


ITC 4/51 Information Technology and Control Vol. 51/ No. 4 / 2022 pp. 757-770 DOI 10.5755/j01.itc.51.4.31347	A New Range-based Breast Cancer Prediction Model Using the Bayes' Theorem and Ensemble Learning	
	Received 2022/05/07	Accepted after revision 2022/09/06
	 <a href="http://dx.doi.org/10.5755/j01.itc.51.4.31347">http://dx.doi.org/10.5755/j01.itc.51.4.31347</a>	

**HOW TO CITE:** Khozama, S., Mayya, A. M. (2022). A New Range-based Breast Cancer Prediction Model Using the Bayes' Theorem and Ensemble Learning. *Information Technology and Control*, 51(4), 757-770. <http://dx.doi.org/10.5755/j01.itc.51.4.31347>

# A New Range-based Breast Cancer Prediction Model Using the Bayes' Theorem and Ensemble Learning

## Sam Khozama

Faculty of Information Technology and Bionics; Pázmány Péter Catholic University,  
Budapest, Hungary; e-mail: [khozama.sam@itk.ppke.hu](mailto:khozama.sam@itk.ppke.hu)

## Ali M. Mayya

Department of Computer and Automatic Control Engineering; Tishreen University, Syria;  
e-mail: [ali.mayya@tishreen.edu.sy](mailto:ali.mayya@tishreen.edu.sy)

**Corresponding author:** [khozama.sam@itk.ppke.hu](mailto:khozama.sam@itk.ppke.hu)

Breast cancer prediction is essential for preventing and treating cancer. In this research, a novel breast cancer prediction model is introduced. In addition, this research aims to provide a range-based cancer score instead of binary classification results (yes or no). The Breast Cancer Surveillance Consortium dataset (BCSC) dataset is used and modified by applying a proposed probabilistic model to achieve the range-based cancer score. The suggested model analyses a sub dataset of the whole BCSC dataset, including 67632 records and 13 risk factors. Three types of statistics are acquired (general cancer and non-cancer probabilities, previous medical knowledge, and the likelihood of each risk factor given all prediction classes). The model also uses the weighting methodology to achieve the best fusion of the BCSC's risk factors. The computation of the final prediction score is done using the post probability of the weighted combination of risk factors and the three statistics acquired from the probabilistic model. This final prediction is added to the BCSC dataset, and the new version of the BCSC dataset is used to train an ensemble model consisting of 30 learners. The experiments are applied using the sub and the whole datasets (including 317880 medical records). The results indicate that the new range-based model is accurate and robust with an accuracy of 91.33%, a false rejection rate of 1.12%, and an AUC of 0.9795. The new version of the BCSC dataset can be used for further research and analysis.

**KEYWORDS:** Machine Learning, Ensemble Learning, Breast Cancer, Probability Estimation, Risk Factors.

---

## 1. Introduction

Breast cancer is one of the most common and challenging diseases that has received much attention either in medical or biomedical domains. The main problem of cancer diagnosis and prediction is the huge amount of data that cannot be dealt with in the traditional manual method (physician's observations), and a more powerful speed approach is needed [14, 15, 22]. Fortunately, the rapid development in the computer science field, especially in machine learning methodologies, has revealed the hidden information inside those datasets and provided health organizations with useful tools for diagnosing and predicting cancer [1, 2, 5, 27].

Many previous breast cancer prediction tools were designed; some used well-known breast cancer datasets, while others used their own collected data. In some researches, the designed tools used known built-in models like Kaplan-Meier [13], while others used some known machine learning models (Support Vector machines (SVM), K-Nearest Neighbour (K-NN), Random Forests (RF), Decision Trees (DT), Neural Networks, Naïve Bayes and Logistic Regression (LR)) [3, 4, 11, 17, 22]. Some researchers used deep learning methodologies and fused them with image models, obtaining mammography breast image features, with the textual information of risk factors to improve the prediction model's accuracy [29]. Some of these researches got benefit of the parameter optimizations and ensemble learning methods to enhance the performance significantly [16, 21, 23].

---

## 2. Related Work

The conditional probability of the Bayes theorem was introduced by Ramkumar et al. [25] for liver cancer prediction. They used a dataset collected from 20 patients from the BUPA research lab. There were seven attributes in the data set including the Mean corpuscular volume, Alkaline phosphate, alkaline aminotransferase, aspartate aminotransferase, gamma transpeptidase, number of half-pints equivalent to alcohol and a selector to split the dataset into training and validation. For the used dataset, different probabilities were computed (the probability that an individual will suffer or not from liver cancer, and the

test's conditional probability will be positive/negative given that the disease is present/absent). The researcher used the Weka tool in order to analyse their dataset and apply the required Naïve Bayes classifier. The obtained results were not so good, and the accuracy was only 50%. Their results indicated that the proposed methodology had no pre-processing steps. The size of the used dataset was very small.

For the aim of Breast Cancer Surgery Survivability Prediction, Al-Jawad et al. [3] used the Bayesian Network and the Support Vector Machines SVM. Haberman's survival dataset contained 306 cases (225 confirmed cancer cases survived 5 years after cancer surgery). The Weka tool was used to apply the SVM and BN classifiers in their research. The research also computed five different statistical features of the used dataset's three attributes: mean, median, standard deviation, maximum and minimum values). They also computed the correlation coefficients of the three features pairs (Age and survival status 0.067, Year and survival status -0.00477, Positive nodes and survival status 0.28677). They chose fixed values of the optimizable parameters of SVM and BN models, which was why their methodology achieved low performance. Their results indicated that SVM outperformed the Bayesian Network by 6.88%. The SVM achieved 73.78% and 74.77% for Recall and Precision metrics, while the BN achieved 78.22% and 64.47% for Recall and Precision, respectively. The main problems of their research were the small dataset size and the fixed learning parameters.

In 2018, Annemieke et al. [4] compared logistic regression with different Bayesian Networks BNs. They selected a subset of data from the Netherlands Cancer Registry, including 37,320 samples. The selected dataset was between 2003 and 2006, related to women with early-stage breast cancer. The Bayesian network classifiers, the correlation coefficients, the constraint-based learning methodologies, and the score-based learning models were used to support the BNs architectures to get better performance. AUC (Area Under Curve) evaluation metrics were used to evaluate those different models, and in order to apply those validations, an external validation set was obtained from the NCR from 2007 and 2008 (N = 12,308). Although logistic regression indicat-

ed the best performance on most experiments of the sub-dataset analysis, BNs exceeded the performance of regression for SP prediction for the high and low-risk subsets. The authors concluded that in the case of BNs, the value of coefficient estimators had no relationship with the changes of the other variables' values.

