IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11343

Advanced Random Forest Ensemble for Stroke Prediction

Dipita Paul[1], Gobinda Gain[2], Sujit Orang[3], Priteeranjan Das[4], Avijit Kumar Chaudhuri[5]

[1-5]Computer Science and Engineering, Techno Engineering College Banipur,

Banipur College Road, Banipur, Habra, West Bengal 743233

Abstract: India's stroke rate is rising much faster than that of other developing countries. A small percentage of patients die as a result of the initial trauma of a stroke. The adjusted average frequency of stroke is 84-262 per 100,000 in rural areas and 334-424 per 100,000 in urban areas. According to the most recent population research[1,], the incidence rate is 119-145 per 100,000. Initial ischemic infarction, recurrent ischemic stroke, pneumonia, recurrent hemorrhagic stroke, pulmonary embolism, coronary artery disease, and other vascular or nonvascular causes are among the leading causes of death. Machine Learning Techniques focus on predicting the risk of having a stroke or the possible survival of patients who survived the initial stroke. Therefore, the goal of this research work is to apply the principles of machine learning on the data set collected from population of 5110 people are involved in this study with 2995 females and 2115 males. The dataset for this study is extracted from Kaggle data repositories (https://www.kaggle.com/datasets) to predict whether a patient is likely to get stroke based on the dataset attribute information. To test the reliability of the proposed model in dealing with stroke data, a variety of training and testing partitions were used - i.e., 50-50 percent, 66-34 percent, 80-20 percent, and 10-fold cross-validations. The results were then
compared with previous studies on the same dataset, where the proposed classifier was found to be the best in all performance measures.

Keywords: Advanced Random Forest Ensemble ,Stroke prediction, ischemic stroke, Machine Learning , Kappa Statistic, ROC-AUC.

1.INTRODUCTION

In a country like India we are very busy to be a part of crowd people don't know what will happen next, like a stroke. Adjusted average frequency of stroke range, 84-262 / 100,000 in rural areas and 334-424 / 100,000 in urban areas. The incidence rate is 119-145 / 100,000 based on the latest population research[1].Disruption of blood supply to the brain causes brain damage and that's how Stroke happens. Although stroke seems to happen suddenly, but like many other diseases, it too takes time to develop with continuous high blood pressure in the vain. Generally, people are unknown about the build ups of the stroke or don't realize the symptoms that may have appeared from the start. They don't even feel the urgency or they're tend to be hesitant to check up the symptoms in the hospitals & do further examinations. This is one of the main reasons behind the increasing number of cases of stroke.

Stroke prediction is a challenge in the current healthcare Industry, This is not just for proving the existence of illness but also for ruling out disease in healthy subjects. The mainstream approach to Stroke prediction assessment uses sensitivity and specificity as indicators of test accuracy in add to the good standing status of stroke prediction.[2,3,] When the prediction findings are reported on an ordinal scale or on the ongoing scale, the prediction may be measured over all possible threshold values. As a result, Predictions vary depending on the threshold.

This study has been commonly used in clinical epidemiology for evaluating the Stroke prediction ability. The sensitivity structure is compared to specification and is known as the receiver operating character (ROC) curve, and the sub-curve field(AUC) is considered to be an important indicator of the accuracy of the definitions. This curve is important in assessing the ability to evaluate the actual status of a subject, determine the best values, and compare two diagnostic tasks as each activity is performed on the same topic. According to a Pubmed study, this study was commonly used in clinical epidemiology to assess the diagnostic ability of biomarkers (e.g., serum markers) and imaging studies in distinguishing patients from healthy subjects .[4,5,6,7,8] This mathematical method is often used in clinical trials to quantify the risk of adverse effects based on a patient's risk profile. This paper discusses the benefits of the ROC curve, the accuracy tests using the ROC curve and their predictive behavior, as well as the bias and confusing problems in the ROC analysis.

