Prediction of Malignant Breast Cancer Cases using Ensemble Machine Learning: A Case Study of Pesticides Prone Area

Nishtha Hooda, Ruchika Gupta, and Nidhi Rani Gupta

***Abstract*—Cancer of the female breast is one of the leading types of cancers worldwide. This paper presents a case study of Malwa Belt in India that has witnessed the proliferation in the overall mortality rate due to breast cancer. The paper researches mortality aspect of the disease and its association with the various risk parameters including demographic characteristics, percentage of pesticides residue present in the water and soil, life style of the women in the affected area, water intake, and the amount of pesticide exposure to the patient. The levels of organochlorine pesticides like DDT and its metabolites and isomers of HCH in blood, tumor and surrounding adipose are estimated. Additionally, an extent of exposure of the subjects to environmental pollutants like heavy metals (Lead, Copper, Iron, Zinc, Calcium, Selenium, and Chromium etc.) are also examined. For the obtained experimental data, an efficient ensemble machine learning based framework called *Bagoost* is proposed to predict the risk of breast cancer in Malwa women. The performance of the proposed machine learning model results in an accuracy of 98.21%, when empirically tested using K-fold cross validation over the real time data of malignant and benign cases and is established to be efficacious than the existing approaches.**

***Index Terms*—Machine Learning, Breast Cancer, Prediction, Risk, Ensemble, Malignant, Benign.**

I. Introduction

Breast cancer is highly prevalent cancer in females and accounts for the second highest number of deaths worldwide [1]. Punjab was selected to be the first state for the Green Revolution in the late 60's by the Indian government against producing the highest harvest yield. High amount of synthetic fertilizers and pesticides were used by the farmers to enhance the overall crop yield. Consequently, Green Revolution got heavily backed up with the high usage of chemical fertilizers, pesticides, and herbicides that unquestionably increased the production rate of the harvest yield by more than double in count, however gave a frightening by-product in terms of numerous hazardous health disorders that subsequently faced by later generations. The pesticides were used frequently during green revolution however, after green revolution some of them got banned but found persisting in human tissues showing either banned chemicals are still being used without license or due to some bio accumulation the concentration of them have been magnifying in human body generation by generation.

Malwa region includes the city of Patiala, Ferozepur, Nabha, Faridkot, and Ludhiana and is also known as the *cancer belt of India* due to the maximum of cancer patient reported in the region caused by the imprudent use of pesticides by the farmers for growing more cotton that in turn gradually emanated into largely uncurbed water contamination [2]. This specific region of Punjab reported a sharp increase in the number of cancer patients in recent years and the severity of the stance can be observed by the fact that a train runs from Bathinda to Bikaner is called as 'Cancer Express' by the natives as it brings a large number of cancer patients from Malwa to Rajasthan for the treatment of this dreaded disease. Cumulative exposure to pesticides may come from the food, water, air, dust, soil, and the like while pesticides can be absorbed through skin contact, inhalation, or accidental ingestion.

Existing cumbersome, expensive, and time consuming diagnosis techniques for the detection of breast cancer potentiates the need for the development of a novel, deterministic framework to predict the risk of breast cancer. With the significant influx of attention towards the prediction problems and the high activity of advancements in the machine learning field in recent years, the methods based on machine learning are considered to be the most suited to design the prediction model [3].

In this paper, we seek to develop a framework that predicts the risk of breast cancer by assessing the impact of organochlorine and heavy metals exposure on the malignant/ benign breast disease cases. Machine Learning models are proposed to train with the data collected from the malignant and benign cancer patients of Malwa belt, Punjab in India.

In particular, the research is carried out to build an ensemble machine learning model to predict the risk of the disease with legitimate level of accuracy while gathered high-dimensional complex data supports the best feature rankers implementation. The main contribution of the study is as follows:

- The levels of organochlorine pesticides like DDT and its metabolites and isomers of HCH in blood, tumor, and surrounding adipose are estimated for the women suffering from benign and malignant growth of the disease.
- The extent of exposure of the subjects to environmental pollutants like heavy metals (lead, copper, iron, zinc,

Dr. Nishtha Hooda is working with the School of Computing, Indian Institute of Information Technology (IIIT), Una, India. e-mail: 27nishtha@gmail.com

Dr. Ruchika Gupta is with the Department of Computer Science and Engineering, Chandigarh University and Postdoctoral Fellow at Indian Institute of Technology (IIT), Guwahati, India. e-mail: rgupt009@gmail.com

Dr. Nidhi Rani Gupta is working with the Department of Chemistry, Multani Mal Modi College, Patiala, India. e-mail: nidhigupta0508@gmail.com

calcium, Selenium, Chromium etc.) are estimated by determining their blood/breast tissue levels.
- Machine learning techniques are explored for the obtained data sets and an ensemble based prediction model for predicting the risk of breast cancer is designed and implemented.
- The effectiveness of proposed prediction framework is validated using K fold cross validations technique.
- A web-based application that helps the common people, researchers, and research community while providing the cancer risk information is proposed to be developed.

The presented study extends the systematic identification and prediction of the risk of breast cancer by incorporating the effect of metal exposure, pesticides, and other risk factor. The novelty of the research work lies in the fact that the training data employed for building the ensemble machine learning based prediction model is generated by determining the extent of exposure of the subjects to heavy metals, DDT, metabolites of DDT, and isomers of HCH in blood/ breast tissue levels, thus contributing into the advancement in the state-of-the-art breast cancer prediction techniques.

The rest of the paper is organized as follows: Section 2 highlights the related work. Section 3 discusses the data set with the description of experimental setup used for the empirical evaluation. Section 4 presents the proposed framework and the methods used in the proposed framework. Section 5 summarizes the performance evaluation and cross validation results. Finally, Section 6 discusses the conclusion and future scope of the work.