In Yang et al. [30] study, three different classifiers were fused to get the best efficiency (Bayesian and Markov models and the artificial neural network). Bayesian and Markov models were first used to establish a connection between the previous and current incidence of cancer. The outputs of these two classifiers fed back into the Neural Network classifier. A pre-processing step was applied to the used dataset. They applied normalization and missed data manipulation steps to prepare the cancer dataset. Twenty attributes were used from an entire dataset consisting of 36,000 cases. Those cancer cases included 10,500 patients with lung cancer, 13,500 with liver cancer, and 12,000 with stomach cancer, respectively. The researchers split the dataset into 75% training and 25% test. The experimental results showed that the overall training accuracy was 73.55%, 76.07% and 75.63% for the ANN, CBM and the proposed fusion methodology. For the test set, they got 68.78%, 70.63%, 72.47% as accuracies for the same previous settings. The main problem of their results was that the F1-score was rather small, which indicated that their proposed approach suffered from false positive and false negative results. This might have been due to the fact that they collected their data from different data sources. They also compared their results with other classifiers like Random Forests (RF), SVM and ELM. While their approach's performance exceeded the ELM, the performance of SVM and RF was better.

Recently, in 2021, Khozama and Mayya [17] developed a breast cancer prediction model based on risk factors. They used the BCSC dataset containing 12 risk factors and applied three different types of sampling in order to unbiased the dataset. They designed a weighting system of the risk factors depending on special medical questionnaires and information retrieved from the international medical reports. They got an accuracy of 95.8% of the oversampled dataset. Their system also indicated improvement in the false rejection and false discovery errors rates against the unweighted version of the dataset.

Another research used the BCSC dataset to predict breast cancer using 154899 records [22]. They used many machine learning algorithms like Logistic Regression, SVM, Naïve Bayes, and Bayesian Network. Their results confirmed that the Naïve Bayes classifier had the best accuracy in predicting the likelihood of breast cancer while the SVM and BNs had the lowest performance.

A further type of research used Next-Generation Sequencing (NGS) methodology together with machine learning algorithms for the aim of breast cancer prediction [19]. The NCBI (National Centre for Biotechnology Information) dataset was used to extract the NGS data samples forming 4 different categories (1580 samples). The researchers extracted the sequence features and then used different machine learning classifiers (K-NN, SVM, Naïve Bayes, AdaBoost, Decision Trees, Random Forests and gradient boosting). The evaluation proved that the decision tree was the best classifier with 94.30% accuracy.

Miloš et al. [26] analyzed the machine learning models to predict life's quality for breast cancer patients. They used two different datasets; the BcBase early breast cancer prediction dataset and the ORB prostate cancer dataset. The researcher evaluated different machine learning algorithms like RF, SVM, Naïve Bayes, K-NN and decision trees. They examined two different types of QoL models (centrally-trained and federated). The results indicated that both models gave accurate predictions in the case of short term predictors while the centrally-trained models overperformed in the case of long predictors. The results also indicated low values of precision and recall in both models.

Recently in 2022, Guo et al. [16] proposed an MLP-based cancer prediction model. In this case, ensemble learning was also used to improve the performance of the MLP classifier by applying the optimization process for tuning some particular parameters (number of input features, number of hidden layers, number of neurons of each layer and weight values). The experiments were applied on the Wisconsin Breast Cancer Database WBCD. The obtained accuracy was 98.79% when using the MLP classifier and the parameter optimization.

Some previous studies used the well-known machine learning models, and few of them used the idea

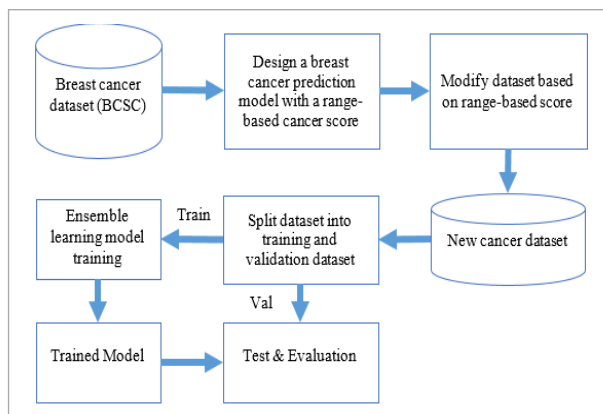
of ensemble learning or using a mix of many models. Many studies used the BCSC dataset (partially or completely), but none of them analyzed the probabilistic distribution of this dataset. All previous studies introduced the cancer prediction problem using specific cancer prediction results (yes or no). In our study, a range-based cancer score will be computed based on a probabilistic model. The probabilistic analysis of a selected subset of the BCSC dataset will be performed, so that previous knowledge of all risk factors, their likelihoods and the general cancer statistics will be extracted and used along with a weighted system (previous work) to compute the final cancer score. To evaluate our methodology, an ensemble model will be trained using the new version of the BCSC dataset. The selection of hyperparameters of the ensemble model will be applied to get the best performance and avoid the limitations of the previous studies.

### 3. Materials and Methods

Whereas previous research achieved great result in the field of predicting cancer in patients, for more efficient and accurate prediction, we suggest using a range-based cancer score value instead of using a scalar value (0 or 1) so that the decision will not be either cancer or not. Rather, it will be a range value between 0% and 100%, indicating the potential breast cancer risk. Figure 1 includes the proposed methodology.

**Figure 1**

The proposed range-based cancer score methodology

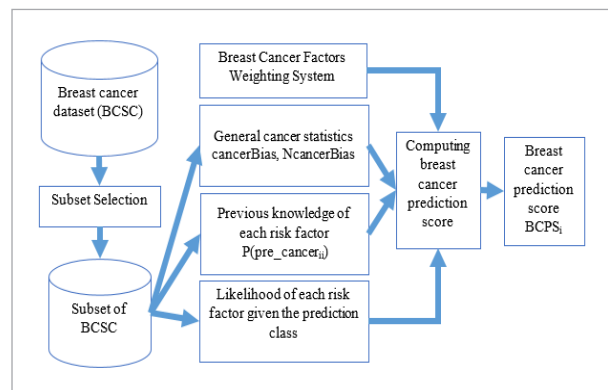


### 3.1. Designing the Breast Cancer Range-Based Model

The proposed methodology must include two main systems to design our model. The first is the breast-cancer factors weighting system (acquired from the previous work [17]), while the other is the proposed statistical system to compute the breast cancer statistics required for the final cancer score. The range-value cancer prediction model is described in Figure 2.

**Figure 2**

The range-value cancer prediction score model



### 3.2. Computing Breast Cancer Prediction Score

Our breast cancer prediction system has the following inputs:

- 1 Breast cancer factors' weights were obtained from our previous research (previous paper [17]): each risk factor had been assigned a scalar weight (1, 2, 3, 4...), indicating the degree of importance of each risk factor. These numbers will be used in our prediction system to get an accurate prediction score taking into account the importance of each risk factor.
- 2 General cancer statistics: The general probability of cancer "cancerBias" and non-cancer "NcancerBias", which are computed from the breast cancer dataset.

