International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.161

Machine Learning-Based Classification andPredictionfor Patients withStrokes

SwethaManjariDasari
Student, Department of Data Science and Business Systems, School of Computing,
SRM Institute of Science and Technology, Kattankulathur.

Hemavathi D
Associate Professor, Department of Data Science and Business Systems, School of Computing, SRM Institute of Science and Technology,
Kattankulathur.

***Abstract-* A stroke is a condition in which the blood vessels in the brain are ruptured, harming the brain. Symptoms may emerge when the brain's blood and ot her nutrient flow is disrupted. The leading cause of death and disability world wide, according to the World Health Organization (WHO), is stroke. The severity of a stroke can be lessened by early detection of the numerous warning symptoms. Many machine learning (ML)models have been created to forecast the probability of a brain stroke. This study uses four distinct models for accurate prediction using avariety of physiological indicators and machine learning techniques including Support Vector Machine (SVM), Decision Tree (DT) Classification, Random Forest (RF) Classification, and K-Nearest Neighbors (KNN).With an accuracy of almost 95.1%,RandomForestwasthemostaccuratealgorithmforthisanalysis.Theopen-accessStrokePredictiondatasetwasutilizedinthemethod'sdevelopment.Theirrobustnesshasbeendemonstrated by several model comparisons, andtheschememaybeinferredfromthestudyanalysis.**

***Keywords*–ML,SVM,DT,RF,KNN**

I. Introduction

A. General

According to the CDC, an estimated 12% ofall deaths are caused by strokes, a chronic disease inthe United States. The negative effects of a stroke [1]are regularly felt by more than 795,000 people in theUnitedStates.Thefourthmajorcauseofdeathin Indiaisdue tothis.

InanupgradingMedtechfield,MachineLearning is one of the best approaches for foretellingthe onset of stroke. Detailed searches and results canbeachievedbytheuseofappropriatedataandmethods. Brain stroke prediction researches are veryfew whencomparedwithheartstroke.

The steps used during this analysis help inpredictingthechancesofastrokeinthebrain.RFbroughtoffthefinestresultsamidstavarietyofmethodsthatareutilized,byacquiringthebest-resulting metric.

This representation has a shortcoming because it wasperformed on documented inputs as a substitute foractual computer Tomography (CT). Execution of themachinelearningClassificationapproachisdemonstrated in thestudy.

To move forward with this work, a Kaggle dataset [2]is chosen that has different physical characteristics asits attributes. Following analysis, the closing resultsdepend on these characteristics. The input data file isfirstputtogetherforthemodelbycleaningandpreparingforunderstanding.

Data preprocessing is the process that follows. To fillin any null values, the dataset is first examined forthem.Ifrequired, LabelEncodingisusedtotransformcharacter variables into integers. Data is cleaved intotrainingalongwithtesting sets.

Afterward, the newly acquired information is utilizedto generate a model employing different classification techniques. The resultsoftheseapproachesarecomputed and collated to ascertain which one yieldsthemost precise predictionmodel.

B. Purpose

The maingoaloftheproposalisto developaMachine Learning Classification and prediction forpatients with strokes. The input data file is taken fromthe"Healthcaredatasetstrokedata"sectionoftheKagglewebsite[3].

To comprehend the data better, qualitativedata, quantitative data, and multicollinearity analysiswillbecarriedout.Considerthemodels:SVM,Decision Tree,RandomForest,andK-NearestNeighbor. Finally, a better method will be selected topredict stroke.

The main purpose is to expose stroke in theinfancy stage, which helps in aiding the patient andalsopreventsdeathscausedbystrokes.

II. Literature Survey

AccordingtoTasfiaIsmailShoilyetalcomparison.'softhetakenmethods,theNaiveBayes has higher precise results. The input data file,[4]which was cross-indexed by many professionals,wasobtainedbyobservingvarious medicalreports.

Theproposedmodelwillaidpatientsinunderstandingtheprobabilityofhavingastroke.4distinctmodelsweretrained.Themodels'resultswerevalidated. Machine learning models are applied to thedataset.

Inordertopredictstroke, JoonNyungHeo etal.tookinto consideration three approaches: DNN, RF, andLR. From readings, theDeep Neural Network(DNN) is frequently utilized for ischemia or acutestroke patients[5]. By utilizing the given input datathe DNN model approaches an 87% accuracy whichsurpasses the other models. It is improved by usingautomatedcalculationsthataremoreaccurate,whichreducestheneedforsimpler models.