II. Related Work

Breast cancer is a malignant growth disease in which out of control cell growth begins in the tissue of breast. Malignancy in biological terms described as the ability of a group of cells to divide progressively, free of homeostatic control, invade, form distant metastasis, and eventually kills the host. Breast cancer is reported as one of the most frequent cancers among women especially in India. Mortality rate due to the breast carcinoma among Indian females reached to 12.7 per 1,000,000 females and upon observing the rapid increase it is expected that the number of deaths is likely to reach as high as 1,797,900 by end of the year 2020 [4]. According to the latest global cancer data by World Health Organization (WHO) 2019, 18.1 million new cancer cases and 9.6 million deaths in 2018 have been estimated. WHO analysis also reported that one out of six women develop cancer during their lifetime and one in eleven women die from the disease globally. Mentioned scenario is pretty alarming in terms of its severity not only for the developing nations but also for the developed ones [5] [6]. There are various factors considered to be the potential contributors in this deadly disease including age of menarche/ menopause, parity, age at first child birth, duration of lactation, BMI, and the like, however besides them co-founders environmental pollutants like organochlorines such as DDT, HCH and their metabolites [8], and polychlorinated biphenyles (PCBs) [8] have been shown adverse effect on human health. These chemical pollutants frequently been used in vector control and in agriculture to increase the overall crop production. Due to the lipophilicity and estrogen receptor nature these chemicals accumulate in the various body tissues and their concentration increase at faster rate due to the biomagnifications which results into breast carcinoma [9]. Evidently, organochlorines are well known xenoestrogens mimicking estrogen activity and promote breast tumors [9] [10]. Machine learning offers computational methods to study and analyze the potential contributors more intelligently and such methods can instrumentalize the prediction against the likelihood of discussed deadly disease occurrence beforehand. Researchers around the globe employing machine learning algorithms to perform insightful analytics that can predict the persistence of significant pain in the person [11]. Different clinical and psychological parameters over 1000 demography are required to be studied in the analysis [12]. Such big and complex data can also be efficiently handled by deep learning models in the research [13].

III. Experimental Investigation

This section presents details of the data set studied and the experimental setting used for investigation.

A. Data Set

The data in study is collected from the women suffering from benign and malignant growth of breast disease of Malwa Region, Punjab. The subjects are examined by an expert and a filled questionnaire is taken from 86 subjects of malignant growth and 42 subjects of benign growth class of the disease. Levels of organochlorine pesticides like DDT and its metabolites and isomers of HCH in blood, tumor and surrounding adipose in the women suffering from benign and malignant growth of disease are estimated. Additionally, the extent of exposure of the subjects to the environmental pollutants like heavy metals (Lead, Copper, Iron, Zinc, Calcium, Selenium, and Chromium etc.) are estimated by determining their blood/ breast tissue levels.

B. Experimental Setting

The R 'caret' package is used to implement the various pre-processing and model building techniques. One shot training and testing technique is adopted in the process. Experiments are designed to use 10 fold cross validation method in which data set is divided into 10 equal sized subsets. Different machine learning algorithms are then selected as the base classifier to train the remaining nine subset folds and testing is performed on the last fold. To examine the robustness of designed framework the same process is iterated. In order to evaluate the proposed framework six different parameters namely accuracy, sensitivity, specificity, MCC, F-Measure, and area under curve (AUC) are used as performance matrices.

IV. Proposed Framework

The prediction of malignant and benign cases of breast cancer is performed by considering various parameters using

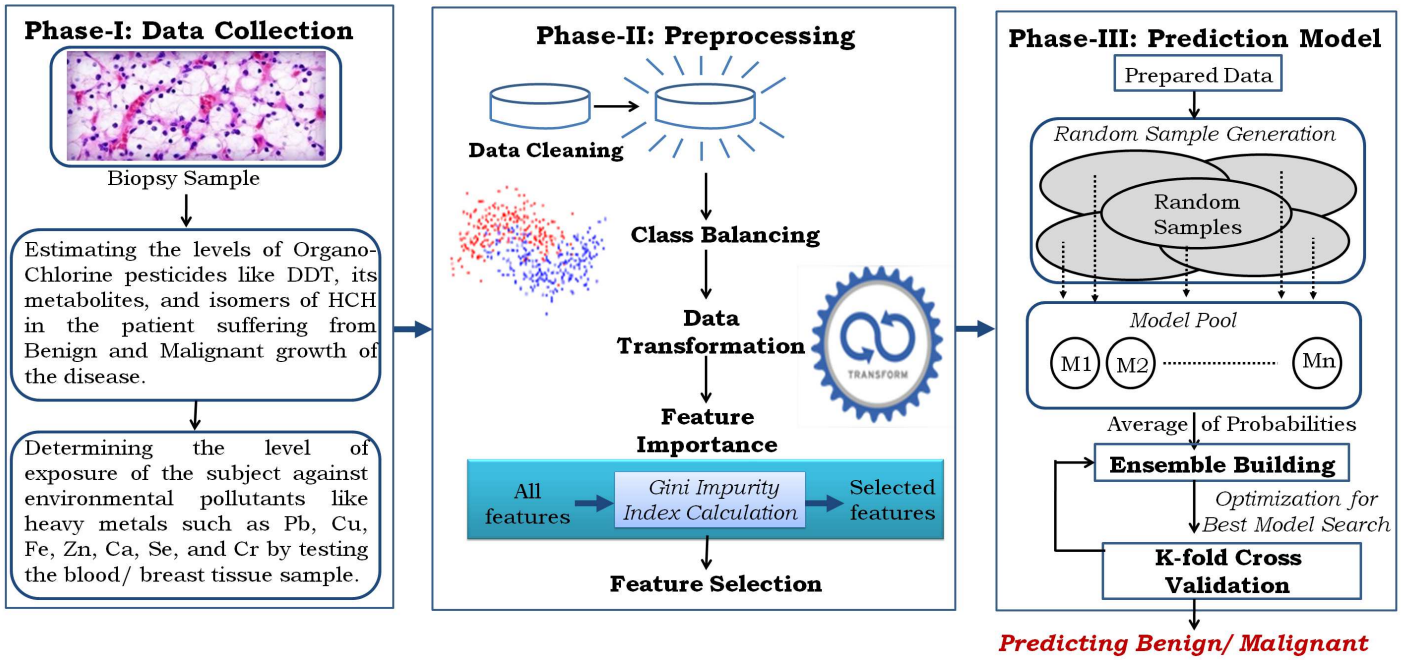


Fig. 1. Bagoost Framework for Prediction of Malignant and Benign Cases of Breast Cancer

an ensemble machine learning based Bagoost Framework presented in Fig. 1 where the proposed framework is described in three phases. The first phase focuses on the data collection and management of the experimental data of biopsy samples. The second phase works on applying machine learning techniques for pre-processing of the questionnaire and experimental data received from phase 1, while the third phase details out the design and development of the prediction model. Ensemble machine learning technique is proposed to get implemented in the final phase to optimize the prediction results. Details of all three of the phases are explained further below.

