Prediction of Onset of Diabetes using Adaptive Boosting

Pushpa S. K, Manjunath T N, Bhavya G, Vinutha K, Anupchandra Rao M C

Abstract*: Diabetes is one of the most common diseases, as per the survey in 2015, 30 million people in US are suffering from this disease, i.e about 90-95 percent of the population. If diabetes is untreated at the early stages, high blood glucose in the body leads to various other health problems like: eye problems, stroke, nerve damage, heart disease, stroke etc. Technology has seen an explosive growth in the development and use of Artificial Intelligence in various domains. The increased sophistication and capabilities of these tools are unlocking new possibilities in fields of Medicine, Agriculture, Manufacturing and Automobiles. The goal of this work is to predict the onset of diabetes using Machine Learning namely Adaptive Boosting. Boosting is a technique wherein a series of low accuracy classifiers are combined to create a high accuracy classifier. In many areas the problems are so complicated that simple algorithms such as KNN, Decision Tress are incapable of making predictions. Hybrid algorithms such as Random Forests and Gradient Boosting are gaining popularity due to these reasons are used by multinational companies one example being Netflix. In this work Decision Tree and Support Vector Machine methods has been considered with eight important attributes namely, Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age and predicts if a person has diabetes. Multiple models are built using decision tree and support vector machine without Adaptive Boosting and with Boosting technique and the results are compared and evaluated. Result shows that support vector machine gives an improved overall accuracy of 80%.

***Keywords*: Machine Learning, Ensemble Learning, Ada boaster..**

I. INTRODUCTION

Medicinal services organizations all things considered, types, and strengths are ending up progressively inspired by how man-made brainpowerr can bolsterr betterr patient consideration while lessening costs and improving efficiencies. Overr a generally brief time frame, the accessibility and complexity of computer based intelligence has detonated, leaving suppliers, payers, and different partners with a confounding exhibit of instruments, advancements, and procedures to browse. Man-made rreasoning (simulated intelligence), machine learning, and profound learning are overwhelming the human services industry. They are not la-la-land advances any more; they are down to earth devices that can help organizations upgrade theirr administration arrangement, improve the standard of consideration, create more income, and diminishing danger. About every rreal organization in the medicinal services space have just stated to utilize the innovation by and by.

Gathering models have been utilized broadly in credit scoring applications and different territories since they are increasingly steady and, all the more significantly, perform superiorr to single classifiers. They are likewise known to lessen model predisposition and difference. Therapeutic informational indexes comprise of an enormous number of highlights. The Exhibition of the classifier will be deceased if the informational indexes contain uproarious highlights. Diabetes is a sickness which happens when the blood glucose level turns out to be high, which eventually prompts other wellbeing. Diabetes is caused predominantly because of the utilization of profoundly prepared sustenance, terrible utilization propensities and so on. As indicated by WHO, the quantity of individuals with diabetes has expanded throughout the years. Early identification and treatment of diabetes is a significant advance toward keeping individuals with diabetes sound. It can diminish the dangerr of genuine confusions, for example, untimely coronary illness and stoke, visual impairment, appendage rremovals, and kidney disappointment. About one in seven U.S. adults has diabetes now, according to the Canters forr Disease Control and Prevention. But by 2050, that rate could skyrocket to as many as one in three. High blood glucose levels can damage the body's organs. Possible complications include damage to large (macrovascular) and small (microvascular) blood vessels, which can lead to heart attack, stroke, and problems with the kidneys, eyes, gums, feet and nerves. Early detection on the onset of Diabetes can improve the process of treatment to a great extent possibly even lead to a cure. Furthermore, machine learning algorithms that predict outcomes using a single model do not provide adequate accuracy. This calls for the use of Hybrid Algorithms such as Bagging and Boosting in-order to improve accuracy[8]-[12].

II. LITERATURE SURVEY

Shaowen [4] observed that as the iterative numberr increases, it leads to the Degeneration Phenomenon wherein the generalization ability of the classifierr is reduced. In order to circumvent this problem, he proposes an LWE-Adaboost algorithm which limits weight expansion.

