RESEARCH ARTICLE OPEN ACCESS

Predicting Stroke Using Data Mining and Machine Learning Methods: A Comparative Analysis

Vaishali Sarde [1], Pankaj Sarde [2]

Department of Computer Application, Govt. J. Yoganandam Chhattisgarh College Raipur

Department of Mathematics, Rungta College of Engineering & Technology, Bhilai (CG), India

ABSTRACT

Stroke is same to the brain as a heart attack is to the heart. In case of a stroke, part of the brain loses the blood supply, and that brain area stops getting oxygen. Without oxygen supply, the affected brain cells become oxygen-starved and stop working properly.If the brain cells don't get oxygen for a long, they will die. If adequate brain cells in particular brain area die, the damage may become permanent, and person may lose the capacities related with that area. However, restarting blood flow may limit the severe damage. Time is very critical factor in treating a stroke.With the availability of huge volume of datasets of medical records, data mining techniques can be applied to unfold the trends hidden in the dataset. Analysis of such data may help medical practitioners in doing prognosis of any critical medical conditions. Data mining and machine learning techniques for medical records can lead to major improvements in healthcare system. Earlier stage diagnosis of disease can save the life.

This paper uses five different techniques from data mining and machine learning - KNN, Support Vector Machine, decision Tree, Naive Bayes and Artificial Neural Network to predict the stroke. Comparative performance analysis of algorithms, is presented. The measures used for the performance analysis are Accuracy, Precision, Recall, f1-score and Support. For experiment stroke dataset is taken from Kaggle's data repository[4] . It has 5110 records of patients. Result shows that SVM and ANN performs well as compared to other three data mining and machine learning algorithms.

Keywords — KNN, Support Vector Machine, decision Tree, Naive Bayes and Artificial Neural Network, Machine Learning

I. INTRODUCTION

According to WHO Annually, 15 million people worldwide suffers with stroke. Of these, 5 million die and another 5 million are left permanently disabled,

Centers for Disease Control and Prevention (CDC) declares stroke as the fifth-leading cause of death in the United States. Stroke is responsible for around 11% of total deaths. Consistently, over 795,000 people in United States gone through the adverse effect of stroke [1][2][3].

A stroke is a neurological condition, occurs when blockage or bleed of the blood vessels either stops or reduces the blood supply to an area of the brain. This results in lower oxygen availability to the the bain and brain cells start to die. Strokes fall into three main categories. Ischemic stroke involves, the arteries supplying blood to the brain get either narrow or blocked. A transient ischemic attack occurs when blood flow to the brain is blocked temporarily. A hemorrhagic stroke happens of blood through the arteries in the brain. According to the American Heart Association, about 13 percent Trusted Source of strokes are hemorrhagic. This paper predicts the occurrence of stroke using various classification methods from machine learning and data mining. Dataset is taken from Kaggle's data repository [4] with 5110 records. Dataset includes 11 parameters with 2 binary classes stroke or no stroke. Five techniques KNN, decision Tree, Support Vector Machine, Naive Bayes and Artificial Neural Network are applied to the dataset for the prediction of stroke. Experiment's result shows SVM and ANN performs well as compared to other three data mining and machine learning algorithms. Accuracy, Precision, Recall, f1-score and Support are taken as measures for performance analysis of algorithms.

II. RELATED WORK

Minhaz et al. [5] used the ten algorithms to train the model. Data is taken from hospitals of Bangladesh. To improve the performance of all classifiers Weighted voting classifier is used. After optimization best model is discovered. [9] Bandi V. et. al proposed an improvised random forest method for the task of stroke prediction. Different levels of risks associated with stoke have been analysed. Research is conducted for limited types of strokes. Hung et al. in [12] focused on comparison of deep learning models and machine learning models for prediction of stroke using electronic medical claims database.

