An Intelligent System for Predicting Breast Cancer (ISPBC) using a Novel Feature Selection Technique

Akhil Kumar Das[1], Saroj Kr. Biswas[2], Ardhendu Mandal[3], Arijit Bhattacharya[1,*] & Debasmita Saha[4]

[1]Department of Computer Science, Gour Mahavidyalaya, Mangalbari, Malda-732142, WB, India
[2]Department of Computer Science and Engineering, NIT Silchar, Assam-788010, India,
[3]Department of Computer Science and Technology, University of North Bengal, Darjeeling-734013, India
[4]Department of Computer Science, University of Gour Banga, Malda-732101, WB, India
*E-mail: barijit@hotmail.com

Abstract. Breast cancer (BC) is becoming a global epidemic, largely affecting women. Breast cancer cases keep climbing steadily. Thus, early detection technologies or systems that notify patients to this disease are essential. Individuals can start treatment for this life-threatening illness, so that patients may be cured or given longer lives. To achieve this, in this study, an expert intelligence system named Intelligent System for Predicting Breast Cancer (ISPBC) was developed. The proposed system utilizes an innovative feature selection technique known as Enriched Feature Set (EFS) in order to identify the most appropriate and significant features. The proposed EFS employs the advantages of heuristic search techniques and stochastic hill climbing to select the most significant and important features. The Decision Tree and Random Forest techniques are employed for breast cancer diagnosis, distinguishing between malignant and benign types. The suggested model's performance was evaluated by comparing measures such as accuracy, precision, and recall through the utilization of tenfold cross-validation. To measure the efficacy of the suggested model, ISPBC's performance was compared to that of base classifiers and models published in the literature. A maximum accuracy of 96.09% was attained by ISPBC according to the results.

Keywords: *breast cancer; enriched feature set; heuristic search techniques; intelligent system; random forest; stochastic hill climbing.*

1 Introduction

Breast cancer constitutes around 23% of the total cancer cases, rendering it a prevalent ailment among women on a global scale [1]. In rare cases, it sometimes also occurs in men, who account for roughly 0.5 to 1% of all BC cases. BC is not an infectious or transmissible disease. It is typically a critical illness for women between the ages of 40 and 50. Breast cancer (BC) is a malignant neoplasm that has metastasized from breast cells. A malignant tumor is characterized by the

Received April 26th, 2024, Revised October 15th, 2025, Accepted for publication November 24th, 2025.
 ISSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2025.19.2.2

presence of cancerous cells that possess the capacity to invade neighboring tissues and metastasize to far anatomical sites, including the brain, bones, and lungs [2].

In 2008, 1.38 million women were diagnosed with BC, representing 50% of all BC patients and about 60% of all deaths [3]. In 2012, there were 1.7 million new cases reported [4]. In 2013, around 232,340 women were identified as having BC, and among them 39,620 women died due to BC in the USA [3]. In 2015, India recorded approximately 156,100 new cases of BC and 76,000 women were expected to die from the disease according to the WHO [5]. In the USA, an estimated 316,100 new instances of breast cancer were reported, with an estimated 40,500 individuals projected to succumb to this ailment [6]. In the year 2018, 627,000 females died as a result of this devastating condition [7]. The ACS estimates that there were 3.1 million BC survivors in the USA. According to an ACS press release from 2019, invasive BC has been detected in 268,500 women and around 2,600 men, while 62,900 women had been diagnosed with non-invasive BC. In 2019, it was anticipated that 41,760 women and 500 men would die from BC [8]. According to the WHO, there will be 2.3 million females diagnosed with BC and 685,000 deaths worldwide in 2020. BC impacts around 255,000 women and 2,300 men annually in the USA. According to the CDC, breast cancer kills roughly 42,000 women and 500 men in the USA each year. Every year, 150,000 Indian women are diagnosed with BC, with 70,000 of them dying as a result, according to the ICMR. In the United Kingdom, 1 in every 12 females between the ages of 1 and 85 is diagnosed with breast cancer [9]. As to the WHO, 2.1 million women are impacted by BC annually. In 2021, around 284,200 women with BC were identified and 44,130 were expected to pass away from the disease [10]. Among women in the USA, it was projected that in 2022 there would be around 287,850 new instances of invasive BC and 51,400 cases of DCIS. Additionally, it was estimated that 43,250 people would succumb to BC [11].

As a result, BC is quickly becoming the most life-threatening disease in the world. BC cannot be avoided [12], but it can be cured if it is caught early enough, before it spreads to any other part of the body. If detected early and treated correctly, the mortality rate of BC will decrease. Finding preventative treatment is crucial given the severity of the life-threatening challenges that patients encounter. It is crucial to recognize BC early on in order to provide appropriate treatment, avoid complications, and lower BC mortality. Several studies have been conducted in order to develop an intelligent system for the prediction of BC using different methodologies like WNBC, (AR + NN), AdaBoost ELM model, and others. However, it is commonly observed that the majority of expert systems exhibit deficiencies in effectively managing data preprocessing and systematically select features.

In order to surmount these constraints, this research paper describes an intelligent system named Intelligent System for Predicting Breast Cancer (ISPBC) using a novel feature selection technique to diagnose BC based on symptomatic aspects. In the proposed ISPBC system, the EFS feature selection approach is utilized to identify the most pertinent features inside a BC data set. To get efficient features, EFS uses a heuristic search technique (HST) and stochastic hill climbing (SHC). HST approaches have the advantages of greater efficiency and effectiveness, reduced time complexity by reducing the search space to find an optimal solution. The SHC method makes the whole search space more likely to be explored and raises the chance of escaping local optima and discovering more relevant responses.

This study evaluated the accuracy of different stand-alone and ensemble machine learning algorithms, as well as numerous models found in the literature by comparing them to the proposed system based on a BC data set. To achieve optimal predictions, a tenfold cross-validation procedure was applied to validate the model. Hence, the proposed model offers a precise breast cancer detection system. According to the evaluation, the ISPBC demonstrated a peak accuracy of 96.09% when employed in conjunction with several single-classifier models, ensemble models, and models derived from the literature. The execution of the ISPBC system was also compared to the accuracy, precision, and recall of simple Decision Tree (DT) and simple Random Forest (RF).

The rest of this paper is organized as follows: Section 2 introduces related works in breast cancer prediction and their methodology. Section 3 describes the proposed model for breast cancer prediction. In Section 4, the data set description and data preprocessing methods used in this study are provided. Section 5 discusses the proposed feature selection method, the results, and appropriate commentary based on the models. Section 6 follows with the conclusion.