Inadditiontoprovidinginformationonpotentialdisabilities broughtonbystroke,JaehakYuetalpreferred.'sC4.5DTmodel[6]leveragestheNIHSSscore,itclassifiesstrokeintensityinto 4 categories.

Thecapacitytopredictthepotentialtimingofastrokeanditsa ssociatedhandicapenablestheuseofadditionaldrugsandtheappropriatesafetymeasures. [7] Random Forest and Naive Bias both have highaccuracyratingsof88.9%and85.4%, respectively.

SVM was employed by Jeena R.S. and Dr. SukeshKumar with an approach that takes data as input andconvertsittoarequiredformforresearchpurposes.[8]350 inputs for the prediction were taken after pre-processing to remove redundant and conflicting data.91%accuracywasachievedthankstoMATLABsoftware.

ChutimaJalayondejahasstatedthatwhenusingdemographicdatatomakepredictions,DecisionTrees, Naive Bayes, and Neural Networks were thethree models that were taken into consideration. Thedecision Tree was found to have the highest accuracyandthelowFPrate.SinceFNpredictsthecontrarybutcausesmortalitybecausethepatientexperiencesastroke, FN is harmful. The decision Tree was takenintoconsiderationforaccuracy,[9]whileNeuralNetwork was chosen for safety because it had a highFPvalue anda lowFNvalue.

ABayesianRuleList(BRL)waspredictedbyBenjamin Letham etal.,and itbuildsadistribution ofpermutations from data. [10] The algorithm scales theinput data sets with complex features. High levels ofaccuracy, precision, and tractability can be attainedwiththe BRLapproach.

Pei-WenHuang1etal.usedphysiologicaldatatopredict stroke using the multimodal analysis method.Thisinformationincludesphotoplethysmography,arterialbloodpressure,andelectrocardiography(EKG) (PPG). [11] Each of these signals has beenexamined for accuracy. Additionally, they combinedthe signals and claimed it has the highest accurateresults.

Artificial neural networks may be used to forecastthromboembolicstrokedisease,accordingtoresearch.TheBackpropagationalgorithmwastakenastheapproaching method. The accuracy achieved by thismodelwas88%.[12]However,duetothecomplexityofinternalstructuresandthelargenumberofneurons,it takes an extended period of time to analyze theinformation.

III. Proposed Methodology

The suitable input data set for the model development has been taken from all the different data sets available in Kaggle after a lot of consideration, this dataset is further moved into the implementation part.

The steps involved in making the input data ready for machine understanding begin once the input data is taken and this process is called data preparation. This deals[13] specifically with dataset's label encoding where categorical data is encoded into numerical data, treatment of missing values by replacing them with the mean of the available data of that respective attribute in the dataset, and management of data that is imbalanced. The preprocessed data is now ready for model construction.

Exploratory Data Analysis is performed on the preprocessed data for getting relevant inferences and observations. Various visualizations like Graphical pie charts are used for obtaining the inferences.

Feature selection is also performed to ensure the essential features are only used for the developed model, which helps in maintaining the performance of the model and it also helps in solving the overfitting problem.

The following fig.1 ,the model building is shown by using various methods which help in the best prediction