Wu et al. [15] proposed a model for stroke prediction. In this study authors applied regularized logistic regression, support vector machine, and random forest models to balanced and imbalanced dataset both for stroke prediction. Best results from both datasets are identified and compared.

Studies given in [19,20] stated that identification of important features can improve the performance of machine learning algorithm. It is necessary to identify combination of features affects most the classification instead of using all features. This shows the interdependency of features in prediction of stroke.

In [21] Li, X et. al proposed a methodology to find out the different symptoms of stroke disease and various preventive measures for a stroke from social media resources. They provided a method for clustering tweets using spectral clustering based on the content iteratively.

In [22] patients' electronic health records are analysed to recognize the effect of risk factor involve in prediction of stroke. Random forest, Neural Network are applied over the dataset of electronic health records.

III. DATASET

3.1 Source of Dataset

Stroke dataset (Comprehensive) is taken from Kaggle repository [4]. This dataset consists of 11 parameters like gender, age, hypertension, and heart_disease etc. Based on parameters it can be predicted whether a patient is likely to get stroke not.

3,2 Features used in Dataset

11 parameters given in the table 1 are used for the experimental purpose.

Table1: Features used in experiment

S. No.	Feature
1	gender
2	age
3	hypertension
4	heart_disease
5	ever_married
6	work_type
7	Residence_type
8	avg_glucose_level
9	bmi
10	smoking_status
11	stroke

Table 1 shows the various features used for the prediction of strokes. Description of features is as follows- id: unique identifier, gender, age: age of the patient, hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension, heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease, ever_married, work_type: "children", "Govt_job", "Never_worked", "Private" or "Self-employed", Residence_type: "Rural" or "Urban", avg_glucose_level: average glucose level in blood, bmi: body mass index, smoking_status, stroke: 1 if the patient had a stroke or 0 if not.

3.3 Data Preprocessing

Data preprocessing is most important task for data mining and machine learning process. Data preprocessing improves the quality of data that can give better result. For the from Kaggle's data repository [4] data preprocessing is performed in two steps.

1. **Converting text data into numerical form.**

Some of feature values are in the textual form which has been converted into numerical form.

2. **Replacing missing Values**

BMI values are missing in many places which have been replaced by the average value of BMI for the same age group.

3. **Splitting dataset**

Dataset is separated between training and testing data set. 67% data is used for training purpose and remaining 33% data is used for testing purpose.

3.4 Dataset Statistics

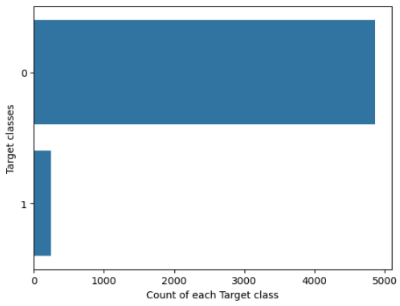

Fig 1: Counting of each Target Class- Stroke(class-1), Not Stroke (class 0)

Fig 1 represents number of records corresponding to Stroke and not stroke class.

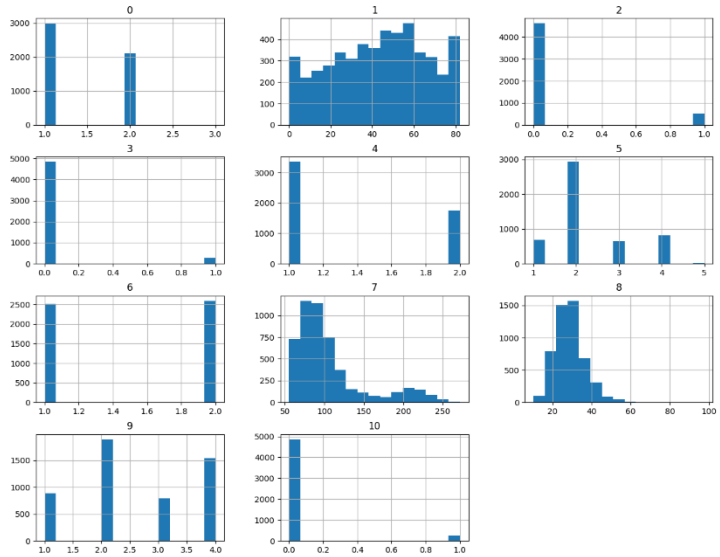

Fig 2 : Histogram of features

Fig 2 represents the distribution of data for each of the feature.

