


Breast Cancer Prediction using Machine Learning Models

Orlando Iparraquirre-Villanueva¹, Andrés Epifanía-Huerta²,
Carmen Torres-Ceclén³, John Ruiz-Alvarado⁴, Michael Cabanillas-Carbonell⁵
Facultad de Ingeniería y Negocios, Universidad Norbert Wiener, Lima, Perú¹
Facultad de Ingeniería de Sistemas, Universidad Nacional de San Martín, Perú²
Facultad de Ingeniería, Universidad Católica los Ángeles de Chimbote, Perú³
Facultad de Ingeniería, Universidad Tecnológica del Perú, Lima, Perú⁴
Facultad de Ingeniería, Universidad Privada del Norte, Lima, Perú⁵

Abstract—Breast cancer is a type of cancer that develops in the cells of the breast. Treatment for breast cancer usually involves X-ray, chemotherapy, or a combination of both treatments. Detecting cancer at an early stage can save a person's life. Artificial intelligence (AI) plays a very important role in this area. Therefore, predicting breast cancer remains a very challenging issue for clinicians and researchers. This work aims to predict the probability of breast cancer in patients. Using machine learning (ML) models such as Multilayer Perceptron (MLP), K-Nearest Neighbor (KNN), AdaBoost (AB), Bagging, Gradient Boosting (GB), and Random Forest (RF). The breast cancer diagnostic medical dataset from the Wisconsin repository has been used. The dataset includes 569 observations and 32 features. Following the data analysis methodology, data cleaning, exploratory analysis, training, testing, and validation were performed. The performance of the models was evaluated with the parameters: classification accuracy, specificity, sensitivity, F1 count, and precision. The training and results indicate that the six trained models can provide optimal classification and prediction results. The RF, GB, and AB models achieved 100% accuracy, outperforming the other models. Therefore, the suggested models for breast cancer identification, classification, and prediction are RF, GB, and AB. Likewise, the Bagging, KNN, and MLP models achieved a performance of 99.56%, 95.82%, and 96.92%, respectively. Similarly, the last three models achieved an optimal yield close to 100%. Finally, the results show a clear advantage of the RF, GB, and AB models, as they achieve more accurate results in breast cancer prediction.

Keywords—Prediction; models; machine learning, cells; breast cancer

I. INTRODUCTION

Breast cancer can be classified as a type of cancer that occurs in the cells of the breast. Both men and women can get it, although women are more likely than men to suffer from it. The process of breast cancer begins with the uncontrolled

growth of cells in the lining of the breast [1]. At first, there are no symptoms of pain or cancerous growth, and has a low potential for metastatic growth and is limited to the lobe where it grows without generating any symptoms [2],[3]. Symptoms of breast cancer can include anything from a small lump in the breast to changes in the shape of the breast or changes in the color of the skin [4], to identify breast cancer early, it is important to undergo early detection tests, as there are many types of breast cancer and many of them do not cause symptoms at first. Lobular carcinoma in situ, for example, is a type of cancer that occurs in the area of abnormal milk-producing cells of the breast. Invasive lobular carcinoma, which develops in the lobules of the milk-producing mammary glands, people with this symptom experience thickening of the breast tissue, swelling of the breast, and change in skin texture. Ductal carcinoma in situ, this type of cancer usually does not cause symptoms, it is discovered through mammography and invasive ductal is the most common type of cancer accounting for approximately 80% of cases [5]–[7]. There is solid evidence that alcohol consumption, growing older, having dense breasts, family history, radiotherapy treatments, obesity and exposure to radiation increase the risk of breast cancer [2], [8] in turn, it has been shown that prolonged breastfeeding, the development of the physical activity, avoiding harmful consumption of alcoholic beverages and refraining from smoking save, avoiding prolonged use of hormones reduce the risk of breast cancer [8], [9], [10]. Also, mortality from breast cancer in 2020 was 684,996 worldwide, representing 24% of all cancers. While it is true, in recent years the rates of breast cancer events and mortality have been decreasing worldwide [11]. For example, China has the highest rate of breast cancer, with 17.1%; Africa reached 2.5%; the United States at 4%, Japan at 7%; Morocco at 12.5%; Hungary at 2.1% [12]. As shown in Fig. 1, the countries with the highest rates of breast cancer are present in all continents; the continent of Asia concentrates the highest number of deaths from breast cancer.

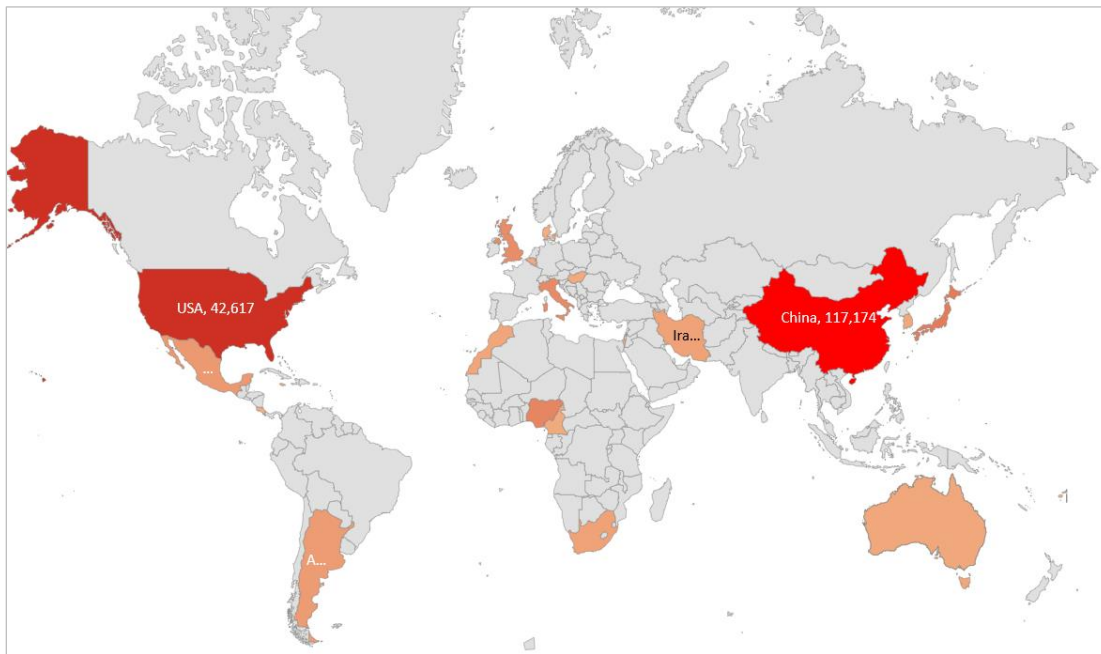


Fig. 1. Breast cancer by country 2022.

In the last decade, technology has undergone impressive development, and with it, ML models are becoming increasingly popular for breast cancer prediction. These models can be used to analyze large patient data sets, such as mammograms, to identify patterns associated with breast cancer development [13]. From these data, ML models can accurately predict a person's risk of developing breast cancer. The accuracy of these models can be further improved by incorporating additional factors such as lifestyle, diet, and family history [14], [15]. With the increasing availability of high-quality datasets and technological advances, ML models are becoming increasingly reliable for breast cancer prediction [16], [17]. There are many types of ML models that can be used to predict the probability that a person will develop breast cancer [18] in this paper we use the classification models such as MLP, KNN, AB, Bagging, GB, and RF, considering that they have excellent performance and performance to analyze and correlate the measurements of the established features. Using features associated with cancer cell imaging, breast cancer can be predicted using ML models. This field of action is in constant development from two deans to after [19], [20].

This paper uses the Wisconsin breast cancer diagnostic dataset to predict and diagnose the likelihood of breast cancer in patients by analyzing six ML models. The dataset is composed of digitized mammogram images and consists of 569 observations and 31 attributes [21]. It also incorporates nine parameters set on a scale of 1 to 10 with values categorized into "benign" or "malignant" tumors.

The article's organization is divided into the following sections. In section II, we describe the most important works that have been done in the area of models of ML. In section III, you will find a description of the method and examples of its application. A summary of the results and discussion of the study can be found in Section IV. Lastly, in Section V, we will present the conclusions that have been reached.

II. PREVIOUS STUDIES

WHO, American Cancer Society, and scholars have published work related to breast cancer. For example, in [22], [23] they analyzed six ML models with the aim of determining the degree of accuracy of each of them. For this, they used three parameters such as age, cell type with cancer, and cell interface receptors. Also, in [24] developed a predictive model to categorize people with breast cancer using the logistic regression (LR) model, GB model, decision tree (DT), and RF model. Obtaining the following results for the LR model 81.9%; GBT with 82%; RF with 82.8%, respectively. Similarly, in [25] they proposed a model to detect breast cancer using ML models. The tests were performed on a dataset consisting of 317,880 clinical observations. The proposed model achieved an accuracy of 91.22%, and a false rejection rate of 112%. Also, in [26] they used a strategy with feature selection, extraction, and classifier algorithms for breast cancer diagnosis. This study included 762 patients with breast cancer and 138 people without cancer. ML algorithms were used a: 1) LR; 2) SVM; 3) Bagging; 4) GNB; 5) DT; 6) GB; 7) K-NN; 8) BNB; 9) RF; 10) AB, 11) Extra Trees (ET) and 12) MLP. The models that achieved the best results were: LR+MLP with 94%. ML models have demonstrated their contribution to the prediction and early diagnosis of cancer. For example, in [27] they conducted a study to predict and diagnose breast cancer using ML models, for which they used parameters such as specificity, sensitivity, precision, accuracy, precision, and F1 score. The GBDT model obtained a score of 96.77 outperforming all other models. The advancement of Artificial Intelligence (AI) has allowed ML techniques and algorithms to become increasingly efficient in prediction, as evidenced in [28] where they developed a model using ML algorithms to identify and classify different types of cancer. They applied the RF, SVM, and RF models to correctly classify breast cancer cases, obtaining a result: sensitivity of 97.12%, specificity of

96.14%, and accuracy of 97.11%. Artificial intelligence has played a very important role in clinical fields, so much so that, in [29] they evaluated the repeatability of ML model types such as re-regressive, multiclass classification, binary rating, and ordinal classification. The results indicated that classification accuracy improved significantly in most environments. Breast cancer negatively affects the quality of life of patients. In view of this, in [28] they selected an appropriate model to classify and predict the causes that lead to contracting breast cancer, for this purpose they used 970 people with breast cancer. As a result, the SVM model showed the highest sensitivity and an accuracy of 91%, demonstrating that the application of ML algorithms helps the classification of characteristics and the optimization of the genetic algorithm. Accurately distinguishing malignant and benign tumors in patients is crucial to saving lives. That is why in [30] they developed a technique for binary classification of malignant tumors of breast cancer, for which they used three pre-trained convolutional neural network (CNN) models such as RestNet-50, EfficientNetb0, and Inception-v3, applying transfer learning and fine-tuning. The proposed method achieved an accuracy of 98.92%, a sensitivity of 99.87%, a specificity of 97.97%, and an F1 score of 0.9987. In the same line, [30] developed an algorithm based on artificial neural networks (ANN), with the purpose of predicting breast cancer, achieving the following results: accuracy of 98.74%, and an F1 score of 98.02%. Computer-assisted breast cancer screening improves the chances of early detection and diagnosis. So, in [31], [32] proposed a breast cancer screening technique to assess the probability of recurrence of individuals with cancer. The model was trained with 6447 patients diagnosed with breast cancer, the data features were classified with conventional ML and CNN. The best accuracy yielded 88.8%, accuracy 89%, and an F1 score of 0.5. The rapid growth of ML models such as CNNs has promoted the massive use of these technologies in biomedical image classification. For example, in [33] they developed an ML technique to classify breast cancer from histopathological images. The model has been tested with the publicly available BreakHis dataset and has obtained significant accuracy.

III. METHODOLOGY

This section presents the theoretical basis of the MLP, KNN, AB, Bagging, GB, and RF models and the development of the work to predict and diagnose breast cancer.

A. Multi-layer Perceptron

The MLP is an ANN type. It uses backpropagation to train the network [34]. The MLP is composed of multiple layers, each of which is connected to all the others, forming a directed network [35]. The MLP learns a feature from a set of inputs and combines the various features into a set of outputs [36]. The layers usually have weights and polarization units that are adjusted during training. It should be noted that, with the exception of the input nodes, each node in the network is a neuron using a nonlinear activation function, and its equation is given by the following equation and is represented by the following Eq. (1).

$$h_{1j} = f\left(\sum_{i=1}^n w_{ij}x_i + b_j\right) \quad (1)$$

MLP is widely used in supervised learning, where it can learn to classify and predict data. In equation (1), h_{1j} is defined as node j of the hidden layer h_1 , w_{ij} represents the input gate of node j of the hidden layer h_1 and b_j is the bias. In MLP network training, loss functions play an important role. The feature vectors are modeled by the network using loss functions, which are evaluated based on how well the architecture models them. As shown in Fig. 2, the multilayer perceptron model.

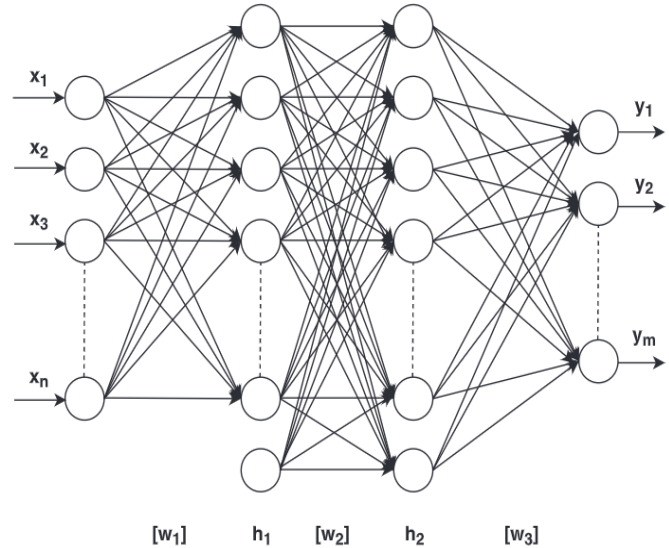


Fig. 2. MLP architecture.

MLPs are limited by their structure, as they are not as flexible as deep learning networks, but they can still be powerful classifiers. Moreover, they do not require large amounts of data, which makes them suitable for many applications [34]. This is the number of training epochs that increases the loss function and gradually reduces its error through optimization.

B. K-Nearest Neighbor

As a nonparametric supervised learning classifier, the K-NN algorithm uses proximity to perform classifications and predictions to perform classifications and predictions, respectively [35]. The algorithm stores the attribute vectors and labels used during its training phase so that the algorithm can be retrained [36]. To label the unlabeled vector, K is set as a user-defined variable, and a label is assigned among the training attributes that are considered most relevant to classify the vector [37]. As for distance metrics for continuous variables, Euclidean distance is used, which is limited to real-valued vectors, for which Eq. (2) is used, and for discrete variables, the overlap metric is used [38].

$$d(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2} \quad (2)$$

The use of the K-NN model in ML mainly has a better performance in classification and prediction. For example, in data processing, estimating values, automatic recommendations, finance, credit data, in health, its best results have been in predicting the risk of heart attacks, breast cancer, and prostate cancer [39].

C. AdaBoost

AB is an ML classification algorithm; its principle is based on building strong classifiers by combining basic or weak classifiers. This classification algorithm works on adaptive sampling to select the between samples [40]. This algorithm iteratively trains the weak classifiers, for which it uses weighted data to incorporate it into an ensemble, to then have the strong classifier [41], as shown in Fig. 3.

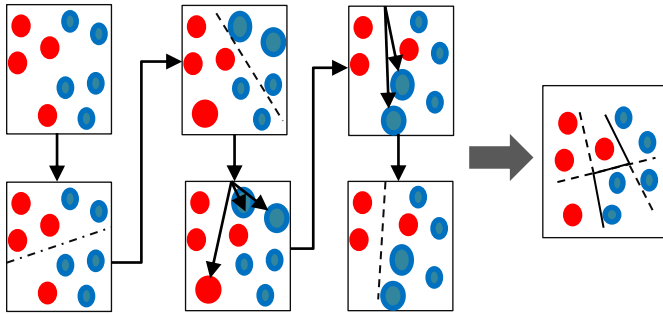


Fig. 3. AB algorithm scheme of work.

Fig. 3 shows that the AB algorithm generates several weak classifiers, where each of the classifiers has a set weight in its performance. Finally, the prediction is obtained by combining the weak classifiers and voting by weight.

D. Bagging

The bagging model is an ML technique used to improve the accuracy and stability of classification algorithms. It works by combining multiple weak classifiers to form a more robust prediction model [42]. The idea is to create multiple versions of the classifier, each with a different set of parameters, and then combine the results from all of them to produce a better overall prediction [43]. These types of algorithms are run in parallel and seek to take advantage of the independence that exists between single-classifier algorithms, given that the best classifier is chosen by the majority. The Bagging implementation process follows the following steps: Step 1: multiple subsets are created from the data set; Step 2: the base model is created in each of the training subsets; Step 3: each model learns in parallel with each training set; Step 4: the final predictions are determined by combining the predictions of all models.

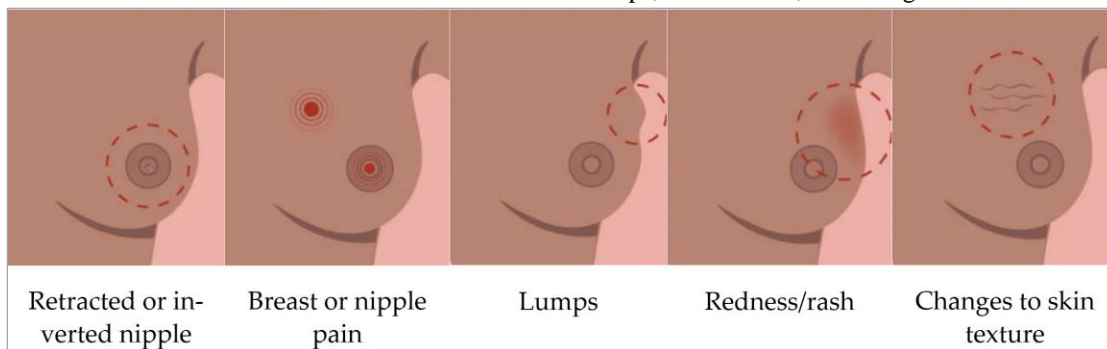


Fig. 5. Signs of breast cancer.

E. Gradient Boosting

This classifier combines several weak predictors into a single strong predictor [44]. Using this method, the accuracy of the predictors can be increased by adding predictors sequentially to a set of predictors, each of which corrects the previous one [44]. Basically, the goal of this technique is to find the best predictor for a given problem by iteratively training the model using weak predictors, and gradually improving them until they become strong learners just before solving the problem [45]. This technique has many applications, from data mining to ML or IA.

F. Random Forest

In the ML field, RF is an algorithm that works as an ensemble. To make predictions, a large number of decision trees are used together to create the decision tree [46]. A decision tree is created using a random subset of the data, and then the results of each tree are combined to make a final prediction, based on the results of all the trees [47]. In terms of classification and regression areas, RF is an extremely powerful algorithm. It can handle large data sets and can be used for both supervised and unsupervised learning. Fig. 4 shows what the model prediction looks like for a new observation.

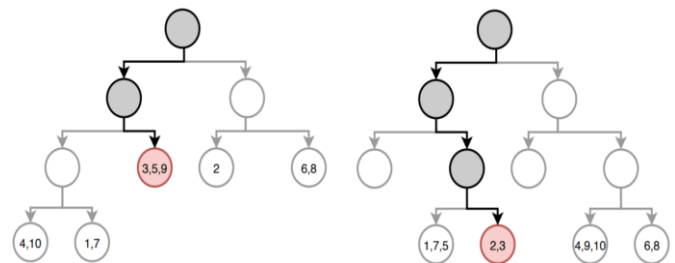


Fig. 4. RF algorithm flowchart.

G. Understanding Data

According to the American Cancer Society (ACS), a lump is one of the most common symptoms of breast cancer [1]. Several benign breast conditions can cause cancer-like symptoms. However, some of these disappear with time and others require medical treatment. These conditions include cysts, mastitis, hyperplasia, sclerosing adenosis, intraductal papilloma's, fibroadenoma, radial scar, fatty necrosis, and phyllodes tumors. Fig. 5 shows some signs of breast cancer, such as: retracted or inverted nipple, breast or nipple pain, lumps, redness/rash, and changes in skin texture.

For this work, the Wisconsin Breast Cancer Diagnostic Dataset was used to identify and predict breast cancer. For this purpose, six classification models were used: MLP, K-NN, AB, Bagging, GB, and RF. In addition, univariate analysis, bivariate analysis, and correlation analysis are used for exploratory data analysis (EDA). To evaluate the accuracy of the model, the following methods are used: confusion matrix, classification report, and AUC. The dataset corresponds to digitized images of samples and is composed of 569 observations and 31 attributes: diagnosis, Radius-mean(R-ME), Texture-mean(T-ME), pe-perimeter-mean (P-ME), area-mean(A-ME), smoothness-mean(S-ME), compact-ness-mean(C-ME), concavity-mean(CO-ME), concave points-mean(CP-ME), sym-metry-mean(S-ME), fractal dimension-mean(FD-ME), radius-se(R-SE), tex-ture-se(T-SE), perimeter-se(P-SE) area-se(A-SE), smoothness-se(S-SE), compact-ness-se(C-SE), concavity-se(CO-SE), concave points-se(CP-SE), symmetry-se(S-SE), fractal-dimension-se(F-D-SE), radius-worst(R-WO), texture-worst(T-WO), perimeter-ter-worst(P-WO), area-worst(A-WO), smoothness-worst(S-WO), compact-ness-worst(CO-WO), concavity-worst(C-WO), concave points-worst(CP-WO), sym-metry-worst(S-WO) and fractal-dimension-worst(F-D-WO).

H. Data Cleansing

The data cleaning process, for this case study, was performed using Python programming language was performed using a variety of libraries and techniques. Among the libraries used were Pandas, NumPy, SciPy, Scikit-learn, and NLTK. The Pandas library was used to read data, clean it and manipulate it. It is useful for dealing with missing values, outliers, and other problems. The NumPy library was used to perform calculations on the data, such as mean, median, mode and standard deviation. SciPy and Scikit-learn are declared for the use of ML and statistical analysis. Also, it is used to perform regression, clustering, and other types of analysis. NLTK library is declared for further use for data processing. Also, it will be used to extract text features, such as sentiment analysis and keyword extraction. We then proceeded with loading the dataset and identifying each of the variables, as shown in Table I. The number of variables and the type of data for each of the variables. In addition, in this section, we try to eliminate all duplicate data, handle outliers and deal with incorrect data.

TABLE I. DATASET VARIABLES AND DATA TYPES

Column	not empty Count	Dtype
[diagnosis]	569 (not empty)	Blob
[R-ME]	569 (not empty)	Float 64
[T-ME]	569 (not empty)	Float 64
[P-ME]	569 (not empty)	Float 64
[A-ME]	569 (not empty)	Float 64
[S-ME]	569 (not empty)	Float 64
[C-ME]	569 (not empty)	Float 64
[CO-ME]	569 (not empty)	Float 64
[C-P-ME]	569 (not empty)	Float 64
[S-ME]	569 (not empty)	Float 64
[FD-ME]	569 (not empty)	Float 64

[R-SE]	569 (not empty)	Float 64
[T-SE]	569 (not empty)	Float 64
[P-SE]	569 (not empty)	Float 64
[A-SE]	569 (not empty)	Float 64
[S-SE]	569 (not empty)	Float 64
[C-SE]	569 (not empty)	Float 64
[CO-SE]	569 (not empty)	Float 64
[CP-SE]	569 (not empty)	Float 64
[S-SE]	569 (not empty)	Float 64
[F-D-SE]	569 (not empty)	Float 64
[R-WO]	569 (not empty)	Float 64
[T-WO]	569 (not empty)	Float 64
[P-WO]	569 (not empty)	Float 64
[A-WO]	569 (not empty)	Float 64
[S-WO]	569 (not empty)	Float 64
[CO-WO]	569 (not empty)	Float 64
[C-WO]	569 (not empty)	Float 64
[CP-WO]	569 (not empty)	Float 64
[S-WO]	569 (not empty)	Float 64
[FD-WO]	569 (not empty)	Float 64

I. Exploratory Data Analysis

EDA is an approach to data analysis for organizing key features. Primarily, EDA is used to see what the data can say beyond the formal task of modeling or hypothesis testing. EDA is also used to check the data for interesting features or outliers that may suggest the need for further examination. In addition, EDA can be used to evaluate the assumptions of a model before fitting it to the data. In order to visualize the data graphically, the diagnosis column first had to be enumerated so that Malignant(M)=1, Benign(B)=0. Then, the ID column was set for the dataset, the ID column will not be used for ML. For this, the countplot(), plt.figure() and print() functions were used. As shown in Fig. 6.

Now, for a better understanding of the content of Table II, it is important to have basic knowledge about variance, standard deviation, number of samples, or the maximum and minimum values. This type of information provides a better understanding of what is happening with the data. Therefore, before visualization understands standardization, feature extraction, and feature selection.

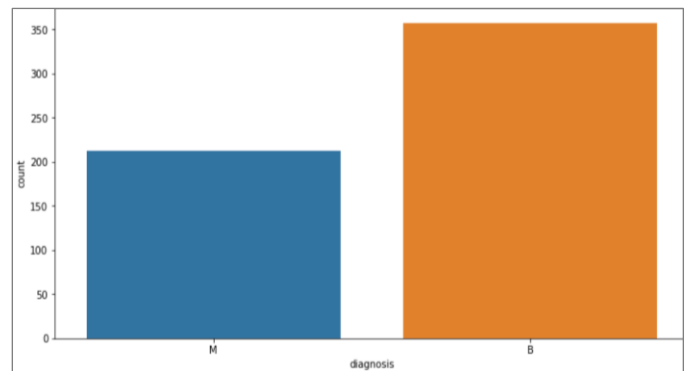


Fig. 6. M and B cancer diagnosis count.

TABLE II. STANDARDIZATION, EXTRACTION, AND SELECTION OF CHARACTERISTICS

	Radius mean	medium texture	Perimeter mean	middle zone	Smoothness mean	Compactness mean
count	[569.00000]	[569.00000]	[569.00000]	[569.00000]	[569.00000]	[569.00000]
mean	[14.127292]	[19.289649]	[91.969033]	[654.889104]	[0.0963600]	[0.1043410]
std	[3.5240490]	[4.3010360]	[24.298981]	[351.914129]	[0.0140640]	[0.0528130]
min	[6.9810000]	[9.7100000]	[43.790000]	[143.500000]	[0.0526300]	[0.0193800]
25%	[11.700000]	[16.170000]	[75.170000]	[420.300000]	[0.0863700]	[0.0649200]
50%	[13.370000]	[18.840000]	[86.240000]	[551.100000]	[0.0958700]	[0.0926300]
75%	[15.780000]	[21.800000]	[104.10000]	[782.700000]	[0.1053000]	[0.1304000]
max	[28.110000]	[39.280000]	[188.50000]	[2501.00000]	[0.1634000]	[0.3454000]

For better visualization of the data, we used the seaborn library, but we classified the features into three groups because the differences between the feature values were so high that it was impossible to observe them, as shown in Fig. 7. Each group includes 10 features for a more effective presentation of the data.

Fig. 7 can be seen. For example, that the T-ME features, the median of M and B appear separate, so it can be very useful for classification. The FD-ME feature, however, does not separate the median of the M and B, so the median in this case cannot be used to classify the data. For reasons of space, the following groups are not shown. In the classification, it was also shown that the variables C-WO and CP-WO are very similar. However, it cannot be stated that they are correlated with each other, in the case of being correlated; one of the two variables is eliminated. To compare the two characteristics more deeply, the joint plot is used.

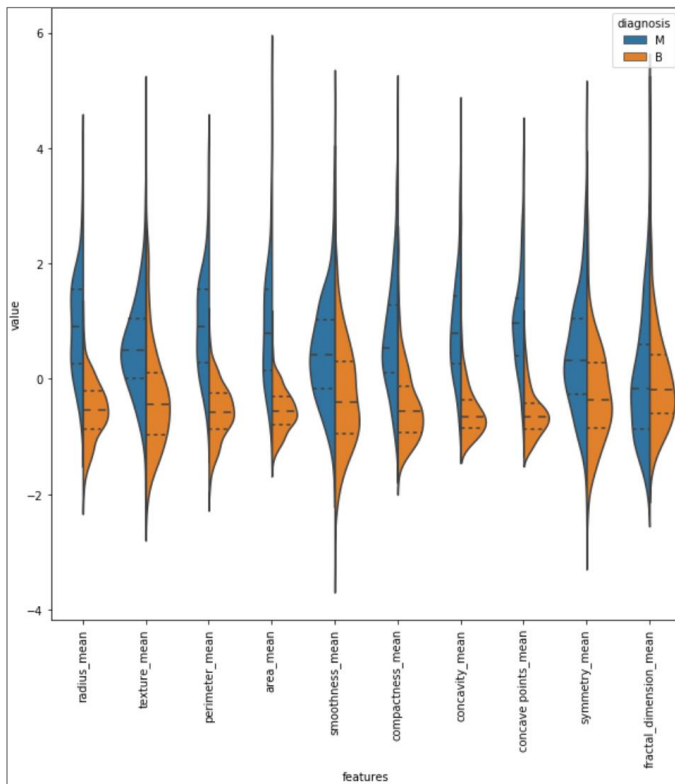


Fig. 7. Standardization and classification of characteristics.

In the next step, features are selected using correlations, univariate features are selected, recursive feature elimination with cross-validation is performed, and attribute categorization is performed. MLP, K-NN, AB, Bagging, GB, and RF classification are used to train the model and predict. As shown in Fig. 8, the R-ME, M-ME, and A-ME features are correlated with each other, so only the A-ME feature will be used. In this way, the features that are correlated are found, with support of the classifiers. C-ME, CO-ME, and CP-ME are correlated with each other, so only CO-ME is chosen. In addition, R-SE, P-SE, and A-SE are correlated, so only A-SE was used. R-WO, P-WO, and A-WO are correlated, so I use A-WO. CO-WO, C-WO, and CP-WO are correlated, so C-WO was used. C-SE, CO-SE, and CP-SE are correlated, so I use CO-SE, T-ME, and T-WO are correlated so I use T-ME, A-WO, and A-ME are correlated so I use A-ME. Specifically, X and Y are not correlated at all; the correlation seen in Fig. 8 is such a strong correlation by chance.

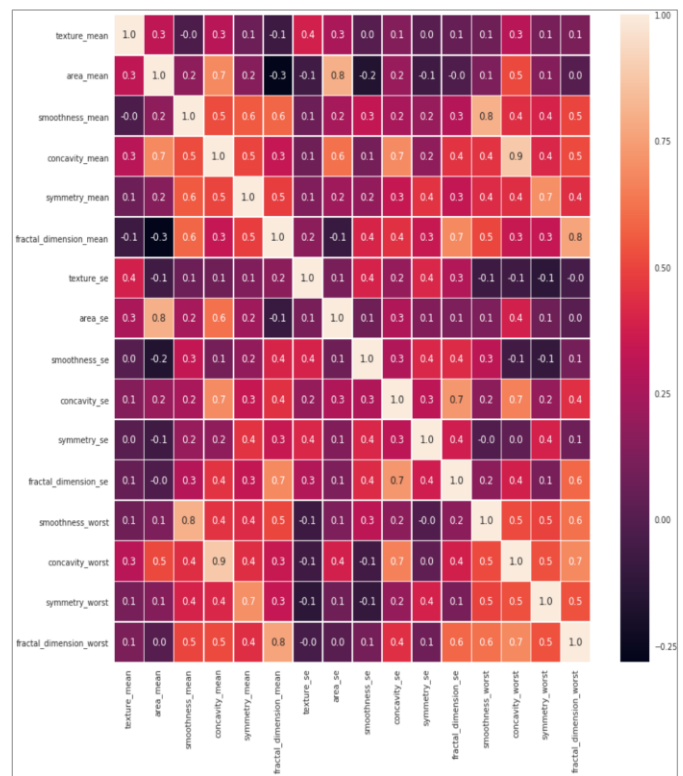


Fig. 8. Numerical correlation of variables.

As part of this work, we use the normalization technique for feature scaling to convert feature values into a mean-centered distribution with unit standard deviation, and this feature scaling method has been widely used in ML algorithms. There are several types of neural networks, such as MLP, K-NN, AB, Bagging, GB, and RF. For example, there is a requirement to normalize features in algorithms such as K-NN and MLP. As a result of the different properties measured by the dataset at each scale, there are heterogeneous features among the datasets at each scale.

J. Model Training and Testing

In univariate feature selection, SelectKBest is used which eliminates all features except those with high scores. This method allows choosing the number of features to use. For example, the number of features(k)=5, which means that the model must find the 5 best features, this is achieved with the following function: SelectKBest(arg, k=5).fit(x_train, y_train). The results are presented in Table III.

The next step consists of preparing the MLP, K-NN, AB, Bagging, GB, and RF models for training and validation using the train_test_split(), project_data.drop(),

X_train.select_dtypes() and Pipeline() functions. The latter allows training the model with the data by adjusting its parameters to create a model that can accurately predict the result while evaluating the model to ensure its accuracy and reliability.

Then the prepare_model() function is used to compile the model with a given number of features. It takes the features as an argument and returns a compiled model as its output. Also, the function prepare_confusion_matrix(y_true, y_pred) is used to print the confusion matrix, as shown in Fig. 9. Similarly, the function prepare_classification_report() is used to generate the classification report for the given results. Finally, the prepare_roc_curve() function allows preparing the receiver operating characteristic (ROC) curve and calculates the false positive rate and the true positive rate, which allows for measuring specificity, and sensitivity, among others. After the evaluation, the following results were obtained [Bagging: 99.78021978021978%; K-NN: 96.7032967032967%; RF: 100.0%, AB: 99.56043956043956%; GB: 100.0% and MLP: 96.26373626373373626%]. It should be noted that only four models have been presented in Fig. 9: Bagging, K-NN, AB, GB.

TABLE III. SELECTION OF UNIVARIATE CHARACTERISTICS

list: [
6.06916433e	3.66899557e	1.00015175e	1.30547650e	1.95982847e
3.42575072e	4.07131026e	6.12741067e	1.32470372e	6.92896719e
1.39557806e	2.65927071e	2.63226314e	2.58858117e	1.00635138e
1.23087347e]				
List of features: Index([
texture_mean	area_mean	smoothness_mean	concavity_mean	menmetry_fraction_mean
texture_rease	suavidad_se	concavidad_se	simetría_se	fractal_dimension_se
suavidad_peor	concavidad_peor	simetría_peor	fractal_dimension_peor)]

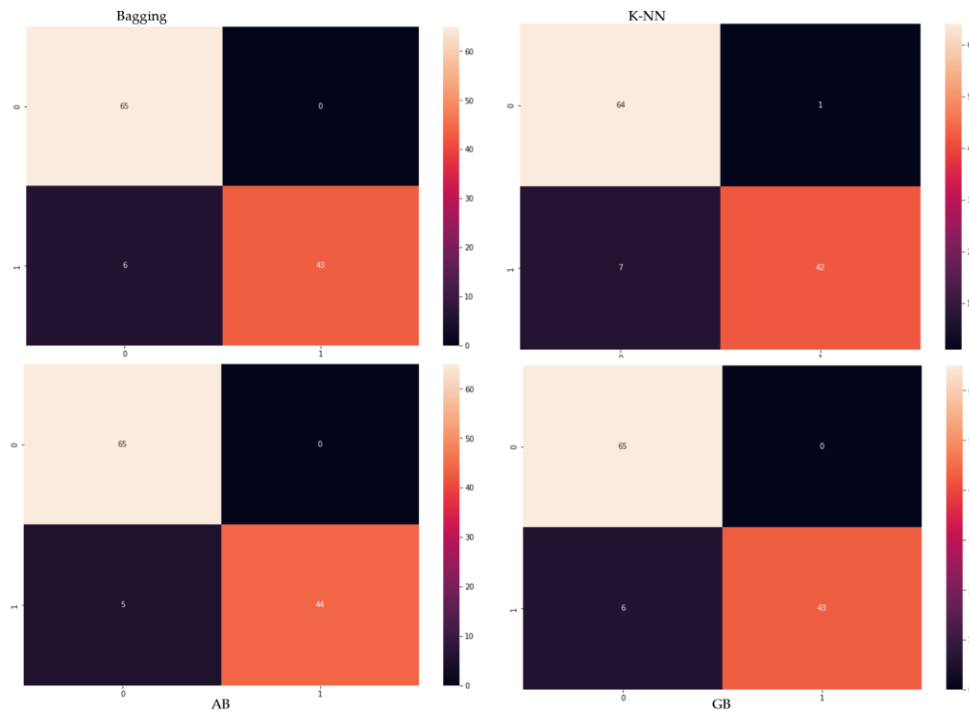


Fig. 9. Matrix of confusion.

