International Journal of Advanced Technology and Engineering Exploration, Vol 9(87)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2021.874555

Research Article

Machine learning techniques with ANOVA for the prediction of breast cancer

Bharti Thakur[1*], Nagesh Kumar[2] and Gaurav Gupta[1]
Shoolini University, Yogananda School of AI, Computing and Data Sciences, Solan, Himachal Pradesh, India[1]
Chitkara University, School of Engineering & Technology, Chitkara University, Himachal Pradesh, India[2]

Received: 05-August-2021; Revised: 15-February-2022; Accepted: 18-February-2022

Abstract
Breast cancer is one of the most common cancer among females. In this paper, machine learning techniques are applied to a molecular taxonomy of breast cancer international consortium (METABRIC) dataset to extract prime clinical attributes. Analysis of variance (ANOVA), is used for clinical feature selection. Five different machine learning algorithms are implemented, which are support vector machine (SVM), decision tree, random forest, AdaBoost and artificial neural network (ANN). Among all the machine learning classifiers, ANN gives the highest accuracy of 87.43%. This statistical technique is helpful for the detection of breast cancer, and it will increase the survival rate of females.

Keywords
Breast cancer, Genes, ANOVA, ANN, SVM, Machine learning, Healthcare.

1.Introduction
Breast cancer is the most common disease all around the globe [1]. Humanity grappled with recognizing and handling breast cancer in the last 3500 years [2]. Alongside the past civilizations and between the 18th and 19th centuries [2], it was acknowledged that breast cancer is difficult to control in the worst stage. Breast cancer continued to be a significant female health matter more than 3500 years ago [2]. In 2020 India's new breast cancer cases were 2261419, death number was 684996 as per Globocan report [3]. Advanced identification of breast cancer will enlarge the possibility of retrieval and longevity of life. Breast Cancer is a complex disease that is activated due to unchecked break-up of cells inside the terminal vessel and lobular of the breast [4]. Breast cancer mainly occurs due to mutation in genes [5]. Due to day-to-day habits of urban population, more breast cancer cases are found in cities than villages [6]. Genes are small pieces of deoxyribonucleic acid (DNA) that develop in chromosomes. DNA carries the specification for structuring proteins. DNA orders contain four nucleotides: adenosine, cytidine, guanosine, and thymine (A, C, G, and T). The precise order of nucleotides, when combined they form a DNA sequence [7].

*Author for correspondence

DNA conversion negatively influences fitness which is called mutations. These mutations are the reason for breast cancer in females [8]. Breast cancer gene 1 and breast cancer gene 2 (BRCA1 and BRCA2) are the genes that are highly responsible for breast cancer. Along with these genes, partner and localizer of BRCA2 (PALB2), phosphatase and tensin homolog (PTEN), tumour protein (TP53), ataxia telangiectasia mutated (ATM), Cadherin 1 (CDH1), Checkpoint kinase 2 (CHEK2) is responsible for breast cancer [9]. Machine learning is an approach of artificial intelligence, which gives the system capability to spontaneously absorb and refine from circumstance without being organized [10]. Machine learning follows two groups such as supervised and unsupervised. Supervised learning follows the procedure where whole data is labeled, and output are based on that labeled data. In unsupervised learning, the information is not classified or categorized. No plan of fixed result is examined in this type of learning [11]. In India, every 2 minutes one female is identified with breast cancer and every 9 minutes one death is reported [12]. Breast cancer in India is graded as the leading cancer among females, which is 25.8% per 10000 females, and the death rate is 12.7% per 10000 females [13]. Prior monitoring is essential to decrease breast cancer deaths. Breast cancer is the uttermost familiar cancer among women [14]. The identification strategy for breast cancer includes clinical examination, mammography, ultrasound, and biopsy. As per the American Cancer Society, the

following features increase the possibility of growing breast cancer [15].

1.1Common possibility characteristic

As the female ages, spreading breast cancer increases. Being female is the most outstanding significant possibility for growing breast cancer.

1.2Generative possibility characteristic

Having a family record of breast cancer increases the chances of breast cancer. Mutation record is another feature mutation in specific genes like BRCA1 and BRCA2.

1.3Body characteristic

Females who get pregnant later in their life and have no children come under the high-risk zone of breast cancer. Females with thick breast tissue have also an extreme possibility of breast cancer.

However, with the advancement of technology, chances of early detection of breast cancer have increased. With the introduction of machine learning and its various algorithms, breast cancer can be detected accurately too [16]. Machine learning is very helpful where personal skill is missing for illustration like exploring the mars, handling missing data, recognizing trends and samples. When mutation [17] occurs in these eight genes shown in *Table 1* with their location and exon count, breast cancer occurs in a female body. The detection is possible through the machine learning techniques [18].

Key contributions of this paper include the following points:

- The most significant clinical features are picked out with the help of analysis of variance (ANOVA) from the molecular taxonomy of breast cancer international consortium (METABRIC) dataset, which is highly responsible for the death of females because of breast cancer.
- Machine learning algorithms are implemented with cross-validation to check the model's accuracy.
- To determine which machine learning algorithm will give the highest accuracy on the METABRIC dataset.

A detailed related study was discussed and presented in section 2 in the form of a literature survey. After this, a detailed methodology is defined in section 3, which explains about dataset, training, and testing of models. The results obtained after implementation is described in section 4, and section 5 presents a discussion on the same. Finally, the paper concluded in section 6, which summarizes this paper's major points and future directions.

Table 1 Location of eight genes that are highly responsible for breast cancer with exon count

Genes	Location	Exon count
BReast CAncer gene1	17q21.31	21
BReast CAncer gene2	13q13.1	27
Tumor protein p53	17q13.1	12
Ataxia telangiectasia mutated	11q22.3	69
Cadherin-1	16q22.1	16
Checkpoint kinase 2	22q12.1	22
Partner and Localizer of BRCA2	16q12.2	14
Phosphatase and tensin homolog	10q23.31	10

2.Literature survey

The researchers implemented various machine learning algorithms for breast cancer to reveal important details. Different algorithms were implemented on the images and gene datasets to determine breast cancer. Kothari et al., in their research, implemented three machine learning algorithms, namely decision tree, random forest, and set covering machine, to recognize which gene is responsible for triple-negative breast cancer and non-triple negative breast cancer. The decision tree classifier got 0.522 accuracies while training on the METABRIC dataset [19].

