Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 20 No. 11 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v20i11.49139

PAPER

Improving the Accuracy of Oncology Diagnosis: A Machine Learning-Based Approach to Cancer Prediction

Michael Cabanillas-Carbonell[1], Joselyn Zapata-Paulini[2](✉)

[1]Faculty of Engineering, Universidad Privada del Norte, Lima, Peru

[2]Graduate School, Universidad Continental, Lima, Peru

70994337@continental.edu.pe

ABSTRACT

Cancer ranks among the most lethal illnesses worldwide, and predicting its onset can be a crucial factor in enhancing people's quality of life by taking preventive measures to improve treatment and survival. This study conducted comparative research to determine the machine learning model with the highest accuracy for tumor type classification, distinguishing between malignant (cancer) and benign tumors. The models evaluated include decision tree (DT), naive bayes (NB), extra trees classifier (ETM), random forest (RF), K-means clustering (K-means), logistic regression (LR), adaptive boosting (AdaBoost), gradient boosting (GB), light gradient boosting machine (LightGBM), and extreme gradient boosting (XGBoost) to identify the one with the best accuracy. The models were trained using a dataset of 569 records and a total of 32 variables, containing patient information and tumor characteristics. The study was structured into sections, such as related studies, descriptions of the models, case study development, results, discussion, and conclusions. The models' performance was evaluated based on metrics of precision, sensitivity, accuracy, and F1 score. Following the training, the results positioned the XGBoost model as having the best performance, achieving 98% precision, accuracy, sensitivity, and F1 score.

KEYWORDS

machine learning (ML), cancer, prediction, tumor, models

1 INTRODUCTION

Each year, millions of people worldwide are diagnosed with cancer, and slightly more than half of those diagnosed die from the disease [1]. Currently, cancer, along with cardiovascular disease, is the leading cause of death in approximately 127 countries [2]. Lifestyle, environmental factors, and genetic variations are believed to influence the development of cancer, which is present in more than 90% of diagnosed cases [3], [4]. Cancer is generally characterized by the abnormal growth of cells and can occur in any body structure or organ [5]. It is estimated that in 2020, there will be approximately 19 million new cases of cancer and about 10 million

Cabanillas-Carbonell, M., Zapata-Paulini, J. (2024). Improving the Accuracy of Oncology Diagnosis: A Machine Learning-Based Approach to Cancer Prediction. *International Journal of Online and Biomedical Engineering (iJOE)*, 20(11), pp. 102–122. https://doi.org/10.3991/ijoe.v20i11.49139

Article submitted 2024-03-16. Revision uploaded 2024-05-05. Final acceptance 2024-05-08.

deaths [6]. Moreover, a higher incidence of cancer is observed in countries with a high socioeconomic level and advanced stages of development, particularly breast, colon, prostate, and uterine cancer [7]. People under 75 years of age have a 20% risk of developing cancer and a 10% risk of dying from the disease [8]. In some countries, the incidence level is 400 diagnosed cases per 100,000 men and 300 per 100,000 women [9].

In recent years, non-melanoma skin cancer, tracheal, bronchus, and lung cancer, colon and rectal cancer, breast cancer, prostate cancer, stomach cancer, and other malignant neoplasms have reached the highest mortality rates, with an incidence rate of 79.1%, 27.66%, 26.71%, 24.17%, 17.39%, 15.59%, and 10.45%, respectively [10]. These cancers account for about half of the mortality worldwide, with the most deaths occurring in East Asia at 36.2%, followed by South Central Asia at 12%, Eastern Europe at 7.3%, and North America at 7.3% of deaths [11]. The incidence rate was 49.2% in Asia, 22.4% in Europe, 13.4% in North America, 7.8% in South and Central America, 5.9% in Africa, and 1.3% in Oceania [12].

In the United States, some of the most common cancers affecting the male population include prostate cancer at 27%, lung and bronchus cancer at 12%, and colon and rectal cancer at 8%. For women, breast cancer accounts for 31%, lung and bronchus cancer for 13%, and colon and rectal cancer for 8% [13], [14]. Conversely, China exhibits a lower cancer incidence compared to the United Kingdom and the United States. However, the mortality rate in China is notably higher, ranging from 30% to 40%, with over 36% of deaths attributed to liver, stomach, and esophagus cancers [15].

Much study has been done with mathematical models to determine and classify cancer, as in [16], where a novel technique is proposed for the treatment of tumor models with a power-law kernel using the Sumudu transform. Machine learning (ML) models are an important tool as they use large datasets to identify patterns that can predict the development of diseases [17], [18]. ML has been used in various fields of study, such as aiding in the discovery of disease-related genes [19], word analysis and classification [20], [21], and price prediction [21], among others. By using these models, it is possible to estimate the probability that a person will develop a disease in the future [22], [23].

This study aims to compare the accuracy of ML models for tumor type classification, distinguishing between malignant (cancer) and benign tumors. The models evaluated include decision tree (DT), Naive Bayes (NB), extra trees classifier (ETM), random forest (RF), k-means clustering (K-Means), logistic regression (LR), adaptive boosting (AdaBoost), gradient boosting (GB), light gradient boosting machine (LightGBM), and extreme gradient boosting (XGBoost) to determine which of the 10 models provides better accuracy. This paper is structured into six parts. The first part contextualizes the study problem. The second part details related work, while the third part describes the ML models used and analyzes the data before training the models. The training results are presented in part four, followed by discussions in part five. Finally, the conclusions of the study are presented in the sixth part.

2 RELATED WORK

The following are studies that aim to predict the most prevalent cancers globally, such as lung cancer, breast cancer, and colon cancer, among others. In a study by the authors [24], a comparative investigation of five ML models focused on breast cancer prediction was conducted using the Wisconsin dataset. The study identified artificial neural networks (ANN) as the most accurate model with 0.9857 accuracy and 0.9782 precision, followed by support vector machines (SVM) with 0.9714

