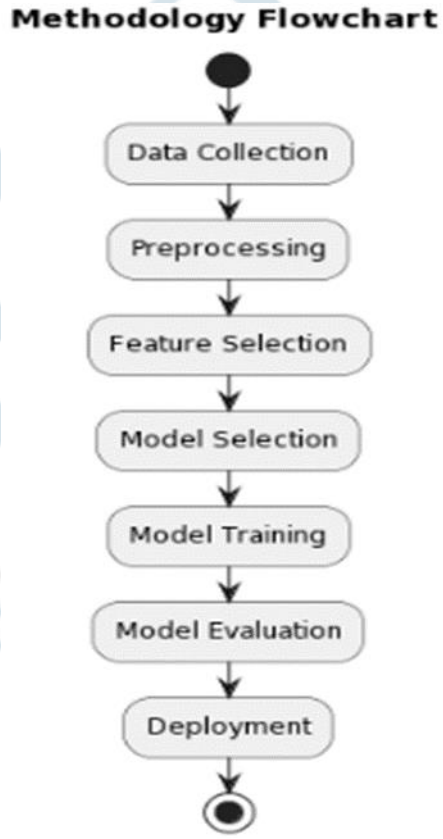

Fig. 1. Methodology of the project.

Let's go through every face of the project looking at what we have done in our project:

A. Data Collection

During the project's data gathering phase, we obtained a dataset that included information relevant to stroke prediction. With 5110 elements (rows) and 12 columns, the dataset is organized tabularly. Every row denotes a unique patient, and every column relates to a particular characteristic or property that is important for determining stroke risk. This dataset acts as a thorough archive of clinical, lifestyle, and demographic data, offering insightful information on variables affecting the risk of stroke.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 6 June 2024

Table I: Small Sample of Data Used

S.no	Gender	Age	Hypertension	Heart_disease	Ever_married	Work_type	Residence_type
1	Male	67	0	1	Yes	Private	Urban
2	Female	61	0	0	Yes	SE	Rural
3	Male	80	0	1	Yes	Private	Rural
4	Female	49	0	0	Yes	Private	Urban
5	Female	79	1	0	Yes	SE	Rural

The Table I gives as a sample data from the data set used, this table do not contain all the columns from the real dataset.SE indicates Self-Employed in the table.

B. Pre-Processing

In the pre-processing step of the project, we aimed to prepare the dataset for subsequent analysis and modelling by addressing various data quality issues and ensuring compatibility with machine learning algorithms. The pre-processing steps included handling missing values and encoding categorical variables.

First, we identified missing values in the dataset and decided to address them by removing rows with missing BMI values. This decision was made after observing that the BMI column had 201 missing values, which constituted a small proportion of the overall dataset as shown in Table II.

Table II: Finding The Missing Values

id	0
Gender	0
Age	0
Hypertension	0
Heart_disease	0
Ever_married	0
Work_type	0
Residence_type	0
Avg_glucose_level	0
Bmi	201
Smoking_status	0
Stroke	0
Dtype:int64	

Next, we encoded categorical variables using label encoding, which involves converting categorical values into numerical representations. This step ensures that categorical variables can be properly processed by machine learning algorithms.

By completing these pre-processing steps, we ensured that the dataset was clean, properly formatted, and ready for model training. This pre-processing lays the foundation for developing accurate and reliable stroke prediction models.

C. Feature Selection

In feature selection relevant features are to be selected from the data set by exploring the data set. We need to find the relevant features, connections between the target value and the features. To make this happen we did plot many types of graphs lets go through them.

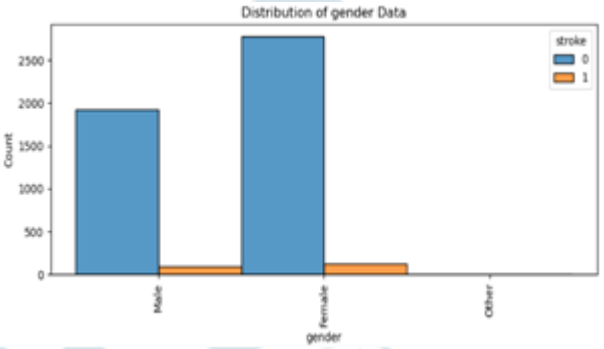

Fig. 2. Distribution of gender in the data set.