Revised Manuscript Received on January 15, 2020
***** Correspondence Author
Pushpa S. K, ISE department, BMSIT&M, Bengaluru, India. Email: murtheppa16pushpa@gmail.com.
Manjunath T N, ISE department, BMSIT&M, Bengaluru, India. Email: manju.tn@bmsit.in.
Bhavya G, ISE department, BMSIT&M, Bengaluru, India. Email: bhavyasati@bmsit.in
Vinutha K, ISE department, BMSIT&M, Bengaluru, India. Email: vinuthak_ise2014@bmsit.in.
Anupchandra Rao M C, ISE department, BMSIT&M, Bengaluru, India. Email: anupchandra7@gmail.com

Results obtained from experiments indicate that this algorithm can rrestrain the occurrence of the

Degeneralization Phenomenon. He proposes to modify the weight update method by not just considering the overall classification to the error of classification but also according to the specific conditions of each sample which are added to the predicted set forr each sample to determine uncertainty. Thereby achieving lower false detection rate and higher prediction accuracy. Prior to the training process all samples are normalized to rreduce the impact of the image itself due to grey level distribution. OpenCV is used to detect the effect of human face detection.

Moon-Hyun Kim [7] has pointed out that the performance of the ensemble depends on the diversity among the memberr classifiers as well as the performance of each memberr classifiers. According to Kim existing Adaboost algorithms are focused on error minimization problems. He thus proposes to inject diversity into the boosting process in order to improve the performance of the Adaboost Classifierr which outperforms the rregularr unmodified algorithm. AdaBoost selects a member classifier to minimize the error in each cycle in comparison, AdaBoost selects a member classifier to minimize error and to maximize the diversity among the member classifiers in each cycle. AdaBoost selects candidate classifiers by measuring the difference between minimum error and error of each weak classifier in each cycle. If the difference is smaller than a threshold value, the weak classifier is selected as a candidate classifier. The diversity between the last generated ensemble classifier and each candidate classifier. Diverse AdaBoost can thus improve the generalization performance of ensemble classifiers by considering diversity while sacrificing accuracy of a weak classification in each cycle.

Kisang Kim[3] proposed to assign different weights to initial datum based on statistical properties of attributes. He first identifies the problem in Adaboost wherein it chooses some set of features that are most effective forr classifying the training data. At inception all training data is treated equally howeverr, when the next feature is to be determined it assigns a different weight to each training data, so that the misclassified data in the previous stage gets classified correctly in the next iteration. It also makes use of a threshold value to be used when an input needs to be classified. The optimal threshold value is the one that minimizes misclassification rate. He furtherr identified that positive training data tend to have a well concentrated distribution of feature names while negative training samples have a scattered distribution of attribute values. Using this difference in polarity in data samples we assign different initial weights. Namely negative samples are assigned an equal initial weight whereas positive samples are assigned a higher initial weight which rresults in the attribute value being closer to the peak of the overall distribution of total positive training data as shown in Fig.1.

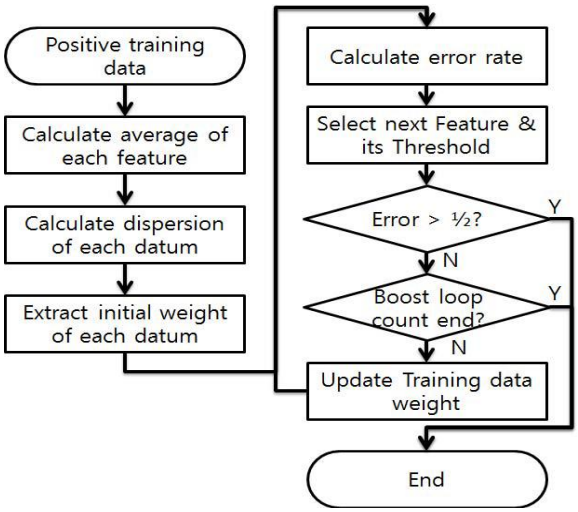

Fig. 1. Adaboost optimization proposed by Kisang.