There are multiple factor reported by the Framingham Study[9,10], which increases stroke risks which includes gender, age, hypertension, Heart diseases, ever married, work type, BMI, avg glucose level, smoking status, stroke. Many more numbers of studies in previous years. Discoveries have been made that factories risk like creatinine level , the time required to walk 15 feet ,and more the model of prophecy preceding have adopted the risk factors features that are

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

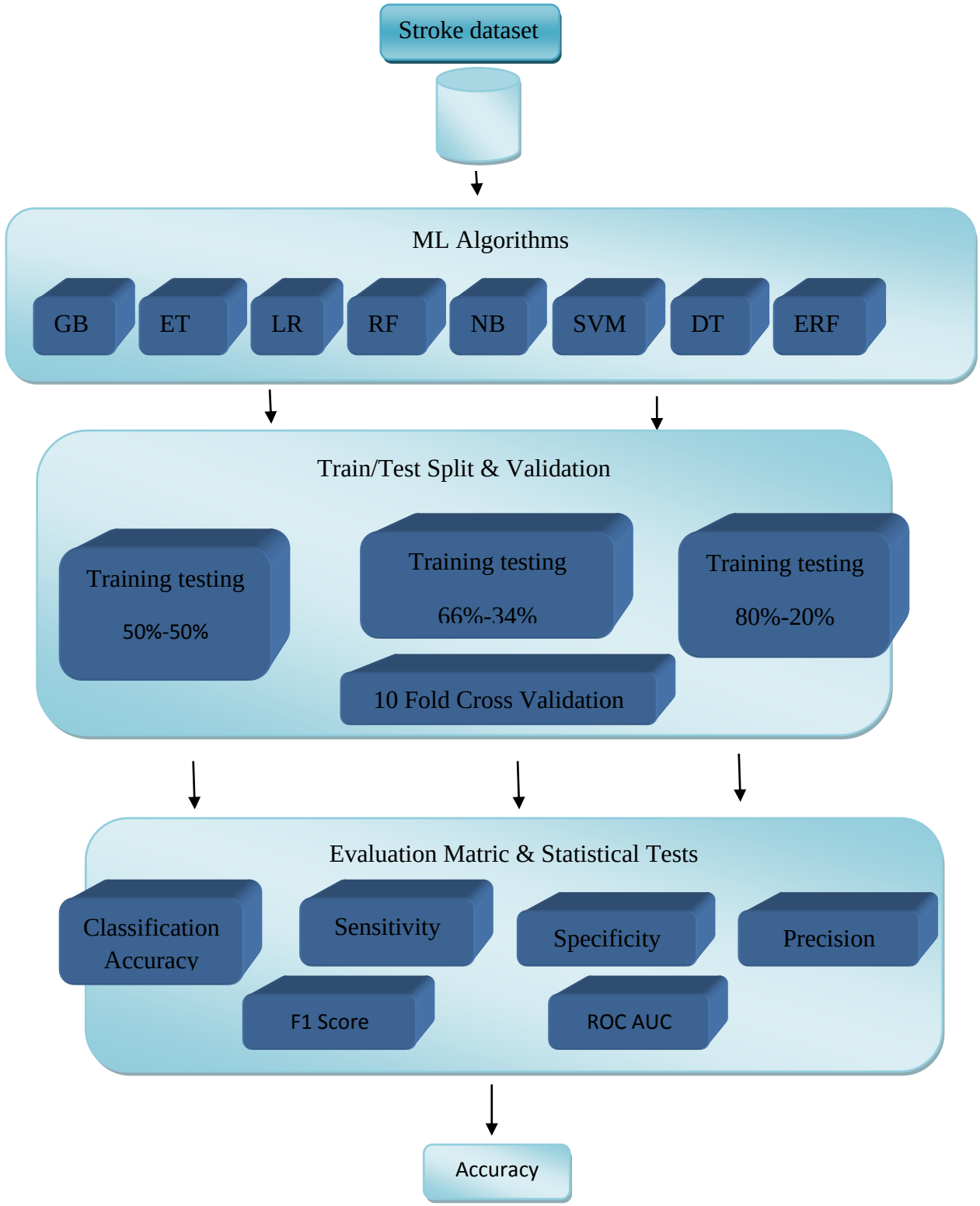
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11343

conformed by selecting manually by specialists or clinical trials. On the other hand machine learning methods can features to identify stroke highly related with the occurrence dexterously from the huge numbers of features ;So ,we can use machine learning to (i) predict new risk factors, and (ii) improve precision prophecy of stroke risk. Our paper combines with machine learning approach shows stroke risks. By machine learning approach, we explored to improve prophecy correctness and conducted considerable in our result.

