RFSVM: A Novel Classification Technique for Breast Cancer Diagnosis

Badal Soni, Angshuman Bora, Arpita Ghosh, and Anji Reddy

Abstract: Cancer is a disease, which develops, in human body due to gene mutation. Due to various factor cells turn into cancerous cell and grow rapidly while damaging normal cells. Many women get affected by breast cancer, which might even cause death if not treated at early stage. Early detection of breast cancer is highly important to increase the survival rate. Machine learning methods and technologies are making it possible to classify and detect the class in an accurate manner. Among other classifiers, random forest and support vector machine are two classifiers that have a good classification power. In this, research a combination of these two classifier i.e. Random Forest and Support Vector Machine (RFSVM) is proposed for early diagnosis of breast cancer cell using Wisconsin Breast Cancer Dataset (WBCD). Using different train-test data ratio experiments are performed and an average of more than 98percentage accuracy is achieved using this hybrid classifier. This paper overcomes the over-fitting problem of random forest and the need of tuning the parameters of Support Vector Machine. Even with limited data available, the classifier tunes its parameters so well to give a highly accurate result.

***Keywords: Random Forest, Support Vector Machine, Breast Cancer, Classification, Feature Extraction*.**

I. INTRODUCTION

Cancer cell can develop in human body and keep growing if it is not identified at initial stage. Various factors are responsible for increasing number of cancer cell like pollution of environment, degrading food quality, smoking habitsetc [5]. In recent years, the growing number of patient of breast cancer is high and the survival rate is less. The researchers are developing various kinds of techniques for identification of breast cancer at early stage. In India, due to the less awareness of breast cancer and lack of medical facilities survival rate of breast cancer is very less compare to United State [15]. Demand of developing new technologies and improvement of the existing technology became important due to the growing rate of breast cancer death. Early detection of breast cancer can only increase the probability of complete treatment to a patient.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author

Badal Soni*, Computer Science and Engineering,National Institute of Technology Silchar, Assam, India. Email: badal@nits.ac.in

Angshuman Bora, Computer Science and Engineering,National Institute of Technology Silchar, Assam, India. Email: angshu.btf@gmail.com

Arpita Ghosh, Computer Science and Engineering,National Institute of Technology Silchar, Assam, India. Email:arpita.csphd@gmail.com

Anji Reddy,Computer Science and Engineering, Lendi Institute of Engineering and Technology, Andhra Pradesh, India. Email: anjisoftware@hotmail.com

Among various cancer cell detection and classification method accurate prediction and classification became an important and challenging task among researchers. The data generation in the field of bio medical is very fast. This data can provide very useful information for research in medical area. Various machine-learning algorithms are useful for retrieval of information based on the previous data and experience. ML makes it possible to extricate data and information from the premise of past encounters and identify difficult design from vast data set. Prediction and classification of cancer cell is possible by information extraction from the experience. In this research, various supervised learning techniques are compared with the proposed technique *i.e.* Random Forest-Support Vector Machine (RFSVM) based method, which is useful for increasing the prediction of the disease. In this research, comparative analysis has been done using various supervised machine-learning methods, which is useful for predicting the increasing rate of breast cancer. Machine learning application in the field of bio medical research is growing day by day because of the effectiveness of the process of prediction and classifying data. The machine learning approaches here compared based on Precision, Recall, F1-score and Accuracy. Previously researchers have done various analysis using different machine learning approaches as linear regression, Random Forest, Multilayer perceptron, Decision Tree. Transfer learning methodology, which is mainly based on the convolution neural network model, is also used to classify the histopathological images to benign and malignant class. Here in terms to increase the accuracy of classification, a new approach has been proposed based on Random Forest classifier and Support Vector Machine, which includes testing of various train-test data ratio and achieved the accuracy more than 98 percentages. Related works are highlighted in the Section 2, in Section 3 background studies have been discussed, the dataset collection and pre-processing is given in Section 4, proposed method is given in Section 5, in Section 6 experimental results and analysis are discussed, Section 7 is about the conclusion and future scope about the work.

II. LITERATURE SURVEY

Paper [16], given by Madhuri Gupta and Bharat Gupta they used machine learning algorithms to the Wisconsin dataset. The algorithms they used are linear regression, Random Forest, Multilayer perceptron and Decision Tree. They compared the result based on “10 fold cross validation method” and identified that multilayer perceptron is more accurate than other machine learning algorithms.

In paper [6], authors compared the results and analyzed it based on Ensemble Learning techniques i.e. Bagging, Boosting and Voting based Ensemble classifier on the Wisconsin Breast Cancer dataset. Comparison of these three ensemble techniques proved that if accuracy were the highest concern than bagging based Ensemble learning model would provide 95.6 percent accuracy, which is better than the other two methods. In paper [8], Jongwon Chang *et. al*. used Break -His dataset and transfer learning methodology which is mainly based on the convolution neural network model and used it to classify the histopathological images to benign and malignant class. The usage of data augmentation and transfer learning method achieved highest training accuracy, which is 0.89. In paper [17], authors used ensemble method and claimed that the method is better than a single classification method. Sequential least squares programming method is used for assigning weight and for combining the results soft voting process is applied. Based on the accuracy, F-score, R2 and 10-fold cross validation parameters the ultimate evaluation of the algorithms is done. Paper [20], applied various machine learning algorithms as Naïve Bayes, Random Forest, KNN, SVM and selected the most accurate algorithm among all this, which is SVM. The accuracy achieved by SVM is 97.9 percent. In the comparative study, it is shown, that SVM classifier correctly classifies 569 cases out of 699 cases where 357 cases are benign and 212 are malignant. Paper [2], applied random forest classification algorithm that is mainlya decision tree algorithm as a classifier to the FNA (Fine Needle Aspiration) biopsy dataset. Receiver operating characteristics or ROC is used here for evaluation process. Here the evaluation of random forest is done based on the sensitivity, accuracy and specificity percentage, which are 75, 72 and 70percent respectively. Paper [29], Zhiqiong Wang *et. al*. proposed a computer aided diagnosis method including the usage of convolution neural network. The identification of cancer cell here is divided into three steps. First step includes the unsupervised extreme learning machine clustering process (US-ELM) and computer aided diagnosis (CNN) for mass detection. Second part includes modeling of the feature set. Third step describes the classification of cancer cell where ELM classifier along with fused feature set is used to distinguish between the benign and malignant cells. In paper [14], authors do the comparative study with relevance vector machine (RVM) and other classification algorithms. In the paper, they proposed that the RVM method is better than the other classification algorithms as the computational cost of RVM is less. Paper [28], given by Laith R. Sultan MD *et. al.* proposed a method of machine learning along with ultrasound model which include Doppler and gray-scale effects for identification of breast cancer. In paper [27], the author used different types of classifiers along with different types of biomarkers. Paper [21], the author has done a comparative study, which includes the different types of classifiers and predicted that the SVM without the fast co-relation based filter is providing highest accuracy which is97.9 percent. In paper [23], the author for classification purposes performs Logistic regression. Paper [4], proposed the comparison of back propagation neural network method and logistic regression model. Paper [7], used three classification algorithms that are SVM, Bayesian network and random forest where SVM is showing the highest performance. In the paper [18], given by Erwin Halim *et. al.* used for mammography purpose atechnique called parallel method which is based on the multi-resolution MRF(MMRF) segmentation, for histological verification they used MLP and for gene identification they used KNN-SVMRFE process and claim that their method will provide automatic digital data and facts which will be useful for diagnosis of breast cancer. Liu Lei used a method in [24] using Linear Regression algorithm and learn modules of Machine learning and classified breast cancer dataset. They got a classification accuracy of 96.5 percent by selecting two features of maximum perimeter and maximum texture. Here, they mention by choosing better feature the classification accuracy can be improved. The paper [9], presents a diagnosis system for detecting breast cancer based on Rep Tree, RBF Network and Simple Logistic. Here 10-fold cross validation method was applied to evaluate the proposed system performances. The correct classification rate of proposed system is 74.5%. This paper showed that the Simple Logistic could be used for reducing the dimension of feature space and proposed Rep Tree and RBF Network model can be used to obtain fast automatic diagnostic systems for other diseases. Joana Diz *et. al.* [13], used data mining based techniques that is useful for cancer cell classification. Two feature extraction techniques are applied here. In addition, for classification purposes various data mining classifiers are applied. The results of classifiers are tested under the parameters called area under curve, sensitivity and specificity. Gopal K. Dhondalay *et. al.* [12], mentioned in their paper about the gene signatures labeled with ER positive and ER negative classes which are using artificial neural network. Their model has showed the efficiency in terms of selecting genes compared to other model. Yassin *et. al.* [30], have given a review paper based on machine learning techniques for diagnosis of breast cancer. Their systematic review is mainly based on to identify various studies which are related to breast cancer CAD system and MLT classifiers. Pramanik *et.al.*[26], proposed a sequential hybrid intelligence system, which is useful for segmentation of the abnormal regions. The research work is mainly based on segmentation of breast thermal image as it provides the information about the importance of breast cell. In paper [3], experiments have performed using SVM classifier with different train-test ratio. The performance of the research is evaluated by the parameters accuracy, ROC curves, specificity, sensitivity, and confusion matrix. The accuracy achieved sing SVM is almost 99.51 percentages. The paper [19], present a review of deep learning methods in breast screening. It is mentioned in this paper that machine learning mechanisms are providing satisfactory results to detect malignancy and determining the density but deep learning approaches are self sufficient and independent learner. Meriem A. *et. al.* [25], present a comparative analysis of two supervised machine learning approaches as Naive Bayes classifier and KNN classifier and using K–fold cross validation the accuracy is predicted for the classifier.

