Predictive Model for Brain Stroke in CT using Deep Neural Network

Maya B S, Asha T

***Abstract:** The increasing in the incidence of stroke with aging world population would quickly place an economic burden on society. In proposed method we use different machine learning classification algorithms like Decision Tree, Deep Neural Network Learning, Maximum Expectization , Random Forest and Gaussian Naïve Bayesian Classifier is used with associated number of attributes to estimate the occurrence of stroke disease. The present research, mainly PCA (Principal Component Analysis) algorithm is used to limit the performance and scaling used to be adopted to extract splendid context statistics from medical records. We used those reduced features to determine whether or not the patient has a stroke disorder. We compared proposed method Deep neural network learning classifier with other machine-learning methods with respect to accuracy, sensitivity and specificity that yields 86.42%, 74.89 and 88.49% respectively. Hence it can be with the aid of both patients and medical doctors to treat viable stroke.*

Keywords : Stroke, Deep Neural Network Learning, Gaussian Naïve Bayesian Classifier, Principal Component Analysis, Machine Learning Algorithm.

I. INTRODUCTION

The stroke causes health disorders in long term and it is the major cause for death. If the affected person has heart related hassle like mitral regurgitation, heart failure, pulmonary stenosis, then stroke will easily make the condition worse [1]. It is brought when flow of blood to an region of area is reduced due to a clot or a blood vessel splits due to oxygen and nutrient deficiency[5].Cerebrovascular incident (CVA) or stroke categorized into different forms [2]. This disease can be lead to paralysis, intense chest pain, speech impairment, memory loss and ability to think, coma or death. Stroke affects any age on the body via medical therapy it can be avoided and essential are modifiable risk factors. Early identification of stroke is still a challenge and is very much significant in medical field.

Machine learning techniques are indeed worth exploring in predicting the probability of stroke. Machine learning is a method of data analysis that automates methodical model building [4].

Revised Manuscript Received on May 21, 2020.
***** Correspondence Author

Maya B S , Assistant Professor in department of computer science & Engineering, Bangalore Institute of Technology, Bangalore.

Dr. Asha. T, Assistant Professor in department of computer science & Engineering, Bangalore Institute of Technology, Bangalore.

Due to the risk development of machine intelligence techniques based on various mathematical approaches, it is possible to implement artificial intelligence significantly in medical decision making . throughout medicine, medical data are stored , among others,
throughout centralized databases and comprehensive clinical and laboratory datasets[3]. These data can be used to forecast the incidence and prognosis of diseases.

The key objective of this work is to use classification techniques for machine learning to predict patients at risk of developing stroke. Therefore, five predictive classification algorithms like Decision Tree, Random Forest, Gaussian Naïve Bays Classifier, Maximum Expectization and Deep NN Learning is used on patient stroke data collection obtained from different hospitals. The classifier is compared with each one to find the best results in order to find the best predictive model. As a result , a system has been created to recognize patients with stroke using the correct decision support tool to help achieve these objectives :i) Reduces stroke impact on patient life ii) Increase community life expectancy wellbeing iii) Decrease human service cost.

The rest of subsections of this exploration article are composed as follows: The literature assessment on machine-leaning technique to predict stroke is mentioned in Section two. Section three discusses the approach while the outcomes and discussion are mentioned in section four. At last conclusion and potential research are addressed in section 5.

II. REVIEW CRITERIA

Ashfaq Ahmed et.al. [7] proposed a study using the techniques of machine learning classifiers. Those have been used to describe and compare key data sets of different kernels and kernel parameters for cancer, liver and heart disease. Random forest and SVM has been tested for different collection of data on malignant disease, liver disease and coronary illness. The outputs are obtained by implementing various kernels were adapted to the proper identification of parameter. Outputs were analyzed to develop statistical methods of learning. The contrasting results were observed with SVM classification methods for linear, polynomial and quadratic kernel functions.

Oskar et. al. [6] applied variety of classification methods to classify ischemic stroke and results of this study concluded that RDF s and CNN results outperforms all formerly published methods, human observer accuracy is not yet reached.