A. Collection and Management of Biopsy Samples

Environmental exposure of organochlorine pesticides has been suggested to be significant determinants in the development of breast cancer tissue. Despite the extensive amount of study, only a small number of consistent findings between organochlorine pesticides and breast cancer have emerged. Maintaining the adequate balance between merits and demerits, the framework is designed with the following steps:

To begin with, medical collaborators are associated to obtain the desired samples. Standardization of the analytical parameters is decided while recruitment of the study subjects and assessment of analytical parameter is selected. Collection, processing, and analysis of breast tissue samples is performed thereafter. In parallel, demographic data is also collected both from the control group (benign case of the disease) and the study group (malignant case of the disease). Analysis of various pollutants (pesticides and metals) is done further.

1) Sample Size and Selection of Subjects: The study of around 86 samples from malignant group and 42 from benign group is statistically investigated. All the subjects in both the groups namely; control and study who gave the consent for surgery are included for the investigation. Subjects are interviewed and aligned to a structured questionnaire. The inclusion criteria for the study are females with a palpable lump in the breast where a biopsy/ surgical excision treatment is planned and consented by the subject.

2) Survey Tools: Questionnaire for socio-demographic status and clinical examination including Age, Weight, Height, BMI, Residence (Rural/ Urban), Addiction (Smoking/ Tobacco/ Alcohol), Menstrual Status, Age of Menarche, Age of Menopause (if achieved), Hyper estrogenic state, Duration of lactation, Number of children, and Dietary habit (veg/ non-veg) etc. are studied.

3) Sample Collection, Transportation, and Storage: A blood sample (approx. 3 ml) is taken by the medical collaborator and collected in the pre-heparinized vials. Further a 1 gm of tissue from the excised breast lump is collected in high-density bottles. Samples are then transported in the ice-cold condition and stored at -200°C till analysis.

4) Determination of Environmental Pollutants: In order to determine the environment pollutants following steps are taken:

a. *Extraction and Cleanup of Samples*: Extraction of pesticide residues are carried out by the modified method [14] described by [15] using n-hexane as extraction solvent.
b. *Quantification of Pesticides*: Samples are then analyzed by Gas Liquid Chromatograpy under the conditions described by [16].

c. *Identification of Pesticides*: Identification of pesticide residue extracted from the samples are done using Gas Chromatography Mass Spectrometry (GCMS)/ Dual column gas chromatography.

5) Determination of Metals:

a. Digestion of samples is performed initially.
b. *Bio-monitoring of Lead on AAS (GTA):* In Blood samples Lead level are estimated by using Graphite furnace atomic absorption spectrophotometer.
c. *Analysis of Metals on AAS (Flame):* Analysis of Fe, Ca, Cu, and Zn in the breast tissue, tumor, and blood are done on Flame atomic absorption Spectrometer.

B. Pre-processing of Data

The hypothesis of the research is to check the relevance of machine learning techniques in predicting risk of the malignancy of breast cancer by training machine learning model with both malignant and benign cases. To achieve the goal, benign and malignant groups are assessed statistically. Following are the steps taken for pre-processing phase and is implemented using machine learning techniques:

1) *Missing Value Imputation*: To get satisfactory training results, missing values in the biopsy data are replaced by the mean value for rough approximation.
2) *Class Balancing*: It is observed that the collected data consists of more samples of malignant cases that makes the data set imbalanced. The prediction model based on the clinical data usually need to be trained on a much larger training cohort which is dealt using data augmentation. Synthetic minority over-sampling technique (SMOTE) algorithm is a popular technique to address the issue of imbalanced class and balances the classes by over-sampling of the minority class [17]. SMOTE algorithm is implemented using Random oversampling involves randomly selecting examples from the minority class, with replacement, and adding them to the training dataset, hence increasing the size of the dataset.
3) *Data Transformation*: The main intent is to transform the data into an enhanced version in order to increase the likelihood of better prediction performance of machine learning classifiers. With minimum distortion of the numeric values in the collected data, normalization is applied here to change the numeric values to a common scale.
4) *Feature Selection*: Feature importance of various risk factors studied as survey tools is calculated using Gini Importance in the proposed Bagoost framework. Gini Importance measure calculates the mean decrease gini, a statistical measure representing the contribution of feature to the homogeneity in the data. Mean decrease in Gini Index with lowest value shows least contribution in the homogeneity in the data and hence can easily be removed.

C. Prediction Model

For training the machine learning models, initially random samples of the data is generated. Random sampling is done to eliminate biasing from the training data. In the next step, different models are trained with random samples. Performance of the models is tested using K fold cross validation method. To optimize the accuracy of the prediction, an ensemble model

Algorithm 1 Adaboost Pseudocode

Considering: $(x_1, y_1), \ldots, (x_m, y_m)$ *where* $x_i \in \mathcal{X}, y_i \in -1, +1$
Initialize: $\mathcal{D}_1(i) = 1/m$ *for* $i = 1, \ldots, m$
1. For t = 1 to $\mathcal{T}$
2. Weak learner training using distribution $\mathcal{D}_r$
3. Obtain h_t: $\mathcal{X} \rightarrow \{-1, +1\}$ //where h_t is weak hypothesis
4. Objective: Choose low weighted error h_t
$\mathcal{E}_t = Pr_{i \sim D_t}[h_t(x_i) \neq y_i]$
5. Select α_t = $\frac{1}{2}$ ln ($\frac{1-\mathcal{E}_t}{\mathcal{E}}$)
6. For i = 1 to m // update $\mathcal{D}_t$
$\mathcal{D}_{t+1}$(i) = $\frac{\mathcal{D}_t(i) exp(-\alpha_t y_i h_t(x_i))}{\mathcal{Z}_t}$
//where $\mathcal{Z}_t$ specifies factor of normalization
7. Obtain $\mathcal{H}(x)$ = sign ($\sum_{t=1}^{\mathcal{T}} \alpha_t h_t(x)$)

Algorithm 2 Bagoost Pseudocode

1. Collect training sampled data consisting of R features.
2. Missing values in the biopsy data are replaced by mean value for rough approximation.
3. Over-sampling of the minority class is performed using SMOTE algorithm.
4. Rank the features using Gini index value to measure the importance of features.
5. Split the data into bootstrap samples for independently sampling with replacement from initial sampled data.
6. Call Algorithm 1 to train Adaboost Model using n bootstrap samples.
7. Combine the prediction of n models using majority voting.
8. Outcome of the prediction is winning class in the majority voting.