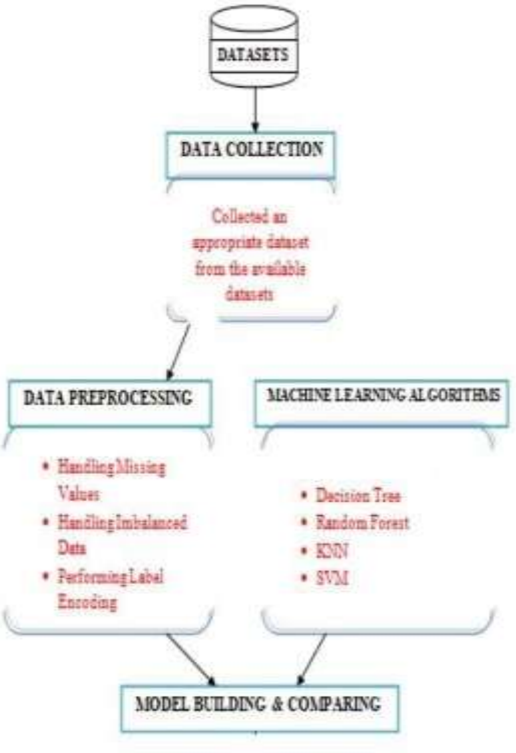

DATASETS
DATA COLLECTION
Collected an appropriate dataset from the available datasets
DATA PREPROCESSING
MACHINE LEARNING ALGORITHMS
• Handling Missing Values
• Handling Imbalanced Data
• Performing Label Encoding
• Decision Tree
• Random Forest
• KNN
• SVM
MODEL BUILDING & COMPARING

Fig. 1. ML model building Flow Diagram

For the model creation, the preprocessed datasets and the ML methods are taken into consideration. Among the algorithms utilized are DT Classification, RF Classification, KNN, and SVM Classification, 5 accurate metrics are used to compare the six distinct models that were built.

IV. Implementation

A. Dataset

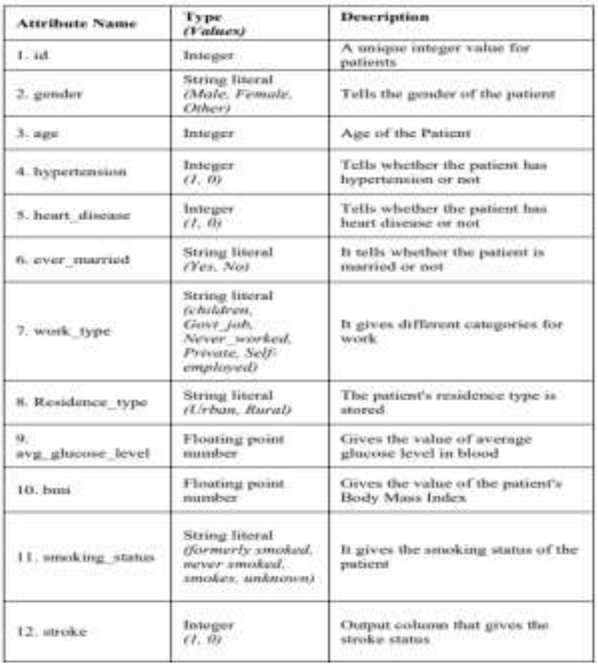
The Kaggle dataset was used to predict strokes. The input data files consist of twelve columns and five thousand hundred and ten rows. The columns that are taken into consideration are: "id," "gender," "age," "hypertension," "heart disease," "ever married," "work type," "Residence type”. "avg glucose level," "BMI" "smoking status”, "stroke."

The column “stroke” has the output value as a binary value which is either “1” or “0”.If a patient has a risk of stroke, the value is denoted by 1 and if the patient does not have any risk of stroke has the value of 0.

Mostly the column stroke has a value of 0 compared to the value of 1, due to which the data input file is mostly unbalanced.The next step to balance the unbalanced data preprocessing is done for the best results.

The following table 1 contains the summary of the dataset mentioned earlier.

TABLE 1. DATASET DESCRIPTION



Attribute Name	Type *(Values)*	Description
1. id	Integer	A unique integer value for patients
2. gender	String literal *(Male, Female, Other)*	Tells the gender of the patient
3. age	Integer	Age of the Patient
4. hypertension	Integer *(1, 0)*	Tells whether the patient has hypertension or not
5. heart_disease	Integer *(1, 0)*	Tells whether the patient has heart disease or not
6. ever_married	String literal *(Yes, No)*	It tells whether the patient is married or not
7. work_type	String literal *(children, Govt_job, Never_worked, Private, Self-employed)*	It gives different categories for work
8. Residence_type	String literal *(Urban, Rural)*	The patient's residence type is stored
9. avg_glucose_level	Floating point number	Gives the value of average glucose level in blood
10. bmi	Floating point number	Gives the value of the patient's Body Mass Index
11. smoking_status	String literal *(formerly smoked, never smoked, smokes, unknown)*	It gives the smoking status of the patient
12. stroke	Integer *(1, 0)*	Output column that gives the stroke status

B. *Preprocessing*

Preprocessing is one of the important stepsbeforemodelbuilding.Theundesirablenoiseandoutliers are removed from the input data by using thepreprocessing method, if not it will cause a deviationfromnormaltraining.Thisstepinvolvesmostlyfixingtheerrorsthatpreventtheoperationofthemodeleffectively.