Fig.1. The Architecture for the Proposed System



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11343

To identify the stroke disease, an expert system is needed, to asset the community , that is able to identify the possibility of stroke based on the symptoms felt. The system is built based on the Random Forest method. This is one of the Expert System methods with a high accuracy value. It is expected that the public will be provided the knowledge about the symptoms that might lead to a possible stroke, the user might suffer by using the system. In our case the system can perform calculations with 94% accuracy. It helps to identify whether the person has the tendency of having a stroke. Inclusion, In our work we find the potential risk indicators which has been explored by traditional applications. And last, we note that this methods can have missing data which can lead to error in risk factors and not show well recognized result. this architecture diagram which define the work flow of the scenario for proposed model is down below.

2.RELEVANT LITERATURE

Stroke prediction is a challenge in the current healthcare industry, this is not just for proving the existence of illness but also for ruling out disease in healthy subject. The mainstream approach to Stroke prediction assessment uses sensitivity and specificity as indicators of test accuracy in add to the good standing status of stroke prediction. When the prediction findings are reported on an ordinal scale or on the ongoing scale, the prediction may be measured over all possible threshold values. As a result, predictions vary depending on the threshold. This study has been commonly used in clinical epidemiology for evaluating the Stroke prediction ability. The sensitivity structure is compared to specification and is known as the receiver operating character (ROC) curve, and the sub-curve field(AUC) is considered to be an important indicator of the accuracy of the definitions. This curve is important in assessing the ability to evaluate the actual status of a subject, determine the best values, and compare two diagnostic tasks as each activity is performed on the same topic. According to a Pubmed study, this study was commonly used in clinical epidemiology to assess the diagnostic ability of biomarkers (e.g., serum markers) and imaging studies in distinguishing patients from healthy subjects . This mathematical method is often used in clinical trials to quantify the risk of adverse effects based on a patient's risk profile. This paper discusses the benefits of the ROC curve, the accuracy tests using the ROC curve and their predictive behavior, as well as the bias and confusing problems in the ROC analysis.

Chaudhuri et al.[16] estimate the illnesses using the recursive feature elimination (RFE) approach, which select a surest selection of characteristics, and an ensemble algorithm, the enhanced decision tree (EDT). The results received in the look at display that the accuracy level of EDT is not tormented by the removal of less relevant characteristics, permitting choice-makers to cognizance on a few capabilities to lower remedy time and blunders. EDT achieves a very good stage of consistency in forecasting the contamination, with or without feature choice[11].

Chaudhuri et al[17]. in comparison proven tactics and proposed a framework for integrating findings from diverse DMT to keep away from type 2 and kind 1 errors. To predict the ailment, sets of statistics had been used: ailment and remedy datasets, in addition to functions recognized as sizable by means of the ensemble method – the random forest. The results show that traditional methods, which include LR, outperformed RF in terms of sizable features. This approach, however, fails while the data dichotomy (i.e., ailment or no sickness) isn't wonderful. The DT analysis became accomplished continually throughout all editions of the dataset used on this paper[12].

We have engendered a table provided with the precedent result as well as the one we get along in our research. In this result we virtually give most of the previous year results which are from different paper and all values are not provided in different paper but the our result we have provided all the values in a single result.

Table 1 Comparison of Relevant studies

SL. No	Auther name	Accuracy	Sensitivity	Specificity	kappa	ROC_AUC	F1-score	Recall	Precision
1.	Olga Lyashevska , Fiona Malone , Eugene MacCarthy , Jens Fiehler , Jan-Hendrik Buhk , Liam Morris	0.90	0.44	0.88	X	0.733	X	0.71	0.78
2.	Aditya Khosla, Yu Cao, Cliff Chiung-Yu Lin, Hsu Kuang Chiu, Junling Hu, Honglak Lee	X	X	X	X	0.774	X	X	X
3.	BENJAMIN LETHAM, CYNTHIA RUDIN, TYLER H.	Svm(0.99) Rf(0.99)	X	X	X	Svm(0.767) Rf(0.75	X	X	X