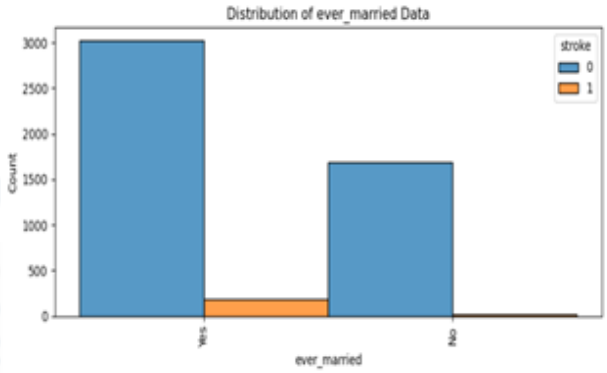

Fig. 3. Distribution of data for married and unmarried.

We found that Female are affected by stroke more than men's.

From figure 3. we can see that people who are married mostly suffer with stroke when compared with Unmarried people.

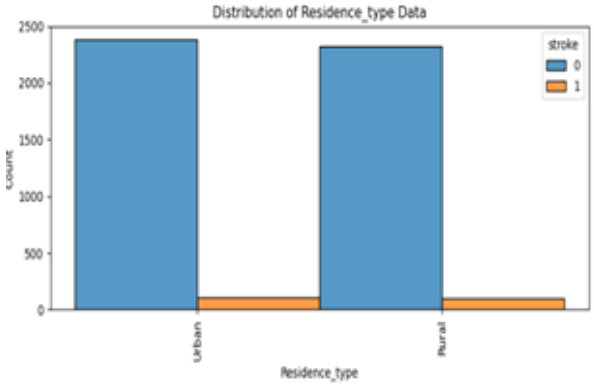

Fig. 4. Distribution of residence_type data.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 6 June 2024

From Figure 4. we were able to check the distribution of stroke in urban and rural people. We found that it is equally divided.

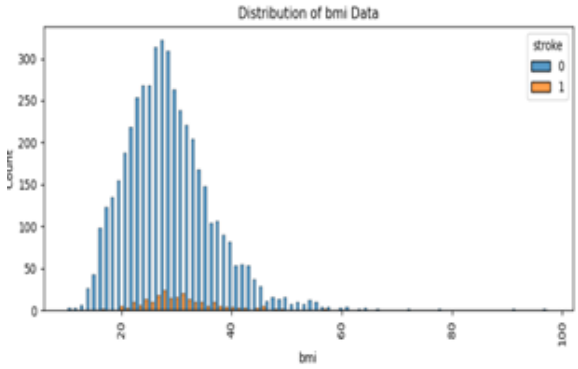

Fig. 5. Distribution of BMI Data.

We found that mostly strokes takes place from BMI value 20 to 40.

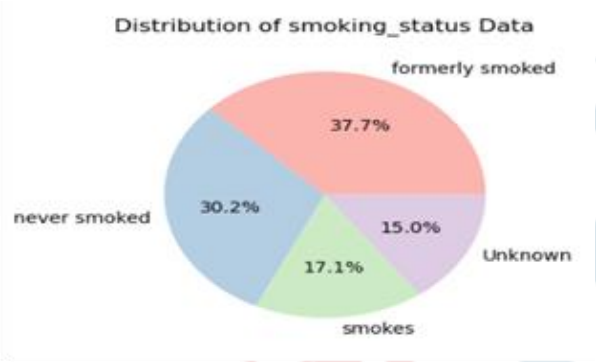

Fig. 6. Distribution of smoking status data.

This flow chart is the distribution of smokers and other features related to it. We found that most of the people formerly smoke.

Then we did check how is the division of data between of people suffering from stroke and people who are not. We found that there are 4700 people who are normal and 209 are suffering from stroke.

Then we have plotted scatter plots for checking the relation and outliers in the data The following graphs are which we did plot.

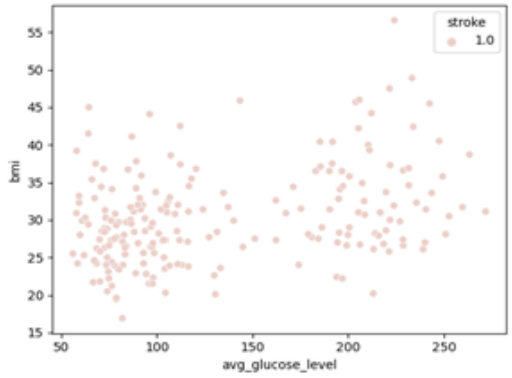

Fig. 7. Scatter plot between average glucose level and BMI.

