Original Article

BREAST CANCER DATA FEATURE SELECTION USING ENSEMBLE LIGHT GRADIENT BOOSTING TECHNIQUE

Mr. S. Vibinchandar[1], Dr V. Krishnapriya[2]

[1] Research Scholar, Department of Computer Science, Sri Ramakrishna College of arts and Science, Tamilnadu, India.
[2] Associate Professor and Head, Department of Computer Science with Cognitive Systems, Sri Ramakrishna College of Arts and Science, Tamilnadu, India.

DOI: 10.47750/pnr.2022.13.S08.398

Abstract

Breast cancer is one of the highly dangerous diseases among the females around the world. The efficient and correct detection of BC is big medical issue and many researchers proposed different diagnostic methods for detection of this disease, however these existing methods still needed further improvement to correct and efficient detection of this disease. In this paper, proposed a Ensemble Light Gradient Boosting Technique (LightGBM) new BC identification method by using machine learning algorithm and clinical data. Feature Selection Algorithm is one more important step in the machine learning classification process, as most of the time, there are many features in the dataset which are irrelevant or have the least correlation with the output classes for example serial or ID number in any dataset. Such features affect the performance of the machine learning classifiers. LightGBM Feature selection improves classification accuracy and reduces model computation time.

Keyword: Breast cancer, Machine learning, Feature Selection, Light, Gradient.

Introduction

Breast Cancer (BC) is a hazardous disease and is endured by numerous ladies across the world. In 2021 there were 2 and half-million new cases revealed. The fifth integral justification behind female demise is BC contrasted with cancers concerning different types. The dangerous growth of BC which delivered inside breast cells. A gathering of dividing cells that structure a knot or mass of additional tissue is called Tumors and these growths can be cancerous (dangerous) or non-cancerous (harmless). In various nations with the high level created clinical innovation aggregate the 5-year endurance pace of first stages BC is (80-90%), and diminishing up to 24% for recognizable proof of BC at the primary stages.

Learning from exceptionally enormous databases is really difficult for a large number of the present-day data mining and machine learning algorithms. This issue is for the most part refers to utilizing the expression "enormous data," which shows the difficulties and disadvantages of completing the processing and examination of intricate pieces of data. It has attracted a lot of revenue countless regions like bioinformatics, medication, showcasing, and monetary organizations, because of the broad arrangements of crude data, which are put away. Feature Selection (FS) procedures can be used for diminishing the dimensionality, before the use of any data mining strategies like classification, affiliation rules, clustering and relapse.

Feature selection (FS) is an essential advance in the machine learning process and because of proper feature selection the machine learning (ML) model execution increments and the computational season of the model lessening. The feature selection process has incredible ramifications for the classification aftereffects of the model. The selection of reasonable feature selection algorithms is a muddled process for the selection of additional suitable features from the dataset. The objective of feature selection is to take out features which give practically no extra information past that given by the leftover features. Feature selection is a troublesome assignment since it isn't difficult to predict which features or mix of features can bring about accomplishing great classifier accuracy. In addition, a feature which seems, by all accounts, to be pointless or insignificant without help from anyone else can deliver great prediction results when joined with other insignificant features.

Importance of feature selection

Feature selection has been demonstrated to be successful and productive in taking care of high dimensional data. Here are a portion of the justifications for why feature selection is critical:

Enhanced generalization by reducing over fitting: abundance variables in the data can add commotion to a model, which prompts overfitting. By disposing of the noisy features in the data, we can considerably further develop the speculation capacity of a machine learning model. **Reduces training times:** diminishing the quantity of variables to fabricate a machine learning model will decrease the computational expense and in this way accelerate model structure. Increment model interpretability: we some of the time lose explainability in a machine learning model when we have many features. By decreasing the quantity of features, the model becomes more straightforward and simpler to. A model with 50 features has preferable explainability over a model with 200 features. **Variable redundancy:** features inside data are frequently profoundly related, making them repetitive. By eliminating these associated features, the model will be less inclined to make commotion based predictions. **Reduces prediction time:** decreasing the quantity of features lessens the calculation cost - less complex models will more often than not have quicker prediction times.