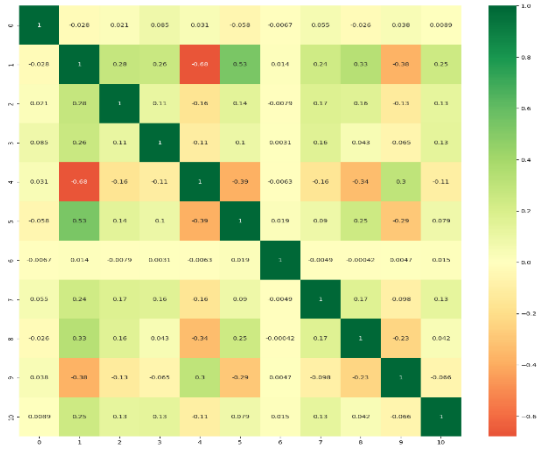

Fig 3: HeatMap

Fig 3 represents the heat map to relate feature values.

IV RESEARCH METHODOLOGY

4.1 Overall Process Model

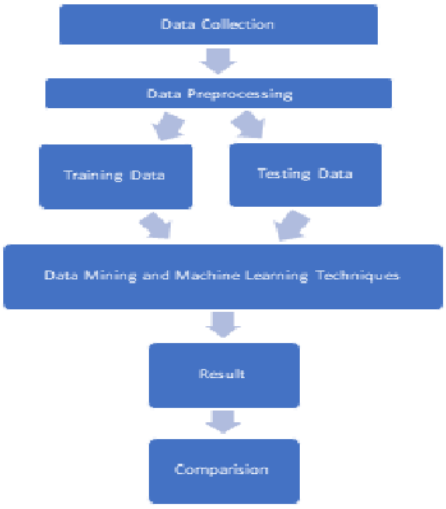

Fig 4: Process Model

Fig 4 shows the overall process model of the proposed system. Firstly, data is collected from Kaggle's data repository [4]. Next the preprocessing of data takes place by converting textual data into numerical form. Missing values are replaced with the appropriate values. In the next step data is divided into training and test data. Further various techniques from data mining and machine learning have been applied on the real dataset for the prediction of strokes. Outcomes of different algorithms are compared to find the best result.

4.2 K-Nearest Neighbour (KNN)

KNN is very simple supervised Machine learning algorithm. It is based on the concept that new data is assigned to the category which has the maximum number of similar data available. K-NN is a non-parametric algorithm. Which means for the distribution of data no underlying assumptions is made.

The K-NN can be described using the following steps:

Step 1 – Firstly Choose the value of K for nearest neighbours.

Step 2 – For each test data point do the following steps-

- Calculate the distance between current test data point and all rows of training data. Distance can be: Manhattan, Euclidean or Hamming distance.
- Sort the rows based on distance value.
- Choose top K sorted rows.
- Test data point will be assigned to the class which is having maximum nearest training data.

Step 3-End

4.3 Decision Tree

Decision Tree is a Supervised learning technique. It is used for classification and Regression both. It has a tree-like structure, where internal nodes are represented by the features available in the dataset, branches are represented by the decision rules and leaf nodes are represented by the result.

It is a graphical way to represent the all possible solutions of given problem on the basis of the feature values. Figure 5 represents the decision tree structure.