Yang et. al. [8] applied kNN, multiple linear regression (MLR), and a regression tree model to expect the stroke severity index (SSI) and verified that kNN is greater reliable than different models. Sasi et. al. [9] offered a predictive model with DT, ANN, SVM, Logistic Regression (LR), Generalized Boosted Model (GBM) to predict ICU transmission of stroke patients and concluded that GBM was the most accurate. Several studies have employed methods of deep learning to solve various problems [14–16]. There have been several computer-aided diagnosis systems, in particular that use deep learning to detect various diseases [17-23]. Dhar R et.al. [10] used Machine-learning and deep learning , different methods and datasets have been used to identify or predict those diseases.

Arslana. et.al. [12] proposed stroke prediction algorithms like SVM, Stochastic Gradient Boosting and Penalized Logistic Regression. SVM has reached an precision of 97.2%. Reza et al. [11] , use of K-nearest neighbor and C4.5 decision tree, the precision of the stroke prediction was high compared to others. ANN model achieved stroke accuracy as 88.2% as shown in Eibe Frank et al. [13]. Many of the previous studied used techniques of machine learning to identify different diseases. However, there was no attempt to predict stroke with these tools.

III. METHODOLOGY

Stroke disease prediction that focuses on increasing people's consciousness and studies are focused on recent trends. The model proposed uses predictive classification algorithms such as Decision tree, Maximum Expectization ,Gaussian Naïve Bayesian, Random Forest and Deep Neural Networks to predict the existence of stroke and Principle Component Analysis algorithm to diminish dimensionality. Therefore decreased subset of features may be used as inputs. Figure 1 reflects the Stroke Disease Prediction.

A. Dataset Collection

The first phase stroke dataset was obtained from many Bangalore hospitals and medical centers with patients diagnosed as stroke or non stroke. The data set is composed of groups. The one class contains medical records for patients confirmed to have stroke and the another class contains patients with stroke mimics, that were typically misdiagnosed as stroke due to sign and symptoms similarity. we are consider 1,500 instance of stroke patients dataset. It involves both hemorrhagic stroke and ischemic stroke .Table 1 provides statistical information about patients data . The median age of subject with stroke patients was 65 years. Among these patients 53.7% were male, 43.3% were female.

B. Pre-processing Dataset

The data collection for the condition is initially pre-processed to make it appropriate for feature extraction. It is used for remove unnecessary information, incomplete data, noise and incoherent data.