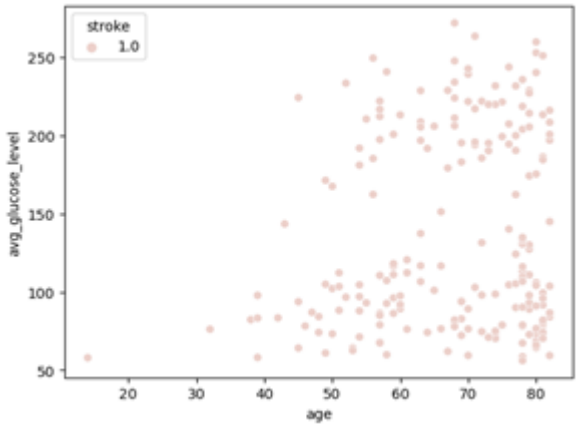

Fig. 8. Scatter plot between age and average glucose level.

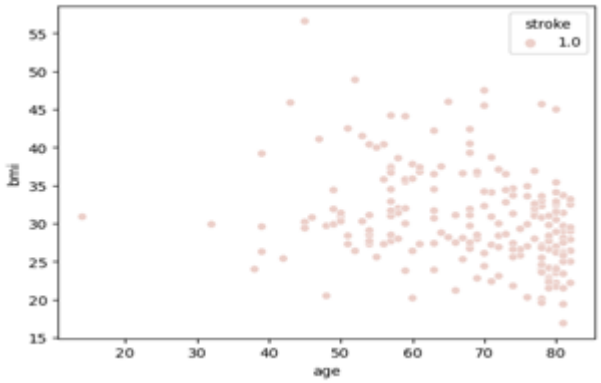

Fig. 9. Scatter plot between age and BMI.

We found that the data is randomly scattered and there are very less relationships between all these elements.

Then we did plot group graphs for a better understanding of the data using Histography.

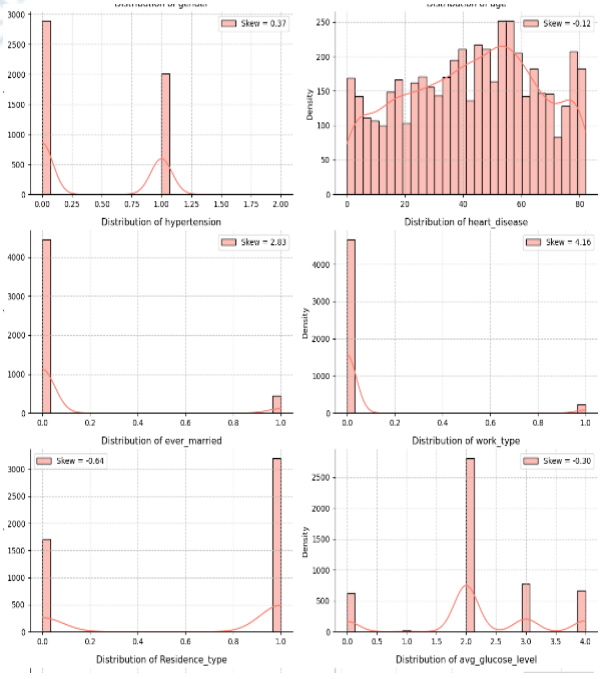

Fig. 10. Plotting histrographs for the data.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 6 June 2024

We did find relations in the features and the key value. As now we have known the relations we will go with feature engineering.

D. Model Selection

The model selection part of the project involves choosing the appropriate machine learning algorithms for training and evaluation. In the provided code, seven classification algorithms were selected for this purpose: Logistic Regression, Support Vector Classifier (SVC), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC), AdaBoost Classifier (ABC), and K-Nearest Neighbours Classifier (KNC). These algorithms were instantiated and trained using the training data (X_train and y_train).

Table III: Accuracy scores of the model.

S.No	Model Name	Accuracy
1	Logistic Regression	0.97
2	SVM	0.97
3	Decision Tree	0.94
4	Random Forest	0.97
5	Gradient Boosting	0.97
6	Ada Boost Classifier	0.97
7	KNN	0.97

From the accuracy scores we can tell that most of the machines have same accuracy score of 0.97 only Decision Tree has the least accuracy score of 0.94.