Mirsadeghi et al. [20] in their work, studied 450 patients with metastatic breast cancer from the cBio cancer genomics portal. They applied four software tools for characteristic removal. Artificial neural network (ANN), random forest, On-linear support vector machine (SVM) were considered to estimate feasible genes for metastatic breast cancer.

Amrane et al. [21] in their research, used two separate classifiers. Naïve Bayes (NB) and K-nearest neighbour (KNN) on Wisconsin dataset for breast cancer, where KNN gives the highest accuracy for predicting breast cancer.

Wu and Hicks [22] in their research, applied various machine learning algorithms and indicated breast cancer. KNN, NB, SVM, and decision tree classifier were executed for classification.

Divyavani and Kalpana [23] estimated the effectiveness of ANN and SVM on the Wisconsin diagnostic dataset. An accuracy of 98% and 99% were obtained by ANN and SVM respectively. However, their research did not analyse other important machine learning algorithms.

Ak [24] applied the Wisconsin dataset with 32 clinical attributes. Logistic Regression, KNN, SVM, NB, decision tree, random forest was implemented to predict whether a tumor is benign or malignant. They found that 62.7% females suffered from benign tumour, and 37.3% sustained the malignant tumour.

In their research, Thottathyl et al. [25] implemented a K-means clustering algorithm on the Wisconsin dataset for early detection of breast cancer.

Ahmed et al. [26] in their study, implemented NB, SVM, Multilayer Perceptron, J48, and random forest on the Wisconsin dataset. They included parameters for classification were accuracy, recall, precision, and receiver operating characteristic (ROC) area.

Teixeira et al. [27] in their research, worked on the Wisconsin dataset. They used deep neural network, and 10 attributes were considered, which displayed accuracy of 92%.

Magboo and Magboo [28] compared four machine learning models, comprising Logistic Regression, NB-KNN, and SVM. In their research, they applied Wisconsin prognostic breast cancer dataset. Logistic Regression gave the best output, which includes (precision, recall, accuracy, F1 score, Area under the receiver operating characteristics (AUROC), and Cohen Kappa statistics.

Naji et al. [29] implemented SVM, random forest, Logistic Regression, decision tree, and KNN on the Wisconsin dataset. SVM gives the highest accuracy of 97.2%.

Lahoura et al. [30] in their research, implemented an extreme learning machine (ELM) on the Wisconsin dataset for breast cancer detection. ELM with 100 hidden nodes was implemented to detect breast cancer.

From the literature survey, shown in *Table 2*, it can be depicted that the highest accuracy of SVM is 98%, a decision tree is 95.61%, and random forest is 97.2% on the Wisconsin dataset. SVM, decision tree, and arbitrary forest accuracy are defined, but past researchers assumed no specific criteria of features selection. Accuracy is not represented in the METABRIC dataset. Cross-validation is also not implemented on these machine learning algorithms as cross-validation is essential to check the usefulness of the model on unseen data, which researchers in their previous work do not implement.

Table 2 Dataset used by various researchers

Ref.	Dataset	Machine learning classifier/technique	Accuracy
[19]	Gene Dataset	Decision Tree	0.522
		Random Forest	0.754
[20]	Gene Dataset	ANN	____
		Random Forest	____
		On-linear SVM	____
[21]	Wisconsin Dataset	Naïve Bayes	96.19
		KNN	97.51
[22]	Wisconsin Dataset	KNN	87%
		Naïve Bayes	85%
		SVM	90%
		Decision Tree	87%
[23]	Wisconsin Dataset	ANN	99%
		SVM	98%
[24]	Wisconsin Dataset	Logistic regression	98.06%

Ref.	Dataset	Machine learning classifier/technique	Accuracy
[25]	Wisconsin Dataset	KNN	96.49
		SVM	96.49
		Naïve Bayes	94.73%
		Random Forest	95.61%
		Decision Tree	95.61%
		K means clustering	_____
[26]	Wisconsin Dataset	Naïve Bayes	97.2%
		Support Vector Machine	96.13%
		Multilayer Perceptron	96.13%
		J48	94.26%
		Random Forest	95.5%
[27]	Wisconsin Dataset	Deep Neural Network	92%
[28]	Wisconsin Dataset	Logistic Regression	0.80%
[29]	Wisconsin Dataset	Naïve Bayes	0.60%
		KNN	0.60%
		SVM	0.75%
		Random Forest	____
		SVM	97.2%
		Logistic Regression	____
		Decision Tree	_____
[30]	Gene Dataset	ELM Technique	0.98%

Most of the work is performed on the Wisconsin dataset by various researchers mentioned in the literature survey (*Table 2*). Very little research is performed on the gene dataset. We learned that literature, observation performed no feature scaling on the gene dataset earlier. So, there is still a need for future work on gene datasets. Feature scaling must be done to get the best results for the survival of females from breast cancer. As statistical methods are essential for finding out the best features to predict the overall death of patients in the METABRIC dataset. For the Wisconsin dataset, researchers have taken into consideration the literature survey [21−29] in *Table 2*. For gene dataset graph, researcher have considered the references [19, 20, 30].

3.Methods

Our research used an openly accessible METABRIC-ribonucleic acid (RNA) mutation dataset from Kaggle [31]. The dataset contains 1904 patients with 31 clinical attributes wherein 331 genes and protein have been considered in the dataset. These genes and proteins correlate with death from breast cancer. In the mentioned dataset, mutation in 175 genes has been analysed on the basis of 1904 breast cancer patients, out of which, females from the age group of 29 to 87 are diagnosed with breast cancer.

3.1Data pre-processing

In data pre-processing, data are normalized [32]. Data pre-processing is performed to identify missing values and remove the missing values from the dataset. Missing categorical and numerical values are removed from the dataset. Scaling all the attributes is performed to get the best result and to identify the overall death of patients with breast cancer. Standard scalar is used for attribute scaling. In traditional scalar, values are fixed in the range of 0 and 1.

3.2Feature selection and methodology

Feature selection decreases input variables and increases the model's performance. ANOVA is also known as the analysis of variance. It is a statistical method used for feature reduction based on a correlation between features and labels. With the help of ANOVA, six best clinical features are selected out of 31 clinical characteristics to predict the overall death from breast cancer [33]. The columns determined by ANOVA shown in *Table 3* are age at diagnosis, inferred menopausal state, lymph nodes examined-positive, Nottingham prognostic index, overall survival months, and overall survival to predict death from breast cancer. Clinical data attributes are defined in *Table 3*.