III. BACKGROUND STUDY

For handling the cancer data set, supervised machine learning algorithms are used here. Supervised learning is a process where labeled data is given as an input to the algorithm and based on that classification process has performed.

There are many supervised learning algorithms, which are better for classification purposes.

A. Support Vector Machine:

This is the most used classification algorithm. The working principal Support vector machine is mainly based on marginal calculations [11]. By a hyper-plane, it divides the data points into different subclass. In case of linear SVM, the mathematical expression of the hyper-plane is written below:

$$\mathbf{x_i w + b = 0} \quad \mathbf{(1)}$$

For minimal classification error, the distance of the hyper plane and classes should be maximum. The hyper-plane to have the most extreme edge, which can amplify the separation of the hyper-plane and closest focuses from the two classes. The mechanism of optimal hyper-plane for division of training data without error can be examined using soft margin which will permit an interpretive technique of learning with errors. The optimal hyper-plane can be described as below, if there are *n* numbers of training patterns say

$$\mathbf{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),\ x_i \text{ € } [-1, 1]} \qquad \mathbf{(2)}$$

can be linearly separable if there exist a scalar *b* and vector *w* than,

$$\boldsymbol{w \times y_i + b\ + 1, if\ x_i = 1\ and}$$
$$\boldsymbol{w\ \times y_i\ + b - 1, if\ x_i =\ -1} \qquad \mathbf{(3)}$$

Than the optimal hyper-plane Equation (4) should divide the training data with ultimate margins.

$$\boldsymbol{w_0 \times x\ + b_0 = 0} \qquad \mathbf{(4)}$$

Development of the mechanism of Support vector machines were done for training data separation with minimal or no error. Later the support vector approach is extended to overlaying the idea of division without error on the training vectors is un-achievable. As robust and ground breaking as neural networks the support vectors can be, consider as a new learning machine with this new extension.

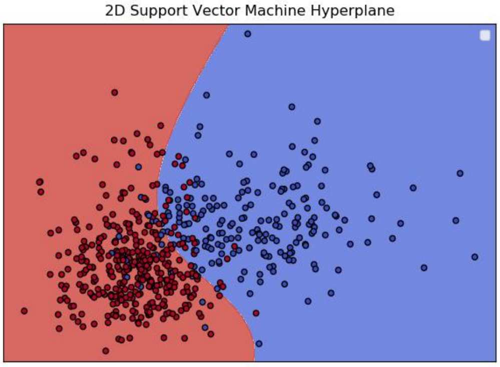

Fig 1: Example of SVM classification

B. K-Nearest Neighbor:

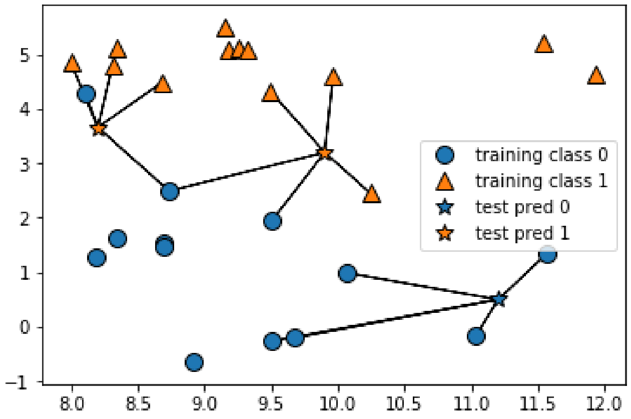

Fig 2: Example of KNN classification

KNN algorithm is based on a methodology that the specimen within the dataset should be in nearest region to other specimen with similar properties [22]. The presence of any specimen without any classification label can be determined by observing the properties of its nearest class. For measurement of Euclidean distance, which can be used to measure nearest neighbor following equation can be applied.

$$\boldsymbol{d(p,q) = \sqrt{(x_i - x)^2 + (y_i - y)^2}\ , \forall\ i\ \in (1,n)} \, \mathbf{(5)}$$

The execution of a KNN classifier is essentially controlled by the decision of K just as the application of separation matrices. The scope is influenced by the affectability of the choice of the area of size K, as the fact that the sweeps of the nearby points are dictated by the separation of theK^{th}nearest neighborto diverse theKyields distinctive contingent class probabilities. In case ifKisextremely less, the local estimation will in generally produce be exceptionally poor attribute consists of mislabeled and noisy points. So for further equalize the area consider the K as large value and consider an expansive locally around the region. However, a huge estimation of K effectively makes the area over equalization and the performance of classification deteriorate which includes the anomaly from different classes. The capacity of KNN algorithm is proved in some real domain. However, there is some limitation of this algorithm as the storage requirement is large, incase to search the similar specimen from neighbors the choice of similarity function is very responsive.