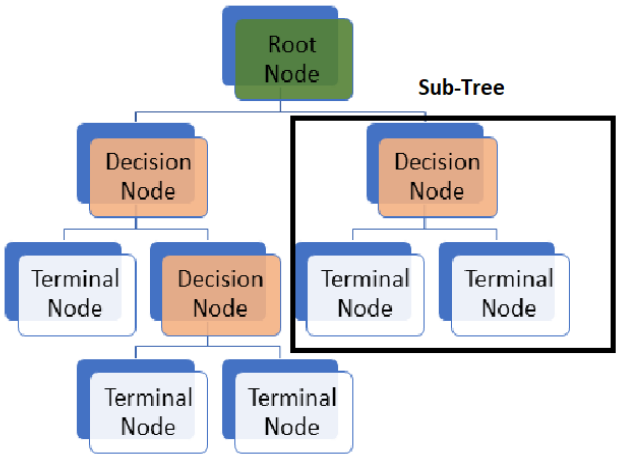

Fig 5: Decision Tree.

.

4.4 Support Vector Machine

Support Vector Machine (SVM) is used for classification. It is a supervised machine learning algorithm. SVM algorithm is used to find the optimal hyperplane to separate the data points in different classes. SVM tries to maintain gap between the nearest points of different classes to be as large as possible. Following figure shows how the hyperplane separates the data point.

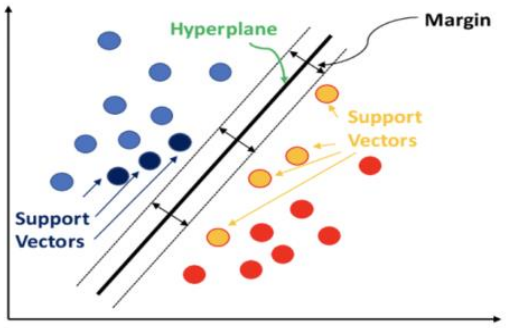

Fig 6: Support Vector Machine

4.5 Navie Bayes

Naive Bayes algorithm is used for classification problem. It is a supervised learning algorithm. It uses the concept of Bayes theorem for classification problem.

Bayes theorem calculate the probability of the current event using the given probability of already occurred event.

It is called as a probabilistic classifier, that means prediction is done on the basis of the probability of an object.

The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)\mathrm{P(A)}}{\mathrm{P(B)}}$$

Where,

P(A|B) -Posterior probability- Probability of hypothesis A on the observed event B.

P(B|A) -Likelihood probability-Probability of the evidence

4.6 Artificial Neural Network (ANN)

Artificial Neural Networks consists of various layers of artificial neurons also called units. ANN has three types of layers, input layer, hidden layers and output layers. Input layer takes the real world data and passed it to multiple hidden layers which process and transform the data and passed it to the output layer. Output layer holds the units belongs to the different classes.

The proposed model for this paper is as follows

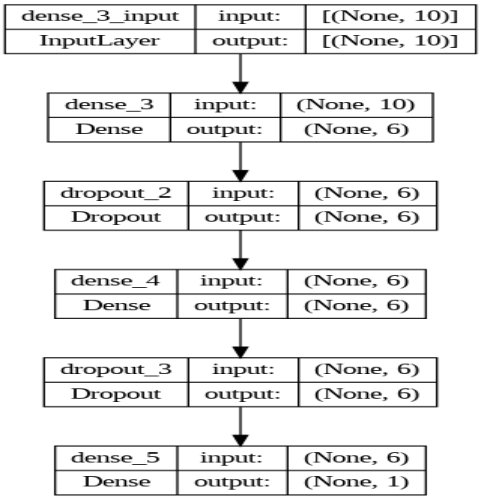

Fig 7: Proposed ANN Model

V. RESULTS AND DISCUSSION

5.1 K-Nearest Neighbour (KNN)

Table 2 : KNN Performance

	precision	recall	f1-score	support
0	0.96	0.86	0.91	4861
1	0.12	0.37	0.18	249
accuracy			0.84	5110
macro avg	0.54	0.61	0.55	5110
weighted avg	0.92	0.84	0.88	5110

Above table shows the performance of the KNN algorithm on the basis of precision, recall f1-score and support. For class 0 and 1 i.e. not stroke and stroke class, all the four measures are given in the table. Table shows the accuracy of KNN algorithm is 84%.