2 Literature Review

Numerous recent studies have been undertaken with the aim of forecasting the occurrence of breast cancer. Even though many researchers have worked on this topic using ML algorithms, this section summarizes previous research on BC diagnosis.

Dai *et al.* [13] discusses diagnosing of BC using RF. To achieve high prediction accuracy, the RF approach incorporates several eigenvalue features as well as the outputs of many DTs. The researchers used BC data from the UCI Repository with 569 instances. From the experimental analysis, they acquired a 95.56% prediction accuracy for BC. They also determined specificity, sensitivity, and precision. Gupta *et al.* [14] investigated various ML techniques for BC prediction,

including K-NN, LR, DT, RF, and SVM, using a radial basis function kernel. They used a BC data set from the UCI Machine Learning Repository and compared the results of the various techniques. Deep Learning with ANN achieved the highest level of accuracy with a score of 98.9 percent. The best result came from Adam's gradient descent learning, which seeks to minimize errors while also training data as efficiently as possible. Kabiraj *et al.* [15] presents two ML algorithms, RF and XGBoost, to detect BC using a BC data set with 275 instances and 12 features. They compared the results in terms of accuracy, sensitivity, precision, F1-Score, and specificity to the mention classifiers. From the experimental analysis, they got a 74.73% prediction accuracy for BC using RF and a 73.63% accuracy using XGBoost. Aroef *et al.* [16] studied the classification of BC using RF and SVM. In addition, hold-out validation was utilized to validate and determine the performance of the abovementioned models. According to the results, RF achieved 90.9% accuracy and SVM achieved 95.4% accuracy. Therefore, SVM gave better results than RF.

Wang *et al.* [17] proposed the Improved Random Forest-based Rule Extraction (IRFRE) technique for diagnosing BC. This method uses a DT ensemble to develop precise and comprehensible classification rules for the diagnosis of BC. Three BC data sets were analyzed to assess method accuracy and interpretability. The empirical findings demonstrated that the IRFRE technique surpassed several widely used individual techniques, ensemble learning techniques, and rule extraction techniques in terms of precision and comprehensibility, thereby significantly enhancing the performance of cancer detection. Bharati *et al.* [18] used a variety of classification algorithms to detect BC, including NB, RF, LR, MLP, and K-NN. For this purpose, they used the WEKA data mining tool. They obtained a BC data set from the UCI machine learning library with 286 instances. The BC data set was explored in terms of Kappa statistics, FP rate, TP rate, and precision. The incidence of BC was predicted using a variety of approaches, and the results of each technique were compared. The performance of the K-NN classifier algorithm was 97.9021, which was the highest number of correctly classified items. Montazeri *et al.* [19] recommend a model named Trees Random Forest for the prediction of various types of BC survival using different machine learning methods. They also used a rule-based classification approach for this purpose. When compared to other methodologies, the TRF technique produced better outcomes in the investigation, with a 96% accuracy rate. Octaviani *et al.* [20] discusses breast cancer prediction using a Random Forest ensemble learning method. The BC data was taken from the UCI repository. The result of this experiment was more than 99% accuracy.

Table 1 Accuracies of previous works using Decision Tree and their limitations.

Ref.	Year	Methodology	Limitation	Accuracy of DT (%)
[21]	2015	**DT + SVM** model	No proper preprocessing and no mention feature selection technique	91
[22]	2016	NN, **DT(J48**), ANN, SVM, etc.	No proper data preprocessing method and does not check for outliers in the BC data set.	94.56
[23]	2017	**J48**, RF, Random tree, REP Tree Priority based	Only missing value handling.	95.43
[24]	2017	**DT**	No proper preprocessing.	94.3
[25]	2018	NB, RBFN and **J48**	Only missing value handling.	93.41
[26]	2018	**DT** and ANN	No preprocessing.	94.0
[27]	2019	K-NN, SVM, **DT**, RF, and MLP.	No proper data preprocessing method and does not check for outliers in the BC data set.	92.85
[28]	2020	NB, **J48**, RF, SMO and MLP	No feature selection and data preprocessing	94.27
[29]	2020	BN, SVM, **DT(J48)**, LG, RF, MLP	No proper data preprocessing method.	94.99
[30]	2022	RF, SVM, MLP and **DT**	Only missing data handling	93.41
[31]	2022	**J48**, NB, LR, SVM and KNN	No proper data preprocessing method and does not check for outliers in the BC data set.	92
[32]	2022	SVM, LR, **DT**, RF and K-NN	Only missing value handling.	94.29
[33]	2022	ESBCP system	Does not consider the overfitting problem.	94.01

Table 2 Accuracies of previous works using Random Forest and their limitations.

Ref.	Year	Methodology	Limitation	Accuracy of RF (%)
[34]	2017	LR, DT and **RF**	No proper data preprocessing method and does not check for outliers in the BC data set.	88.14
[35]	2018	DT, **RF**, SVM, NN and LR.	No proper preprocessing and no mention feature selection technique.	96.1
[36]	2019	PCA+RF, **RF**, KNN, NB, ANN and PCA+ANN	Does not check for outliers in the BC data set.	95.0
[37]	2019	**RF**, Randon Tree, NB,etc	No proper data preprocessing method and does not check for outliers in the BC data set.	96. 63
[38]	2019	K-NN, SVM, DT, **RF**, and MLP.	No proper data preprocessing method and does not check for outliers in the BC data set.	96.42
[28]	2020	NB, J48, **RF**, SMO and MLP	No feature selection and data preprocessing.	95.56
[29]	2020	BN, SVM, DT(J48), LG, **RF**, MLP	No proper data preprocessing method.	95.57
[39]	2020	KNN, **RF,** ANN, SVM, and LR	Only missing value handling.	95.71
[40]	2020	**RF**, SVM, KNN and LR	No proper data preprocessing.	91.66
[41]	2021	fuzzy-ID3 (FID3) model	No missing value handling and tenfold CV.	94.36
[42]	2022	LR, DT, **RF**, KNN and NB	No proper data preprocessing method and does not check for outliers in the BC data set.	95.32
[32]	2022	SVM, LR, DT, **RF** and K-NN	Only missing value handling.	93.81

Research Gap: Although numerous researchers have used DT and RF methods to predict breast cancer, there is always room for improvement to make them more accurate. To improve BC prediction accuracy, the suggested work developed a unique feature selection technique named Enriched Feature Set (EFS).