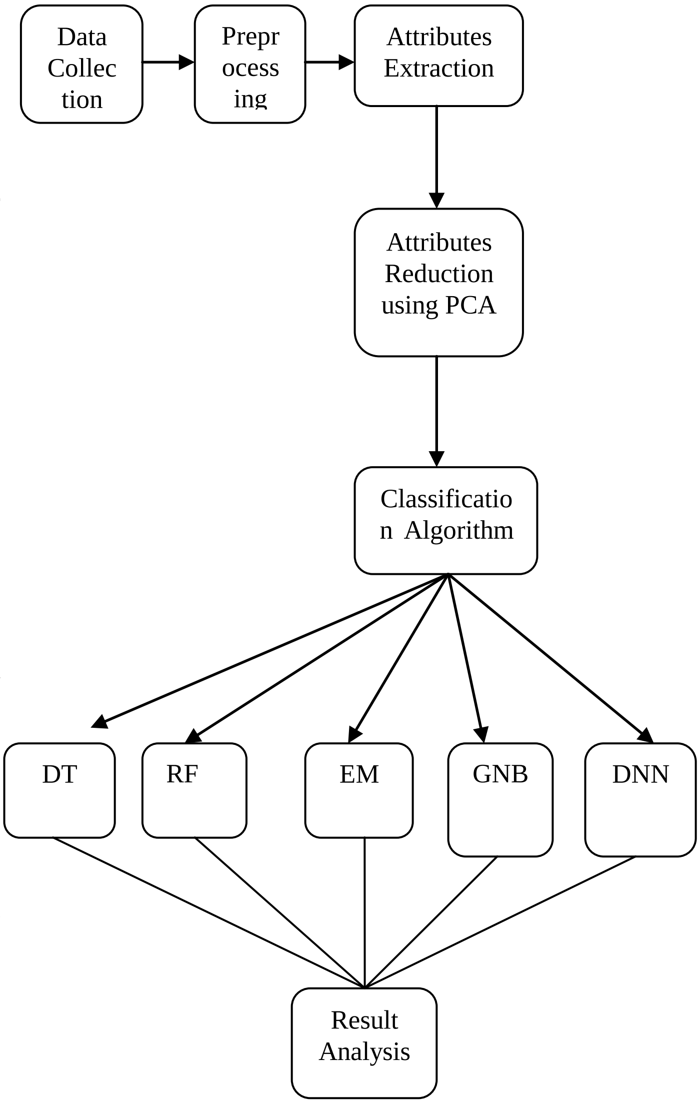

Fig. 1. Proposed diagram for Prediction of Brain Stroke Disease.

DT: Decision Tree, RF: Random Forest, EM: Expectation Maximization, GNB: Gaussian Naïve Bayesian Classifier, DN N: Deep Neural Network

Table 1. The Patient dataset selection process

Variables	Values	Number
Gender	Male	873(58.2%)
	Female	627(41.8%)
Stroke Type	Ischemic	934(62.22%)
	Hemorrhagic	566(37.73%)

C. Attributes Selection and Reduction using PCA

The data set contained different attribute like class stroke, age , gender, work type, married status, hyper tension, Glucose level, Residence, Body mass index, treatment name (done, not-done), and different lab test, Smoking status, Heart disease, etc . Selection of the attributes was then applied to the dimensionality of the results. Selection of attribute is used for choosing the most appropriate attributes[24].

PCA is a plain for extracting useful data from different datasets. PCA defines new features effectively reducing the dimensions to cover or simplify structure using different algorithms [25]. The former attributes should not be inserted into a DNN classifier, since precise knowledge is missing. To avoid data discretization and transform binary or categorical attributes into continuous variables using PCA maximum attribute filter.

PCA is constructed on the basis of an important property of eigen vector decomposition. In practice PCA computing a data set X requires (1) Extracting the mean of each form of measurement and (2) computation of eigenvectors . This approach is expressed in Matlab code and Figure 2 shows the corresponding graph. The algorithm has the potential to reduce attribute disturbance by translating it into the PCA space, extracting and transforming the bad eigen vector and converting into the original space. The below table 2 shows the final attribute sets for stroke prediction. In addition , data are split into two different data sets: 64% training data set for model building and 36% test data set for model evaluation.

D. Proposed Prediction Model for Brain Stroke with Reduced Attributes

Prediction model composed of two models such as classification and assessment. The classification model uses training data set to construct a predictive model for classification. The test data set is used to test classification output. Patient dataset is compiled from healthcare institutes who have stroke symptoms. Then classification algorithm such as Decision Tree, Deep Neural Network Learning, Maximum Expectization , Random Forest and Gaussian ,