C. Random Forest:

It is a type of ensemble learning process where the predictions of an instance can be determined by combination of different trees, which are trained via isolate training [10]. In this method, the trees are trained independently and through averaging the prediction of trees are combined. Random Forest generally use ID3 algorithm for decision tree training. There are three principle decisions to be made

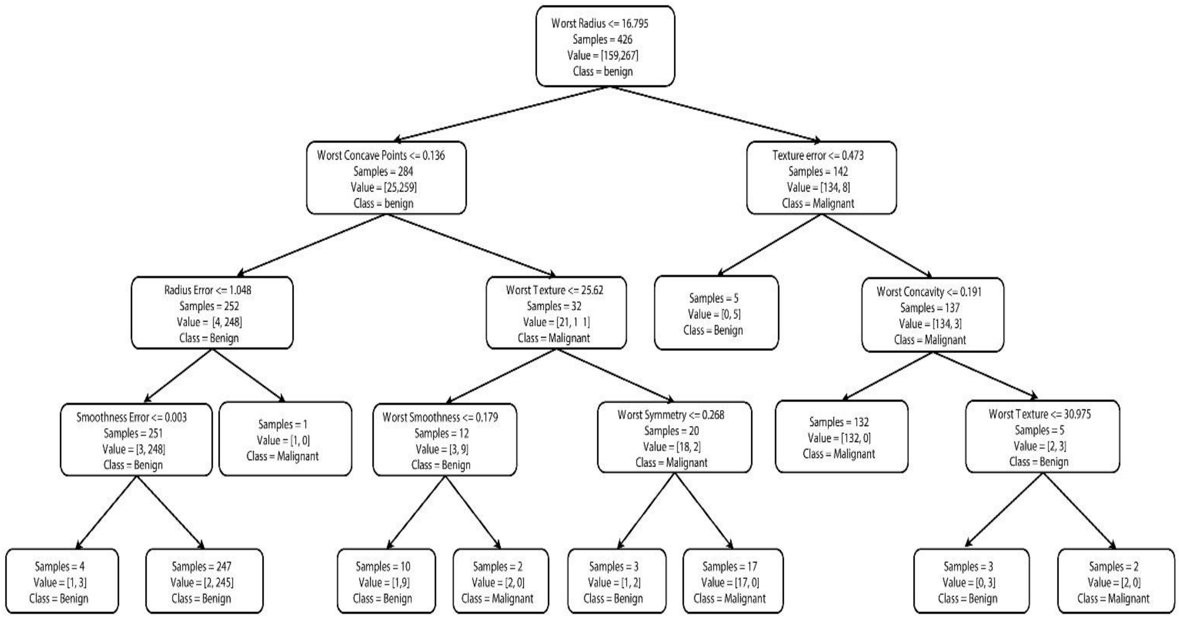

Fig-3: Flow diagram of a tree from Random Forest

while constructing a random tree. These are the strategy for part the leafs, the kind of indicator to use in each leaf, and the technique forinf using haphazardness into the trees. In addition, Gini measure is used for calculating the utilization of split feature. The mathematical explanation of Gini can be given here,

$$\boldsymbol{Gini(P_m) = \sum_n Q_{mn}(1 - Q_{mn})} \qquad \textbf{(6)}$$

An element choice dependent on Gini significance, not with standing and may go before a regularized direct grouping to recognize this ideal subset of highlights. And to procure a two-fold advantage of both dimensional decrease and the disposal of noise from the order task.

IV. DATA-SET COLLECTION AND ANALYSIS

A. Data-set Collection:

We collected Wisconsin Breast Cancer Dataset (WBCD) from Kaggle [1]. This dataset contains 569 patient details, out of which there are 212 Malignantpatients and 357 Benign patients. The dataset contains 32 columns. Here, one unnecessary column is Patient IDalso present. So, this column is removed from the dataset.

After this, there are remaining total31columns in the dataset, where30 features are the independent parameters (Input parameters) and the last column named Diagnosis contains the target values of input parameters. Therefore, now this dataset becomes a processed dataset with 569rows and31columns. After the data in the dataset has been organized, the next task is to gain some information. In order to do that, we start accessing the dataset. This dataset consists of following 10main features: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Each feature here consists of 3sub-features *i.e.* mean of feature, standard error of feature and worst parameter of feature.

B. Processing the Data-set:

To get meaningful information from data-set, we performed the following procedure with the dataset.

1. Take the original dataset and divide the dataset into 2 parts, Benign and Malignant, depending upon the diagnosis value (target value) of dataset last column.

- The Benign table will consist of 357 patient details.
- The Malignant table will consist of 212 patient details.

2. Calculate the minimum, mean, maximum and standard deviation of the features present.
3. So pre-processing of the data present in Train-Test data is required next.

In Table 1 an example of first six feature properties for class benign is given. Here, Benign_min column holds the minimum value of each feature present in the data-set of class Benign.

Table-1: First 6 properties of class Benign

Features	**Benign_min**	**Benign_mean**	**Benign_max**	**Benign_SD**
Radius_mean	6.921	12.146	17.85	10.95
Texture_mean	9.71	17.914	33.81	3.99
Perimeter_mean	43.79	78.05	114.60	11.80
Area_mean	143.5	462.79	992.10	134.287
Smoothness_mean	0.0526	0.924	0.1634	0.0134
Compactness_mean	0.01938	0.080085	0.22390	0.033790