Kuldeep Randhawa [2] compared ensemble machine learning methods with standard methods and concluded that hybrid models namely majority voting method achieves good accuracy rates when used with publicly available card data and real-world credit card data from a financial institution to detect fraudulent transactions. A total of 12 machine learning algorithms ranging from Neural Networks to Deep Learning Models along with hybrid models such as Adaboost and Majority Voting. Furthermore, noise was added to the dataset to test the robustness and reliability of all models. Majority voting is one of the most frequently used methods in data classification, which involves a combined model with at least two algorithms. Each algorithm makes its own prediction for every test sample. The final output is for the one that receives the majority of the votes. The MCC metric has been adopted as a performance measure, as it takes into account the true and false positive and negative predicted outcomes. The best MCC score is 0.823, achieved using majority voting.

Ting Zhang [5] proposed a face detection algorithm as mentioned in Fig.2 that combines skin color segmentation and Adaptive Boosting. The algorithm sets up the skin Gaussian model in YCbCr color space using skin color clustering characteristics it then sorts skin color regions using the Adaboost classifier to detect the face in the image. Face detection using skin color segmentation has the advantage that it can detect faces with high accuracy but the disadvantage is that it also detects the body of the human along with the background when it is complex. The algorithm builds a cascade classifier to detect faces, with the advantage of having improved detection speed. However, the downside is that the error detection also increases with the increase in detection rate. The algorithm first inputs the color image into the detection module to get candidate face region.

Next the classifier is fed the detected region for more accurate location so that color information can be used more quickly to eliminate irrelevant background content and to increase speed of detection and to minimize the error detection rate.

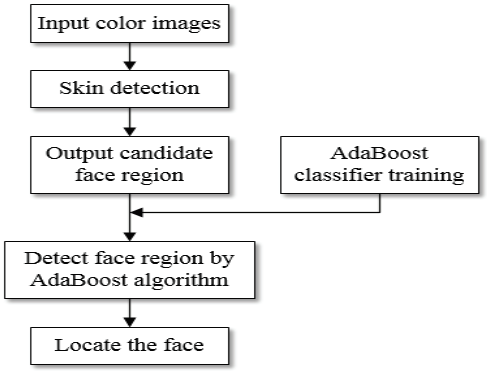

Fig 2. Face Detection using Adaboost

Astina Minz's [1] proposal makes use of Machine Learning classification methods for Magnetic Resonance Imaging (MRI) as shown in Fig.3. MRI is one of those reliable imaging techniques used for medical diagnosis. Manual assessment of these images is a tedious job as the amount and granularity of the data are too complex to be properly analyzed by humans. Minz hence proposes to use the Adaboost classification technique for Brain Tumor Detection. The proposed system consists of three phases namely Pre-processing, Feature Extraction and Classification.

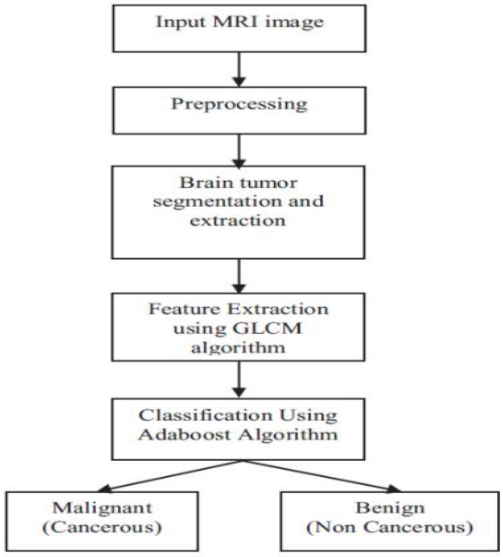

Fig 3. MRI Analysis and Classification using Adaboost

The MRI brain images are acquired and are given as input to the preprocessor. The RGB MR image is transformed to grayscale image and then a median filter is applied in order to eliminate noise from the images. The brain MRI is segmented by using the thresholding technique. In feature extraction the important features required for image classification are extracted. The segmented brain MR image is used, and texture features are extracted from the segmented image which illustrate the texture property of the image. Extraction of features is done using GLCM(gray level co-occurrence matrix) algorithm. The Machine learning algorithms (Adaboost) classify the MR brain image either as normal or abnormal.