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11343

	MCCORMICK, DAVID MADIGAN	Lr(0.98)				3) Lr(0.774)			
4.	Cathy M. Stinear, Marie-Claire Smith, Winston D. Byblow	X	(0.63-0.83)	X	X	0.75	X	X	X
5	Cemil Colak, Esra Karaman, M.Gokhan Turtay	Acc(81.82) Svm(80.38)	X	X	X	0.905 0.899	X	X	X
6.	Tianyu Liu, Wenhui Fan, Cheng Wu	71.6	67.4	32.6	X	X	X	X	X
7.	Sahar Adil , Tanvir Anwar and Adel, Al Jumaily	Svm(0.93)	X	X	X	X	X	X	X
8.	Nazar Zaki, Elfadil A Mohamed, Tetiana habuza	Svm(0.97) Rf(0.97) Lr(0.98) Knn(0.93) DT(0.97)	X	X	X	X	Svm(0.97) Rf(0.97) Lr(0.98) Knn(0.93) DT(0.97)	Svm(0.97) Rf(0.97) Lr(0.98) Knn(0.93) DT(0.97)	Svm(0.97) Rf(0.97) Lr(0.98) Knn(0.93) DT(0.97)
9.	Induja S.N, Raji G.C	Knn(98%) DT(99%) NB(95%)	Knn(0.998) DT(1.0) NB(0.972)	Knn(0.007) DT(0.001) NB(0.467)	X	X	X	X	X
10.	Javaria Amin, Muhammad Sharif, Muhammad Almas Aljum, Mudassar Raza,Syed Ahmad Chan Burkhi	0.9778	0.9787	0.9770	X	X	X	X	X
11.	Our paper	ERF(94.46)	ERF(94.50)	ERF(0.96)	ERF(3.166)	ERF(76.28)	ERF(92.80)	ERF(0.04)	ERF(91.18)

3.METHODOLOGY

3.1 Dataset:

The authors collected stroke details in a few hospitals and also in contact with peopleprevious experience with each side using online and offline questionnaire methods.The authors collected data on a total of 268 participants in the study of 131 women and 137 men.A database of this study was released to predict that a patient may experience a stroke based onthe following are the attributes of information namely high blood pressure, diabetes, age, heart disease and previous history of stroke etc.

Table 2 Description of Stroke Dataset

Sl no.	Features	Description	Range of Values
1.	Id	Id number	

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11343

2.	Gender	Male or Female	0 = Female ; 1 = Male;
3.	Age	Age at exam time in years	Continuous
4.	Hypertension	a feeling of lightheadedness or dizziness	0= no ; 1= yes;
5.	heart_disease	Previous record of heart diseases	0= no ; 1= yes;
6.	ever_married	Marital status	0=yes; 1=no;
7.	work_type	Work type of the patient	1=private; 2=self-employed; 3=other;
8.	Residence_type	Residence type of the patient	1=urban; 2=Rural;
9.	avg_glucose_level	Glucose level at exam time in years	Continuous
10.	Bmi	Body mass index exam time in years	
11.	smoking_status	Smoking status of patient	1=formerly smoked; 2 =never smoked; 3 =smokes; 4 = Unknown;
12.	Stroke	Previous record of stroke	0= no ; 1= yes;

3.2. Algorithm for ERF:

Input: Sequence of X examples,

M={(A_1,B_1).....(A_n,B_n)} where, $A_i \in a$ with labels $B_j \in b$ = $\{\omega_1+\omega_2+\omega_3+\ldots\ldots\ldots+\omega_c\}$, where $\omega_i$is the total number of classes and number of iterations for learning = j.

Initialization: Distribution, $N_i^j=\frac{1}{X}$, i = 1, 2, 3, …, X

Neighbour(g):P=m, 1≤m≤X
For j = 1 to j, perform the following –
Step 1: Select the subset for training the dataM_{Set}, peaked from the distributionN_i.

Step 2: Train the base classifier with n_{Set} and obtain the hypothesish_t, where H_j: $a \longrightarrow b$.
Step 3: Compute the error of H_j.

$$H_j\text{-}\varepsilon_j=\sum_{i}^{X} N_i^j e_i^j \quad \text{where} \quad e_i^j \begin{cases} (H_j(A_i)=B_i)=1 \\ (H_j(A_i)\neq B_i)=0 \end{cases}$$

Step 4: If ε_j > 0.5, then set j = j -1 and exit from the loop.

Step 5: Set weight, $\beta_j=\frac{\varepsilon_j}{1-\varepsilon_j}$.