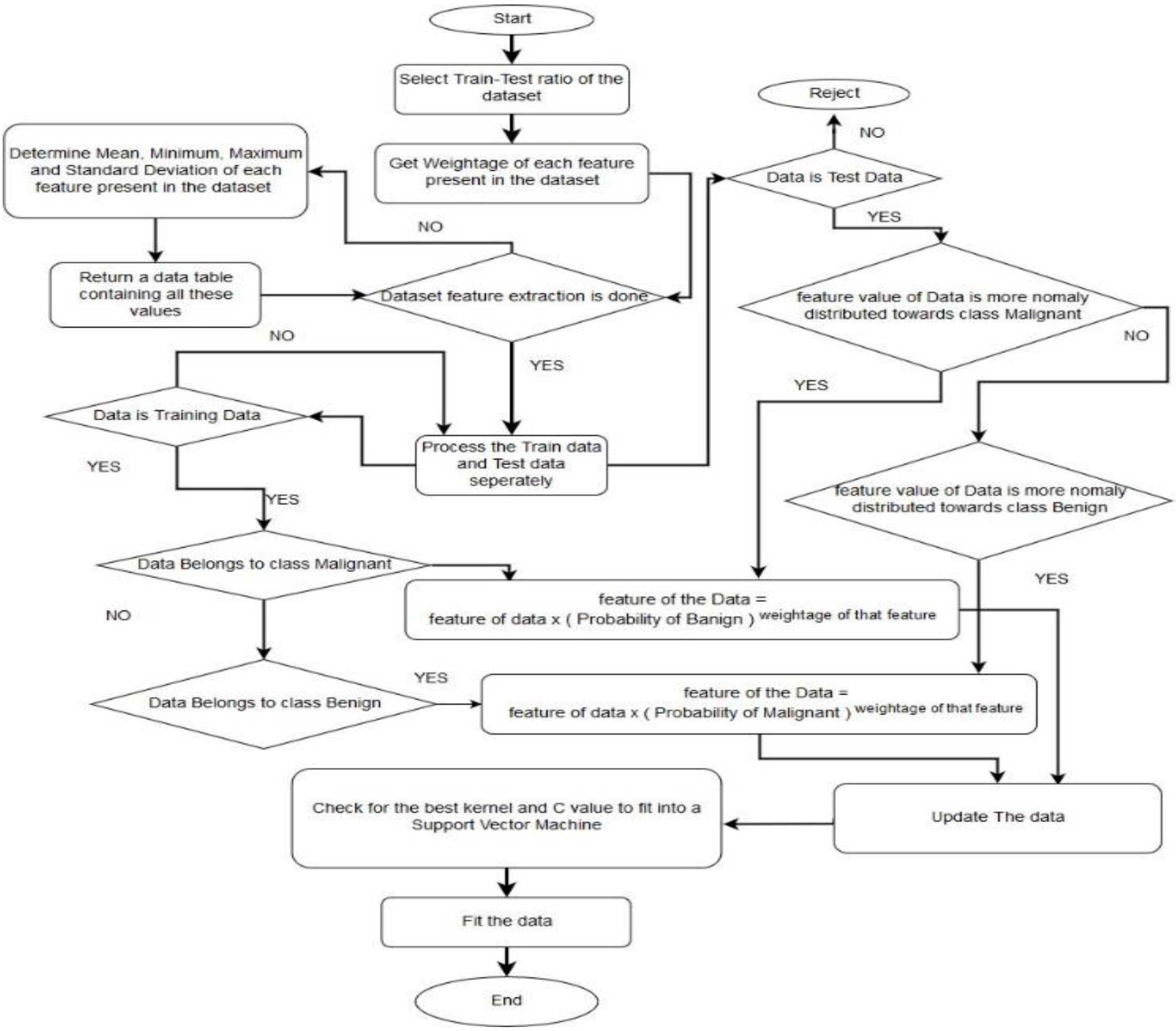
Start
Select Train-Test ratio of the dataset
Get Weightage of each feature present in the dataset
Determine Mean, Minimum, Maximum and Standard Deviation of each feature present in the dataset
Return a data table containing all these values
NO
Dataset feature extraction is done
NO
YES
Data is Training Data
Process the Train data and Test data seperately
YES
YES
Data Belongs to class Malignant
NO
Data Belongs to class Benign
YES
Reject
NO
Data is Test Data
YES
feature value of Data is more nomaly distributed towards class Malignant
NO
YES
feature value of Data is more nomaly distributed towards class Benign
YES
feature of the Data = feature of data x (Probability of Banign) weightage of that feature
feature of the Data = feature of data x (Probability of Malignant) weightage of that feature
Update The data
Check for the best kernel and C value to fit into a Support Vector Machine
Fit the data
End

Fig 4: Proposed Architecture of Classification Method

Benign_mean column holds the mean value of each feature present in the data set of class Benign.

Benign_max column holds the maximum value of each feature present in the data set of class Benign.

Benign_SD column holds the standard deviation value of each feature present in the data set of class Benign.

1. The training data is passed through a random forest classifier. For this architecture, we have selected 100 numbers of Decision trees for the Forest. There after we collect the random forest generated, feature weights of all the features present in the data set.
2. Since, in this dataset, all the data are nearly normally distributed. Therefore, the data is pre-processed with the assumption that the data is normally distributed.
3. Train data contains it's labels, so we can process them directly depending on their target values. But the test data cannot be processed in the same way. So we will apply some transformations in test data in order to get more accurate pre-processed data
4. Train Data Processing: First, check the train data whether it is Benign or Malignant. If the train data

V. PROPOSED ARCHITECTURE

The proposed architecture of the classification model is given in Figure 4. Based on the availability of data, in the proposed architecture; we can choose any kind of Train-Test data ratio. The data set is then divided after selection
of Train-Test ratio.

belongs to class Malignant, then update the Train Data in following ways:

$$New\_train = Train\_data \times (Probability\ of\ Benign)^{(Weight\ of\ feature)} \qquad (7)$$

If the train data belongs to class Benign, then update the train data in the following way,

$$New\_train = Train\_data \times$$

$$(Probability\ of\ Malignant)^{(Weight\ of\ feature)} \quad (8)$$

Here, feature values of train data are multiplied with their inverse probability to gain equality in the outcome. In this process, we get a stabilized data.

5. Test Data Processing: For mathematical representation, we assume the followings:
   C represents Class it will be either B for Benign and M for Malignant. i^{th} Feature of C is given by$f_{i,c}$. Mean of i^{th}feature of C is represented by $m_{i,c}$. Deviation from Mean of i^{th} feature from its class C is given by following Equation.

$$\Delta m_{i,c} = f_{i,c} - m_{i,c} \quad (9)$$

Standard Deviation (SD) of$f_{i,c}$ is represented by $SD_{i,c}$ and Standard Normal Value (SND) is given by the following equation,

$$SNV_{i,c} = \frac{\Delta m_{i,c}}{SD_{i,c}} = \frac{f_{i,c} - m_{i,c}}{SD_{i,c}} \quad (10)$$

Maximum and minimum of $f_{i,c}$ are given by $max_{i,c}$ and $min_{i,c}$ respectively. Data distribution range of i^{th} feature of class C is given by the following equation,

$$DDR_{i,c} = max_{i,c} - min_{i,c} \quad (11)$$

Standard class deviation of a feature from mean is given by the following equation,

$$SCDM_c = DDR_{i,c} \times SNV_{i,c}$$
$$SCDM_c = \left(max_{i,c} - min_{i,c}\right) \times \frac{\Delta m_{i,c}}{SD_{i,c}} \quad (12)$$
$$SCDM_c = (max_{i,c} - min_{i,c}) \times \frac{(f_{i,c} - m_{i,c})}{SD_{i,c}} \quad (13)$$

As in Figure 6, it is shown that data distribution follows normal distribution, the distribution mainly depends on the $SNV_{i,c}$ value. However, we can seethe data distribution range also affects the outcome of processed data. In other words, the outcome also depends on the range of the data in which data is distributed. So we will multiply the $DDR_{i,c}$ with $SNV_{i,c}$ value.This will give us a new parameter $SCDM_c$, whereCis Class of cancer. $SCDM_{i,c}$Value will be calculated for each feature and for each of the class Malignant and Benign. The next steps will be followed as: If$SCDM_B$<$SCDM_M$, we will consider the test case belongs to the class Benign sincethe value lies near to mean of Benign feature.