accuracy and 0.9565 precision. Similarly, a study [25] compared different ML models for breast cancer prediction. The results showed that AdaBoost, GB, and RF achieved 0.1 accuracy, followed by k-nearest neighbor (KNN), bagging, and the multi-layer perceptron (MLP), with accuracies of 0.9956, 0.9582, and 0.9692, respectively. In study [26], a comparative analysis of 13 ML models for breast cancer prediction was developed using the Wisconsin Breast Cancer Original (WBCO) dataset. The study concluded that the MLP model achieved the highest accuracy at 0.9876. On the other hand, a study [27] presented a study of ML models for colon cancer prediction and survival using a dataset from Chang Gung Memorial Hospital, Taiwan, with 4021 records. The study found that the RF model achieved the highest accuracy at 0.84. Similarly, the authors of the study [28] conducted a comparative analysis of multiple ML models focused on colon cancer prediction. They employed feature selection and classification techniques for data processing, with the results positioning RF as the most accurate model with 0.951. Study [29] performed a comparative study of 6 ML models for the prediction and classification of colon and lung cancer, utilizing feature engineering techniques for data processing. The study concluded that the XGBoost model outperformed all others with 0.99 accuracy. In contrast, a study [30] compared different ML models for ovarian cancer prediction, using Pearson's skewness and correlation coefficient to process the dataset. The findings indicated that the RF model achieved an accuracy of 0.8872. In study [31], the authors evaluated different ML models focused on lung cancer prediction, with the results positioning the ANN model as the most accurate with 0.1 accuracy. Similarly, a study [32] examined different ML models for lung cancer prediction, concluding that the ANN model is the most accurate with 0.813 accuracy. The authors of the study [33] conducted an investigation to identify the optimal ML model for lung cancer prediction using a dataset containing hundreds of grayscale images. The training results positioned the Adaboost model as the best, achieving 0.9074 in accuracy, 0.8180 in sensitivity, and 0.9399 in specificity. In the study [34], multiple ML models were analyzed for colorectal cancer prediction, with RF achieving the best metrics with 0.75 in accuracy and 0.76 in sensitivity. Compared various [35] ML models for cervical cancer prediction, with the NB model being identified as the most accurate, achieving 0.9638 in accuracy. Similarly, a study [36] evaluated three ML algorithms for cervical cancer prediction, with all three methods achieving the best metrics at 0.9333 in accuracy. Finally, in the study [37], RF, XGBoost, BN, and convolutional neural networks (CNN) models were analyzed and trained for cervical cancer prediction. The study concluded that the CNN model is the best predictor, achieving 0.1 in accuracy.

3 METHODOLOGY

In this part of the paper, we will develop the case study, which is divided into two sections. In Section 3.1, we describe the ML models (NB, DT, RF, ETM, K-means, logistic regression model, adaptive boosting model, gradient boosting model, LigthGBM, and XGBoost). In Section 3.2, we perform a detailed analysis of the dataset to subsequently train the models.

3.1 Description of the ML models

Naive bayes model. NB is a well-known probabilistic classification algorithm, known for its simplicity and effectiveness in various real-world applications [38]. NB operates under the assumption that all attributes in a dataset are independent of

each other, which streamlines the training and learning process [39]. The model is founded on Bayes' rule or theorem, represented by the following notations: [40], [41]. The model can be mathematically described by equations (1) and (2).

$$P(A/B) = \frac{P(A\,and\,B)}{P(B)} \tag{1}$$

And

$$P(B/A) = \frac{P(A\,and\,B)}{P(A)} \tag{2}$$

Where $P(A)$ represents the likelihood of event A, $P(B)$ denotes the probability of event B, and P (A and B) indicates the joint probability of both events A and B. $P(A\backslash B)$ is the conditional probability of event A assuming that B has occurred.

Decision tree model. DT is a popular tool in supervised learning, as it can be used for classification and prediction [42]. The model is structured by recursively dividing the dataset into smaller subsets until a level of homogeneity is reached [43]. Furthermore, it is hierarchically structured with internal nodes and leaves, where leaves represent decisions or class labels [44]. DT can be applied in various fields, such as asthma prediction and financial risks, among others [45], [46]. Equation (3) presents the mathematical representation of the model.

$$E(s) = \sum_{k=0}^{n} \binom{n}{k} - Py * \log 2Pn \tag{3}$$

Where s is the sample, E represents the entropy, Pn is the probability of NO, and Py is the probability of YES.

Random forest model. RF is one of the most widely used algorithms because it can be applied to both data regression and classification [47]. Typically, the model trains thousands of DT using a random subset of data, aggregates the results of each tree, and generates predictions [48], [49]. Additionally, the model is utilized for feature selection metrics, data classification, and assessing the proximity between data [50]. The model's architecture is illustrated in Figure 1.

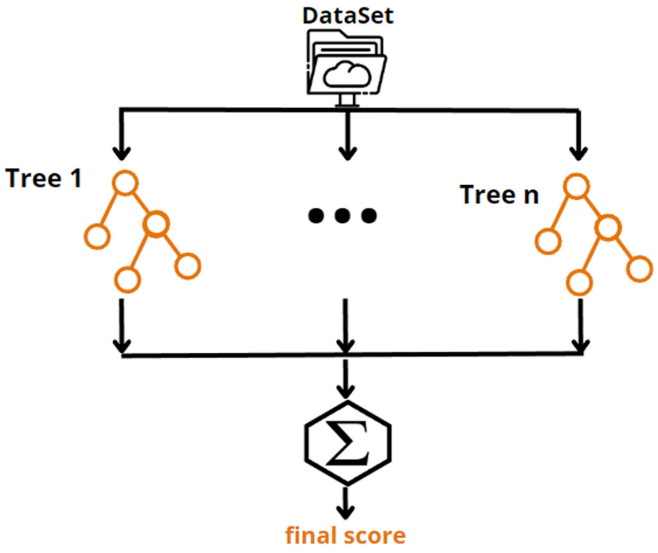

Fig. 1. Architecture of the RF model

Extra trees model. ETM is very similar to the RF model, but it employs different data selection methods. The model is constructed with multiple decision or regression trees that are not pruned to prevent overfitting [51]. ETM uses randomization to split the nodes based on cutoff points and leverages the entire training dataset to build the trees without employing bootstrap replication [52]. The model finds applications in various fields such as linear regression in ML, classifying cardiac signals, and predicting epileptic seizures, among others [53], [54], and [55]. Similar to RF, the model aggregates all trees to average and predict the final outcome using the Breiman equation, as described below [56]. The mathematical representation of the model is shown in equation (4).

$$G(x,\theta_1,\ldots,\theta_r)=\frac{1}{R}\sum_{r=1}^{R}G(x,\theta_r) \tag{4}$$

K-Means model. K-means is one of the most widely used clustering algorithms today, as it is quick to learn and simple to apply [57]. The model utilizes the value of k throughout the clustering process, sequentially assigning each data point to the center of the corresponding cluster and updating at each new assignment until convergence is reached [58]. The algorithm can be used in parallel for data processing acceleration and can be combined with other data segmentation techniques [59], [60]. The model can be expressed as equation (5).

$$\arg\min_{S}\sum_{i=1}^{K}\sum_{x\in S_i}\left\|x-\mu_i\right\|^2 \tag{5}$$

Where K is the number of groups, S is the set of observations, x is the observation point, and μ_i mean of the points in S_i.

