*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 5*
*DOI: https://doi.org/10.17762/ijritcc.v10i5.5552*
***Article Received**: 20 March 2022 **Revised**: 04 April 2022 **Accepted**: 25 April 2022 **Publication**: 31 May 2022*

# Tree Based Boosting Algorithm to Tackle the Overfitting in Healthcare Data

**Blessa Binolin Pepsi M[1], Vidhya S[2],Ashwini A[2]**
[1]Assistant Professor (Senior Grade),[2]UG Student
[1,2] Department of Information Technology
[1,2]Mepco Schlenk Engineering College
Sivakasi, Tamilnadu.
[1]mblessa@mepcoeng.ac.in
[2]vidhyasrinivasan020_it@mepcoeng.ac.in
[2]ashwininila2000_it@mepcoeng.ac.in

**Abstract:** Healthcare data refers to information about an individual's or population's health issues, reproductive results, causes of mortality, and quality of life. When people interact with healthcare systems, a variety of health data is collected and used. However, these healthcare data are noisy as well as it prone to over-fitting. Over-fitting is a modeling error in statistics that occurs when a function is too closely aligned to a limited set of data points. As a result, the model learns the information and noise in the training data to the point where it degrades the model's performance on fresh data. The tree-based boosting approach works well on over-fitted data and is well suited for healthcare data. Improved Paloboost performs trimmed gradient and updated learning rate using Out-of-Bag mistakes collected from Out-of-Bag data. Out-of-Bag data are the data that are not present in In-Bag data. Improved Paloboost's outcome will protect against over-fitting in noisy healthcare data and outperform all tree baseline models. The Improved Paloboost is better at avoiding over-fitting of data and is less sensitive, according to experimental results on health-care datasets**.**

***Keywords**:* Healthcare Data, Missing value, Data Preprocessing, Classification, Gradient Boosting Machine, XGBoost.

## I. Introduction

Healthcare data[1] normally has strange and unlike data. We say that healthcare data[21] has unlike data because it has less number of samples compared to their features, for example, sometimes in hospitals or any healthcare facilities, people will not often collect the blood tests at each visit of the patient because of this the column in their record remains null. So this feature is called “missing values”. Sometimes the data has noisy data also which means it has meaningless data. This occurrence of noisy data is due to the mistake made by the human or incomplete information of a particular patient. So many healthcare data are complex which leads to over-fitted data while simple data does not

suffer from much over-fitting. Since our healthcare data was collected from many sources it has many over-fitted data so to overcome this, we propose the algorithm that is tree-based [2] or tree-boost algorithm which is best suited for healthcare data. In trees, base trees are fitted with the errors at each stage because trees are processed in a stage-wise fashion. An insensitive data type is another added advantage of a tree. The tree-boost algorithm, Stochastic Gradient Tree Boost [3] is widely used because it produces robust performance over other Classification, Regression machine learning techniques.[4] SGTB tackles the over-fitting very easily because it introduces the randomness so because of randomness the computation time also gets speeds up.

Gradient Boosting Machine [5] combines the predictions from multiple decision trees to generate the final predictions[18]. Generally, SGTB provides a default hyper-parameter for better results, and to get accurate results, hyper-parameter tuning is required to tune the data Therefore SGTB provides four hyper-parameter to tune the data. The four hyper-parameters are 1) depth of the tree, 2) the number of trees required, 3) default learning rate, and 4) sampling rate. Increasing trees increases the chance of over-fitting. So to reduce over-fitting, lower learning rates are used but usage of lower learning rates may require more trees to achieve better performance.

So optimal hyperparameters are required to achieve accurate results and also to tackle the over-fitting. Since healthcare data has “noisy labels “, “ meaningless data”, “a large number of features “, tuning is done using Improved Paloboost [6] where Improved Improved Paloboost extends Stochastic Gradient Tree-Boost which helps to tune the hyperparameters to achieve the better results.” Improved Paloboost” stands for Pruning and adaptive learning with Out-Of-Bag Samples”.[7]

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 5*
*DOI: https://doi.org/10.17762/ijritcc.v10i5.5552*
***Article Received**: 20 March 2022 **Revised**: 04 April 2022 **Accepted**: 25 April 2022 **Publication**: 31 May 2022*

Out-Of-Bag sample means where the data are not chosen for the sampling process. Since the trees are processed in a stage-wise fashion, in Improved Paloboost also at each stage, to determine the properties of trees the Out -Of -Bag errors are used.

