EARLY BRAIN STROKE PREDICTION USING MACHINE LEARNING ALGORITHMS

Nagaraju. Sonti[1], Ch. Karthik Reddy[2], D. Venkata Siva Reddy[3], G. Narendra[4] and Ch. Nageswara Rao[5]

[1] Assistant Professor, *Dept of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India*

[2,3,4,5] *UG Students, Dept of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India*

ABSTRACT

A brain stroke is a serious medical emergency that can have life-altering or permanently crippling effects. Cell death due to inadequate blood supply to the brain causes it to happen. In India, the annual incidence of stroke stands at 141 per 100,000 people. Predicting early brain strokes has become increasingly challenging, requiring time-consuming assessments. Given the life-or-death nature of stroke diagnoses and prognoses, precision and accuracy are crucial. Machine learning techniques offer a means to predict stroke issues by analyzing extensive medical data. By leveraging a substantial dataset for training and testing, the study assesses the predictive capabilities of various machine learning methods. A few examples of these techniques are K-nearest neighbour, decision trees, logistic regression, SVMs, Random Forest, AdaBoost, and Bernoulli navie bayes. The research assesses the efficacy of the model by means of the F1 score, Accuracy, Precision, and Recall, which are extracted from the confusion matrix. A web application will be developed, enabling users to input relevant parameters. Using this Flask-based application, the model processes these parameters. This approach, powered by the most accurate and effective method, can predict the likelihood of strokes.

Keywords*: K-nearest neighbour, decision trees, logistic regression, SVM, Random Forest, AdaBoost, and Bernoulli navie bayes*

1. INTRODUCTION

Acute focused injury to the central nervous system due to a vascular problem is the main cause of stroke, a neurological impairment. It is among the top causes of death and disability on a worldwide basis [1]. The total frequency in the United States is believed to be 2.5%, and over 7 million Americans over the age of 20 have experienced a stroke.

A patient's health and well-being are severely diminished by the condition. The estimated damage to the US economy from 2014 to 2015 was $351.2 billion [2]. Hospital services and bed availability are also negatively affected. The two most common types of strokes are ischemic and hemorrhagic. In contrast to an ischemic stroke, which occurs when blood vessels in the brain get blocked, a hemorrhagic stroke occurs when a brain vessel breaks. Between eighty-five and ninety percent of strokes are caused by blocked arteries [3].

A healthier population and more knowledge of the variables that put people at risk can avert this disease. Obesity, poor nutrition, excessive alcohol consumption, and insufficient physical activity are just a few of the numerous lifestyle-related risk factors [4]. Several preexisting conditions, such as diabetes, hypertension, and cardiovascular

problems, increase the risk of stroke. Stroke risk may be mitigated through the adoption of a healthy lifestyle and the self-management of certain illnesses.

We released a guideline in 2019 from the American College of Cardiology and the American Heart Association. If a patient is at high risk of having an artery blockage, which could cause a heart attack, stroke, or death, the guideline suggests that they have a thorough evaluation and examination [5]. The availability of clinical data has greatly improved in recent years, allowing doctors to more accurately identify patients at high risk through methods such as comprehensive patient histories and comprehensive physical examinations. Factors in a patient's lifestyle (such as their food and level of physical activity), as well as their demographics (such as their age and gender), and any preexisting medical conditions (such as diabetes or hypertension) that could cause a stroke are all included in their medical records [5].
Arterioles can get blocked and damage to blood arteries can build up over a long period of time, both of which increase the risk of stroke. It would be much easier to avoid strokes in their early stages if doctors could quickly and simply evaluate the risks of stroke. Potentially lowering the financial strain on health care systems, this strategy might save lives.
To help doctors diagnose patients at high risk of stroke in this age of artificial intelligence and machine learning, a clinical decision support system has been developed.
The cardiovascular sector holds great promise for the application of machine learning methods, which could lead to improvements in areas such as stroke risk assessment [6,7] and post-treatment patient outcome prediction [8,9, 10]. The bulk of these studies rely on neuroimaging techniques, such as computed tomography and magnetic resonance imaging, or health habits and lifestyle variables, such as smoking or alcohol intake, to categorize or forecast the condition [11]. Conditions that are predisposing to strokes include hypertension and diabetes mellitus.

2. LITERATURE REVIEW

Stroke is a major public health concern since it affects both sexes equally, reducing quality of life and straining public health resources. Because of its widespread usage in illness prevention, AI is playing an essential part in the scientific community's top priority: developing models for predicting strokes so that they can be avoided. Stroke diagnosis models, treatment outcome and patient response prediction models, and tailored rehabilitation procedure designs have all been the subject of extensive study [12], [13], [14], [15], [16].

For instance, in their data mining system for ischemic stroke prediction, The Support Vector Machine (SVM) classifier achieved the best results, with an accuracy of 97.89% and an area under the curve of 97.83%, according to Arslan et al. [17], who analyzed data from 80 ischemic stroke patients and 112 healthy persons. Finding the most important risk factors for ischemic stroke was another focus of the research. The summaries of relevant articles are categorized based on their publication year, the dataset used, the algorithm applied, and the achieved accuracy, as illustrated in Table 1.