Logistic regression model. LR is a statistical model that illustrates the relationship between variables and is used to predict the probability of an event occurring based on independent variables [61]. This model is widely utilized in fields such as finance, marketing, and the social sciences [62]. Moreover, the model utilizes the likelihood function for optimization and subsequent training [63], [64]. The equation of the model is detailed in equation (6). The event *Y occurring* has a probability denoted as $P(Y)$.

$$P(Y)=\frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\cdots+b_nX_n)}} \tag{6}$$

Adaptive boosting model. AdaBoost is an ML algorithm that enhances the accuracy of other classification models by amalgamating multiple weak classification algorithms into a robust one. This technique assigns weights to each data point in the training set to train weak classifiers [65], [66]. The model is widely recognized as one of the most popular, being the first one dedicated to practical application [67]. AdaBoost is extensively used in both studies and industry due to its capability to enhance other classification models [68], [69]. The model can be mathematically represented in equation (7), where T denotes the count of weak models, $F_T(x)$ is the final prediction of x, $f_t(x)$ is the prediction of the weak model, and α_t is the weighting coefficient.

$$F_T(x)=\sum_{t=1}^{T}\alpha_t f_t(x) \tag{7}$$

Gradient boosting model. Similar to AdaBoost, GB focuses on enhancing the accuracy of classification and regression algorithms. The model sequentially

trains multiple weak learners to correct the errors of previous learners, resulting in a more precise model [70]. Due to its capability to enhance the accuracy of other models, AdaBoost is extensively utilized in various fields of study and industries [71], [72]. The model is optimized in function space using gradients, based on Friedman's statistical development [73]. AdaBoost can be represented by the following equation (8), where $f(x)$ is the prediction function, $\hat{y}$ is the final model accuracy, γ is the learning coefficient, and $h(x)$ is the prediction of the i-th weakest model.

$$\hat{y} = f(x) = \sum \gamma * h(x) \tag{8}$$

LigthGBM model. The LigthGBM model is mainly focused on classification and regression. Compared to other models, it is faster and more efficient [74]. The model utilizes the gradient-based one-sided sampling (GOSS) technique to reduce the amount of data used in the training process, enhancing the accuracy and speed of the model [75], [76]. The model architecture is illustrated in Figure 2.

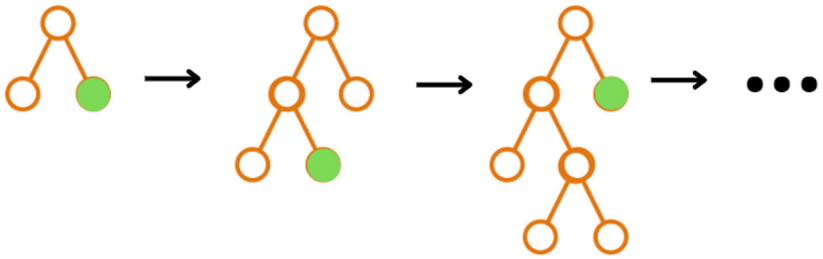

Fig. 2. Architecture of the LigthGBM model

Extreme gradient boosting model. XGBoost is a highly popular classification and regression algorithm. It sequentially incorporates weak learners to enhance the model's accuracy and utilizes regularization techniques to mitigate overfitting issues [77], [78]. Equation (9) outlines the formula used by the model to compute the prediction of each tree. Here, $f(x)$ represents the prediction generated by the i-th decision tree, and y denotes the final model prediction.

$$\hat{y}_i = \sum_{t=1}^{m} f_t(x_i) \tag{9}$$

3.2 Case study

Understanding the dataset. For this study, we used a dataset provided by Kaggle, which included a total of 33 variables. These variables consist of patient id, diagnosis (B = benign, M = malignant), and various tumor characteristics such as mean radius, mean perimeter, mean texture, mean area, mean smoothness, mean concavity, mean concave points, mean symmetry, mean fractal dimension, mean compactness, radius se, perimeter se, texture se, area se, smoothness se, concavity se, compactness se, concave points se, fractal dimension se, symmetry se, radius worse, texture worse, perimeter worse, area worse, smoothness worse, compactness worse, concavity worse, concave points worse, symmetry worse, and fractal dimension worse. The dataset comprises 569 patient records. The process of developing the case study is outlined in Figure 3.

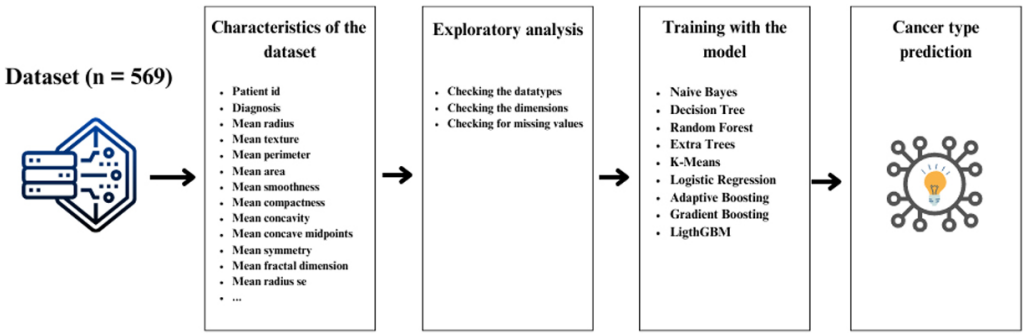

Fig. 3. Case study development process

Preparation of the case study. In this section, we carry out a content analysis of the dataset before proceeding with the analysis and training of the models. Initially, we imported the necessary libraries for data manipulation. During the initial analysis, we observed that the dataset consists of continuous and categorical variables. It is important to note that no null values are present, as indicated in Table 1. Subsequently, we examined the types of data stored in each column of the dataset, as outlined in Table 2. To streamline the training process, we opted to remove the 'Unnamed: 32' column, as it will not be used in the process.

Table 1. Content of the data set

	0	1	2	3	...	565	566	567	568
id	842302	842517	84300903	84348301	...	926682	926954	927241	92751
diagnosis	M	M	M	M	...	M	M	M	B
radius_mean	17.99	20.57	19.69	11.42	...	20.13	16.6	20.6	7.76
texture_mean	10.38	17.77	21.25	20.38	...	28.25	28.08	29.33	24.54
perimeter_mean	122.8	132.9	130	77.58	...	131.2	108.3	140.1	47.92
area_mean	1001	1326	1203	386.1	...	1261	858.1	1265	181
smoothness_mean	0.1184	0.08474	0.1096	0.1425	...	0.0978	0.08455	0.1178	0.05263
compactness_mean	0.2776	0.07864	0.1599	0.2839	...	0.1034	0.1023	0.277	0.04362
concavity_mean	0.3001	0.0869	0.1974	0.2414	...	0.144	0.09251	0.3514	0
concave points_mean	0.1471	0.07017	0.1279	0.1052	...	0.09791	0.05302	0.152	0
...	...	...	...	...	...	...	...	...	...
texture_worst	17.33	23.41	25.53	26.5	...	38.25	34.12	39.42	30.37
perimeter_worst	184.6	158.8	152.5	98.87	...	155	126.7	184.6	59.16
area_worst	2019	1956	1709	567.7	...	1731	1124	1821	268.6
smoothness_worst	0.1622	0.1238	0.1444	0.2098	...	0.1166	0.1139	0.165	0.08996
compactness_worst	0.6656	0.1866	0.4245	0.8663	...	0.1922	0.3094	0.8681	0.06444
concavity_worst	0.7119	0.2416	0.4504	0.6869	...	0.3215	0.3403	0.9387	0
concave points_worst	0.2654	0.186	0.243	0.2575	...	0.1628	0.1418	0.265	0
symmetry_worst	0.4601	0.275	0.3613	0.6638	...	0.2572	0.2218	0.4087	0.2871
fractal_dimension_worst	0.1189	0.08902	0.08758	0.173	...	0.06637	0.0782	0.124	0.07039
Unnamed: 32	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN

Table 2. Summary information of the data set

#	Column	Dtype	Non-Null Count
0	id	int64	569 non-null
1	diagnosis	object	569 non-null
2	radius_mean	float64	569 non-null
3	texture_mean	float64	569 non-null
4	perimeter_mean	float64	569 non-null
5	area_mean	float64	569 non-null
6	smoothness_mean	float64	569 non-null
7	compactness_mean	float64	569 non-null
8	concavity_mean	float64	569 non-null
9	concave points_mean	float64	569 non-null
10	symmetry_mean	float64	569 non-null
11	fractal_dimension_mean	float64	569 non-null
12	radius_se	float64	569 non-null
13	texture_se	float64	569 non-null
14	perimeter_se	float64	569 non-null
15	area_se	float64	569 non-null
16	smoothness_se	float64	569 non-null
17	compactness_se	float64	569 non-null
18	concavity_se	float64	569 non-null
19	concave points_se	float64	569 non-null
20	symmetry_se	float64	569 non-null
21	fractal_dimension_se	float64	569 non-null
22	radius_worst	float64	569 non-null
23	texture_worst	float64	569 non-null
24	perimeter_worst	float64	569 non-null
25	area_worst	float64	569 non-null
26	smoothness_worst	float64	569 non-null
27	compactness_worst	float64	569 non-null
28	concavity_worst	float64	569 non-null
29	concave points_worst	float64	569 non-null
30	symmetry_worst	float64	569 non-null
31	fractal_dimension_worst	float64	569 non-null

Exploratory analysis of the data. In Figure 4, we present an analysis of the target variable, which involves classifying tumors as benign (B) or malignant (M). It is evident that there is an imbalance in the class distribution, with a higher number of benign tumor records compared to malignant tumors. While this imbalance is not substantial, it is a crucial factor to consider in the dataset analysis.

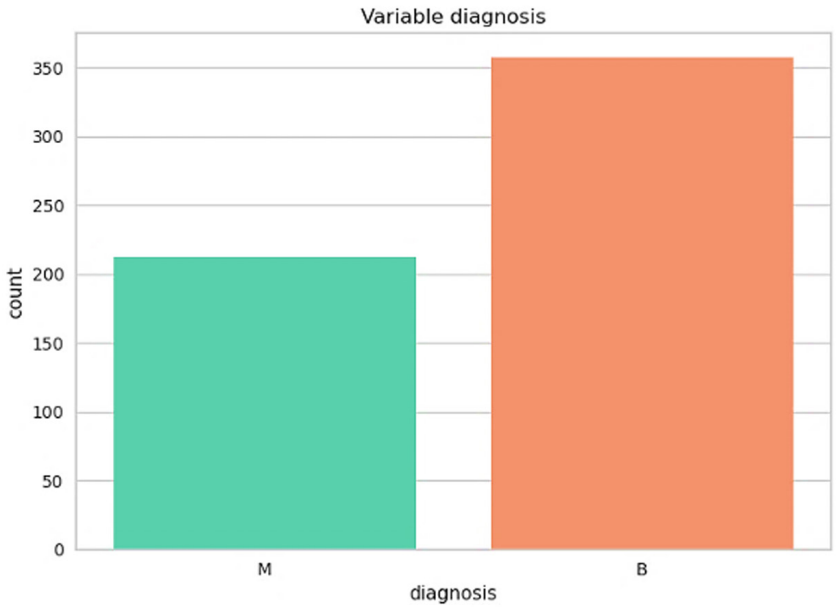

Fig. 4. Analysis of the target variable

In the bivariate analysis presented in Figure 5, the relationships between specific visual characteristics of tumors and the likelihood of developing cancer were examined. The results illustrated in Figure 5a indicate that tumors with a mean radius ranging from 10 mm to 15 mm are more likely to be benign and less likely to develop cancer. Conversely, in Figure 5b, it was discovered that tumors with a mean texture size between 20 mm and 25 mm have an increased probability of developing cancer. Additionally, Figure 5c demonstrates that tumors with a mean perimeter exceeding 70 mm have a reduced risk of being cancerous.

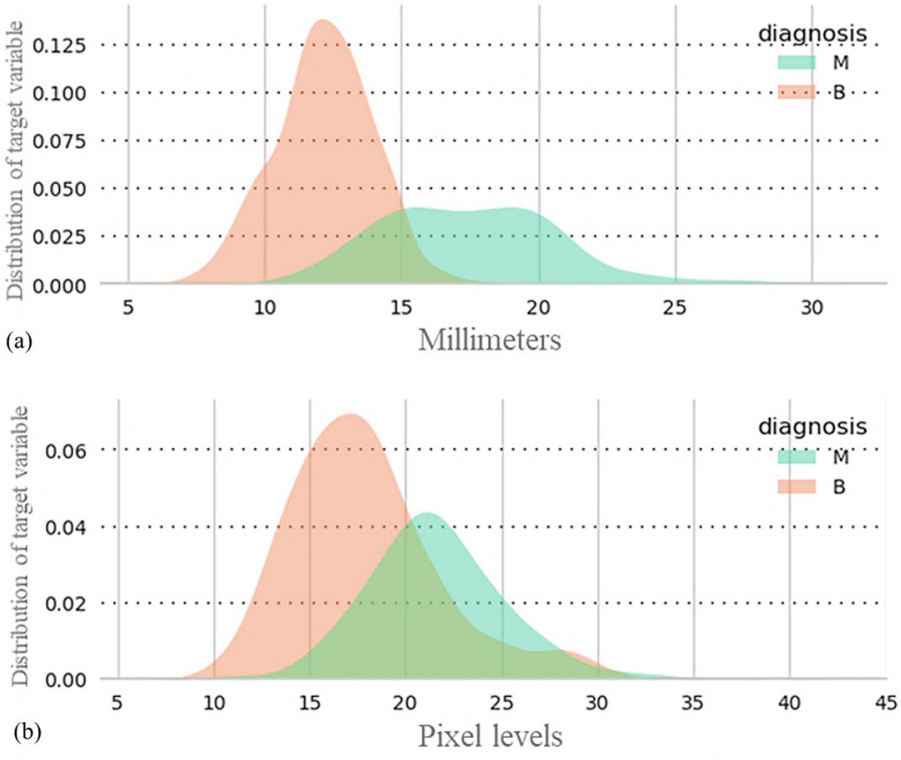

Fig. 5. *(Continued)*

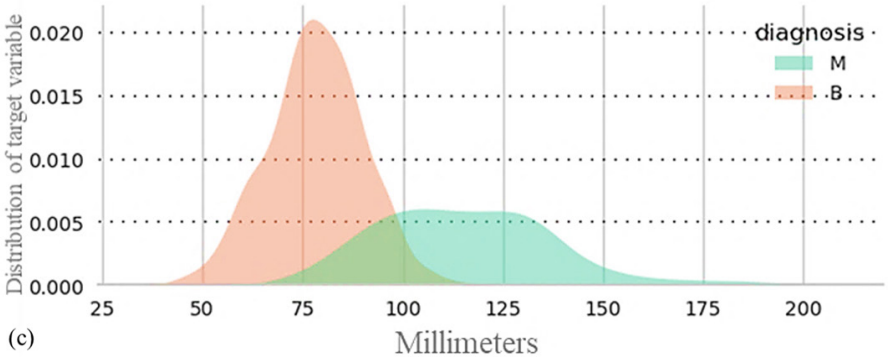

Fig. 5. Analysis of the objective variable with the visual characteristics of the tumor: (a) Objective variable and mean radius of the tumor, (b) Objective variable and average texture of the tumor, (c) Objective variable and average perimeter