Improved Paloboost uses Out-Of-Bag-Samples to deduct Out-Of Bag error is also called Out-of-bag estimate when this error increases the tree gets over-fitted and measuring the error in machine learning model like random forests, boosted decision trees, it's utilizing bootstrap aggregating. So to reduce this over-fitting the tree leaves are pruned to reduce the complexity. Pruning is the removing of tree leaves which does not suit the base tree to provide accurate results. Because of this size of the decision tree gets reduced. This is called "Pruning the tree ". Improved Paloboost uses the in-bag samples because it prunes the data so that learning rates will be adjusted. So because of Pruning, hyper-parameter tuning occurs at each stage with different Out-Of-Bag-Samples in Improved Improved Paloboost. The hyper-parameter is not independent of the others.

To find the optimal hyper-parameter with help of 'Grid Search' which calculates the performance for each combination of all the supplied hyper-parameters and their values, and then chooses the optimum value for the hyper-parameters and it is robust one and it is not much more sensitive to data types. Improved Paloboost performance curves are fairly consistent regardless of hyper-parameter settings. While Improved Improved Paloboost's important features are the results of both the regularization techniques like Adaptive learning rate and Gradient aware pruning. The Adaptive learning rate method is an optimizer and is used to minimize the objective function of the gradient descent method. Gradient aware pruning is used for classification and Regression to filter the pruning. It dynamically adjusts the tree depth and saves the computation resource and research time. The result of Improved Improved Paloboost illustrates the adaptive learning rates that guard against over-fitting.

## II. Objective

The main objective of this work is to

- To perform classification techniques using Improved Improved Paloboost.
- To tackle over-fitting in noisy healthcare data.
- To compare Improved Improved Paloboost with different tree baseline models.

It tackles the missing values in healthcare data using a tree-based boost algorithm which produces robust performances over missing values in data. Boosting [10],[20] converts weak learners to strong learners by reducing bias and variance, and many algorithms were used. One of the most used algorithms is the Tree-based algorithm. Trees are used because they are very much helpful in reducing over-fitting in data. Some tree-based algorithms are "Decision-Tree", "XGBoost", "Improved Improved Paloboost", "Gradient Boosting Machine " and "Random Forest ". All the data s are derived from multiple data s and multiple sources. Here "Length-Of-Stay" data set was used which was derived from the source "Physionet-2012 " [8] which contains some missing values which can be solved or handled using this Tree-based boosting algorithm using a machine learning technique called "Classification " [9].

Normally healthcare data has missing as well as noisy data. The healthcare data has been gathered from online sources. So we propose the algorithm called "Improved Improved Paloboost", which is based on supervised learning i.e. classification model to tackle noisy data in healthcare. The classification algorithms are Gradient Boosting Machine, XGBoost, Decision Tree, and Random Forest. Binary label classification is proposed with the methods above since the class labels fall under a binary category like 0 's and 1's.

The result of Improved Paloboost will guard against over-fitting in noisy healthcare data and gives better performance overall tree baseline models.

## III.Related works

Boosting algorithms create models in stages, with basic learners learning sequentially to produce a strong final model.

**Gradient Boosting Machine(GBM):**

Gradient boosting Machine[15],[5] combines the predictions from multiple decision trees to generate the final predictions. Boosting is a technique that combines multiple weak learners and predicts strong learners,so that speed and accuracy will be high, especially for large and complex data. The gradient is used to minimize the loss function, basically Gradient Boosting Machine is the partial derivative of the loss function. It describes the steepness of error function and minimizes the overall prediction error. It have a lots of flexibility so can optimize on different loss function and provides several hyper parameter tuning. Finally model will improve the performance of an algorithm.

$$F* = \operatorname{argmin} \sum_{i=1}^{N} L(y, x) \text{ ---(1)}$$

Gradient Boosting Machine F-function is estimate the minimum value of the given data.. N-gives pairs of target and L-loss function.

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m) \text{ ---(2)}$$

Gradient derivative of loss function of each parameter, calculate Step size, learning rate and new parameters based on equation(1).