This study concentrated on three main points:

1. Using two tree-based classifiers – a decision tree classifier and a random forest classifier.
2. An innovative and effective approach of the feature selection model – EFS.

3. A novel and efficient breast cancer prediction expert system – ISPBC.

3 Methodology

ISPBC is divided into three stages: a) data preparation; b) feature selection (FS); and c) classification. The term 'raw data' refers to a jumbled collection of several types of information. The BC data set is typically insufficient, inconsistent, lacking in specific patterns, and prone to various errors. The irrelevant and unneeded features are deleted during this data preparation phase in order to generate a data set with optimal features for BC prediction. Therefore, the raw BC data set is translated into a suitable and understandable format that can be easily understood. Feature selection is the process of picking a subset of relevant features. The next stage, i.e., feature selection, is used to select the best features from the BC data set. Classification is a basic task in data mining that has been extensively researched in statistics, machine learning, neural networks, and expert systems over the years. Here, DT and RF employ the proposed system to detect BC. The proposed ISPBC is depicted schematically in Figure 1.

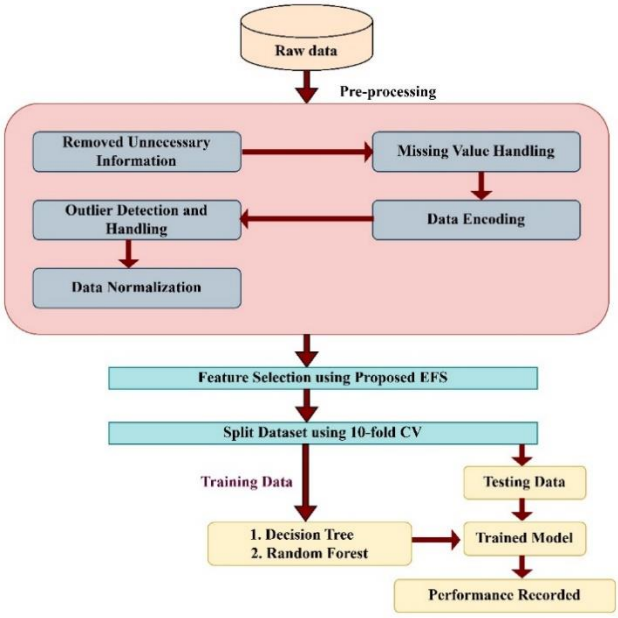

Figure 1 Intelligent System for Predicting Breast Cancer (ISPBC).

4 Data Description

In this study, we used the publicly available breast cancer data set created by Dr. W. H. Wolberg of the University of Wisconsin, which was obtained from the UCI

ML repository [43]. It has 699 cases derived from fine-needle aspiration articulations of human breast tissue. The data set contained 458 and 241 benign and malignant cases, respectively. Because 16 instances of the data set contained missing information, we employed 683 examples in our experiment, with 444 and 239 instances belonging to the 'benign' or 'not harmful' and 'malignant' or 'may be dangerous' classes, respectively. Every instance had eleven attributes, as shown in Table 3.

Table 3 Raw Data Description.

SL NO	id	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
0	1000025	5	1	1	1	2	1	3	1	1	2
1	1002945	5	4	4	5	7	10	3	2	1	2
2	1015425	3	1	1	1	2	2	3	1	1	2
3	1016277	6	8	8	1	3	4	3	7	1	2
4	1017023	4	1	1	3	2	1	3	1	1	2
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
694	776715	3	1	1	1	3	2	1	1	1	0
695	841769	2	1	1	1	2	1	1	1	1	0
696	888820	5	10	10	3	7	3	8	10	2	1
697	897471	4	8	6	4	3	4	10	6	1	1
698	897471	4	8	8	5	4	5	10	4	1	1

Figure 2 describes the features' correlations with each other in the BC data set.

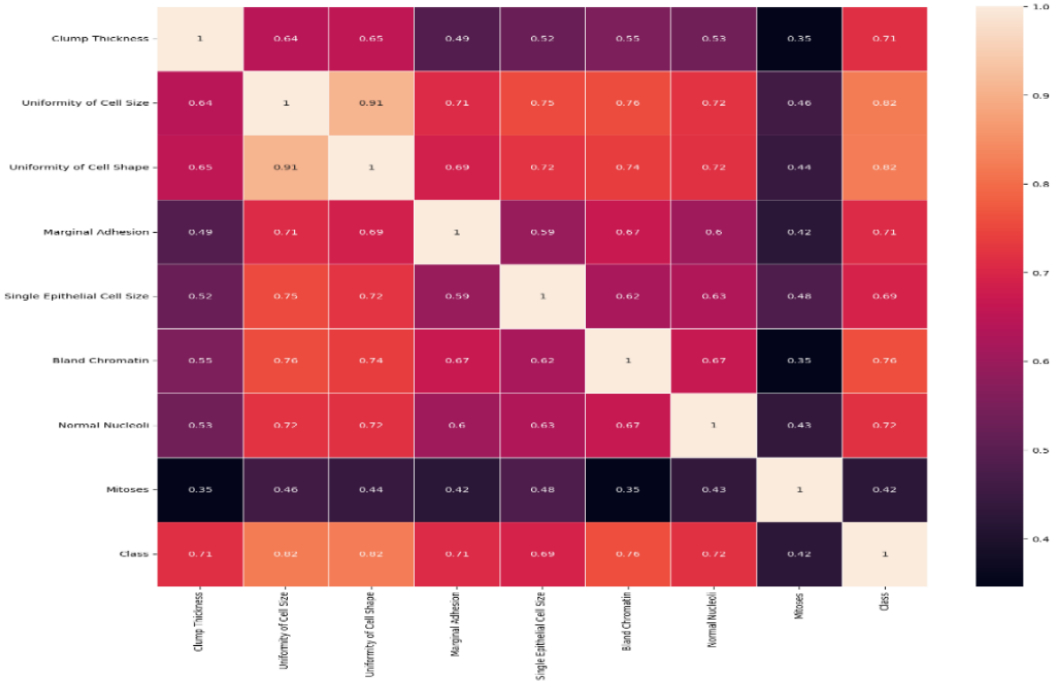

Figure 2 Heatmap of BC data set.

4.1 Data preprocessing

Data preparation is essential, because it transforms the data set into a usable format that the method can handle. It contains the following sub-phases.

4.1.1 Remove Unnecessary Information

One attribute was an extraneous feature. In order to generate homogeneous data collection, the item 'id' was removed in this stage.