Literature survey

Amin Ul Haq, Jian Ping Li, Abdus Saboor 1, Jalaluddin Khan1, Samad Wali, Sultan Ahmad, Amjad Ali, Ghufran Ahmad Khan, And Wang Zhou (2020) proposed supervised (Relief) algorithm and unsupervised (Auto encoder and PCA) algorithms have been utilized for related feature selection from the data set and afterward these chose features have been utilized for the training and testing of the classifier SVM for exact and on-time location of BC. Relief (RF) is an administered learning feature selection calculation which involves a filter system for feature selection from the data set. The hypothetical and mathematical knowledge of the RF calculation has been introduced for a superior comprehension of the calculation. Relief is a practically distance-based filter FS calculation which positions features that separate classes in light of how to make coordinated features that can isolate classes. The proposed technique has achieved high outcomes as far as accuracy on chose features chose by the Relief calculation. **A. Sanjay, H. V. Nair, S. Murali and K. S. Krishnaveni (2018)** proposed Medical data mining models are transcendent in medical data investigation. Applied the Naive Bayes (NB) classifier calculation alongside another better feature selection strategy towards an exact identification of breast cancer in its beginning phases and exhibition of the NB classifier supposedly was discernibly improved when unimportant features were screened out from the modeling process. Observational confirmation reaffirms that our half and half feature-selection approach, utilizing an insignificant arrangement of properties, outflanks the outcomes acquired from singular feature selection methods. Proposed another half and half feature selection strategy for prior prediction of breast cancer. Their data mining model purposes a cross breed feature selection technique that is profoundly productive, exact and outperforms the outcomes acquired from individual feature selection methods.

H. Hasan and N. M. Tahir, et.al proposed Feature selection of breast cancer in view of Principal Component Analysis. Machine learning is a piece of electronic thinking that uses a variety of authentic, probabilistic and streamlining techniques and grants PCs to acquire from past models and identify designs from enormous data sets, which is especially appropriate to help medical specialists in the finding of disease in light of an assortment of experimental outcomes. Consequently, in this exploration, we considered further by fostering a feature extraction calculation in light of Principal Component Analysis (PCA) and Artificial Neural Network (ANNs) as classifiers as the ideal device to upgrade the classification of harmless or threatening in view of the Wisconsin Breast Cancer Database. the capacity of PCA alongside three principles of thumbs as feature selection is assessed and affirmed effective in view of the accuracy pace of ANN for conclusion of breast cancer cells. Introductory outcome got is promising to exhibit that this strategy can segregate among typical and breast cancer patients. **A. U. Haq et al** proposed Detection of Breast Cancer through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques. Breast cancer is quite possibly the most basic disease and is endured by many individuals all over the planet. The proficient and right location of breast cancer is as yet expected to guarantee this medical issue in spite of the fact that analysts all over the planet are proposed different symptomatic methods for recognition of this disease; in any case, these current methods actually required further improvement to right and productive discovery of this disease. In this review, we proposed another breast cancer recognizable proof strategy by utilizing machine learning algorithms and clinical data. In the proposed strategy managed (Relief calculation) and unaided (Autoencoder, PCA algorithms) procedures have been utilized for related feature selection from the data set and afterward these chose features have been utilized for training and testing of classifier support vector machine for precise and on-time discovery of breast cancer.

G. Li et.al proposed Effective Breast Cancer Recognition Based on Fine-Grained Feature Selection. Early location and analysis of breast cancer are pivotal to further developing the endurance paces of patients. Henceforth, pathologists and radiologists need a PC supported analysis framework to help their clinical findings really and productively. To lighten the example shortage issue, a straightforward, compelling model called "refinement, correlation, adaptive" (RCA) for breast cancer recognition is proposed according to the perspective of fine-grained feature selection.