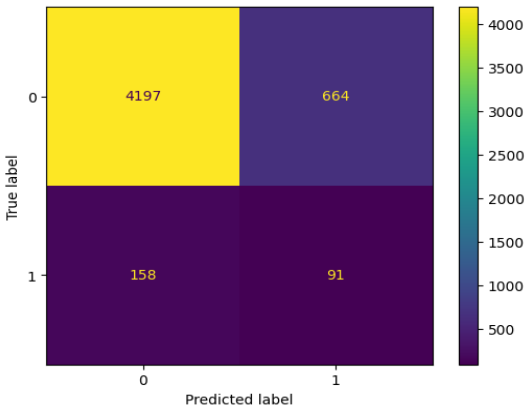

Fig 8: KNN Confusion Matrix

Fig 8 shows the confusion matrix for KNN algorithm. Diagram shows out of 4861 not stroke data 4197 are correctly classified and out of 249 stroke data 91 are correctly classified.

5.2 Decision Tree

Table3: Decision Tree

	precision	recall	f1-score	support
0	0.96	0.95	0.95	1606
1	0.16	0.19	0.17	81
accuracy			0.91	1687
macro avg	0.56	0.57	0.56	1687
weighted avg	0.92	0.91	0.92	1687

Above table shows the performance of the Decision Tree algorithm on the basis of precision, recall f1-score and support. For class 0 and 1 i.e. not stroke and stroke, all the four measures are given in the table. Table shows the accuracy of Decision Tree algorithm is 91%.

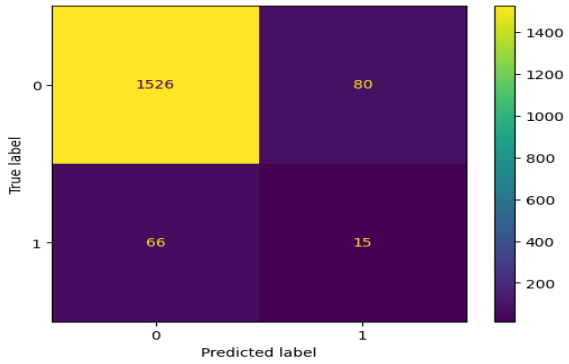

Fig 9: Decision Tree Confusion Matrix

Fig 9 shows the confusion matrix for Decision Tree algorithm. Diagram shows out of 1606 not stroke test data 80 are correctly classified and out of 81 stroke test data 15 are correctly classified. It shows the best result in comparison with other algorithms.

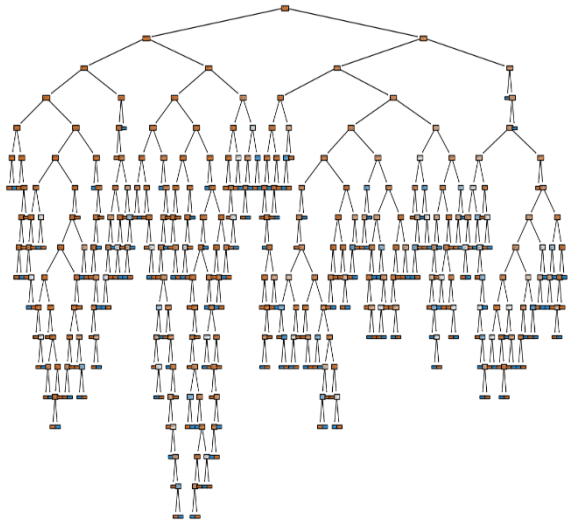

Fig 10: Decision Tree Model

Fig 10 represents the decision tree structure for the real dataset[4] used in the experiment.