4.1.2 Missing Data Handling

The collected BC data sets from the UCI repository contained a number of features to represent the data set. In the BC data sets, the bare nuclei attribute had 16 rows, with missing values denoted by '?'. There are several methods for dealing with missing values, such as imputation with the mean, mode, and so on. We removed these from the data set for simplicity. After removing them, the final data set contained 683 records, 444 of which were classified 'benign' and 239 of which were classiefied 'malignant'. The following table shows the data distribution of after deletion.

4.1.3 Data Encoding

Because the raw data of breast cancer (BC) consisted of 11 attributes, there was only one object data type among the most basic nuclei. To facilitate processing, the feature was encoded using label encoding, resulting in six labels, ranging from 0 to 6. The attribute class under consideration was classified into two distinct categories, namely benign ('2') and malignant ('4'). Furthermore, the encoding process involved assigning a value of 0 to benign tumors and a value of 1 to malignant tumors.

4.1.4 Outlier Detection and Handling

Outliers are elements that cause difficulties for learning and prediction. Detection and removal of outliers present in the data set is a challenging issue [44]. In our work, we utilized the z-score to find outliers in the data set. An absolute value of the z-score of less than 3 was taken into account; 73 records were identified as outliers.

4.1.5 Data Normalization

The features in the breast cancer data sets were converted in such a way that each characteristic contributed equally. This was mostly done to organize and analyze large amounts of data. It also converted the data from one format to another to enable further processing in this stage. For standardization to work, all of the

input variables are adjusted independently by taking the standard deviation and subtracting the mean. This changes the distribution so that the standard deviation is one and the mean is zero.

4.2 Feature Selection

In this study, we present EFS to pick the most important BC data set features. This method uses heuristic search and stochastic hill climbing to choose the most important and promising features. A fitness proportionate selection approach (fitness function) eliminates insignificant features. Here, the search for features was carried out using a fixed number of iterations. The proposed algorithm uses total accuracy as its scoring function and modified DT is utilized as the learning model. Every feature in every data set is categorized according to the number of classes it belongs to and its distance from the centroid of the cluster determines its score. It is the value of this score that is utilized by the fitness proportionate selection procedure (FS). The modification is retained if the improved learning model's overall performance is a result of the additional subset of features. If it does not, it is ignored and another neighbor in the feature space takes its place. The largest number of predetermined iterations is used to execute this procedure. The EFS feature selection approach is shown schematically in Figure 3.

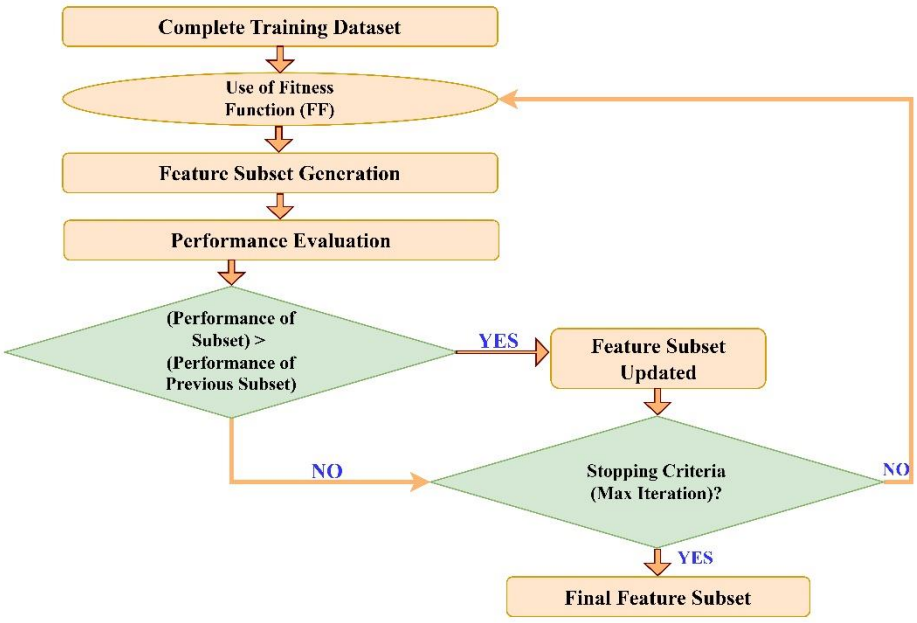

Figure 3 Enriched Feature Set (EFS).

The algorithmic form of the feature selection (FS) technique is shown below.

EFS algorithm for FS

Input: Initial set feature
Output: Feature subset
Notations:

$IFS \rightarrow initial\ set\ of\ features$
$m \rightarrow total\ samples$
$K \rightarrow number\ of\ classes$
$Min \rightarrow minimum\ feature\ value$
$Max \rightarrow maximum\ feature\ value$
$Dist \rightarrow distance\ between\ the\ feature\ and\ the\ cluster's\ centroid$
$SF \rightarrow selected\ feature$
$P_{ac} \rightarrow performance\ of\ the\ DT\ for\ IFS$
$P_{acI} \rightarrow performance\ of\ the\ DT\ for\ the\ preferred\ feature$
$FF \rightarrow final\ subset\ of\ features$
Procedure:

$\boldsymbol{S1}$: $Determine\ the\ classification\ accuracy\ \ (P_{ac})\ \ of\ a\ Decision\ Tree\ (DT)\ considering\ IFS.$
$\boldsymbol{S2}$: $Calculate\ the\ score\_value\ for\ each\ feature$
$\boldsymbol{S2.1}$: $Select\ centroid\ of\ K\ clusters\ randomly\ (initially)$
$Then\ select\ a\ random\ number\ within\ (Min., Max.)$
$\boldsymbol{S2.2}$: $For\ i\ to\ IFS$

$$dist = \sqrt{\sum_{k=1}^{m}(x_{ik} - y_{ik})^2}$$

$Assign\ i-th\ feature\ to\ the\ cluster\ with\ minimum\ \boldsymbol{Dist}\ value$
$\boldsymbol{S2.3}$: $Update\ the\ centroid\ of\ K\ clusters$

$$Centroid = \frac{\sum\ mean\ values\ of\ the\ features\ \ present\ in\ the\ cluster}{number\ of\ features\ in\ the\ cluster}$$

$\boldsymbol{S2.4}$: $For\ i\ to\ IFS$
$score\_value\ (i) = distance\ of\ i\ to\ the\ centroid\ of\ its\ cluster.$
$\boldsymbol{S3}$: $Repeat\ EFS\ for\ max\ iteration$
$\boldsymbol{S3.1}$: $SF = EFS\ (IFS, score\_value)$
S3.2: if $P_{acI} > P_{ac}$
$FS = SF$
$P_{ac} = P_{acI}$
$IFS = SF$
Return FF

4.3 Classification

RF is one of the most prominent supervised ML methods. It may be utilized for both classification and regression methods in solving problems. It is based on the ensemble learning technique, which integrates several classifiers to solve the problem and also increases the model's performance. It collects predictions from each DT and predicts the final output based on the majority of votes. In a RF, each DT is increased by utilizing a bootstrap sample of the training data. As a result, some cases are not utilized in the method of increasing a tree. This is called out-of-bag (OOB) and is used for evaluating variable importance and predictive performance [45-47].