Proposed Methodology

Data Set

The dataset utilized in this paper is Wisconsin Breast Cancer (Diagnostic) data (WDBC) from the standard UCI Machine learning vault. The dataset has 569 examples and 32 qualities. The dataset comprises of two classes, in particular Begnin and Malignant, 357 and 212 examples. They are Mean, Standard Error and most awful mean. They are characterized as:

$$Mean = \frac{\sum f(x)}{f}, Standard\ Error = \frac{Standard\ deviation}{Square\ root\ (N)}$$

The following table shows the attributes of WDBC along with their description.

Table 1. Attributes of WDBC

S. No	Name of the attribute	Description of the attribute
1	Radius	Mean of the distances from the center to the perimeter
2	Texture	Standard deviation of gray scale value
3	Perimeter	Total distance between continues snake points
4	Area	Measure of the number of each pixels inside the snake points and add half of the pixels to the parameter
5	Smoothness	Variation in radius lengths

6	Compactness	Perimeter^2/area-1.0
7	Concavity	Severity of the portions of the contour
8	Concave_points	Number of concave points of the contour
9	Symmetry	The longest harmony through the middle ought to be found and the lines which are perpendicular to the significant axis limits to the atomic limit toward each path are estimated
10	Fractal_dimension	“costline approximation”-1

This paper performs feature selection of WDBC by gradient boosting technique that is LightGBM. The general accuracy of the learning model can be expanded by taking care of the model with enhanced features. At first, the classifiers were run autonomously and the best 10 significant features from every calculation were noted. Among each of the three feature subsets, this research select 7 features that are normal in the three feature subsets.

Gradient boosting

Outfit learning joins the predictions of numerous algorithms to create a superior prediction esteem. The fundamental explanation ensembling was utilized is that the quantity of models attempting to predict a similar objective variable as a solitary predictor gives a superior prediction. Stowing and Boosting are two group methods. In Bagging, we make free learners and join them utilizing model normal strategies (model: casting a ballot), while in helping predictors were developed consecutively.

Slope boosting calculation is a kind of ensembling strategy in which classification and prediction are made by joining more fragile models. This procedure has three components: 1) the misfortune function, which should be enhanced. The misfortune function is a measurement for working out how great the model coefficients are at fitting the data. It has no definition and relies upon what the programmer needs to advance. 2) A Weak learner, for performing classification or prediction (choice trees were utilized as a feeble learner here), 3) additive model, which adds each tree in turn. After the calculation of the misfortune function, the new tree gets added to the current model to play out an angle plunge technique to limit the misfortune function. Asgradient supporting models work exceptionally on optimizing the misfortune, greatest accuracy will be achieved.

Gradient Boosting Algorithm Procedure

Step i: Calculate the loss function and minimize it. Here loss function is the mean squared error of target and predicted values.

$$Loss = \sum (x_i - x_i^p)^2$$

Whereas, x_i *is the* i^{th} Target value and $x_i p$ *is the* i^{th} prediction value.

Step ii: By applying gradient descent and Update the predictions

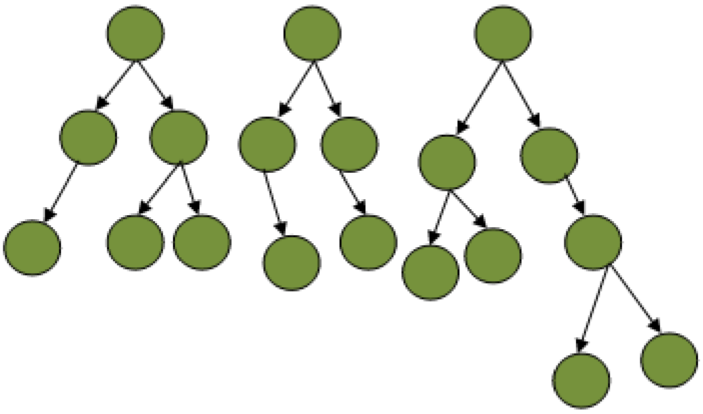
$$x_i^p = x_i^p + \alpha \times \delta \sum \frac{(x_i - x_i^p)^2}{\delta x_i^p}$$

Whereas α is the learning rate, $\sum (x_i - x_i^p)^2$ is the loss function. Consequently by refreshing the predictions, the misfortune function is limited and predicted values were around near close to values.