In their study, the authors [28] investigate the challenges and potential biases of medical picture processing using deep learning systems. They propose many measures to enhance the explainability and trustworthiness of these algorithms, including visualization tools, feature attribution methods, and interpretable models. Reading this article will help you understand why it's crucial to make sure that deep learning algorithms used for medical image analysis are open and easy to understand for everyone involved.

Table -1: Performance comparison of different models

Ref	Year	DL Algorithm	Dataset	Performance
[18]	2022	Eli5, LIME, AGB	Kaggle	Accuracy: 80%
[19]	2022	NB, LR, K-NN, SGD, DT, MLP, RF, Stacking	Kaggle	Accuracy: 80%

[20]	2022	LR, RF, KNN, SVM, MLP	Cerebral Stroke Prediction- Imbalanced Dataset (Kaggle)	false-negative rate (18.60%), Accuracy (73.52%)
[21]	2021	AGB, LR, RF, KNN, SVM, MLP	Hospitals in Bangladesh	Accuracy: 98%
[22]	2021	LR, RF, KNN, SVM, MLP	Geisinger Health Open- Source Data Set	Accuracy: 95%
[23]	2020	DT and ID3.	HealthCare Dataset	Accuracy: 98%
[24]	2021	LR, RF, SVM	The open-access Stroke Prediction dataset	Accuracy: Random forest: 96%
[25]	2021	Rf, LR, and DT	HealthCare Dataset	Accuracy: KNN: 95%
[26]	2022	NN, DT, and RF	EHRs by McKinsey & Company	Accuracy: NN: 77%
[27]	2019	RF, DT, and RF	Healthcare Dataset Stroke	RF:90%, DT:79%, SVM: 77%, LR:77%.

3. METHODOLOGY

3.1 Proposed system

Figure 1 illustrates the proposed system's architecture.

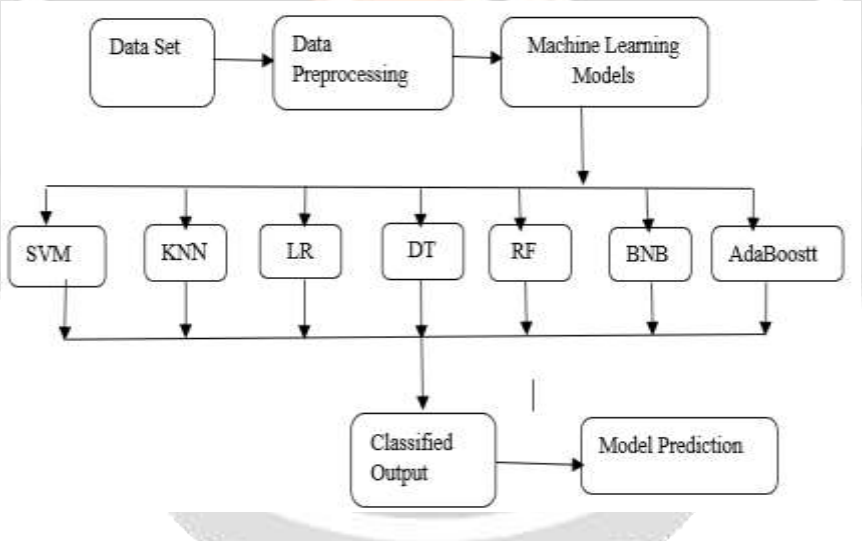

Fig -1: Block diagram of proposed system

The operational concept of the proposed system for identifying brain strokes is depicted in Figure 2.

A.Dataset

We acquired a dataset from Kaggle initially comprising 5110 samples; however, it contained nonrelevant entries unrelated to strokes. Subsequently, we curated a new dataset, totaling 159 kilobytes, which consists of 2870 samples. This dataset encompasses 11 features such as gender, age, hypertension, smoking status, and more. Among these 2870 samples, 996 instances are associated with strokes, while the remaining 1874 samples are categorized as non-stroke cases.

B. Data preprocessing

After compiling the dataset, an evaluation was conducted to identify and address the presence of null values. These null values were subsequently replaced with the mean value corresponding to each respective feature. Additionally, an analysis focusing on outliers in the BMI feature was carried out, and any outliers detected were substituted with

the mean value to ensure data consistency. Subsequent to this, a scrutiny of the data types for each feature was performed, leading to the conversion of object data types to integer types.