Table 3 Clinical data description

S. No.	Clinical features
1	Age at diagnosis
2	Inferred menopausal state
3	Lymph nodes examined positive
4	Nottingham prognostic index
5	Overall survival months
6	Overall survival

3.3Machine learning algorithms

We have used five machine learning algorithms,
(1) SVM (2) decision tree (3) random forest (4) Ada Boost (5) ANN.
These algorithms are selected because of the following reasons.

1. SVM algorithm is implemented to create the best line or a decision boundary that can segregate n-dimensional space into classes to quickly put the new data point in the correct category in the future. This best decision boundary is called a hyperplane [34].
2. In the decision tree, the data are continuously divided according to a specific parameter defined in *Table 4*. The leaves are the decisions or the outcomes [35, 36].
3. In random forest, the "forest" it builds is an ensemble of decision trees, trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result [37, 38].
4. The AdaBoost algorithm is a boosting technique used as an ensemble method in machine learning. It is called AdaBoost as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified model the biological analogy. ANN consists of nodes representing neurons connected by arcs. It corresponds to dendrites and synapses. Each turn is associated with a weight while at each node. Values are applied as received input by the node, and the activation function is defined along the incoming arcs, adjusted by the weights of the hooks.
5. These algorithms are implemented with METABRIC dataset and ANOVA statistical function (*Figure 1*).

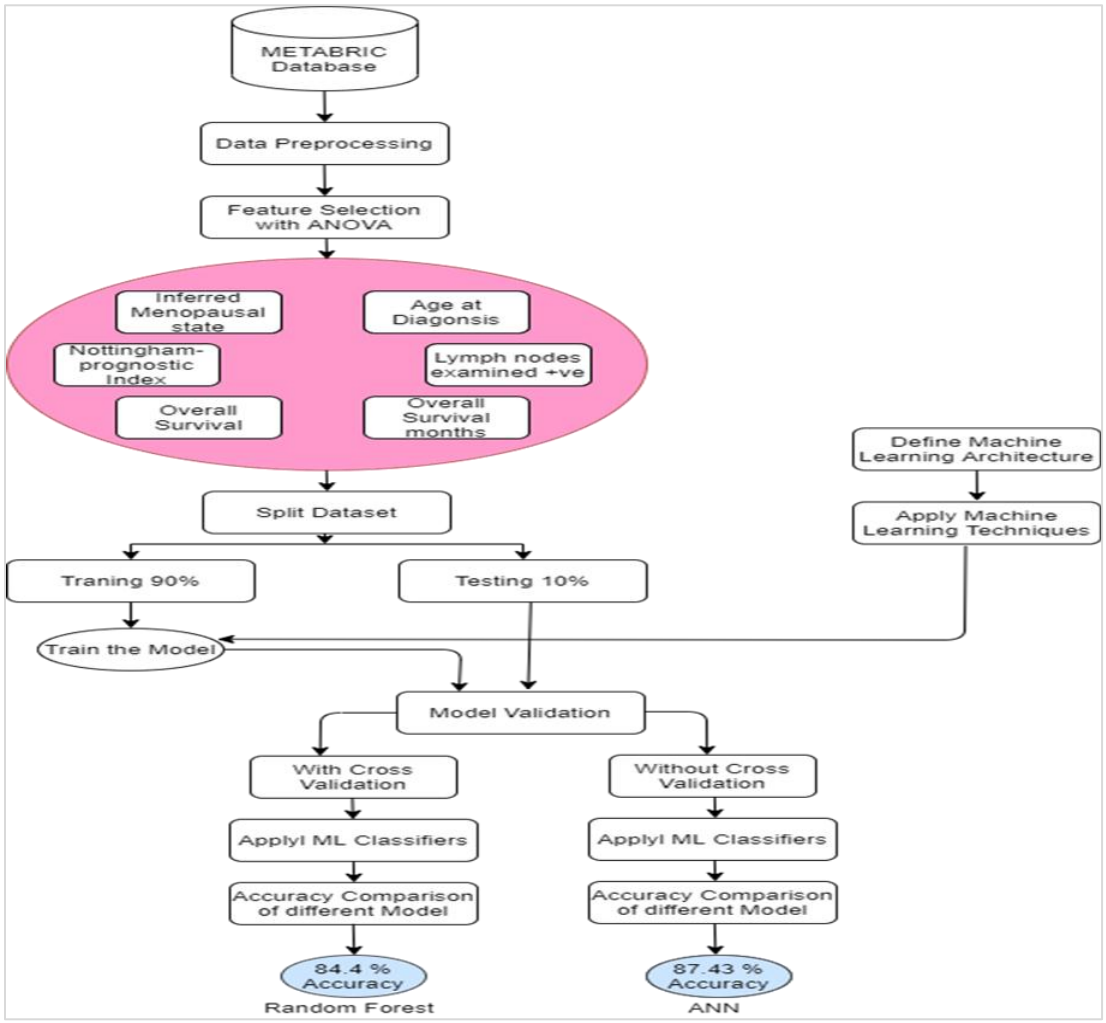

Figure 1 Machine learning techniques on METABRIC dataset

3.3.1SVM

In SVM, we aim to get the authentic hyperplane that amplifies the boundaries. We have used C as a hyperparameter set at the beginning of training, the model mentioned in *Table 5*. Hyperplane equation diving the points in SVM [39].

$$\vec{\omega} \cdot \vec{x} - b = 0 \quad (1)$$

Where in Equation 1

$\vec{\omega}$ is the normal direction of a plane

b is a form of entrance

If $\vec{\omega}.\vec{x}$ is calculated to be large, then it is owned by a class. If it is less than b, it is owned by a second class.

3.3.2 Decision tree

In the decision tree, we have calculated the entropy and information gain mentioned in *Table 4*, entropy manages how the decision tree chooses the data.

$$Entropy = -E = -\sum_{i=1}^{c} -\text{Pi}.\log 2\ \text{Pi} \quad (2)$$

Where in Equation 2

Pi= probability of an element

After that, information gain is calculated. This technique analyses the best split in a decision tree [40].

$$IG = (T, A) = Entropy\ (T) - \sum_{V \in A} \frac{|Tv|}{T}.\text{Entropy(Tv)} \quad (3)$$

Where in Equation 3

T = Target column

A = The variable (column) we are testing.

v = all value in A

3.3.2Random forest

This algorithm, can be implemented on regression and classification problems. In our research, we use the Gini index on classification data [41].

$$Gini = 1 - \sum_{i=1}^{c} (\text{Pi})^2 \quad (4)$$

Where in Equation 4

c = classes in the target variable

Pi = ratio of class

This formula decides in our random forest implementation which branches will appear the most in random forest.