Light Gradient Boosting Model

This is an extremely quick model and henceforth it is named the Light Gradient Boosting Mode. LIGHTGBM is one of the angle helping structures that utilize tree-based learning algorithms. It is utilized for classification and feature positioning and can perform many machine learning assignments. This model depends on a choice tree calculation and parts the tree utilizing a leaf wise methodology.

Figure 5.Leaf-wise growth of the Tree



This generally brings about higher accuracy than other boosting algorithms. Like XGBoost, the significance of each feature can be acquired from the feature_importances_ quality embedded in the calculation. LIGHTGBM is profoundly proficient, gives better accuracy, can deal with data on an extremely enormous scope and utilizations lower memory. The classifiers LIGHTGBM is accessible as bundles in sklearn in python. LIGHTGBM classifier was applied to the given dataset to create the significant features of the dataset. Given a training dataset$X = \{(x_i, y_i)\}_{i=1}^{n}$, the goal of LightGBM is to find an approximation $\hat{f}(x)$ to a certain function$f^*(x)$ that minimizes the expected value of particular loss function $L(y, f(x))$ as follows:

$$\hat{f} = \arg min E_{y,,x} L(y, f(x))$$

LightGBM integrates a number of T regression trees $\sum_{t=1}^{T} f_t(X)$ for approximating the final model, which is $f_T(X) = \sum_{t=1}^{T} f_t(x)$

Ensemble LightGBM algorithm

Input:

Training data: $T = \{(X_\iota, y_\iota), (X_2, y_2) \ldots, (X_n, y_n)\}, X_i \; E_{x,x} \in R, y; €\{-1, +1\}; \; loss\ function: L(y, 0\ (x));$

Iterations: M;

Ratio of sampling the Big gradient data: f; slight gradient data sampling ratio: z;

1: Combine features that are mutually exclusive (i.e., features not simultaneously accept nonzero numbers) of $x_i, i = \{1 \ldots, N\}$ *by the exclusive feature bundling (EFB) technique;*

2: Set $\theta_o(x) = arg \min_c \sum L(yi, c);$

3: For m= 1 to M do

4: *Calculate gradient absolute values:*

$$r^i = \left| \frac{\partial L\left(y_{i,}\theta(x_i)\right)}{\partial \theta(x_i)} \right|_{\theta(x)=\theta_{m-1(x)}} \quad , i = \{1, \dots, N\}$$

5: *Resample dataset using gradient-based one-side sampling process:* $highN = f \times len(T);\ randN = z \times len(T);$

$sorted = GetSortedIndices(abs(r));$

$F = sorted[1{:}HighN];\ Z = Random\ Pick(sorted[HighN{:}len(T)], randN);$

$$\acute{T} = F + Z;$$

6: *Calculate IG:*

$$V_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i\in F_l} r_i + \frac{1-a}{b}\sum_{x_i\in Z_l} r_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i\in F_r} r_i + \frac{1-a}{b}\sum_{x_i\in Z_r} r_i\right)}{n_l^j(d)}\right)$$

7: *Develop a novel decision tree* $\theta_m(x)'$ *on set* T'

8: *Update* $\theta_m(x) = \theta_{m-1}(x) + \theta_m(x)$

9: *End of for*

10: *Return* $\tilde{\theta}(x) = \theta_m(x)$

As a result, we considered the top 10 features that are listed as follows:

TOP 10 FEATURES	
1.	Texture_mean
2.	Concavity_worst
3.	Texture_worst
4.	Area_worst
5.	Smoothness_mean
6.	Concavepoints_worst
7.	Symmetry_se
8.	Perimeter_worst
9.	Smoothness_worst
10.	Concavepoints_mean