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 5*
*DOI: https://doi.org/10.17762/ijritcc.v10i5.5552*
***Article Received**: 20 March 2022 **Revised**: 04 April 2022 **Accepted**: 25 April 2022 **Publication**: 31 May 2022*

$F_m(x) = F_0 + \sum_{m=1}^{m} \beta_m h_m(x)$ ---(3)

New weak learners are created, keep repeating step(3), keep generating new learners until step size is very small [19]. Finally update the prediction and minimize the loss function.

### A. XGBoost

XGBoost[12] and Gradient boosting machine[15] both follows the principle of gradient boosting. Parallel tree boosting was provided by XGBoost and used in many leading machine learning library. It uses more accurate approximations to find the best tree model. It provides more information about the direction of the gradient and minimize the loss function.

XGBoost delivers more accurate approximations by using the strengths of second order derivative of the loss function L1 and L2. The goal of regularization is to reduce variance while raising bias in order to reduce over-fitting models. Ridge(L2) adds the sum of the beta coefficients squared, while Lasso(L1) adds the sum of the absolute beta coefficients.

L1 regularization reduces the number of features in a large dimensional data set by producing binary weights from 0 to 1 for the model's features. L2 regularization spreads out the error terms over all weights, resulting in more precise customized final models..XGBoost is a more regularized model formalization which control over-fitting and improves model generalization.

L1,L2 parameters are used to find the value of the leaf node tree(O_value).

Formula used,

$$O\_value = 2 * \text{sum of residuals} +\_\alpha/\text{number of residuals} + \lambda$$

## IV. Proposed System Design

The proposed algorithm includes different steps executed sequentially. They are listed below with the flow diagram of it in Fig. 1

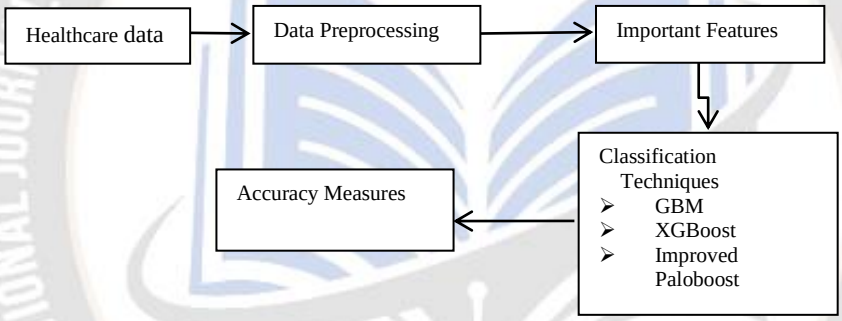



Fig. 1: Overall Workflow of Proposed System

The proposed system is implemented with the following modules :

A. Problem Definition
B. Data Preprocessing
C. Important Feature using Genetic Algorithm
D. Classification techniques
E. Improved Paloboost

### A. Problem Definition

Healthcare data is notorious for having bizarre and unusual data. Healthcare data is unlike data because it includes fewer samples relative to its features. For example, blood tests are not always collected at each visit of a patient in hospitals or other healthcare facilities, therefore the column in their record remains null. This feature is known as "missing values". Due to missing values over-fitting occurs. Over-fitting is a modeling error in statistics that occurs when a function is too closely aligned to a limited set of data points.

As a result, the model learns the information and noise in the training data to the point where it degrades the model's performance on fresh data. The tree-based boosting approach works well on over-fitted data and is well suited for healthcare data. The overall goal is to eliminate missing values and deal with data over-fitting. Supervised learning techniques are used to tackle the missing and over-fitted data.

Improved Paloboost uses Out-Of-Bag-Samples to calculate Out-Of-Bag error, also known as Out-of-bag estimate, which occurs when the tree becomes over-fitted and the error is measured in a machine learning model. So to reduce this over-fitting the tree leaves are pruned to reduce the complexity. In comparison to other tree baseline models, the results of Improved Paloboost will defend against over-fitting and also provide solid and robust performance where the accuracy of the model will be high compared to other tree baseline models.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 5*
*DOI: https://doi.org/10.17762/ijritcc.v10i5.5552*
***Article Received**: 20 March 2022 **Revised**: 04 April 2022 **Accepted**: 25 April 2022 **Publication**: 31 May 2022*

**A. Data Preprocessing**

The first step is to preprocess the healthcare data to examine and omit missing values, noisy labels, and outliers. Healthcare data contains many noisy labels and missing values which are removed using a method called Data Preprocessing. Data Preprocessing is a technique where missing values or noisy labels are removed or replaced with the normalization technique like mean or median values which takes the average value of columns. In the case of multi-class labels,[14] is converted to binary class labels by replacing the values or by removing the values. Multi-class labels are replaced with 0's and 1's. For example, if a class label contains repeated values like 1,2,3 in a data, comparatively class label value "3" contains lesser information than the other two, that class label is removed or replaced. After the data is preprocessed, it proceeds to the next analysis.