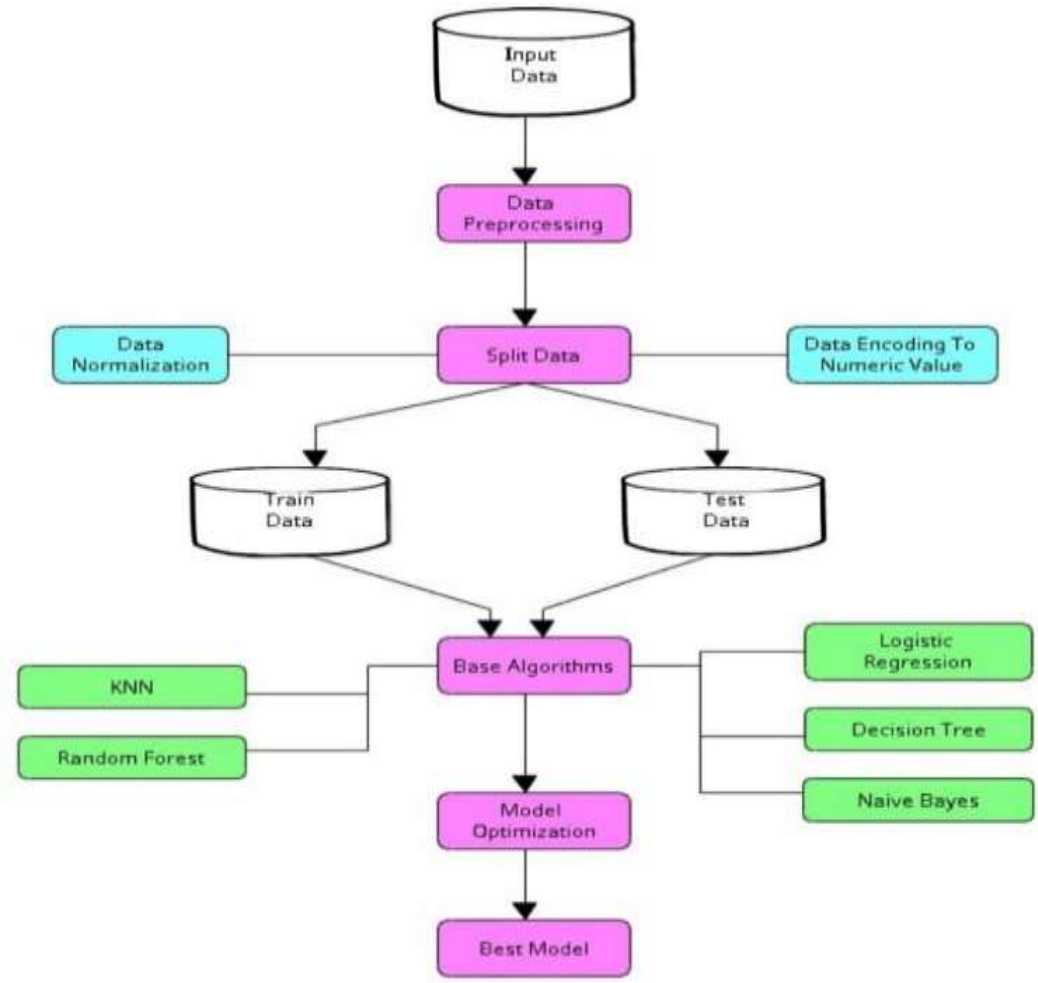

Fig -2: Flow chart for proposed method

C. Model Training

Trained the following machine learning algorithms: KNN, Naïve Bayes, Decision Tree, Logistic Regression, and Random Forest. KNN, a proximity-based technique, was used throughout the model training phase. Finding the most common class among a data point's k-nearest neighbours in the feature space is how this method sorts them. In order to train the models, we used the Random Forest algorithm, which is an effective ensemble method. The method relies on training a network of decision trees and then producing a single class that is the mean of all the classes predicted by those trees. The training method included the widely used linear model, Logistic Regression. If you want to know what the odds are that a given sample belongs to a certain class, this approach can handle binary classification problems well.

The Decision Tree algorithm played a crucial role in model training. This method recursively splits the dataset into subsets based on the most discriminative features, constructing a tree-like structure to facilitate classification.

Naive Bayes, a probabilistic algorithm, was included in the model training phase. Leveraging Bayes' theorem, it calculates the probability of each class for a given set of features, assuming independence among features—a simplifying yet effective assumption in practice.

4. RESULTS AND DISCUSSION

The results are shown in Table 2, and it is evident that Random Forest algorithm achieves the highest level of accuracy.

Table -2: Classification accuracy for data leakage and no data leakage

Algorithms	Data Leakage	No Data Leakage
Random Forest	90.36	82.23
Logistic Regression	80.18	74.35
Support Vector Machine	80.18	74.65
K Nearest Neighbours	86.74	81.61
Naïve Bayes	76.03	71.26
Decision Tree	89.02	83.43

In an ideal world, it would have a score of 90.36 percent; in an actual data leak scenario, it would be 82.23%. Among all methods, Decision Tree has the second-best performance, with accuracy ratings of 83.43% when data leakage is present and 89.02% when it is not.

