spr

brainstroke prediction.docx

SR University

Document Details

Submission ID trn:oid:::3618:74940138	11 Pages
Submission Date Dec 29, 2024, 12:39 PM GMT+5:30	4,695 Words
Download Date Dec 29, 2024, 12:40 PM GMT+5:30	26,975 Characters
File Name brainstroke prediction.docx	
File Size 515.5 KB	

33% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

84 Not Cited or Quoted 28%
Matches with neither in-text citation nor quotation marks

6 Missing Quotations 2%
Matches that are still very similar to source material

4 Missing Citation 3%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

26%	Internet sources
25%	Publications
28%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **84** Not Cited or Quoted 28%
  Matches with neither in-text citation nor quotation marks
- **6** Missing Quotations 2%
  Matches that are still very similar to source material
- **4** Missing Citation 3%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

26%	Internet sources
25%	Publications
28%	Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.mdpi.com	9%
2	Internet	formative.jmir.org	3%
3	Internet	mdpi-res.com	3%
4	Publication	Yuelong Zhang, Qi Zhang, Jia Lv, Dahong Zhang, Meriem Kaddouri, Saeed Hamee...	1%
5	Submitted works	Universitat Internacional de Catalunya on 2023-01-09	1%
6	Internet	www.medrxiv.org	1%
7	Publication	Raymond Farah, Nava Samra. "Mean platelets volume and neutrophil to lymphoc...	1%
8	Submitted works	Liverpool John Moores University on 2023-06-12	1%
9	Submitted works	Federal University of Technology on 2024-05-10	1%
10	Internet	www.researchgate.net	1%

11	Publication	Churchman, Emma. "Understanding Mechanisms Underlying Changes in Parental...	0%
12	Publication	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challeng...	0%
13	Internet	dokumen.pub	0%
14	Internet	link.springer.com	0%
15	Internet	www.igi-global.com	0%
16	Internet	easychair.org	0%
17	Submitted works	University of Greenwich on 2024-09-06	0%
18	Internet	docplayer.net	0%
19	Submitted works	Bournemouth University on 2024-08-20	0%
20	Internet	www.ncbi.nlm.nih.gov	0%
21	Publication	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Co...	0%
22	Submitted works	Liverpool John Moores University on 2023-09-06	0%
23	Publication	Niranjan Sapkota. "The crypto collapse chronicles: Decoding cryptocurrency exch...	0%
24	Submitted works	The University of Memphis on 2024-12-03	0%

25	Publication	Jickling, Glen C, DaZhi Liu, Bradley P Ander, Boryana Stamova, Xinhua Zhan, and F...	0%
26	Internet	assets-eu.researchsquare.com	0%
27	Publication	"Proceedings of International Conference on Recent Innovations in Computing", ...	0%
28	Publication	Parvathaneni Naga Srinivasu, Uddagiri Sirisha, Kotte Sandeep, S. Phani Praveen, ...	0%
29	Submitted works	RMIT University on 2024-08-18	0%
30	Publication	Thangavel Murugan, W. Jai Singh. "Cybersecurity and Data Science Innovations fo...	0%
31	Submitted works	City University of Hong Kong on 2019-04-09	0%
32	Submitted works	Tilburg University on 2024-12-02	0%
33	Internet	heca-analitika.com	0%
34	Submitted works	University of Greenwich on 2024-12-20	0%
35	Internet	arxiv.org	0%
36	Internet	www.herminahospitals.com	0%
37	Publication	Hatice Nizam-Ozogur, Zeynep Orman. "Chapter 55 A Comparative Study ofPrepro...	0%
38	Submitted works	Malta College of Arts,Science and Technology on 2024-06-10	0%

39	Submitted works	Universiti Teknologi MARA on 2021-07-31	0%
40	Submitted works	University College London on 2024-10-11	0%
41	Internet	ruj.uj.edu.pl	0%
42	Submitted works	Napier University on 2024-08-14	0%
43	Submitted works	Queen's University of Belfast on 2023-10-15	0%
44	Submitted works	University of Hertfordshire on 2024-04-28	0%
45	Submitted works	University of Westminster on 2023-01-14	0%
46	Internet	www.theseus.fi	0%
47	Submitted works	CSU, San Jose State University on 2021-12-17	0%
48	Submitted works	George Bush High School on 2023-07-15	0%
49	Submitted works	Liverpool John Moores University on 2022-03-22	0%
50	Submitted works	Liverpool John Moores University on 2024-09-15	0%
51	Submitted works	Sheffield Hallam University on 2023-09-07	0%
52	Internet	fastercapital.com	0%

53	Internet	
kuey.net		0%
54	Internet	
rosap.ntl.bts.gov		0%
55	Internet	
thefinancialexpress.com.bd		0%
56	Publication	
"Advances in Human Factors and Ergonomics in Healthcare and Medical Devices",...		0%
57	Submitted works	
Arts, Sciences & Technology University In Lebanon on 2024-08-22		0%
58	Publication	
Pawan Singh Mehra, Dhirendra Kumar Shukla. "Artificial Intelligence, Blockchain,...		0%
59	Submitted works	
Asia Pacific University College of Technology and Innovation (UCTI) on 2023-03-01		0%
60	Submitted works	
Leeds Beckett University on 2022-09-09		0%
61	Submitted works	
Liverpool John Moores University on 2022-06-03		0%

Brain Stroke Prediction using Machine Learning and Neural Networks Models: A Comparative Study

Kola Vennela, Devireddy Vinisha, Nampelli Sai Vivek
School of Computer Science and Artificaial Intelligence,
SR University,Warangal - 506371, Telangana, India
Kolavennela90@gmail.com
Devireddyvinisha28@gmail.com
Saivivek729@gmail.com