$$New\_test = Test\_data \times (Probability\ of\ Benign)^{(Weight\ of\ feature)} \quad (14)$$

And if $SCDM_B > SCDM_M$, we will consider the test case belongs to the class Benign sincethe value lies near to mean of Benign feature.

$$New\_test = Test\_data \times (Probability\ of\ Malignant)^{(Weight\ of\ feature)} \quad (15)$$

6. Scaling Data in Train-Test Division: The Train and Test data are distributed in a logarithmic scale. Therefore, we need to convert them to a scalar state so that the Support Vector Machine can easily classify the data. So, we will apply this transformation to all the data available here:

Minimum of a feature in *New_Train* is given as$min_{f,NTR}$, Maximum of a feature in *New_Train* is given by $max_{f,NTR}$, Minimum of a feature in *New_Train* with class C is given by $min_{f,C,NTR}$, Maximum of feature of in *New_Train* with class C is given by $max_{f,C,N,T,R}$. We first take the minimum and maximum of each feature of the training data present in the modified data set. Minimum of Train data as:

$$min_{f,NTR} = \min\left(min_{f,Benign,NTR}, min_{f,Malignant,NTR}\right) \quad (16)$$

Maximum of train data,

$$max_{f,NTR} = \max\left(max_{f,Benign,NTR}, max_{f,Malignant,NTR}\right) \quad (17)$$

After that we find the distribution value of feature f,

$$DR_f = (max_{f,NTR} - min_{f,NTR}) \times (max_{f,NTR}) \quad (18)$$

Then we would scale *New_Train* data using the formula below,

$$Scaled\_Train = \frac{New\_Train - min_{f,NTR}}{DR_f} \quad (19)$$

$$Scaled\_Test = \frac{New\_Test - min_{f,NTR}}{DR_f} \quad (20)$$

7. Process Scaled Data using SVM: Since, now the data has been scaled, so we will proceed for SVM kernel selection. We will pass the data through four kernels available in SVM (Linear kernel, Poly Kernel, RBF Kernel, Sigmoid Kernel). The kernel that yields a maximum output is chosen. However, since the outcome depends on test data accuracy, so we need to give weights to SVM's output. Let, Accuracy of *Scaled_Train* with SVM (STR) is given a weight of test data ratio (TeS). Accuracy of *Scaled_Test* with SVM (STE) is given a weight of train data ratio (TrS). Here,

$$TeS = \frac{size(Scaled\_Test)}{size(Scaled\_Train) + size(Scaled\_Test)} \quad (21)$$

$$TrS = \frac{size(Scaled\_Train)}{size(Scaled\_Train) + size(Scaled\_Test)} \quad (22)$$

The equation for kernel selection weight = *Kernel_Score* can be written as,

$$Kernel\_Score = (STR) \times (Tes) + (STE) \times (TrS) \quad (23)$$

Because, higher the training data, higher will be the chance that the data will over fit. Therefore, we have to reduce its effect in the selection of kernel. With this weighing system, the kernel selection becomes more reliable. The best kernel is chosen depending on the value of the highest *kernel_Score* graded to each kernel.

8. We apply similar procedure in the selection of C value (Penalty Parameter C of the error term), we repeat the same procedure of weighing outcomes. We pass the Train data as well as the Test data to a SVM with the selected kernel and check for the best fit provided by the classifier by varying the input parameter C in the range [0.01, 100].

9. Each SVM is rated for selected kernel where the value of C is varying. Kernel_Score_with C. denotes the score. This is more likely a hit and trial method for the selection of best C value. Each SVM is again scored using the following formula,

$$Kernel\_Score\_withC = (STR_k) \times (Tes) + (STE_k) \times (TrS) \quad (24)$$

Here, $STR_k$is the SVM accuracy with *Scaled_Train* data with selected Kernel k and STE_k is the SVM accuracy with *Scaled_Test* data with the selected kernel *k*.

Depending on the highest*Kernel_Score_with C* for different value of C, C value will be selected.

10. RFSVM has decided the most suitable value of C with the suitable kernel k. C is decided by the best value of Kernel_Score_with C, and k is decided by the best value of Kernel_Score. With this step, the selection procedure of parameters of SVM is complete.
    The classifier will return the result (accuracy) of the given data with the best-estimated kernel and C value that fits the data with very high accuracy.

VI. RESULT AND ANALYSIS

In experimental work here we used the set up of Intel®$Core^{TM}$ i5-7200U processor with 8-GB DDR4-2133 SDRAM and windows 10 operating system. For simulation work, we used Python 3.6.7 with scikit-learn, pandas, matplotlib and numpy modules. We used confusion matrix to calculate accuracy of the classifier. The format of confusion matrix is as given below: There are 4 terms associated with calculation of accuracy of a classifier. Their mathematical definitions are also given with them. Here, each term has different meaning.

	Predicted True	**Predicted False**
Actual True	True Positive [31]	False Positive [32]
Actual False	False Negative	True Negative

False Positive(FP) = Predicts Benign as Malignant
True Negative(TN) = Predicts Benign as Benign
False Negative(FN) = Predicts Malignant as Benign

$$Accuracy = \frac{(TN + TP)}{(TN + TP + FP + FN)} \times 100$$

$$Recall = \frac{(TP)}{(TP + FN)}$$

$$True\ Negative\ Rate(TNR) = \frac{(FP)}{(TN + FP)}$$

$$Precision = \frac{(TP)}{(TP + FP)}$$

- The F-beta score weights recall more than precision by a factor of beta. The value beta = 1.0 means recall and precision are equally important.
- The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

$$F1 = \frac{2(Precision\ \times Recall)}{(Precision + Recall)}$$

- The support is the number of occurrences of each class in test and target Data.

Table 2: Experimentation Results proposed RFSVM architecture of Confusion Matrix Parameters for different Train-Test Ratio

Parameter	**Train-Test Ratio**			
	50-50	**60-40**	**70-30**	**80-20**
TP	107	77	63	40
TN	6	4	1	1
FP	168	146	106	73
FN	4	1	1	0

Table 3: Experimentation Results of proposed RFSVM architecture for different Train-Test Ratio

Ratio	**50-50**		**60-40**		**70-30**		**80-20**	
Class	**Malignant**	**Benign**	**Malignant**	**Benign**	**Malignant**	**Benign**	**Malignant**	**Benign**
Precision	0.96	0.97	0.98	0.97	0.98	0.99	1.00	0.99
Recall	0.95	0.98	0.98	0.99	0.98	0.99	0.98	1.00
F1-Score	0.96	0.97	0.98	0.98	0.98	0.99	0.99	0.99
Support	113	172	81	147	64	107	41	73

Table 4: Experimentation Results of Support Vector Machine for different Train-Test Ratio

Ratio	**50-50**		**60-40**		**70-30**		**80-20**	
Class	**Malignant**	**Benign**	**Malignant**	**Benign**	**Malignant**	**Benign**	**Malignant**	**Benign**
Precision	0.99	0.90	1.00	0.90	1.00	0.93	1.00	0.96
Recall	0.80	0.99	0.82	1.00	0.84	1.00	0.92	1.00
F1-Score	0.89	0.95	0.90	0.95	0.92	0.96	0.96	0.98
Support	101	184	87	141	58	113	37	77