Likewise, Figure 6 shows the results of the analysis of the target variable with the metric characteristics of the tumor. In Figure 6a, it is observed that when the smoothness of the tumor exceeds 0.005, the probability of the tumor being malignant or benign is nearly equal. Similarly, in Figure 6b, it is evident that a tumor compactness of 0.012 correlates with a lower probability of cancer development. Additionally, in Figure 6c, it is noted that a range of concave points from 0.05 mm to 0.10 mm is associated with a higher probability of developing a benign tumor, as opposed to a range from 0.15 mm to 0.10 mm, where it could be cancerous.

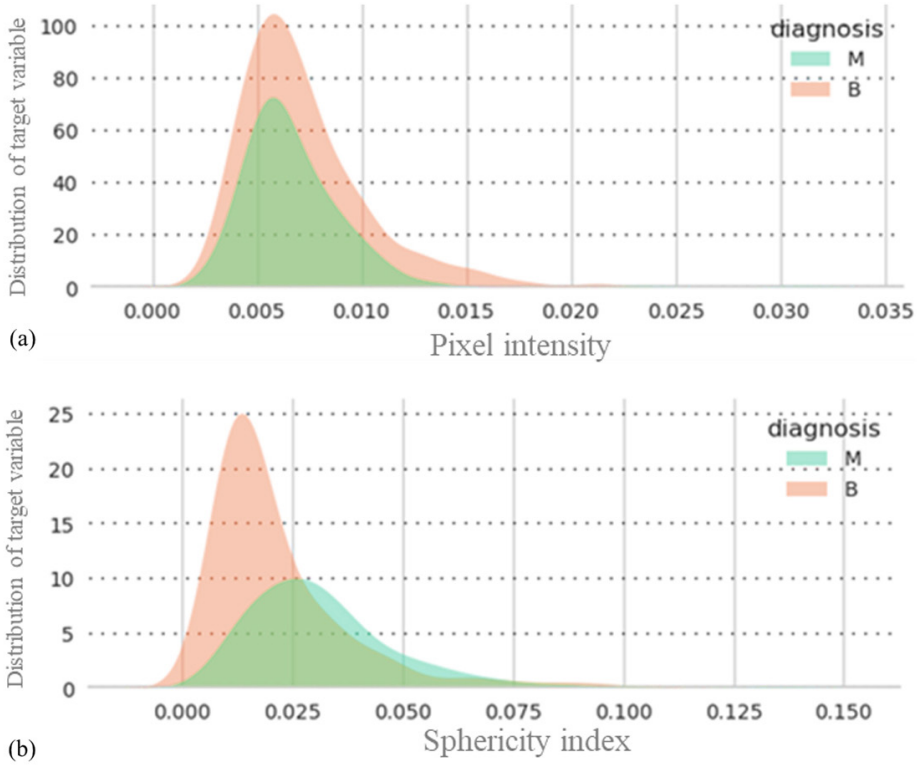

Fig. 6. *(Continued)*

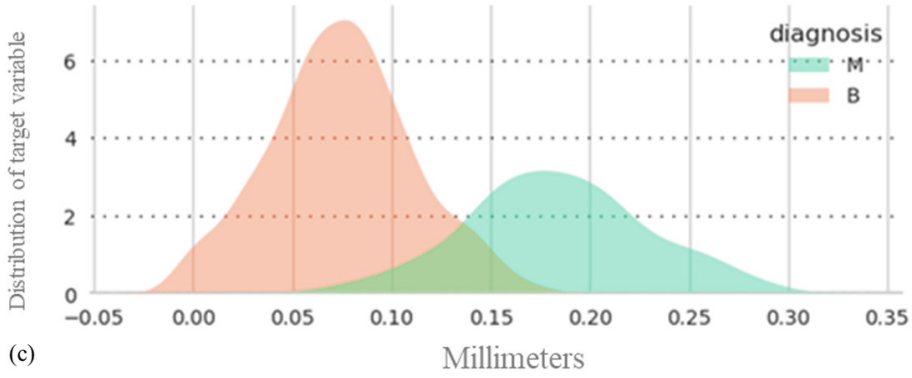

Fig. 6. Analysis of the target variable with the metric characteristics of the tumor: (a) Target variable and tumor smoothness, (b) Target variable and tumor compactness, (c) Target variable and tumor concave points

In Figure 7, the distribution of the data as a function of the target variable is presented to analyze the probabilities associated with the development of cancer according to additional tumor characteristics. According to Figure 7a, if the measure of tumor concavity (concavity worst) is pronounced, the odds of the tumor being malignant also increase considerably. Similarly, in Figure 7b, it is shown that a high number of concave points (concave points worst) is related to a higher probability of developing a cancerous tumor. Tumor smoothness (smoothness worst), depicted in Figure 7c, is also an important factor, as higher smoothness is associated with a higher probability of cancer. Additionally, an increase in tumor symmetry (the worst symmetry) also increases the probability of malignancy, according to Figure 7d. On the other hand, the fractal dimension (fractal dimension worst) does not seem to be a relevant indicator, as the probability of the tumor being benign or malignant is almost the same, as shown in Figure 7e. In contrast, tumor compactness is shown to be an important factor, as its increase correlates with a higher probability of cancer, as shown in Figure 7f.