Abstract

This paper discusses in detail some of the various machine learning and neural network models applied in predicting the risk of brain stroke. Brain stroke remains one of the major global health concerns because of the number ranking for various causes of mortality and disability. The ability to predict the risk of stroke before symptoms occur would save lives, reduce complications during treatment, and enhance the quality of life for patients if interventions could be made earlier. We explore the predictive capability of four models—Random Forest, Support Vector Machine (SVM), XGBoost, and Artificial Neural Networks (ANN)—trained on a dataset of demographic, clinical, and lifestyle variables. Results show that ensemble methods, Random Forest and XGBoost, yield the best accuracy scores of 94% and 93%, respectively. The results allow for possible application of machine learning techniques in strengthening clinical decision-making and identifying susceptible individuals. Further work in this area would include the inclusion of machine learning algorithms in healthcare infrastructures for real-time risk assessment and monitoring of acute stroke events and use of all-encompassing datasets, genetic and imaging types to further improve predictive abilities.

Keywords. Brain stroke, machine learning, neural networks, ANN, SVM

1.Introduction

Stroke, or cerebrovascular accident (CVA), is a severe medical condition resulting from the interruption of blood supply to a part of the brain occurring either by blockage (ischemic stroke) or a rupture in one of the blood vessels (hemorrhagic stroke). This can immediately cut off the oxygen and nutrient supply to the brain cells, causing them to die off in as little as minutes. Stroke is one of the leading causes of death worldwide and a main reason for long-term disability. According to reports by the World Health Organization, around 15 million people are affected by a stroke every year, while 5 million die, and another 5 million are permanently disabled. This burden weighs not just on the individuals and their families but also exercises critical economic pressure on the healthcare systems of the world.

Early intervention for high-risk stroke patients happens to be the most important issue in the management of stroke. Early detection with immediate intervention reduces the severity of sequelae of brain injury, lowers the grades of disability, and chances of recovery. The term "golden hour" is a nomenclature often used for the critical period, which underscores the rapidity and precision needed for stroke prediction. However, accurate identification of stroke risk in people is usually difficult because stroke is a condition influenced by a combination of genetic, physiological, and lifestyle factors that differ widely across populations. Many conventional statistical approaches used within stroke risk prediction are often too weak to portray complex, nonlinear interactions that exist in clinical data.

Advances in artificial intelligence and learning algorithms have opened up new avenues for boosting the robustness and consistency of stroke risk predictions. For example, whereas the statistical models of old require classical computation, machine learning models can be generated after examining massive data sets to derive their subtle patterns with adaptation to new information. Such models are highly suitable for complex medical predictions. It is even possible to train machine learning algorithms on diverse health datasets such as clinical records, demographic factors, lifestyle information, and even genomic data to provide an integrated assessment of an individual's risk with stroke prediction. Such models are powerful and flexible since they can handle and integrate multiple data sources about a patient.

This kind of approach is portrayed by different machine learning models, which include Random Forest, Support Vector Machines (SVM), XGBoost, and even Artificial Neural Networks (ANN). These models portray varied pathway approaches in predictive healthcare, where from them, one finds that Random Forest models are capable of combining multiple decision trees to reduce the variance and improve accuracies in predictions. XGBoost is an ensemble learning technique based on gradient boosting, known for efficiency and accuracy and often capable of handling large, complex datasets with great efficacy. SVM is a robust classifier that builds the classification hyperplane and performs especially well with high-dimensional data, besides having always consistent decision boundaries. These are ANN models: although computationally expensive, they can fit nonlinear relationships in data and thus find high applications in complex classification tasks involving unclear patterns.

1.2.Objectives and Scope:

It will carry out a comparative analysis of four machine learning models based on stroke prediction: Random Forest, SVM, XGBoost, and ANN. It will be interesting to see which model is the most accurate, robust, and consistent when applied for the final prediction of stroke risk from the validation set. Each model has its strengths-ensemble models like Random Forest and XGBoost are very robust and generalize well on diverse datasets. SVM has been found to be a consistent performer as it can take high dimensional data but ANN's adaptability to fit complex, non-linear data patterns into it requires considerable computations more than in SVM.

In this comparative analysis, we will answer the following critical questions in predictive healthcare. First, we will determine which model generates the most robust performance on evaluation metrics: accuracy, precision, recall, and F1-score. Second, we will assess how consistent the models are in their performances when trained and tested on stroke datasets which reflect real-world variability in data. Third, we will study the key contributing factors in each model's decision-making process and if possible ascertain what these could be interpreted in clinically meaningful ways.

The practical implications the findings might have are outstanding for healthcare providers. If there existed a reliable prediction model of a stroke, that would also be helpful in early diagnosis with the capacity to identify the high-risk individuals and provide timely preventive measures. Such interventions might include lifestyle modification or adjusting medication to reduce the incidence of stroke and thereby reduce healthcare costs. Furthermore, such predictive models can be directly integrated with EHRs so as to monitor the risk of a patient in real time and thereby enable real-time proactive management of healthcare according to the need of the individual.

1.3.Significance of Research

This approach could aid in the transformation of healthcare by providing data-based insights and proactively enabling care. Thereby, from a reactive to preventive approach in medicine, the study evaluates a predictive healthcare model. With algorithms embedded within machine learning, healthcare can better identify at-risk patients and then deploy targeted interventions designed to work toward improved overall patient outcomes. This is especially critical in stroke prevention where a difference between full recovery and permanent disability may be the deciding factor in acting early.

