The Bioscan AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES 21(2): S.I (2), 645-659, 2026 www.thebioscan.com

A FFM-based Feature Selection for Breast Cancer Classification: Machine Learning Approach

Ankita Sinha[1], Ranjan Banerjee[2], Ananya Smruti Snigdha Ojha[3] Sulagna Basu[4], Swarnabha Chakraborty[5], Abhik Bhattacharya[6], Shankar Prasad Mitra[7], Kaushik Paul[8], Debmalya Mukherjee[9]

[1,2,3,4,5,6,7,8,9]Assistant Professor

[1,2,3,4,7,8] Computer Science & Engineering, Brainware University

[5,6,9] Computational Sciences, Brainware University

ethosankita@gmail.com,rbkpccst@gmail.com,ananyaojha1006@gmail.com , sbd.cse@brainwareuniversity.ac.in,swc.cs@brainwareuniversity.ac.in, abh.cs@brainwareuniversity.ac.in, spmitra2016@gmail.com, kkp.cse@brainwareuniversity.ac.in kkp.cse@brainwareuniversity.ac.in,dbml.mukherjee@gmail.com

Orcid Id: 0009-0002-3179-227X, Orcid Id: 0009-0003-1950-7530, ORCID ID: 0009-0005-3844-1965, Orchid Id:0009-0004-2624-8889, Orcid Id.: 0009-0007-1372-6096, Orcid Id: 0009-0003-2392-2040, Orcid Id:0009-0003-5457-5534, Orcid Id: 0009-0009-6929-7651, 0009-0006-9946-0964

DOI: https://doi.org/10.63001/tbs.2026.v21.i02.S.I(2).pp645-659

KEYWORDS

Breast Cancer; Computer Aided Diagnosis; Frequent feature-set mining; Classification

Received on:
14-03-2026

Accepted on:
11-04-2026

Published on:
09-05-2026

Abstract

Breast cancer is a major health issue for women worldwide. Early detection of breast cancer can improve the possibility of curative treatment as well as lower the mortality rate. The aim of this research work is to build a classification model with relevant features from the {Wisconsin breast cancer dataset (WBCD). In this research work, we have applied two distinct methodologies to the pre-processed dataset. The first methodology contains the existing classification model (without feature selection method), while the second methodology includes the proposed feature selection, i.e. Frequent Feature-set Mining (FFM) method based on frequent itemset mining concept along with different learning classifiers like Support Vector Machine (SVM), Naive Bayes, Random Forest, Logistic Regression, Adaptive Boosting (Adaboost), Decision Tree, and k-Nearest Neighbor (k-NN), and finally evaluated on different statistical parameters like accuracy, balanced accuracy, sensitivity, specificity, precision, jaccard similarity, dice coefficient, Receiver Operating Characteristic curve (ROC-Curve) and precision-recall curve (PR-Curve) without and with proposed feature selection method. Because of these factors, the Naive Bayes followed by SVM model did the best without the proposed feature selection method (97.05% and 96.47% accuracy), and the Naive Bayes followed by k-NN model did the best with the proposed feature selection method on the WBCD (99.41% and 98.82% accuracy) for the breast cancer classification model.

1. Introduction

Cancer is a huge public health issue around the world. Breast cancer is currently the most common cancer in females worldwide. Early diagnosis or detection, however, can greatly increase the likelihood of receiving appropriate treatment and recovery. Treatment methods are improving along with the growing population, and datasets with a variety of attributes are growing as well. Keeping records up to date improves the likelihood of survival, hence serves the purpose of developing the classification model. In this case, learning models may be utilized for detecting cancer, especially breast cancer, in a simplified way. According to the World Health Organization (WHO), [1], three out of ten women worldwide

have been diagnosed with breast cancer (in 2020). Breast cancer morale can be low when

the cancer is diagnosed at an early stage and treated immediately. Breast cancer is commonly found in the milk glands and ducts, expanding cancerous cells in an uncontrolled manner [2,3]. Self-breast examination (SBE) of the breast is the primary method of self-diagnosis and increases awareness among women as they are more familiar with their breasts and notice things like redness, nipple discharge, dumpling on the nipple, irritation, or pain in the breast [4]. However, in some cases, women are diagnosed with breast cancer despite the absence of any visible symptoms. Following SBE, there are numerous Computer Aided Diagnostics (CAD) or screening options, including mammography, thermography, and ultrasound. Every diagnostic test has its pros and cons, [3,4], which are discussed as follows:

- **Mammogram:** Easily diagnosing an abnormal area or lump in the breast, a mammogram is a more elaborate version of an X-ray.
- **Magnetic resonance imaging (MRI):** Scans the whole body, especially the breasts, by using a magnetic computer and generating elaborate images of the breasts, especially the inside of the breasts.
- **Ultrasound:** Basically, it uses sound waves to generate images of inside the breast, also known as sonograms.
- **Thermography:** Good for early detection by using a special digital camera with infrared thermal technology for image generation by measuring breast surface temperature.

However, mammography is the best screening option despite having certain drawbacks. Due to the abundance of features, there is a risk of false benign and false malignant conditions. The second limitation is not good for cost or for women with dense breasts [5].