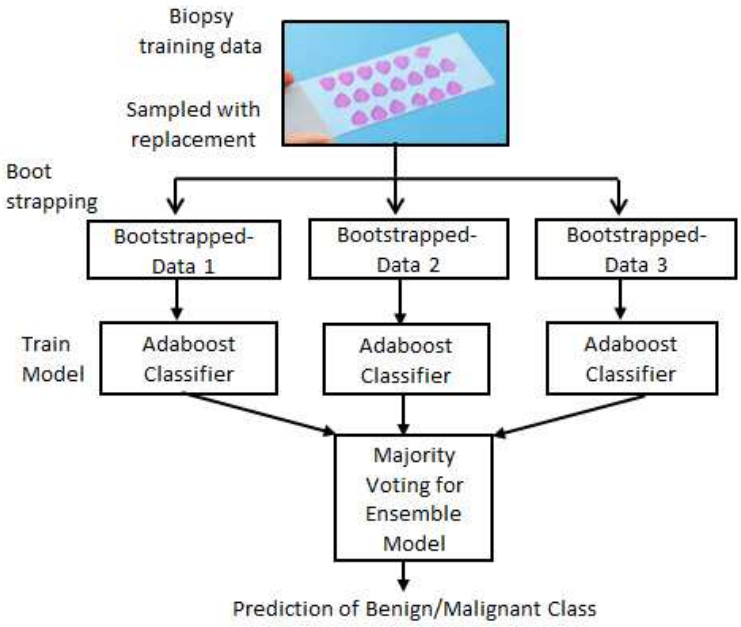


Fig. 2. Bagoost Ensemble Model for the Prediction of Malignant and Benign Cases of Breast Cancer

building is performed. Different combinations of the models are created using the pool of ten machine learning models namely, Bayes Net, Naive Bayes, Support Vector Machine, Neural Network, Logistic Regression, Adaboost, Decision Tree, C 4.5, Decision Stump, and Random Forest. These ensembles are trained using random samples of the data. During experimentation, the performance of built ensemble is evaluated using various performance metrics like accuracy, true-positive rate, false positive rate, area under the curve, etc. After experimentation, the combination of bagging based Random Forest and Adaboost [18] performed with the best performance and give with the name Bagoost ensemble model as shown in the Fig. 2. The steps of Adaboost and Bagoost Models are presented in Algorithm 1 and 2, respectively.

Adaboost (stands for adaptive boosting) machine learning algorithm is a meta-mechanism. It has a strong merit to be able to integrate itself with various other learning models in order to optimize the overall performance. It allows to achieve the same by gradual addition of relatively weaker models in a sequential fashion that are trained with weighted data. The process of addition carried out till weak learners specified count is reached and no subsequent optimization can be further introduced over the given training set. Let R represents the total number of risk factors used for building the breast cancer prediction model. Class represents the benign and malignant outcome of the prediction. An 'm' sample training set is taken with $\mathcal{X}$ be the total set of x inputs while y be the output described with the set having two values, -1 and +1. Weights of all the samples given by $\mathcal{D}$ are initialized by 1/m. Upon defining the hypothesis, for all the classifiers ranging from 1 to $\mathcal{T}$ selection of those classifiers are done that have the lowest weighted classification error (i.e. error rate must be above 0.5). Selection of the weights are done as per classifier α and weight updation is performed in the subsequent step. It is described that in case of wrong classification, the exponential entity becomes greater than 1 while in the right classification case it comes lower than 1.

V. Results and Discussion

In this section, results of feature importance and selection are presented at the beginning. To validate the proposed framework, experimental results are highlighted using K fold cross validation technique and the performance of proposed ensemble model is also compared with the state-of-the-art algorithms.

A. Feature Importance and Analysis

Results of feature importance on the basis of their Gini Importance are presented in the Table I. It can be observed that the histology risk factor has contributed with the maximum importance while other risk factors like the number of residing years in the region, age of the women, and marriage also contribute with relatively higher importance. The prediction performance of the proposed model is calculated using different feature subsets (top 3, top 5, top 10, top 15, top 20, top 30 of the Table I) and presented in Table II. The top three features in the Table I gives the best performance. Performance of ensemble model after feature selection comes out to be 98.20% which is bit improved by employing random sampling at the time of ensemble building. This shows that employing feature selection helps in improving the complexity of model without much compromising the predictive performance. This is also useful in improving the complexity of the web service and simplifying the graphical interface of the breast cancer predictor. The detailed analysis of the risk factors is presented as follows.

TABLE I
Feature Importance of various risk factors studied in questionnaire for the malignancy prediction of breast cancer

Feature	Gini Importance
Histology	7.36
Residing Years	2.73
Age	2.02
Age of Marriage	0.86
Husband Income	0.45
Age of Menopause	0.48
No of children	0.67
Work From Hand	0.20
Side of lump	0.21
Age at First Child Birth	0.38
Height	0.57
Occupation	0.06
Healthy Dietary Habits (HDH)	0.17
Rapid Weight Loss (RWL)	0.03
Blood Donation	0.17
BMI	0.74
Pesticide Exposure (PE)	0.11
Abode	0.11
Lump Size	0.45
Time to Diagnosis Months	0.40
Contraceptive Pills (CP)	0.03
Source of Drinking Water	0.23
Wear Bra (WB)	0.16
Duration of Lactation Months	0.25
Weight in Kg.	0.66
Marital Status	0.09
Parity	0.11
Chulha Usage (CU)	0.08
Duration	0.14
Heavy Breast (HBR)	0.11
Chronic	0.05
No. of Miscarriages (MC)	0.10
Age of Menarche	0.39
Smoking Habit	0.02
Family history of Breast Lesions	0.02

TABLE II
Comparison of performance of different feature subsets trained using Bagoost model