**B. Important Feature using Genetic Algorithm**

A genetic algorithm is an optimization technique based on natural selection. It uses parameters like Initial Population, Fitness Function, Crossover, and Mutation.

The initial Population is the number of generations chosen randomly. The fitness function is the function that is used to find the maximum value among the overall values. Off-springs are produced by the Crossover technique. Exchanging the off-springs by flipping any one of the bits among individual strings using the Mutation technique. As a Genetic algorithm uses all these parameters for evaluation, the selection of features will be more accurate for given input features of data. Some of the columns from the input dataset are extracted based on their scores where scores are the method of ranking driver packages based on the features that they support. Scores with the highest values consider an important feature. The main purpose of extracting important features is to reduce the dimensionality of the model and also to improve the performance of the model. Finally, the model will give an accurate performance for the selected important features using a Genetic algorithm. The result will be plotted as a bar graph.

**C. Classification Techniques**

Many healthcare data contain missing values and noisy labels which can be tackled using tree boosting algorithms. Boosting combines multiple weak learners into strong learners and trees proceed in a stage-wise manner so that the over-fitting of data will be reduced. To tackle and produce a better performance of data, machine learning contains types of learning like supervised and unsupervised learning algorithms are used. The supervised learning algorithm contains a classification technique and the unsupervised learning algorithm contains a regression technique. The proposed system works with classification tasks. Classification has class labels that contain multi-class and binary class labels. Class labels are used to predict the root node of the tree. Many tree-based boosting algorithms are used. Some of the most commonly used algorithms are "Decision Tree"," Random Forest"," Gradient Boosting Machine"," XGBoost" and "Improved Paloboost.

**D. Improved Paloboost**

The main function of the Improved Improved Improved Paloboost algorithm is to build a strong model in the step by step process by learning the base model. As many healthcare data contain missing values. To handle these missing values "Improved Improved Improved Paloboost algorithm" [6] was proposed. Normally data contain hyper-parameter which should be tuned while using. Therefore grid search are used which calculates the performance for each combination of all the supplied hyper-parameters and their values, and then chooses the optimum value for the hyper-parameters. Normally Improved Paloboost uses samples that are not included in the overall samples so that Out-Of-Bag error will be calculated for the samples which are not included. The dataset is divided into training, testing, and validation. Here we have used training and testing pairs to train our model. Thus the model can be trained multiple times by changing the learning rate value and increasing the tree depth. Sometimes tuned hyper-parameters may lead to over-fitting in data. So to overcome this, Improved Paloboost which is not sensitive to hyper-parameter provides robust performance of the trained model. Improved Paloboost uses Out-Of-Bag Samples to tune the hyper-parameter. Thus these samples are used to estimate the learning rate and depth of the tree model at every stage. Thus trees will be trained or data are fitted in the model, the same out-of-bag samples can be used as they do not provide much impact, and also learning rate and size of the trees are specific parameters at each stage. Additionally, another two optimizations are used called "updated learning rate" and "Pruned Gradient", where the updated learning rate is used to minimize the objective function of the gradient descent method and the pruned gradient is used for classification and Regression to filter the pruning which dynamically adjusts the tree depth and save the computation resource and research time. Both methods guard against over-fitting and give better-predicted accuracy of the model. As the proposed system gives considerable accuracy, introducing a new optimization technique called Genetic Algorithm which gives more accuracy compared to the accuracy of the existing technique. This produces a better predictive and robust model.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 5*
*DOI: https://doi.org/10.17762/ijritcc.v10i5.5552*
***Article Received**: 20 March 2022 **Revised**: 04 April 2022 **Accepted**: 25 April 2022 **Publication**: 31 May 2022*

**(i) Pruned Gradient**

The pruned gradient is used to reduce the error and Out-Of-Bag samples are compared between children and parent nodes. It utilizes the Out-Of-Bag samples to achieve flexible trees. The pruned gradient does not generalize well with other samples as it removes the gradient estimates. The more stable gradient estimates can be achieved with the merging of regions with high variance. Therefore in Improved Paloboost, after the trees are fitted to the gradient, the Out-Of-Bag samples are applied to the tree which is not included in the In-Bag samples. A new leaf is generated by associating with the loss function and the gradient is multiplied by the learning rate. Pruned Gradient is used as it merges the child nodes to the root node having a loss function less than the loss function multiplied by the learning rate.