III. EXISTING SYSTEM

Astina Minz's proposal makes use of Machine Learning classification methods for Magnetic Resonance Imaging (MRI). Ting Zhang [] proposed a face detection algorithm that combines skin color segmentation and Adaptive Boosting. The algorithm sets up the skin Gaussian model in YCbCr color space using skin color clustering characteristics it then sorts skin color regions using the Adaboost classifier to detect the face in the image. Kuldeep Randhawa [] compared ensemble machine learning methods with standard methods and concluded that hybrid models namely majority voting method achieves good accuracy rates when used with publicly available card data and real-world credit card data from a financial institution to detect fraudulent transactions. Kisang Kim[] proposed to assign different weights to initial datum based on statistical properties of attributes. Shaowen [] observed that as the iterative number increases, it leads to the Degeneration Phenomenon wherein the generalization ability of the classifier is reduced. In order to circumvent this problem, he proposes an LWE-Adaboost algorithm which limits weight expansion. Moon-Hyun Kim [] has pointed out that the performance of the ensemble depends on the diversity among the member classifiers as well as the performance of each member classifiers.

IV. PROPOSED SYSTEM

To predict the onset of diabetes using eight important attributes i.e. Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age using the Adaptive Boosting algorithm with Decision Trees and Support Vector Machines as base estimators and evaluate the results.

The proposed system predicts the onset of diabetes using Adaptive Boosting. It makes these predictions by building a model on eight important parameters. The data set is first analyzed and cleansed, this formatted data is sent to four different models built using Decision Trees, Support Vector Classifier, DT with Adaptive Boosting, SVM with Adaptive Boosting. These models are evaluated using two methods namely Train_Test_Split and K-Fold Cross Validation. The results obtained are compared and observed.

A. Objectives

- To research into the working of Ensemble Machine Learning methods, in areas where traditional machine learning methods are insufficient.
- To build a highly accurate machine learning classifier using eight important parameters that predicts if a person has diabetes using the Adaptive Boosting Algorithm.
- To compare two machine learning model evaluation methods (Train Test Split vs K Fold Cross Validation).

From the above Literature Survey, we can observe that in most cases Hybrid algorithms aka Ensemble Learning methods such as Adaboost and Majority Voting outperformed simple algorithms such as KNN, Decision Trees etc. Adaptive Boosting was used in many different scenarios ranging from Medical Diagnosis,

Face Detection, Credit Card Fraud Detection to general classification applications. Modifications to the algorithm either through updation of weights or injection of diversity into the boosting process were made to improve the speed and classification accuracy.

V. IMPLEMENTATION

The program begins with opening the dataset as a panda's dataframe. The training attributes and class values are stored in separate variables. The data is cleansed by removing mis-classified instances, null values and instances having zero as value. This cleansed data is passed to the Train-Test-Split function where the two sets, namely training set and test set of data are obtained this in-turn is passed to A Decision Tree Classifier, A Support Vector Classifier, An Adaptive Boosted Decision Tree Classifier and last but the least An Adaptive Boosted Support Vector Classifier. The model is built for the four different classifier's and the results are evaluated. Evaluation is also done using the K-Fold Cross-Validation Technique since it acts as a more reliable metric to measure accuracy. The Precision and Recall are displayed for each of the classifiers as well.

A. Module Description

Scikit-Learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use. The library is built upon the SciPy (Scientific Python) that must be installed before use. This stack that includes:

- **NumPy:** Base n-dimensional array package
- **SciPy:** Fundamental library for scientific computing
- **Matplotlib:** Comprehensive 2D/3D plotting
- **IPython:** Enhanced interactive console
- **Sympy:** Symbolic mathematics
- **Pandas:** Data Structures and analysis

Extensions or modules for SciPy care conventionally named SciKits. As such, the module provides learning algorithms and is named Scikit-Learn. The vision for the library is a level of robustness and support required for use in production systems. This means a deep focus on concerns such as ease of use, code quality, collaboration, documentation and performance. Some popular groups of models provided by Scikit-Learn include:

- Clustering: for grouping unlabeled data such as KMeans.
- Cross Validation: for estimating the performance of supervised models on unseen data.
- Datasets: for test datasets and for generating datasets with specific properties for investigating model behavior.
- Dimensionality Reduction: for reducing the number of attributes in data for summarization, visualization and feature selection such as Principal component analysis.
- Ensemble methods: for combining the predictions of multiple supervised models.
- Feature extraction: for defining attributes in image and text data.
- Feature selection: for identifying meaningful attributes from which to create supervised models.
- Parameter Tuning: for getting the most out of supervised models.
- Manifold Learning: For summarizing and depicting complex multi-dimensional data.
- Supervised Models: a vast array not limited to generalized linear models, discriminate analysis, naïve-bayes, lazy methods, neural networks, support vector machines and decision trees.