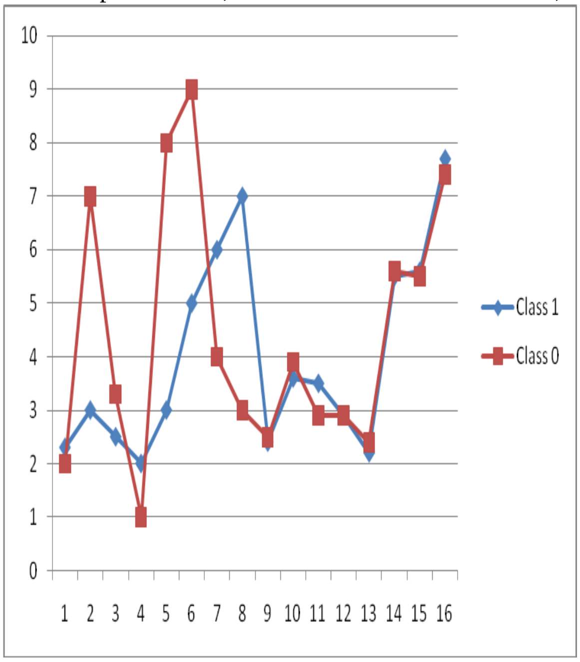

Fig. 2. Two-dimensional graph

(class 0 shows without stroke patients and class 1 shows with

Table 2: Attributes with risk factor for stoke prediction

Patient Dataset								
ID	**Gender**	**Age**	**Hyper Tension**	**Heart Disease**	**Married**	**Work_Type**	**Residence**	**Smoking Status**
36305	Man	81	0	0	Yes	Private	Town	formerly smoked
61826	Woman	75	0	1	Yes	Self	Rural	formerly smoked
14158	Woman	13	0	0	No	Children	Rural	
12997	Man	25	0	1	No	Private	Town	
40801	Woman	62	0	0	Yes	Govt_job	Rural	never smoked
9348	Woman	65	1	0	Yes	Private	Town	never smoked
51550	Woman	49	0	0	Yes	Self	Rural	
60512	Man	45	0	0	Yes	Govt_job	Town	never smoked
31307	Woman	76	0	0	Yes	Self	Rural	never smoked
39164	Man	77	0	0	Yes	Self	Town	smokes

15161	Woman	16	0	0	No	Private	Rural	

Naïve Bayesian Classifier are used to determine whether patients suffer from stroke disorder with patient dataset shown in table 2. Then performance assessment is performed with different algorithms and accuracy is compared with different models.

The objective of this research is to determine the effectiveness of selected machine learning techniques for predicting the occurrence of stroke, as well as to identify those features of the patient that have the greatest impact on the occurrence of stroke. It will also be analyzed how the reduction of attributes describing patients affects the accuracy of the classification.

- **Decision Tree**

Decision trees are data structure with root node, branches and nodes on the chain. Internal node represents a test on attribute or function of data, each branch represents the examination results and each leaf node contains a class mark. The root of a tree is located at topmost. A decision tree is a class discriminator which remedies the training set [26].

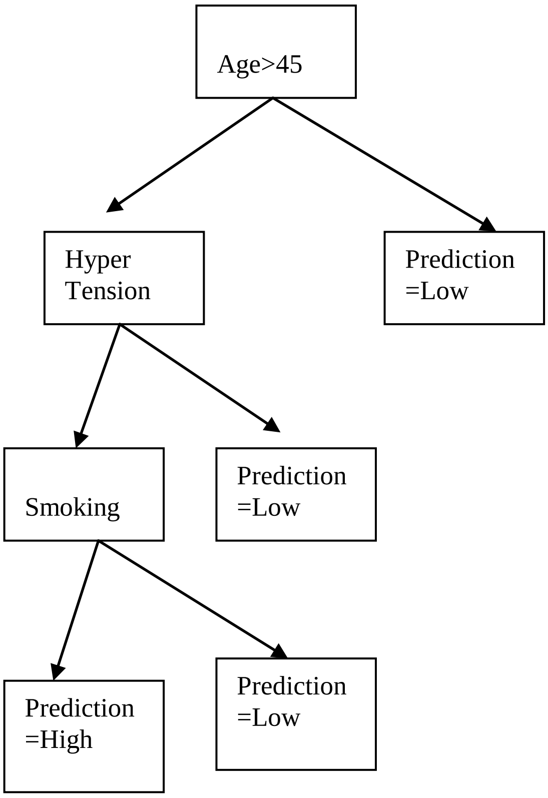

Fig. 3. Prediction of Stroke using Decision tree

Each tree branch and root node includes a split point that is used for to check one or more attributes or features which decides how the data is divided and split to predict the disease from given data shown in above Figure 3.

- **Random Forest**

Random Forest is a method of classification (and regression) involving the different decision trees were generated based on a random collection of data. The idea of this algorithm is to create a team of experts from random decision trees, where random trees are constructed on the concepts of random selection of a subset of the analyzed featured in the node unlike classic decision trees [29]. In addition, individual trees from random forest trees are built according to the concept of bugging. Random forests are considered one of the best classification methods. Single random forest classifiers are decision trees. The Random Forest algorithm is very well suited for trial testing, where the observation vector is a large dimension. Their additional advantage is the ability to use the learned random forest for other issues than just for classification. For example, based on trees from the forest, one can determine the ranking of variables, and thus determine which variables have better predictive properties.