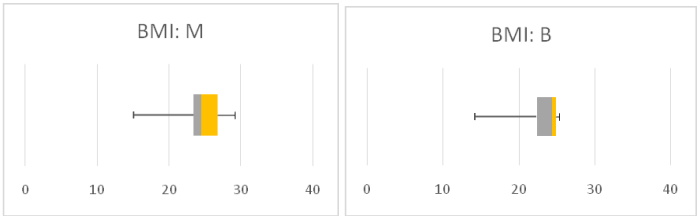
Feature set	TP Rate	FP Rate	F-Measure	MCC	ROC	Accuracy
Top 3	**0.982**	**0.019**	**0.98**	**0.96**	**0.98**	**98.02**
Top 5	0.975	0.026	0.97	0.94	0.95	97.54
Top 10	0.967	0.045	0.967	0.92	0.96	96.72
Top 15	0.967	0.045	0.967	0.92	0.96	96.72
Top 20	0.967	0.045	0.967	0.92	0.96	96.72
Top 30	0.951	0.069	0.951	0.882	0.973	95.12
Bagoost	**0.987**	**0.018**	**0.98**	**0.96**	**0.98**	**98.21**

TABLE III
ANALYSIS OF RISK FACTORS STUDIED IN QUESTIONNAIRE OF MALIGNANT CASES FOR THE BREAST CANCER PREDICTION

S.No	HDH	Abode	OW	Smoke	Tobacco	Alcohol	CU	MP	Married	Child	MC	FHBC	PE	RWI	CP	Educated	BD	PWD	WB	HBR
1	✗	✗	✓	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗
2	✗	✓	✓	✗	✓	✗	✓	✓	✓	✓	✗	✗	✓	✗	✓	✗	✗	✓	✗	✗
3	✓	✓	✗	✓	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
4	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗
5	✓	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓	✗	✓	✗	✗	✓	✗	✓	✓	✓
6	✓	✓	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
7	✗	✓	✗	✗	✗	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗
8	✗	✗	✓	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✓	✓
9	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
10	✓	✗	✓	✗	✗	✗	✗	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗
11	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	✓	✓	✓
12	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	✗	✓	✗	✗	✓	✗	✓	✓	✓
13	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
14	✓	✗	✓	✗	✗	✗	✓	✗	✓	✓	✗	✗	✗	✗	✗	✓	✗	✓	✗	✗
15	✓	✗	✓	✗	✓	✗	✓	✗	✓	✓	✓	✓	✗	✓	✗	✓	✗	✗	✓	✓
16	✓	✗	✓	✓	✓	✗	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗
17	✗	✓	✓	✗	✗	✗	✗	✓	✓	✓	✗	✗	✓	✗	✗	✓	✗	✓	✗	✗
18	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓	✓	✗	✗	✓	✗	✓	✓	✓
19	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗	✗	✓	✗	✓	✗	✗
20	✗	✓	✓	✗	✗	✗	✗	✓	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓	✓
21	✓	✗	✗	✗	✗	✗	✓	✗	✓	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓	✓
22	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	✓	✓	✓
23	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗
24	✓	✗	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓	✓	✗
25	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓
26	✗	✗	✓	✗	✗	✗	✓	✓	✓	✗	✓	✗	✓	✗	✗	✗	✓	✗	✓	✗
27	✓	✗	✗	✗	✗	✗	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
28	✓	✓	✓	✗	✗	✗	✗	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓	✓
29	✓	✓	✗	✗	✗	✗	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓
30	✗	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✓	✓
31	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗	✗	✓	✓	✓	✗	✗
32	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗
33	✓	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓	✗	✗	✗	✓	✗	✓	✗	✓	✓
34	✓	✓	✓	✗	✗	✗	✗	✓	✓	✓	✗	✗	✓	✗	✗	✗	✓	✗	✓	✓
35	✓	✓	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
36	✗	✓	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	✗	✓	✓	✓	✓	✓
37	✓	✓	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
38	✓	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓	✗	✓	✗	✗	✗	✗	✓	✓	✗
39	✗	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✓	✗
40	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓
41	✗	✓	✗	✗	✗	✗	✗	✓	✓	✗	✓	✗	✓	✗	✗	✓	✗	✓	✓	✗
42	✓	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✓	✗
43	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗
44	✓	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓	✓
45	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗
46	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗
47	✓	✗	✓	✗	✗	✓	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
48	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗
49	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗
50	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗

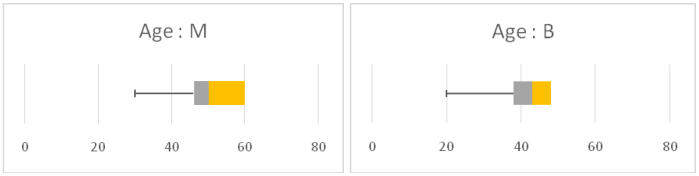
Table III describes the analysis of the malignant cases against those attributes that have observed prominently in the sufferers. In reference to the attribute pesticide exposure (PE), it is statistically registered that 74% of malignant cases received exposure to the usage of pesticides for a lengthier time-span, while the women cases here pesticides exposure are registered lesser have displayed some irregularities in other attributes. It is also prominently displayed in the recorded data that another major factor is non-education of the cases under study. Total of 66% of the women are found uneducated that contributed in their hindrances showing the case to doctors for examination at early stage of the disease development. 72% of women exhibit the healthy dietary habits (HDH) and still suffer from the disease, however 28% of the women cases displayed non-healthy dietary habits that contributes into the non-prevention of disease. As per the report, sufferers have the high risk probability if some family history of breast cancer is present, however on the contrary in studied data, it is clearly noticed that the malignant cases are dominated by suffers exposure to the pesticide. In order to analyze it further, Table III showcases the analyzed study with a red and green color variation against every entry in order to incorporate the heat map effect of the results obtained to enhance its readability. It is clearly observed that the majority of the sufferers are married and mother of at least one child. The red color concentration over the column PE exhibits the profound prominence of the attribute to develop the disease at first place. Other publicly proclaimed factors like smoking, tobacco, and alcohol hold almost negligible impact as a cause to induce the disease.