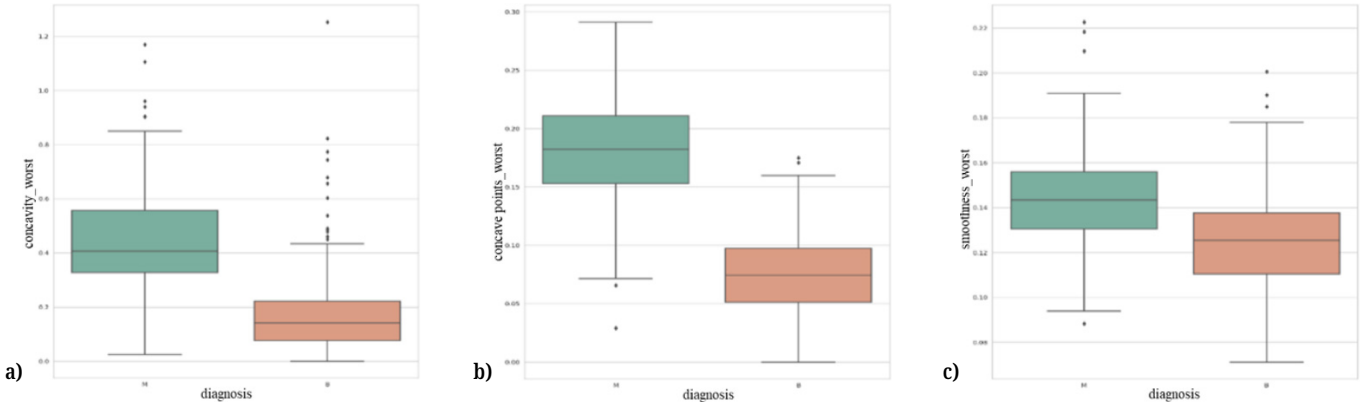

Fig. 7. *(Continued)*

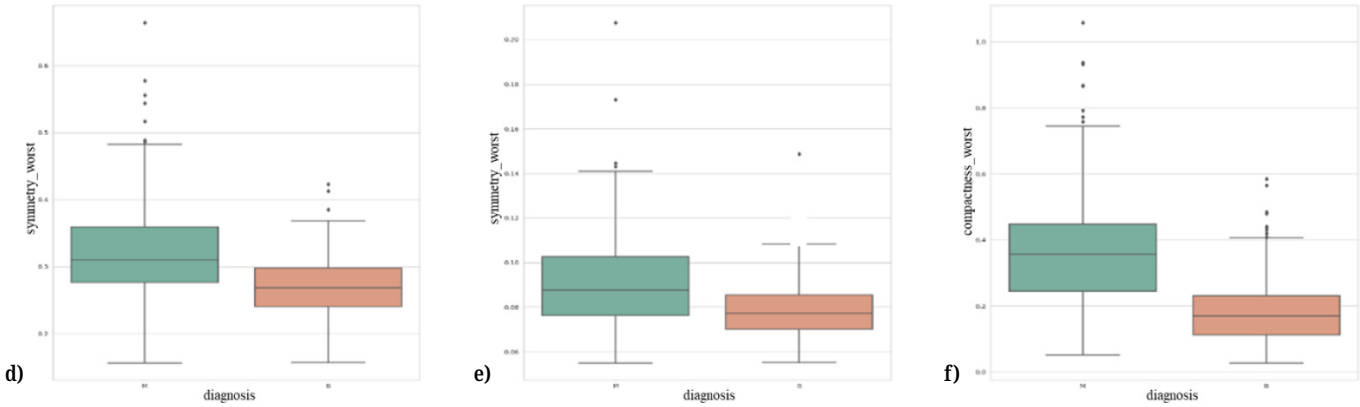

Fig. 7. Box plot of the target variable with additional variables: (a) Target variable and tumor concavity, (b) Target variable and tumor concavity points, (c) target variable and tumor smoothness, (d) Target variable and tumor symmetry, (e) target variable and fractal dimension, (f) Target variable and tumor capacity

Data processing and modeling. Before training the models, the LabelEncoder class from the scikit-learn library was used to convert the categorical variables into continuous variables. Next, the target variable (diagnosis) was separated from the other variables. Subsequently, the StandardScaler class was employed to standardize all the data. Finally, the dataset was divided into two groups, with 30% allocated "0" of the data to the test group and the remaining 70% allocated "1" to the training group.

4 RESULTS

In this study, NB, DT, RF, ETM, K-Means, LR, AdaBoost, GB, LightGBM, and XGBoost models were focused on tumor type prediction, distinguishing between malignant (cancer) and benign tumors. These models were analyzed and trained. The dataset, consisting of 32 variables and 569 patient records, was extracted from the Kaggle platform. This dataset was analyzed and processed to subsequently train ML models. The results of these trainings are shown in Table 3.

Table 3. Model training results

Naive Bayes			
	F1-Score (%)	Recall (%)	Precision (%)
0	0.93	0.92	0.94
1	0.88	0.9	0.86
macro avg	0.91	0.91	0.9
weighted avg	0.91	0.91	0.91
accuracy	0.91		
Decision Tree			
	F1-Score (%)	Recall (%)	Precision (%)
0	0.93	0.91	0.96
1	0.89	0.94	0.86
macro avg	0.91	0.92	0.91
weighted avg	0.92	0.92	0.92
accuracy	0.92		

(Continued)

Table 3. Model training results *(Continued)*

Random Forest			
	F1-Score (%)	Recall (%)	Precision (%)
0	0.98	0.97	0.98
1	0.96	0.97	0.95
macro avg	0.97	0.97	0.97
weighted avg	0.97	0.97	0.97
accuracy	0.97		
Extra Trees			
	F1-Score (%)	Recall (%)	Precision (%)
0	0.96	0.99	0.93
1	0.92	0.87	0.98
macro avg	0.94	0.93	0.96
weighted avg	0.95	0.95	0.95
accuracy	0.95		
K-Means			
	F1-Score (%)	Recall (%)	Precision (%)
0	0.95	0.97	0.92
1	0.9	0.86	0.95
macro avg	0.92	0.91	0.93
weighted avg	0.93	0.93	0.93
accuracy	0.93		
Logistic Regression			
	F1-Score (%)	Recall (%)	Precision (%)
0	0.98	0.98	0.97
1	0.96	0.95	0.97
macro avg	0.97	0.97	0.97
weighted avg	0.97	0.97	0.97
accuracy	0.97		
AdaBoost			
	F1-Score (%)	Recall (%)	Precision (%)
0	0.98	0.97	0.98
1	0.96	0.97	0.95
macro avg	0.97	0.97	0.97
weighted avg	0.97	0.97	0.97
accuracy	0.97		

(Continued)

Table 3. Model training results *(Continued)*

Gradient Boosting			
	F1-Score (%)	Recall (%)	Precision (%)
0	0.97	0.97	0.96
1	0.94	0.94	0.95
macro avg	0.96	0.95	0.96
weighted avg	0.96	0.96	0.96
accuracy	0.96		
LightGBM			
	F1-Score (%)	Recall (%)	Precision (%)
0	0.97	0.97	0.96
1	0.94	0.94	0.95
macro avg	0.96	0.95	0.96
weighted avg	0.96	0.95	0.96
accuracy	0.96		
XGB Boost			
	F1-Score (%)	Recall (%)	Precision (%)
0	0.98	0.98	0.98
1	0.97	0.97	0.97
macro avg	0.97	0.97	0.97
weighted avg	0.98	0.98	0.98
accuracy	0.98		

In training the NB, DT, RF, ETM, and AdaBoost models, we used entropy and Gini calculations, along with GridSearch, to determine which metric is more effective for optimizing the models. The results indicate that the NB, DT, RF, ETM, K-Means, LR, AdaBoost, GB, LightGBM, and XGBoost models managed to achieve an accuracy of 91%, 92%, 97%, 95%, 93%, 97%, 97%, 96%, 96%, and 98%, respectively.

All 10 models achieved exceptional metrics, reaching accuracies above 90%. The model that stands out the most is the XGBoost, which achieved the best performance with 98% precision, 98% sensitivity, a 98% F1 score, and 98% accuracy. This was followed by the RF, LR, and AdaBoost models, which achieved 97% precision, 97% sensitivity, a 97% F1 score, and 97% accuracy. In third place, we have the GB and LightGBM models with 96% accuracy and 96% in F1 count, except for sensitivity, where GB obtained 96% and LightGBM 95%. In fourth place, the ETM model achieved 95% accuracy, 95% sensitivity, and a 95% F1 score. Finally, the K-means, DT, and NB models achieved 93%, 92%, and 91% accuracy, respectively.