- **Expectation Maximization**

The expectation maximization (EM) algorithm is an method in presence of latent variables to achieve maximum likelihood estimates. It is an recursive technique to estimate maximization for parameter .The maximum-likelihood model is used when random values are chosen to find the best fit for a data set. The recursive algorithm estimates the best match for an incomplete or missing data point in an unsupervised manner [27, 28]. The best match is determined for a collection of complete data with $\theta = \xi$ as shown in Figure 4 by estimating maximum likelihood. To capture the parameters of the model in a missing data using complex EM algorithm. The missing data points are arbitrarily identified by an algorithm and these are used to estimate the old data set.

- **Gaussian Naïve Bayesian Classifier**

Gaussian Naïve Bayes (GNB) is a supervised learning classification of algorithm that uses Bayes' theorem as a basis for classifying observations into one of a pre-defined set of class based on predictor variables. The Gaussian Naïve Bayesian model for predicting stroke diseases is shown in Figure 5.

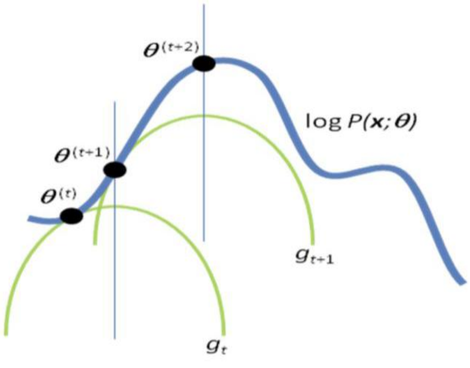

Fig. 4. The convergence of EM algorithm [23].

GNB classifiers estimate the conditional probabilities that an event belongs to a particular class provided the values of the predictor variables are presumed to be class conditionally independent and do not take the covariance between the predictor variables naively into account [30]. The classifier includes ease of use and quick dealing with multiple attributes of the dataset.

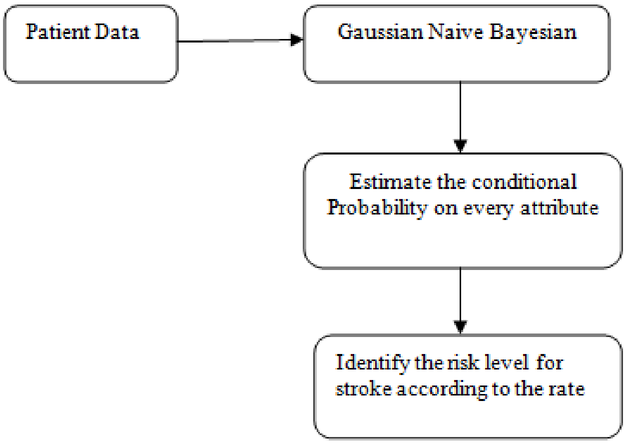

Fig. 5. Gaussian Naïve Bayesian model for predict stroke diseases

- **Proposed Model using Deep Neural Network**

A deep neural network is a multilayered ANN between the layers of input and output. The DNN decides the correct statistical mechanism for converting the input into the output whether it is a linear relationship or a non-linear relation. The introduction of more hidden layers to the network enables a deep architecture to communicate more complex hypotheses as the hidden layers catch the nonlinear relationships .The network travels through the layers estimating the probability of each result shown in below Figure 6.

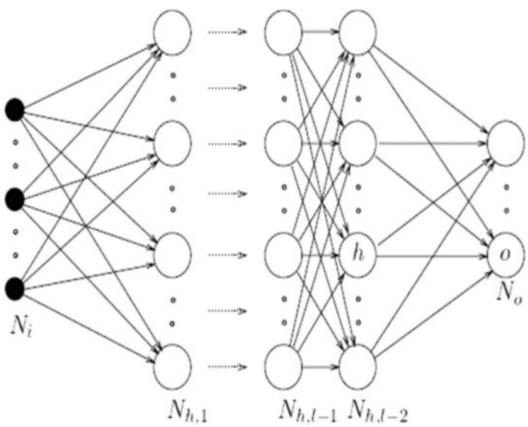

Fig. 6. Deep Neural Network architecture

Traditionally, the detection efficiency is calculated by measuring accuracy (Acc) for all classifiers. In the stroke datasets with 1500 data on survival , an difference was evident. We used extra metrics: sensitivity (S_t), specificity (S_f), and accurate predictive value (APV). The $\mathbf{S_t}$ reflects the likelihood of detecting stroke data, $\mathbf{S_f}$ reflects the likelihood of detecting non stroke data and **APV** is the probability that non-survival or occurrence of stroke status has been corrected for adequate. To test model efficiency we used rue-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN) as a parameter . Accuracy is the fraction of correct classifications to the total number of classifications. The S_t , S_f , APV and Accuracy was determined using below formulas.