Experiment Result

Precision

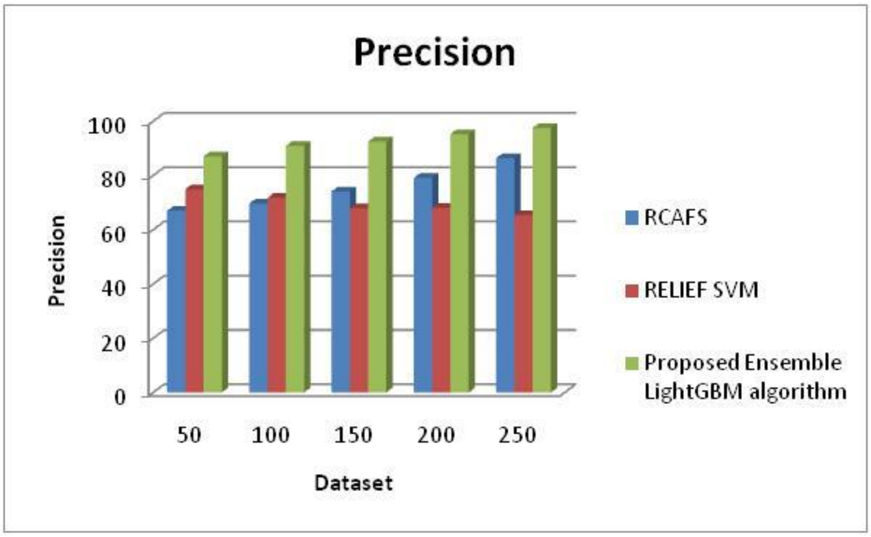
$$Precision = TruePositives\ /\ (TruePositives + FalsePositives)$$

Table 2.Comparison tale of Precision

Dataset	RCAFS	RELIEF SVM	Proposed Ensemble LightGBM algorithm
50	66.94	74.91	87.01
100	69.66	71.77	90.87
150	74.12	67.93	92.48
200	79.09	68.05	95.23
250	86.38	65.39	97.52

The Comparison table 2 of Precision Values explains the different values of existing RCAFS, RELIEF SVM and proposed Ensemble LightGBM algorithm. While comparing the Existing algorithm and proposed Ensemble LightGBM algorithm, provides the better results. The existing algorithm values start from 66.94 to 86.38, 65.39 to 74.91 and proposed Ensemble LightGBM algorithm values starts from 87.01 to 97.52. The proposed method provides the great results.

Figure 6.Comparison chart of Precision



The Figure 6 Shows the comparison chart of Precision demonstrates the existing RCAFS, RELIEF SVM and proposed Ensemble LightGBM algorithm. X axis denote the Dataset and y axis denotes the Precision ratio. The proposed Ensemble LightGBM algorithm values are better than the existing algorithm. The existing algorithm values start from 66.94 to 86.38, 65.39 to 74.91 and proposed Ensemble LightGBM algorithm values starts from 87.01 to 97.52. The proposed method provides the great results.

Recall

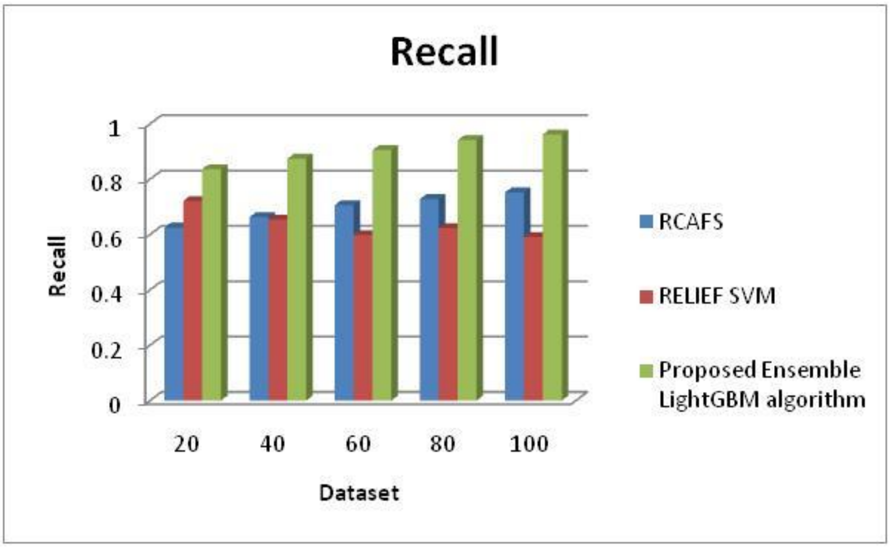
$$Recall = TruePositives / (TruePositives + FalseNegatives)$$