3.3.3Ada boost

This algorithm is an adaptive boost classifier and a boosting algorithm. The main advantage of the Ada Boost classifier is a weak beginners change towards the powerful learner [42].

$$H(x) = sign(\sum_{i=1}^{C} \alpha_t h_t(x)) \quad (5)$$

Where in Equation 5

$H(x)$ denotes the weightage of input training data

$h_t(x)$ relates to the weak classifier t output where x is input

α_t represents weight shared with the classifier.

3.3.4ANN

ANN comprises of input, hidden, and output layer. In ANN we follow the formula mentioned below [43].

$$F(b + \sum_{i=1}^{n} X_i W_i) \quad (6)$$

Where in Equation 6

b = bias

X = input to neuron

W =weights

^{n} = number of inputs from incoming layer

$_{i}$ = a counter from 1 to n

In our work, ANN gives the highest accuracy among all the machine learning classifiers

3.4Training and testing

Training and testing are performed to calculate the accuracy of served on the METABRIC dataset. The dataset is divided into two parts, one portion is for training activity and another piece is for testing:

- In first split, 70% of data (1334 patients) for training and 30% (571 patients) for testing.
- At the second split, 80% data (1524 patients) for training and 20% (361 patients) for testing.
- In third split, 90% data (1714 patients) for training and 10% (191 patients) for testing.

To check the accuracy level, which is mentioned in *Table 5*. The 10-fold cross-validation is also performed on the same dataset to check whether splitting the dataset into equal numbers will affect the accuracy level or not [44]. Cross-validation is implemented in machine learning to evaluate the expertise of the machine learning model on hidden data.

Table 4 Parameters used in all the machine learning classifiers

Machine learning classifier	Parameters	Parameters used
SVM	Optimizer	C
	Kernel	RBF
Decision Tree	No of nodes	7
	Entropy	7.55
	Information Gain	7.43

Machine learning classifier	Parameters	Parameters used
Random Forest	No. of estimators	100
	Criteria followed	GINI
	Min samples split	2
AdaBoost Classifier	Learning Rate	0.01
	No of estimators	50
	Base estimators	None
ANN	Number of Neurons	200
	Activation Function	Rectified linear unit (ReLu)
	Multiclass Classification	Sigmoid
	Optimizer	Adam

Table 5 Train/test split

No. of the patients for training the data	No. of patients for testing the data
1334	571
1524	361
1714	191

3.5Confusion matrix

Each column of the confusion matrix denotes the copy of the estimated class. A confusion matrix can be implemented on binary and multiclass classification problems [45]. In this dataset, we have used multi types, namely died of disease, died of other causes, and living. To get the accuracy in the confusion matrix formula is [46].

$$Accuracy = \frac{TN+TP}{TN+FP+FN+TP} \quad (7)$$

where in Equation 7
(*TN*= True Negative, *TP* = True Positive, *FP* = False Positive, *FN* = False Negative)

As the confusion matrix is a method, to sum up, the execution of classification algorithms.

4.Results

With the implementation of ANOVA on the METABRIC dataset, we choose the six clinical features out of 31 parts. Selected attributes are shown in *Table 3*. Machine learning algorithms are implemented on the METABRIC dataset providing the highest accuracy. We get the following results on these algorithms based on 90% training and 10% testing data.

4.1Support vector machine

C is a hypermeter in SVM to control error mentioned on the X-axis in *Figure 2*. SVM gives the accuracy of 0.8691% on the tested dataset, and a 10-fold cross-validation method is also implemented, subdividing actual samples into 10 equal-sized subsamples. Each subsample is considered validation data for testing the model and repeating the process 10 times.

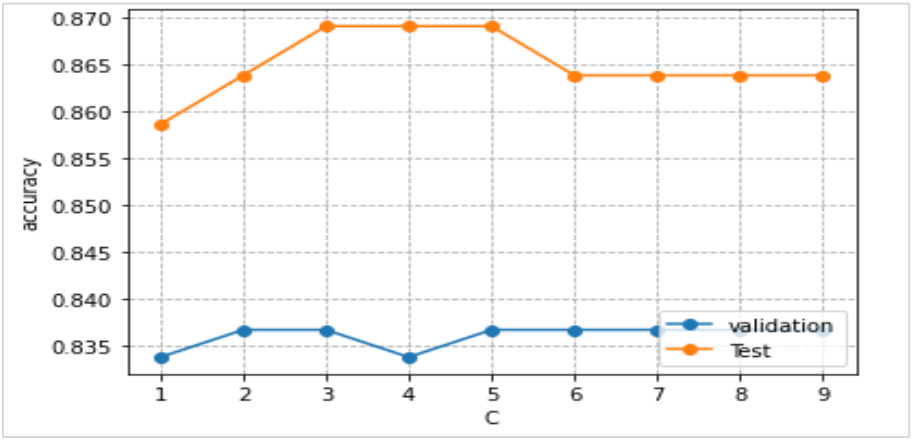

Figure 2 Accuracy of support vector machine

4.2Decision tree

It has a tree-like structure used for classification and prediction. The data are divided into defined parameters [47]. Six parameters are given, which are highly correlated with death from cancer.

Format of features in the decision tree is mentioned below

(x, y) = (x1, x2, x3 ___ x6, Y)
Y is the target variable dependent on X features that contain (x1-----x6). We have considered the max-depth, the length of the longest path from the root to leaf, and twelve iterations are performed as it affects the accuracy rate which is mentioned in *Figure 3*. The decision tree gives an accuracy of 0.853%.