$$S_t = TP/(TP + FN) \quad (1)$$

$$S_f = TN / (TN + FP) \quad (2)$$

$$APV = TP / (TP + FP) \quad (3)$$

$$Accuracy= (TP+TN) /(TP+TN+FP+FN) \quad (4)$$

IV. RESULTS AND DISCUSSION

The diverse settings tested, five hidden layers were featured in the best DNN architecture, each with 15 neurons, 25 training epochs and a trial and error dependent batch size of 5. Table 3 display the confusion matrix resulting from preprocessing of scaled PCA.

Table 3. Confusion Matrix for DNN method.

Confusion Matrix	Predicted (T)	Predicted (F)
Actual (T)	173	1123
Actual (F)	146	58

The computational results provided by the classification algorithms are summarized in table 4. We list the values of the confusion matrix and the four output parameters. Compared to commonly used performance metrics sensitivity, specificity, actual predicted value and accuracy with better reflects algorithm performance [31].Considering all predictive model performance parameters, DNN model yields with optimal stroke probability Accuracy of 86.4%, sensitivity of 74.89%, specificity of 88.49%, and APV of 54.23%. Optimum DNN with PCA algorithm followed by Gaussian Naive Bayesian classifier. Table 4 shows the performance parameter evaluation using 12-fold cross validation to authenticate outcomes shown in Figure 7. These outputs strongly support the conclusion that the PCA algorithm with DNN outperforms with other algorithms.

We applied different procedures to stop normalization of the over fitting, dropout and batch. The above prevents the lack of feed-forward data in a manner close to the weighting of initialization and the dropout uses weighting to reduce the effect of any hidden nodes.

Our research predicts stroke using the history and health behavior dataset of emergency service uses. This will raise potential medical expenses and make diagnosis easier. We have used survey data that is binary and used PCA to boost data resolution. The limitations of our work not included MRI image dataset. Furthermore, our results only apply to subjects who could in the future suffer from stroke which could reduce the overall precision of our predictive model.

Table 4. Comparison and performance with metrics of different classifier

Classifier	TP	TN	FP	FN	Accuracy(%)	Sensitivity (%)	Specificity (%)	APV (%)
DT	92	1008	352	48	73.32	65.7	74.1	73.3
EM	136	1041	245	78	78.46	63.54	80.1	78.46
RFC	152	1056	230	62	80.53	71.02	83.4	80.53
GNB	185	1060	198	57	82.46	73.9	84.16	82.46
DNN	**173**	**1123**	**146**	**58**	**86.42**	**74.89**	**88.49**	**86.42**

V. Conclusions

We developed a deep learning model based on data from 1500 participants, featuring PCA to predict stroke based on history of medical usage and health habits. Our research enables early detection of patients at high dangerous of stroke who needs extra testing and suitable therapy until the disorder exacerbation. We used a deep neural network to test the interest variables and PCA to produce better continuous inputs for the deep neural network. Our method's accuracy, sensitivity, and specificity value was 86.42%, 74.89% and 88.49%, respectively. Using the dataset, we will use our approach to predict stroke as well as additional diseases. In future, we will change and extend approach for evaluating other applications of medical facilities and cancer data sets on health behavior. Also, to obtain more accurate DNN performance , we will use comprehensive indices and physiological signals as input data. Finally, we will use automated tuning methods to minimize training time and boost efficiency.