The variable importance or feature importance of a decision tree can be determined by examining the prediction accuracy of the tree before and after random permutations of the actual feature. Variable importance or feature importance is the capacity to measure the significance of explanatory variables in prediction over all the decision trees in RF. The significance of explanatory variables is determined by the decrease in predictive accuracy when their values are randomly permuted. In this way, RF gives a more accurate and stable prediction.

Thus, an RF classifier is basically just a group of DTs chosen at random from the training set; the final prediction is then derived from the sum of all DTs' votes. However, during the data preparation process, data sets may include some irrelevant features that decrease the DTs' performance in building the RF.

5 Experimental Result and Discussion

Python is used to test the proposed intelligence expert model, ISPBC, in a Windows environment. The suggested model employs the proposed EFS feature selection method. Tenfold CV was used to validate the suggested model. It used a total of 683 estimators to analyze the data sets for BC prediction.

We used EFS in this work to select the most significant and promising features by removing irrelevant and redundant features from the existing feature set. To eliminate a feature, a fitness proportionate selection technique was used, with a probability of selecting a feature based on its score value. To validate the findings, the proposed approach was compared to simple DT and simple RF. The comparisons were based on accuracy, precision, and recall. A confusion matrix was used to describe the performance of the ISPBC model in this case. Table 4 shows the confusion matrix.

Table 4 Confusion Matrix.

		Predicted: NO	Predicted: YES
Actual:	**NO**	True Negative (Tn)	False Positive (Fp)
	YES	False Negative (Fn)	True Positive (Tp)

This ISPBC model was used to determine the accuracy, precision and recall [51-53] using Eqs. (1) to (3):

$$\text{Accuracy} = (Tp + Tn) / (Tp + FP + Tn + Fn) \quad (1)$$

$$\text{Precision(Pr)} = Tp / (Tp + Fp) \quad (2)$$

$$\text{Recall(Rc)} = Tp / (Tp + Fn) \quad (3)$$

Table 5 shows the results for the ISPBC system's accuracy tested with DT and RF. The ISPBC system is contrasted with the simple DT and RF methods. Compared to their simple version, the suggested ISPBC with DT and RF exhibited a greater classification accuracy. This is due to the fact that the proposed ISPBC improves classifier performance by handling missing values and removing outliers. The suggested ISPBC with RF outperformed the competition in terms of accuracy, according to further analyses. According to the results, for the breast cancer data set, the accuracy of ISPBC was 2.43%, i.e., 0.94% more accurate than that of the simple DT and simple RF, proving ISPBC's capacity to improve the performance of the proposed model. For this experiment, ISPBC was found to be more accurate than simple DT and simple RF, respectively.

Table 5 Performance Comparison (in %).

Metric	Simple DT	Simple RF	ISPBC using DT	ISPBC using RF
Accuracy	93.17	95.12	95.60	**96.09**
Precision	94	95	94	**96**
Recall	92	94	92	**95**

Figure 4 is a visual representation of the accuracy comparison among the suggested ISPBC, simple DT and RF, which helps with viewing and understanding.

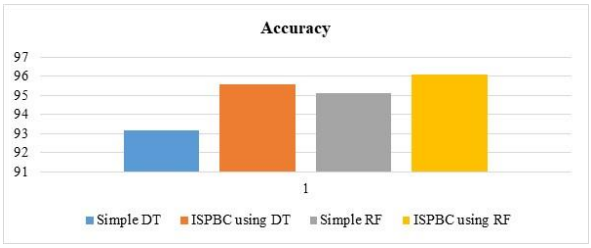

Figure 4 Accuracy comparison graph for BC.

Table 5 displays the precision of the ISPBC model compared with simple DT and simple RF. According to the results, ISPBC's precision for the BC data set was greater than that of simple RF but the same as that of simple DT. Although higher precision improves the effectiveness of the suggested model, there are still cases where actual positive malignant cases were predicted incorrectly. Enhanced precision performance demonstrates a reduced occurrence of false positives and a high level of accuracy in predictions.

Figure 5 is a visual representation of the precision comparison between the proposed ISPBC, simple DT and RF, which helps with viewing and understanding.

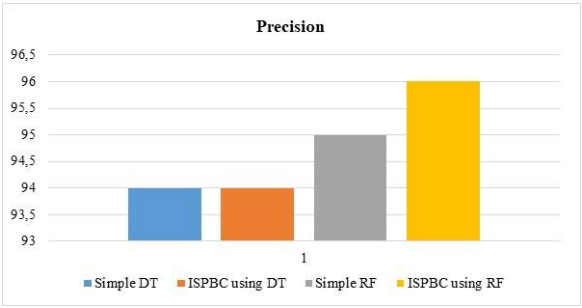

Figure 5 Precision comparison graph for BC.

Better performance in terms of recall illustrates that the number of false negatives is comparatively lower and the prediction is almost fully accurate. Table 5 displays the recall of the ISPBC system. Recall was taken into account to enhance this work. According to the results for the BC data set, the recall of ISPBC was 3% more accurate than that of simple RF, but the same as that of simple DT, proving ISPBC's capacity to improve the performance of the proposed model.

In order to better grasp the performance of the proposed ISPBC, simple DT, and RF, Figure 6 provides a graphic representation of the comparison of their recall.

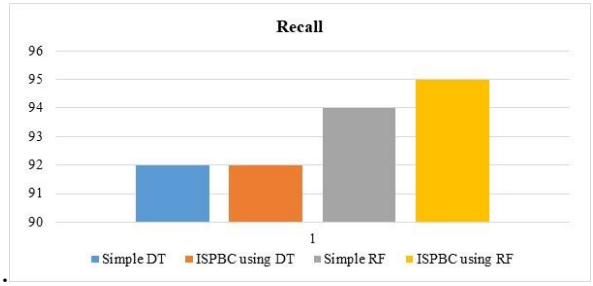

Figure 6 Recall comparison graph for BC.