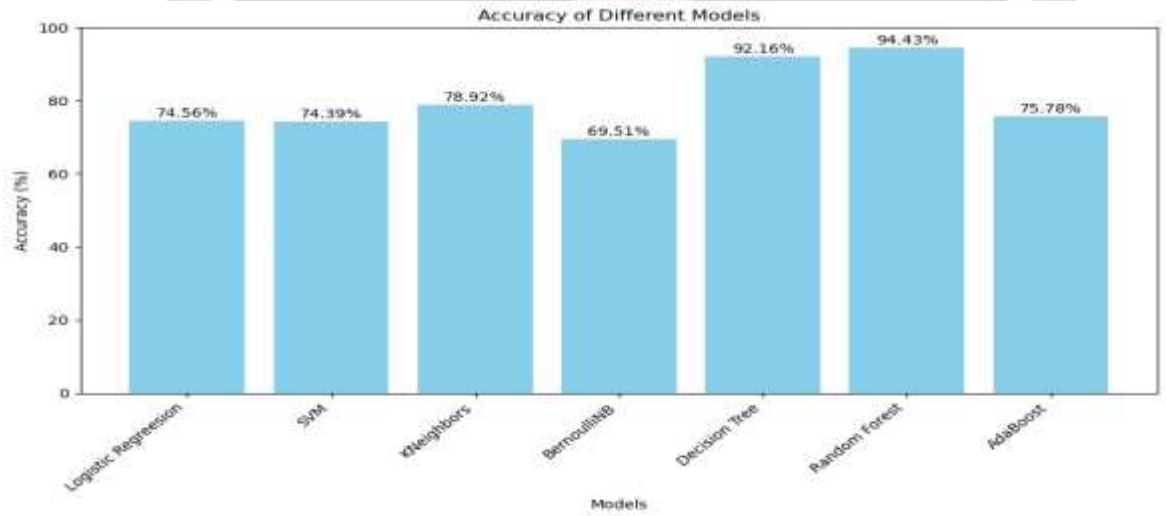

Fig -3: Accuracy Graph for all ML models.

Random Forest and Decision Tree consistently beat the other approaches, as shown in Figure 3, independent of the presence or absence of data leaking. Naive Bayes has the worse accuracy ratings in both cases. Since the difference between accuracy ratings with and without data leakage is often minor, it might not be a huge concern for this particular dataset and combination of algorithms.

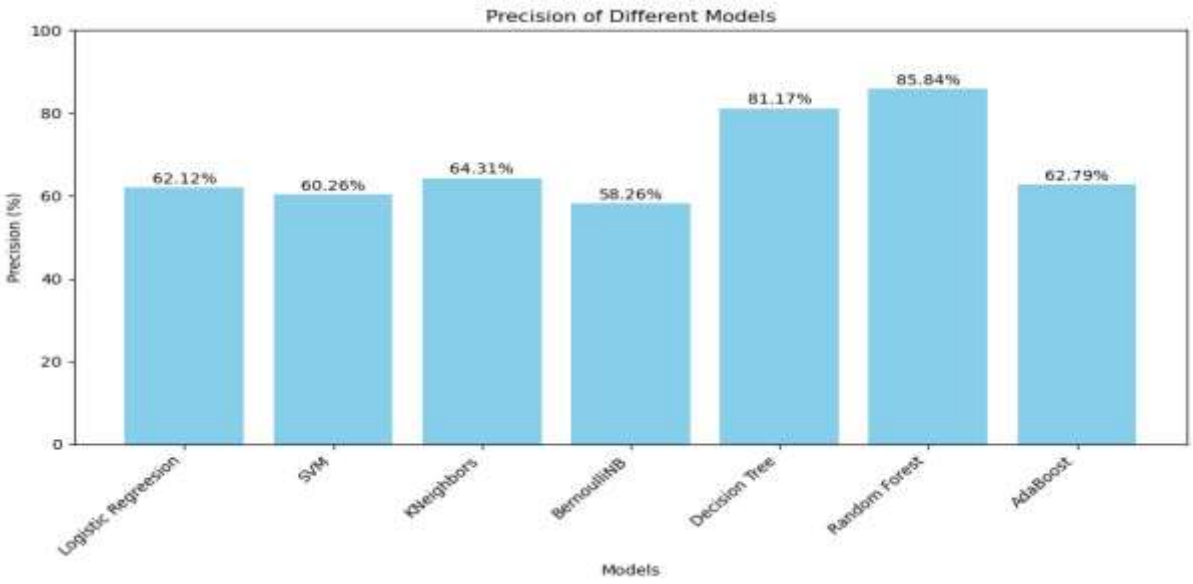

Fig -4: Precision Value for all ML models.

As demonstrated in Figure 4, RF possesses the greatest precision values for both classes, with a 0.93 for class 0 (No) and 0.88 for class 1 (Yes). The precision values of LR and KNN are similarly high for class 0 (No): 0.81 and 0.92, respectively. When it comes to class 1 (Yes), NB's precision of 0.74 is the lowest of the two classes.