Aftertakingthedesireddataintoconsideration the second stage is performed which isto clean the data and make sure it is in developing themodel. The dataset used comprises twelve properties.First off,” id” is discarded because it does not add anyvalue. Following the dataset is checked if it has anyzerovaluesandfilledifanyarediscovered.Thecolumn” BMI” has a zero value which is replaced bythemeanvalue.

As the zero values from the input dataset areremoved, the following Label Encoding process takesplace.

C. *LabelEncoding*

Label Encoding is a process that is used tomakethecomputerunderstandthestringvaluespresent in the input dataset, thus it converts the datainto integer values. Strings need to be translated tointegerssincemachinesareofteneducatedonnumerical values. The input dataset contains stringtypein5columns.WhenLabelEncodingisapplied,the total string values in the entire input dataset areencoded, turninginto numericalvalues.

D. HandlingImbalancedData

Data scaling helps in improving the model'saccuracy as imbalanced data creates bias when themodelistrainedwhichinturnresultsinpooraccuracy.Min-maxdatascalingtechniqueisusedforscalingthestroke dataset.

V. MODEL DEVELOPMENT

A. Dividingthedata

Aftersucceedingindealingwiththeunbalanced dataset and completing data preparation,thenextstageiscreatingthemodel.Thebalanceddataiscleavedintotrainandtestgroups,thetraininggroupconsistsof 80%whilethetestgroup consistsofa20%ratio,whichisusedtoincreaseofprecisionandproductivityofthe activity.

After dividing the balanced data manyclassificationmethodswillbeperformed.Theclassification techniques used for this purpose includeSVMClassification,DTClassification,RFClassification, andKNNClassification.

A. *Algorithms*

1) *DTClassification(DecisionTree)*

TheclassificationandregressioncomplicationsmaybesolvedusingasupervisedlearningsupervisedtechniquewhichisDTclassification, however, it canbe typically usedinsolvingproblemswith Classification.

ThisclassifierhasastructurelikeTree,withinputdatarepresentingtheirowncharacteristics,rules,andclassification will be represented by branches and theresultsofclassificationwillberepresentedbyterminalnodes.

Theterminalandnon-terminalnodesformaDecisiontree.ContrarytoaLeafnode,whichrepresents the outcome of the decision and has no extrabranches, a Decision node allows for the making of achoice andcontainsnumerousbranches.

Torunthetestsorformopinions,theprovided dataset's characteristics are used. It is a visualrepresentationofallpossibilitiesforresolvingaconundrumorselectingacourseofactioninconsiderationofspecific criteria.

In developing a Tree, the Classification andRegressionTreemethodsareusedwhichareoftenknown as CART. It creates a question followed by asub-treewithbinaryanswersi.e.,yesorno.

2) *RandomForestClassification*

The well-known random forest classifier isused by combining several classifiers to handle severalissues and to increase the productivity of the model.This method mainly depends on ensemble learning.The regression and classification complications aresolved bythisclassifier.

BytakingthegiveninformationintoconsiderationtheRandom Forest classifier uses many decision trees ondifferent subgroups to increase the predicted outcomeaccuracy of the data. This algorithm uses each decisiontree in foretelling the results based on the majority ofvotesthendependingononedecisiontree.TheOverfittingproblemcanbesolved usingmoretrees.

Some decision trees will anticipate the correctoutcome when compared to others and the reason isRandomForestclassifierusesadifferentdistinctdecisiontreeto forecastthetypeofinputdataset.However, all the trees provide reliable forecasts whentakenasawhole.

Thefollowingtwotheoriesareputoutinaneffort to improve the RF classifier. It should containreal values for RF to foresee the correct outcome asopposed to a speculative outcome. There must be averyminimalconnectionbetweentheforecastsofeachtree.

3) K-NearestNeighbor

Baseduponsupervisedlearningthis isoneofthesimplestMLtechniques.Thisassumessimilaritybetweenthealreadyusedcasestonewcasesandisfollowedbyastepwherethealgorithmtakesthenewcaseto aplace in analreadyusedcategory.