Table 6 presents a performance comparison between the proposed ISPBC**,** single-classifier, ensemble models and other models in the literature.

Table 6 Comparison with previous works.

Reference	Methodology single-classifier models	Accuracy (%)
[22]	J48	94.56
[24]	DT	94.30
[25]	J48	93.41
[26]	DT	94.0
[27]	DT	92.85
[30]	DT	93.41
[31]	J48	92.00
[48]	LR	88.14
	Simple DT	93.17
Accuracy of the proposed Model ISPBC with DT is 95.60%, that is, better than the above works.		
Ensemble models		
[28]	RF	95.56
[29]	RF	95.57
[32]	RF	93.81
[34]	RF	88.14
[36]	RF	95.00
[39]	RF	95.71
[40]	RF	91.66
[41]	RF	94.36
Various models present in literature		
[21]	DT + SVM model	91.00
[33]	ESBCP system	94.01
[49]	WNBC	92
[50]	Firefly Algorithm based Expert System	94.81
	Simple RF	95.12
Accuracy of proposed Model ISPBC with RF is 96.09%, that is, better than the above works.		

Table 8 shows that when ISPBC was compared to the following models: (DT + SVM) model [21], ESBCP [33], WNBC [49], and Firefly Algorithm-based Expert System models [50], the proposed ISPBC model stood head and shoulders above the others.

Also when compared to several single-classifier and ensembles of models, the proposed ISPBC model outperformed them all in terms of accuracy. These models included single-classifier models ([22],[24-27],[30-31],[48]) and ensemble models ([28-29],[32],[34],[36],[39], [40-41]).

Evaluating the performance in relation to current cutting-edge models, it was seen that the prior models did not properly employ feature selection and data preprocessing. So, from Table 6, it can be seen that the proposed model ISPBC performed better in terms of accuracy than the single-classifier-based models, ensemble models, and various models present in the literature.

A visual representation of the accuracy of the proposed ISPBC compared to models found in the literature is shown in Figure 7.

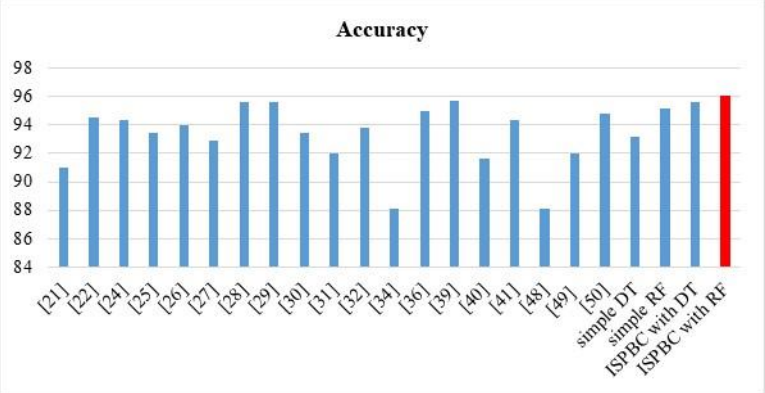

Figure 7 ISPBC performance compared to SOTA models.

6 Conclusion

Breast cancer is a life-threatening disease that has exploded into a global epidemic in recent decades. As a result, early detection and treatment of BC are critical. Even though all features of BC are not required for BC prediction, the proposed model employs DT, RF, and the EFS algorithm, demonstrating that such a lazy learning approach outperforms RF and DT, as shown above in Table 1 and 2. The Enriched Feature Set (EFS) method was developed to select the most significant features. The data preprocessing phase in the proposed model (ISPBC) takes the raw BC data set and preprocesses it by removing irrelevant features using Remove Superfluous Information, Missing Data Handling, and Normalized Data stages. Tenfold CV was used to validate the suggested model. Once the model had been trained, it was tested, and the results showed that the accuracy of ISPBC was superior to that of simple DT and simple RF by 2.43%, and 0.94%, respectively. Additionally, the recall and precision performance metrics were used to confirm the expert model's performance and it was further compared to simple DT, simple RF, and the proposed model ISPBC. It was found that the recall and precision performances of ISPBC were better than those of simple DT and simple RF, respectively. ISPBC uses symptomatic features to diagnose breast

cancer, saving time and money while also detecting breast cancer at an early stage. Because the ISPBC model outperforms simple DT, and simple RF, it can be summarized as a substantial, user-satisfying intelligent system for detecting BC early on.

The ISPBC model's performance may be improved in the future by training it on a larger set of data and then using various preprocessing approaches to remove all irrelevant and superfluous data.

Abbreviations

ACS: American Cancer Society; **ANN:** Artificial Neural Network; **ESBCP:** Expert System for Breast Cancer Prediction; **FP:** False Positives; **FS:** Fitness Function; **HST:** Heuristic Search Technique ; **ICMR:** Indian Council for Medical Research; **IRFRE:** Improved Random Forest-based Rule Extraction; **ISPBC:** Intelligent System for Breast Cancer Prediction; **K-NN:** K-Nearest Neighbors; **LR:** Logistic Regression; **MLP:** Multi-Layer Perceptron; **NB:** Naïve Bayesian; **PCA:** Principal Component Analysis; **PID:** Patient Identification Phase; **RAFN:** Radial basis function network; **REP Tree**: Reduced Error Pruning Tree; **RF:** Random Forest; **SHC:** Stochastic Hill Climbing; **SVM:** Support Vector Machine; **TP:** True Positives; **TRF:** Tree Random Forest; United Kingdom; **USA:** United States of America; **UCI:** University of California Irvine; **WHO:** World Health Organization; **WNBC:** Weighted Naive Bayes Classifier; **XGBoost:** Extreme Gradient Boosting;

Acknowledgement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E. & Forman, D., *Global Cancer Statistics*, CA: A Cancer Journal for Clinicians, **61**(2), pp. 69-90, 2011. DOI: 10.3322/caac.20107.

[2] Chowdhury, S., & Sultana, S., *Awareness on Breast Cancer among the Women of Reproductive Age*, Journal of Family and Reproductive Health , **5**(4), pp. 127-134, 2011. https://www.sid.ir/paper/320831/en.

[3] Akram, M., Iqbal, M., Daniyal, M., & Khan, A. U., *Awareness and Current Knowledge of Breast Cancer*, Biological Research, **50**(1), pp. 1-23, 2017. DOI 10.1186/s40659-017-0140-9.