Table 3.Comparison tale of Recall

Dataset	RCAFS	RELIEF SVM	Proposed Ensemble LightGBM algorithm
20	0.625	0.721	0.836
40	0.663	0.654	0.874
60	0.706	0.598	0.905
80	0.728	0.623	0.941
100	0.752	0.591	0.962

The Comparison table 3 of Recall Values explains the different values of existing RCAFS, RELIEF SVM and proposed Ensemble LightGBM algorithm. While comparing the Existing algorithm and proposed Ensemble LightGBM algorithm, provides the better results. The existing algorithm values start from 0.625 to 0.752, 0.591 to 0.721 and proposed Ensemble LightGBM algorithm values starts from 0.836 to 0.962. The proposed method provides the great results.

Figure 7.Comparison chart of Recall



The Figure 7 Shows the comparison chart of Recall demonstrates the existing RCAFS, RELIEF SVM and proposed Ensemble LightGBM algorithm. X axis denote the Dataset and y axis denotes the Recall ratio. The proposed Ensemble LightGBM algorithm values are better than the existing algorithm. The existing algorithm values start from 0.625 to 0.752, 0.591 to 0.721 and proposed Ensemble LightGBM algorithm values starts from 0.836 to 0.962. The proposed method provides the great results.

F -Measure

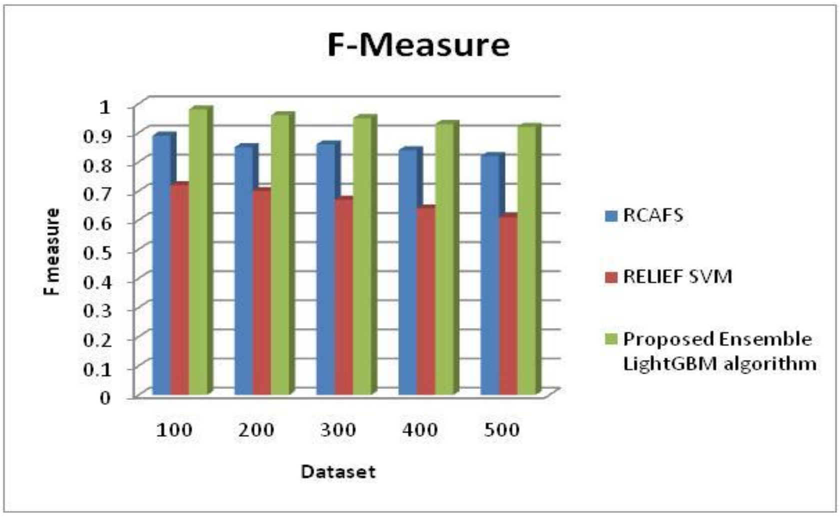
$$F - Measure = (2 * Precision * Recall) / (Precision + Recall)$$

Table 4.Comparison tale of F -Measure

Dataset	RCAFS	RELIEF SVM	Proposed Ensemble LightGBM algorithm
100	0.89	0.72	0.98
200	0.85	0.70	0.96
300	0.86	0.67	0.95
400	0.84	0.64	0.93
500	0.82	0.61	0.92

The Comparison table 4 of F -Measure Values explains the different values of existing RCAFS, RELIEF SVM and proposed Ensemble LightGBM algorithm. While comparing the Existing algorithm and proposed Ensemble LightGBM algorithm, provides the better results. The existing algorithm values start from 0.82 to 0.89, 0.61 to 0.72 and proposed Ensemble LightGBM algorithm values starts from 0.92to 0.98. The proposed method provides the great results.