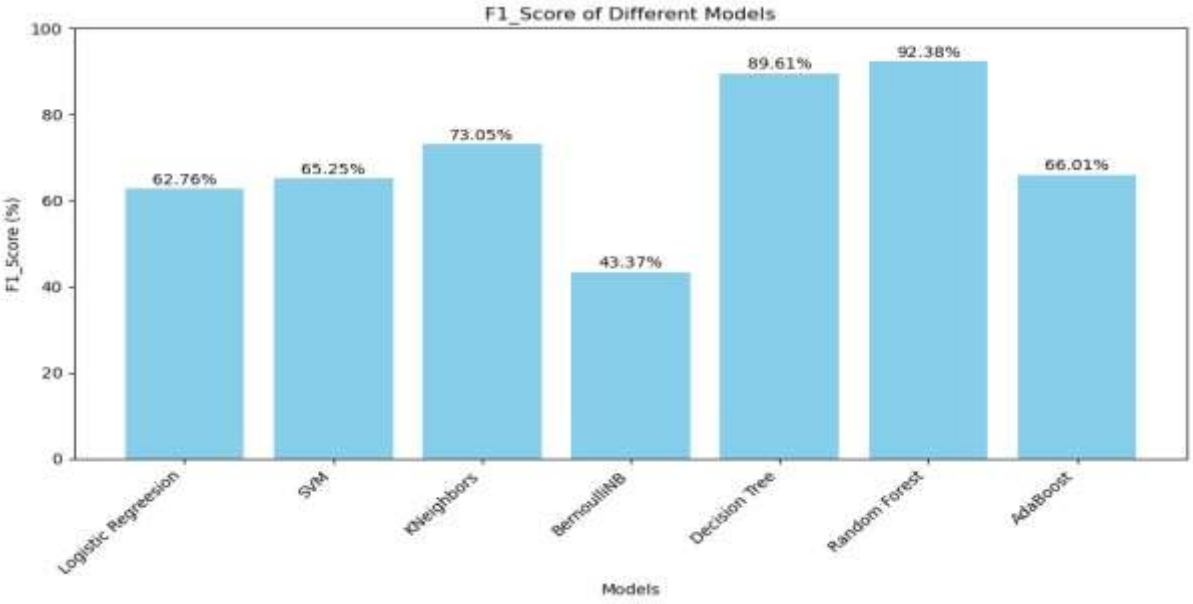

Fig -5: F1-score for all ML models.

Both DT and RF had the highest F1-scores in every class (Figure 5), with DT topping Class 1 and RF topping Class 0. Both categories rank NB bottom in terms of F1-score. Plotting the results for each class allows for an easy visual comparison of the models' performance.

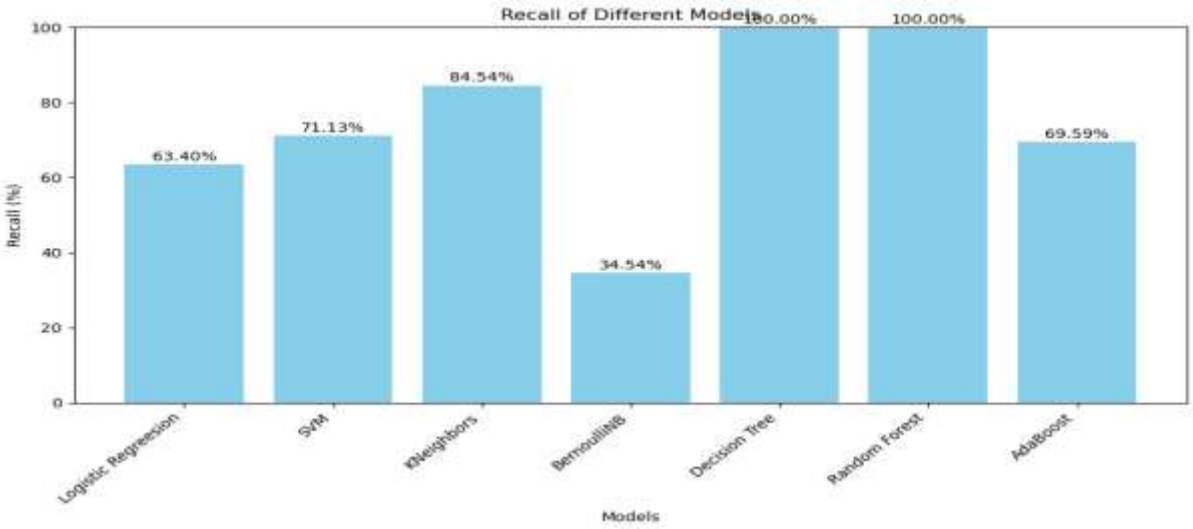

Fig -6: Recall Values for all ML models.

Recall scores for both classes are highest for the RF and DT models (Figure 6), with the former having the best score for Class 0 and the latter for Class 1. Recall ratings for both classes are lower for KNN and SVC models compared to LR and NB, which show higher variation. Also figure 7 shows the ROC of different models. The comparative analysis of different proposed models is shown in table 3.