The cancerBias and NcancerBias represent the previous knowledge obtained by the BCSC sub dataset in which the class "Cancer" constitutes 68.88% of all samples while the "Ncancer" class has less than 31.12% of the entire samples.

- 3 Likelihood of each risk factor given that the prediction result is cancer or non-cancer: The sum of all other likelihoods of its inner values given the same prediction, as described in Equation (1).

$$P(\text{Risk\_Factor}_i | \text{Prediction}) = \sum_k P(\text{Inner\_Value}_{ij} | \text{Prediction}) \quad (1)$$

Where k is the total number of risk factor's inner values.

- 4 Previous knowledge of each risk factor: The medical opinion about the probability of the effect of each risk factor on the final breast cancer score. This part is formulated as  $P(\text{pre\_cancer}_{ij})$  and obtained from the analysis of some medical questionnaires delivered to specialist physicians in breast cancer.

The final breast cancer prediction score as a range-value is calculated using the previous inputs (Equation (2)).

$$\text{BCPS}_i = \text{cancerBias} * \text{BCPS}_{\text{cancer}} + \text{NcancerBias} * \text{BCPS}_{\text{Ncancer}} \quad (2)$$

BCPS<sub>cancer</sub> and BCPS<sub>Ncancer</sub> are the post probabilities of cancer and non-cancer, given the risk factors formulated as Equations 3 and 4 suggest based on Bayes' theorem.

$$\text{BCPS}_{\text{cancer}} = \sum_n P(\text{prediction} = \text{cancer} | \text{Risk\_factor}_i) \times (\text{STW}(j) / \sum_n \text{STW}(j)) \quad (3)$$

$$\text{BCPS}_{\text{Ncancer}} = \sum_n P(\text{prediction} = \text{Non-cancer} | \text{Risk\_factor}_i) \times (\text{STW}(j) / \sum_n \text{STW}(j)) \quad (4)$$

STW<sub>j</sub> is the suggested training weight (of our previous paper) and n is the number of all risk factors. The post probability of each risk factor is calculated as shown in Equation (5).

$$P(\text{Prediction} = \text{Cancer} | \text{Inner\_value}_i) = \sum_k (P(\text{Inner\_value}_{ij} | \text{Prediction} = \text{Cancer}) \times P(\text{Pre\_cancer}_{ij}) / P(\text{Inner\_value}_{ij})) \quad (5)$$

Where K is the number of the risk factor's inner values (For example, the menopause risk factor has three different values (K=3), which are Pre-menopause (0), Post-Menopause (1) and Unknown (9)).  $P(\text{Pre\_cancer}_{ij})$  values are the pre-probabilities of the previous knowledge of cancer related to the

risk factor's inner value  $ij^{\text{th}}$ .  $P(\text{Innervalue}_{ij})$  is the evidence of each risk factor information and can be formulated as Equation (6) shows.

$$P(\text{Inner\_value}_{ij}) = P(\text{Inner\_Value}_{ij} | \text{Prediction} = \text{Cancer}) \times P(\text{Pre\_Cancer}_{ij}) + P(\text{Inner\_Value}_{ij} | \text{Prediction} = \text{Non-Cancer}) \times (1 - P(\text{Pre\_Cancer}_{ij})) \quad (6)$$

### 3.3. Modifying BCSC Dataset Based on Risk Range-based Score

The BCSC dataset is modified by adding three new attributes in this step. The added attributes are the cancer score, the non-cancer score and the final prediction. Future studies can use these attributes to predict and analyze the BCSC dataset. The final prediction of our proposed methodology will use this new version of the BCSC dataset.

### 3.4. Ensemble Learning Model Training

Ensemble learning is a method in which many classifiers (models) are fused to build a huge powerful model. It has the advantage of using many classifiers to improve performance. A fusion of ensemble learning and hyperparameters optimization has been given a lot of attention in the last few years [21].

Many hyperparameters are selected to be optimized. Those parameters are the maximum number of splits, number of learners and learning rate. The used ensemble method is the AdaBoost algorithm [31], while the learner type is the decision trees algorithm [18].

A Decision Tree (DT) is a machine learning model in which the internal nodes represent features, while branching denotes one of the possible results [10]. At the first step, the best promising feature is chosen as the root node, then the splitting process is applied based on a specific criterion (varying from one method to another).

Many learners are created and learned sequentially (fitting the model using the dataset) [24, 28]. Thus, in each step, a decision tree learner is chosen and fitted so that the error is forwarded to the next step and used to learn the next step learner. In this way, the miss-classified samples of the previous model will be correctly classified by the next one [24].



## 4. Results and Discussion

### 4.1. Dataset Description

We use a balanced version of the BCSC [6, 7] containing 317880 records and 13 risk factors for the experimental part. The risk factors are menopause, age group (agegrp), race, Hispanic factor, Body Mass Index (BMI), age at first birth (agefirst), number of first relatives with breast cancer (nrelbc), breast procedure (brstproc), last mammogram before the index mammogram (lastmamm), surgical menopause (surgmeno) and current hormone therapy (Current\_hor). The last attribute in BCSC dataset is the “count” column which holds the frequency of each case in the dataset.

### 4.2. Training Scenarios

The machine learning model will be trained using the modified version of the BCSC dataset. Two training scenarios are suggested for the learning process. In the first scenario, the selected subset of the entire BCSC dataset is used to build and train the ensemble learning model. For the second scenario, the whole BCSC dataset is used. In both scenarios, the datasets are split into 80% training and 20% validation.

### 4.3. Computing the Probabilistic Model Using the Subset Training Dataset

For each risk factor in the subset, the pre and post probabilities are computed using the 67633 records of the subset training dataset. The post probabilities computed according to equation 5 are illustrated in Table 1.