Figure 8.Comparison chart of F -Measure



The Figure 8 Shows the comparison chart of F -Measure demonstrates the existing RCAFS, RELIEF SVM and proposed Ensemble LightGBM algorithm. X axis denote the Dataset and y axis denotes the F -Measure ratio. The proposed Ensemble LightGBM algorithm values are better than the existing algorithm. The existing algorithm values start from 0.82 to 0.89, 0.61 to 0.72 and proposed Ensemble LightGBM algorithm values starts from 0.92to 0.98. The proposed method provides the great results.

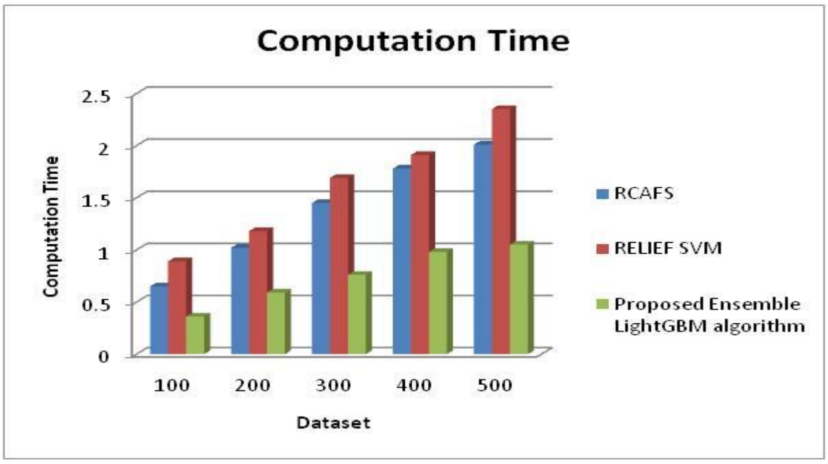
Computation Time

Table 5.Comparison tale of Computation Time

Dataset	RCAFS	RELIEF SVM	Proposed Ensemble LightGBM algorithm
100	0.65	0.89	0.36
200	1.02	1.18	0.59
300	1.45	1.69	0.76
400	1.78	1.91	0.98
500	2.01	2.35	1.05

The Comparison table 5 of Computation Time Values explains the different values of existing RCAFS, RELIEF SVM and proposed Ensemble LightGBM algorithm. While comparing the Existing algorithm and proposed Ensemble LightGBM algorithm, provides the better results. The existing algorithm values start from 0.82 to 0.89, 0.61 to 0.72 and proposed Ensemble LightGBM algorithm values starts from 0.92to 0.98. The proposed method provides the great results.

Figure 9.Comparison chart of F -Measure



The Figure 9 Shows the comparison chart of Computation Time demonstrates the existing RCAFS, RELIEF SVM and proposed Ensemble LightGBM algorithm. X axis denote the Dataset and y axis denotes the Computation Time. The proposed Ensemble LightGBM algorithm values are better than the existing algorithm. The existing algorithm values start from 0.65 to 2.01, 0.89 to 2.35 and proposed Ensemble LightGBM algorithm values starts from 0.36to 1.05. The proposed method provides the great results.

Conclusion

In this paper we proposed Ensemble Light Gradient Boosting Technique (Ensemble LightGBM) has extraordinary reference an incentive for planning underlying variety utilizing different kinds of essential learners, for example, brain organizations and backing vector machines or other ensemble learning calculations. The proposed Ensemble LightGBM method could be additionally applied in recognizing cancers of different kinds and give specialists direction for early diagnosis, which has numerous valuable clinical applications in clinical breast growth diagnosis. For patients who are with a breast cancer history, such a model can prompt a more quick mediation with the most proper therapy. Further, there is examination to be done toward pre-handling of big data calculations for preparing better delegate feature sets for better predictions.