B. Process

1. Initially, all observations are given equal weights.
2. A model is built on a subset of data.
3. Using this model, predictions are made on the whole dataset.
4. Errors are calculated by comparing the predictions and actual values.
5. While creating the next model, higher weights are given to the data points which were predicted incorrectly.
6. Weights can be determined using the error value. For instance, the higher the error the more is the weight assigned to the observation.
7. This process is repeated until the error function does not change, or the maximum limit of the number of estimators is reached.

C. Function Usage

train_test_split(): Split arrays or matrices into random train and test subsets

- read_csv(): Reads a csv file as a pandas dataframe.
- metrics.confusion_matrix(): Returns the confusion matrix given the predicted and observed output.
- metrics.accuracy_score(): Returns the accuracy of the model given the predicted and observed output.
- metrics.precision_score(): Calculates and returns the precision value given the predicted and observed output.
- metrics.recall_score(): Calculates and returns the recall value given the predicted and observed output.
- classifier.fit(): Builds a specific model given the training data set
- classifier.predict(): Predicts the output of a given input instance.

D. Evaluation Methods

Train/Test Split

This method split the data set into two portions: a training set and a testing set. The training data set is used to build i.e. train the model. And the testing set is used to test the model and evaluate its accuracy as mentioned in Fig.4.

Pros: But, train/test split is still useful because of its flexibility and speed.

Cons: Provides a high-variance estimate of out-of-sample accuracy.

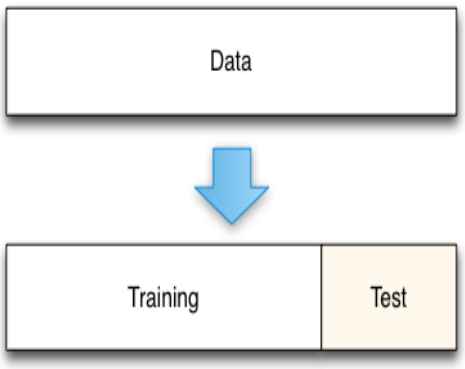

Fig 4. Evaluation via Train-Test-Split

K-Fold Cross Validation

This method splits the data set into K equal partitions ("folds"), then use 1-fold as the testing set and the union of the other folds as the training set. Then the model is tested for accuracy. The process will follow the above steps K times as shown in Fig.5, using different fold as the testing set each time. The average testing accuracy of the process is the testing accuracy.

Pros: More accurate estimate of out-of-sample accuracy. More "efficient" use of data (every observation is used for both training and testing)

Cons: Much slower than Train/Test split.

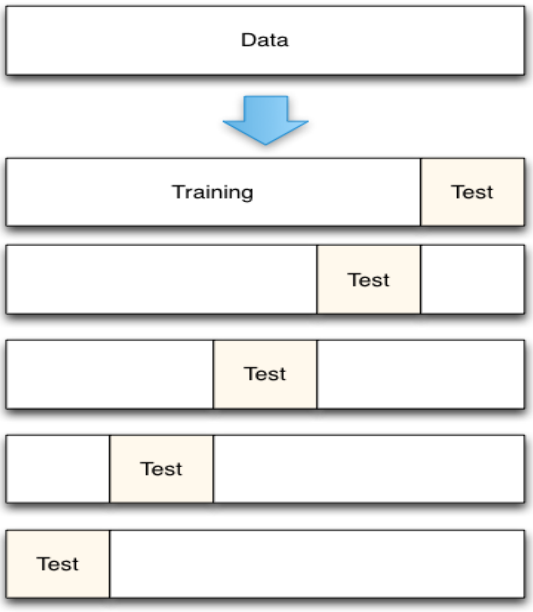

Fig 5. Evaluation via K-Fold Cross Validation

VI. RESULTS AND DISCUSION

The Table below shows the results of two machine learning algorithms one is decision tree and another support vector machine as base estimators and with adaptive boosting method.