IV. RESULTS AND DISCUSSION

After training the MLP, K-NN, AB, Bagging, GB, and RF models, on the data set, a learning algorithm is created and used for training. The performance of the models with

unobserved data is then evaluated. The evaluation of each of the models was performed by testing their performance on unseen data. Metrics such as accuracy, precision, recall, F1 score, and ROC curve are used to determine model performance as shown in Table IV.

TABLE IV. MODEL EVALUATION RESULTS

bagging classifier Report				
	accuracy [%]	recall [%]	f1-score [%]	support
B	93	97	95	71
M	95	88	92	43
accuracy			94	114
macro avg	94	93	93	114
weighted avg	94	94	94	114
KNN classifier Report				
	accuracy [%]	recall [%]	f1-score [%]	support
B	92	99	95	71
M	97	86	91	43
accuracy			94	114
macro avg	95	92	93	114
weighted avg	94	94	94	114
RF classifier Report				
	accuracy [%]	recall [%]	f1-score [%]	support
B	93	93	93	71
M	88	88	88	43
accuracy			91	114
macro avg	91	91	91	114
weighted avg	91	91	91	114
AB classifier Report				
	accuracy [%]	recall [%]	f1-score [%]	support
B	97	92	94	71
M	87	95	91	43
accuracy			93	114
macro avg	93	93	93	114
weighted avg	93	93	93	114
GB classifier Report				
	accuracy [%]	recall [%]	f1-score [%]	support
B	96	94	95	71
M	91	93	92	43
accuracy			94	114
macro avg	93	94	93	114
weighted avg	94	94	94	114
MLP Report				
	accuracy [%]	recall [%]	f1-score [%]	support
B	95	97	96	71
M	95	91	93	43
accuracy			95	114
avg	95	94	94	114
weighted avg	95	94	94	114

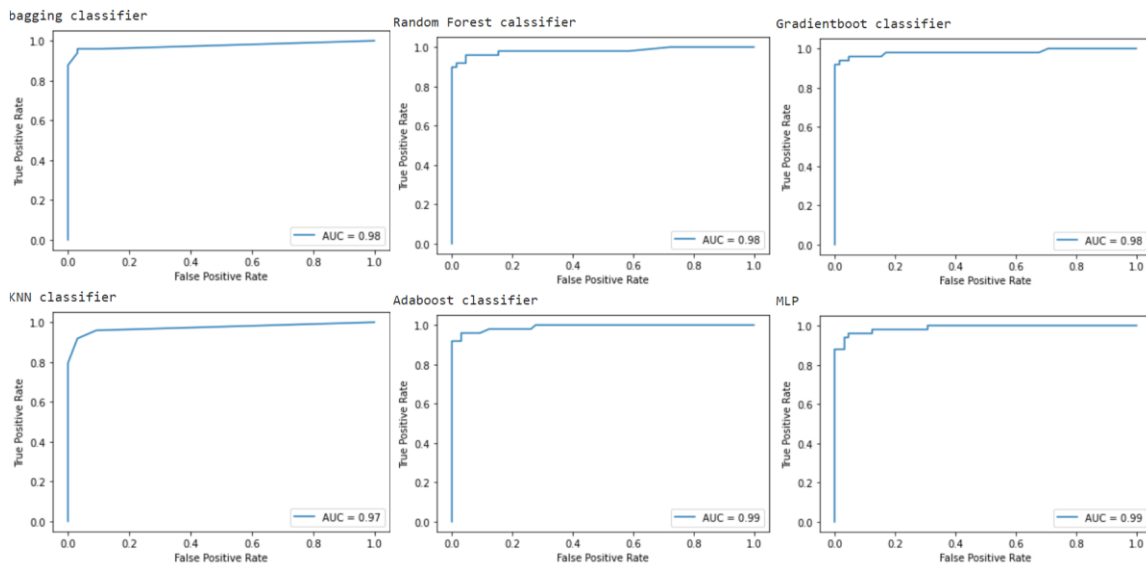


Fig. 10. Performance curve of ML models.

The false positive rate for each model is calculated in a similar way. For example, in the case of the Bagging technique, it helps to improve stability and accuracy by creating various models; in the K-NN model it is the ratio of false positives to the total number of points; in the RF model, it is the rate of false positives that the classifier incorrectly identifies a negative event (e.g., a "no" or a "0") as a positive event (e.g., a "yes" or a "1"); in the AB model the "false positive" rate and the true positive rate depend on the algorithm, the complexity of the data analyzed and the parameters used for the classifier. In general, the false positive rate is quite low, and the "true positive" rate is quite high for AB and in the MLP model, the false positive rate is the probability of misclassifying a true negative case as a positive case. In much the same way the true positive rate is calculated. For example, in the Bagging model, the true positive rate is the proportion of correctly classified positive cases divided by the total number of positive cases; similarly, in the RF and GB model, the true positive rate is the rate at which the classifier correctly identifies a positive event (e.g., a "yes" or a "1"); in the MLP model, the true positive rate is the probability that it correctly classifies a true positive case. Now, for each model, we evaluated the AUC (AUC) performance curve. For example, the models used in this work; Bagging, K-NN, RF, AB, GB, and MLP, obtained the following performance: 98%, 97%, 98%, 98%, 99%, 98%, and 99%, respectively. Fig. 10 shows that the performance curve of each of the models is optimal, reaching practically on average 98%, this makes it possible to opt for any of the models used in this work to classify and predict breast cancer.

For the training and validation of each model used, we worked with an adequate data set. The results shown in Fig. 10 and Table IV show that the performance of each model was successful in cancer prediction accuracy. These results showed superiority in the same ML models in [24] and [26] where the Bagging and K-NN models achieved a performance of 96.47% and 96.40% in predicting Breast Cancer. These results do not determine that one is better than the other, on the contrary, the

performance rate varies according to different factors, and one of them is the volume of data with which it is trained. On the other hand, in [26] they developed a model to predict breast cancer, for which they used the RF model, with which they achieved an accuracy performance of 97.1%, very similar to the 98% accuracy obtained in this work. AI has played a very important role in clinical fields, and models such as AB have contributed a great deal in this field, since it is the model that has achieved the best results, reaching 99% in this study. Likewise, in [29] in the binary classification of malignant tumors of breast cancer, it reached 99.92% accuracy, which makes it the best model for classifying and predicting breast cancer. Similarly, the GB model, which is an excellent classifier by adding predictors sequentially, achieved a 98% performance rate in training, which is also in agreement with the results obtained in [30], where they used the GB model for the purpose of predicting breast cancer, where it achieved a 98.74% performance rate. Finally, the MLP model is characterized as one of the best predictors, this predictor learns a feature from a set of inputs and combines the different features in a set of outputs, the performance rate of this model has been 99%, and it is a result with a high pre-accuracy rate, which allows it to be a reliable option for the prediction of breast cancer. Also, [20], [21] used this model with three clinical factors: age, cancer cell type, and cell surface receptors, obtaining satisfactory results, with a performance rate of 98%. The summary of the analysis of the 6 models used in this work to predict breast cancer is presented in Table V.

TABLE V. SUMMARY OF THE ANALYSIS

Model	Train Accuracy	AUC SCORE
Bagging	99.56	0.97
KNN	95.82	0.97
RF	100	0.98
Adaboost	100	0.96
GB	100	0.97
MLP	96.92	0.98

V. CONCLUSIONS

Prediction of different types of cancer is one of the most complex fields of medical engineering and AI. In this work, 6 ML models were trained for breast cancer prediction, for which the Wisconsin breast cancer diagnostic dataset was used, with the purpose of predicting and diagnosing in patients the probability of having breast cancer. The dataset corresponds to digitized images of samples and is composed of 569 observations and 31 attributes. Also, the performance of the results of each of the models was analyzed, as shown in Fig. 10. Also, the behavior was compared in the context of the work developed: the Random Forest classifier, Adaboost, and Gradientboot, achieved the best results of 100%, more accurate in terms of breast cancer prediction. The normalization technique was used for feature scaling with the purpose of converting the feature values into an input distribution at the mean with a unit standard deviation. This can be seen in the numerical correlation of variables in Fig. 8, also, Table III shows the univariate characteristics. Table V shows the accuracy of each model: Bagging 99.56%; KNN 95.82%; Random Forest 100%; Adaboost 100%, Gradientboot 100%; and MLP 96.92%. The main contributions of this work consist of the evaluation of 6 ML models for breast cancer prediction. Likewise, the results keep a clear originality of this work, and at the same time confirm that the results obtained in this work are related to other similar works that used ML techniques applied to breast cancer prognosis.

In the future, a possible development that would complement the use of the models would be the development of a mobile application based on services to consume the implemented model. The most important contribution of this work is that doctors through ML models can analyze the data of breast cancer patients in a personalized way to predict their effectiveness, constituting a support tool for health. Limitations of this work include: 1) The data used for training may be biased, which means that there may be biases between terms; 2) the quality of ML model data depends on the quality and volume; if the data is limited, the results will be inaccurate; 3) in terms of resources, training ML models requires a processor with a high responsiveness.

REFERENCES

- [1] American Cancer Society, "What Is Breast Cancer?," 2020. <https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html>.
- [2] World Health Organization, "Breast cancer," 2021. <https://www.who.int/news-room/fact-sheets/detail/breast-cancer> (accessed Jan. 02, 2023).
- [3] World Cancer research Fund International, "Breast cancer statistics International." <https://www.wcrf.org/cancer-trends/breast-cancer-statistics/>.
- [4] I. Mihaylov, M. Nisheva, and D. Vassilev, "Application of machine learning models for survival prognosis in breast cancer studies," *Information (Switzerland)*, vol. 10, no. 3, 2019, doi: 10.3390/INFO10030093.
- [5] A. Sharma and P. K. Mishra, "Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis," *International Journal of Information Technology (Singapore)*, vol. 14, no. 4, pp. 1949–1960, Jun. 2022, doi: 10.1007/S41870-021-00671-5/METRICALS.
- [6] S. S. Yadav and S. M. Jadhav, "Thermal infrared imaging based breast cancer diagnosis using machine learning techniques," *Multimed Tools Appl.*, vol. 81, no. 10, pp. 13139–13157, Apr. 2022, doi: 10.1007/S11042-020-09600-3/METRICALS.
- [7] Z. Zeng et al., "Identifying Breast Cancer Distant Recurrences from Electronic Health Records Using Machine Learning," *J Health Inform Res.*, vol. 3, no. 3, pp. 283–299, Sep. 2019, doi: 10.1007/S41666-019-00046-3/METRICALS.
- [8] A. Alzu'bi, H. Najadat, W. Doulat, O. Al-Shari, and L. Zhou, "Predicting the recurrence of breast cancer using machine learning algorithms," *Multimed Tools Appl.*, vol. 80, no. 9, pp. 13787–13800, Apr. 2021, doi: 10.1007/S11042-020-10448-W/METRICALS.
- [9] S. Rani, M. Kaur, and M. Kumar, "Recommender system: prediction/diagnosis of breast cancer using hybrid machine learning algorithm," *Multimed Tools Appl.*, vol. 81, no. 7, pp. 9939–9948, Mar. 2022, doi: 10.1007/S11042-022-12144-3/METRICALS.
- [10] N. Arya and S. Saha, "Multi-modal advanced deep learning architectures for breast cancer survival prediction," *Knowl Based Syst.*, vol. 221, p. 106965, Jun. 2021, doi: 10.1016/J.KNOSYS.2021.106965.
- [11] S. Lei et al., "Global patterns of breast cancer incidence and mortality: A population - based cancer registry data analysis from 2000 to 2020," *Cancer Commun.*, vol. 41, no. 11, p. 1183, Nov. 2021, doi: 10.1002/CAC2.12207.
- [12] OMS, "Breast cancer." <https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer> (accessed Jan. 02, 2023).
- [13] P. Xuan, L. Jia, T. Zhang, N. Sheng, X. Li, and J. Li, "LDAPred: A method based on information flow propagation and a convolutional neural network for the prediction of disease-associated lncRNAs," *Int J Mol Sci.*, vol. 20, no. 18, Sep. 2019, doi: 10.3390/IJMS20184458.
- [14] A. Turcu-Stiolica et al., "Diagnostic Accuracy of Machine-Learning Models on Predicting Chemo-Brain in Breast Cancer Survivors Previously Treated with Chemotherapy: A Meta-Analysis," *Int J Environ Res Public Health.*, vol. 19, no. 24, Dec. 2022, doi: 10.3390/IJERPH192416832.
- [15] O. Iparraguirre-Villanueva et al., "The Public Health Contribution of Sentiment Analysis of Monkeypox Tweets to Detect Polarities Using the CNN-LSTM Model," *Vaccines* 2023, Vol. 11, Page 312, vol. 11, no. 2, p. 312, Jan. 2023, doi: 10.3390/VACCINES11020312.
- [16] M. F. Aslan, "A hybrid end-to-end learning approach for breast cancer diagnosis: convolutional recurrent network," *Computers and Electrical Engineering.*, vol. 105, p. 108562, Jan. 2023, doi: 10.1016/J.COMPELECENG.2022.108562.
- [17] P. Wang et al., "Cross-task extreme learning machine for breast cancer image classification with deep convolutional features," *Biomed Signal Process Control.*, vol. 57, p. 101789, Mar. 2020, doi: 10.1016/J.BSPC.2019.101789.
- [18] Y. Kaya and F. Kuncan, "A hybrid model for classification of medical data set based on factor analysis and extreme learning machine: FA + ELM," *Biomed Signal Process Control.*, vol. 78, p. 104023, Sep. 2022, doi: 10.1016/J.BSPC.2022.104023.
- [19] A. Ahuja, L. Al-Zogbi, and A. Krieger, "Application of noise-reduction techniques to machine learning algorithms for breast cancer tumor identification," *Comput Biol Med.*, vol. 135, p. 104576, Aug. 2021, doi: 10.1016/J.COMPBIOMED.2021.104576.
- [20] O. Iparraguirre-Villanueva et al., "Search and classify topics in a corpus of text using the latent dirichlet allocation model," *Indonesian Journal of Electrical Engineering and Computer Science.*, vol. 30, no. 1, pp. 246–256, Apr. 2023, doi: 10.11591/IJEECS.V30.I1.PP246-256.
- [21] A. Ahuja, L. Al-Zogbi, and A. Krieger, "Application of noise-reduction techniques to machine learning algorithms for breast cancer tumor identification," *Comput Biol Med.*, vol. 135, p. 104576, Aug. 2021, doi: 10.1016/J.COMPBIOMED.2021.104576.
- [22] K. N. Chitrala, M. Nagarkatti, P. Nagarkatti, and S. Yeguvapalli, "Analysis of the TP53 deleterious single nucleotide polymorphisms impact on estrogen receptor alpha-p53 interaction: A machine learning approach," *Int J Mol Sci.*, vol. 20, no. 12, Jun. 2019, doi: 10.3390/IJMS20122962.
- [23] H. Y. Tsai et al., "Integration of Clinical and CT-Based Radiomic Features for Pretreatment Prediction of Pathologic Complete Response to Neoadjuvant Systemic Therapy in Breast Cancer," *Cancers (Basel)*, vol. 14, no. 24, Dec. 2022, doi: 10.3390/CANCERS14246261.

- [24] F. Xiong, X. Cao, X. Shi, Z. Long, Y. Liu, and M. Lei, "A machine learning-Based model to predict early death among bone metastatic breast cancer patients: A large cohort of 16,189 patients," *Front Cell Dev Biol*, vol. 10, Dec. 2022, doi: 10.3389/FCELL.2022.1059597.
- [25] S. Khozama and A. M. Mayya, "A New Range-based Breast Cancer Prediction Model Using the Bayes' Theorem and Ensemble Learning," *Information Technology and Control*, vol. 51, no. 4, pp. 757–770, Dec. 2022, doi: 10.5755/J01.ITC.51.4.31347.
- [26] E. Taghizadeh, S. Heydarheydari, A. Saberi, S. JafarpourNesheli, and S. M. Rezaei, "Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods," *BMC Bioinformatics*, vol. 23, no. 1, Dec. 2022, doi: 10.1186/S12859-022-04965-8.
- [27] R. R. Kadhim and M. Y. Kamil, "Comparison of machine learning models for breast cancer diagnosis," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 1, pp. 415–421, Mar. 2023, doi: 10.11591/IJAI.V12.II.PP415-421.
- [28] S. A. Mortazavi, "Machine Learning Models for Predicting Breast Cancer Risk in Women Exposed to Blue Light from Digital Screens," *J Biomed Phys Eng*, Apr. 2022, doi: 10.31661/JBPE.V010.2105-1341.
- [29] A. Lemay et al., "Improving the repeatability of deep learning models with Monte Carlo dropout," Feb. 2022, doi: 10.1038/S41746-022-00709-3.
- [30] D. Clement, E. Agu, J. Obayemi, S. Adeshina, and W. Soboyejo, "Breast Cancer Tumor Classification Using a Bag of Deep Multi-Resolution Convolutional Features," *Informatics*, vol. 9, no. 4, p. 91, Oct. 2022, doi: 10.3390/INFORMATICS9040091.
- [31] H. Wang, Y. Li, S. A. Khan, and Y. Luo, "Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network," *Artif Intell Med*, vol. 110, p. 101977, Nov. 2020, doi: 10.1016/J.ARTMED.2020.101977.
- [32] S. M. and J. Joy, "A machine learning based framework for assisting pathologists in grading and counting of breast cancer cells," *ICT Express*, vol. 7, no. 4, pp. 440–444, Dec. 2021, doi: 10.1016/J.ICTE.2021.02.005.
- [33] S. Chattopadhyay, A. Dey, P. K. Singh, D. Oliva, E. Cuevas, and R. Sarkar, "MTRRE-Net: A deep learning model for detection of breast cancer from histopathological images," *Comput Biol Med*, vol. 150, p. 106155, Nov. 2022, doi: 10.1016/J.COMPBIOMED.2022.106155.
- [34] Y. Zhou, Y. Niu, Q. Luo, and M. Jiang, "Teaching learning-based whale optimization algorithm for multi-layer perceptron neural network training," *Mathematical Biosciences and Engineering*, vol. 17, no. 5, pp. 5987–6025, Sep. 2020, doi: 10.3934/MBE.2020319.
- [35] O. Iparraguirre-Villanueva et al., "Convolutional Neural Networks with Transfer Learning for Pneumonia Detection," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, p. 2022, Accessed: Jan. 04, 2023. [Online]. Available: www.ijacsa.thesai.org.
- [36] Fix Evelyn; Hodges Joseph, "Discriminatory Analysis. Nonparametric," 1951.
- [37] P. A. Jaskowiak and R. J. G. B. Campello, "Comparing Correlation Coefficients as Dissimilarity Measures for Cancer Classification in Gene Expression Data", Accessed: Nov. 07, 2022. [Online]. Available: <https://www.researchgate.net/publication/260333185>.
- [38] F. Nigsch, A. Bender, B. van Buuren, J. Tissen, E. Nigsch, and J. B. O. Mitchell, "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization," *J Chem Inf Model*, vol. 46, no. 6, pp. 2412–2422, 2006, doi: 10.1021/CI060149F/SUPPL_FILE/CI060149F-FILE002.XLS.
- [39] P. Kasemsumran and E. Boonchieng, "EEG- Based Motor Imagery Classification Using Novel String Grammar Fuzzy K-Nearest Neighbor Techniques with One Prototype in Each of Classes," 2020 International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2020, pp. 742–745, Feb. 2020, doi: 10.1109/ICAIIIC48513.2020.9065236.
- [40] R. Saini, "Integrating Vegetation Indices and Spectral Features for Vegetation Mapping from Multispectral Satellite Imagery Using AdaBoost and Random Forest Machine Learning Classifiers," *Geomatics and Environmental Engineering*, vol. 17, no. 1, pp. 57–74, 2023, doi: 10.7494/GEOM.2023.17.1.57.
- [41] T. H. Nguyen and A. T. Vu, "An Efficient Differential Evolution for Truss Sizing Optimization Using AdaBoost Classifier," *CMES - Computer Modeling in Engineering and Sciences*, vol. 134, no. 1, pp. 429–458, 2023, doi: 10.32604/CMES.2022.020819.
- [42] N. H. A. Malek, W. F. W. Yaacob, Y. B. Wah, S. A. Md Nasir, N. Shaadan, and S. W. Indratno, "Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, p. 598, Jan. 2022, doi: 10.11591/IJEECS.V29.II.PP598-608.
- [43] A. Almomani et al., "Age and Gender Classification Using Backpropagation and 焔 agging 焔 lgorithms," *Computers, Materials & Continua*, vol. 74, no. 2, pp. 3045–3062, 2023, doi: 10.32604/CMC.2023.030567.
- [44] M. Fan, K. Xiao, L. Sun, S. Zhang, and Y. Xu, "Automated Hyperparameter Optimization of Gradient Boosting Decision Tree Approach for Gold Mineral Prospectivity Mapping in the Xiong'ershan Area," *Minerals*, vol. 12, no. 12, Dec. 2022, doi: 10.3390/MIN12121621.
- [45] S. Priya, N. K. Karthikeyan, and D. Palanikkumar, "Pre Screening of Cervical Cancer Through Gradient Boosting Ensemble Learning Method," *Intelligent Automation and Soft Computing*, vol. 35, no. 3, pp. 2673–2685, 2023, doi: 10.32604/IASC.2023.028599.
- [46] S. Das, M. S. Imtiaz, N. H. Neom, N. Siddique, and H. Wang, "A hybrid approach for Bangla sign language recognition using deep transfer learning model with random forest classifier," *Expert Syst Appl*, vol. 213, Mar. 2023, doi: 10.1016/J.ESWA.2022.118914.
- [47] M. Imane, C. Rahmoune, M. Zair, and D. Benazzouz, "Bearing fault detection under time-varying speed based on empirical wavelet transform, cultural clan-based optimization algorithm, and random forest classifier," *JVC/Journal of Vibration and Control*, Jan. 2021, doi: 10.1177/10775463211047034.