References

1. Al-Ani, Ahmed, Mohamed Deriche, and Jalel Chebil. "A new mutual information based measure for feature selection." Intelligent Data Analysis 7.1 (2003): 43-57.

2. Alba, Enrique, et al. "Gene selection in cancer classification AnimeshHazra, Subrata Kumar Mandal, AmitGupta, "Study and Analysis of Breast Cancer

3. Bertoni, A.; Folgieri, R.; Valentini, G. Bio-molecular cancer prediction with random subspace ensembles of support vector machines. Neurocomputing **2005**, 63, 535–539.

4. Marvi-Khorasani, H.; Usefi, H. Feature Clustering Towards Gene Selection. In Proceedings of the 2019 18th IEEE International Conference on Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; pp. 1466–1469.

5. Cheadle, C.; Vawter, M.P.; Freed, W.J.; Becker, K.G. Analysis of Microarray Data Using Z Score Transformation. J. Mol. Diagn. **2003**, 5, 73–81.

6. BiplobDey, Arun Kumar, "A Review Article on Breast Cancer", *International Journal of* Breast Cancer Diagnosis", *International Journal of Computer Sciences and Engineering*,

7. Bry, X.; Trottier, C.; Verron, T.; Mortier, F. Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. J. Multivar. Anal. **2013**, 119, 47–60.

8. Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms*", International journal of* Center, M.M.; Jemal, A.; Smith, R.A.; Ward, E. Worldwide Variations in Colorectal Cancer. CA A Cancer J. Clin. **2009**, 59, 366–378.

9. Coates, A.; Ng, A.Y. Learning Feature Representations with K-Means. In Pattern Recognition. ICPR International Workshops and Challenges; Springer: New York, NY, USA, 2012; pp. 561–580.

10. Ferlay, J.; Ervik, M.; Lam, F.; Colombet, M.; Mery, L.; Piñeros, M.; Bray, F. Global Cancer Observatory: Cancer Today; International Agency for Research on Cancer: Lyon, France, 2018.

11. Ganesh N. Sharma, Rahul Dave, JyotsanaSanadya, Piush Sharma and K. K Sharma,"Various

12. García S, Luengo J, Herrera F. Data Preprocessing in Data *journal of Advanced Engineering and Research Development*, Vol.4, No.11, pp. 607-611,

13. *Journal of Cancer Research*,Vol.7, No.3,pp610-627,2017. Ajay Kumar, R. Sushil, A.K. Tiwari, "Comparative Study of Classification Techniques for

14. Lee, K.S.; Geem, Z.W. A new meta-heuristic algorithm for continuous engineering optimization: Harmony search theory and practice. Comput. Methods Appl. Mech. Eng. **2005**, 194, 3902–3933.

15. Machine Learning Techniques", *International Journal of Management, Technology And* Martinez, A. M. and Kak, A. C. Pca versus lda. IEEE

16. MedisettyHari Krishna, Dr. Kunjam. NageswaraRao, "Prediction of Breast Cancer Using Mining. Berlin: Springer; 2015.

17. Puneetyadav, rajatVarshney, Vishan Kumar Gupta,"Diagnosis of Breast Cancer using Decision.

18. Shin, S.-Y.; Lee, I.-H.; Kim, D.; Zhang, B.-T. Multiobjective Evolutionary Optimization of DNA Sequences for Reliable DNA Computing. IEEE Trans. Evol. Comput. **2005**, 9, 143–158.

19. T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys, *technology and research,* Vol. 1, No. 2, pp.109-126, 2010.

20. Transactions on Pattern Analysis and Machine Intelligence, Tree Models and SVM", *International Journal of Engineering and Technology*, Vol. 5, No.03.