Table- I: Name of the Table that justify the values

Name of the Algorithm	Accuracy (Run1)	Accuracy (Run2)
Decision Tree	73%	68%
Support Vector Machine	68.9%	64%
With Adaptive Booster		
Decision Tree	73%	66%
Support Vector Machine	82.7%	71%

A. Accuracy Graphs

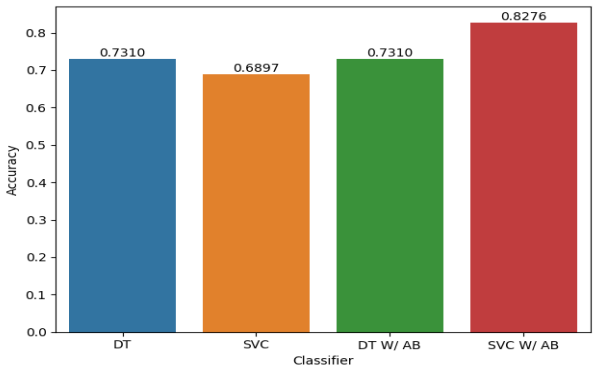

Fig 6. Comparison of Accuracy (Run1)

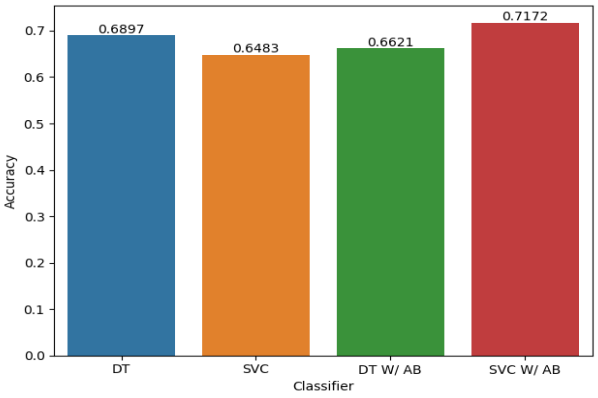

Fig 7. Comparison of Accuracy (Run2)

B. Observations

☐ Adaptive Boosting applied to the same dataset using two different base estimators yielded some interesting results as mentioned in Fig.6 and Fig.7:

☐ The machine learning model that was built using Decision Trees as base estimators predicted if a person has diabetes or not with accuracy close to 75%.

☐ The model that was built using Support Vector Machines as base estimator gave an improved overall accuracy of 80%.

☐ It was also noticed that the model sometimes overfitted the data points thereby predicting with reduced accuracy.

Furthermore, the accuracy of the classifier improved to a great extent when the value of learning rate was decreased gradually.

Learning rate is a hyper-parameter that controls how much we are adjusting the weights of our network with respect the loss gradient. The lower the value, the slower we travel along the downward slope. The learning rate affects how quickly our model can converge to a local minima (aka arrive at the best accuracy). Thus, getting it right from the get go would mean lesser time for us to train the model. However, since the application we are using the model requires highest accuracy and can bear to offset performance. Hence, we proceed to choose a smaller learning rate.

VII. CONCLUSION

As we are aware that diabetes is one of the most common disease, and we have sophisticated techniques to early predict these disease at the earliest. The use of Machine Learning in the field of Medical sciences is increased exponentially over the past few years. However traditional algorithms which build single models aren't capable of high accuracy in this particular domain. This calls for hybrid algorithms such as Adaboost or XGBoost including algorithms such as Random Forests. Since Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Hence it is evident that a SVM algorithm shows overall improvement with adaptive boost method.