The fact that machine learning can easily process large datasets and associated complex inter-relationships among variables speaks to one of its strengths in healthcare. While clinical data like blood pressure and cholesterol levels may be combined with lifestyle factors like smoking habits and physical activity levels to produce a nuanced risk profile in stroke prediction, the clinician using a comparison of an exemplar case will ask patients with acute ischemia about their cigarette smoking history, rate of alcohol consumption, and other pertinent lifestyle factors. Additionally, predictiveness with which these algorithms could deal with new patient data promises that predictive accuracy can be maintained at high levels over time, in tandem with the actual evolution of information in health care. This feature allows machine learning models to be updated and periodically retrained so that they keep pace with shifting demographics of patients and emerging trends in medicine.

However, a number of challenges will need to be addressed before machine learning can assure reliability of outcomes in healthcare. A major one of these is the interpretability of machine learning models, especially deep learning models such as ANN that function much more like a "black box." In applications involving clinical expertise, it is important for healthcare professionals to understand how and why a prediction was given so they are making decisions based on informed reasoning. Although models like Random Forest and XGBoost provide increased interpretability through feature importance metrics, ANN may demand additional tools like SHapley Additive exPlanations (SHAP) values to explain individual predictions. Ethical considerations also entail considerations about patient privacy and any potential biases present in the data used for training that the final model should guarantee fairness and reliability in the predictions.

This paper contributes to the growing literature on AI-based stroke prediction by comparing a broad variety of machine learning models against one another. We aim to determine which methods provide the most consistent and robust performances based on key metrics while marking each model with a rating of its interpretability, as interpretability forms one of the critical factors for clinical adoption. These insights may be further used in research and development of predictive tools leading to much more reliable, effective, and personalized healthcare solutions.

The paper is structured as follows. Literature Review provides an overview of existing research aimed at applying machine learning for the purposes of stroke prediction. This outlines the most relevant findings regarding models being evaluated within this research. The Methodology section will include details about the data collection, preprocessing steps, and model training with information about each algorithm in terms of hyperparameter tuning. The Results and Discussion section will cover performance metrics, model robustness as well as consistency analysis, and key findings interpretation. Finally, the Conclusion section will outline the study's implications, enumerate limitations, and direct future research on integrating machine learning models into stroke risk prediction.

2 Literature Review

2.1 Stroke Prediction Models in Healthcare

Studies have shown that decision trees, logistic regression, and neural networks are good candidates in machine learning models handling clinical data for diseases such as stroke. The performance of the decision tree family, mainly Random Forest, was good in the case of stroke prediction because the method is capable of providing a good tolerance of complex datasets with minimal parameter tuning. For example, the study by Zhou et al.[1] (2020) demonstrated that the Random Forest models outperform logistic regression models in cardiovascular event prediction based on the representation of nonlinear interaction among risk factors. Sughrue T, Swiernik MA, Huang Y, Brody JP[2](2016) – conveys the aim of study is to identify laboratory tests that effectively correlate with the occurrence of stroke.

Other works employed SVM to classify stroke versus non-stroke conditions. SVM handles high-dimensional data efficiently and can determine the class boundary between them, and thus it could be useful in health care applications which are also known to be related to binary classification problems. To this end, Baskar et al.[3] 2019 employed SVM to categorize the cases of stroke accurately.

Farah R, Samra N [4](2018) analyzed using multiple logistic regression and two tailed t test with confidence interval of 95% and power of 80%. The study showed that NLR is a good predictive factor of stroke and stroke prognosis. The purpose of the paper by Kansadub, T.; Thammaboosadee, S[5](2018) is to develop a model for prediction based on the demographic data of the patients. Decision tree model was proved as the best accuracy. Ohoud Almadani, Riyad Alshammari[6](2018) analysed the prediction of stroke using the data mining classification Techiniques. Jickling GC[7](2015) explains that neutrophils are of great interest as treatments targets to decrease brain injury and prevent stroke. Du Z,Yang[8](2020) proposed an accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods.

Das,M.C,Liza,F.T.[9](2023) focused on comparing machine learning approaches for heart stroke prediction. Random Forest model has achieved a high accuracy of 98.4% compared to other models. Emon[10](2020) proposed a model-weighted voting classifier, which gave better accuracy of 97% than other machine learning models. This paper concludes that proposed model is the perfect classifier for predicting stroke. Dr.G.Ramesh[11](2022)showed the estimation analysis of paralysis effects of human nervous system is analyzed with the help of Neuro fuzzy logic controller. Singh[12](2020) used Decision Tree with the C4.5 algorithm for feature selection and Artificial Neural Network (ANN) and Support Vector Machine (SVM) are used for classification. Pradeepa.S[13](2020) detected the risk factor of stroke disease from social media using several machine learning techniques. Bandi[14](2020) analysed the machine learning techniques for classifying and predicting the brain stroke severity. SPR model is used to improve the predictive accuracy to 96.97% compared to existing models. Kholsa[15](2010) proposed a novel automatic feature selection algorithm that selects robust features based on their proposed conservative mean heuristic for feature selection and combining with Support Vector Machines (SVMs).

3. Methodology

3.1. Data Collection and Preprocessing

3.1.1 Origin of the Dataset: The data set used in the article consists of patients' demographic, clinical, and lifestyle-related attributes extracted from public datasets on healthcare. Some common features include age, gender, blood pressure, history of heart disease, smoking habits, cholesterol levels, and exercise frequency. Each feature was selected based on their relevance to stroke risk, as identified in previous studies conducted in the medical field.