**Table 1**

Cancer and non-cancer post probabilities of BCSC risk factors

No.	Risk Factor	P(Prediction=Cancer  Innervalue ij)	P(Prediction=No cancer Innervalue ij)
1	Menopause	Pre=78.34%, Post (age>55)= 30.29%, Unknown= 21.89%	Pre=21.66%, Post (age>55)= 69.71%, Unknown= 78.11%
2	Age group	35-39=4.67%; 40-44=11.81%; 45-49=22.47%; 50-54=41.57%; 55-59=29.2%; 60-64 =22.2%; 65-69=12.43%; 70-74=13.4%; 75-79=14.56%; 80-84=6.79%.	35-39=95.33%; 40-44=88.19%; 45-49=77.53%; 50-54=58.43%; 55-59=70.8%; 60-64 =77.8%; 65-69=87.57%; 70-74=86.6%; 75-79=85.44%; 80-84=93.21%.
3	Density	Almost entirely fatty: 9.99%, Scattered fibro-glandular: 45.88%, Heterogeneously dense: 52.68%, Extremely dense: 38.24%, Unknown: 20.97%	Almost entirely fatty: 90.01%, Scattered fibro-glandular: 54.12%, Heterogeneously dense: 47.32%, Extremely dense: 61.76%, Unknown: 97.03%
4	Race	White: 72.85%; Asian/Pacific Islander: 36.36%; Black: 10.62%; Native American: 7.77%; Other/mixed:28.1%; Unknown: 19.66%.	White: 27.15%; Asian/Pacific Islander: 63.64%; Black: 89.38%; Native American: 92.23%; Other/mixed: 71.9%; Unknown: 80.34%.
5	Hispanic	No: 28.33%;Yes: 81.6%; Unknown: 28.17%.	No: 71.67%;Yes: 18.4%; Unknown: 71.83%.
6	BMI	10-24: 18.94%; 25-29.99: 23.34%; 30-34.99: 31.41%; 35 or more: 41.58%; Unknown: 59.57%.	10-24: 81.06%; 25-29.99: 76.66%; 30-34.99: 68.59%; 35 or more: 58.42%; Unknown: 40.43%.
7	Age at first birth (agefirst)	Age<30: 30.54%; Age 30 or greater: 60.11%; Nulliparous: 60.24%; Unknown: 23.83%.	Age<30: 69.46%; Age 30 or greater: 39.89%; Nulliparous: 39.76%; Unknown: 76.17%.
8	Number of first degree relatives with breast cancer (nrelbc)	Zero: 21.75%; One: 49.02%; 2 or more: 96.99%; Unknown: 24.68%.	Zero: 78.25%; One: 50.98%; 2 or more: 3.01%; Unknown: 75.32%.
9	Previous breast procedure (brstproc)	No: 18.11%; Yes:87.41%; Unknown: 36.34%.	No: 81.89%; Yes:12.59%; Unknown: 63.66%.
10	last mammogram before the index mammogram (lastmamm)	Negative: 63.47%; False positive: 88.77%; Unknown: 20.25%.	Negative: 36.53%; False positive: 11.23%; Unknown: 79.75%.
11	Surgical menopause	Natural: 31.38%; Surgical: 81.27%; Unknown or not Menopausal: 32.57%.	Natural: 68.62%; Surgical: 18.73%; Unknown or not Menopausal: 67.43%.
12	Hormone therapy	No: 29.08%; Yes: 82.6%; Unknown: 31.92%.	No: 70.92%; Yes: 17.64%; Unknown: 68.08%.

For the “count” attribute and by using the distribution of cancer and non-cancer samples, the post probabilities are computed as follows:  $P(\text{Cancer}|\text{count}<2)=0.3$ ,  $P(\text{NCancer}|\text{count}<2)=0.3$ .

$P(\text{Cancer}|\text{count}\geq 2\ \&\ \text{count}<50)=0.8$ ,  
 $P(\text{NCancer}|\text{count}\geq 2\ \&\ \text{count}<8)=0.8$ .

$P(\text{Cancer}|\text{count}\geq 50\ \&\ \text{count}<1000)=0.95$ ,  
 $P(\text{NCancer}|\text{count}\geq 8\ \&\ \text{count}<16)=0.95$ .

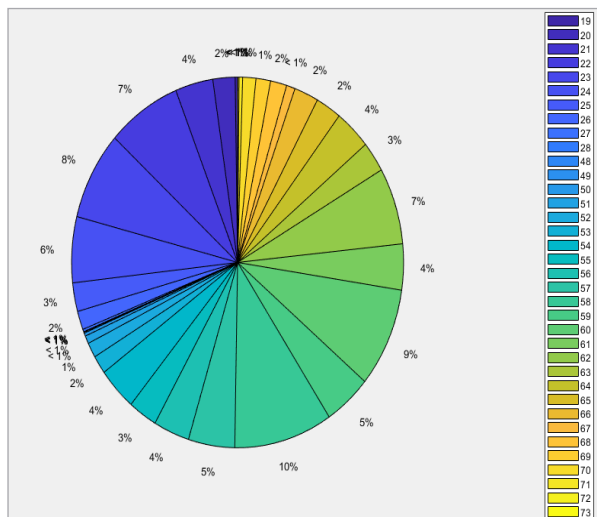
$P(\text{Cancer}|\text{count}\geq 1000)=1$ ,  
 $P(\text{NCancer}|\text{count}\geq 16)=1$ .

#### 4.4. Computing the Final Prediction Range-based Cancer Score Using the Subset Training Dataset

In this step, we use the probabilistic statistics of the previous stage and the risk factors weights obtained from previous work [17]. The calculations in this step aim to evaluate the probabilistic model and compute the final prediction score  $BCPS_{\text{cancer}}$  and  $BCPS_{\text{Non-cancer}}$  according to equations 3 and 4.

Figure 3 illustrates the distribution of the “result prediction score” of the subset dataset where the main remark is that the “non-cancer” class is divided into the subclasses (‘19’, ‘20’, ‘21’, ‘22’, ‘23’, ‘24’, ‘25’, ‘26’, ‘27’, ‘28’) representing the low-predicted percentages of breast cancer instead of using only one class to describe the presence or absence of breast cancer. On the other hand, the “cancer class” is di-

**Figure 3**  
 Distribution of new subclasses of the “cancer” and “non-cancer” original classes in case of using the sub dataset

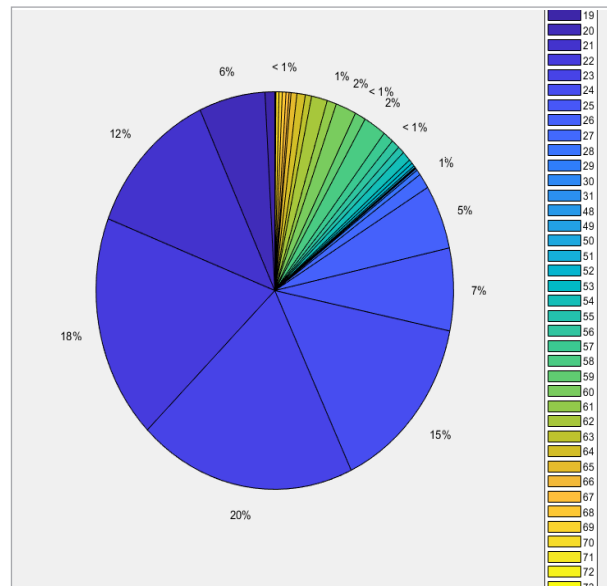


vided into 26 subclasses (‘48’, ‘49’, ‘50’, ‘51’, ‘52’, ‘53’, ‘54’, ‘55’, ‘56’, ‘57’, ‘58’, ‘59’, ‘60’, ‘61’, ‘62’, ‘63’, ‘64’, ‘65’, ‘66’, ‘67’, ‘68’, ‘69’, ‘70’, ‘71’, ‘72’, ‘73’) which are the high-predicted percentages of breast cancer scores.