[4] Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., ... & Bray, F., *Cancer Incidence and Mortality Worldwide: Sources, Methods and Major Patterns in GLOBOCAN 2012*, International Journal of Cancer, **136**(5), pp. E359-E386,2012. DOI: 10.1002/ijc.29210.

[5] Das, A.K., Biswas, S.K., Bhattacharya, A. & Alam, E., *Introduction to Breast Cancer and Awareness*, in 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS),1, pp. 227-232. IEEE,2021. DOI: 10.1109/ICACCS51430.2021.944168.

[6] DeSantis, C. E., Ma, J., Goding Sauer, A., Newman, L. A., & Jemal, A., *Breast Cancer Statistics, Racial Disparity in Mortality by State.* CA: A Cancer Journal for Clinicians, **67**(6), 439-448, 2017. DOI: 10.3322/caac.21412.

[7] Jaikrishnan, S. V. J., Chantarakasemchit, O. & Meesad, P., *A Breakup Machine Learning Approach For Breast Cancer Prediction*, in 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 1-6, IEEE.2019. DOI: 10.1109/ICITEED.2019.8929977.

[8] Nelson, H. D., Fu, R., Zakher, B., Pappas, M. & McDonagh, M., *Medication Use For The Risk Reduction of Primary Breast Cancer in Women: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force*, Jama, **322**(9), pp. 868-886, 2019. DOI: 10.1001/jama.2019.5780.

[9] Han, S. J., Guo, Q. Q., Wang, T., Wang, Y. X., Zhang, Y. X., Liu, F., ... & He, Y., Prognostic significance of interactions between ER alpha and ER beta and lymph node status in breast cancer cases, Asian Pacific Journal of Cancer Prevention, **14**(10), pp. 6081-6084, 2013. DOI: 10.7314/APJCP.2013.14.10.6081.

[10] Miller, K.D., Ortiz, A.P., Pinheiro, P.S., Bandi, P., Minihan, A., Fuchs, H. E., ... & Siegel, R.L., *Cancer Statistics for the US Hispanic/Latino Population*, CA: A Cancer Journal for Clinicians, **71**(6), pp. 466-487, 2021. DOI: 10.3322/caac.21660.

[11] Giaquinto, A.N., Sung, H., Miller, K.D., Kramer, J.L., Newman, L.A., Minihan, A., ... & Siegel, R.L., *Breast Cancer Statistics,* CA: A Cancer Journal for Clinicians, **72**(6), pp. 524-541, 2022. DOI: 10.3322/caac.21754

[12] Sharma, S., Aggarwal, A., & Choudhury, T., *Breast Cancer Detection Using Machine Learning Algorithms*, in 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems

(CTEMS), pp. 114-118,2018, IEEE. DOI: 10.1109/CTEMS.2018.8769187.

[13] Dai, B., Chen, R. C., Zhu, S. Z., & Zhang, W. W., *Using Random Forest Algorithm for Breast Cancer Diagnosis*, in 2018 International Symposium on Computer, Consumer and Control (IS3C), pp. 449-452, IEEE, 2018. DOI: 10.1109/IS3C.2018.00119.

[14] Gupta, P., & Garg, S., *Breast Cancer Prediction Using Varying Parameters of Machine Learning Models*, Procedia Computer Science, **171**, pp. 593-601, 2020. DOI: 10.1016/j.procs.2020.04.064.

[15] Kabiraj, S., Raihan, M., Alvi, N., Afrin, M., Akter, L., Sohagi, S.A., & Podder, E., *Breast Cancer Risk Prediction Using XGboost and Random Forest Algorithm*, in 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-4. IEEE, 2020. DOI: 10.1109/ICCCNT49239.2020.9225451.

[16] Aroef, C., Rivan, Y. & Rustam, Z., Comparing random forest and support vector machines for breast cancer classification, TELKOMNIKA (Telecommunication Computing Electronics and Control), **18**(2), pp. 815-821,2020. DOI: 10.12928/TELKOMNIKA.v18i2.14785.

[17] Wang, S., Wang, Y., Wang, D., Yin, Y., Wang, Y. & Jin, Y., *An Improved Random Forest-Based Rule Extraction Method for Breast Cancer Diagnosis*, Applied Soft Computing, **86**, 105941, 2020. DOI: 10.1016/j.asoc.2019.105941.

[18] Bharati, S., Rahman, M.A. & Podder, P., *Breast Cancer Prediction Applying Different Classification Algorithm with Comparative Analysis Using WEKA*, In 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), pp. 581-584, IEEE, 2018. DOI: 10.1109/CEEICT.2018.8628084.

[19] Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A., *Machine Learning Models in Breast Cancer Survival Prediction*, Technology and Health Care, **24**(1), pp. 31-42, 2016. DOI: 10.3233/THC-151071.

[20] Octaviani, T.L. & Rustam, D.Z., *Random Forest for Breast Cancer Prediction*, in AIP Conference Proceedings. AIP Publishing LLC, **2168**(1), p. 020050, 2019. DOI: 10.1063/1.5132477.

[21] Sivakami, K. & Saraswathi, N., *Mining Big Data: Breast Cancer Prediction Using DT-SVM Hybrid Model*, International Journal of Scientific Engineering and Applied Science (IJSEAS), **1**(5), pp. 418-429,2015.

[22] Godara, S. & Singh, R., *Evaluation of Predictive Machine Learning Techniques as Expert Systems in Medical Diagnosis*, Indian Journal of Science and Technology, **9**(10), pp. 1-14, 2016. DOI: 10.17485/ijst/2016/v9i10/87212.

[23] Hamsagayathri, P., & Sampath, P., Performance analysis of breast cancer classification using decision tree classifiers, *Int J Curr Pharm Res*, *9*(2), 19-25, 2017.

[24] Yi, L. & Yi, W., *Decision Tree Model In The Diagnosis Of Breast Cancer*. in 2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC), pp. 176-179, IEEE, 2017. DOI: 10.1109/ICCTEC.2017.00046.

[25] Chaurasia, V., Pal, S. & Tiwari, B. B., *Prediction of Benign and Malignant Breast Cancer Using Data Mining Techniques*, Journal of Algorithms & Computational Technology, **12**(2), pp. 119-126, 2018. DOI: 10.1177/1748301818756225.

[26] Higa, A., *Diagnosis of Breast Cancer Using Decision Tree and Artificial Neural Network Algorithms*, Cell, **1**(7), pp. 23-27, 2018.