We choose Linear Regression by watching the performance of the model. As linear regression predicts the output better as the data is numerical, we will be going with Linear Regression.

E. Model Training

As we have chosen Logistic Regression, we will be training the model on train and test set.

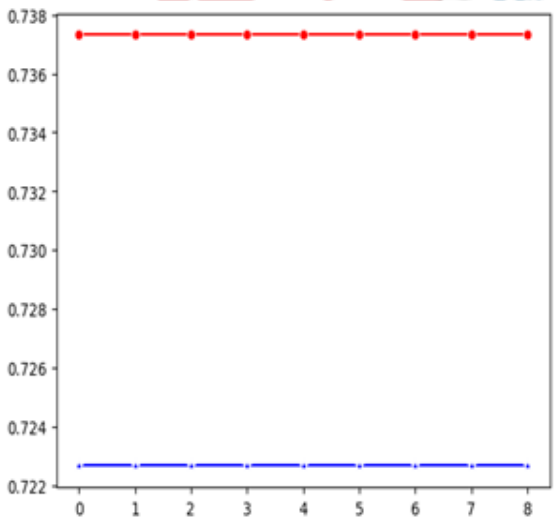

Fig. 11. Graph for accuracy of train and test set for LR.

We have observed that the graph is a straight line which indicates that our model performed well.

F. Model Evaluation

We have plotted graphs for ROC which provides a visual representation of the trade-off between sensitivity (TPR) and specificity (1-FPR) for different threshold values.

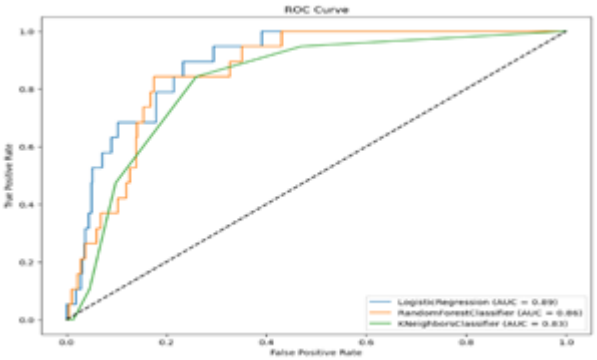

Fig. 12. ROC curves for the model.

We can see that Logistic Regression has the highest score of 0.89 from the graph plotted in Figure 12.

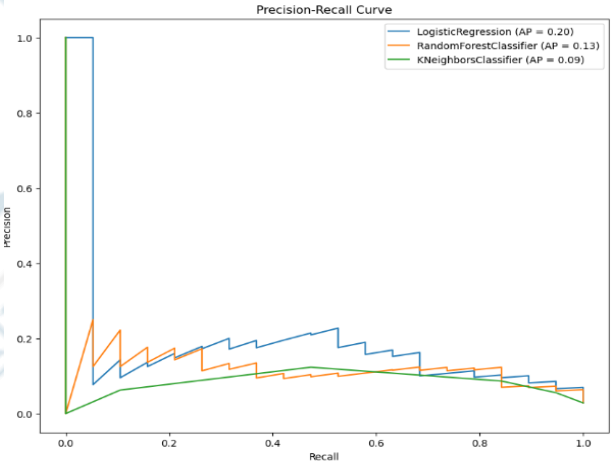

Fig. 13. Recall value graph for the models.

We can see that recall value for LR model is more than the other two values.

IV. CONCLUSION

The comprehensive analysis and evaluation conducted in this study provide valuable insights into the efficacy of machine learning-based models for stroke prediction. Through the utilization of a diverse dataset encompassing various demographic, lifestyle, and clinical attributes, we explored the performance of multiple classification algorithms, including Logistic Regression, Support Vector Classifier (SVC), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC), AdaBoost Classifier (ABC), and K-Nearest Neighbors Classifier (KNC). Our findings reveal that while several models exhibit high accuracy rates, such as Logistic Regression (97.16%), SVC (97.16%), RFC (97.16%), and

KNC (97.16%), each algorithm presents unique strengths and considerations. For instance, RandomForest and KNN demonstrate robust predictive performance, with accuracy scores of 97.16% each, underscoring their efficacy in identifying individuals at heightened risk of stroke. However, it's crucial to consider other metrics such as precision, recall, and F1-score to obtain a holistic understanding of model performance. Notably, Logistic Regression stands out with a recall value of 0.2, indicating its ability to correctly identify 20% of all actual positive instances. This suggests its potential utility in scenarios where minimizing false negatives is critical. Overall, our study underscores the importance of leveraging machine learning techniques and comprehensive datasets to enhance stroke prediction, enabling timely interventions and personalized care plans to mitigate the burden of stroke-related morbidity and mortality on a population scale.