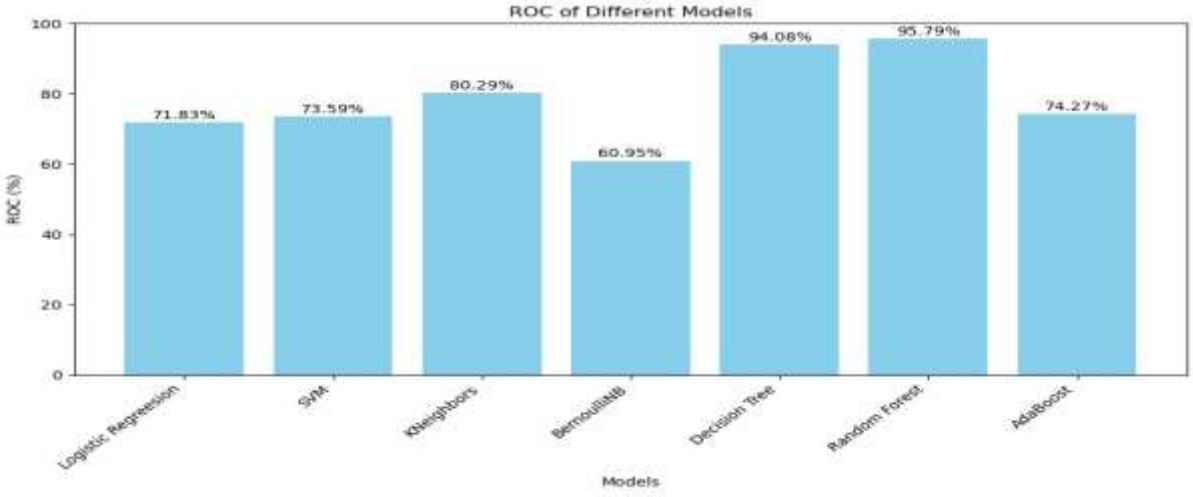

Fig -7: ROC of different models

Table -3: Comparative analysis of proposed models

Model Name	Accuracy (%)	Precision (%)	F1-Score (%)	Recall (%)	ROC (%)
Logistic Regression	74.56	62.12	62.76	63.40	71.83
SVM	74.39	62.26	65.25	71.13	73.59

KNN	78.92	64.31	73.05	84.54	80.29
Decision Tree	92.16	81.17	89.61	100	94.08
Random Forest	94.43	85.84	92.38	100	95.79
Logistic Regression	74.56	62.12	62.76	63.40	71.83
SVM	74.39	62.26	65.25	71.13	73.59
AdaBoost	75.78	62.79	66.01	69.59	74.27
BNB	69.51	58.26	43.37	34.54	60.95

Fig -8: Early prediction

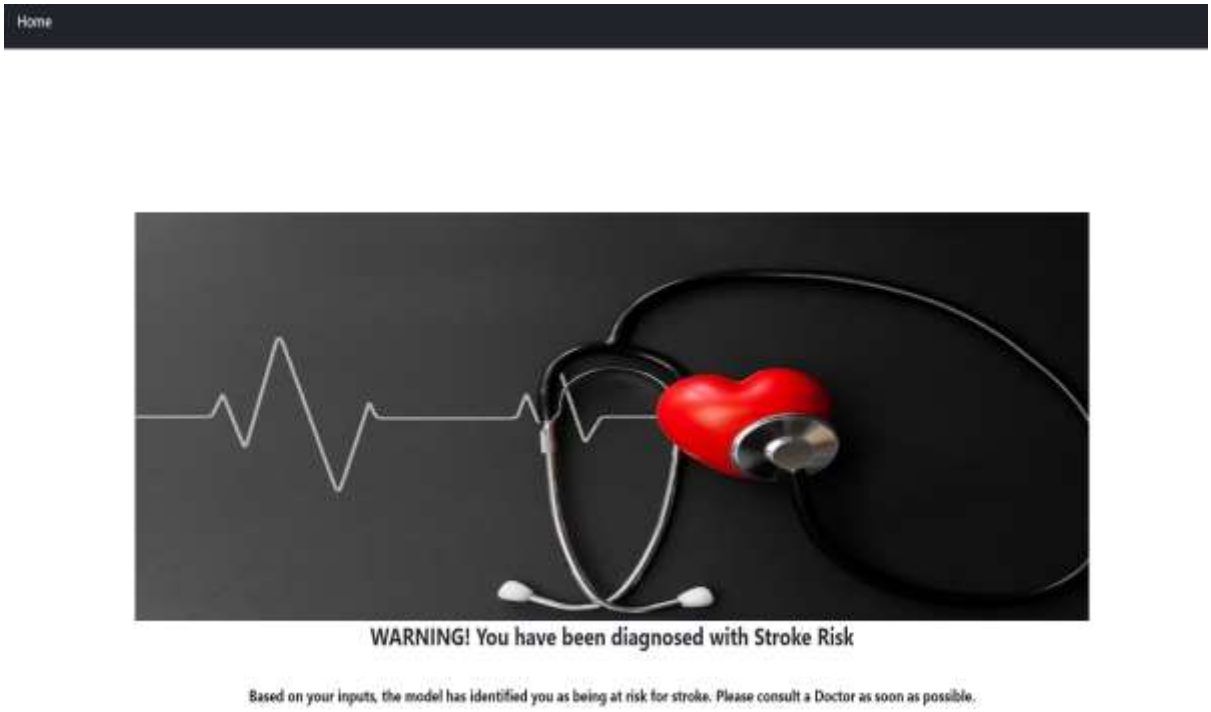

Fig -9: Stroke risk

Fig -10: No Stroke risk

Figure 8 shows Early detection of brain using machine learning methods. And figure 9 shows the warning when you have diagnosed with stroke risk. When there was no risk, it will become don't worry and was shown in figure 10.