5.3 Support Vector Machine (SVM)

Table 4: SVM Performance

	precision	recall	f1-score	support
0	0.95	1.00	0.98	1606
1	0.00	0.00	0.00	81
accuracy			0.95	1687
macro avg	0.48	0.50	0.49	1687
weighted avg	0.91	0.95	0.93	1687

Above table shows the performance of the Support Vector Machine (SVM) algorithm on the basis of precision, recall f1-score and support. For class 0 and 1 i.e. not stroke and stroke all the four measures are given in the table. Table shows the accuracy of SVM algorithm is 95%.

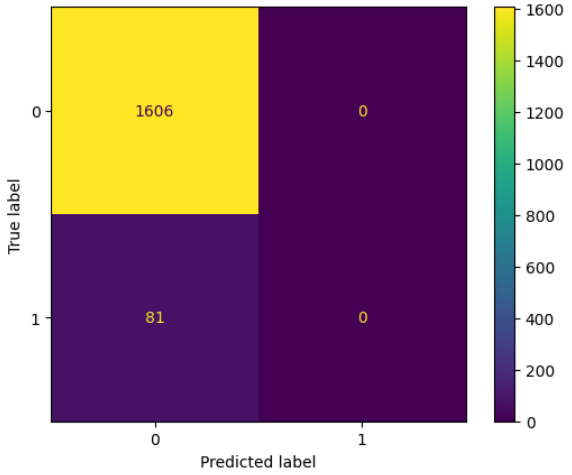

Fig 11: SVM Confusion Matrix

Fig 11 shows the confusion matrix for Support Vector Machine (SVM) algorithm. Diagram shows out of 1606 not stroke test all are correctly classified and out of 81 stroke data no one are correctly classified.

5.4 Navie Bayes

Table 5: Navie Bayes Performance

	precision	recall	f1-score	support
0	0.96	0.90	0.93	1606
1	0.11	0.26	0.16	81
accuracy			0.87	1687
macro avg	0.54	0.58	0.54	1687
weighted avg	0.92	0.87	0.89	1687

Above table shows the performance of the Navie Bayes algorithm on the basis of precision, recall f1-score and support. For class 0 and 1 i.e. not stroke and stroke class all the four measures are given in the table. Table 5 shows the accuracy of Navie Bayes algorithm is 87%.

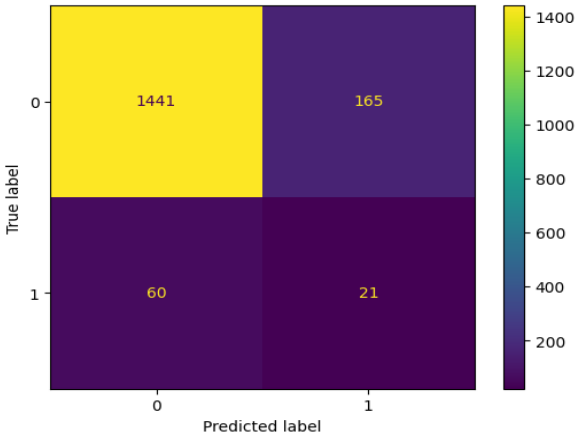

Fig 12: Navie Bayes Confusion Matrix

Fig 12 shows the confusion matrix for Navie Bayes algorithm. Diagram shows out of 1606 not stroke test data 1441 are correctly classified and out of 81 stroke test data 21 are correctly classified.

5.5 Artificial Neural Network (ANN)

Table 6: Proposed ANN Model Performance

	precision	recall	f1-score	support
0	0.95	1.00	0.98	1606
1	0.00	0.00	0.00	81
accuracy			0.95	1687
macro avg	0.48	0.50	0.49	1687
weighted avg	0.91	0.95	0.93	1687

Above table shows the performance of the Artificial Neural Network (ANN) algorithm on the basis of precision, recall f1-score and support. For class 0 and 1 i.e. not stroke

and stroke all the four measures are given in the table. Table shows the accuracy of Artificial Neural Network (ANN) algorithm is 95%.