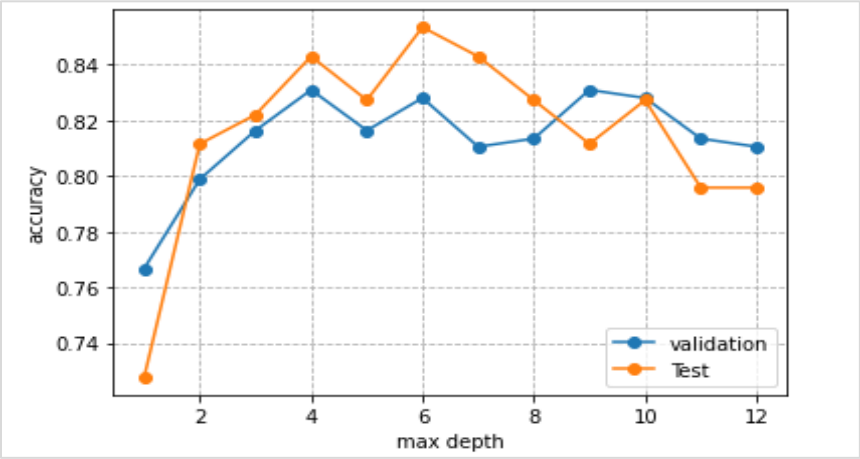

Figure 3 Accuracy of decision tree

4.3Random forest

Random forest consists of a single decision tree, which works as an ensemble [48]. This algorithm is found to be accurate than other classifiers and works expertly on massive datasets. It can determine which variable is primary in classification and gives the result accordingly. To get the highest accuracy, the max depth of random forest is considered, which describes each tree's deep in a forest. The highest accuracy of random forest is 0.863% mentioned in *Figure 4.*

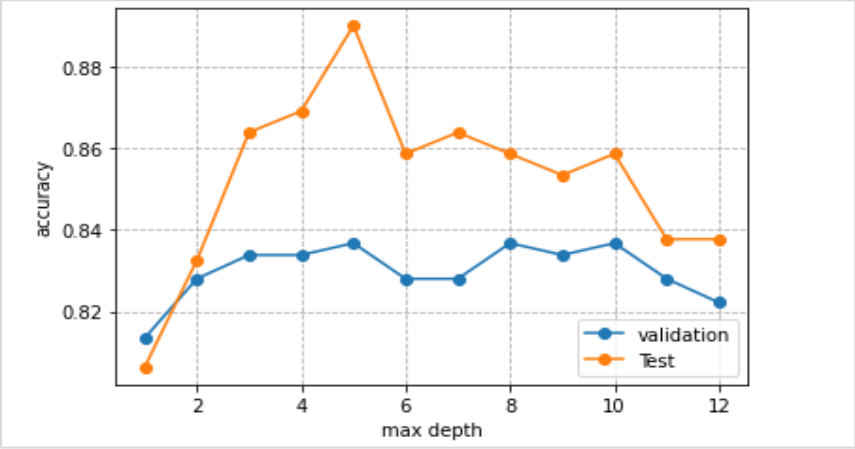

Figure 4 Accuracy of random forest

4.4Ada boost classifier

It is a boosting technique. Weights are reallocated to every occurrence and excessive weights to wrong classified events [49]. In our experimental result, the Ada Boost gives an accuracy of 0.685 with a learning rate of 0.01. The learning rate is mentioned in *Figure 5.*

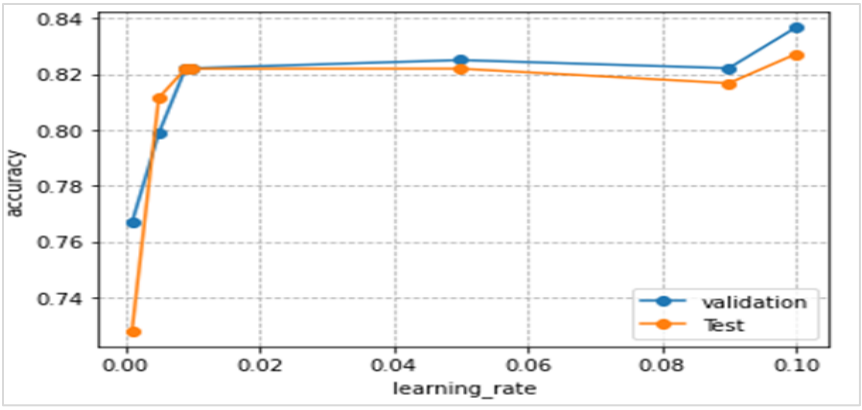

Figure 5 Accuracy of AdaBoost classifier

4.5ANN

The input layer contains components of the dataset; the output layer includes a single node. Weights joining the layers are rearranged by utilizing training data of METABRIC database. It further applies feedforward ANN with Adam optimizer. This classifier gives an accuracy of 87.43% with 200 neurons and ReLu activation function *Table 6.*

Table 6 ANN parameters

No of neurons	Activation function	Optimizer
200	ReLu	Adam

4.6Confusion matrix

Confusion Matrix is created of all the classifiers as confusion matrix is used to assess the execution of the classification model. The confusion matrix talks about the errors built by the classifier. In our research, we have multi classes in the confusion matrix first class died of disease, the second died of another cause, and the third was living. Confusion matrix gives knowledge regarding misconception made by the classifier and variety of errors that are being made. It indicates how a classification model is unorganized, while building prognosis. Confusion matrix results are shown below in *Table 7.*

Table 7 Performance of classification model with the help of confusion matrix

Machine learning classifier	Class	Died of disease	Died of other cause	living
Decision Tree	Died of disease	50	10	0
	Died of other cause	18	34	0
	Living	0	0	79
Ada Boost Classifier	Died of disease	0	60	0
	Died of other cause	0	52	0
	Living	0	0	79
Random Forest	Died of disease	51	9	0
	Died of other cause	17	35	0
	Living	0	0	79
SVM	Died of disease	51	9	0
	Died of other cause	16	36	0
	Living	0	0	79
ANN	Died of disease	54	6	0
	Died of other cause	38	14	0
	Living	33	20	26

4.7Accuracy results

In *Figure 6*, we have mentioned the accuracy level of (70%/30%), (80%/20%), and (90%/10%) as the train and test split. The accuracy of (90%/10%) train and test split has been taken from the cross-validation accuracy that is elaborated in *Table 8*. Further, the research depicts that this split gives the highest accuracy. From *Table 8*, we can say that ANN gives the highest accuracy among all the classifiers. When cross-validation is applied on all the classifiers, then

among all, the random forest gives the highest accuracy [50]. 80-20 and 70-30 split accuracy results are shown in *Table 8*.

We combined the different measurement aspects to check the effectiveness of all machine learning algorithms [51]. F1 scores are implemented to evaluate the standard of multiple classes. Hamming Loss is the portion of the labels that are wrongly anticipated. It is used to check the model performance. The standard error is a statistical word that calculates the accuracy. Kappa value is used to compute the inter-rater authenticity for the qualitative module. The score of all the classifiers is mentioned in *Table 9* below.

By implementing all the above algorithms, we got the following results.