Update the distribution, $N^j:N_i^{(j-1)}=\frac{N_i^j}{Z_j}\times\Phi_j$

where $\Phi_j=e_i^j \begin{cases} (H_j(A_i)=B_i)=\beta_j \\ (H_j(A_i)\neq B_i)=1 \end{cases}$ and $Z_t=\sum_i^X N_i^j$ is a constant for normalization, so that N_j^{-1} becomes a proper distribution.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11343

Output: Given an unlabeled instance a, select the class that has the maximum total vote as the optimum classification.

$$h_f(x) = \arg\max_{y=Y} \sum_{t=1}^{r} \log \frac{1}{\beta_t} \times v, \text{ where } V = \begin{cases} (h_t(x) = \omega_j) = 1 \\ (h_t(x) \neq \omega_j) = 0 \end{cases}.$$

3.3. Assessment of Performance of Machine Learning Algorithms:

In this paper, mathematical metrics are used to evaluate the performance of phases of machine learning algorithms. Metrics include (1) Accuracy, (2) Kappa statistics for each model and (3) receiver operating characteristic (ROC) curve and area under the curve (AUC), precision, (4) sensitivity, (5) accuracy, (6) remember, (7) f1-score values .The methods used are (1) Gradient Boosting (GB), (2) Extra Tree (ET), (3) Logistic Regression (LR), (4) Random Forest (RF), (5) Nave Bayes, (6) Support. vector machine (SVM), (7) Decision Tree (DT) and (8) Ensembled Random Forest (ERF) etc.

3.3.1. Naive Bayes[13]: is an algorithm, a supervised learning algorithm, based on the vision of the Bayes and widely used to solve problems.It is widely used in text classification and includes high-quality training databases.Naive Bayes Classifier is a simple and effective differentiating algorithms that helps create faster machine learning models that can make faster predictions .Other popular examples are Sentimental analysis, as well as article separation, spam filtering for empty ports.

3.3.2. Support Vector Machine: is one of the most popular Supervised Reading algorithms, used for Planning and Backing problems. However, it is often used for partition problems in machine education.The goal of the SVM algorithm is to create the best line for determining the boundary that can divide n-dimensional space into classes. So that we can easily place a new data point in the appropriate category in the future. This best decision limit is called the hyper plane.[18]

3.3.3. Extra Trees: is a machine learning algorithm that incorporates speculations from many decision trees. It is closely related to the random forest algorithm used. It usually gains better performance than the random forest algorithm, although a simple algorithm for building decision trees is used as members of the collection.

3.3.4. Gradient Boosting: is one of the most popular boosting algorithm. In Gradient Boosting, each prediction corrects a previous error.In contrast, the weights of training conditions do not change, instead, each prediction is trained using the rest errors that precede it as labels.There is a method called Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees).

3.3.5. Decision tree: It's a most powerful and famous kit, for classification and prediction. A Decision tree is a flowchart-like structure that represents a "test" on an attribute. where every internal node can denote a test on an attribute. Every branch can represent an outcome of the test.

3.3.6 Random Forest: Description: Random forest is the machine learning algorithm. Random forest's main work is solve to regression and classification problems. It's used ensemble learning, which is a technique for mixing many classifiers to provide solutions to hard problems. The random forest algorithm is makes with many decision trees. The "forest" generated by the random forest algo. This algorithm set up the outcome based on the guess of the decision trees. A random forest eliminate the limitations of a decision tree algorithm. **[14]**

3.3.7 Accuracy: accuracy measures the level of well-planned events, sensitivity to the level of well-classified cases with stroke, and the specificity of the level of well-organized events without stroke.

3.3.8 Kappa Statistic: Cohen's kappa statistics enable the accuracy of classification accuracy. The kappa school provides a measure of the accuracy of the division in width.[15]

3.3.9 ROC Curve and AUC Values :AUC is a performance metric. It measures the level at which the curve is up in the northwest corner by comparing the ROC curve with the area below the curve.[6,7]

4.RESULTS AND DISCUSSION

We built the proposed model using machine learning algorithms. In this model, we did a comparative study of Logistic regression, Random forest , Nave Bayes, Support Vector Machine, and Dicision tree, Extra tree , Ensembled Random forest, Ensembled Logistic regression , Ensembled Nave Bayes, Ensembled Dicision tree. Among these popular machine learning methods , some show better accuracy, while the performance of others is lower. The machine learning strategies used in the collected database we get 94% accuracy. all the results will be specifically provided in the table below .Our advised model provides the best accuracies in diverse disease diagnostics, however it efficiently handles the lacking value problem in datasets. Noise, missing values, and inconsistency are common place functions of medical datasets discovered in diverse repositories. Researchers use numerous pre-processing steps to resolve those issues, inclusive of data cleaning, data integration, data transformation, data reduction, and so forth.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11343

For all elements, the proposed type version generates numbers at random between the minimal and maximum values. This allows to check any viable mixture of values from different factors, and this method additionally overcomes the problem of noise, lacking cost, and inconsistency.