Derivative(gradient) of the loss function(error) of each parameter is taken.Residual error, Step size, and the learning rate are calculated based on the loss function.

**(ii) Adaptive Learning Rate**

The learning rate controls the error at an early stage. But setting up of learning rate leads to a tedious process so learning rates are updated at each stage. Improved Paloboost can assign different learning rates at every stage to build the tree model. Adaptive learning rate uses two functions Gaussian and Bernoulli for classification as well regression. Moreover, the main function of the adaptive learning rate is that it predicts the class labels correctly based on the trained model, and the whole tree is built.

## V.Results and Discussion

The data was collected from publicly available sources.

| Source | Data set | Techniques | Missing Values | Features |
|---|---|---|---|---|
| Physionet 2012 | Length of stay | Classification | Yes | 53 |
| | Diabetes | Classification | No | 9 |

"Physionet 2012" of dataset "Length of the stay in ICU" for analysis and all the steps included in the proposed design is executed. They are listed below with the graph of it in Fig. 1.

Any health care datasets can be given as an input to handle the data with the noisy labels as well as to tackle the over-fitting in data. The proposed system was implemented in the Anaconda platform with Jupyter notebook under Windows OS. Among the overall data with 3941 rows and 52 columns for the Length of stay dataset, a total of 70% was taken as a training set and 30% was taken as a testing set. The evaluation measures like accuracy, precision, and recall were calculated based on the selection of important features from the whole dataset.

**Length_of_stay** It is a publicly available source and easily accessible for iterations. To reduce the missing values present in the data, select the most common measurements: Urine_tot, GCS_tot, GCS_mean, Temp_tot, Temp_mean, NiDiasABP_tot, etc., using a Genetic Algorithm that gives higher accuracy. For each measurement to get a total of 35 features. The total missing value in Length of stay is 210.3%. The main objective of the task is to predict the total length of stay of ICU patients in hospitals. Accurate predictions based on class label ICU available can lead to the performance of the model.

Figure 1 shows different algorithm model that represents the accuracy prediction on the dataset. The row shows the different tree baseline models(Improved Paloboost, XGBoost, Gradient Boosting Machine). The column represents the range of accuracy. Higher the accuracy values, the most accurate the predictions are. Improved Paloboost will achieve better accuracy. Other tree baseline models(GBM, XGBoost) over-fit to the training data and hardly produce better predictions. Moreover, Improved Paloboost achieves robust results across the range of hyperparameters. Thus proposed system provides better and robust performance and tackles over-fitting of data in health care datasets.

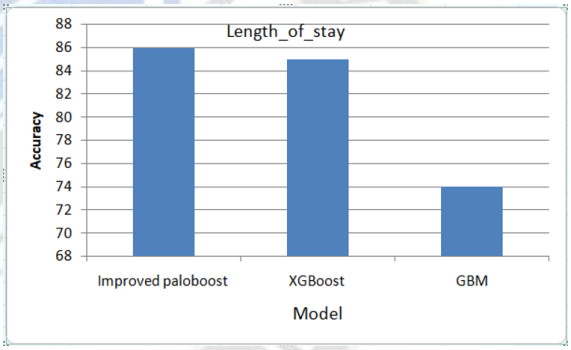



Fig. 1: Accuracy measure for length_of_stay

**Diabetes** Data is collected from an online source[16]. The objective is to predict whether the patient has diabetes or not based on the certain diagnostic measurements included in the dataset like Glucose, BMI, Age, Blood pressure, Skin thickness, Pregnancies, and Insulin. These features are selected using a Genetic Algorithm. Figure 2 shows the accuracy prediction for different algorithm models.The row shows the different tree baseline models(Improved Paloboost, XGBoost, GBM ). The column represents the range of accuracy. Accuracy will be high for more predicted values in Improved Paloboost. Other tree baseline models(GBM, XGBoost) over-fit so much to the training data and hardly produce better predictions. Moreover, Improved Paloboost achieves robust results across the range of hyperparameters.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 5*
*DOI: https://doi.org/10.17762/ijritcc.v10i5.5552*
***Article Received**: 20 March 2022 **Revised**: 04 April 2022 **Accepted**: 25 April 2022 **Publication**: 31 May 2022*