- Six best clinical attributes are selected which are positively correlated with one another and highly responsible for the death of a female, with the help of ANOVA statistical function.
- The decision tree algorithm is implemented on the METABRIC dataset. The number of iterations performed on the decision tree is 12, which increased accuracy to 85%.
- In random forest validation score is 0.776, and the test score is 0.770 when the first iteration is performed. At the 12th iteration, the validation score increased to 0.831, the test score increased to 0.827, and accuracy reached 86%. We considered the best score [52].
- Tunning of the SVM algorithm on the METABRIC dataset with nine iterations gives the validation score of 0.837, test score of 0.86, and accuracy of 86.9%.
- By implementing the ANN at epoch 100, the loss (prediction error) is 0.2507 with 80 per accuracy, and at epoch 200, the loss is 0.1722 with 87.4%. So, the minimum failure is considered with maximum precision.

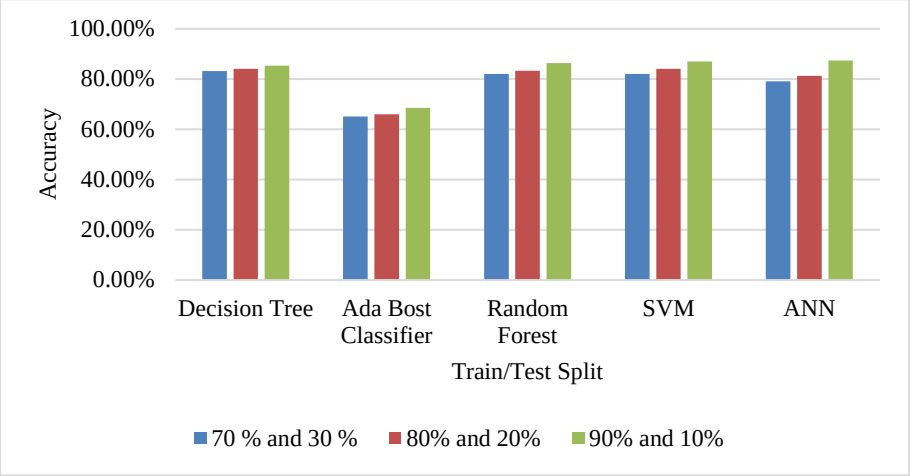

Figure 6 Comparison based on a train-test split of the dataset

Table 8 Highest accuracy with cross validation in different split ration 90-10, 80-20 and 70-20

ML classifier	90-10 split	80-20 split	70-20 split	Cross validation
SVM	86.9	85.2	84.1	84.1
Decision Tree	85.3	84.0	82.2	82.4
Random Forest	86.3	85.1	83.0	84.4
Ada Boost	68.5	63.1	63.0	67.9
ANN	87.4	85.1	84.0	83.3

Table 9 Different measurement aspects of machine learning classifiers

Classifier	F1Macro	Hamming loss	Standard error	Kappa value
Ada Boost	0.544	0.314	0.033	0.53
Decision Tree	0.829	0.146	0.025	0.77
Random Forest	0.854	0.125	0.023	0.80
SVM	0.848	0.130	0.024	0.80
ANN	0.536	0.465	0.313	0.77

5.Discussion

Our first goal of the present study is to examine the clinical features that are correlated with death and breast cancer. Our findings propose that with the help of the ANOVA statistical function, we can select the best clinical features when the number of elements is enormous. ANOVA is used to predict the factors that are responsible for the detection of breast cancer in women.

The second aim is to implement the machine learning algorithms on the METABRIC dataset and check the various algorithms' accuracy. In SVM, to check the accuracy of the SVM classifier, nine iterations are performed to avoid misclassification in all training samples. In the decision tree, seven nodes have been considered and further 12 iterations have been performed on the decision tree classifier algorithm. Random forest is implemented to increase the anticipated accuracy. We have taken 100 no of estimators, and the number of iterations performed for tunning the random forest classifier is 12. The criteria followed in the random forest is Gini.

In the Ada Boost classifier, we have taken 50 estimators wherein the learning rate is 0.01. This classifier aims to place the mass of classifiers and train the facts specimen in all repetition. The final classifier we have used for prediction is ANN; this classifier works the same way persona intelligence examines procedure details. In ANN, we have taken 200 neurons. As in our research, we have used a standard scalar for scaling the attributes where values are fixed in 0 and 1. ANOVA is used for making the best conclusion of features. Our data lies between 0 to 1, therefore we have used the sigmoid function for multiclass classification and the ReLu activation function in ANN. SVM gives the accuracy of 86.9%; decision tree 85.3%, random forest 86.3%, Ada Boost 68.5%, and ANN gives 87.43% accuracy. The third aim is to examine the model's validity on obscured data. We implemented 10-fold cross-validation on all machine learning classifiers and checked for accuracy.

As per the research, it has been found that the random forest gives the minimum hamming loss and standard error by performing comparative analysis on various classifiers. Research says that Kappa value of random forest and SVM is the same. Kappa value is used to measure the inter-rater dependability for a definite unit. A confusion matrix shows the accuracy and defects of each class in the concerned model. As mentioned in *Table 7*, three classes are: died of disease, other causes, and living. For the living class, all the classifiers give the same output as 1.0, but for the ANN, the value is 0.727, shown in *Table 10*. ANN cross-validation overall output is less accurate in the confusion matrix. But as per *Table 8*, it can be justified that the ANN accuracy of the METABRIC dataset is above other classifiers. The random forest consists of countless decision trees that work as an ensemble and large datasets. In random forest, we have tried to increase the numbers of estimators in every node break to get a more accurate result. For measuring the impurity, researchers have followed the Gini criteria to deal with the classification problem.

A complete list of abbreviations is shown in *Appendix I.*

Table 10 Accuracy of classification model by confusion matrix

ML Classifier	Died of disease	Died of other cause	Living
SVM	0.869	0.869	1.0
Decision Tree	0.853	0.853	1.0
Random Forest	0.874	0.874	1.0
Ada Boost	0.685	0.681	1.0
ANN	0.534	0.806	0.727

Our study has some limitations.

- After performing the cross-validation on the same data with identical algorithms, the level of accuracy falls apart.
- Eight genes most responsible for breast cancer can predict death from breast cancer.

Despite these restrictions, our dataset contains 6 best correlated clinical features with 331 genes /proteins and predicts the death from breast cancer. Nevertheless, our study calls attention to the use of machine learning to predict breast cancer.

6.Conclusion and future work

Machine learning classifier SVM, decision tree, Ada Boost, random forest, and ANN were used in this paper for detecting breast cancer. We have used the dataset of 1904 patients with 6 clinical attributes and

331 genes and proteins. When the dataset is extensive, machine learning uses statistical functions for feature selection. With the help of statistical procedures and machine learning, breast cancer can be effectively diagnosed. Various researchers related to the research area have not opted for standard scalar and ANOVA functions in traditional machine learning methods, but in our research, we have scaled the dataset from 0 to 1 range with the help of a standard scalar. An ANOVA statistical function is used for feature scaling. Cross-validation is also implemented on all machine learning classifiers to check the model's usefulness on invisible data. A 10-fold cross-validation technique is implemented to train and test the classifiers.

In our approach, the author(s) has combined statistical methods with machine learning to get better accuracy for the detection of breast cancer. This machine learning model can assist doctors and researchers in assembling better techniques for breast cancer identification in females. As we see the improvement in technology every day, new algorithms are being implemented to solve the problem of detection of breast cancer. So, researchers can implement deep learning by taking care of all these things. This dataset also consists of all the genes responsible for breast cancer. In the future, genetic algorithms and deep learning can also be implemented to predict breast cancer with more accuracy.

Acknowledgment
None.

Conflicts of interest
The authors have no conflicts of interest to declare.

Authors contribution statement
Bharti Thakur:Conceptualization, investigation on challenges, data analysis, design, writing-original draft. **Nagesh Kumar**: Analysis, interpretation of results, review and supervision. **Gaurav Gupta**: Study conception, review and supervision.

References
[1] Priyanka KS. A review paper on breast cancer detection using deep learning. In conference series: materials science and engineering 2021 (pp. 1-9). IOP Publishing.
[2] Lukong KE. Understanding breast cancer–the long and winding road. BBA Clinical. 2017; 7:64-77.
[3] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians. 2021; 71(3):209-49.
[4] Mahdi KM, Nassiri MR, Nasiri K. Hereditary genes and SNPs associated with breast cancer. Asian Pacific Journal of Cancer Prevention. 2013; 14(6):3403-9.
[5] Shiovitz S, Korde LA. Genetics of breast cancer: a topic in evolution. Annals of Oncology. 2015; 26(7):1291-9.
[6] Gupta A, Shridhar K, Dhillon PK. A review of breast cancer awareness among women in India: cancer literate or awareness deficit?. European Journal of Cancer. 2015; 51(14):2058-66.
[7] Pyingkodi M, Thangarajan R. Informative gene selection for cancer classification with microarray data using a metaheuristic framework. Asian Pacific Journal of Cancer Prevention: Asian Pacific Journal of Cancer Prevention. 2018; 19(2):561-4.
[8] Sun Y, Zhu S, Ma K, Liu W, Yue Y, Hu G, et al. Identification of 12 cancer types through genome deep learning. Scientific Reports. 2019; 9(1):1-9.
[9] El RSA, Al-montasheri A, Al-hazmi B, Al-dkaan H, Al-shehri M. Machine learning model for breast cancer prediction. In international conference on fourth industrial revolution 2019 (pp. 1-8). IEEE.
[10] Le NQ, Yapp EK, Nagasundaram N, Yeh HY. Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous FastText N-grams. Frontiers in Bioengineering and Biotechnology. 2019:1-9.
[11] Carbonell JG, Michalski RS, Mitchell TM. An overview of machine learning. Machine Learning. 1983:3-23.
[12] Vaka AR, Soni B, Reddy S. Breast cancer detection by leveraging machine learning. ICT Express. 2020; 6(4):320-4.
[13] Malvia S, Bagadi SA, Dubey US, Saxena S. Epidemiology of breast cancer in Indian women. Asia-Pacific Journal of Clinical Oncology. 2017; 13(4):289-95.
[14] Momenimovahed Z, Salehiniya H. Epidemiological characteristics of and risk factors for breast cancer in the world. Breast Cancer: Targets and Therapy. 2019:151-64.
[15] Oeffinger KC, Fontham ET, Etzioni R, Herzig A, Michaelson JS, Shih YC, et al. Breast cancer screening for women at average risk: 2015 guideline update from the American cancer society. JAMA. 2015; 314(15):1599-614.
[16] Gupta P, Garg S. Breast cancer prediction using varying parameters of machine learning models. Procedia Computer Science. 2020; 171:593-601.
[17] Feng Y, Spezia M, Huang S, Yuan C, Zeng Z, Zhang L, et al. Breast cancer development and progression: risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. Genes & Diseases. 2018; 5(2):77-106.
[18] Musumeci F, Rottondi C, Nag A, Macaluso I, Zibar D, Ruffini M, et al. An overview on application of machine learning techniques in optical networks.