Box plot is one of the best depiction techniques to show specific features for the data analysis. In this study three of the features from the captured data are shown namely; BMI, Age, and Number of years person residing in the affected region in the Fig. 3, Fig. 4, and Fig. 5 respectively. As per the study of the National Institutes of Health (NIH) a person is



(a) BMI of Malignant Cases (b) BMI of Benign Cases

Fig. 3. BMI comparison of Malignant and Benign cases of Breast Cancer



(a) Age of Malignant Cases (b) Age of Benign Cases

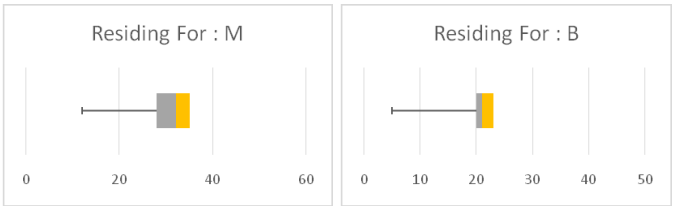
Fig. 4. Age comparison of Malignant and Benign cases of Breast Cancer

considered underweight if the BMI is lesser than 18.5, whereas the person with BMI between 18.5 and 24.9 is considered to be the ideal case. BMI beyond 25 is claims to be the overweight case. It is observed in the Fig. 3 that by and large majority of the women (Malignant and Benign both) in the sample are above 20 BMI and lesser number of woman fall into 'underweight' category. Moreover, the Malignant cases demostrate the inclination towards higher range of BMI. Benign cases exhibit the inclination towards ideal side of the BMI while the diseased cases are found in the extreme end of the range shown through the error bar in the Fig. 3 (a).

Second profound parameter 'Age' also demonstrates its significance and plotted in Fig. 4. Ageing is found to be one of the most prominent risk factors for various type of cancer disease. According to the statistical analysis depicted, it is observed that the median age of the breast cancer diagnosis is 50 years and one quarter of the registered breast cancer are diagnosed in the women aged from 50 to 60, however, the disease can occur at any age between as low as 30 to as high as 60 years.

It is noticed that the majority of women underwent menopaused (MP) state and the attribute miscarriage (MC), Rapid Weight Loss (RWL), and Contraceptive Pill (CP) usage do not contribute significantly. It is also seen that almost 52% of the malignant population are the women consuming purified water (PWD) while the women with heavy breast (HBR) and wears bra (WB) attributes stand half way. Overall, it is critically analyzed that the major attribute responsible for the disease in subject is PE.

The global breast cancer pattern are matched with the study this paper presents and it is alarming to observe that in India the median age of the breast cancer diagnosis dropped down to 50 years while the median age is registered to be 60 years globally. Third feature is 'residing for', describes for how many years the sufferers have been staying in the affected region. Upon the analysis it is turned out to be highly influential parameter among breast cancer sufferers. It can be



(a) Residing years of Malignant Cases (b) Residing years of Benign Cases

Fig. 5. Number of residing years comparison of Malignant and Benign cases of Breast Cancer

TABLE IV
COMPARISION OF PERFORMANCE OF BAGOOST WITH STATE-OF-THE-ART CLASSIFIERS

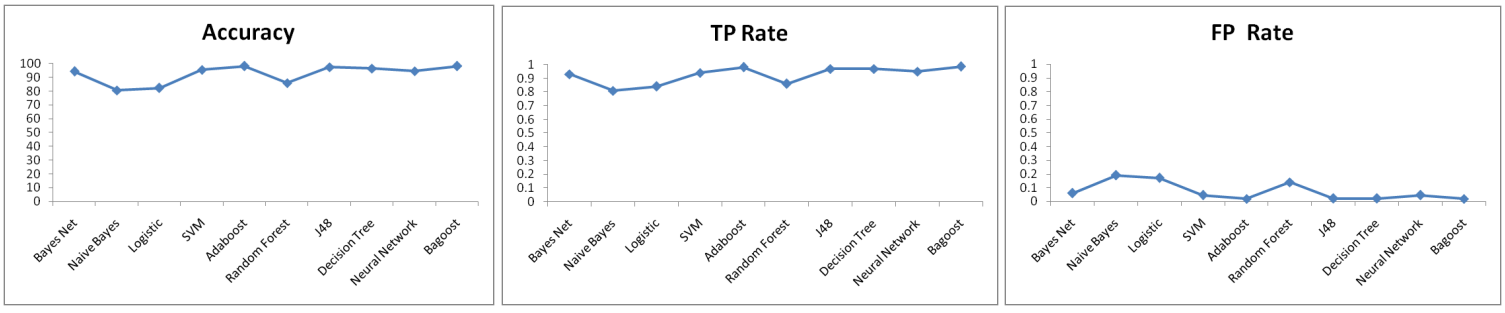
	TP Rate	FP Rate	F-Measure	MCC	ROC	Accuracy
Bayes Net	0.930	0.060	0.94	0.88	0.97	94.20
Naive Bayes	0.810	0.190	0.80	0.61	0.87	80.71
Logistic	0.840	0.170	0.82	0.66	0.89	82.36
SVM	0.940	0.045	0.94	0.90	0.95	95.53
Adaboost	0.982	0.019	0.98	0.96	0.97	98.06
Random Forest	0.860	0.140	0.86	0.74	0.96	85.89
J48	0.970	0.020	0.970	0.94	0.95	97.47
Decision Tree	0.970	0.020	0.97	0.94	0.94	96.47
Neural Network	0.950	0.045	0.95	0.91	0.94	94.53
Bagoost	**0.987**	**0.018**	**0.98**	**0.96**	**0.98**	**98.21**

clearly observed from the box plot (shown in the Fig. 5) that the women who exposed to the region for more than 32 years are more likely to be diagnosed under a malignant case of the disease while the women residing for lesser than 23 years are safe from the disease.