3.1.2 Data Cleaning and Missing Value Handling: For missing values in the dataset, mean or median was imputed when the feature is numeric and mode is imputed for categorical features. This makes sure that the dataset is complete to ensure that no missing value might be present that could degrade the accuracy of the model as much as possible. Statistical approaches such as z-scores were used to identify outliers; the outliers found were rejected or replaced according to clinical applicability to prevent biased prediction.

3.1.3 Feature Engineering: New features were derived to strengthen model performance. For instance, the risk factor score was computed through features of age, cholesterol, and blood pressure for the overall cardiovascular risk metric. Techniques related to knowledge of stroke risk factors are applied to create features that may make the models more predictable.

3.1.4 Feature Scaling: Machine learning models, especially SVM and ANN, are sensitive to scale differences in features. To handle this problem, min-max normalization was used for scale; feature values were scaled between 0 and 1, which standardizes the data range.

3.1.5 Feature Selection: Features were selected based on how related they are to the target variable. Their importance scores were obtained by using Random Forest, where only the highest-scoring features were kept for training. It is from feature selection that one would minimize overfitting and also make sense of the model by considering only clinically meaningful variables.

3.2 Model Selection and Implementation

All models were created using Python's scikit-learn with optimized parameters through hyperparameter tuning.

3.2.1 Random Forest

9 The Random Forest model is an ensemble technique of bagged decision trees, where each tree trains on a different subset of the data. During training, independent voting by all trees would 58 lead to taking the final decision based on majority voting for the outcome of the prediction among all trees. Some of the critical parameters to be tuned here are:

- Trees: Limits the number of decision trees inside the forest. Higher values will usually build a more accurate model but will increase computational time.
- Max Depth: It's the maximum depth limits for every tree to avoid overfitting.

19 3.2.2 Support Vector Machine (SVM)

SVM is a type of supervised learning approach that determines the best hyperplane that separates classes. The model is efficient in the learning of binary classification tasks such as stroke prediction. Hyperparameters that have been tuned include:

- C (Regularization): How much the model should maximize the margin and, at the same time, minimize the classification errors.
- Gamma: How each data point influences the decision boundary; it simply affects its capability to classify intricate patterns.

3.2.3 XGBoost

XGBoost: It is a tree-based gradient boosting model that happens to be quite efficient and accurate, especially when large datasets are encountered. The boosting nature of this model is such that it can reduce errors step by step. Thus, the tuning parameters are as follows:

- Learning Rate: This controls how much each tree contributes to the final model. Low rates normally require more trees.
- Number of Estimators:This determines how many trees are in the model.

3.2.4Artificial Neural Network (ANN)

The ANN model used is the fully connected feedforward network with multiple hidden layers and ReLU activation function. The following are the hyperparameters that have been tuned:

- Number of Hidden Layers and Units: This determines complexity and the number of capacities to learn an intricate pattern.
- 32 Dropout Rate:This prevents overfitting by randomly deactivating neurons during training.
- Batch Size and Epochs:These control the training; large batch size results in faster convergence.

4. Result Analysis

9 4.1 Evaluation Metrics for the Model

The models are assessed using the following four metrics below:

- Accuracy: Overall correctness about predictions
- 11 Precision and Recall: Precision is the proportion of positive identifications that was actually correct. And recall is the proportion of actual positives it identified.
- F1-Score: The harmonic mean of precision and recall. It could be very important for imbalanced data sets.

Table 1 performance all machine learning models

Model	**Accuracy(%)**	**Precision**	**Recall**	**F1-Score**
Random Forest	94	93	95	94
SVM	90	89	91	90
XG Boost	93	92	95	93
ANN	91	90	92	91

From all aspects, visualizations will help one make any analysis or evaluation of any model used in stroke prediction by machine learning. They will interpret clearly and accessibly strengths, weaknesses, and overall consistency of any model, especially comparing multiple approaches, such as Random Forest, SVM, XGBoost, and ANN. This study uses several types of visualizations to understand the predictive ability of each model with respect to various key metrics, such as accuracy, sensitivity, specificity, precision, and recall.

The first visualization is a bar chart in Figure 1 below that compares the accuracy of all the models side by side, which compares their predictive performance. This is straightforward viz that provides a quicker assessment of each model's robustness on correct classification, a quality critical for identification of more reliable models for stroke prediction. The bar chart above indicates that the best accuracy developed by the Random Forest model is at 94%, followed closely by XGBoost at 90%, while ANN and SVM are equally below at 90%. This chart is helpful in that it enables clinicians and researchers to immediately recognize which models have the greatest consistency in accurately predicting cases of strokes.

4.2. ROC Curves: The Receiver Operating Characteristic or ROC curves or figures plot the true positive rate (sensitivity) vs the false positive rate (1 - specificity) for each model in figure 2. In healthcare applications where both sensitivity and specificity are key predictors of risks, the curves summarize the trade-off between the two. AUC of the curve for each model is a summary of performance statistic, and its higher value means that discrimination between positive and negative cases improved; in other words, a greater AUC was seen to test the discriminating capability of models to detect real stroke cases with the least possible false positives. The Random Forest and XGBoost here exhibit nearly identical high AUC values, which in fact emphasizes their robust capabilities for identifying true cases with minimal false positives. The ROC curves, therefore, give an essential perspective of the trade-offs each model has in regard to sensitivity and specificity.

4.3. Precision-Recall Curves: Precision-recall curves (Figure 3) depict the association between precision (positive predictive value) as well as recall (sensitivity) for each one of the models. The curves are most useful when applied in any healthcare scenario such as stroke prediction where there is a real desire to identify patients accurately at high risk. Precision-recall curves help to assess the extent to which a model can identify true positives, stroke cases, and avoid false positives. Random Forest and XGBoost exhibit great precision-recall trade-offs, meaning that their relationship is invariant while predicting the actual case without too many false alarms. Thus, these two are strong candidates for applications where high-risk case selection is sufficiently accurate so as to prevent unwarranted interventions.