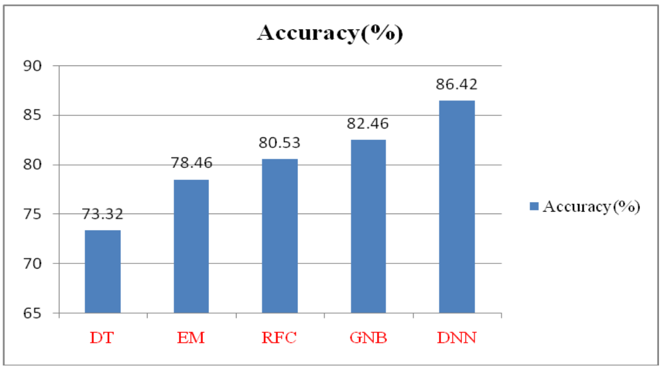

Fig. 7. Accuracy Performance with different Machine Learning Classifier

REFERENCES

1. Kohn L. T., Corrigan J. M., and Donaldson M. S., "To Err is Human:
2. Building a Safer Health System" , Institute of Medicine (IOM). National Academies Press, Washington, 1999.
3. Edward C Jauch, et al, "Ischemic Stroke ". Medscape, Mar 2019.
4. Liebeskind D.S., Albers G.W., Crawford K., Derdeyn C.P., George
5. M.S., Palesch Y.Y., Toga A.W., Warach S., Zhao W., Brott T.G., et al., "Imaging in Stroke Net: Realizing the potential of Big Data" , Journal of Stroke,Vol. 46, Issue 7, pp. 2000–2006, 2015.
6. Liebeskind D.S., "Artificial intelligence in stroke care: Deep learning
7. or superficial insight?", Journal of EBioMedicine, Vol. 35, pp.14-15, 2018.
8. Benjamin E. J., Blaha M. J., Chiuve S. E., et al, " Heart Disease and
9. Stroke Statistics 2017 update: a report from the American Heart Association", Circulation 2017. https://doi.org/10.1161/cir.0000000000000485.
10. Oskar Maier, Christoph Schröder, Nils Daniel Forkert , Thomas
11. Martinetz , Heinz Handels, "Classifiers For Ischemic Stroke Lesion Segmentation: A Comparison Study", Journal of PLOS ONE 10(12), doi:10.1371/journal, 2015.
12. Ashfaq Ahmed K., Aljahdali S. and Hussain S.N. "Comparative
13. Prediction Performance with Support Vector Machine and random Forest Classification Techniques", International Journal of Computer Applications, Vol. 69, No. 11, pp.12–16,2013.
14. Yang Y H ,Sung S F, Hsieh C Y, Lin H J, Chen C H, Chen Y W, Hu Y
15. H,"Developing a stroke severity index based on administrative data was feasible using data mining techniques", Journal of Clinical Epidemiology, Vol. 68, No.11,pp.1292–1300,2015 10.1016/j.jclinepi.2015.01.009.
16. Sasi S ,Alotaibi N N, ," Stroke in-patients' transfer to the ICU using
17. ensemble based model", In: IEEE international conference on electrical, electronics, and optimization techniques (ICEEOT), pp 2004–2010,2016.
18. Dhar R, Chen Y, An H, Lee JM "Application of machine learning to
19. automated analysis of cerebral edema in large cohorts of ischemic stroke patients", Journal of Frontiers Neurology, 9:687, doi: 10.3389/fneur.2018.00687, 2018.
20. Leila Amini, et al.,"Prediction and Control of Stroke by Data
21. Mining", International journal of preventive Medicine, Vol 4,S245–S249, May 2013.
22. Ahmet K. Arslana, Cemil Colaka, and Ediz Sarihanb, "Different
23. Medical Data Mining Approaches Based Prediction of Ischemic Stroke", Journal of Computer Methods and Programs in Biomedicine, March 2016.
24. Eibe Frank, Mark A. Hall, and Ian H. Witten ,The WEKA
25. Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