Table 5: Experimentation Results of K-Nearest Neighbor for different Train-Test Ratio

Ratio	**50-50**		**60-40**		**70-30**		**80-20**	
Class	**Malignant**	**Benign**	**Malignant**	**Benign**	**Malignant**	**Benign**	**Malignant**	**Benign**
Precision	0.97	0.93	0.99	0.99	0.98	0.92	0.97	0.97
Recall	0.87	0.98	0.97	0.99	0.87	0.99	0.95	0.99
F1-Score	0.92	0.96	0.98	0.99	0.93	0.95	0.96	0.98
Support	103	182	76	152	71	100	38	76

Table 6: Experimentation Results of Random Forest Classifier for different Train-Test Ratio

Ratio	**50-50**		**60-40**		**70-30**		**80-20**	
Class	**Malignant**	**Benign**	**Malignant**	**Benign**	**Malignant**	**Benign**	**Malignant**	**Benign**
Precision	0.93	0.98	0.91	0.97	0.90	0.95	0.87	1.00
Recall	0.96	0.96	0.94	0.95	0.93	0.93	1.00	0.92
F1-Score	0.94	0.97	0.92	0.96	0.92	0.94	0.93	0.96
Support	104	181	78	150	70	101	39	75

Table 7: Accuracy of Proposed RFSVM architecture for different Train-Test ratio

	Train-Test ratio			
	50-50	**60-40**	**70-30**	**80-20**
Accuracy	**96.5**	**97.8**	**98.8**	**99.1**

Table 8: Accuracy comparison of Proposed RFSVM architecture with other classifiers for different Train-Test ratio

	Train-Test ratio			
Classifiers	**50-50**	**60-40**	**70-30**	**80-20**
SVM	93.0	93.0	95.0	97.0
KNN	94.0	99.0	94.0	97.0
RF	96.0	95.0	93.0	95.0
RFSVM	**96.5**	**97.8**	**98.8**	**99.1**

Table 9: Accuracy of Proposed RFSVM architecture for different Train-Test ratio

	Train-Test ratio			
Instances	**50-50**	**60-40**	**70-30**	**80-20**
1	98.596	98.947	99.415	99.714
2	98,245	98.596	98.947	99.441
3	97.895	98.246	98.830	99.123
4	97.544	97.895	98.596	98.596
5	97.193	97.192	98.246	98.246
Accuracy	**96.5**	**97.8**	**98.8**	**99.1**

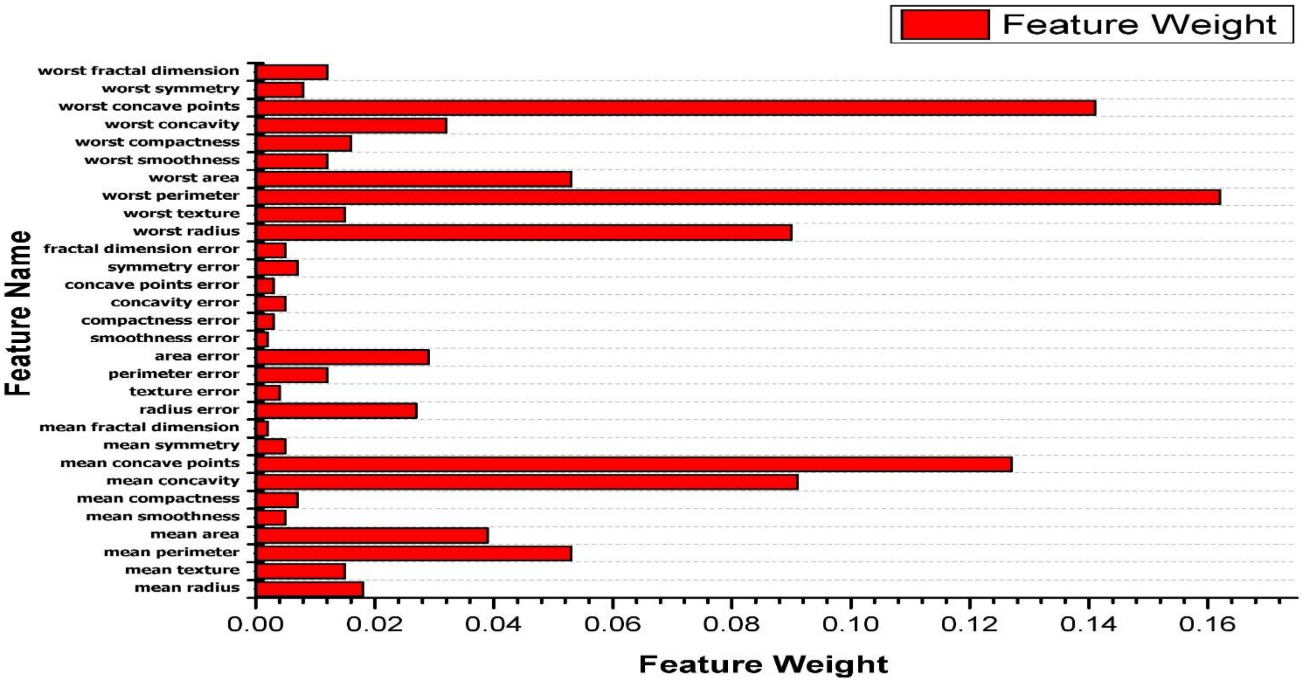

Fig 5: The weightage provided by random forest

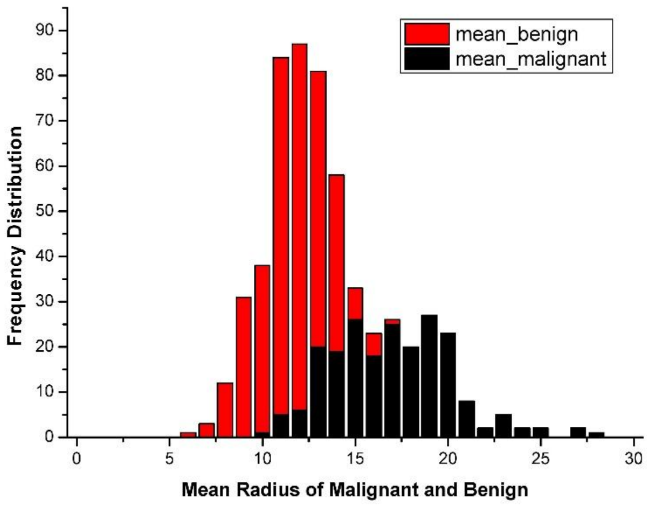

Fig 6: Figure showing Normal Distribution of mean for both Benign and Malignant class

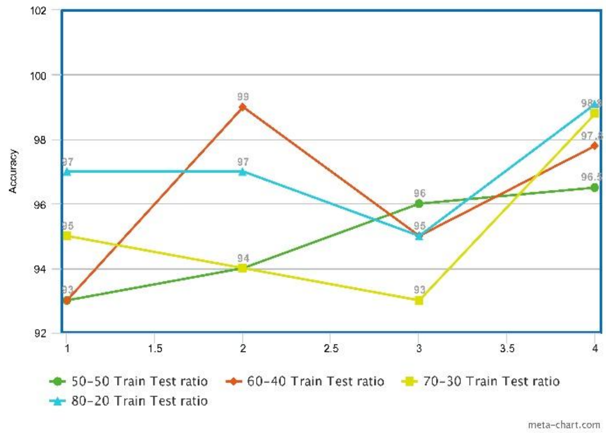

Fig 7: Classifier accuracy with different Train-Test ratio

Above mentioned performance measures, values are given in Table 2 and Table 3.Here Table 2 is showing the true positive, true negative, false positive, false negative of RFSVM classifier with different train-test ratio of data. Table 3 is showing the precision, recall, F1-score and support value of RFSVM classifier for each Benign and Malignant class. It can be seen that the classifier very accurate while dealing with the data. In Table 4, Precision, Recall, F1-score and Support value for Support Vector Machine can be seen. With normalized data, we can see a relatively positively growing accuracy with the increase in the size of training data (Table 8). Similarly, Table 5 and Table 6 show the output of traditional methods K-Nearest Neighbor and Random Forest respectively. From Table 8, we can see that with increasing training data size, accuracy of KNN fluctuates. This is because KNN vastly depends upon the distribution of data point near the test sample. Random forest can fit the data very well, if the height of tree is not specified. This will lead to over fitting of sample data. To avoid this, height of the tree is kept in control. Hence, the accuracy almost lies near 95%. The comparative results are given in Table 8, using the classifiers SVM, KNN, RF and the proposed hybrid classifier RFSVM. Using different train-test ratio we can see in Table 8 that the accuracy of RFSVM classifier is higher than the traditional classifiers. Using 50-50, 60-40, 70-30 and 80-20 train-test ratio experiments are done and in all cases, the accuracy of RFSVM is more than 96 percent.

The Table 9 shows the values of RFSVM with ratio 50-50, 60-40, 70-30and 80-20 using five instances. Each time the classifier called and a train-test ratio is provided the data is randomly selected. Therefore, each time depending on selection of data the accuracy of RFSVM varies. However, using randomly selected data the classifier is able to classify accurately which is an advantage of the proposed classifier over the other. Here data was passed through RFSVM classifier a total of 20 times for 4 different train-test ratio. This means, for each train-test ratio, we called RFSVM 5 times. The output of the classifier is shown in Table 9. The average of the outcome for each train-test ratio is also shown in the table.

VII. DISCUSSION AND ANALYSIS

RFSVM is a hybrid classifier. It is based on both mathematical analysis of data as well as hit and trial method. However, the part that makes is unique is the data processing. Another specialty of RFSVM is that it can be applied to any dataset, whether it is binary classification or not. RFSVM can be used even if the amount of data is very less. Here, with the Wisconsin Breast Cancer Dataset (WBCD), we can see the same output. Even with fewer amounts of training data, the classifier is able to give relatively high accuracy as compared to other classifiers.

Breast cancer is growing very rapidly. In such scenario, a proper prediction can help a patient in survival. In this case, RFSVM is a successful classifier. RFSVM has some advantages, which are given below.

– RFSVM can yield high accuracy with relatively less amount of data.
– Time taken by RFSVM to yield output is relatively lesser than other classifiers with same level of accuracy. Eg: Neural Network.
– Accuracy of other classifiers depends directly on the accuracy of collected information. However, since in case of RFSVM raw data is processed and modified, the chance of getting a better result with high accuracy is more in case of RFSVM.

However, there are also some constraints. While using the RFSVM, these points must be keep under consideration.

– If amount of data is very less, the results will not be much reliable.
– If unnecessary data is provided to the classifier (like patient ID has nothing to do with the output of the classifier), the results may vary.
– Before inserting data to the classifier, one should remove incomplete data from the table. It is not mandatory. But to get a proper result, one shoulder fine the raw data before proceeding

VIII. CONCLUSION AND FUTURE SCOPE

This paper concludes that with detailed analysis of WBCD, the accuracy of aclassifier can be increased to a very high amount. This paper proposed a methodology that combines the pros of Random Forest Classifier and Support Vector Machine and yields result with a very high accuracy. Due to the proper processing and prepossessing of data, our proposed classifier provides us accuracy upto 99.714 percent. In addition, the data distribution and selection of train data as well as test data also matters. However, RFSVM classifier helps us to reduce the effect of data size in its accuracy. Due to proper feature selection method, RFSVM can give 97.895 percent average accuracy even with 50-percentage train data size. Larger the train data, larger the chance to get a higher accuracy. However, we cannot keep increasing the train data size because at a certain time, it will lead towards over fitting. Since, tree height cannot be pre-determined for a better classification of data, direct feeding of data into Random Forest leads to over-fitting of data. In addition, to get a satisfactory result from Support Vector Machine, one has to tune the perimeters of Support Vector Machine many times. The combine methodology helps to get rid of this problem. Data is fitted well using this method. SVM is tuned itself again and again until the result meets a standard outcome. If this technique is applied with different data but same train-test ratio, the overview of the parameters of SVM can be found. This will tell us about the property of distribution of data in the dataset. In this way, it can be seen that RFSVM classifies data of data-set with much less difficulty and high accuracy.

Hybridized classifiers are generally better than most other classifiers. This same can be observed with RFSVM classifier. It performs very well with the given dataset. Due to the feature extraction method and processing train-data and test data in different manners, this hybrid classifier outperforms the other classifiers. This classifier is not limited to binary dataset only. It can be applied to classify other datasets with a more number of classes. Further exploration of the dataset can lead us to identify a better feature weight selection and a better train-test data processing method. These results will be focused on our future work.

Compliance with ethical standards: Yes

Conflict of interest: The authors declare that they have no conflict of interest.