5. CONCLUSION

Finally, there have been encouraging outcomes from the creation and testing of machine learning models for making early predictions of brain strokes. In this work, we used a variety of algorithms to identify potential risk factors for stroke. These algorithms included KNN, Random Forest, Logistic Regression, Decision Tree, and Naive Bayes. The careful preprocessing of the dataset, addressing null values, outliers, and data type considerations, contributed to the robustness of our models. As evidenced by metrics such as F1 score and ROC analysis, our models showcase a meaningful step forward in enhancing stroke prediction capabilities.

6. REFERENCES

[1] Sacco RL, Kasner SE, Broderick JP, Caplan LR, Connors JJB, Culebras A, Elkind MSV, George MG, Hamdan AD, Higashida RT, Hoh BL, Janis LS, Kase CS, Kleindorfer DO, Lee JM, Moseley ME, Peterson ED, Turan TN, Valderrama AL, Vinters HV, American Heart Association Stroke Council, Council on Cardiovascular Surgery and Anesthesia. Council on Cardiovascular Radiology and Intervention. Council on Cardiovascular and Stroke Nursing. Council on Epidemiology and Prevention. Council on Peripheral Vascular Disease. Council on Nutrition, Physical Activity and Metabolism An updated definition of stroke for the 21st century: A statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. 2013 Jul;44(7):2064–2089. doi: 10.1161/STR.0b013e318296aeca.STR.0b013e318296aeca [PubMed] [CrossRef] [Google Scholar]

[2] Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP: A report from the American Heart Association. Circulation. 2019 Mar 05;139(10):e56–e528. doi: 10.1161/CIR.0000000000000659. https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000000659?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%3dpubmed . [PubMed] [CrossRef] [Google Scholar]

[3] European Stroke Initiative Executive Committee. EUSI Writing Committee. Olsen TS, Langhorne P, Diener HC, Hennerici M, Ferro J, Sivenius J, Wahlgren NG, Bath P. European Stroke Initiative Recommendations for Stroke Management – Update 2003. Cerebrovasc Dis. 2003;16(4):311–337. doi: 10.1159/000072554. https://www.karger.com?DOI=10.1159/000072554 . [PubMed] [CrossRef] [Google Scholar]

[4] Boden-Albala B, Sacco RL. Lifestyle factors and stroke risk: Exercise, alcohol, diet, obesity, smoking, drug use, and stress. Curr Atheroscler Rep. 2000 Mar;2(2):160–166. doi: 10.1007/s11883-000-0111-3. [PubMed] [CrossRef] [Google Scholar]

[5] Arnett D, Blumenthal R, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, Himmelfarb CD, Khera A, Lloyd-Jones D, McEvoy JW, Michos ED, Miedema MD, Muñoz D, Smith SC, Virani SS, Williams KA, Yeboah J, Ziaeian B. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol. 2019 Sep 10;74(10):e177–e232. doi: 10.1016/j.jacc.2019.03.010.S0735-1097(19)33877-X [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[6] Manuel DG, Tuna M, Perez R, Tanuseputro P, Hennessy D, Bennett C, Rosella L, Sanmartin C, van Walraven C, Tu JV. Predicting stroke risk based on health behaviours: Development of the Stroke Population Risk Tool (SPoRT) PLoS One. 2015;10(12):e0143342. doi: 10.1371/journal.pone.0143342. https://dx.plos.org/10.1371/journal.pone.0143342 .PONE-D-15-24430 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[7] Lee J, Lim H, Kim D, Shin S, Kim J, Yoo B, Cho K. The development and implementation of stroke risk prediction model in National Health Insurance Service's personal health record. Comput Methods Programs Biomed. 2018 Jan;153:253–257. doi: 10.1016/j.cmpb.2017.10.007. https://linkinghub.elsevier.com/retrieve/pii/S0169-2607(16)31470-5 .S0169-2607(16)31470-5 [PubMed] [CrossRef] [Google Scholar]

[8] Nielsen A, Hansen MB, Tietze A, Mouridsen K. Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning. Stroke. 2018 Jun;49(6):1394–1401. doi: 10.1161/strokeaha.117.019740. [PubMed] [CrossRef] [Google Scholar]