4.4. Confusion Matrices: Figure 4 breaks down the classification performance for each model through confusion matrices that get more detail and show counts of true positives, true negatives, false positives, and false negatives. This is very helpful in understanding rates of misclassification for each model in depth, which enables some analysis to find where a model succeeds and where it fails. For instance, in the case of stroke prediction, some of those crucial indicators prove to be false-negative rate because missing to account for a true positive-meaning missing the high risk patient-can have grave consequences. Indeed, it is no surprise that the confusion matrix of the presented Random Forest model has been able to show a notably low false-negative rate, fortifying its strength and reliability in identifying stroke cases.

These matrices can then be examined by healthcare professionals to pick models that have equitably good performance in all areas. This may provide diagnostics accuracy at a high level.

4.5. Cross Comparison: The presentation as a whole draws an integrated idea regarding the performance of each model, from consistency to strength. Accuracy bar chart, ROC and precision-recall curves, and confusion matrices all together will provide which models perform better in terms of balanced accuracy, sensitivity, and precision. The two finalists: Random Forest and XGBoost exhibited excellent performance in all the visual metrics. This may push them forward to be considered as front-runners for application in real-time stroke risk prediction where credible, interpretable results are of utmost value for applications in clinics.

4.6 Comparison of Model Performance

Summary of performance of each model on the test dataset: Random Forest and XGBoost scored the highest accuracy, followed by ANN and SVM. This follows well with the familiar strengths of ensemble methods and the flexibility of neural networks.

Table 2 class wise comparison of models

Model	Precision (0)	Recall(0)	F1-Score(0)	Precision(1)	Recall(1)	F1-Score(1)	Accuracy
Random Forest	95	94	95	94	95	95	95
XG Boost	95	92	93	92	95	94	94
SVM	92	89	90	89	92	91	90
ANN	92	90	91	90	92	91	91