We provide a specific table with all the method we used in this, which are Accuracy, Sensitivity, Specificity, precision ,F1 score, Kappa, ROC_AUC, Recall.

The various method shown in table no 3 provided different accuracy values such that LR and SVM showed 93%, and ERF 94% accuracy.

The provided values considering all the selected features across train-test split . In table no 4 and 5 provided the different values of sensitivity and specificity which are calculated by confusion matrix. In sensitivity the LR and SVM getting the same value 93% and in ERF we get 94%. In specificity table the value in GB, ET and NB is 99% and in the ERF we get is 96%.

In table no 6,7,8,9 we find the respected value as shown in table of precision,f1 score, recall and ROC AOC which is 91% ,92%, and 76% in ERF,

Table 3. Comparison of Accuracy

Train - Test Split	GB	ET	LR	RF	NB	SVM	DT	ERF
50-50	0.95	0.94	0.95	0.95	0.89	0.95	0.90	0.95
66-34	0.94	0.93	0.94	0.94	0.87	0.94	0.91	0.94
80-20	0.94	0.93	0.94	0.94	0.87	0.94	0.91	0.94
10-fold Cross Validation	0.44	0.55	0.93	0.66	0.85	0.93	0.42	0.94

Table 4. Comparison of Sensitivity

Train - Test Split	GB	ET	LR	RF	NB	SVM	DT	ERF
50-50	0.25	0.13	0	0.08	0.69	0	0.29	0.95
66-34	0.67	0.05	0	0.14	0.19	0	0.16	0.17
80-20	1	0.15	0	0.5	0.37	0	0.19	0.5
10-fold Cross Validation	0.91	0.92	0.93	0.94	0.85	0.93	0.91	0.94

Table 5. Comparison of specificity

Train - Test Split	GB	ET	LR	RF	NB	SVM	DT	ERF
50-50	0.95	0.95	0.95	0.95	0.97	0.95	0.96	0.95
66-34	0.95	0.94	0.94	0.94	0.96	0.94	0.95	0.94
80-20	0.94	0.94	0.94	0.94	0.96	0.94	0.95	0.94
10-fold Cross Validation	0.99	0.99	0.92	0.98	0.99	0.94	0.95	0.96

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11343

Table 6. Comparison of Precision

Train - Test Split	GB	ET	LR	RF	NB	SVM	DT	ERF
50-50	0.25	0.13	0.00	0.08	0.18	0.00	0.13	0.29
66-34	0.67	0.05	0.00	0.14	0.19	0.00	0.16	0.17
80-20	1.00	0.15	0.00	0.50	0.23	0.00	0.19	0.50
10-fold Cross Validation	0.82	0.86	0.90	0.88	0.94	0.90	0.81	0.91

Table 7. Comparison of f1-score

Train - Test Split	GB	ET	LR	RF	NB	SVM	DT	ERF
50-50	0.03	0.05	0.00	0.01	0.25	0.00	0.15	0.07
66-34	0.04	0.02	0.00	0.02	0.26	0.00	0.15	0.02
80-20	0.03	0.05	0.00	0.03	0.31	0.00	0.17	0.06
10-fold Cross Validation	0.57	0.66	0.91	0.76	0.89	0.92	0.55	0.92

Table 8. Comparison of ROC_AUC

Train - Test Split	GB	ET	LR	RF	NB	SVM	DT	ERF
50-50	0.50	0.50	0.50	0.50	0.66	0.50	0.55	0.51
66-34	0.51	0.49	0.50	0.50	0.65	050	0.54	0.50
80-20	0.50	0.51	0.50	0.50	0.68	0.50	0.55	0.51
10-fold Cross Validation	0.44	0.28	0.95	0.34	0.85	0.28	0.30	0.76

5.CONCLUSION

Comparison of performance of popular machine learning models with that of our proposed model requires estimation. Our proposed separator is compared to assess whether the proposed model is the best and whether it improves the performance and accuracy of the sections. Accuracy is determined by the number of strategies for selecting features and outcomes produced by other models in many research papers. In selecting the features to be used, our proposed category had no limitations. The best results are obtained by considering all the features found in the database in this model. The results obtained using our fully integrated database from the site about to confirm that we are more effective in accurately predicting the incidence of Stroke compared to other available machine learning algorithms available. Therefore, the main contribution of this research paper is not only the development of an integrated learning model but also the reorganization of the fixed structure of the development algorithm by changing the base rate. The test results show that our integrated class model is effective in improving performance metrics and accuracy of grades compared to others.