#### 4.5. Computing the Probabilistic Model Using the Whole (Original) Training Dataset

The same experiments are repeated to compute the distribution of the subclasses of the entire BCSC dataset (Figure 4). The same number of subclasses for the “cancer” class is obtained but with different distribution since the original dataset has a different distribution of “cancer” and “non-cancer” classes. On the other hand, three subclasses, “29”, “30”, and “31” of the “non-cancer” class are presented.

**Figure 4**  
 Distribution of new subclasses of the “cancer” and “non-cancer” original classes in case of using the whole dataset



There is an essential distribution difference in the case of the sub dataset and the whole dataset. Figure 4 shows this significant difference where the “non-cancer” categories (from “19” until “31”) have higher percentages than the “cancer” categories. This counts as normal since the original dataset has almost 84% of its samples as “non-cancer” class.

#### 4.6. Ensemble Model Training Results

The sub dataset and whole dataset are given to the ensemble classifier to get two learned models; one for the subset and another for the whole dataset.

The hyperparameters optimization is applied in both training scenarios using the AdaBoost ensemble method and the Bayesian optimization. The Minimum Classification Error (MCE) of the training process on the sub and the whole dataset is computed. Through the iterations from 10 to 30, the MCE value of the subset is lower by 0.1 than the whole dataset.

#### 4.7. Ensemble Model Evaluation

Many evaluation metrics are computed to evaluate the trained ensemble model that has been trained using the sub dataset and the entire dataset. Those metrics include the True Positive Rate (TPR), the False Negative Rate (FNR), the Positive Predictive Rate (PPR) and the False Discovery Rate (FDR) [8, 20]. All these metrics are calculated using four statistics (TP: the true positives which describe the correctly classified samples of the whole positive ones, FN: the false positives which express the incorrectly classified samples of the whole positive ones, TN: the true positives which calculate the correctly rejected samples of the whole negative ones, and the FP: the false positives representing the incorrectly accepted samples of the whole negative ones).

TPR ( $TP/(TP+FN)$ ) is the proportion of the correctly classified samples per predictive class, while the FNR ( $FN/(TP+FN)$ ) is the proportion of the incorrectly classified samples per true class [12]. Likewise, PPR ( $TP/(TP+FP)$ ) is the proportion of the correctly classified samples per predictive class, while FDR ( $FP/(TP+FP)$ ) is the proportion of the incorrectly classified samples per predictive class [12]. On the other hand, the accuracy is calculated as  $((TP+TN)/(TP+TN+FP+FN))$  representing the proportion of the correctly classified samples of all data samples. Table 2 includes the evaluation results of the trained ensemble model using the new version of the sub dataset and the whole dataset.

Table 2 statistics show that the average TPR are 94.61% and 92.52% for both sub and whole datasets. Similarly, the average PPRs are 92.28% and 85.55% for sub and whole datasets. The total accuracy of both sub and whole datasets is 95.5% and 85.3%, respectively. To measure the ability to distinguish between different subclasses, the Area Under Curve (AUC) [9] is used for all trained ensemble models. Table 3 includes detailed AUC results of all subclasses ("19"- "73") of the sub and whole datasets.

**Table 2**

Evaluation results of the ensemble model using the sub and whole dataset

Class	TPR_sub	TPR_Whl	FNR_sub	FNR_Whl	PPR_sub	PPR_Whl	FDR_sub	FDR_Whl
19	64.1	77.9	35.9	22.1	83.33	86.7	16.7	13.3
20	86.58	83.9	13.42	16.1	87.53	87.5	12.47	12.5
21	85.85	84.4	14.15	16.6	82.41	85.3	17.59	14.7
22	85.88	84.4	14.12	16.6	88.12	85.6	11.88	14.4
23	88.61	83.9	11.39	16.1	87.8	85.9	12.2	14.1
24	87.14	82.9	12.86	17.1	87.41	85.2	12.59	14.8
25	77.73	75.3	22.27	24.7	77.73	80.6	22.27	19.4
26	84.62	83.9	15.38	16.1	83.4	85.3	16.6	14.7
27	77.78	72.3	22.22	27.7	87.5	78.6	12.5	21.4
28	75	64.3	25	35.7	75	72.3	25	27.7
29	-	64.5	-	35.5	-	69.6	-	30.4
30	-	69.4	-	30.6	-	73.5	-	26.5
31	-	0	-	100	-	-	-	100
48	100	100	0	0	100	100	0	0
49	100	100	0	0	100	94.1	0	5.9
50	100	100	0	0	100	82.1	0	17.9
51	100	100	0	0	95.9	86	4.1	14.0
52	100	100	0	0	100	81.9	0	18.1
53	98.16	100	1.84	0	100	83.6	0	16.4
54	100	96.6	0	3.4	100	83.1	0	16.9
55	100	99.2	0	0.8	100	84.0	0	16.0
56	100	100	0	0	99.16	85.8	0.84	14.2
57	100	99.6	0	1.4	100	86.8	0	13.2
58	99.38	99.2	0.62	0.8	100	86.4	0	13.6
59	100	100	0	0	98.73	84.9	1.27	15.1
60	99.66	98.6	0.34	1.4	100	84.8	0	15.2
61	100	99.3	0	0.7	100	87.2	0	12.8
62	100	100	0	0	100	87.1	0	12.9
63	100	99.1	0	0.9	100	85.3	0	14.7
64	100	99.3	0	0.7	100	85.8	0	14.2
65	100	98.1	0	1.9	100	89.8	0	10.2
66	100	100	0	0	100	89.5	0	10.5
67	100	100	0	0	100	83.4	0	16.6
68	100	100	0	0	98.17	93.0	1.83	7.0
69	97.92	100	2.08	0	97.9	90.1	2.1	9.9
70	97.7	100	2.3	0	100	90.4	0	9.6
71	100	100	0	0	100	90.5	0	9.5
72	100	100	0	0	100	79.2	0	20.8
73	100	100	0	0	100	100	0	0



**Table 3**

AUC calculations of the subclasses of the sub and whole datasets: CN: class number, SUB: sub dataset, WHL: whole dataset, NoS: number of samples