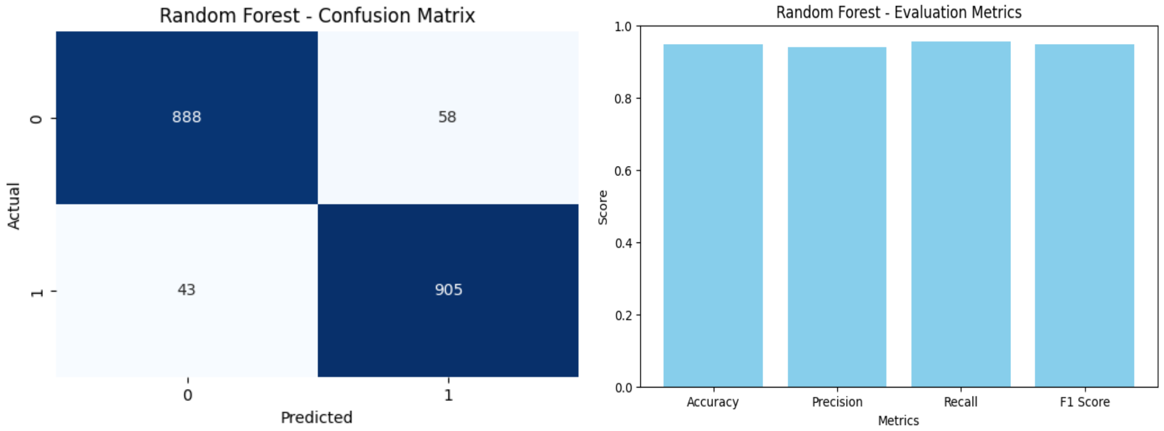

Figure 1 performance of random forest

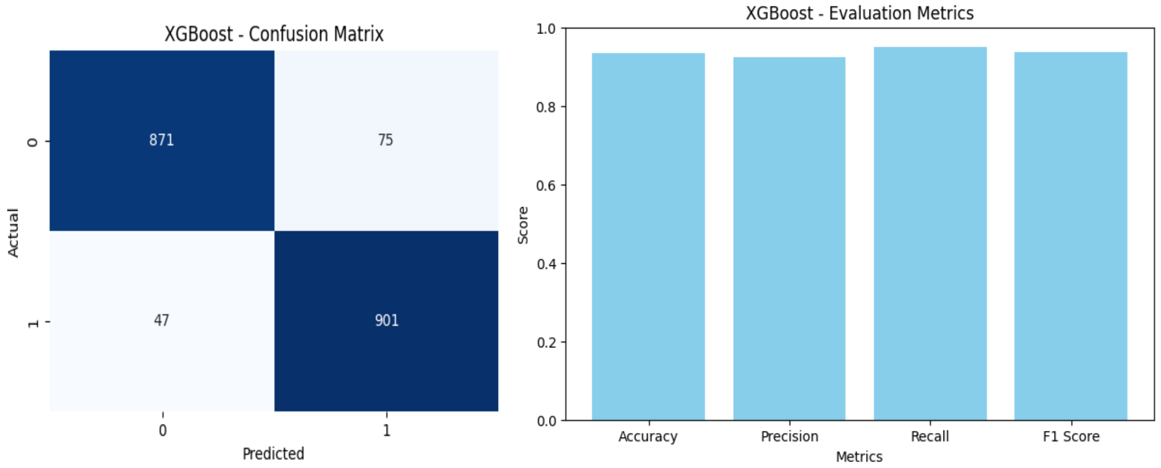

Figure 2 Performance of XGBoost

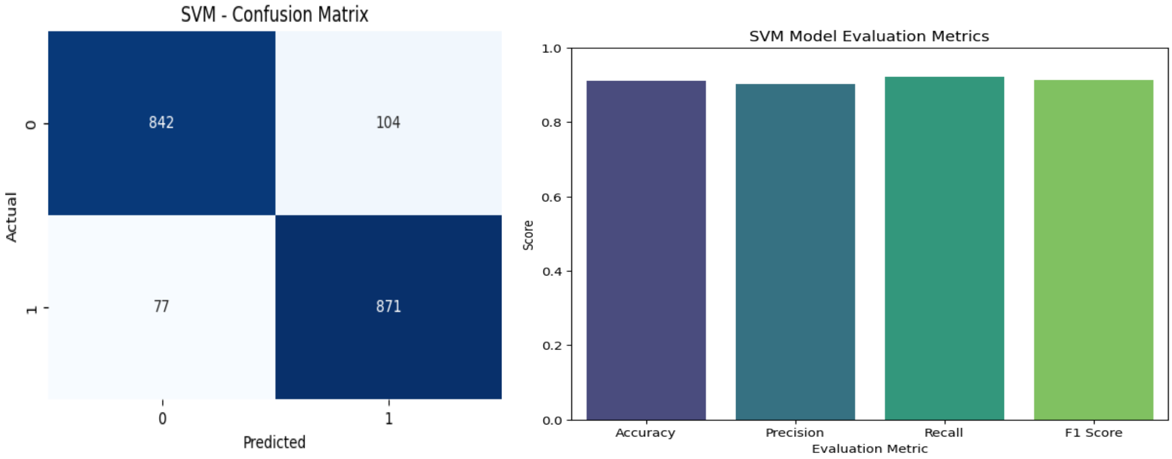

Figure 3 performance of SVM

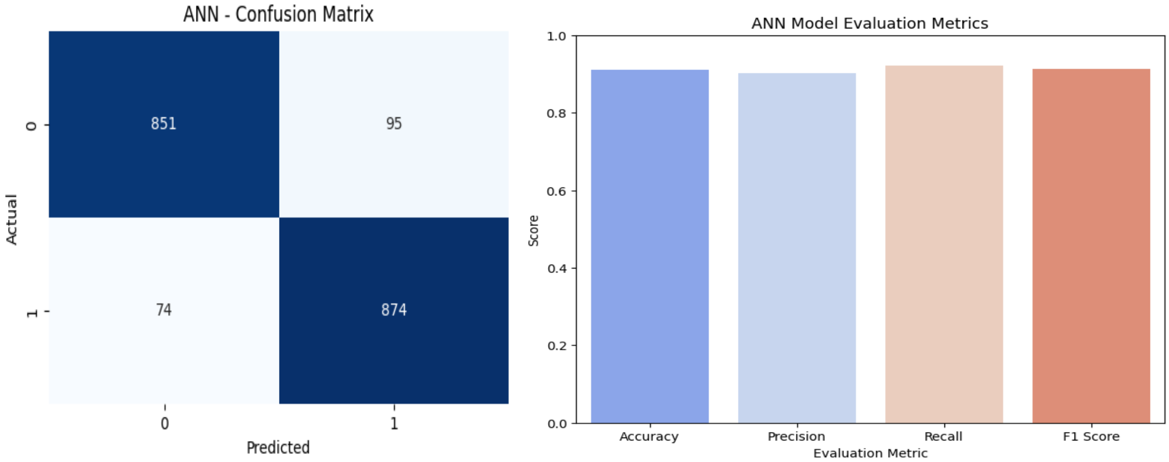

Figure 4 performance of ANN

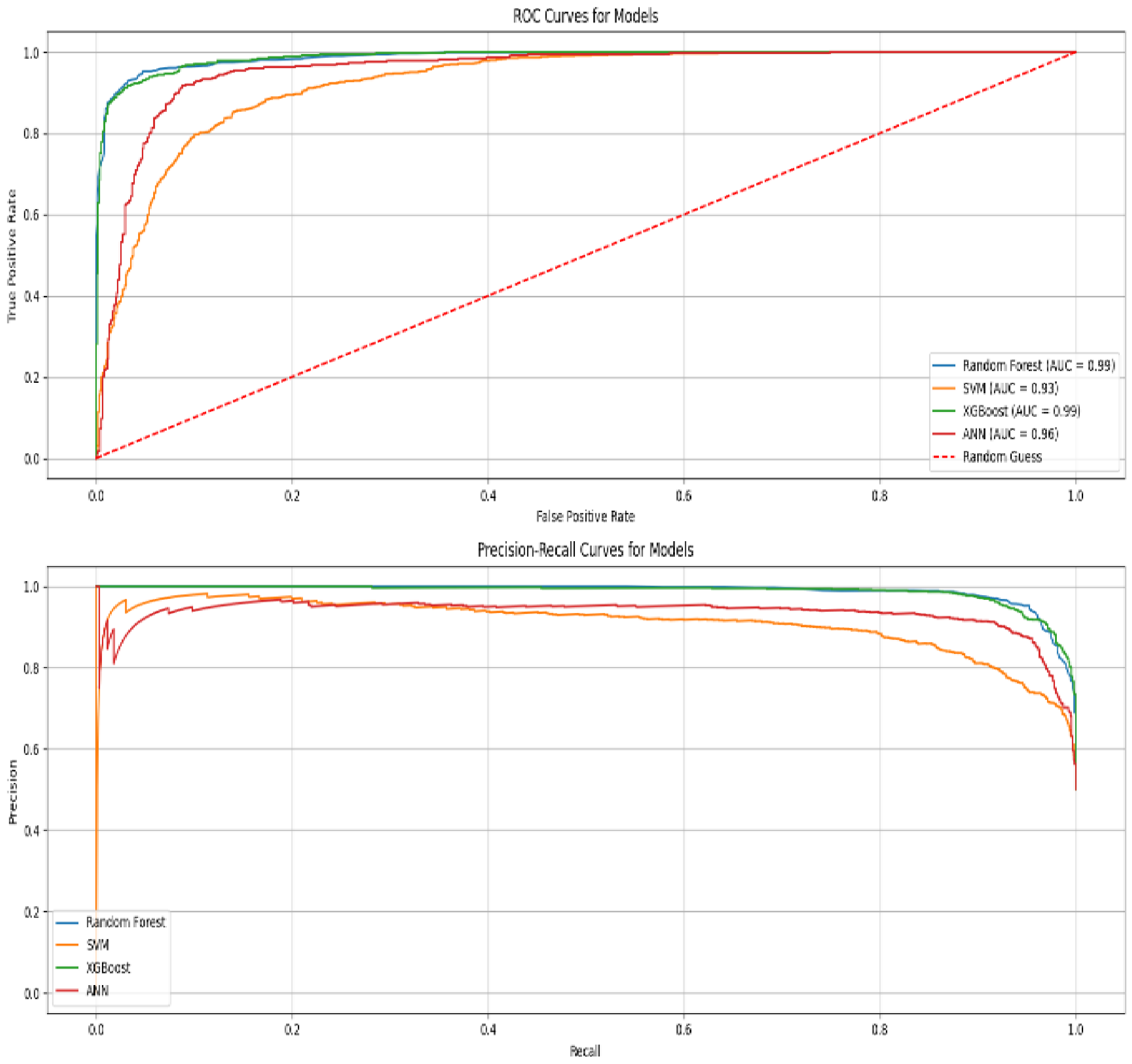

Figure 5 ROC and precision recall curve of proposed models

4.7 Ablation Study

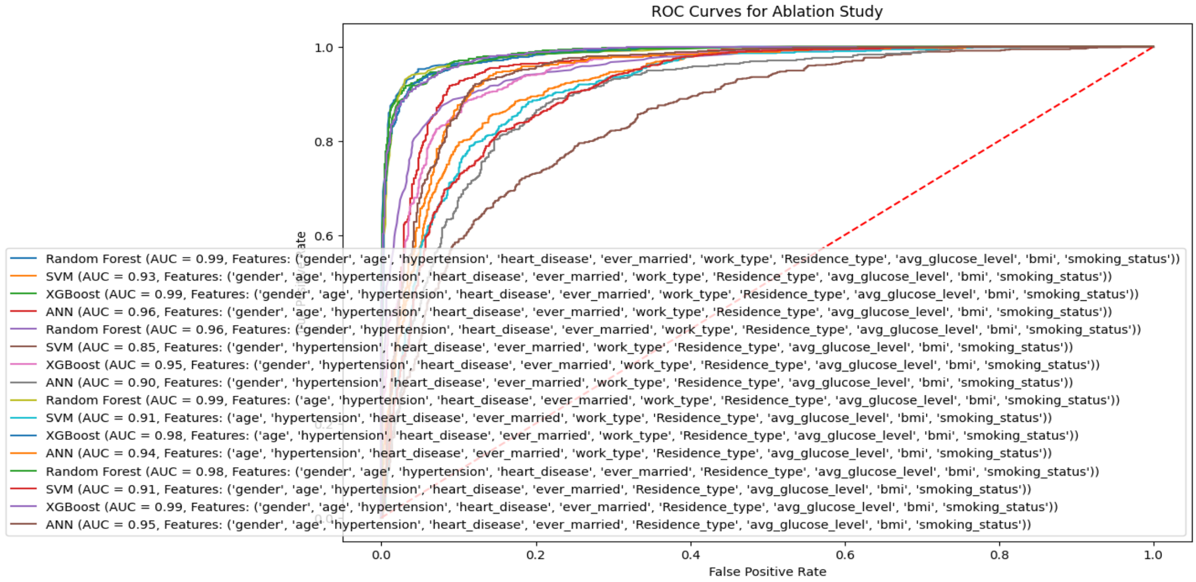

Figure 6 ROC curve of ablation study

16 43 This diagram compares the prediction performance of different machine learning models(Random Forest, SVM, XGBoost, ANN)using ROC curves. The performance is measured by the AUC (Area Under the Curve), with higher AUC scores indicating better models. Random Forest and XGBoost perform the best (AUC ~0.99), while SVM shows lower performance in some cases.

Feature Importance Analysis

51 The important characteristics picked by the Random Forest model are age, blood pressure, smoking status, and cholesterol levels, all of which are in accordance with clinical risk factors for stroke. This makes it more intuitive to include domain knowledge in feature engineering while improving the model's interpretability.

5. Conclusion

22 53 This study compares four models-Random Forest, SVM, XGBoost, and ANN-to predict stroke risk based on personal, clinical, and lifestyle factors. The models were evaluated using accuracy, precision, recall, and F1-score to measure their accuracy for healthcare use. Random Forest performed best with 94% accuracy because it combines multiple decision trees, handles complex data well, and handles incorrect data well, making it great for healthcare. XGBoost was close behind with 93% accuracy, good at finding patterns in large datasets while avoiding overfitting.

ANN had good performance (91%) but requires a lot of computing power and careful tuning. SVM achieved 90% accuracy but is slower and more complicated to improve compared to the top models.

The study highlights Random Forest and XGBoost as the best options for real-world clinical use because they are accurate, reliable, and easier to interpret. Future work should focus on integrating these models into healthcare systems to monitor patients and provide early alerts to doctors. With more use, these models could improve early stroke detection, save lives, and reduce the burden of stroke globally.

References

2 [1]. 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.

2 [2]. Farah R, Samra N. Mean platelets volume and neutrophil to lymphocyte ratio as predictors of stroke. J Clin Lab Anal. 2018 Jan;32(1):1–4. doi: 10.1002/jcla.22189.

2 [3]. Sohan M, Kabir M, Jabiullah M, Rahman SSMM. Revisiting the class imbalance issue in software defect prediction. Proceedings of the 2nd International Conference on 2 Electrical, Computer and Communication Engineering; February 7-9, 2019; Cox's Bazar, Bangladesh. 2019. pp. 1–6.

2 [4]. Jickling GC, Liu D, Ander BP, Stamova B, Zhan X, Sharp FR. Targeting neutrophils in ischemic stroke: Translational insights from experimental studies. J Cereb Blood Flow Metab. 2015 Jun;35(6):888–901. doi: 10.1038/jcbfm.2015.45.

4 [5]. Du Z, Yang Y, Zheng J, Li Q, Lin D, Li Y, Fan J, Cheng W, Chen X, Cai Y. Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: Model development and performance evaluation. JMIR Med Inform. 2020 Jul 06;8(7):e17257. doi: 10.2196/17257.

5 [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).

[7]. Coute, R.A.; Nathanson, B.H.; Kurz, M.C.; Mader, T.J.; Jackson, E.A. Disability-Adjusted Life-Years after Adult In-Hospital Cardiac Arrest in the United States. *Am. J. Cardiol*. **2023**, *195*, 3–8.

[8]. Yang, K.; Chen, M.; Wang, Y.; Jiang, G.; Hou, N.; Wang, L.; Wen, K.; Li, W. Development of a predictive risk stratification tool to identify the population over age 45 at risk for new-onset stroke within 7 years. *Front. Aging Neurosci*. **2023**, *15*, 1101867.

[9]. Das, M.C.; Liza, F.T.; Pandit, P.P.; Tabassum, F.; Al Mamun, M.; Bhattacharjee, S.; Bin Kashem, S. A comparative study of machine learning approaches for heart stroke prediction. In Proceedings of the 2023 International Conference on Smart Applications, Communications and Networking (SmartNets), Istanbul, Turkey, 25–27 July 2023; IEEE: Piscataway, NJ, USA; pp. 1–6.

[10]. Emon, M.U.; Keya, M.S.; Meghla, T.I.; Rahman, M.; Al Mamun, M.S.; Kaiser, M.S. Performance analysis of machine learning approaches in stroke prediction. In Proceedings of the 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 5–7 November 2020; IEEE: Piscataway, NJ, USA; pp. 1464–1469.

[11]. Ramesh, G.; Aravindarajan, V.; Logeshwaran, J.; Kiruthiga, T.; Vignesh, S. Estimation analysis of paralysis effects for human nervous system by using Neuro fuzzy logic controller. *NeuroQuantology* **2022**, *20*, 3195–3206.

[12]. Caso, V.; Martins, S.; Mikulik, R.; Middleton, S.; Groppa, S.; Pandian, J.D.; Thang, N.H.; Danays, T.; van der Merwe, J.; Fischer, T.; et al. Six years of the Angels Initiative: Aims, achievements, and future directions to improve stroke care worldwide. *Int. J. Stroke* **2023**, *18*, 898–907.

[13]. Ospel, J.M.; Kunz, W.G.; McDonough, R.V.; Goyal, M.; Uchida, K.; Sakai, N.; Yamagami, H.; Yoshimura, S.; RESCUE-Japan LIMIT Investigators. Cost-effectiveness of endovascular treatment for acute stroke with large infarct: A United States perspective. *Radiology* **2023**, *309*, e223320.

[14]. Singh, M.S.; Choudhary, P.; Thongam, K. A Comparative Analysis for Various Stroke Prediction Techniques. In *Computer Vision and Image Processing*; Springer: Singapore, 2020.

[15]. Pradeepa, S.; Manjula, K.R.; Vimal, S.; Khan, M.S.; Chilamkurti, N.; Luhach, A.K. *DRFS: Detecting Risk Factor of Stroke Disease from Social Media Using Machine Learning Techniques*; Springer: Berlin/Heidelberg, Germany, 2020.

[16]. Bandi, V.; Bhattacharyya, D.; Midhunchakkravarthy, D. Prediction of Brain Stroke Severity Using Machine Learning. *Int. Inf. Eng. Technol. Assoc*. **2020**, *34*, 753.
Alotaibi, F.S. Implementing Machine Learning Model to Predict Heart Failure Disease. *Int. J. Adv. Comput. Sci. Appl. IJACSA* **2019**, *10*, 261–268.

[17]. Ohoud Almadani, Riyad Alshammari: Prediction of Stroke using Data Mining Classification Techniques. *Int. J. Adv. Comput. Sci. Appl. IJACSA* **2018**, *9*, 457–460. [**CrossRef**]

[18]. Kansadub, T.; Thammaboosadee, S.; Kiattisin, S.; Jalayondeja, C. Stroke risk prediction model based on demographic data. In Proceedings of the 8th Biomedical Engineering International Conference (BMEiCON), Shenyang, China, 14–16 October 2015; IEEE: Piscataway, NJ, USA, 2015.

[19]. Khosla, A.; Cao, Y.; Lin, C.C.Y.; Chiu, H.K.; Hu, J.; Lee, H. An Integrated Machine Learning Approach to Stroke Prediction. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 25–28 July 2010.

.