DOI: 10.17148/IJARCCE.2022.11343

REFERENCE

[1].Stroke Epidemiology and Stroke Care Services in India[Professor and Head, Department of Neurology, Christian Medical College, Ludhiana, Punjab, India.BResearch Co-ordinator, Department of Neurology, Christian Medical College, Ludhiana, Punjab, India.correspondingauthorCorrespondingauthor.Correspondence: JeyarajDurai Pandian. Department of Neurology, Head of Research, Betty Cowan Research and Innovation Center, Christian Medical College, Ludhiana, Punjab]

[2]. 2. Ben-Shakhar, G., Lieblich, I., & Bar-Hillel, M. (1982). An evaluation of polygraphers' judgments: A review from a decision theoretic perspective. Journal of Applied Psychology, 67(6), 701.

[3] . Hanley, J. A. (1989). Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging, 29(3), 307-335.

[4]. Darmon, M., Vincent, F., Dellamonica, J., Schortgen, F., Gonzalez, F., Das, V., ... & Schlemmer, B. (2011). Diagnostic performance of fractional excretion of urea in the evaluation of critically ill patients with acute kidney injury: a multicenter cohort study. Critical care, 15(4), 1-8. 6. Daubin, C., Quentin, C., Allouche, S., Etard, O., Gaillard, C., Seguin, A., ... & Du Cheyron, D. (2011). Serum neuron-specific enolase as predictor of outcome in comatose cardiac-arrest survivors: a prospective cohort study. BMC cardiovascular disorders, 11(1), 1-13. 7. Hajia Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian journal of internal medicine, 4(2), 627.

[5]. Daubin, C., Quentin, C., Allouche, S., Etard, O., Gaillard, C., Seguin, A., ... & Du Cheyron, D. (2011). Serum neuron-specific enolase as predictor of outcome in comatose cardiac-arrest survivors: a prospective cohort study. BMC cardiovascular disorders, 11(1), 1-13.

[6]. Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian journal of internal medicine, 4(2), 627.

[7]. 1Vandewiele, G., Dehaene, I., Kovács, G., Sterckx, L., Janssens, O., Ongenae, F., ... &Demeester, T. (2021). Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. Artificial Intelligence in Medicine, 111, 101987.

[8]. Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation, 115(5), 654-657.

[9]Dawber, T. R., Meadors, G. F., & Moore Jr, F. E. (1951). Epidemiological approaches to heart disease: the Framingham Study. American Journal of Public Health and the Nations Health, 41(3), 279-286.

[10] Wolf, P. A., D'Agostino, R. B., Belanger, A. J., & Kannel, W. B. (1991). Probability of stroke: a risk profile from the Framingham Study. Stroke, 22(3), 312-318.

[11].Chaudhuri, A. K., Sinha, D., Banerjee, D. K., & Das, A. (2021). A novel enhanced decision tree model for detecting chronic kidney disease. Network Modeling Analysis in Health Informatics and Bioinformatics, 10(1), 1-22.

[12].Chaudhuri, A. K., Sinha, D., Bhattacharya, K., & Das, A. An Integrated Strategy for Data Mining Based on Identifying Important and Contradicting Variables for Breast Cancer Recurrence Research.

[13].Naive (Bayes) at forty: The independence assumption in information retrieval

[14]. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[15].19 Ben-David, A. (2008). Comparison of classification accuracy using Cohen's Weighted Kappa. Expert Systems with Applications, 34(2), 825-832.

[16]. Chaudhuri, A. K., Ray, A., Banerjee, D. K., & Das, A. Selection of Variables in Logistic Regression Model with Genetic Algorithm for Stroke Prediction.

[17]. Chaudhuri, A. K., Ray, A., Banerjee, D. K., & Das, A. Selection of Variables in Logistic Regression Model with Genetic Algorithm for Stroke Prediction.

[18] https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm