RESEARCH ARTICLE

Editorial Process: Submission:11/23/2021 Acceptance:12/25/2022

Comparison of Classification Success Rates of Different Machine Learning Algorithms in the Diagnosis of Breast Cancer

Irem Ozcan[1], Hakan Aydin[2], Ali Cetinkaya[3]*

Abstract

Objective: To identify which Machine Learning (ML) algorithms are the most successful in predicting and diagnosing breast cancer according to accuracy rates. **Methods:** The "College of Wisconsin Breast Cancer Dataset", which consists of 569 data and 30 features, was classified using Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbor (KNN), Logistic Regression (LR), Multilayer Perceptron (MLP), Linear Discriminant Analysis (LDA), XgBoost (XGB), Ada-Boost (ABC) and Gradient Boosting (GBC) ML algorithms. Before the classification process, the dataset was preprocessed. Sensitivity, accuracy, and definiteness metrics were used to measure the success of the methods. **Result:** Compared to other ML algorithms used in the study, the GBC ML algorithm was found to be the most successful method in the classification of tumors with an accuracy of 99.12%. The XGB ML algorithm was found to be the lowest method with an accuracy rate of 88.10%. In addition, it was determined that the general accuracy rates of the 11 ML algorithms used in the study varied between 88-95%.**Conclusion:** When the results obtained from the ML classifiers used in the study are evaluated, the efficiency of the GBC algorithm in the classification of tumors is obvious. It can be said that the success rates obtained from 11 different ML algorithms used in the study are valuable in terms of being used to predict different cancer types.

Keywords: Machine learning- breast cancer- classification- data management- information systems

Asian Pac J Cancer Prev, **23 (10)**, 3287-3297

Introduction

Breast cancer is a very common problem worldwide. This disease is seen in the form of small tumors or masses in the breast tissue, especially in the milk ducts and glands. The fact that these masses have smooth and clear borders indicates that they are benign, while their irregular borders and rough structures indicate that they are malignant, that is, they carry the risk of cancer. In recent years, it is seen that the incidence of breast cancer has increased all over the world, especially among women (Butler, 2020; Yassin et al., 2018). Early diagnosis of this disease, which also causes death, is important for successful treatment. Early and accurate diagnosis of breast cancer can be made with mammographic examination or physical examination (Ashareef et al., 2020). Especially considering the importance of human health, the extent, nature, and risk of the disease should be determined at the early stage by studying the cost of diagnosis and treatment of individuals suffering from a disease (Jouirou et al., 2019; Chaurasia et al., 2018). Mammography is one of the most important screening methods recommended for the mass diagnosis of breast cancer (Zeru et al., 2019). In addition to mammography, breast examination by a specialist physician is important for the early detection of breast cancer. Although rare, some breast cancers cannot be diagnosed with mammography and ultrasound, so physical examination is recommended in addition to mammography. In the process of interpreting and diagnosing the test results obtained in the examinations performed in the diagnosis of breast cancer, expert human knowledge is needed. However, with the developing ML techniques, successful studies are carried out in the diagnosis of breast cancer, as in other types of cancer.

ML is a branch of Artificial Intelligence (AI) that enables computers to quickly detect patterns in complex and large data sets by learning from existing data. The use of ML as an aid to healthcare professionals is increasing rapidly. ML techniques, the use of which is rapidly increasing in the diagnosis of different types of cancer, are also used in the diagnosis of breast cancer. In the literature, it is seen that ML algorithms are used in classification processes for breast cancer detection. With

*[1]Department of Computer Engineering, Faculty of Engineering and Architecture, Istanbul Gelisim University, Istanbul, Turkey. [2]Department of Computer Engineering, Faculty of Engineering, Istanbul Topkapı University, Istanbul, Turkey. [3]Department of Electronics Technology, Istanbul Gelisim Vocational School, Istanbul Gelisim University, Istanbul, Turkey. *For Correspondence: alcetinkaya@gelisim.edu.tr*

increasing success in classification and identification or analysis using data science methods, computer technology has gained decision-making power and developed analysis steps (Pinker et al., 2018). Classifying tumor types occurring in the breast as benign, malignant, or normal tissue and minimizing misdiagnosis is an important part of the reliable treatment process for this disease (Sadhukhan et al., 2020). Data on breast cancer is critical for studies on early detection, rapid and accurate classification as benign or malignant using computerized systems, and evaluation of factors affecting diagnosis (Toğaçar et al., 2020). This classification can be done with ML algorithms. With ML computers can quickly detect patterns in complex and large data sets by learning from existing data. Breast cancer is the second most common cancer among all cancer types and can be fatal if not diagnosed correctly and early. For this reason, it is of great importance that the diagnosis of breast cancer is made accurately and with high performance. The purpose of this study is to predict breast cancer with different ML algorithms, to compare the success results obtained, and to determine the ML algorithm with the highest success rate. For this purpose, the University of Wisconsin breast cancer dataset, each containing 30 features and consisting of 569 samples, were classified using ten different ML techniques. As ML algorithms, Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbor (KNN), Logistic Regression (LR), Multilayer Perceptron (MLP), Linear Discriminant Analysis (LDA), XgBoost (XGB), Ada-Boost (ABC) algorithms were used in the study. As a result of the study, the highest success rate was obtained in the GBC algorithm with an accuracy rate of 99.12%.

The paper is structured as follows: Section 2 provides an overview of related studies. Section 3 describes the research methodology used. Then, in Section 4 experimental design and analysis were done. Section 6 concludes the study.