Data Classification model, [6], based on learning classifiers are extensively used in data analytic applications such as data mining in the healthcare sector. There are numerous issues like nature, comprehensibility, missing data, noise in datasets, low computation cost, the capability to deal with irrelevant features, optimal solution, and generalization ability. The solution relies heavily on data science. Data science, a subset of artificial intelligence, considers both Deep Learning (DL) and Machine Learning (ML) pertaining to computer science. Learning classifiers allows us to dig out valuable information through an enormous quantity of data. Without data, a learning classifier is nothing. The more useful data, more will be the accuracy. These datasets contain a lot of information, but the concern is how to extract meaningful knowledge insights from such a huge collection. Learning classifiers can assist in extracting information and rules, as well as training learning models to perform complex tasks and use collected data and experience to become smarter, even if not all of the information in a dataset is necessarily useful. Learning classifiers are used in a variety of industries, including business, healthcare, science, and medicine. In the healthcare sector, learning techniques are utilized for feature filtering, classification, estimate, and decision-making.

ML based classification models and diagnostic models have gained quick and impressive popularity in the Healthcare sector, especially in clinical data. CAD is still a highly challenging and active study area because it doesn't always provide the promised accuracy in early detection or accurately pinpointing the various stages of cancer. Some popular machine learning classifiers, nevertheless, have demonstrated superior performance in a variety of applications, including Naive Bayes, Support Vector Machine (SVM), Decision Tree, k-Nearest Neighbor (k-NN), Adaptive Boosting (AdaBoost), Random Forest (RF), and Logistic Regression.

This research paper is divided into several sections. Section 1.1 i.e related work section, discussed some of existing work followed by

contribution. Section 2, describes the existing and proposed methodology, including data pre-processing, a proposed feature selection algorithm, classification classifiers, and performance statistical measure evaluation. Section 3, describes the experimental result and discussion, including the dataset description followed by the result, statistical analysis i.e ROC curve and PR curve, and a comparative analysis of accuracy with the existing framework. Section 4, concludes the research paper with the future scope of this research.

1.1 Related Work

The significance of feature selection techniques using deep and machine learning in the healthcare industry has previously been noted by numerous studies. Certain research has already been carried out in the breast cancer field; most of it used machine learning classifiers to identify breast cancer at an initial stage; some of it used machine learning algorithms to predict recurrence; and others attempted to enhance the performance of learning models using feature selection methods. An overview of such works is briefly discussed in this section.

Ahmad et al., [7], proposed an GA-based feature selection along with an artificial neural network classifier and implemented it on three different backpropagation techniques: resilient, Levenberg-Marquardt, and gradient descent with momentum. Resilient back propagation produces beat-correct classification.

Perze et al., [8], presented the importance of feature selection and used Mann-Whitney-Wilcoxon statistical test to discriminative features with SVM, LDA and feedforward back-propagation classifier on two different breast cancer dataset and achieved 83.9%, 83.5% and 89% accuracy.

Aalaei et al., [9], presented a wrapper approach using GA-based feature selection and Particle swam-classifier on different datasets: the Wisconsin breast cancer dataset (WBC) and the Wisconsin diagnosis breast cancer (WDBC), with 96.9 and 97.2 accuracy, respectively.

Alickovic et al., [10], proposed system has two stages. In the first stage, to eliminate insignificant features, genetic algorithms are used for feature extraction. Second, construct an accurate system for breast cancer classification using Rotation forests. The proposed model produces the best result with only 14 features.

Khourdifi et al., [11], suggested a prediction model that combines a Fast Correlation-based Feature Selection (FCBF) with classifiers like Random Forest, Naive Bayes, Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), and Multilayer Perception to provide the most effective classifier with and without FCBF. Finally, SVM without FCBF achieved 97.9% accuracy followed by MLP.

Gopal et al., [12], proposed a methodology to detect breast cancer at an early stage by using IOT with ML along with the PCM feature selection method and obtained the best accuracy by using MLP along with PCM, i.e., 98%.

Sakri et al, [13], measured the performance of Naive Bayes and k-NN for the breast cancer recurrence prediction model along with the particle swarm optimization feature selection method and obtained the accuracy of 81.3% and 75%, respectively.

Algherairy et al., [14], suggested a classification model based on Support Vector Machine, Logistic Regression, and XGBoost as well as four different feature selection algorithms, including the Forward Selection, Mutual Information, Pearson's Correlation, and Univariate ROC-AUC methods. and used SVM with correlation feature selection to reach the best result, which was 98%.

Punitha et al., [15], proposed the two proposed hybrid approaches were used in concurrent feature selection and parameter optimization of an ANN model including resilient back-propagation (HABC-RP and Hybrid ABC-RP), Levenberg Marquart (HABC-LM and Hybrid ABC-LM), and momentum-based gradient

descent (HABC-MGD and Hybrid ABC-GD) for parameter tuning of ANN on Wisconsin breast cancer dataset and achieved HABC-RP with 99.14% and 99.54% for Hybrid ABC.

1.2 Contribution

We propose breast cancer classification model for early diagnosis or detection, which is motivated by the limitations identified in previous techniques (as stated in Section 1.1). The suggested methodology comprises a step for feature selection based on frequent itemset mining along with classifiers on" Winsoncin Breast Cancer" dataset. The following are the article's main contributions:

- **To minimize the dimension of the collected data:** The appropriate proposed Frequent Feature-set Mining (FFM) algorithm is employed for identifying a relevant feature-subset from a feature-set.
- **To build breast cancer classification model:** The proposed model easily classify the data into its specific classes.