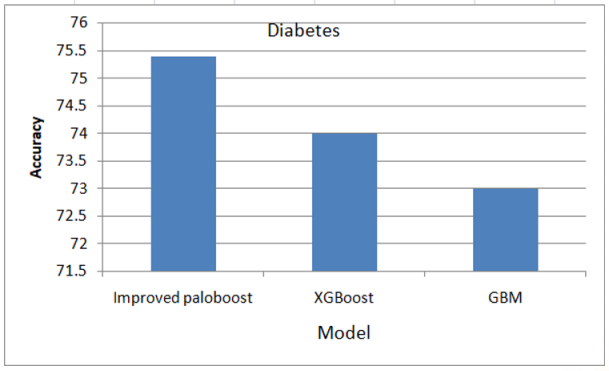



Fig. 2 : Accuracy measure for diabetes

**F1 Measure**

This measure is used for testing the accuracy. Here the proposed system is used to check whether the data can be added to the training set. The data included in the training set should be accurate to perform analytics and therefore the accuracy of the model can be calculated using the important machine learning evaluation metrics.

Normally F1 measure is calculated using Precision and recall. Where precision is several correct predictions made by the model and recall is the percentage of certain classes correctly identified. It is denoted using p1 and p2, where x1 and x2 denote the number of labels for each dataset. Let the common labels between the two be denoted as (L).

Let p1 = L / x1 and p2 = L / x2 ,

F1 measure is calculated by,

$$F1 = \frac{1}{N}\sum_{i=0}^{N}\frac{2p1.p2}{p1+p2} \qquad \text{----(1)}$$

Where N represents the number of data.

If the value of F1 is almost equal to 1 then the data can be accepted, else it can't be accepted to be added to the training set.

**Evaluation Measures**

The performance of the proposed system is measured using the three common evaluation measures in machine learning accuracy, precision, and recall. The overall prediction is based on a confusion matrix,

| | Actual : Yes | Actual : No |
|---|---|---|
| Predicted :Yes | True positive(TP) | False positive(FP) |
| Predicted : No | False negative(FN) | True negative(TN) |

In a tree-based algorithm, each data in different datasets were evaluated using evaluating measures like accuracy, precision, and recall can be calculated. The overall accuracy of the model is predicted for the data which are correctly classified.

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn} \qquad \text{---(2)}$$

$$precision = \frac{tp}{tp+fp} \qquad \text{---(3)}$$

$$recall = \frac{tp}{tp+fn} \qquad \text{---(4)}$$

**Feature importance**

Important features in Improved Paloboost is calculated based on the score where scores are the method of ranking driver packages based on the features that they support.To improve the model, feature importance is used and model dimensionality is reduced. The important features for different health care datasets are shown below :

| Data set | | Features | Feature Selection using GA | Accuracy |
|---|---|---|---|---|
| Length_of_stay | | 53 | 35 | 85% |
| Diabetes | | 9 | 9 | 83% |

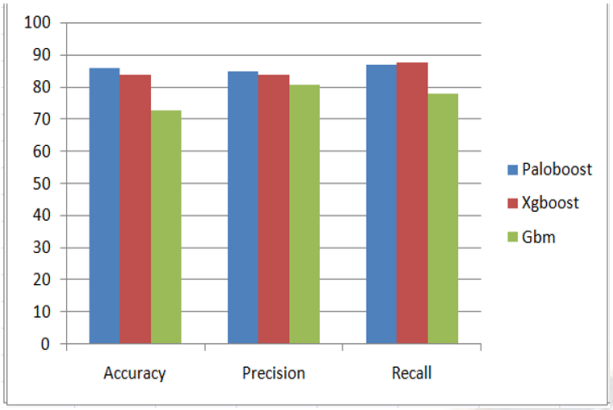
The results show that features that have the highest importance are captured based on their scores. This prediction of features is helpful in further prediction of a model using classification techniques as well as tree-based algorithms. Therefore this proposed system will give robust performance and handle over-fitting of data.