[9] Rondina JM, Filippone M, Girolami M, Ward NS. Decoding post-stroke motor function from structural brain

imaging. Neuroimage Clin. 2016;12:372–380. doi: 10.1016/j.nicl.2016.07.014. https://linkinghub.elsevier.com/retrieve/pii/S2213-1582(16)30134-6 .S2213-1582(16)30134-6 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
[10] Sughrue T, Swiernik MA, Huang Y, Brody JP. Laboratory tests as short-term correlates of stroke. BMC Neurol. 2016 Jul 21;16:112. doi: 10.1186/s12883-016-0619-y. https://bmcneurol.biomedcentral.com/articles/10.1186/s12883-016-0619-y .10.1186/s12883-016-0619-y
[11] Nagaraju Sonti, "COVID-19 Detection from X-rays using CNN-based Graph with Fast Localization spectral filters" Date of Conference: 01-02 September 2023 Date Added to IEEE Xplore: 17 October 2023 ISBN Information: DOI: 10.1109/NMITCON58196.2023.10275820 Publisher: IEEE Conference Location: Bengaluru, India
[12] Y. Zhao, S. Fu, S. J. Bielinski, P. A. Decker, A. M. Chamberlain, V. L. Roger, H. Liu, and N. B. Larson, ''Natural language processing and machine learning for identifying incident stroke from electronic health records: Algorithm development and validation,'' J. Med. Internet Res., vol. 23, no. 3, Mar. 2021, Art. no. e22951.
[13] B. McDermott, A. Elahi, A. Santorelli, M. O'Halloran, J. Avery, and E. Porter, ''Multi-frequency symmetry difference electrical impedance tomography with machine learning for human stroke diagnosis,'' Physiological Meas., vol. 41, no. 7, Aug. 2020, Art. no. 075010.
[14] A. Bivard, L. Churilov, and M. Parsons, ''Artificial intelligence for decision support in acute stroke—Current roles and potential,'' Nature Rev. Neurol., vol. 16, no. 10, pp. 575–585, Oct. 2020.
[15] W. Wang, M. Kiik, N. Peek, V. Curcin, I. J. Marshall, A. G. Rudd, Y. Wang, A. Douiri, C. D. Wolfe, and B. Bray, ''A systematic review of machine learning models for predicting outcomes of stroke with structured data,'' PLoS ONE, vol. 15, no. 6, Jun. 2020, Art. no. e0234722.
[16] M. S. Sirsat, E. Fermé, and J. Câmara, ''Machine learning for brain stroke: A review,'' J. Stroke Cerebrovascular Diseases, vol. 29, no. 10, Oct. 2020, Art. no. 105162.
[17] A. K. Arslan, C. Colak, and M. E. Sarihan, ''Different medical data mining approaches based prediction of ischemic stroke,'' Comput. Methods Programs Biomed., vol. 130, pp. 87–92, Jul. 2016.
[18] M. S. Islam, I. Hussain, M. M. Rahman, S. J. Park, and M. A. Hossain, ''Explainable artificial intelligence model for stroke prediction using EEG signal,'' Sensors, vol. 22, no. 24, p. 9859, Dec. 2022.
[19] E. Dritsas and M. Trigka, ''Stroke risk prediction with machine learning techniques,'' Sensors, vol. 22, no. 13, p. 4670, Jun. 2022.
[20] C. Kokkotis, G. Giarmatzis, E. Giannakou, S. Moustakidis, T. Tsatalas, D. Tsiptsios, K. Vadikolias, and N. Aggelousis, ''An explainable machine learning pipeline for stroke prediction on imbalanced data,'' Diagnostics, vol. 12, no. 10, p. 2392, Oct. 2022.
[21] R. Islam, S. Debnath, and T. I. Palash, ''Predictive analysis for risk of stroke using machine learning techniques,'' in Proc. Int. Conf. Comput., Commun., Chem., Mater. Electron. Eng. (ICME), Dec. 2021, pp. 1–4.
[22] N. Darabi, N. Hosseinichimeh, A. Noto, R. Zand, and V. Abedi, ''Machine learning-enabled 30-day readmission model for stroke patients,'' Frontiers Neurol., vol. 12, Mar. 2021, Art. no. 638267.
[23] Y. Choi and J. W. Choi, ''Stroke prediction using machine learning based on artificial intelligence,'' Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 5, pp. 1–6, Sep. 2020.
[24] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. M. Khan, ''Stroke disease detection and prediction using robust learning approaches,'' J. Healthcare Eng., vol. 2021, pp. 1–12, Nov. 2021.
[25] H. K. Gupta, ''Stroke prediction using machine learning algorithms,'' Int. J. Innov. Res. Eng. Manage., vol. 8, no. 4, pp. 6–9, Jul. 2021, doi: 10.21276/ijirem.2021.8.4.2.
[26] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, ''A predictive analytics approach for stroke prediction using machine learning and neural networks,'' Healthcare Analytics, vol. 2, Nov. 2022, Art. no. 100032.
[27] A. A. Ali, ''Stroke prediction using distributed machine learning based on Apache spark,'' Stroke, vol. 28, no. 15, pp. 89–97, 2019.
[28] K. Mridha, S. Kumbhani, S. Jha, D. Joshi, A. Ghosh, and R. N. Shaw, ''Deep learning algorithms are used to automatically detection invasive ducal carcinoma in whole slide images,'' in Proc. IEEE 6th Int. Conf. Comput., Commun. Autom. (ICCCA), Arad, Romania, Dec. 2021, pp. 123–12.