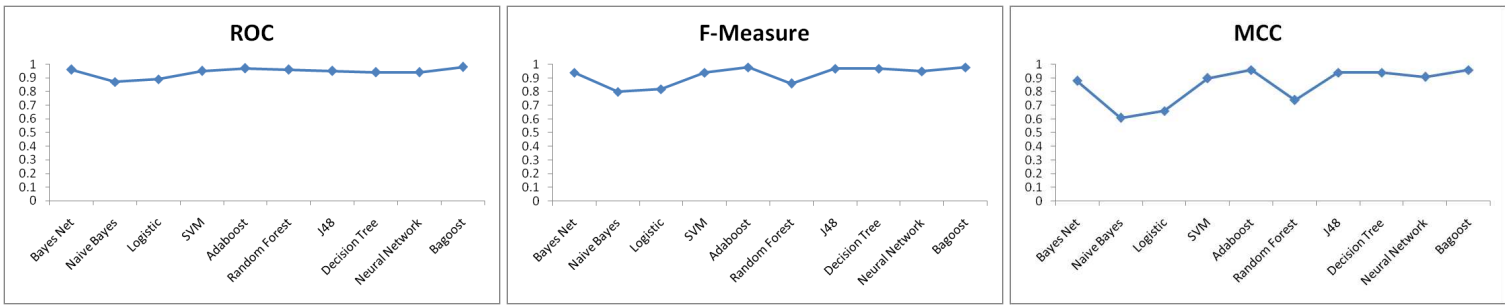
Breast Cancer is not only limited to the ill health burden but also advances the nation to an economic loss. Mostly cancers are caused by an unhealthy lifestyle, stress, and environmental pollutants. Copious risk factors are critically analyzed in the malignant patients as shown in Table III. It is observed in the samples that most of the malignant cases have been exposed to the pesticides for a longer time duration. Although, such women did not consume alcohol and tobacco and tried to adopt an healthy lifestyle but the exposure of pesticides for a longer duration caused a significant loss of health to the patient.

B. K fold Cross Validation

There are various techniques of testing the performance of machine learning algorithms. K fold cross validation method is a statistical based technique that skillfully estimates the performance of model usually with a lower bias than the other methods. Here, K represents the number of splits performed on the data during testing. In this work, value of K=10 is assumed that divides the data into ten equal parts. In the first iteration, first subpart is considered as the test data set and the models are trained using rest of the nine parts. Process is



(a) Accuracy (b) True Positive Rate (c) False Positive Rate



(d) Area under ROC Curve (e) F Measure (f) MCC

Fig. 6. Accuracy, TP rate, FP rate, AUC, MCC, and F Measure results using K fold cross validation using Bagoost ensemble classifier

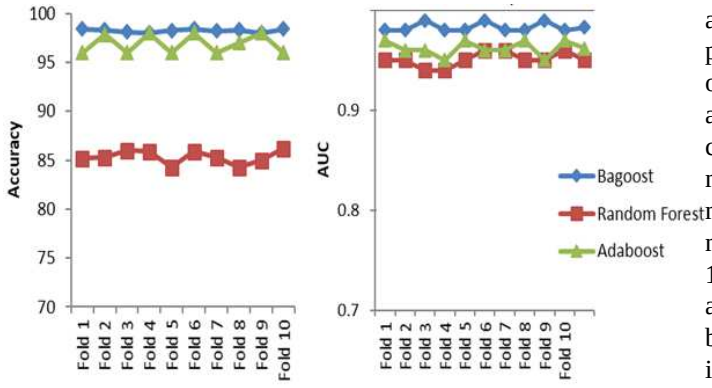


Fig. 7. Performance Comparison of Accuracy and AUC of Bagoost ensemble classifier with Adaboost and Random Forest Ensembles on ten folds of K fold cross validation

repeated in a view that each part gets a chance to be taken as a testing data set. After ten iterations, an average of the individual performance run is computed.

Average performance comparison of all the ten classifiers along with the built ensemble using K fold cross validation through various evaluation metrics is presented in the Table IV. Comprehensive comparative performance analysis of the proposed model with the traditional methods is shown in Fig. 6. Fig. 6 (a) compares the accuracy of models where it can be noticed that Bagoost and Adaboost observe pretty close to each other in terms of prediction accuracy while efficiently predicting the breast cancer cases with an accuracy of 98.21 %. Upon focusing the TP rate, Fig. 6 (b) shows the comparison of True Positives Rate. FP Rate is observed minimal shown in Fig. 6 (c). Although ensemble based Adaboost algorithm also worked with an accuracy of 98.06% but the overall performance of the proposed model is efficacious than the other state-of-the-art methods as when the data is complex and the classes are not completely balanced, Area under ROC curve values helps in selecting the best machine learning model. Observing AUC values of various machine learning models, random forest and Bagoost framework exhibit the more optimized AUC values, closer to the maximum value of 1 as shown in Fig. 6 (d). F-Measure is also observed close to 1 as shown in Fig. 6 (e). To compare the quality of prediction in binary classification, Matthews correlation coefficient (MCC) is used shown in Fig. 6 (f).

From Table IV it is observed that the Bagoost framework report the highest True Positive Rate equivalent to 0.987. The comparison is graphically presented in Fig. 6 (b). Upon empirically examining with numerous models, the most promising candidates to build an ensemble for the prediction of breast cancer are found to be Random Forest and Adaboost. The stability of proposed Bagoost model on ten folds cross validation method is compared with Random forest and Adaboost algorithm as depicted in Fig. 7. The figure shows the performance comparison of Accuracy and AUC of Bagoost ensemble classifier with Adaboost and Random Forest Ensembles on ten folds of K fold cross validation. If the accuracy is focused, it can be analyzed that the accuracy of the Bagoost model is better than the random forest and Adaboost models. Similarly, AUC value of Bagoost model is better than the other two classifiers. Notice that the performance of Bagoost model appears to be stable on all ten folds of K fold cross validation method.

Diverse nature of two ensemble machine learning models employed to build the model can be considered as of the

profound reasons for the outperforming results. In one hand, Adaboost is a popular boosting algorithm supports in reducing the bias and variance. It combines many weak learning decision tree models to make a strong prediction model. It is very sensitive to noise and outliers and specialized for handling high dimensional data. On the other hand, Random is an ensemble of many decision trees and employs bootstrapping. It is specialized in improving the accuracy of the predictions along with maintaining the stability at all fold of K fold testing. The combination of Adaboost and Random Forest gives a flavor of double ensemble machine learning approach. The high predictive performance of double ensemble serve as a proof of utilizing the proposed framework for solving the complex prediction problems.