5 DISCUSSION

Cancer ranks among the most lethal illnesses worldwide; every year, thousands of people die from this disease. Predicting its development can be a crucial factor in improving the quality of life for people and taking preventive actions to enhance treatment and survival rates. This study conducted a comparative investigation to

determine the model with the best accuracy for classifying tumor types in future individuals, distinguishing between malignant (cancer) and benign tumors. The models were trained with a dataset of 569 records and a total of 32 variables, containing patient information and tumor characteristics. After applying data processing and training, the models achieved accuracy levels above 90%. The XGBoost model achieved the best metrics with 98% accuracy, sensitivity, and F1-count. Similar to a study [29], where XGBoost was identified as the best predictor of lung and colon cancer with 99% accuracy and a 98% F1 count, this analysis used histopathological images for predicting 5 types of lung and colon cancer tissues, unlike this study, which did not utilize images for cancer prediction. Additionally, the RF model achieved one of the best metrics in prediction with 97%, similar to the results obtained in the other studies [25], [28], where the RF model achieved 100% and 95.16%, respectively, for predicting breast and colon cancer. The difference with the first study lies in the optimization techniques used to achieve 100% accuracy. On the other hand, studies [27], [30], and [34] for the prediction of colon, ovarian, and colorectal cancer achieved lower metrics than this study, with the RF model achieving 84%, 88.72%, and 75%, respectively. Regarding the AdaBoost model, in this study, the model achieved 97% in all its performance metrics, somewhat similar to the results obtained in [25], [33], and [79] for the prediction of lung cancer, breast cancer, and the classification of autism spectrum disorder, where the model achieved 100%, 90.74%, and 99.8%, respectively. Finally, the NB model obtained 91% accuracy, being the model with the lowest performance, which is lower than the accuracy obtained by [35], where NB achieved an accuracy of 96.38% for cervical cancer prediction. In conclusion, the results of the models are very similar to those obtained in other studies; the main difference lies in the use of different datasets and optimization techniques. For all these reasons, ML models can be a crucial tool in improving the treatment or life prognosis of patients by predicting the formation of cancerous tumors years in advance. However, these models are severely limited by the quality of the dataset used to achieve ideal accuracy.

6 CONCLUSIONS

The use of ML models is becoming increasingly common in the medical field for predicting diseases such as cancer. However, one of the main challenges we face is the quality of the datasets used to train these models. In this study, we compared the accuracy of ML models for tumor type classification, distinguishing between malignant (cancer) and benign tumors. A dataset of 569 records and 32 variables provided by Kaggle was used. After contrasting the training results, it was concluded that the XGBoost model delivered outstanding results, achieving a remarkable 98% accuracy, sensitivity, and F1 score. The other models also achieved exceptional results, with accuracies exceeding 90%.

Additionally, the visual and metric characteristics of tumors are important factors in determining whether they are malignant or not. This can help enhance the diagnosis of oncology patients, thereby improving their quality of life and prognosis in the future.

Finally, ML models for cancer detection are rapidly developing and improving. Therefore, it would be beneficial to examine various types of data, including genomics, proteomics, and medical imaging data, and to study different types of cancer. Additionally, it is important to train the models with diverse datasets to assess the accuracy of the training.

7 REFERENCES

[1] World Health Organization, "Cancer," 2022. https://www.who.int/news-room/fact-sheets/detail/cancer

[2] F. Bray, M. Laversanne, E. Weiderpass, and I. Soerjomataram, "The ever-increasing importance of cancer as a leading cause of premature death worldwide," *American Cancer Society*, vol. 127, no. 16, pp. 3029–3030, 2021. https://doi.org/10.1002/cncr.33587

[3] N. Waespe *et al.*, "Cohort-based association study of germline genetic variants with acute and chronic health complications of childhood cancer and its treatment: Genetic Risks for Childhood Cancer Complications Switzerland (GECCOS) study protocol," *BMJ Open*, vol. 12, no. 1, 2022. https://doi.org/10.1136/bmjopen-2021-052131

[4] A. W. Kurian *et al.*, "Gaps in incorporating germline genetic testing into treatment decision-making for early-stage breast cancer," *Journal of Clinical Oncology*, vol. 35, no. 20, pp. 2232–2239, 2017. https://doi.org/10.1200/JCO.2016.71.6480

[5] P. S. Roy and B. J. Saikia, "Cancer and cure: A critical analysis," *Indian Journal of CANCER*, vol. 53, no. 3, pp. 441–442, 2016. https://doi.org/10.4103/0019-509X.200658

[6] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021. https://doi.org/10.3322/caac.21660

[7] N. L. Renna Junior and G. de Azevedo e Silva, "Socioeconomic status and cancer survival in Brazil: Analysis of population data from the municipalities of Aracaju and Curitiba, 1996–2012," *Cancer Epidemiol,* vol. 85, p. 102394, 2023. https://doi.org/10.1016/j.canep.2023.102394

[8] J. Ferlay *et al.*, "Cancer statistics for the year 2020: An overview," *International Journal of Cancer*, vol. 149, no. 4, pp. 778–789, 2021. https://doi.org/10.1002/ijc.33588

[9] L. A. Torre, R. L. Siegel, E. M. Ward, and A. Jemal, "Global cancer incidence and mortality rates and trends—an update," *Cancer Epidemiol Biomarkers Prev.*, vol. 25, no. 1, pp. 16–27, 2016. https://doi.org/10.1158/1055-9965.EPI-15-0578

[10] L. Lin, Z. Li, L. Yan, Y. Liu, H. Yang, and H. Li, "Global, regional, and national cancer incidence and death for 29 cancer groups in 2019 and trends analysis of the global cancer burden, 1990–2019," *J. Hematol. Oncol.*, vol. 14, no. 1, 2021. https://doi.org/10.1186/s13045-021-01213-z

[11] J. Ferlay *et al.*, "Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods," *Internation. Journal of Cancer*, vol. 144, no. 8, pp. 1941–1953, 2019. https://doi.org/10.1002/ijc.31937

[12] International Agency for Research on Cancer of WHO, "Absolute numbers, incidence, both sexes, in 2022: All cancers," 2024. https://gco.iarc.fr/today/en/dataviz/pie?mode=population&group_populations=0

[13] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *CA Cancer Journal for Clinicians*, vol. 72, no. 1, pp. 7–33, 2022. https://doi.org/10.3322/caac.21708

[14] K. D. Miller *et al.*, "Cancer treatment and survivorship statistics, 2022," *CA Cancer Journal for Clinicians*, vol. 72, no. 5, pp. 409–436, 2022. https://doi.org/10.3322/caac.21731

[15] R. M. Feng, Y. N. Zong, S. M. Cao, and R. H. Xu, "Current cancer situation in China: Good or bad news from the 2018 Global Cancer Statistics?" *Cancer Communications*, vol. 39, no. 1, pp. 1–12, 2019. https://doi.org/10.1186/s40880-019-0368-6