2. Proposed approach

Traditional methodology for breast cancer classification uses end-to-end classification architecture employing machine learning classifier. In end-to-end architecture, there was no feature selection phase. Typically, only redundant can be removed but now features that are weakly related to classes i.e irrelevant features can also be removed easily [16]. Feature selection step plays a vital role in making a classification model. As the data accumulates, the features also grow [17].

In this research paper, two different methodologies are designed one is a traditional model with traditional statistical evaluation parameters and another is a suggested approach that includes a feature selection algorithm with additional statistical parameters to enhance the accuracy of the traditional classification model. The Figure. 1 shows the traditional method of classification model and the Figure. 2 shows our suggested proposed methodology.

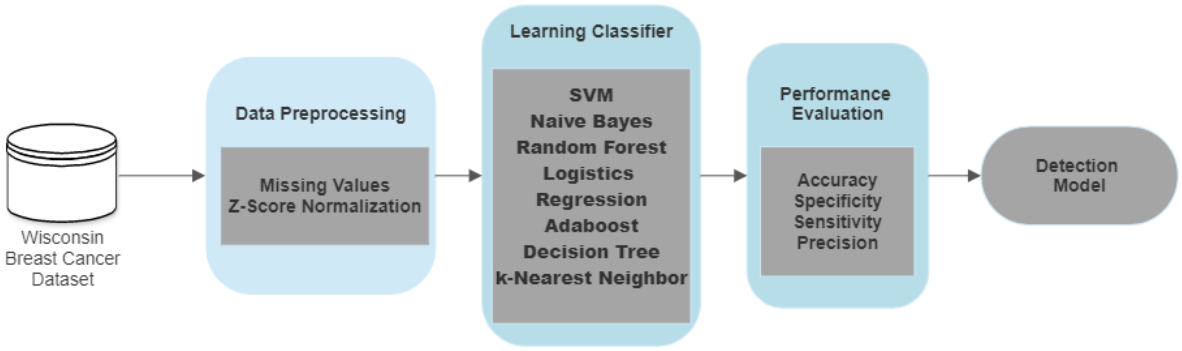

Figure 1. Traditional classification Model

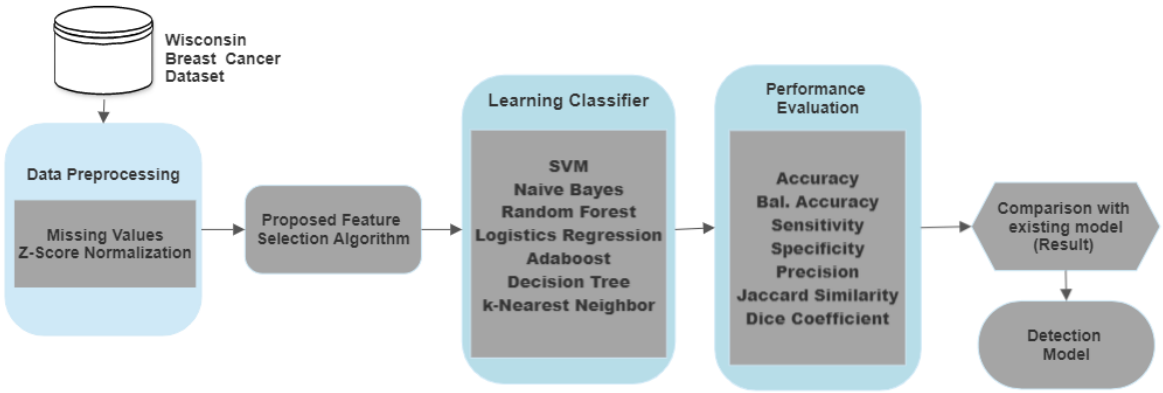

Figure 2. Proposed Model For Breast Cancer Classification

In the proposed methodology Figure 2, the proposed feature-set mining algorithm based on the frequent itemset mining concept is implied for the selection of relevant feature subsets to aid in the early identification, prediction, or diagnosis of malignant tumours. In this proposed algorithm only relevant features are extracted as a subset for the classification step. Each step of the suggested methodology is covered in the subsections that follow.

2.1 Dataset Acquisition

Firstly, the dataset is collected from Kaggle's "WBCD Dataset" (https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data) (Wisconsin Breast Cancer Dataset). It consists of 569 instances with 31 features and 1 binary class label for accurate evaluation. The class variable is either benign (357 instances) or malignant (212 instances). For evaluation, considered 70% of the instances for training and the remaining 30% for testing purposes. Features of WBCD are discussed in the description table, shown in Table 1.

Table 1. Dataset Description

Features	Description
Radius	Averaging the lengths of the radial segments of a line defined by the cell's centroid and individual cell points.
Perimeter	The nuclei perimeter of tumor cell is calculated by adding the total distance between the cell locations.
Area	Calculated by adding 1/2 of the pixels in the perimeter to the number of pixels in the cell's interior.
Compactness	Measured using the formula perimeter2/area. A circular disc reduces the dimensionless number while increasing the irregularity of the cell boundary. We can compensate the fact of irregularity by measuring the shape features.
Smoothness	Calculated by subtracting the length of a radial line from the mean length of the surrounding lines.
Concavity	Draw chords connecting non-adjacent locations and measure the amount to which the actual boundary of the cell is inside each chord. Smaller chords capture minor concavities better than bigger ones, and they also capture irregularities. Count the number of concavities and their severity.
Concave Points	Similar to Concavity, but simply measures the number of concavities rather than their magnitude.

Symmetry	The major axis or longest chord via the center is found and measured. Measure the length difference between lines parallel to the major axis and perpendicular to the cancer cell boundary.
Fractal Dimension	Uses coastal approximation, which greatly increases the size of the 'rulers' while decreasing the precision value and, as a result, the observed perimeter.
Texture	The variation of grayscale intensities in the component pixels is calculated.

2.2 Data Pre-processing

Dataset pre-processing is one of the foremost steps in data analytics to control the data quality by removal of noise, handling missing values in clinical datasets, and normalizing datasets [18]. The format of the data must be suitable for Machine Learning. Sometimes, misleading data(always noise) can lead to bad accuracy in classification models. A classification model will be able to produce a relevant diagnosis since data transformation and feature extraction are employed to increase the performance of classifiers. Only features that are pertinent to the specific disease are chosen and extracted [19]. To gain good performance accuracy, we need a large dataset with a good number of features. Moreover, accuracy can deteriorate with fewer instances in the dataset or fewer features due to over-fitting. It means the learning model will perform well with test data but deteriorate with training data. To overcome this problem first and foremost step of pre-processing is handling missing values, which is '0' in this case so the next priority is a scaling of the dataset and the Z-score normalization technique is used to balance the features on a similar scale. Implies formula:

$$Z - Score = \frac{\chi - \mu}{\sigma} \quad (1)$$

where,

χ= Original data, μ= Mean Value, σ= Standard deviation of data

2.3 Feature Selection

We used frequent feature itemset mining (FFM) based on the frequent item set mining concept to extract the relevant features. Frequent feature set mining is a common algorithm used in data analysis to discover associations and correlations among features in a relational and large dataset [20]. The algorithm iteratively checks the relevant features by generating feature subsets and pruning those that do not meet certain criteria. Although this algorithm takes more time compared to other frequent item set mining algorithms, it gives an optimal solution and accurately performs feature connectivity testing with a low error rate. Each and every step of frequent feature set mining is explained in the proposed feature selection algorithm , refer Algorithm 1.

2.3.1 Preliminaries

1. **Principle:** If an featureset is frequent, then all of its sub featuresets must also be frequent. Principle is used to prune the search space and make the algorithm more efficient.
2. Let Database D = { T_1, T_2 . . . T_N } be a set of instance T , each instance is a set of feature T ⊆ I.
3. Features I = {i_1, i_2, . . . i_M } is a set of features.
4. Frequent itemset X = Set of selected Features X ⊆ I.
5. Instances T contains an featureset X: X ⊆ T
6. The support of an itemset X is defined as:

$$support\ (X) = \frac{T \in D}{X \subseteq T} \quad (2)$$

7. Frequent itemset: an itemset X is called frequent for database D iff it is contained in more than minSup many transactions: support(X) ≥ minSup

8. It consists of 659 instances and 32 features denoted as (T) to (I). The minSup threshold is initially set as 0.4.
9. Generate and Prune : Generate the X (Frequent itemset) by pruning irrelevent feature from I (Original Feature set).

Algorithm 1: Frequent feature-set mining

Goal: Given a database D and a threshold minSup , find all frequent featureset $X \epsilon I$

Input: Pre-processed Dataset

Output: Filtered Featureset(X)

1. Initialization: minSup = 0.4, & X (empty featureset)
2. Step 1: Calculate the support(X) equation 2 for each feature I.
   iF support(X) ≥ minSup
   Delete the feature I
   else
   if {Feature I_row ='1' for all T
   Include that feature I in X (empty featureset) and then delete the feature I
   else
   Check for each row (for those whose support(X) equation:2 ≤ minSup)
   If row ='0' for all feature I then
   delete the row
   endIf
   end if
   end if
3. {Step 2:} Calculate Support(X) for all feature I
   if support(X) equation:2 is unique for feature I
   include that feature I into X and delete the feature I and goto Step 1 and continue.
   else
   if Support(X) is not unique then
   Include all the col into X and go to STOP.
   end if
   end if
4. STOP
5. return X

2.4 Learning Classifiers

ML, a subset of Artificial Intelligence (AI), has expanded significantly in recent years in the context of statistical analysis and data computing, which often enables programs to perform intelligently. ML is typically referred to as the most well-liked new technology because it gives systems the ability to automatically learn from experience and improve on it [21,22]. Machine learning methods heavily rely on classification. Different forms of classification analysis have been applied in the proposed methodology. In

this research, we employ a variety of machine learning methods that can be used to improve the intelligence and capacities of an application based on the significance and potential of "Machine Learning" to analyze the WBCD such as Naive Bayes, Support Vector Machines, Decision Trees, Random Forests, Ada-Boost, Logistic Regression, and k-NN have been developed [22].

1. **Support Vector Machine:** Create a hyperplane that divides the input data into two distinct groups to classify the data [23,24]. Implied as:

$$f(x1, x2) = \frac{e^{-||(x1-x2)||^2}}{2\sigma^2}$$

(3)

where,

σ = variance and hyper-parameter, (x1,x2)= Euclidean distance between two points.

2. **Naive Bayes:** Able to predict, requires only a few training data to identify the required parameters, the probability distribution of a class label [25].

$$P(class|data) = \frac{P(data|class) * P(cla}{P(data)}$$

(4)

whereas,

P(class|data) = Prob of a class being true, given that data is true

P(data|class) = Prob of a data being, given class, is true

P(data) = Prob of data is true

P(class) = Prob of a class is being true

3. **Random Forest:** Build individual decision trees for training, consist of the number of features sampled, the number of trees, and the node size. Use each tree's predictions to create the final classification, and support bagging [26,27].

$$Gini : \sum_{i=1}^{c} * f_i(1 - f_i)$$

$$Entropy : \sum_{i=1}^{c} -f_i log(f_i)$$

(5)

where,

fi_i = is the frequency of class i at a node

c= no of unique class

4. **Logistic Regression:** Determine the likelihood of a specific class using independent variables [28]. Calculate the logistic by adding up the features [29].

$$h\Theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * X)}}$$

(6)

Whereas,

hΘ(X)= hypothesis of LR and (β0 +β1 ∗X)=linear regression

5. **Adaptive Boosting (Adaboost)** : Uses Ensemble Method i.e Boosting for classification, set of classifiers i.e multiple classifiers are considered to allocate the class label to instances [30].

$$H(x) = sign(\sum_{t=1}^{T} * \alpha_t * h_t(x))$$

(7)

whereas, ht(x) = weak classifier t for input x, αt= 0.5*ln((1-E)/E), classifier weight

6. **Decision Tree :** Non-Parametric classifier [31], open to no restrictions regarding the sharing of the input data [32]. Every node divides the quantity of samples in half before dividing the length of DT = $\log_2(n)$.

$$Number\ of\ nodes = \sum_{(i=1)}^{log_2(n)} * 2^i$$

(8)

7. k-NN : A distance-based approach, as the name implies, determines the distance from all points in the vicinity of the unknown data and eliminates those with the closest distances [33,34].

$$d(x,y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

(9)

where, (xi −yi)2 = Square of euclidean distance between two points.

2.5 Performance Evaluation

When attempting to solve classification issues, a confusion matrix is a highly common tool [35]. The error matrix is a matrix that displays how a classification model performed on a set of test data. In order to achieve a quantitative evaluation of the suggested methodology, some statistical performance evaluation indicators were applied in this study. We examined accuracy, balanced accuracy, sensitivity, and specificity as the primary statistical criteria for this classification model without and with proposed feature selection method. They are also described as:

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

$$Balanced\ Accuracy\ (Bal\ Acc)\ \frac{1}{N}\sum_{k=1}^{N} Recall\_k$$

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

where, N=2 (B,M)

$$Sensitivity\ (Sen) = \frac{TP}{TP + FN}$$

$$Specificity\ (Spe) = \frac{TN}{TN + FP}$$

Some additional performance metrics, mentioned in my research paper are:

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

$$Jaccard\ Similarity\ \ (JS) = \frac{TP}{TP + FP + FN}$$

$$Dice\ Coefficient\ (DC) = \frac{2TP}{2TP + FP + FN}$$

Table 2. Selected feature-subset by using Proposed Frequent Feature-set Mining Algorithm

	Selected Features
Proposed Algorithm (FFM)	radiusmean, areamean, smoothnessmean, symmetrymean, radiusse, concavityse, radiusworst, areaworst, concavityworst, symmetryworst, diagnosis(class-B, M)

Table 3. Performance Evaluation without proposed feature selection method

Classifier	TP	TN	FP	FN	Acc	Bal Acc	Sen	Spe	P	JS	DC
SVM	84	80	2	4	96.47	96.50	95.45	97.56	97.67	93.33	96.55
NB	85	80	0	5	97.05	97.22	94.44	1	1	94.44	97.14
RF	86	77	4	3	95.88	95.84	96.62	95.06	95.55	92.47	96.08
LR	84	77	3	6	94.70	94.79	93.33	96.25	96.55	90.32	94.91
Adaboost	88	71	2	9	93.52	89.73	90.72	88.75	97.77	88.88	94.11
DT	82	76	4	8	92.94	93.05	91.11	95	95.34	87.23	93.18
k-NN	82	74	5	9	91.76	91.88	90.10	93.67	94.25	85.41	92.13

Acc: Accuracy, Bal Acc: Balanced Accuracy. Sen: Sensitivity. Spe: Specificity. JS: Jaccard Similarity DC:Dice Coefficient. NB: Naive Bayes. SVM: Support Vector Machine. RF: random Forest. LR:Logistic Regression. Adaboost: Adaptive Boosting. DT: Decision Tree. k − NN: k-Nearest Neighbor.

ROC curve, [36], is an analytic test elimination graph that is plotted between sensitivity against the false positive rate i.e (1- specificity). The ROC curve also deliberates the relation between the True Positive rate and False Negative rate. The ROC curve works with all four confusion matrix matrices, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). For patients who are not in the benign stage, the value of TN is particularly high in the confusion matrix. If the testing model is good for benign stage cancer and our testing model includes more malignant patients, there is a defect in our classification model, which also has a high accuracy, refer Figure 3. The PR curve is the solution to this problem because it only deals with positive numbers i.e TP and FP. The PR curve stands for precision-recall Curve. In PR curve, [37], the precision is plotted against recall. Precision measures the number of individuals predicted by the learning classifiers as Benign in the case of total Benign. Whereas recall measures the likelihood that estimates Benign given all the examples whose correct class is Benign. PR curve for the breast cancer classification model without and with proposed feature selection method is shown Figure 4.

3. Result and Discussion

The proposed methodology has been implemented in Python 3.8. Some of the machine learning libraries are pandas, numpy, seaborn, matplotlib, sklearn.metrics, sklearn.feature_selection, sklearn.model_selection, sklearn.svm, sklearn.naive_bayes, sklearn.linear_model, sklearn.tree, sklearn.ensemble. Ran the experiments on a laptop with 11th Gen Intel(R) Core(TM) i3-1125G4 @ 2.00GHz processor equipped and 8GB RAM.

3.1 Analysis

In this proposed methodology, classification method analysis is based on a specific feature set that distinguishes between the benign and malignant classes using 10-fold cross-validation methods. The WBCD dataset consists of 569 data-instances; for this research, we used 170 data-instances, to test our classification models. At the initial stage, the proposed features selection algorithm(FFM) is analyzed and produced a split feature set into relevant feature subset and irrelevant feature subset.

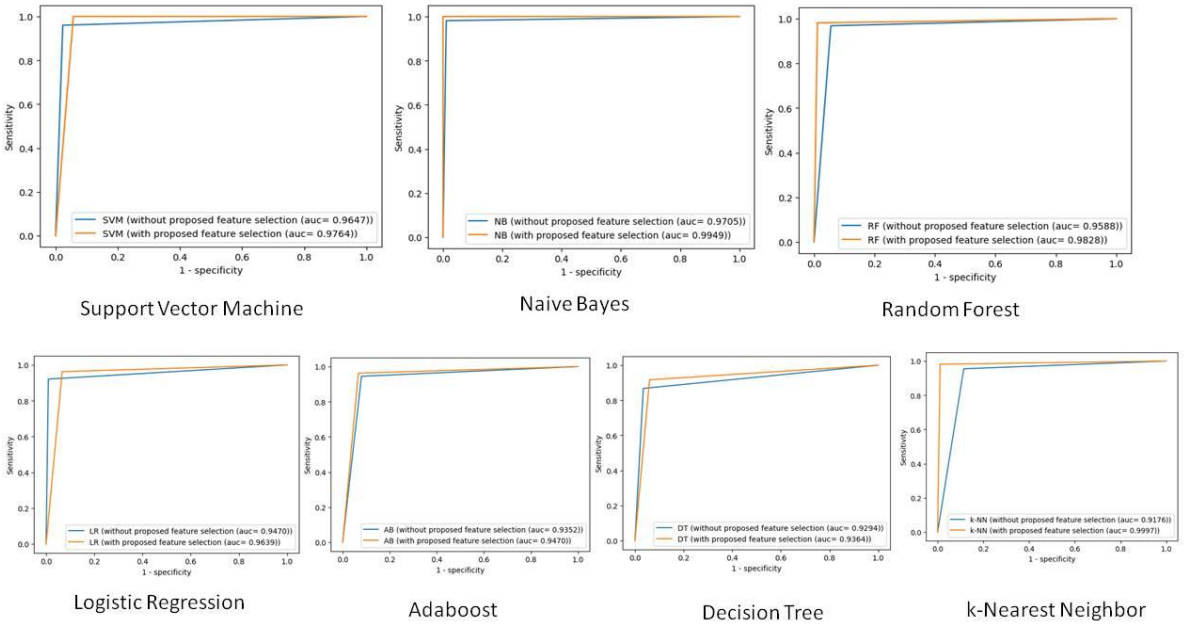

Figure 3. ROC Curve without and with Proposed Feature Selection

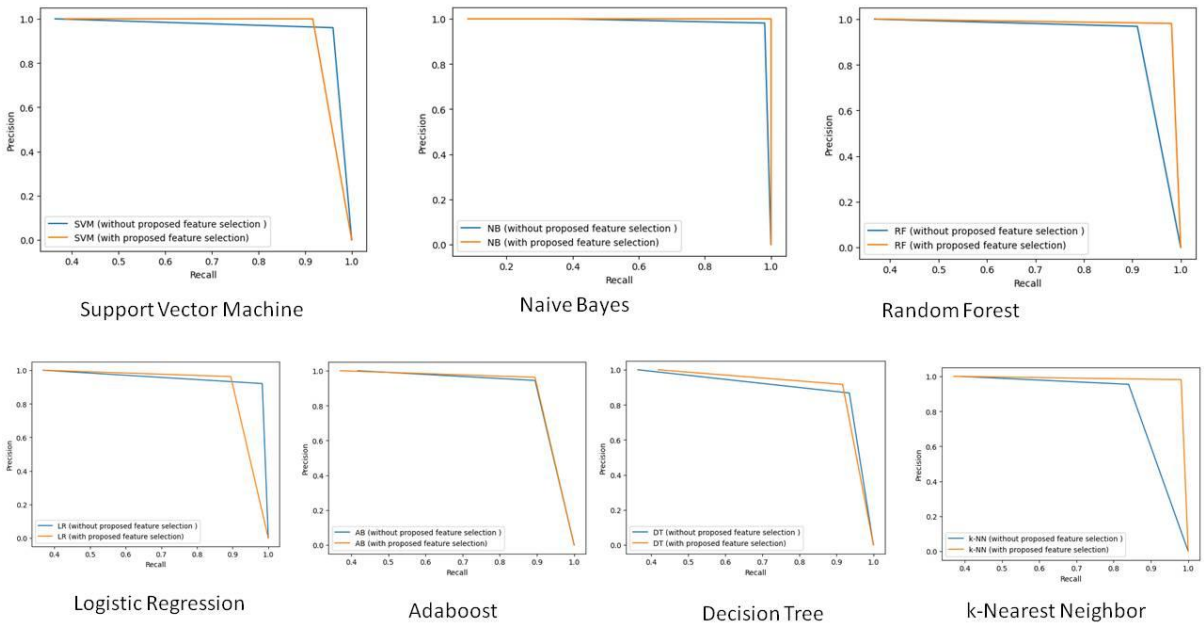

Figure 4. PR Curve without and with Proposed Feature Selection

The relevant feature subset with 11 features, as mentioned in the Table \ref{feature} are trained on different ML classifiers, for example, Naive Bayes, Support Vector Machine, Random Forest, Logistic Regression, Decision Tree, Ada-boosting, and k-NN on different statistical parameters to detect benign and malignant tumors demonstrated in Table 4 and lastly comparison is done with existing framework shown in the Table 5.

Table 4. Performance Evaluation with proposed feature selection method

Classifier	TP	TN	FP	FN	Acc	Bal Acc	Sen	Spe	P	JS	DC
SVM	96	70	2	2	97.64	97.58	97.95	97.22	97.95	96	97.9
NB	98	71	0	1	99.41	99.49	98.98	1	1	98	99.9
RF	97	70	1	2	98.2	98.28	97.97	98.59	98.59	97	98.4
LR	94	70	3	3	96.47	96.39	96.90	95.89	96.90	94	96.9
Adaboost	92	69	4	5	94.70	94.67	94.84	94.50	95.83	91	95.3
DT	91	68	4	7	93.52	93.64	92.85	94.44	95.78	89	94.3
k-NN	97	71	0	2	98.82	99.48	98.97	1	1	97	98.9

Table 5. Performance comparison: Impact of Frequent Feature-set Mining

Author	Dataset	Feature Selection	Classifier	Accuracy
Ahmad et al. [7]	WBCD	GA	ANN	98.99
Aalaei et al. [9	WBCD	GA	PS	96.9
Aalaei et al. [9	WDCB	GA	PS	97.2
Khourdifi et al. [11]	WBCD	Fast Correlation	SVM,MLP	95.17, 91.71
Anter et al. [38]	WBCD	CFCSA	Fuzzy c-mean	93.3
Naseem et al. [39]	WBCD	Stacking	DT, RF,LR, SVM, NB	91.22, 97.07, 98, 98.10, 93.56
Naseem et al. [39]	WBC	Stacking	DT, RF,LR, SVM, NB	76.26, 76.26, 75.24, 78.35, 70.71
Proposed Methodology	WBCD	FFM	SVM, NB, RF, LR, Adaboost, DT, k-NN	97.64, 99.41, 98.23, 96.47, 94.70, 93.52, 98.82

GA: Genetic Algorithm. WBCD: Wisconsin Breast Cancer Dataset WDBC: Wisconsin Diagnostic Breast Cancer. SVM: Support Vector Machine. CFCSA: crow search optimization algorithm integrated with chaos theory and fuzzy c-means algorithm. PCA: Pearson's Correlation. PSO: Particle Swarm Optimization. FFST: Fast finite shearlet transform

Naive Bayes predicts the largest number of True Benign (out of 170) among the five classifiers. Followed by k-NN classifier predicts the second largest True Benign (out of 170) and lowest False Malignant (out of 170).

The performance of our suggested classification model is compared with exiting methodology and also along with some existing work shown in Table 5. It is clear that our proposed methodology has the potential to improve the accuracy of learning models while maintaining a low error rate. Furthermore, in the case of a classification model, a feature selection phase is also important step. However, in order to give a superior classification model, the suggested feature selection approach necessitates slightly more storage and an extra stage than some of the existing methodologies.

4. Conclusion

A breast cancer classification model with proposed frequent feature-set mining algorithm for feature selection were developed to enhanced the performance of model with fewer features. Frequent feature-set mining, filter out irrelevant features that have a weak correlation to class (i.e., benign and malignant). In comparison to the existing classification model, the proposed frequent feature-set mining classification model produced performance outcomes that were somewhat improved. This research is beneficial for the pattern

recognition-based classification model for anomalies since it enables early patient diagnosis with a limited set of features.

In the future, we will be conducting an in-depth analysis of various datasets using a combination of ML and deep learning models. Additionally, a sizable dataset or a dataset of mammography images can be used to test the suggested strategy. To get effective outcomes, future research should focus on optimizing learning classifiers for specific applications in the healthcare sector.

REFERENCES

[1] Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., ... & Zhu, H. P. (2017). Risk factors and preventions of breast cancer. International journal of biological sciences, 13(11), 1387.

[2] El Atlas, N., El Aroussi, M., & Wahbi, M. (2014, November). Computer-aided breast cancer detection using mammograms: A review. In 2014 Second World Conference on Complex Systems (WCCS) (pp. 626-631). IEEE.

[3] Sayegh, H. E., Asdourian, M. S., Swaroop, M. N., Brunelle, C. L., Skolny, M. N., Salama, L., & Taghian, A. G. (2017). Diagnostic

methods, risk factors, prevention, and management of breast cancer-related lymphedema: past, present, and future directions. Current Breast Cancer Reports, 9, 111-121.

[4] Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018, April). ”Breast cancer classification using machine learning”. In 2018 electric electronics, computer science, biomedical engineerings’ meeting (EBBT) (pp. 1-4). IEEE.

[5] Zheng, J., Lin, D., Gao, Z., Wang, S., He, M., & Fan, J. (2020). Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis. IEEE Access, 8, 96946-96954.

[6] Soofi, A. A., & Awan, A. (2017). Classification techniques in machine learning: applications and issues. Journal of Basic & Applied Sciences, 13(1), 459-465.

[7] Ahmad, F., Mat Isa, N. A., Hussain, Z., Osman, M. K., & Sulaiman, S. N. (2015). A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer. Pattern Analysis and Applications, 18, 861-870.

[8] P´erez, N. P., L´opez, M. A. G., Silva, A., & Ramos, I. (2015). Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. Artificial intelligence in medicine, 63(1), 19-31.

[9] Aalaei, S., Shahraki, H., Rowhanimanesh, A., & Eslami, S. (2016). Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. Iranian journal of basic medical sciences, 19(5), 476.

[10] Aliˇckovi´c, E., & Subasi, A. (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and applications, 28, 753-763.

[11] Y. Khourdifi, M. Bahaj, Feature selection with fast correlation-based filter for breast cancer prediction and classification using machine learning algorithms,in: 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), IEEE, 2018, pp. 1–6.

[12] V. N. Gopal, F. Al-Turjman, R. Kumar, L. Anand, M. Rajesh, Feature selection and classification in breast cancer prediction using iot and machine learning, Measurement 178 (2021) 109442.

[13] Sakri, S. B., Rashid, N. B. A., & Zain, Z. M. (2018). Particle swarm optimization feature selection for breast cancer recurrence prediction. IEEE Access, 6, 29637-29647.

[14] Algherairy, A., Almattar, W., Bakri, E., & Albelali, S. (2022, March). The Impact of Feature Selection on Different Machine Learning Models for Breast Cancer Classification. In 2022 7th International

Conference on Data Science and Machine Learning Applications (CDMA) (pp. 91-96). IEEE.

[15] Punitha, S., Stephan, T., & Gandomi, A. H. (2022). A novel breast cancer diagnosis scheme with intelligent feature and parameter selections. Computer Methods and Programs in Biomedicine, 214, 106432.

[16] Gupta, Roopam. "Feature selection techniques and its importance in machine learning: A survey." 2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). IEEE, 2020.

[17] Wang XD, Chen RC, Yan F, et al. Fast adaptive K-means subspace clustering for high-dimensional data. IEEE Access. 2019;7:42639–51.

[18] Al-Jabery, Khalid, et al. Computational learning approaches to data analytics in biomedical applications. Academic Press, 2019.

[19] Singh, Pushpa, et al. "Diagnosing of disease using machine learning." Machine learning and the internet of medical things in healthcare. Academic Press, 2021. 89-111.

[20] Darrab, Sadeq, and Belgin Ergenc. "Vertical pattern mining algorithm for multiple support thresholds." Procedia computer science 112 (2017): 417-426.

[21] Sarker, Iqbal H., Md Hasan Furhad, and Raza Nowrozy. "Ai-driven cybersecurity: an overview, security intelligence modeling and research directions." SN Computer Science 2 (2021): 1-18.

[22] Sarker, Iqbal H. "Machine learning: Algorithms, real-world applications and research directions." SN computer science 2.3 (2021): 160.

[23] Keerthi, S. Sathiya, et al. "Improvements to Platt's SMO algorithm for SVM classifier design." Neural computation 13.3 (2001): 637-649.

[24] Lai, Zhihui, et al. "Maximal Margin Support Vector Machine for Feature Representation and Classification." IEEE Transactions on Cybernetics (2023).

[25] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

[26] Random Forest: https://www.ibm.com/topics/random-forest.

[27] Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.

[28] Nick, Todd G., and Kathleen M. Campbell. "Logistic regression." Topics in biostatistics (2007): 273-301.

[29] Gai, RongLi, and Hao Zhang. "Prediction model of agricultural water quality based on optimized logistic regression algorithm." EURASIP Journal on Advances in Signal Processing 2023.1 (2023): 21.

[30] Subasi, Abdulhamit, and Emir Kremic. "Comparison of adaboost with multiboosting for phishing website detection." Procedia Computer Science 168 (2020): 272-278.

[31] Quinlan, J. Ross. "Induction of decision trees." Machine learning 1 (1986): 81-106.

[32] Han, Xiaoyu, et al. "A three-way classification with fuzzy decision trees." Applied Soft Computing 132 (2023): 109788.

[33] Aha, David W. "Incremental constructive induction: An instance-based approach." Machine Learning Proceedings 1991. Morgan Kaufmann, 1991. 117-121.

[34] Javeed, A., Dallora, A. L., Berglund, J. S., Ali, A., Ali, L., & Anderberg, P. (2023). Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions. Journal of medical systems, 47(1), 17.

[35] Japkowicz, Nathalie, and Mohak Shah. "Performance evaluation in machine learning." Machine Learning in Radiation Oncology: Theory and Applications (2015): 41-56.

[36] Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test

assessment. Journal of Thoracic Oncology, 5(9), 1315-1316.

[37] Ozenne, B., Subtil, F., & Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. Journal of clinical epidemiology, 68(8), 855-859.

[38] Anter, Ahmed M., and Mumtaz Ali. ”Feature selection strategy based on hybrid crow search optimization algorithm integrated with chaos theory and fuzzy c-means algorithm for medical diagnosis problems.” Soft Computing 24.3 (2020): 1565-1584.

[39] Naseem, U., Rashid, J., Ali, L., Kim, J., Haq, Q. E. U., Awan, M. J., & Imran, M. (2022). An automatic detection of breast cancer diagnosis and prognosis based on machine learning using ensemble of classifiers. IEEE Access, 10, 78242-78252.