Thisalgorithmmaintainsthealreadyused data and distinguishes them into different newdata points based on the resemblance. So by utilizingthis method we can obtain new accurate data which ismorecharacterizedandsuitableforthe requirement.

As this algorithm is a distribution-freetechnique,itmakesnoeffortinguessingthedatawhichisunderlyinginthedataset.Thisismostlyusedforsolvingclassificationcomplications.Thisalgorithm is mostly an inactive learner because itstoresthetrain datainplaceoflearning it.

Instead, this algorithm uses the dataset to carry out anactionwhendistinguishingdata.Thismethodstorestheinformationfromthetrainingphasewhenitacquiresthelatestdata, andcategorizesit intoagroupthatistooclosetothe latestdata.

4) SupportVectorClassification

SVM classifier is one of the best methodsin ML which is used in solving complications in bothclassification and regression. This method is beingusedinmanymodelbuildingsforbetterperformance.

The main aim of this method is to get thebestdecisionthatcandifferentiatethen-dimensionalspace into classes. Next, the sub-data points are fastly moved to suitable categories.

“Hyperplane” is defined as the optimal decision boundary. This method selects the extremity points and vectors to generate a “hyperplane”. This

C. OPTIMIZATION

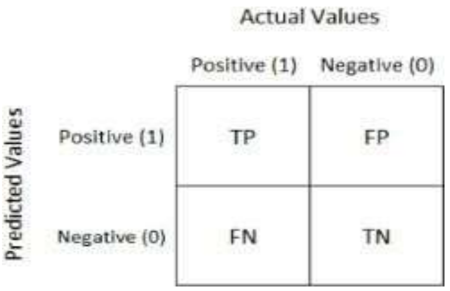
The main objective of machine learning is tobuildmodelsthatperformwellandprovidereliablepredictions for a given set of cases [14]. Machine learning optimization is required to accomplish it. By applying one of the optimization strategies, it alters thehyperparameters for reducing cost function. Since the cost function captures the variation between theapproximatetruevaluesandpredictiveresults.

D. Model Evaluation

Classification Metrics –There are four possible outcomes when making classification predictions.

- False positives (FP) are the cases where the model inaccurately anticipates it to be positive when it was really negative.
- True positives (TP) are cases where the model accurately anticipates it to be positive.
- True negatives (TN) Situations where the model correctly predicts that the negative class is negative.
- False negatives (FN) are situations in which the model expects a negative outcome but shows a positive result.

Confusion Matrix - Accuracy, precision, recall, and F-Measure are the four measures employed to gauge a classification model's performance.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig. 2. Confusion Matrix

exact ones. The obtained percentage is used for testing and is referred to as accuracy.

$$\text{Accuracy} = \frac{TP+TN}{(TP+FP+FN+TN)} \quad (1)$$

2) A proportion of positive cases out of all projected positive cases is called precision.

$$\text{Precision} = \frac{TP}{(TP+F\text{N}} \quad (2)$$

3) A recall is a proportion of instances of positivity out of all real instances of positivity.

$$\text{Precision} = \frac{TP}{(TP+F\text{N})} \quad (3)$$

4) When calculating the score, the F-Metric accuracy measure takes into account the two of precision and recall.

1)The accuracy is defined as the ratio of thetotalnumberofforetellsto thenumberof

It is simple for establishing the method to becharacterizedasapositiveornegativemethodbyusingthesefoursignsasbenchmarkstoconstructtheassessmentcriteria.

VI. Results

A. Comparison Resultsofthefourmethods

Four learning strategies( ) were investigatedin this article to predict stroke. Following a thoroughanalysis, we came to the following conclusions. Thebest-performing model out of the four is taken intoconsideration forprediction.

ROC(ReceiverOperatingCharacteristics)curvesforallthe fourmodelsarecomparedandanalyzedforselectingthemodel withbetterperformance.Modelperformsbetter iftheROCcurveis towards the top left. Following figures show theROCcurves offourmodels.