CN	19	20	21	22	23	24	25	26	27	28
<b>SUB</b>	0.89	0.89	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1
<b>WHL</b>	0.96	0.97	0.97	0.96	0.96	0.96	0.95	0.97	0.96	0.94
<b>NoS</b>	2788	18982	38868	57875	62345	46653	21275	16621	3917	1380
<b>CN</b>	<b>29</b>	<b>30</b>	<b>31</b>	<b>48</b>	<b>49</b>	<b>50</b>	<b>51</b>	<b>52</b>	<b>53</b>	<b>54</b>
<b>SUB</b>	-	-	-	1	1	1	1	1	1	1
<b>WHL</b>	0.94	0.94	0.72	1	1	1	1	1	1	1
<b>NoS</b>	495	146	10	15	60	185	465	885	1085	2560
<b>CN</b>	<b>55</b>	<b>56</b>	<b>57</b>	<b>58</b>	<b>59</b>	<b>60</b>	<b>61</b>	<b>62</b>	<b>63</b>	<b>64</b>
<b>SUB</b>	1	1	1	1	1	1	1	1	1	1
<b>WHL</b>	1	1	1	1	1	1	1	1	1	1
<b>NoS</b>	1905	2370	3055	6475	3105	5890	2710	4540	1805	2390
<b>CN</b>	<b>65</b>	<b>66</b>	<b>67</b>	<b>68</b>	<b>69</b>	<b>70</b>	<b>71</b>	<b>72</b>	<b>73</b>	
<b>SUB</b>	1	1	1	1	1	1	1	1	1	
<b>WHL</b>	1	1	1	1	1	1	1	1	1	
<b>NoS</b>	1685	1595	565	1060	940	865	230	75	10	

All “cancer” subclasses have the AUC=1 in both sub and whole datasets. However, the “non-cancer” subclass “31” has the least AUC value. Table 2 supports this point, since subclass “31” has the highest FNR and FDR values. This issue arises because subclass “31” has too few samples (7 for training and 3 for validation) compared with the other subclasses.

#### 4.8. Variance Results

The new subclasses of the original BCSC datasets have an original containing class, which means that the subclasses (“19” to “31”) belong to the “non-cancer” class. Similarly, the “cancer class” contains the subclasses (“48” to “73”). In order to express the actual results of the modified model, we repeated the performance evaluations in two other new trials; the first one has  $\pm 1$  classes-variance tolerance, while another one represents the  $\pm 2$  classes-variance. The very closed subclasses ( $\pm 1$  or  $\pm 2$ ) introduce similar cancer/non-cancer scores and can be treated as one subclass.

Therefore, if the actual subclass is “21” then the accepted true classes can be “20”, “21” and “22” for  $\pm 1$  classes-variance. On the other hand, for  $\pm 2$  classes-variance, the accepted classes are “19”, “20”, “21”, “22”, and “23”.

In the first trial, two biases of the main classes are allowed so that if the sample has the original true class  $i$ , then the expected valid classes are  $(i-1, i, i+1)$ , while in the second trial the expected valid classes are  $(i-2, i-1, i, i+1, i+2)$ . Tables 4 and 5 include the detailed results of these two described trials for both sub and whole datasets, respectively.

Results of Table 4 indicate that the average TPR of the original confusion matrix (of the sub dataset) is 90.1564%, while it is increased by 4.2% and 5.38% for the ( $\pm 1$  and  $\pm 2$ ) variance scenarios, respectively. In the same way, the PPR of the ( $\pm 1$  and  $\pm 2$ ) variance scenarios has been enhanced by 4.56% and 4.72%, respectively. Similarly, the average TPR of the original confusion matrix (of the whole dataset) is increased by 8.66% and 8.76% for both  $\pm 1$  and  $\pm 2$  variance scenarios, respectively (see Table 5). The average PPR values of the  $\pm 1$  and  $\pm 2$  variance scenarios also increased by 5.33% and 5.55%, respectively. A similar computation of the accuracy also proves the same conclusion where the original accuracy was 85.3%, but it increases by 5.82% and 6.03% for  $\pm 1$  and  $\pm 2$  class-variances, respectively.

**Table 4**

TPR, FNR, PPR and FDR values of the ensemble model using the sub dataset and  $\pm 1$  or  $\pm 2$  class variances.

Class No.	$\pm 1$ classes-variance				$\pm 2$ classes-variance			
	TPR (%)	FNR (%)	PPR (%)	FDR (%)	TPR (%)	FNR (%)	PPR (%)	FDR (%)
19	100	0	100	0	100	0	100	0
20	100	0	99.72	0.28	100	0	100	0
21	100	0	99.85	0.15	100	0	100	0
22	99.8	0.25	100	0	100	0	100	0
23	99.9	0.08	100	0	100	0	100	0
24	100	0	99.69	0.31	100	0	100	0
25	100	0	100	0	100	0	100	0
26	99.6	0.36	100	0	100	0	100	0
27	100	0	100	0	100	0	100	0
28	100	0	100	0	100	0	100	0
48	100	0	100	0	100	0	100	0
49	100	0	100	0	100	0	100	0
50	100	0	100	0	100	0	100	0
51	100	0	95.9	4.1	100	0	100	0
52	100	0	100	0	100	0	100	0
53	100	0	100	0	100	0	100	0
54	100	0	100	0	100	0	100	0
55	100	0	100	0	100	0	100	0
56	100	0	99.16	0.94	100	0	100	0
57	100	0	100	0	100	0	100	0
58	99.7	0.31	100	0	100	0	100	0
59	100	0	100	0	100	0	100	0
60	100	0	100	0	100	0	100	0
61	100	0	100	0	100	0	100	0
62	100	0	100	0	100	0	100	0
63	100	0	100	0	100	0	100	0
64	100	0	100	0	100	0	100	0
65	100	0	100	0	100	0	100	0
66	100	0	100	0	100	0	100	0
67	100	0	100	0	100	0	100	0
68	100	0	100	0	100	0	100	0
69	100	0	100	0	100	0	100	0
70	100	0	100	0	100	0	100	0
71	100	0	100	0	100	0	100	0
72	100	0	100	0	100	0	100	0
73	100	0	100	0	100	0	100	0

**Table 5**TPR, FNR, PPR and FDR values of the ensemble model using the whole dataset and  $\pm 1$  or  $\pm 2$  class variances