Related Works

In (Manju et al., 2021; Tharwat, 2019), SVM is mainly used to optimally separate the parameters of the two classes through optimization. In (Ayer et al., 2010; Dreiseitl et al., 2002; Şamkar et al., 2016; Shipe et al., 2019), LR is used in the analysis of experimental data, meteorology, and medicine. In (Islam et al., 2020) SVM, KNN, RF, ANNs ve LR were compared according to the accuracy, precision, and F1 score. The results of the study show that ANNs obtained the highest accuracy with a score of 98.57%. In (DC et al., 2020), KNN is used in classification problems in the industry. In (Kakileti et al., 2020), the authors evaluated the robustness of multiple ML classifiers for breast cancer risk estimation in the presence of incomplete or inaccurate information. In (Rajaguru et al., 2019), the decision Tree and K-Nearest Neighbor (KNN) algorithm are used for the breast tumor classification. In (Oyewola et al., 2016; Yildiz et al., 2016; Wen et al., 2018; S et al., 2019; Ayesha et al., 2020), studies have been carried out on multiple discriminant analyses and to classify sample states by a prediction equation. In (Hendrickx et al., 2020; Hendriks et al., 2019; Rajaguru et al., 2019; Chand, 2020), different studies have been carried out on DT and complex data sets. In (Komura et al., 2019; Wang et al., 2020), studies have been carried out to obtain more accurate and stable estimations in the health sector with RF and DT. In (Song et al., 2020; Qiu et al., 2021; Yu et al., 2020), model development studies with high speed and performance were carried out to categorize features and targets numerically with XGB. In (Yifan et al., 2021; Huang et al., 2019; Wu et al., 2019), studies were carried out by determining the weighted majority vote for the final estimate with DT. In (Lin et al., 2021), a fuzzy cancer cell colony identification system based on a Fuzzy Inference System (FIS) is proposed. The algorithm works with DT. In (Nassar et al., 2021) different machine learning models were used to select the optimal classification model for the prediction of TMB level according to the patient's receptor status. In (Lu et al., 2019), an incremental learning model has been proposed for predicting the survival of patients with breast cancer. For this purpose, a novel genetic algorithm-based online gradient boosting (GAOGB) model has been proposed. As a result of the study, an improvement was achieved with an increase of 28% in the success rate. In (Takci, 2016), centroid-based classifiers for early detection of breast cancer were compared using centroid classifiers and the other classifiers on the Wisconsin Diagnostic and Prognostic Dataset. As a result of this comparison, the highest classification success rate of 99.04% was achieved in the study. Classification of mammogram images has been performed using multilayer sensor networks (Heidari et al., 2019; Agarap, 2018; J et al., 2020). In their clinical reports, successful results have been obtained by applications in breast cancer diagnosis using artificial intelligence algorithms (Ha et al., 2019; Bharati et al., 2020; Karthik et al., 2018; Zou et al., 2019). As can be seen, there are different studies in the literature on cancer diagnosis. In the information age we live in, it is understood that these studies will increase, especially with the increase in the Internet and Artificial Intelligence studies in the health sector. In (Gopinath et al., 2013), the authors developed an automated computer-aided diagnostic system for the diagnosis of thyroid cancer patterns in fine-needle aspiration cytology (FNAC) microscopic images with a high degree of sensitivity and specificity using statistical texture features and a Support Vector Machine classifier (SVM). In (Nindrea et al., 2018) a total of 1,879 articles were reviewed, of which 11 were selected for systematic review and meta-analysis. Five algorithms for ML able to predict breast cancer risk were identified: SVM, Artificial Neural Networks (ANN); Decision Tree (DT), NB, and KNN. With the SVM, the Area Under Curve (AUC) from the SROC was determined > 90%, therefore classified into the excellent category. It is a fact that during the coronavirus disease (COVID-19) period, the use of information technologies and especially the internet has increased, including in the health sector. In this context, the study by Kamal et al., (2020) explains the importance of looking beyond old protocols for pandemic and post-pandemic cancer care and treatments, prognosis and diagnosis of cancer patients, and starting to embrace a future that can maximize outcomes for patients. In (Patil et al., 2020), it was stated that the advances in

machine learning and artificial intelligence had reached a point where they were included in many disciplines, including medicine, and it was emphasized that more interdisciplinary research was needed to disseminate their clinical applications. The study also highlights that more interdisciplinary research is needed to generalize the clinical application of AI, machine learning, and deep learning across all cancer types and different areas of oncology.

As can be seen, there are studies in the literature on the use of ML algorithms in the diagnosis of breast cancer as well as different types of cancer. In these studies, in which the ML algorithm is used, cancer classification can be made and this disease can be diagnosed with high success rates.

Materials and Methods

Machine Learning (ML) Algorithms

ML algorithms, which include various statistical, probability and optimization techniques, can quickly detect patterns in complex and large data sets by learning from the existing data of computers.

Support Vector Machine (SVM)

SVM algorithm creates the best decision boundary for separating each element in the data on a plane where the points are pointed in the n-dimensional space (Tharwat, 2019). This is called a hyperplane. The goal is for this truth to be the maximum margin for points for both classes (Li et al., 2021). This algorithm is a supervised ML algorithm based on statistical learning theory, which can be used for classification and regression operations, used to separate data belonging to two base classes. The function of this situation can be expressed as follows;

$$D = \{(x_i.c_i) \mid x_i \in \mathbb{R}^P, c_i \in \{-1,1\}_{i=1}^n \quad (1)$$

Naive Bayes (NB)

The NB classifier is based on the theorem of British mathematician Thomas Bayes. The NB classifier is a probabilistic approach to the pattern recognition problem, which can be used with a seemingly very restrictive proposition. This premise is that each descriptive attribute or parameter to be used in pattern recognition should be statistically independent. While this proposition limits the use of the NB classifier, it yields comparable results to methods such as more complex neural networks, even when used while stretching the statistical independence condition. An NB classifier can also be thought of as a Bayesian network in which each feature is conditionally independent of each other and the concept to be learned is conditionally dependent on all these features.

Logistic Regression (LR)

LR is a predictive analysis to explain the relationship between the binary dependent variable and a set of independent variables. LR is a model used in binary classification (0 and 1) where the dependent variable is discontinuous (Ayer et al., 2010; Dreiseitl et al., 2002). The characteristic that distinguishes LR analysis from linear regression is that the dependent variable is categorical and not continuous (H. Şamkar et al., 2016). LR is a representation of the function in both ways and is as follows in the figures (Shipe et al., 2019).

$$\pi(x) = \frac{exp(\beta_0 + B_1 x)}{\perp + exp(\beta_0 + B_1 x)} \quad (2)$$

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + B_1 x \quad (3)$$

The operation of the algorithm computes the probability of each state with a combination of its effects on the outcome for an attribute and classifies the probability value according to the highest. In the general logic of NB, the attributes are assumed to be independent of each other. The general logic is to maximize the set of classes. The following formula is derived from Bayes' theorem.

$$P(y|x_1,\dots,x_n) = \frac{P(x_1|y)P(x_2|y)\dots P(x_n|y)\cdot P(y)}{P(x_1)P(x_1)\dots P(x_\eta)} \quad (4)$$

In the formula, P (Y|X) is the probability that event Y is relative to the given event X. If P (X|Y) is the probability that event X occurs when event G occurs (M. M. Islam et al., 2020).

K-Nearest Neighbor (KNN)

KNN is a classification where the most similar data points are found in the education data and an educated guess is made regarding their classification. K is the number of nearest neighbors used by the classifier to make its prediction. KNN makes predictions based on the results of K's nearest neighbors to that point. Based on which neighborhood gives a better result, all the neighbors are examined. The neighborhood value that has the lowest misclassification error is determined as the optimal k value.

$$d(x,y) = \left(\sum_{k=1}^n x_k - yk|^r\right)^{1/r} \quad (5)$$

Linear Discriminant Analysis (LDA)

LDA is a method developed by the British statistician Roland Fisher (D. Oyewola et al., 2016). LDA is based on searching for a linear combination of variables that best distinguishes between good classes (targets) (S et al., 2019). In the dataset, the classification of patients with known X-grain characteristics into classes according to these characteristics is very important for statistical analysis. The computation of the transformation in the optimal LDA for covariance matrices is used to define attributes that have the largest variance (Wen et al., 2018). The LDA model considers the class-to-class dispersion matrix SM and the class-to-class dispersion matrix SF. If C expresses the average value of X data objects for a data set belonging to a grain class, this function is as follows (Yildiz et al., 2016):

$$s_M = \sum_{i=1}^{c} (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^\tau \quad (6)$$

$$s_F = \sum_{i=1}^{c} \sum_{j=1}^{N_i} (\bar{x}_{i,j} - \bar{x})(\bar{x}_{i,j} - \bar{x})^{\tau} \quad (7)$$

Decision Tree (DT)

DT algorithm aims to build a model consisting of one or more trees to classify the dataset in the first stage (Rajaguru et al., 2019). It is a tree algorithm based on transforming a complex process into a simple set of decisions, starting at the top and dividing into lower branches (Chand, 2020). The information required to classify samples by the decision tree method is calculated as follows:

$$I\{s_1, s_{2,\ldots,} s_N\} = -\sum_{i=1}^{M} p_i \, log(p_i) \quad (8)$$

The dataset {S1, S2,..., SN } is divided into N subsets and the entropy value of each attribute is as follows (Ha et al., 2019):

$$E(A) = \sum_{j=1}^{V} \frac{s_{1j} + s_{2j} + \cdots + s_{nj}}{S} \quad (9)$$

Information values the gain values of the computed elements are as follows. The property with the highest value is placed in the topmost root node (Hendriks et al., 2019).

$$G(A) = I\{s_1, s_{2,\ldots,} s_N\} - E(A) \quad (10)$$

Random Forest (RF)

RF algorithm is one of the community algorithms that consists of a combination of multiple decision trees (Komura et al., 2019). This algorithm is an approach put forward by Leo Breiman in (2001). It is a model consisting of a combination of multiple decision trees. After processing the data on N decision trees, an accurate estimate is tried to be produced by taking the average of the estimates obtained. RF solves the overfitting problem, which is one of the most common problems in traditional decision trees, by dividing both the data set and the features into many parts and processing them on multiple trees. After processing the data through the N decision tree, an attempt is made to produce an accurate estimate by averaging the obtained estimates (Wang et al., 2020). Using a decision tree provides more stability than using a single decision tree. Instead of branching the nodes selected from the best features in the data set, the RF decision tree branches all the nodes by choosing the best features randomly selected at each node. Each dataset is created by displacement from the original dataset. Trees are developed using random feature selection and are not pruned.

XgBoost (XGB)

XGBoost is an ML algorithm used for supervised learning classifications and is based on Decision Trees (Yu et al., 2020). In this study, we used this method to classify benign and malignant tumors (Song et al., 2020). XGBoost has a high running speed and makes the computations easier than other algorithms, which makes it a successful machine learning method. For this reason, it provides successful results on multidimensional data (Qiu et al., 2021). XGB is a high-performance version of the Gradient Boosting algorithm optimized with various tweaks. The most important features of the algorithm are that it can achieve high predictive power, prevent over-learning, manage empty data and do them quickly. Software and hardware optimization techniques have been applied to obtain superior results using fewer resources. Its working logic is quite similar to Gradient Boosting. The first step in XGBoost is to get the base score. This estimate can be any number because, combined with the actions to be taken in the next steps, the correct result will be achieved. How good this prediction is examined by the model's erroneous predictions (residual). Errors are found by subtracting the estimated value from the observed value.

Ada-Boost (ABC)

ABC is the most powerful and widely used method in the field of community learning methods (Wu et al., 2019). This algorithm aims to improve the performance of algorithms mainly used in classification to achieve a better result (Huang et al., 2019). It aims to achieve the best result in learning by importing different algorithms developed before it (Yifan et al., 2021). The general logic of the algorithm is that every time the weight of the incorrect estimates found as a result of the previous stage is increased, the weight of the correctly predicted samples is decreased.

$$\mathcal{e} = \frac{1}{N} \sum_{i=1}^{N} I(y_i \neq 6(x_i)) \quad (11)$$

The operating principle of gradient raising, a collective learning method, turns weak learners into strong learners (Lu et al., 2019). Estimation using the gradient boosting model attempts to fit the loss function found in the previous repetition to the negative gradient vector (Takci, 2016).

Gradient Boosting (GBC)

GBC is a machine learning technique for regression and classification problems. This combines weak predictive models to create a model that typically consists of decision trees. The goal of any supervised learning algorithm is to identify and minimize a loss function. GBC is a type of boosting algorithm. It is based on the intuition that the next best possible model, when combined with previous models, minimizes the overall prediction error. The basic idea is to set target results for this next model to minimize error. Gradient Boost can be used for both Classification and Regression.

Multilayer Perceptron (MLP)

MLP is often preferred in classifications from artificial neural network application domains (Heidari et al., 2019). In this study, a multilayer perceptron is preferred in classifying digitized mammograms with "benign" and "malignant" features (Agarap, 2018). As shown in Figure

1, the multilayer sensor has 30 different input layers, 200 intermediate layers, and one output layer. The input layer receives the data from the multilayer network and is passed to the intermediate layer from where it is connected to the output layer, the final layer, and this is how the algorithm works (J et al., 2020).

Deep Learning (DL)

Deep learning aims to obtain a result by making assumptions based on external data by training the obtained data (Ha et al., 2019). However, machine learning, which is an element of artificial intelligence, is a type of predictive analysis. An artificial neural network is a mathematical model that has been compared to the nervous system in the human body. The artificial neural network consists of functional units called neurons. As shown in Figure 2 the artificial neural module, has five main elements such as inputs, weights, addition and activation function, and outputs (Zou et al., 2019). There can be one or more neurons in the layers. An example of an artificial neural network looks as follows:

The function that converts the given parameter values into outputs in the artificial neural network model is expressed by the following formula (S. Karthik et al., 2018).

$$y_m = f^{(2)}\left\{\sum_{j=1}^{m}\left[f^{(1)}\left(\sum_{i=1}^{n} x_{ij} W_{ij} + b_j\right)\right] W_{ij} + b_m\right\} \quad (12)$$

In the function in Figure 3, X1 indicates the inputs to the network, Wij is the weight of the j input for the i output, bi is the constant bias between the input layer and the middle layer, Σ is the sum function, and f-1(net) is the activation function between the input layer and the middle layer. The value Wjk is the weighting of the output layer and the middleware, and the value bj is the bias term between the output layer and the middleware. F-2(net) is the obtained output (Bharati et al., 2020).

The percentage of correctly classified test patterns is considered a success rate. Accuracy, sensitivity, and determinism are other performance criteria (Wu et al., 2020). In a dataset, there are four different possible outcomes; the positive sample is considered true positive (TP) if it is classified correctly, false negative (FN) if it is classified incorrectly, while the negative sample is considered true negative (TN) if it is classified correctly, and false positive (FP) if it is classified incorrectly (Bektaş et al., 2016). The table listing all these possible states is called the error matrix (Table 1).

The ratio of true positives (TP) to the sum of true positives (TP) and false negatives (FN) is called the sensitivity metric (Senturk et al., 2014; Tapak et al., 2019; Cai et al., 2018).

Accuracy is a widely used metric to measure the success of a model (Equation 13):

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (13)$$

The recall is a metric that shows how many of the operations we should have estimated as positive (Equation 14).

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

Precision shows how many of the values we estimated as positive are positive (Equation 15).

$$Precision = \frac{TP}{TP + FP} \quad (15)$$

Dataset

Dataset used in this study was released by the College of Wisconsin in 1992. The dataset consists of 30 features and 569 samples. The features in the dataset are calculated from the result of a needle aspiration biopsy of a cancer tumor. 10 of the 30 features are measured in the tumor cell. Feature selection is the process of removing irrelevant

Table 1. Complexity Matrix

Estimated values	Real values	
	Positive	Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

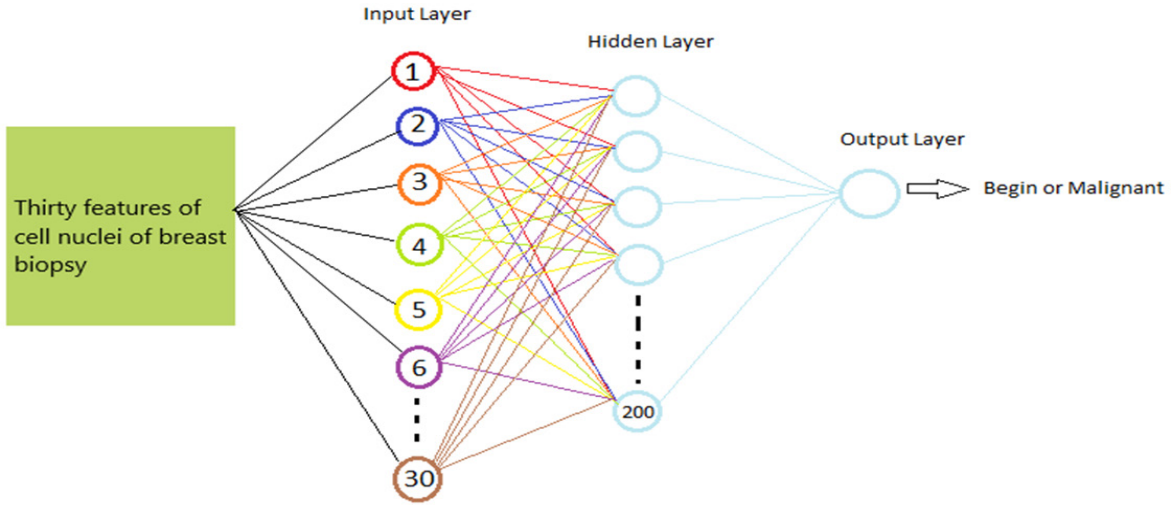

Figure 1. Artificial Neural Network Model

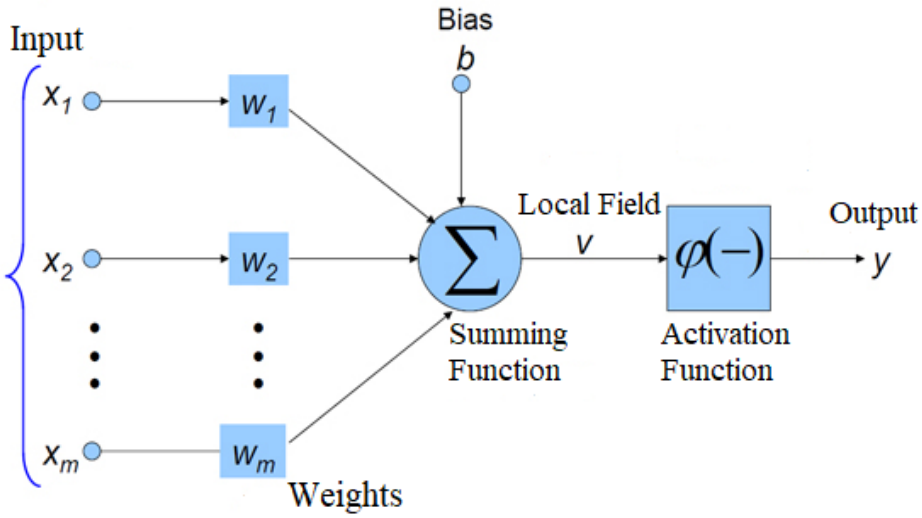

Figure 2. Neural Network Neuron Module

Table 2. Results of All Experiments

Exp. Nu.	Method	Accuracy (%)	Error (%)	Sensitivity (%)	Decision (%)	Precision (%)
1	SVM	97.66	2.34	98.39	97.25	95.31
2	LR	91.22	8.78	92.31	100.00	100.00
3	MLP	95.61	4.39	92.31	97.33	94.74
4	NB	97.37	2.63	97.36	98.67	97.37
5	RF	97.36	2.64	97.37	97.37	94.87
6	DT	97.36	2.64	95.12	98.63	97.50
7	KNN	93.85	6.15	92.11	94.74	89.74
8	LDA	96.49	3.51	100	94.94	89.74
9	XGB	93.86	6.14	88.10	97.22	94.87
10	ABC	98.06	1.94	97.37	98.67	97.37
11	GBC	99.12	0.88	100.00	98.68	97.44
12	SVM	97.36	2.93	97.37	97.37	94.87
13	LR	91.22	8.78	74.36	88.24	74.36
14	MLP	95.61	4.36	97.22	94.87	89.74
15	NB	97.37	2.63	94.87	98.67	97.37
16	RF	97.36	2.93	97.37	97.37	94.87
17	DT	97.36	2.64	97.5	97.30	95.12
18	KNN	93.85	6.15	92.11	94.74	89.74
19	LDA	94.74	5.26	87.18	98.67	97.14
20	XGB	94.74	5.26	94.87	94.67	90.24
21	ABC	97.54	2.46	97.50	97.3	95.12
22	GBC	99.00	1.00	98.66	100.00	100.00
23	SVM	97.36	2.64	95.00	98.65	97.44
24	LR	95.91	5.09	95.33	96.88	98.08
25	MLP	97.66	2.34	99.05	95.45	97.20
26	NB	97.37	2.63	98.67	94.87	97.37
27	RF	96.49	3.51	97.33	94.87	97.33
28	DT	93.85	6.15	94.52	92.68	95.83
29	KNN	85.08	14.95	87.18	80.56	90.67
30	LDA	97.36	2.64	96.15	100.00	100.00
31	XGB	94.74	5.26	92.26	90.24	94.67
32	ABC	96.49	3.51	97.30	92.68	96.00
33	GBC	98.24	1.76	97.40	100.00	100.00

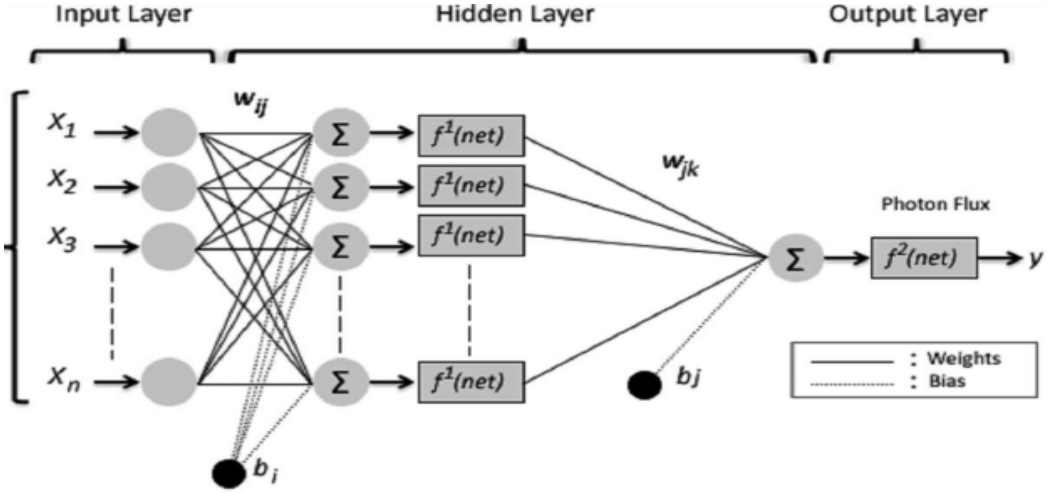

Figure 3. Artificial Neural Network Layer Structure and Main Elements

features or cleaning data from noise. This process directly affects classification success and performance. In this context, the 10 directly measured features used in the study are:

1) Radius,
2) Texture,
3) Diameter,
4) Area,
5) Smoothness,
6) Density,
7) Concavity,
8) Number of concave points,
9) Symmetry,
10) Fractal dimension.

Other features consist of mean, standard error, worst and maximum values obtained from these features. Based on these values, there is a diagnostic class, expressed by labels B (benign) and M (malignant), indicating whether the tumor is benign or malignant. The class distribution of 569 data is 357 benign and 212 malignant. The data were randomly divided into 80% training set and 20% test set and classified according to ML and tested.

Flow Chart of the System

According to this flowchart, we first analyzed the dataset that we used in the study. Then we imported this dataset into the Jupyter platform. By doing this, we visualized the data. The distribution of the features in the dataset is labeled according to the diagnosis. The mean of the feature, "min", "max", "median" and "upper fence" values were displayed as statistical inferences. Similarly, the "standard error" and "worst" columns of the features are plotted and examined according to the diagnosis. Then, a correlation matrix is created to show the relationships between the features of the cancer tumors in the dataset. The correlation matrix is a table of correlation coefficients between several variables. This table shows the binary vector relationship between one variable and every other variable. Each cell in the matrix has a value between "-1" and "1". If this value is close to "1", there is a strong direct correlation between the two data vectors whose correlation is being tested, and if the value is close to "-1", there is a strong inverse relationship. If the correlation value is close to "0", there is no linear relationship between the data. To narrow down the table between the relationships of the characteristics and to facilitate the inferences, they are plotted in a dot plot on a plane. The dot plot is used to examine the positive, negative, or linear relationships between the features. In the data preprocessing steps, it is presented with integer index-based identification. All the "30" characteristics are assigned to the "X" column and the values of the diagnosis column in the data set are assigned to the "y" column. It is a method of splitting the dataset into two "training" and "test" sets. This method

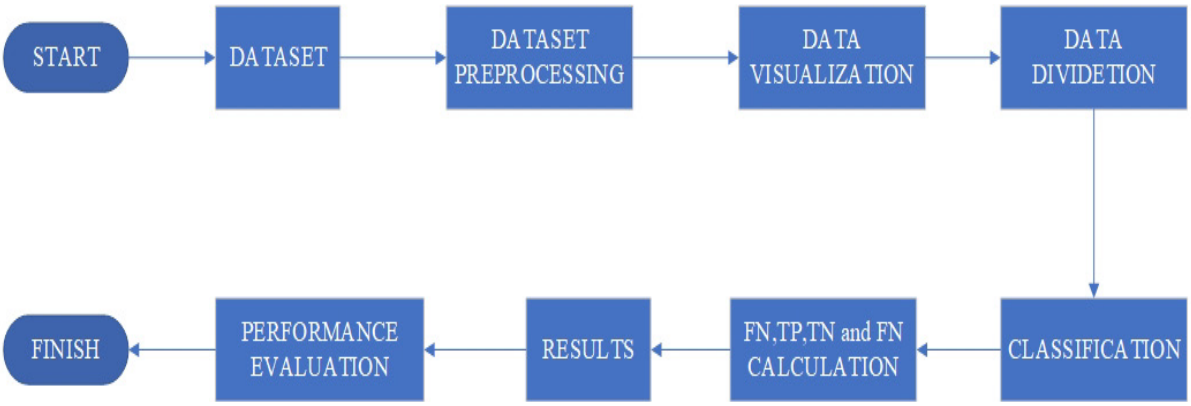

Figure 4. Flow Chart of the Study

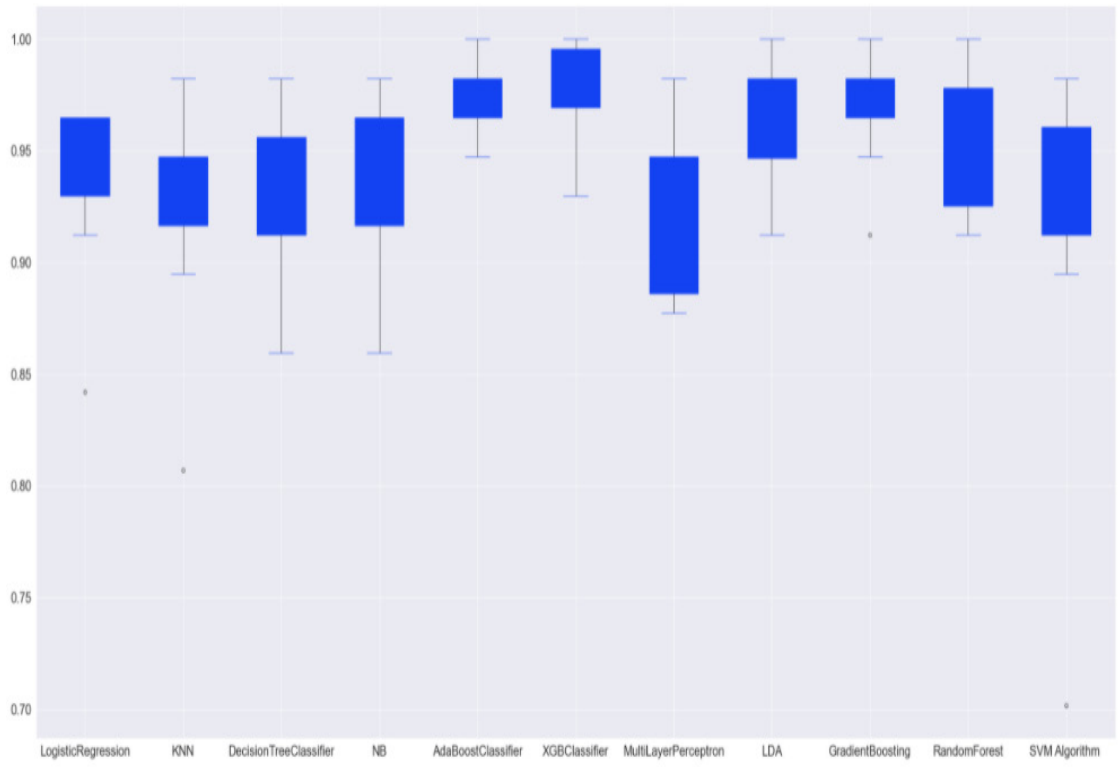

Figure 5. Success Results of All Models Applied to the Study.

is the "hold-out" separation method. While the training set is the data on which the model is trained, the test set is the data used to understand how well the model performs on the untrained data. The next step is to preprocess the data. For different algorithms, the same preprocessing step does not produce good results, but it is necessary to apply it differently. For this scaling, which is called pre-processing, we need to do normalization or standardization depending on the data set. The scaling process only changes the scaling range of our data. The most common normalization method rescales data between "0" and "1". Standardization is centered by subtracting the mean and involves rescaling the distribution of these values so that the mean of the values under study is "0" and the standard deviation is "1". Finally, the classification process begins. Our first machine learning model applied to the dataset is the support vector technique. It is recommended to normalize the data before training this model. The normalization process is applied to all the remaining models. Python programming language is used to create and test the models and eleven different models are applied to the study. The flowchart of the study is shown in Figure 4.

Experimental Design and Analysis

After the training process, the percentage of correct classification is verified using test data. In this study, the metrics of precision, sensitivity, determinism, and classification success are used in the proposed system to compare the performance of machine learning techniques. In the study, sensitivity, accuracy, and definiteness metrics are examined to measure classification performance. As can be seen, in a 30 attribute experiment, XGB, which is one of the upgrade algorithms, is ahead with an accuracy of 99.12% when all metrics are considered. This is followed by the 98.06% success rate of ABC in second place. Compared to the other models, the lowest accuracy is observed for the KNN method with a success

Table 3. Test Results

Exp. Nu.	Experimental Knowledge	Method	Model	Diagnosis	Accuracy (%)	MSE
34	Malignant patient data	Classification	NB	1	97.36	0.0263
35	Benign patient data	Classification	NB	0	97.36	0.0263
36	Malignant patient data	Boosting	GB	1	97.36	0.0263
37	Benign patient data	Boosting	GB	0	97.36	0.0263
38	Malignant patient data	Bagging	RF	1	97.36	0.0263
39	Benign patient data	Bagging	RF	0	97.36	0.0263
40	Malignant patient data	ANN	MLP	1	85.96	0.1403
41	Benign patient data	ANN	MLP	0	85.96	0.1403

rate of 91.11%. When examining the sensitivity metric, the number of correctly predicted positive samples, all models generally perform well, but the LDA and Gradient Boosting methods are the best with a sensitivity of 100%. The worst model is the XGB with 88.10%. LR ranks first with a ratio of 100% and second with an increase in the slope of 98.68% when the crucial criterion, success in correctly classifying the negative samples listed in the table, is examined. It can be observed that logistic regression ranks first with a ratio of 100% when it comes to the elimination rate of false positives, which is a criterion of certainty. The worst models are the LDA and KNN methods with 89.74%. In general, the gradient increase model performed well for all cysts compared to the other models. The accuracy comparisons of the experiments done in the first group are presented in Table 2. The input parameters of this group include 30 features. The data of the experiment is randomly divided into 80% training sets and 20% test sets from 569 samples and tested with twelve different machine learning models, where the creation and testing of the models are done using Python programming language. In the experiments SVM, NB, RF, DT, KNN, LR, MLP, LDA, XGB, ABC, and GBC machine learning algorithms are compared with a total of eleven different classifiers performances testing. The models are compared with both classification accuracy and matrix complexity.

The successful results of our models used in the study are shown in Figure 5. Each range of values in the successful results is shown in blue. The dataset contains 30 features that can be used in this study. In the experiments performed on the dataset, the most successful outcome is sought by performing different stages. In this study, 41 experiments are conducted in four different stages and the results are examined. In these stages, the features that make up our input parameters are reduced and the most successful model prediction is determined using different experiments. The most successful results of these experiments range from 95% to 99%.

Results

Prediction of breast cancer disease with high accuracy rates is an important threshold for the diagnosis and treatment of cancer. Early and accurate diagnosis of this cancer is of great importance in terms of preventing and combating this disease. Today, ML algorithms are also used within the scope of AI in the diagnosis of this and similar types of cancer. Tumors in breast cancer can be benign (non-cancerous) or malignant (cancerous). Since these two different labels are specified in the breast cancer data, cancer estimation is made using these data. In our study, SVM, NB, RF, DT, KNN, LR, MLP, LDA, XGB, ABC, and GBC learning algorithms were used. With these ML algorithms, benign or malignant tumor data were classified, and thus breast cancer prediction success rates were determined. These success rates obtained as a result of the study were compared based on each ML algorithm. As a result of the comparison, the highest success rate was obtained in the GBC algorithm with an accuracy rate of 99.12%. The results of the research are valuable in terms of using this method to predict different types of cancer, as well as containing important information for the diagnosis of breast cancer with machine learning.

Discussion

In future studies, we aim to diagnose benign and malignant tumors with deep learning algorithms (LSTM, CNN, RNN) using the same dataset and compare the success rates achieved.

Author Contribution Statement

All authors contributed equally.

Acknowledgments

Conflicts of interest

The authors declare that there are no conflicts of interest associated with this article.

References

AA Heidari, H Faris, I Aljarah, et al (2019). An efficient hybrid multilayer perceptron neural network with grasshopper optimization. *Soft Computing*, **23**, 7941-58.

AFM Agarap (2018). On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset, In Proceedings of the 2nd international conference on machine learning and soft computing, pp 5-9.

A Jouirou, A Baâzaoui, W Barhoumi (2019). Multi view information fusion in mammograms: A comprehensive overview. *Information Fusion*, **52**, 308-21.

Ashareef B, Yaseen W, Jawa W, et al (2020). Breast cancer awareness among female school teachers in Saudi Arabia: A Population Based Survey. *Asian Pac J Cancer Prev*, **21**, 337-42.

Ayer T, Chhatwal J, Alagoz O, et al (2010). Informatics in radiology: comparison of logistic regression and artificial neural network models in breast cancer risk estimation. *Radiographics*, **30**, 13-22.

B Bektaş, S Babur (2016).Machine learning based performance development for diagnosis of breast cancer. In 2016 Medical Technologies National Congress (TIPTEKNO) IEEE, pp1-4.

BR Manju, V Athira, A Rajendran (2021). Efficient multi-level lung cancer prediction model using support vector machine classifier. In IOP Conference Series: Materials Science and Engineering, 1012, no. 1, p. 012034.

Butler R (2020). Invited commentary: Breast Cancer Risk Assessment and Screening Strategies-What's New?. *Radiographics*, **40**, 937-40.

DCY S Pal (2020). Discovery of hidden pattern in thyroid disease by machine learning algorithms. *Indian J Public Health Res Develop*, **11**, 61-6.

D Oyewola, D Hakimi, K Adeboye, et al (2016). Using five machine learning for breast cancer biopsy predictions based on mammographic diagnosis. *Int J Eng Technol*, **2**, 142-5.

D Yifan, L Jialin, F Boxi (2021). Forecast model of breast cancer diagnosis based on RF-AdaBoost," In 2021 International Conference on Communications, Information System and Computer Engineering (CISCE) IEEE, pp 716-9.

Dreiseitl S, Ohno-Machado L (2002).Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform*, **35**, 352-9.

E Yildiz, Y Sevim (2016). Comparison of linear dimensionality

reduction methods on classification methods," In 2016 National conference on elec-trical, electronics and biomedical engineering (ELECO), pp 161-4.
Gopinath B, Shanthi N (2013). Support Vector Machine based diagnostic system for thyroid cancer using statistical texture features. *Asian Pac J Cancer Prev*, **14**, 97-102.
H Lu, H Wang, SW Yoon (2019).A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis. *Exp Sys Appl*, **116**, 340-50.
H Şamkar, AG Yıldırım, Ö Delibaş (2016). Determining the risk factors causing cancer with logistic regression analysis. *Alphanumeric J*, **4**, 205-22.
H Takci (2016). Diagnosis of breast cancer by the help of centroid based classifiers. *J Facul Eng Architect Gazi Univ*, **31**, 323-30.
Ha R, Chang P, Karcich J, et al (2019). Convolutional neural network based breast cancer risk stratification using a mammographic dataset. *Acad Radiol*, **26**, 544-9.
Hendrickx JO, van Gastel J, Leysen H, et al (2020). High-dimensionality data analysis of pharmacological systems associated with complex diseases. *Pharmacol Rev*, **72**, 191-217.
Hendriks MP, Verbeek X, van Vegchel T, et al (2019). Transformation of the national breast cancer guideline into data-driven clinical decision trees. *JCO Clin Cancer Inform*, **3**, 1-14.
JH, Chiu TH, S Li, et al (2020). Breast cancer–detection system using PCA, multilayer perceptron, transfer learning, and support vector machine. *IEEE Access*, **8**, 204309-24.
J Wen, X Fang, J Cui, et al (2018). Robust sparse linear discriminant analysis. *IEEE Trans Circuits Syst*, **29**, 390-403.
Kakileti ST, Manjunath G, Dekker A, et al (2020). Robust estimation of breast cancer incidence risk in presence of incomplete or inaccurate information. *Asian Pac J Cancer Prev*, **21**, 2307-13.
Kamal V, Kumari D (2020).Use of artificial intelligence/machine learning in cancer research during the COVID-19 pandemic. *Asian Pac J Cancer Care*, **5**, 251-3.
Komura D, Ishikawa S (2019).Machine learning approaches for pathologic diagnosis. *Virchows Arch*, **475**, 131-8.
L Tapak, N Shirmohammadi-Khorram, P Amini, et al (2019). Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clin Epidemiol Global Health*, **7**, 293-9.
Li J, Zhou Z, Dong J, et al (2021). Predicting breast cancer 5-year survival using machine learning: A systematic review. *PLoS One*, **16**, e0250370.
MM Islam, MR Haque, H Iqbal, et al (2020). Breast cancer prediction: a comparative study using machine learning techniques. *SN Comput Sci*, **1**, 1-14.
M Toğaçar, KB Özkurt, B Ergen, et al (2020). BreastNet: A novel convolutional neural network model through histopathological images for the diagnosis of breast cancer," Physica A: Statistical Mechanics and its Applications, **545**, 123592.
Nassar A, Lymona AM, Lotfy MM, et al (2021).Tumor mutation burden prediction model in Egyptian breast cancer patients based on next generation sequencing. *Asian Pac J Cancer Prev*, **22**, 2053-59.
Nindrea RD, Aryandono T, Lazuardi L, et al (2018). Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: a Meta-Analysis. *Asian Pac J Cancer Prev*, **19**, 1747-52.
Patil S, Moafa IH, Mosa Alfaifi M, et al (2020). Reviewing the role of artificial intelligence in cancer. *Asian Pac J Cancer Biol*, **5**, 189-199.
Pinker K, Chin J, Melsaether AN, et al (2018). Precision medicine and radiogenomics in breast cancer: New Approaches toward Diagnosis and Treatment. *Radiology*, **287**, 732-47.
Q Huang, Y Chen, L Liu, et al (2019). On combining biclustering mining and AdaBoost for breast tumor classification. *IEEE Trans Knowl Data Eng*, **32**, 728-38.
R Song, T Li, Y Wang (2020). Mammographic classification based on XG boost and DCNN with multi features. *IEEE Access*, **8**, 75011-21.
Rajaguru H, S RS (2019).Analysis of decision tree and K-nearest neighbor algorithm in the classification of breast cancer. *Asian Pac J Cancer Prev*, **20**, 3777-81.
S RS, Rajaguru H (2019).Comparison analysis of linear Discriminant Analysis and Cuckoo-Search Algorithm in the classification of breast cancer from digital mammograms. *Asian Pac J Cancer Prev*, **20**, 2333-37.
S Ayesha, MK Hanif, R Talib (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, **59**, 44-58.
S Bharati, P Podder, M Mondal (2020). Artificial neural network based breast cancer screening: a comprehensive review. arXiv preprint arXiv2006.01767.
S Chand (2020). A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method. *Machine Vision Appl*, **31**, 1-10.
SF Lin, HT Chen, YH Lin (2021). Automatic counting cancer cell colonies using fuzzy inference system. *J Informat Sci Engin*, **27**, 749-76.
S Karthik, RS Perumal, PC Mouli (2018). Breast cancer classification using deep neural networks In Knowledge computing and its applications. Springer Singapore, pp 227-41.
S Sadhukhan, N Upadhyay, P Chakraborty (2020). Breast cancer diagnosis using image processing and machine learning. In Emerging Technology in Modelling and Graphics Springer, pp 113-127.
S Wang, Y Wang, D Wang, et al (2020). An improved random forest-based rule extraction method for breast cancer diagnosis. *Appl Soft Comput*, **86**, 105941.
Shipe ME, Deppen SA, Farjah F, et al (2019).Developing prediction models for clinical use using logistic regression: an overview. *J Thorac Dis*, **11**, 574-84.
T Cai, H He, W Zhang (2018). Breast cancer diagnosis using imbalanced learning and ensemble method. *Appl Comput Math*, **7**, 146-54.
Tharwat A (2019).Parameter investigation of support vector machine classifier with kernel functions. *Knowledge Information Sys*, **61**, 1269-1302.
V Chaurasia, S Pal, BB Tiwari (2018).Prediction of benign and malignant breast cancer using data mining techniques. *J Algorithms Computational Technol*, **12**, 119-26.
Wu CC, Yeh WC, Hsu WD, et al (2019).Prediction of fatty liver disease using machine learning algorithms. *Comput Methods Programs Biomed*, **170**, 23-9.
Wu N, Phang J, Park J, et al (2020). Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans Med Imaging*, **39**, 1184-94.
Y Qiu, J Zhou, M Khandelwal, et al (2021). Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration. *Engin Comput*, **2021**, 1-18.
Y Zeru, L Sena, Shaweno T (2019). Knowledge, attitude, practice, and associated factors of breast cancer self-examination among urban health extension workers in Addis Ababa, Central Ethiopia. *J Midwifery Reproduct Health*, **7**, 1662-72.
Yassin NIR, Omran S, El Houby EMF, et al (2018).Machine learning techniques for breast cancer computer aided

diagnosis using different image modalities: A systematic review. *Comput Methods Programs Biomed*, **156**, 25-45.
Yu D, Liu Z, Su C, et al (2020).Copy number variation in plasma as a tool for lung cancer prediction using Extreme Gradient Boosting (XGBoost) classifier. *Thorac Cancer*, **11**, 95-102.
ZK Senturk, R Kara (2014). Breast cancer diagnosis via data mining: performance analysis of seven different algorithms," *Comput Sci Engin*, **4**, 35-46.
Zou L, Yu S, Meng T, et al (2019). A technical review of convolutional neural network-based mammographic breast cancer diagnosis. *Comput Math Methods Med*, **2019**, 6509357.