Evaluation measures like accuracy, precision, and recall are calculated for the proposed system upon all three baseline algorithms.

| **Algorithm** | **Accuracy** | **Precision** | **Recall** |
|---|---|---|---|
| GBM | 0.73 | 0.81 | 0.78 |
| XGBoost | 0.84 | 0.84 | 0.88 |
| Improved Paloboost | 0.86 | 0.85 | 0.87 |

The overall accuracy, precision, recall value calculated using classification techniques using tree-based algorithms like Improved Paloboost, XGBoost, and Gradient Boosting Machine is shown below,

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 5*
*DOI: https://doi.org/10.17762/ijritcc.v10i5.5552*
***Article Received**: 20 March 2022 **Revised**: 04 April 2022 **Accepted**: 25 April 2022 **Publication**: 31 May 2022*




The results depict that among all tree-based boosting algorithms, Improved Paloboost performs better than XGBoost and Gradient Boosting Machine tree-based algorithm while selecting the features using the Genetic Algorithm.

## VII.Conclusion

As normally health care data are derived from multiple sources, many data leads to over-fit the model, and also it may contain some missing values, noisy labels, and a large number of features. The proposed system depicts the overall workflow of how data are processed and accuracy is measured. The proposed algorithm tends to tackle the over-fitting of data as it uses two optimization techniques like "Pruned Gradient" and "Adaptive Learning Rate" which increases accuracy by reducing errors and finding optimal learning rates which are used to train and build the model. Classification techniques are used to build the model. Based on comparison Improved Paloboost algorithm performs better than other techniques.

## VI. References

[1]. H. Lee and H.-J. Yoon, "Medical big data: Promise and challenges," Kidney Res. Clin. Pract., vol. 36, no. 1, pp. 3–11, 2017.
[2]. K. E. Goodman et al., "A clinical decision tree to predict whether a bacteremic patient is infected with an extended-spectrum b-lactamase–producing organism," Clin. Infectious Diseases, vol. 63, no. 7, pp. 896–903, 2016.
[3]. J. H. Friedman, "Stochastic gradient boosting," Comput. Statist. Data Anal., vol. 38, no. 4, pp. 367–378, Feb. 2002.
[4]. L. Breiman, Classification and Regression Trees. Abingdon, U.K.: Routledge, 2017.
[5]. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Ann. Statist., vol. 29, no. 5, pp. 1189–1232, Oct. 2001
[6]. Y. Park, "Bonsai-dt - programmable decision tree framework," [Online]. Available: https://yubin-park.github.io/bonsai-dt/
[7]. W. Jiang, "On weak base hypotheses and their implications for boosting regression and classification," The Ann. Statist., vol. 30, no. 1, pp. 51–73, Nov. 2002.
[8]. Healthcare data https://github.com/yubin-park/bonsai-dt/tree/master/research.
[9]. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
[10]. H. Schwenk and Y. Bengio, "Boosting neural networks," Neural Comput., vol. 12, no. 8, pp. 1869–1887, Aug. 2000.
[11]. Y. Park and J. C. Ho, "Improved Paloboost: An over-fitting-robust TreeBoost with out-of-bag sample regularization techniques," 2018. [Online]. Available: https://arxiv.org/abs/1807.08383
[12]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 78579
[13]. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
[14]. R. Bekkerman, "The present and the future of the KDD cup competition," 2015. [Online]. Available: https://www.kdnuggets.com/ 2015/08/kdd-cup-present-future.html
[15]. G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Proc. 31st Int. Conf. Neural Inf. Process. System 2017, pp. 3146–3154.
[16]. Healthcare https://www.kaggle.com/search/Heart and Diabetes.
[17]. Genetic Algorithm: Reviews, Implementations, and Applications,Tanweer Alam,2020.
[18]. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Berlin, Germany: Springer, 2009.
[19]. K. E. Goodman et al., "A clinical decision tree to predict whether a bacteremic patient is infected with an extended-spectrum b-lactamase–producing organism," Clin. Infectious Diseases, vol. 63, no. 7, pp. 896–903, 2016.
[20]. J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," J. Animal Ecology, vol. 77, no. 4, pp. 802–813, Jul. 2008.
[21]. Application of Data Mining Techniques to Healthcare Databy Cambridge University Press: **02 January 2015**