IEEE Communications Surveys & Tutorials. 2018; 21(2):1383-408.
[19] Kothari C, Osseni MA, Agbo L, Ouellette G, Déraspe M, Laviolette F, et al. Machine learning analysis identifies genes differentiating triple negative breast cancers. Scientific Reports. 2020; 10(1):1-5.
[20] Mirsadeghi L, Haji HR, Banaei-moghaddam AM, Kavousi K. EARN: an ensemble machine learning algorithm to predict driver genes in metastatic breast cancer. BMC Medical Genomics. 2021; 14(1):1-19.
[21] Amrane M, Oukid S, Gagaoua I, Ensari T. Breast cancer classification using machine learning. In electric electronics, computer science, biomedical engineerings' meeting 2018 (pp. 1-4). IEEE.
[22] Wu J, Hicks C. Breast cancer type classification using machine learning. Journal of Personalized Medicine. 2021; 11(2):1-12.
[23] Divyavani M, Kalpana G. An analysis on SVM & ANN using breast cancer dataset. Aegaeum J. 2021; 8:369-79.
[24] Ak MF. A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. Healthcare 2020; 8(2):1-23. Multidisciplinary Digital Publishing Institute.
[25] Thottathyl H, Kanadam KP, Panchadula RP. Microarray breast cancer data clustering using map reduce based K-means algorithm. Revue d'Intelligence Artificielle. 2020; 34(6):763-9.
[26] Ahmed MT, Imtiaz MN, Karmakar A. Analysis of wisconsin breast cancer original dataset using data mining and machine learning algorithms for breast cancer prediction. Journal of Science Technology and Environment Informatics. 2020; 9(2):665-72.
[27] Teixeira F, Montenegro JL, Da CCA, Da RRR. An analysis of machine learning classifiers in breast cancer diagnosis. In XLV Latin American computing conference 2019 (pp. 1-10). IEEE.
[28] Magboo VP, Magboo MS. Machine learning classifiers on breast cancer recurrences. Procedia Computer Science. 2021; 192:2742-52.
[29] Naji MA, El FS, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science. 2021; 191:487-92.
[30] Lahoura V, Singh H, Aggarwal A, Sharma B, Mohammed MA, Damaševičius R, et al. Cloud computing-based framework for breast cancer diagnosis using extreme learning machine. Diagnostics. 2021; 11(2):1-19.
[31] Ali HR, Rueda OM, Chin SF, Curtis C, Dunning MJ, Aparicio SA, et al. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. Genome Biology. 2014; 15(8):1-14.
[32] Saoud H, Ghadi A, Ghailani M, Abdelhakim BA. Using feature selection techniques to improve the accuracy of breast cancer classification. In the proceedings of the third international conference on smart city applications 2018 (pp. 307-15). Springer, Cham.
[33] Vrigazova BP. Detection of malignant and benign breast cancer using the Anova-Bootstrap-SVM. Journal of Data and Information Science. 2020; 5(2):62-75.
[34] Abdullah DM, Abdulazeez AM. Machine learning applications based on SVM classification a review. Qubahan Academic Journal. 2021; 1(2):81-90.
[35] Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends. 2021; 2(1):20-8.
[36] Chang CC, Yeh JH, Chiu HC, Chen YM, Jhou MJ, Liu TC, et al. Utilization of decision tree algorithms for supporting the prediction of intensive care unit admission of myasthenia gravis: a machine learning-based approach. Journal of Personalized Medicine. 2022; 12(1):1-16.
[37] Disha RA, Waheed S. Performance analysis of machine learning models for intrusion detection system using Gini impurity-based weighted random forest (GIWRF) feature selection technique. Cybersecurity. 2022; 5(1):1-22.
[38] Schonlau M, Zou RY. The random forest algorithm for statistical learning. The Stata Journal. 2020; 20(1):3-29.
[39] Gaye B, Zhang D, Wulamu A. Improvement of support vector machine algorithm in big data background. Mathematical Problems in Engineering. 2021.
[40] Gulati P, Sharma A, Gupta M. Theoretical study of decision tree algorithms to identify pivotal factors for performance improvement: a review. International Journal of Computer Applications. 2016; 141(14):19-25.
[41] Sarica A, Cerasa A, Quattrone A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Frontiers in Aging Neuroscience. 2017; 9:1-12.
[42] Zhang Y, Ni M, Zhang C, Liang S, Fang S, Li R, et al. Research and application of AdaBoost algorithm based on SVM. In 8th joint international information technology and artificial intelligence conference 2019 (pp. 662-6). IEEE.
[43] Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. Digital Signal Processing. 2018; 73:1-15.
[44] Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Computer Science. 2021; 2(3):1-21.
[45] Battula K. Research of machine learning algorithms using K-fold cross validation. International Journal of Engineering and Advanced Technology. 2021; 8(6S):215-8.
[46] Kumar A, Sushil R, Tiwari AK. Significance of accuracy levels in cancer prediction using machine learning techniques. Technical Communication. 2019; 12(3): 741-7.
[47] Patel HH, Prajapati P. Study and analysis of decision tree based classification algorithms. International

Journal of Computer Sciences and Engineering. 2018; 6(10):74-8.
[48] Octaviani TL, Rustam DZ. Random forest for breast cancer prediction. In conference proceedings 2019 (pp. 1-6). AIP Publishing LLC.
[49] Zheng J, Lin D, Gao Z, Wang S, He M, Fan J. Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis. IEEE Access. 2020; 8:96946-54.
[50] Mohammed SA, Darrab S, Noaman SA, Saake G. Analysis of breast cancer detection using different machine learning techniques. In international conference on data mining and big data 2020 (pp. 108-17). Springer, Singapore.
[51] Easttom C, Thapa S, Lawson J. A comparative study of machine learning algorithms for use in breast cancer studies. In 10th annual computing and communication workshop and conference 2020 (pp. 412-6). IEEE.
[52] Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology. 2018; 12(2):119-26.

Bharti Thakur received her M.Tech degree in Computer Science from Himachal Pradesh Technical University. Currently, She is pursuing PhD from ShooliniUniversity, Solan, H.P. She has attended number of conferences and presented papers in the field of Machine Learning and Artificial Intelligence.
Email: bhartithakur.thakur@gmail.com

Dr. Nagesh Kumar is working as an Assistant Professor at Chitkara University, Himachal Pradesh. He is having more than 6 years of teaching experience. He received his PhD (CSE) from Jaypee University of Information Technology, Waknaghat, Solan (HP) India. He has more than 25 publications in reputed journals and conferences. He remained as an organizing committee member for more than 5 conferences. He is also serving in many capacities like reviewers in several scientific journals in his field and advisor in software companies. His research interests are Wireless Sensor Networks, Network Security, Cyber-physical-human systems, IoT for industries and Smart Agriculture.
Email: nagesh.kumar@chitkara.edu.in

Dr. Gaurav Gupta is a senior IEEE member and received his BE degree (2006) from Dr. B. R. Ambedkar University, Agra, Uttar Pradesh. He received PhD (2020) degree in the stream of Computer Science Engineering from Shoolini University. He attended a number of conferences and presented papers in the field of Machine Learning and Artificial Intelligence.
Email: solan.gaurav@gmail.com

Appendix I

S. No.	Abbreviation	Description
1	A, C, G, T	Adenosine, Cytidine, Guanosine, Thymine
2	ANOVA	Analysis of Variance
3	ANN	Artificial Neural Network
	ATM	Ataxia Telangiectasia Mutated
4	AUROC	Area Under the Receiver Operating Characteristics
5	BRCA1	BReast CAncer gene1
6	BRCA2	BReast CAncer gene 2
7	CDH1	Cadherin
8	CHEK2	Checkpoint Kinase
9	DNA	Deoxyribonucleic Acid
10	ELM	Extreme Learning Machine
11	KNN	K-Nearest Neighbour
12	NB	Naïve Bayes
13	PALB2	Partner and Localizer of BRCA2
14	PTEN	Phosphatase and Tensin Homolog
15	ReLu	Rectified Linear Unit
16	ROC	Receiver Operating Characteristic
17	RNA	Ribonucleic Acid
18	SVM	Support Vector Machine
19	TP53	Tumour Protein