26. Xu K., Feng D, Mi H, “Deep convolutional neural network-based
27. early automated detection of diabetic retinopathy using fundus image”, Journal of Molecules, Vol.22,No.2054,2017. doi:10.3390/molecules22122054.
28. Izadpanahkakhk M., Razavi S., Taghipour-Gorjikolaie, M., Zahiri
29. S., Uncini, “A Deep region of interest and feature extraction models for palmprint verification using convolutional neural networks transfer learning”, Journal of Applied Science,Vol 8,No. 7, 2018. https://doi.org/10.3390/app8071210
30. Steven Eyobu, O., Han D,” Feature representation and data
31. augmentation for human activity classification based on wearable IMU sensor data using a deep LSTM neural network”, Journal on Sensors, Vol.18, No.9, 2018.
32. Jiao Z., Gao X., Wang Y., Li J, “A parasitic metric learning net for
33. breast mass classification based on mammography”, Journal of Pattern Recognition ,Vol. 75, pp.292–301,2018
34. Hu Y., Li J., Jiao Z, “Mammographic Mass Detection Based on
35. Saliency with Deep Features”,In: Proceedings of the International Conference on Internet Multimedia Computing and Service, Xi’an, China, 19–21 Aug 2016;ACM: New York, USA, 2016; pp. 292–297.
36. Shen D., Wu G, Suk H.I,” Deep learning in medical image analysis”, Journal of Annual Review of Biomedical Engineering ,Vol.19, pp.221–248,2017. doi: 10.1146/annurev-bioeng-071516-044442.
37. Suk, H.I., Shen, D, “Deep learning-based feature representation for AD/MCI classification”, In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Nagoya, Japan, 22–26 September 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 583–590.
38. Hua K.L., Hsu C.H., Hidayati S.C., Cheng W.H., Chen Y.J, “Computer-aided classification of lung nodules on computed tomography images via deep learning technique”, Journal of OncoTargets Therapy, Vol. 8, pp.2015–2022,2015.
39. Anthimopoulos M., Christodoulidis S., Ebner L., Christe A., Mougiakakou S, “Lung pattern classification for interstitial lung diseases using a deep convolutional neural network”, IEEE Transaction on Medical Imaging, Vol.35, pp.1207–1216, 2016.
40. Jiao Z., Gao X., Wang Y., Li J, “A deep feature based framework for breast masses classification”, Journal of Neurocomputing , Vol. 197, pp.221–231, 2016.
41. Jiawei Han, et. al, “Data Mining Concepts and Techniques”, Third edition, P84-88,2011
42. Shlens, J,” A tutorial on principal component analysis”, arXiv **2014**, arXiv:1404.1100.
43. B. P. Rimal and E. Choi, "A service-oriented taxonomical spectrum, cloudy challenges and opportunities of cloud computing," International Journal of Communication Systems, vol. 25, pp. 796-819, 2012.
44. Gupta M, Chen Y, “Theory and use of the EM algorithm”, Journal of Foundations and Trends in Signal Processing Vol. 4, issue.3 , pp.223-96, 2010.
45. Carson C, Belongie S, Greenspan H, Malik J. Blobworld, “Image segmentation using expectation-maximization and its application to image querying”, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 24, No. 8, pp.1026-1038, 1999.
46. Małgorzata Z **,”**Attribute Selection For Stroke Prediction”, Journal of Acta Mechanica et Automatica”, Vol.13, No. 3, pp.200-204, 2019.
47. G.Subbalakshmi et. al.,”Decision Support in Heart Disease Prediction System using Naive Bayes “, Indian Journal of Computer Science and Engineering (IJCSE) ,Vol. 2, No. 2, 2011.
48. Chollet, F. Keras ,”The Python Deep Learning Library; Astrophysics Source Code Library”, 2018. Available online: http://adsabs.harvard.edu/abs/2018ascl.soft06022C.

AUTHORS PROFILE

Maya B S working as Assistant Professor in department of computer science & Engineering, Bangalore Institute of Technology, Bangalore. Currently pursuing Ph.D in Computer Science & Engineering at VTU in the area of medical image processing published 4 papers in different International Journal and conference.

Dr. Asha. T is a Professor & HOD in the Department of Computer Science & Engineering, Bangalore Institute of Technology, Bengaluru. She obtained her Ph.D in Computer and Information Science from Visvesvaraya Technological University, Karnataka. She has published around 31 papers in International/National journals and Conference. Her research interests include Data Mining, Medical Informatics, Machine Learning, Pattern Recognition and Big data management etc., email: asha.masthi@gmail.com, asha@bit-bangalore.edu.in