VI. Conclusion

Breast Cancer is not only limited to cause the ill health burden over the sufferers, it also advances the nation to a significant economic loss. This paper presents a case study carried out in the Malwa region known as the cancer belt of India. In the study, initially the levels of organochlorine pesticides like DDT and its metabolites and isomers of HCH in blood, tumor, and surrounding adipose tissues are estimated for the women suffering from benign and malignant growth of the disease. Overall extent of the subjects exposure to the environmental pollutants like heavy metals (Lead, Copper, Iron, Zinc, Calcium, Selenium, and Chromium etc.) are measured and analyzed by determining their blood/breast tissue levels. It is observed that the majority of cancers are caused by the unhealthy lifestyle, high level of stress, and environmental pollutants like pesticides. The importance of different risk factors in the malignant patients are critically examined using a statistical measure called Gini importance. Machine learning techniques are then employed over obtained data sets and an ensemble based prediction model is developed to predict the risk of breast cancer. Upon validating using K fold cross validation technique, Bagoost framework displays the value of TP rate, FP rate, F measure, MCC, ROC, and accuracy as 0.99, 0.018, 0.98, 0.96, 0.98, and 98.21% respectively.

Building and deploying an online web-based cancer prediction application that takes the responses from user against defined questionnaire and predicts the risk of cancer with legitimate accuracy holds a promising scope for the future work. Said application is viewed to recommend the change of life style, food habits, and other prevention for the risk of environmental factors by the medical practitioners if the risk of is high in the individual.

Acknowledgement

This research work is supported by Department of Science and Technology, Government of India with grant number DST/Disha/SoRF/024/2013.

References

[1] Coughlin SS, Ekwueme DU. Breast cancer as a global health concern. Cancer epidemiology. 2009 Nov 1;33(5):315-8.

[2] Blaurock-Busch, Eleonore, et al. "Comparing the metal concentration in the hair of cancer patients and healthy people living in the Malwa region of Punjab, India." Clinical Medicine Insights: Oncology 8 (2014): CMO-S13410.

[3] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015 Jan 1;13:8-17.

[4] Malvia S, Bagadi SA, Dubey US, Saxena S. Epidemiology of breast cancer in Indian women. Asia-Pacific Journal of Clinical Oncology. 2017 Aug;13(4):289-95.

[5] da Costa Vieira RA, Biller G, Uemura G, Ruiz CA, Curado MP. Breast cancer screening in developing countries. Clinics. 2017 Apr;72(4):244-53.

[6] Ennour-Idrissi K, Ayotte P, Diorio C. Persistent Organic Pollutants and Breast Cancer: A Systematic Review and Critical Appraisal of the Literature. Cancers. 2019 Aug;11(8):1063.

[7] Arrebola JP, Belhassen H, Artacho-Cordón F, Ghali R, Ghorbel H, Boussen H, Perez-Carrascosa FM, Expósito J, Hedhili A, Olea N. Risk of female breast cancer and serum concentrations of organochlorine pesticides and polychlorinated biphenyls: a case–control study in Tunisia. Science of the Total Environment. 2015 Jul 1;520:106-13.

[8] Artacho-Cordon F, Fernández-Rodríguez M, Garde C, Salamanca E, Iribarne-Durán LM, Torné P, Expósito J, Papay-Ramírez L, Fernández MF, Olea N, Arrebola JP. Serum and adipose tissue as matrices for assessment of exposure to persistent organic pollutants in breast cancer patients. Environmental research. 2015 Oct 1;142:633-43.

[9] Anand M, Singh J, Siddiqui MK, Taneja A, Patel DK, Mehrotra PK. Organochlorine pesticides in the females suffering from breast cancer and its relation to estrogen receptor status. Journal of Drug Metabolism and Toxicology. 2013

[10] Gray JM, Rasanayagam S, Engel C, Rizzo J. State of the evidence 2017: an update on the connection between breast cancer and the environment. Environmental Health. 2017 Dec;16(1):94

[11] Lötsch, Jörn, et al. "Machine-learning-derived classifier predicts absence of persistent pain after breast cancer surgery with high accuracy." Breast cancer research and treatment. 2018 171.2:399-411.

[12] Khan, SanaUllah, et al. "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning." Pattern Recognition Letters 125 2019: 1-6.

[13] Hussain, Lal, et al. "Automated breast cancer detection using machine learning techniques by extracting different feature extracting strategies." 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering. IEEE, 2018.

[14] Teale FW, Dale RE. Isolation and spectral characterization of phycobiliproteins. Biochemical Journal. 1970 Jan 1;116(2):161-9.

[15] Saxena MC, Siddiqui MK, Seth TD, Murti CK, Bhargava AK, Kutty D. Organochlorine pesticides in specimens from women undergoing spontaneous abortion, premature or full-term delivery. Journal of analytical toxicology. 1981 Jan 1;5(1):6-9.

[16] Siddiqui MA. Sustainable development through beneficial use of produced water for the oil and gas industry (Doctoral dissertation, Texas A&M University).

[17] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002 Jun 1;16:321-57.

[18] Schapire, Robert E, "Explaining adaboost."Empirical inference. 2013,Springer 37–52

Dr. Nishtha Hooda is currently working as a faculty at School of Computing, Indian Institute of Information Technology (IIIT), Una, India. She received her Ph.D. degree from Computer Science and Engineering Department, Thapar University, Patiala, India in 2019. She received her Master of Engineering degree in Software Engineering from Thapar University, Patiala, India in 2014. Her research interests include artificial intelligence, big data analytics, machine learning, and computational biology.

Dr. Ruchika Gupta is an Associate Professor in the Department of Computer Science Engineering at Chandigarh University, Punjab, India. At present she is associated with Indian Institute of Technology (IIT) Guwahati as a Post Doctoral Research Fellow under ISEA project. She received her Ph.D. degree from the Computer Engineering Department, Sardar Vallabhbhai National Institute of Technology, Surat, India in 2018. Her research interests include Location Privacy, Spatial Anonymity, Secured NoC Design, Mobile Computing, and Machine Learning.

Dr. Nidhi Rani Gupta is an Assistant Professor in the Department of Chemistry, Multani Mal Modi College, Patiala. She received her PhD degree from the Department of Chemistry, Thapar University, Patiala, India in 2009. Her area of research includes Analytical Chemistry, Environmental Chemistry, Metal and Pesticides Toxicology, Breast Cancer Studies, Potentiometric Sensors, Water Soluble Carbon Nanotubes, and Biomonitoring of environmental pollutants such as organochlorines.