REFERENCES

1. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data, april 2019.
2. Ahmad, F. K., and Yusoff, N. Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier. In 2013 13th International Conference on Intellient Systems Design and Applications (2013), IEEE, pp. 121–125.
3. Akay, M. F. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst. Appl. 36, 2 (Mar. 2009), 3240–3247.
4. Alarabeyyat, A., Alhanahnah, M., Breast cancer detection using k-nearestneighbor machine learning algorithm. In2016 9th International Conference on Developments in eSystems Engineering (DeSE)(2016), IEEE, pp. 35–39.
5. Arpita Ghosh, ShaluAchamma Sam, A. N. R.Abnormal lung cells detection usingwatershedalgorithm.Research Journal of Pharmacy and Technology(2018).
6. Baneriee, C., Paul, S., and Ghoshal, M.A comparative study of different ensemblelearning techniques using wisconsin breast cancer dataset. In 2017 International Conference on Computer, Electrical & Communication Engineering (ICCECE)(2017), IEEE,pp. 1–6.
7. Bazazeh, D., and Shubair, R.Comparative study of machine learning algorithmsfor breast cancer detection and diagnosis. In2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)(2016), IEEE, pp. 1–4.
8. Chang, J., Yu, J., Han, T., Chang, H.-j., and Park, E.A method for classifyingmedical images using transfer learning: A pilot study on histopathology of breast cancer.In2017 IEEE 19th International Conference on e-Health Networking, Applications andServices (Healthcom)(2017), IEEE, pp. 1–4.
9. Chaurasia, V.Data mining techniques: To predict and resolve breast cancer survivability.International Journal of Computer Science and Mobile Computing Vol. 3(01-2014), pg.10 – 22.
10. Denil, M., Matheson, D., and Freitas, N. D.Narrowing the gap: Random forests intheory and in practice. In Proceedings of the 31st International Conference on MachineLearning(Bejing, China, 22–24 Jun 2014), E. P. Xing and T. Jebara, Eds., vol. 32 ofProceedings of Machine Learning Research, PMLR, pp. 665–673.
11. Dey, A.Machine learning algorithms : A review.(2016), (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (3) , 2016, 1174-1179
12. Dhondalay, G. K., Tong, D.-L., and Ball, G. R.Estrogen receptor status prediction for breast cancer using artificial neural network.2011 International Conference on Machine Learning and Cybernetics 2(2011), 727–731.

13. Diz, J., Marreiros, G., and Freitas, A.Using data mining techniques to support breast cancer diagnosis. In New Contributions in Information Systems and Technologies(Cham, 2015), A. Rocha, A. M. Correia, S. Costanzo, and L. P. Reis, Eds., SpringerInternational Publishing, pp. 689–700.
14. Gayathri, B., and Sumathi, C. Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer. In2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)(2016), IEEE, pp. 1–5.
15. Gogoi, U. R., Bhowmik, M. K., Bhattacharjee, D., Ghosh, A. K., and Majumdar, G.A study and analysis of hybrid intelligent techniques for breast cancerdetection using breast thermograms. InHybrid Soft Computing Approaches. Springer,2016, pp. 329–359.
16. Gupta, M., and Gupta, B.A comparative study of breast cancer diagnosis usingsupervised machine learning techniques. In2018 Second International Conference on Computing Methodologies and Communication (ICCMC)(2018), IEEE, pp. 997–1002.
17. Gupta, M., and Gupta, B.An ensemble model for breast cancer prediction using sequential least squares programming method (slsqp). In2018 Eleventh International Conference on Contemporary Computing (IC3)(2018), IEEE, pp. 1–3.
18. Halim, E., Phoebe Halim, P., and Hebrard, M.Artificial intelligent models for breast cancer early detection. pp. 1–9.
19. Harvey, H., Karpati, E., Khara, G., Korkinof, D., Ng, A., Austin, C., Rijken,T., and Kecskemethy, P.The role of deep learning in breast screening.CurrentBreast Cancer Reports 11, 1 (Mar 2019), 17–22.
20. Khourdifi, Y., and Bahaj, M.Applying best machine learning algorithms for breastcancer prediction and classification. In2018 International Conference on Electronics,Control, Optimization and Computer Science (ICECOCS)(2018), IEEE, pp. 1–5.
21. Khourdifi, Y., and Bahaj, M.Feature selection with fast correlation-based filterfor breast cancer prediction and classification using machine-learning algorithms. In2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)(2018), IEEE, pp. 1–6.
22. Kotsiantis, S. B.Supervised machine learning: A review of classification techniques. In Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applicationsin Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies(Amsterdam, The Netherlands, The Netherlands, 2007), IOS Press, pp. 3–24.
23. Liu, L.Research on logistic regression algorithm of breast cancer diagnose data by machine learning. In2018 International Conference on Robots & Intelligent System (ICRIS)(2018), IEEE, pp. 157–160.
24. Liu, L.Research on logistic regression algorithm of breast cancer diagnose data bymachine learning. In2018 International Conference on Robots Intelligent System(ICRIS)(Los Alamitos, CA, USA, may 2018), IEEE Computer Society, pp. 157–160.
25. Meriem, A., Oukid, S., Gagaoua, I., and Ensari, T. Breast cancer classification using machine learning. pp. 1–4.
26. Pramanik, S., BHOWMIK, M., Bhattacharjee, D., and Nasipuri, M. HybridIntelligent Techniques for Segmentation of Breast Thermograms. 06 2016.
27. Saleh, D. T., Attia, A., and Shaker, O. Studying combined breast cancer biomarkers using machine learning techniques. In2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI)(2016), IEEE, pp. 247–251.
28. Sultan, L. R., Cary, T. W., and Sehgal, C. M. Machine learning to improve breastcancer diagnosis by multimodal ultrasound. In2018 IEEE International Ultrasonics Symposium (IUS) (2018), IEEE, pp. 1–4.
29. Wang, Z., Li, M., Wang, H., Jiang, H., Yao, Y., Zhang, H., and Xin, J.Breastcancer detection using extreme learning machine based on feature fusion with cnn deep features.IEEE Access PP(01 2019), 1–1.
30. Yassin, N., Omran, S., El Houby, E., and Allam, H. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: Asystematicreview. Computer Methods and Programs in Biomedicine 156(12 2017).
31. Soni, Badal, Pradip K. Das, and Dalton Meitei Thounaojam. "Copy-move tampering detection based on local binary pattern histogram fourier feature." *Proceedings of the 7th International Conference on Computer and Communication Technology*. ACM, 2017.
32. Mathur, Prachi, Kanasani Monica, and Badal Soni. "Improved Fusion-based Technique for Underwater Image Enhancement." *2018 4th International Conference on Computing Communication and Automation (ICCCA)*. IEEE, 2018.

AUTHORS PROFILE

Dr. Badal Soni was born in Madhya Pradesh, India. He did M.Tech. From Indian Institute of Information Technology Design and Manufacturing Jabalpur, India in did his Ph.D. in Computer Science and Engineering Department, National Institute of Technology Silchar, India. Currently, he is working as an assistant professor in the department of Computer Science and Engineering, National Institute of Technology Silchar. His research interests include Machine Learning, Image Processing, Image forgery detection and Speech Processing. He has published more than 25 papers in repudiated International Journals and Conferences. He is the professional member of IEEE, ACM, and act as a reviewer in many journals papers.

Angshuman Borais currently pursuing B.Tech in Computer Science and Engineering in NIT Silchar. He is also currently doing deep learning with various dataset available. His current working is related to training of deep learning models mainly for image classification.In addition, some of his works is associated with Computer Vision.

Arpita Ghosh is currently pursuing Ph.D. from computer science department NIT Silchar. Her current research work is related to training of machine learning and deep learning models mainly for medical image classification and detecting diseases.

Anji Reddy is working as an associate professor in department of CSE, Lendi Institute of Engineering and Technology, Andhra Pradesh and currently pursuing Ph.D. from computer science department NIT Silchar. His research area includes deep learning for real time breast cancer cytology images and training of machine learning and deep learning models mainly for medical image classification.