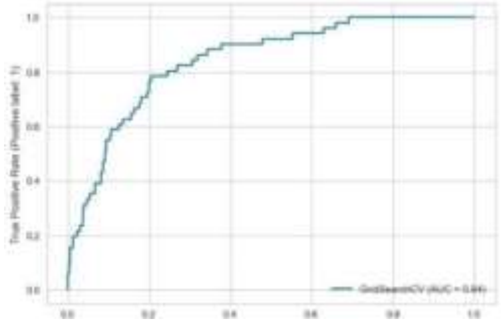


Fig. 3. ROCforRandomForest

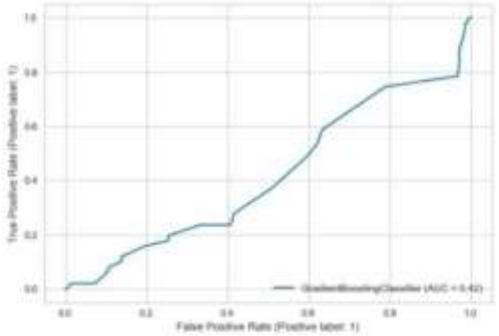


Fig. 4.ROCforSupportVectorMachine

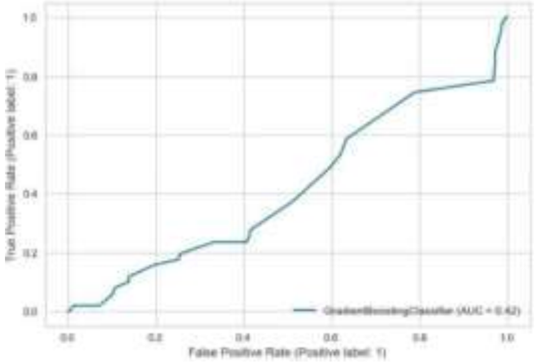


Fig. 5.ROCforK'sNearestNeighbour

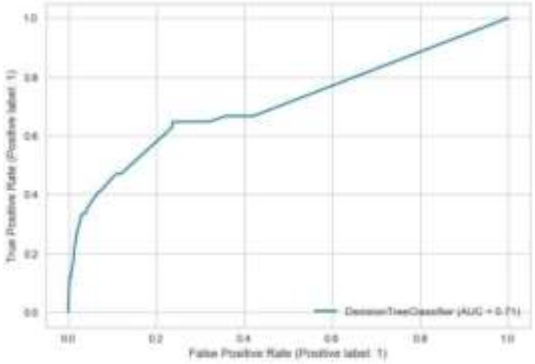
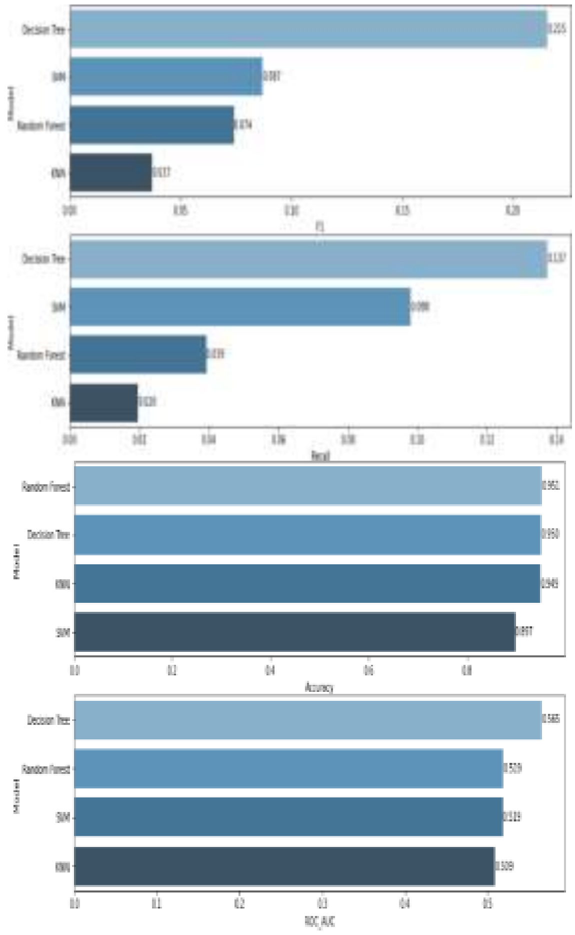


Fig. 6.ROCfor DecisionTree

Thebelowbargraphshowsthecomparisonofthe accuracyscoreofallfourmodels.