Class No.	$\pm 1$ classes-variance				$\pm 2$ classes-variance			
	TPR (%)	FNR (%)	PPR (%)	FDR (%)	TPR (%)	FNR (%)	PPR (%)	FDR (%)
19	98.28	1.72	86.7	13.3	98.28	1.72	86.7	13.3
20	97.3	2.7	99.98	0.02	97.34	2.66	100	0
21	97.54	2.46	99.89	0.11	97.54	2.46	100	0
22	97.49	2.51	99.7	0.3	97.51	2.49	99.86	0.14
23	97.02	2.98	99.91	0.09	97.11	2.89	99.95	0.05
24	96.96	3.04	99.64	0.36	97.17	2.83	99.81	0.19
25	96.99	3.01	99.8	0.2	97.09	2.91	99.88	0.12
26	97.39	2.61	99.39	0.61	97.28	2.72	99.71	0.29
27	96.32	3.68	100	0	96.63	3.67	100	0
28	92.46	7.54	99.35	0.65	95.65	4.35	100	0
29	96.77	3.23	99.13	0.87	96.77	3.23	100	0
30	97.22	2.78	94.12	5.88	97.22	2.78	100	0
31	100	0	-	0	100	0	-	0
48	100	0	100	0	100	0	100	0
49	100	0	94.1	5.9	100	0	94.1	5.9
50	100	0	82.1	17.9	100	0	82.1	17.9
51	100	0	86	14.0	100	0	86	14.0
52	100	0	81.9	18.1	100	0	81.9	18.1
53	100	0	83.6	16.4	100	0	83.6	16.4
54	100	0	83.1	16.9	100	0	83.1	16.9
55	99.32	0.68	85.16	14.84	99.32	0.68	85.16	14.84
56	99.6	0.8	85.8	14.2	99.6	0.8	85.8	14.2
57	99.6	1.4	86.8	13.2	99.6	1.4	86.8	13.2
58	99.2	0.8	86.4	13.6	99.2	0.8	86.4	13.6
59	100	0	84.9	15.1	100	0	84.9	15.1
60	98.6	1.4	84.8	15.2	98.6	1.4	84.8	15.2
61	99.3	0.7	87.2	12.8	99.3	0.7	87.2	12.8
62	100	0	87.1	12.9	100	0	87.1	12.9
63	99.1	0.9	85.3	14.7	99.1	0.9	85.3	14.7
64	99.3	0.7	85.8	14.2	99.3	0.7	85.8	14.2
65	98.1	1.9	89.8	10.2	98.1	1.9	89.8	10.2
66	100	0	89.5	10.5	100	0	89.5	10.5
67	100	0	83.4	16.6	100	0	83.4	16.6
68	100	0	93.0	7.0	100	0	93.0	7.0
69	100	0	90.1	9.9	100	0	90.1	9.9
70	100	0	90.4	9.6	100	0	90.4	9.6
71	100	0	90.5	9.5	100	0	90.5	9.5
72	100	0	79.2	20.8	100	0	79.2	20.8
73	100	0	100	0	100	0	100	0

#### 4.9. Comparison with Previous Studies

Table 6 includes a detailed comparison between our methodology and the previous studies in the field of breast cancer prediction.

Table 6 confirms the fact that our proposed methodology has the unique feature of defining the cancer prediction score as a percentage rather than a fixed binary score (0/1). The results also show the high performance of our probabilistic-based ensemble model against all other previous studies. The big dataset size, the hyperparameters optimization and the range-based score mechanism participate in achieving this high performance.

## 5. Conclusion

Breast cancer prediction is one of the most challenging fields of medical engineering. A novel range-based breast cancer prediction model is designed in the current research. The BCSC dataset is used and analysed using a probabilistic model to define the final prediction value of each case of the dataset. This new final score is used to update the BCSC dataset. The new version of the dataset is used to train an ensemble learning model using the Bayesian hyperparameters optimization method. The training process is performed in two scenarios; one includes the whole data-

**Table 6**

Comparison between our range-based cancer model and previous studies

Reseracher	Methods	Used dataset	Output cancer score	Results/ Limitations
Ramkumar et al. [25]	Naïve Bayes classifier	BUPA research lab 20 cases	Yes or No	Accuracy: 50% Low dataset size, low accuracy. No parameter tuning.
Al-Jawad et al. [3]	SVM, Bayesian network	Haberman's survival 306 cases	Yes or No	SVM: Recall: 73.78% Precision: 74.77% BM: Recall: 78.22% Precision: 64.47% Fixed learning parametes. low datset size.
Yang et al. [30]	Bayesian models Markov models ANN	36,000 cases	Yes or No	Accuracy: ANN: 73.55% CBM: 76.07% Fusion: 75.63% Low F1-score due to the nature of their dataset.
Li et al. [22]	Logistic Regression, SVM, Naïve Bayes, and Bayesian Network	154899 samples	Yes or No	Naïve Bayes has the best accuracy. No prarameter optimization.
Kurian and Jyothi [19]	K-NN, SVM, Naïve Bayes, AdaBoost, Decision Trees, Random Forests, gradient boosting	NCBI dataset 1580 samples	Yes or No	Descion trees accuray: 94.3% No parameter optimization. Low dataset size.
Our Previous Study [17]	Weghting model Decision Trees	BCSC dataset 317880 samples	Yes or No	Accuracy: 95.8% Fixed cancer score
This study	Naive Bayes probabilistic model Ensemble learning Hyperparameter optimization.	BCSC dataset 317880 samples	Range-Based score	Original Accuracy: 85.3%, ±1 variation: 91.12% ±2 Variation 91.33%

set, while the other uses a subset consisting of 67633 samples. In both scenarios, the MCE, TPR, PPR and FDR are computed in three cases; the first case is the 0-variance in which no error-margin is allowed, while for the second and third cases,  $\pm 1$  classes-variance tolerance is applied (The very closed subclasses give similar results). The results indicate TPR, PPR and accuracy improvements for the ( $\pm 1$  and  $\pm 2$ ) variance cases for the sub and whole dataset. Furthermore, the new modified version of the BCSC dataset is more robust and has much detailed information about the prediction of breast cancer, unlike the old version that reveals only the presence of cancer without any percentage.