REFERENCES

1. Astina Minz, Chandrakant Mahobiya MR Image classification using Adaboost for brain tumor type 978-1-5090-1560-3/17 $31.00 © 2017 IEEE DOI 10.1109/IACC.2017.137
2. Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim, Asoke K. Nandi, Credit card fraud detection using AdaBoost and majority voting, 10.1109/ACCESS.2018.2806420, IEEE 2018.
3. Kisang Kim, Hyung-Il Choi Adjusting Initial Weights for Adaboost Learning, 10.1109/CAIPT.2017.8320686, IEEE 2017.
4. Liao Shaowen , Zhang Jianqing, Chen Yong Further improvement of AdaBoost algorithm IEEE 2015.
5. Chongshan Lv, Ting Zhang, Cong Lin, Face Detection Based on Skin Color and AdaBoost Algorithm, 978-1-5090-4657-7/17 2017 IEEE.
6. Zhiquan Qi, Fan Meng, Yingjie Tian, Lingfeng Niu, Yong Shi, and Peng Zhang Adaboost-LLP: A Boosting Method for Learning With Label Proportions 2162-237X © 2017 IEEE
7. Tae-Ki An, Moon-Hyun Kim A New Diverse AdaBoost Classifier 978-0-7695-4225-6/10 $26.00 © 2010 IEEE DOI 10.1109/AICI.2010.82
8. Nongyao Nai-aruna and Punnee Sittidech, Ensemble Learning Model for Diabetes Classification, Advanced Materials Research Vols. 931-932 (2014) pp 1427-1431, doi:10.4028/www.scientific.net/AMR.931-932.1427
9. G. Liang, and C. Zhang, Empirical Study of Bagging Predictors on Medical Data, 9th Australasian Data Mining Conference, 121, 31-40, (2011).
10. J. Abellán, Ensemble of decision tree based on imprecise probabilities an uncertainty measures, Information Fusion, 14,423-430, (2013).
11. I. Syarif, E. Zaluska, A. Prugel-Bennett and G. Wills, Application of Bagging, Boosting and Stacking to Intrusion Detection, MLDM2012, LNAI7376, 513-602, (2012).
12. T.G. Dietterich, An Experimental Comparison of Three Methods for Construction Ensembles Of Decision Trees: Bagging, Boosting, and Randomization, Machine Learning, 40, 139-157, (2000)

AUTHORS PROFILE

Dr. Pushpa S. K is working as Professor in Dept. of ISE, BMSIT&M. She has completed her B.E in Computer Science and Engineering in the Year 1995, M.E in the Year 2004 and Ph. D in the Year 2017. She has 18 Years of Teaching Experience and presently guiding three Ph.D students. Her area of interest is Wireless Sensor Networks, IoT and Machine Learning. She has published 27 research articles in various International Conferences and Journals and delivered 4 expert talks in various faculty develop programme on Machine Learning & organized 3 National Conferences at BMSIT&M. She is a life member of "The Institution of Engineers", "Indian Science Congress Association" and "Indian Society for Technical Education".

Dr. Manjunath T N has completed his B.E from SJCIT, Chickballapur and M. Tech from JNNCE, Shimoga and Ph. D from R & D Centre, Bharathiar University, and Coimbatore. Awarded: April-2015. At present he is guiding 5 Ph. D students. He has published more than 60 technical papers in various International Conferences and Journals. He has actively involving in organizing various technical talks, workshops and conferences. He has teaching and Industrial experience of more than 18 years in various company and Institutions like Global Softech India, Accenture services, Wipro Technologies, Bangalore, SJBIT, Acharya Institute of Technology. He is a Life member of Indian Society for Technical Education, International Association of Engineers and Member of the Society of Digital Information and Wireless Communications. At present he is working as Dean External Relations and Professor, Dept. of ISE, BMSIT&M, Bengaluru.

Mrs, Bhavya G working as Assistant Professor in Department of Information Science and Engineering, BMSIT&M. She has completed her B.E in BTLIT Bangalore in 2011 and M.Tech from AMC Engineering College 2013 in CSE. She is having 6 years of teaching experience and pursuing research in the field of breast cancer prediction using Machine learning. Her area of interest is Machine learning, IOT and Data mining. She has published 9 research papers in various International Conference and Journals and delivered expert talk on Machine learning topic.

Vinutha K is working as Assistant professor in Dept of ISE, BMSIT&M. she has completed her B.E in Channabasaveshwar Institute of Technology Tumkur in 2011 and M.Tech from AMC Engineering College 2013 in CSE. She is having 5 years of teaching experience and pursuing research in the field of prediction models using Machine learning. Her area of interest is Machine learning and Data mining. She has published 7 research papers in various International Conference and Journals and delivered 5 expert talks on various topics of Machine learning, python and data mining.

Anupchandra Rao M C, Student, ISE department, BMSIT&M.