VII. CONCLUSION

Thisresearchisperformedforexposingstroke in the infancy stage, which helps in aiding thepatienttohavealessdetrimentalmedication,reducingthemedicationexpense,knowingtheaccurateprobability of results and it also helps to increase theMedtech level in the healthcare sector. This researchalso helps in saving many lives and to remove strokeriskfrombecomingoneofthedeadliestdeathworldwide.

95.1%isthemaximumpreciseoutcomeacquired by the RF Classifier compared to the othermethodsbyusingthe12variablesand5109data.

RF has the lead over other methods in distinguishingdatabecauseitinvolvesdatawithincompleteattributes. This algorithm is also better at graspinglarge data.

VIII. FUTURE WORK

Forfutureworkinresearch,theimplementationofsmartarrangementsisrecommended to be made in the prognosis of stroke,in addition to the alternative algorithms in ML whichcanbeused forgivingaccurate andbestresults.

A few suggestions can be taken into consideration byaddingtheattributestotheinputdatafile.Forexample, exhausting activities and professions to getbetter results.

EnsembleLearningwhichhelpstohavebetterpredictionperformance.

REFERENCES

[1] AdityaKhosla,YuCao,CliffChiung-YuLin,Hsu-KuangChiu,JunlingHu,HonglakLee, “AnIntegratedMachineLearningApproach to Stroke Prediction,” In: Proceedings of the 16th ACMSIGKDD international conference on Knowledge discovery anddata mining,2010.

[2] D. Shanthi, G.Sahoo, and N.Saravanan, “Designing an artificialneuralnetworkmodelforthepredictionofthrombo-embolicstroke,”Int. J.BiometricBioinform.(IJBB),2009.

[3] Datasetnamed‘StrokePredictionDataset’fromKaggle:https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

[4] M.S.Singh,P. Choudhary,and K.Thongam,“Acomparativeanalysisforvariousstrokepredictiontechniques,”In:Springer,Singapore,2020.

[5] S. Pradeepa,K.R.Manjula,S.Vimal,M.S.Khan,N.Chilamkurti, and &A. K. Luhach, “DRFS: Detecting Risk Factor ofStrokeDiseasefromSocialMediaUsingMachineLearningTechniques, ” InSpringer2020.

[6] VamsiBandi,DebnathBhattacharyya,DivyaMidhunchakkravarthy: “Prediction of Brain Stroke Severity UsingMachine Learning,” In: International Information and EngineeringTechnologyAssociation,2020.

[7] C. S. Nwosu, S.Dev, P. Bhardwaj, B.Veeravalli, and D. John, “Predicting stroke from electronic health records,” In: 41st AnnualInternational Conference of the IEEE Engineering in Medicine andBiologySociety IEEE,2019.

[8] FahdSalehAlotaibi,“ImplementationofMachineLearningModel to Predict Heart Failure Disease,” In: International Journal ofAdvancedComputerScienceandApplications(IJACSA),2019.

[9] Moshika, A., Thirumaran, M., Natarajan, B., Andal, K., Sambasivam, G., &Manoharan, R. (2021).Vulnerability assessment in heterogeneous web environment using probabilistic arithmetic automata. IEEE Access, 9, 74659-74673..

[10] T. Kansadub,S.Thammaboosadee,S.Kiattisin,C.Jalayondeja, “Strokeriskpredictionmodelbasedondemographicdata,”In:8thBiomedicalEngineeringInternationalConference(BMEiCON)IEEE,2015.

[11] TasfiaIsmailShoily,TajulIslam,,SumaiyaJannatandSharminAkterTanna, "Detection of stroke using machine learningalgorithms",10thInternationalConferenceonComputing,Communication and Networking Technologies (ICCCNT), IEEE,July2019.

[12] R.S. Jeena andSukesh Kumar “Stroke predictionusingSVM”, International Conference on Computing, Communicationand NetworkingTechnologies(ICCCNT),IEEE,2016.

[13] Rajesh, M., &Sitharthan, R. (2022). Image fusion and enhancement based on energy of the pixel using Deep Convolutional Neural Network. Multimedia Tools and Applications, 81(1), 873-885.

[14] JoonNyungHeo , Jihoon G. Yoon , Hyungjong Park , YoungDaeKim,HyoSukNamandJiHoeHeo."Strokepredictioninacutestroke",Stroke,AHA Journal, vol. 50, pp. 1263-1265,20Mar2019.