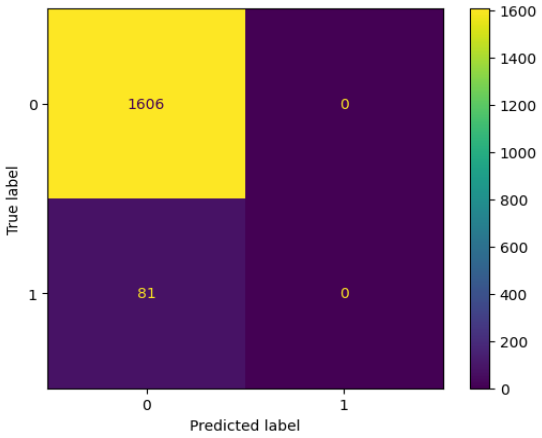

Fig 13: Proposed ANN Model Confusion Matrix

Fig 13 shows the confusion matrix for Artificial Neural Network (ANN). Diagram shows out of 1606 not stroke test data all are correctly classified and out of 81 stroke test data no one are correctly classified.

5.6 Comparative analysis of algorithms

Table 7: Comparison Chart of five Algorithms based on measures Precision, recall, f1-score and Accuracy

	Class	precision	recall	f1-score	Accuracy
KNN	**0**	0.96	0.86	0.91	0.84
	1	0.12	0.37	0.18	
Decision Tree	**0**	0.96	0.95	0.95	0.91
	1	0.16	0.19	0.17	
SVM	**0**	0.95	1.00	0.98	0.95
	1	0.00	0.00	0.00	
Navie Bayes	**0**	0.96	0.90	0.93	0.87
	1	0.11	0.26	0.16	
ANN	**0**	0.95	1.00	0.98	0.95
	1	0.00	0.00	0.00	

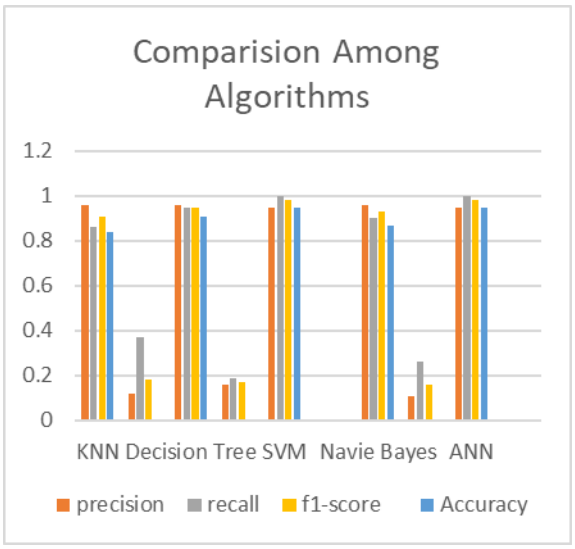

Fig 14: Comparison among algorithms

Fig14 shows the comparative chart of four performance measures of all the five algorithms. As chart shows SVM and ANN performs best among all other three algorithms. Still there is a scope to modify ANN model for the better result.

VI. CONCLUSION AND FUTURE SCOPE

A stroke is a medical condition occurs when blood supply to the part of brain stops. It is very serious life-threatening condition. Prediction of stroke based on medical parameter can save the life in some extent.

In this paper we focused on five different machine learning and datamining techniques KNN, decision Tree, Support Vector Machine, Naive Bayes and Artificial Neural Network. Data has been taken from the Kaggle's data repository[4] for the experimental purpose. As the result shows SVM and ANN gives 95% accuracy for the prediction of strokes which is the best result as compare to other algorithms. However proposed ANN model can be modified for the more accurate result.

REFERENCES

[1] https://www.cdc.gov/stroke/data-research/facts-stats/index.html. Statistics of Stroke by Centers for Disease Control and Prevention.

[2] Roger, V. L., Go, A. S., Lloyd-Jones, D. M., Adams, R. J., Berry, J. D., Brown, T. M., et al. (2011). Heart disease and stroke statistics—2011 update. *Circulation, 123*, e18–e209.

[3] National Stroke Association. (2011). Stroke 101 fact sheet. [Google Scholar].

[4] Kaggle. (n.d.). Stroke prediction dataset. Retrieved from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[5] Emon, M. U., Keya, M. S., Meghla, T. I., Rahman, M. M., Al Mamun, M. S., & Kaiser, M. S. (2020). Performance analysis of machine learning approaches in stroke prediction. In *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA 2020)* (pp. 1464–1469).

[6] Chun, M., et al. (2021). Stroke risk prediction using machine learning: A prospective cohort study of 0.5 million Chinese adults. *Journal of the American Medical Informatics Association, 28*(8), 1719–1727.

[7] Hertel, R., & Benlamri, R. (2022). A deep learning segmentation-classification pipeline for x-ray-based COVID-19 diagnosis. *Biomedical Engineering Advances*, 100041.

[8] Ray, S. (2019). A quick review of machine learning algorithms. In *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Perspectives and Prospects* (Com. 2019) (pp. 35–39).

[9] Bandi, V., Bhattacharyya, D., & Midhun Chakkravarthy, D. (2020). Prediction of brain stroke severity using machine learning. *International Information and Engineering Technology Association*.

[10] Pathan, M. S., Jianbiao, Z., John, D., Nag, A., & Dev, S. (2020). Identifying stroke indicators using rough sets. *IEEE Access, 8*, 210318–210327.

[11] Khosla, A., Cao, Y., Lin, C.-Y., Chiu, H.-K., Hu, J., & Lee, H. (2010). An integrated machine learning approach to stroke prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 183–192).

[12] Hung, C.-Y., Chen, W.-C., Lai, P.-T., Lin, C.-H., & Lee, C.-C. (2017). Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 3110–3113). IEEE.

[13] Nwosu, C. S., Dev, S., Bhardwaj, P., Veeravalli, B., & John, D. (2019). Predicting stroke from electronic health records. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 5704–5707). IEEE.

[14] Yu, J., Park, S., Kwon, S. H., Ho, C. M. B., Pyo, C. S., & Lee, H. (2020). AI-based stroke disease prediction system using real-time electromyography signals. *Applied Sciences, 10*(19).

[15] Wu, Y., & Fang, Y. (2020). Stroke prediction with machine learning methods among older Chinese. *International Journal of Environmental Research and Public Health, 17*(6), 1–11.

[16] Amendolia, S. R., Cossu, G., Ganadu, M. L., Golosio, B., Masala, G. L., & Mura, G. M. (2003). A comparative study of K-nearest neighbour, support vector machine and multilayer perceptron for thalassemia screening. *Chemometrics and Intelligent Laboratory Systems, 69*(1-2), 13–20.

[17] Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *Journal of the American Medical Informatics Association, 24*(1), 198–208.

[18] Alotaibi, F. S. (2019). Implementation of machine learning model to predict heart failure disease. *International Journal of Advanced Computer Science and Applications (IJACSA)*.

[19] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.-F., & Hua, L. (2012). Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems, 36*(4), 2431–2448.

[20] García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems, 98*, 1–29.

[21] Li, X., Bian, D., Yu, J., Li, M., & Zhao, D. (2019). Using machine learning models to improve stroke risk level classification methods of China national stroke screening. *BMC Medical Informatics and Decision Making, 19*, 1–7. https://doi.org/10.1186/s12911-019-0781-4

[22] Nwosu, C. S., Dev, S., Bhardwaj, P., Veeravalli, B., & John, D. (2019). Predicting stroke from electronic health records. In *Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 5704–5707). IEEE.