## References

- Ahmad, A. S., Mayya, A. M. A New Tool to Predict Lung Cancer Based on Risk Factors. *Heliyon*, 2020, 6(2), 1-9. <https://doi.org/10.1016/j.heliyon.2020.e03402>
- Alghunaim, S., Al-Baity, H. H. On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context. *IEEE Access*, 2019, 7, 91535-91546. <https://doi.org/10.1109/ACCESS.2019.2927080>
- Aljawad, D. A., Alqahtani, E., Ghaidaa, A. K., Qamhan, N., Alghamdi, N., Alrashed, S., Olatunji, S. O. Breast Cancer Surgery Survivability Prediction Using Bayesian Network and Support Vector Machines. In *2017 International Conference on Informatics, Health and Technology (ICIHT) 2017*, 1-6. <https://doi.org/10.1109/ICIHT.2017.7899000>
- Annemieke, W., Nane, G. F., Vliegen, I. M., Siesling, S., IJzerman, M. J. Comparison of Logistic Regression and Bayesian Networks for Risk Prediction of Breast Cancer Recurrence. *Medical Decision Making*, 2018, 38(7), 822-833. <https://doi.org/10.1177/0272989X18790963>
- Anusuya, V., Gomathi, V. An Efficient Technique for Disease Prediction by Using Enhanced Machine Learning Algorithms for Categorical Medical Dataset. *Information Technology and Control*, 2021, 50(1), 102-122. <https://doi.org/10.5755/j01.itc.50.1.25349>
- Barlow, W. Breast Cancer Surveillance Consortium dataset 2006, Retrieved from <https://www.bcsc-research.org/data/rfdataset>. Accessed on June 1, 2021.
- Barlow, W. E., White, E., Ballard-Barbash, R., Vacek, P. M., Titus-Ernstoff, L., Carney, P. A., Kerlikowske, K. Prospective Breast Cancer Risk Prediction Model for Women Undergoing Screening Mammography. *Journal of the National Cancer Institute*, 2006, 98(17), 1204-1214. <https://doi.org/10.1093/jnci/djj331>
- Benjamini, Y. Discovering the False Discovery Rate. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 2010, 72(4), 405-416. <https://doi.org/10.1111/j.1467-9868.2010.00746.x>
- Bradley, A. P. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, 1997, 30(7), 1145-1159. [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2)
- Che, D., Liu, Q., Rasheed, K., Tao, X. Decision Tree and Ensemble Learning Algorithms with Their Applications in Bioinformatics. *Software Tools and Algorithms for Biological Systems*, 2011, 191-199. [https://doi.org/10.1007/978-1-4419-7046-6\\_19](https://doi.org/10.1007/978-1-4419-7046-6_19)
- Cruz, J.A., Wishart, D.S. Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics*, 2006, 2(1). <https://doi.org/10.1177/117693510600200030>
- Das, S., Rai, A., Merchant, M. L., Cave, M. C., Rai, S. N. A Comprehensive Survey of Statistical Approaches for Differential Expression Analysis in Single-Cell RNA Sequencing Studies. *Genes*, 2021, 12(12), 1-29. <https://doi.org/10.3390/genes12121947>
- Eroglu, I., Sevilimedu, V., Park, A., King, T. A., Pilewskie, M. L. Accuracy of the Breast Cancer Surveillance Consortium Model Among Women with LCIS, 2021, 190(3), 1-20. <https://doi.org/10.21203/rs.3.rs-873932/v1>
- Fang, R., Pouyanfar, S., Yang, Y., Chen, S. C., Iyengar, S. S. Computational Health Informatics in the Big Data Age: A Survey. *ACM Computing Surveys (CSUR)*, 2016, 49(1), 1-36. <https://doi.org/10.1145/2932707>



15. Greener, J. G., Kandathil, S. M., Moffat, L., Jones, D. T. A guide to Machine Learning for Biologists. *Nature Reviews Molecular Cell Biology*, 2022, 23(1), 40-55. <https://doi.org/10.1038/s41580-021-00407-0>
16. Guo, Z., Xu, L., Asgharzadeholiaee, N. A. A Homogeneous Ensemble Classifier for Breast Cancer Detection Using Parameters Tuning of MLP Neural Network. *Applied Artificial Intelligence*, 2022, 36(2), 1-21. <https://doi.org/10.1080/08839514.2022.2031820>
17. Khozama, S., Mayya, A. M. Study the Effect of the Risk Factors in the Estimation of the Breast Cancer Risk Score Using Machine Learning. *Asian Pacific Journal of Cancer Prevention*, 2021, 22(11), 3543-3551. <https://doi.org/10.31557/APJCP.2021.22.11.3543>
18. Kingsford, C., Salzberg, S. L. What are Decision Trees. *Nature Biotechnology*, 2008, 26(9), 1011-1013. <https://doi.org/10.1038/nbt0908-1011>
19. Kurian, B., Jyothi, V. L. Breast Cancer Prediction Using an Optimal Machine Learning Technique for Next Generation Sequences. *Concurrent Engineering*, 2021, 29(1), 49-57. <https://doi.org/10.1177/1063293X21991808>
20. Larner, A. J. Paired Measures. In *The 2x2 Matrix*, Springer, Cham, 2021, 15-47. [https://doi.org/10.1007/978-3-030-74920-0\\_2](https://doi.org/10.1007/978-3-030-74920-0_2)
21. Lévesque, J. C., Gagné, C., Sabourin, R. Bayesian Hyperparameter Optimization for Ensemble Learning. *arXiv preprint arXiv:1605.06394*, 2016.
22. Li, M., Nanda, G., Sundararajan, R., Evaluating Different Machine Learning Models for Predicting the Likelihood of Breast Cancer. *Advanced Aspects of Engineering Research*, 2021, 2, 132-142. <https://doi.org/10.9734/bpi/aaer/v2/1651C>
23. Mate, Y., Somai, N. Hybrid Feature Selection and Bayesian Optimization with Machine Learning for Breast Cancer Prediction. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2021, 1, 612-619. <https://doi.org/10.1109/ICACCS51430.2021.9441914>
24. Mishina, Y., Murata, R., Yamauchi, Y., Yamashita, T., Fujiyoshi, H. Boosted Random Forest. *EICE Transactions on Information and Systems*, 2015, 98(9), 1630-1636. <https://doi.org/10.1587/transinf.2014OPP0004>
25. Ramkumar, N., Prakash, S., Kumar, S. A., Sangeetha, K.: Prediction of Liver Cancer Using Conditional Probability Bayes Theorem. *International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2017, 1-5. <https://doi.org/10.1109/ICCCI.2017.8117673>
26. Savić, M., Kurbalija, V., Ilić, M., Ivanović, M., Jakovetić, D., Valachis, A., Kosmidis, T. Analysis of Machine Learning Models Predicting Quality of Life for Cancer Patients. In *Proceedings of the 13th International Conference on Management of Digital EcoSystems*, 2021, 35-42. <https://doi.org/10.1145/3444757.3485103>
27. Senerath, J., Don, M., Chinthaka, A., Ganegoda, G. U. Involvement of Machine Learning Tools in Healthcare Decision Making. *Journal of Healthcare Engineering*, 2021, 1-20. <https://doi.org/10.1155/2021/6679512>
28. Xia, Y., Liu, C., Li, Y., Liu, N. A Boosted Decision Tree Approach Using Bayesian Hyper-parameter Optimization for Credit Scoring. *Expert Systems with Applications*, 78, 2017, 225-241. <https://doi.org/10.1016/j.eswa.2017.02.017>
29. Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology*, 2019, 292, 60-66. <https://doi.org/10.1148/radiol.2019182716>
30. Yang, C., Yang, J., Liu, Y., Geng, X.: Cancer Risk Analysis Based on Improved Probabilistic Neural Network. *Frontiers in Computational Neuroscience*, 14, 2020, 14-58. <https://doi.org/10.3389/fncom.2020.00058>
31. Zhang, C., Ma, Y. *Ensemble Machine Learning: Methods and Applications*. Springer Science Business Media, 2012. <https://doi.org/10.1007/978-1-4419-9326-7>