[27] Kaur, P., Kumar, R., & Kumar, M., *A Healthcare Monitoring System Using Random Forest and Internet of Things* (IoT), Multimedia Tools and Applications, **78**(14), pp. 19905-19916, 2019. DOI: 10.1007/s11042-019-7327-8.

[28] Ahmed, M.T., Imtiaz, M.N., & Karmakar, A., *Analysis of Wisconsin Breast Cancer Original Data Set Using Data Mining and Machine Learning algorithms For Breast Cancer Prediction*, Journal of Science Technology and Environment Informatics, **9**(2), pp. 665-672, 2020. DOI: 10.18801/jstei.090220.67.

[29] Ed-daoudy, A., & Maalmi, K., *Breast Cancer Classification with Reduced Feature Set Using Association Rules and Support Vector Machine*, Network Modeling Analysis in Health Informatics and Bioinformatics, **9**(1), 34, 2020. DOI: 10.1007/s13721-020-00237-8.

[30] Chakraborty, S., & Murali, B., *A Novel Medical Prognosis System for Breast Cancer*, in Proceedings of International Conference on Advanced Computing Applications: ICACA 2021, pp. 403-413, Singapore: Springer Singapore, 2021. DOI: 10.1007/978-981-16-5207-3_34.

[31] Dholi, P., & Patil, D. V., *A Prognosis and Prediction of Breast Cancer using Machine Learning Techniques*, in Proceedings of the 3rd International Conference on Contents, Computing & Communication (ICCCC-2022),2022. DOI: 10.2139/ssrn.4043530.

[32] Zhang, Z., & Li, Z., *Evaluation Methods for Breast Cancer Prediction in Machine Learning Field,* in SHS Web of Conferences, EDP Sciences, **144**, 03010, 2022. DOI: 10.1051/shsconf/202214403010.

[33] Das, A. K., Biswas, S. K., & Mandal, A., An expert system for breast cancer prediction (ESBCP) using decision tree, Indian J Sci Technol, **15**(45), pp. 2441-2450, 2022. DOI: 10.17485/IJST/v15i45.756.

[34] Murugan, S., Kumar, B.M. & Amudha, S., *Classification and Prediction of Breast Cancer Using Linear Regression, Decision Tree and Random Forest*, in 2017 International Conference on Current Trends in Computer,

Electrical, Electronics and Communication (CTCEEC), pp. 763-766, IEEE, 2017. DOI: 10.1109/CTCEEC.2017.8455058.

[35] Li, Y. & Chen, Z., *Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction,* Appl Comput Math, **7**(4), pp. 212-216, DOI: 10.11648/j.acm.20180704.15.

[36] Sahu, B., Mohanty, S.N. & Rout, S.K., *A Hybrid Approach for Breast Cancer Classification and Diagnosis*, EAI Endorsed Transactions on Scalable Information Systems, **6**(20), 2019. DOI: 10.4108/eai.19-12-2018.156086.

[37] Mathew, T.E., *Simple and Ensemble Decision Tree Classifier Based Detection of Breast Cancer.* International Journal of Scientific & Technology Research, **8**(11), pp. 1628-1637.

[38] Kaur, P., Kumar, R., & Kumar, M., *A Healthcare Monitoring System Using Random Forest and Internet of Things* (IoT), Multimedia Tools and Applications, **78**(14), pp. 19905-19916, 2019. DOI: 10.1007/s11042-019-7327-8

[39] Islam, M.M., Haque, M.R., Iqbal, H., Hasan, M.M., Hasan, M. & Kabir, M.N., *Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques*, SN Computer Science, **1**(5), 290, 2020. DOI: 10.1007/s42979-020-00305-w.

[40] Pyingkodi, M., Muthukumaran, M., Shanthi, S. & Saravanan, T.M., *Performance Study of Classification Algorithms Using the Microarray Breast Cancer Data Set,* International Journal of Future Generation Communication and Networking, **13**(2), 12381245, 2020.

[41] Idris, N.F. & Ismail, M.A., *Breast Cancer Disease Classification Using Fuzzy-ID3 Algorithm with FUZZYDBD Method: Automatic Fuzzy Database Definition*, PeerJ Computer Science, **7**, e427,2021. DOI: 10.7717/peerj-cs.427.

[42] Mehta, D., Mohite, A., Shinde, V., Khatri, R. & Dokare, I., *Detection of Breast Cancer using Machine Learning Algorithms*, in Proceedings of the 7th International Conference on Innovations and Research in Technology and Engineering (ICIRTE-2022), organized by VPPCOE & VA, Mumbai-22, INDIA, 2022. DOI: 10.2139/ssrn.4108758.

[43] https://archive.ics.uci.edu/ml/data sets/breast+cancer+wisconsin+%28original%29. (20 April 2024)

[44] Boukerche, A., Zheng, L. & Alfandi, O., *Outlier Detection: Methods, Models, and Classification*, ACM Computing Surveys (CSUR), **53**(3), pp.1-37, 2020. DOI: 10.1145/3381028.

[45] Sage, A., *Random Forest Robustness, Variable Importance, and Tree Aggregation,* 2018.

[46] Khourdifi, Y. & Bahaj, M., *Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification*, in 2018 International Conference on Electronics, Control, Optimization and Computer Science

(ICECOCS), pp. 1-5, IEEE, 2018. DOI: 10.1109/ICECOCS.2018.8610632.

[47] Das, A.K., Biswas, S.K., Mandal, A., Bhattacharya, A. & Saha, D., *Machine Learning Based Expert System for Breast Cancer Prediction (MLESBCP)*, in International Conference on Computational Technologies and Electronics, Cham: Springer Nature Switzerland, pp. 275-286, 2023. DOI: 10.1007/978-3-031-81935-3_24.

[48] Verma, D. & Mishra, N., Analysis and prediction of breast cancer and diabetes disease data sets using data mining classification techniques, In 2017 International Conference on Intelligent Sustainable Systems (ICISS), pp. 533-538, IEEE, 2017. DOI: 10.1109/ISS1.2017.8389229.

[49] Kharya, S. & Soni, S., *Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection*, International Journal of Computer Applications, **133**(9), pp. 32-37, 2016. DOI: 10.5120/ijca2016908023.

[50] Alaybeyoglu, A. & Mulayım, N., *A Design of Hybrid Expert System for Diagnosis of Breast Cancer and Liver Disorder*, The Eurasia Proceedings of Science Technology Engineering and Mathematics, **2**, pp. 345-353, 2018.