V. ACKNOWLEDGMENT

The authors would like to thank Assistant Prof. Puneet, Faculty of Computer Science and Engineering, Lovely Professional University for providing the necessary guidance and facilities for the preparation of the paper.

REFERENCES

[1] S. Sirsat, M. S., Fermé, E., & Câmara, J. (2020). Machine Learning for Brain Stroke: A Review. Journal of Stroke and Cerebrovascular Diseases, 29(10), 105162. https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105162

[2] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun and M. S. Kaiser, "Performance Analysis of Machine Learning Approaches in Stroke Prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2020, pp. 1464-1469, doi: 10.1109/ICECA49313.2020.9297525.

[3] Sailasya, Gangavarapu and Gorli L Aruna Kumari. “Analyzing the Performance of Stroke Prediction using ML Classification Algorithms.” (2021).

[4] M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), Bangkok, Thailand, 2017, pp. 158-161, doi: 10.1109/IEMECON.2017.8079581

[5] Dritsas, E., & Trigka, M. (2021). Stroke Risk Prediction with Machine Learning Techniques. Sensors, 22(13), 4670. https://doi.org/10.3390/s22134670

[6] A. N. Tusher, M. S. Sadik and M. T. Islam, "Early Brain Stroke Prediction Using Machine Learning," 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2022, pp. 1280-1284, doi: 10.1109/SMART55829.2022.10046889.

[7] Heo, T. S., Kim, Y. S., Choi, J. M., Jeong, Y. S., Seo, S. Y., Lee, J. H., Jeon, J. P., & Kim, C. (2020). Prediction of Stroke Outcome Using Natural Language Processing-Based Machine Learning of Radiology Report of Brain MRI. Journal of Personalized Medicine, 10(4), 286. https://doi.org/10.3390/jpm10040286

[8] H. Puri, J. Chaudhary, K. R. Raghavendra, R. Mantri and K. Bingi, "Prediction of Heart Stroke Using Support Vector Machine Algorithm," 2021 8th International Conference on Smart Computing and Communications (ICSCC), Kochi, Kerala, India, 2021, pp. 21-26, doi: 10.1109/ICSCC51209.2021.9528241.

[9] Arslan, A. K., Colak, C., & Sarihan, M. E. (2016). Different medical data mining approaches based prediction of ischemic stroke. Computer Methods and Programs in Biomedicine, 130, 87-92. https://doi.org/10.1016/j.cmpb.2016.03.022

[10] K.Srinivas,B.Kavihta Rani, and Dr. A.Govrdhan, Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks, (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255

[11] T. L. Stewart, J. G. Chipperfield, R. P. Perry and J. M. Hamm, "Attributing heart attack and stroke to “Old Age: Implications for subsequent health outcomes among older adults", Journal of Health Psychology, vol. 21, no. 1, pp. 40-49, 2016.

[12] Bentley, P., Ganesalingam, J., Carlton Jones, A. L., Mahady, K., Epton, S., Rinne, P., Sharma, P., Halse, O., Mehta, A., & Rueckert, D. (2013). Prediction of stroke thrombolysis outcome using CT brain machine learning. NeuroImage: Clinical, 4, 635-640. https://doi.org/10.1016/j.nicl.2014.02.003

[13] An Efficient Modified Bagging Method for Early Prediction of Brain Stroke," 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2019, pp. 1-4, doi: 10.1109/IC4ME247184.2019.9036700.

[14] B. Akter, A. Rajbongshi, S. Sazzad, R. Shakil, J. Biswas and U. Sara, "A Machine Learning Approach to Detect the Brain Stroke Disease," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2022, pp. 897-901, doi: 10.1109/ICSSIT53264.2022.9716345.

[15] V. Krishna, J. Sasi Kiran, P. Prasada Rao, G. Charles Babu and G. John Babu, "Early Detection of Brain Stroke using Machine Learning Techniques," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2021, pp. 1489-1495, doi: 10.1109/ICOSEC51865.2021.9591840.