[16] K. S. Nisar, M. Farman, A. Zehra, and E. Hincal, "Numerical and analytical study of fractional order tumor model through modeling with treatment of chemotherapy," *International Journal of Modelling and Simulation*, pp. 1–14, 2024. https://doi.org/10.1080/02286203.2024.2327659

[17] A. N. Ramesh, C. Kambhampati, J. R. T. Monson, and P. J. Drew, "Artificial intelligence in medicine.," *Ann. R. Coll. Surg. Engl.*, vol. 86, no. 5, p. 334, 2004. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1964229/

[18] Q. Bi, K. E. Goodman, J. Kaminsky, and J. Lessler, "What is machine learning? A primer for the epidemiologist," *American Journal of Epidemiology*, vol. 188, no. 12, pp. 2222–2239, 2019. https://doi.org/10.1093/aje/kwz189

[19] D. H. Le, "Machine learning-based approaches for disease gene prediction," *Briefings in Functional Genomics*, vol. 19, nos. 5–6, pp. 350–363, 2020. https://doi.org/10.1093/bfgp/elaa013

[20] O. Iparraguirre-Villanueva *et al.*, "Classification of tweets related to natural disasters using machine learning algorithms," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 14, pp. 144–162, 2023. https://doi.org/10.3991/ijim.v17i14.39907

[21] R. Jáuregui-Velarde, L. Andrade-Arenas, D. H. Celis, R. C. Dávila-Morán, and M. Cabanillas-Carbonell, "Web application with machine learning for house price prediction," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 23, pp. 85–104, 2023. https://doi.org/10.3991/ijim.v17i23.38073

[22] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, 2019. https://doi.org/10.1186/s12911-019-1004-8

[23] A. Banerjee *et al.*, "Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: Systematic review of validity and clinical utility," *BMC Med.*, vol. 19, no. 1, 2021. https://doi.org/10.1186/s12916-021-01940-7

[24] M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir, "Breast cancer prediction: A comparative study using machine learning techniques," *SN Comput. Sci.*, vol. 1, no. 5, 2020. https://doi.org/10.1007/s42979-020-00305-w

[25] O. Iparraguirre Villanueva, A. Epifanía Huerta, C. Torres Ceclén, J. Ruiz Alvarado, and M. Cabanillas Carbonell, "Breast cancer prediction using machine learning models," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 2, 2023. https://doi.org/10.14569/IJACSA.2023.0140272

[26] C. Saini, K. D. Mahato, C. Azad, and U. Kumar, "Breast cancer prediction using different machine learning algorithms: A comparative study," in *2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1)*, Banglore, India, 2023, pp. 1–6. https://doi.org/10.1109/ICAIA57370.2023.10169729

[27] P. Gupta *et al.*, "Prediction of colon cancer stages and survival period with machine learning approach," *Cancers 2019,* vol. 11, no. 12, p. 2007, 2019. https://doi.org/10.3390/cancers11122007

[28] A. S. M. Shafi, M. M. I. Molla, J. J. Jui, and M. M. Rahman, "Detection of colon cancer based on microarray dataset using machine learning as a feature selection and classification techniques," *SN Appl. Sci.*, vol. 2, no. 7, 2020. https://doi.org/10.1007/s42452-020-3051-2

[29] A. Hage Chehade, N. Abdallah, J. M. Marion, M. Oueidat, and P. Chauvet, "Lung and colon cancer classification using medical imaging: A feature engineering approach," *Phys. Eng. Sci. Med.*, vol. 45, pp. 729–746, 2022. https://doi.org/10.1007/s13246-022-01139-x

[30] A. S. Azar *et al.*, "Application of machine learning techniques for predicting survival in ovarian cancer," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, 2022. https://doi.org/10.1186/s12911-022-02087-y

[31] N. Banerjee and S. Das, "Prediction lung cancer- in machine learning perspective," in *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA 2020)*, 2022, pp. 1–5. https://doi.org/10.1109/ICCSEA49143.2020.9132913

[32] R. Patra, "Prediction of lung cancer using machine learning classifier," in *Communications in Computer and Information Science*, vol. 1235, pp. 132–142, 2020. https://doi.org/10.1007/978-981-15-6648-6_11

[33] K. Ingle, U. Chaskar, and S. Rathod, "Lung cancer types prediction using machine learning approach," in *International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2021, pp. 1–6. https://doi.org/10.1109/CONECCT52877.2021.9622568

[34] L. Zheng, E. Eniola, and J. Wang, "Machine learning for colorectal cancer risk prediction," in *2021 International Conference on Cyber-Physical Social Intelligence, (ICCSI),* 2021, pp. 1–6. https://doi.org/10.1109/ICCSI53130.2021.9736248

[35] S. K. Suman and N. Hooda, "Predicting risk of cervical cancer: A case study of machine learning," *Journal of Statistics and Management Systems*, vol. 22, no. 4, pp. 689–696, 2019. https://doi.org/10.1080/09720510.2019.1611227

[36] L. Akter, Ferdib-Al-Islam, M. M. Islam, M. S. Al-Rakhami, and M. R. Haque, "Prediction of cervical cancer from behavior risk using machine learning techniques," *SN Comput. Sci.*, vol. 2, 2021. https://doi.org/10.1007/s42979-021-00551-6

[37] M. Kruczkowski, A. Drabik-Kruczkowska, A. Marciniak, M. Tarczewska, M. Kosowska, and M. Szczerska, "Predictions of cervical cancer identification by photonic method combined with machine learning," *Sci. Rep.,* vol. 12, no. 1, 2022. https://doi.org/10.1038/s41598-022-07723-1

[38] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: Applications, variations and vulnerabilities: A review of literature with code snippets for implementation," *Soft Comput*, vol. 25, pp. 2277–2293, 2021. https://doi.org/10.1007/s00500-020-05297-6

[39] R. Mosquera, O. D. Castrillón, and L. Parra, "Support vector machines, Naïve Bayes classifier and genetic algorithms for the prediction of psychosocial risks in teachers of colombian public schools," *Inf. Tecnol.*, vol. 29, no. 6, 2018. https://doi.org/10.4067/S0718-07642018000600153

[40] C. Bielza and P. Larrañaga, "Discrete Bayesian network classifiers," *ACM Computing Surveys*, vol. 47, no. 1, pp. 1–43, 2014. https://doi.org/10.1145/2576868

[41] O. Takawira and J. W. M. Mwamba, "Determinants of sovereign credit ratings: An application of the Naïve Bayes classifier," *Eurasian Journal of Economics and Finance*, vol. 8, no. 4, pp. 279–299, 2020. https://doi.org/10.15604/ejef.2020.08.04.008

[42] S. B. Kotsiantis, "Decision trees: A recent overview," *Artif. Intell. Rev.*, vol. 39, no. 4, pp. 261–283, 2013. https://doi.org/10.1007/s10462-011-9272-4

[43] V. G. Costa and C. E. Pedreira, "Recent advances in decision trees: An updated survey," *Artif. Intell. Rev.*, vol. 56, pp. 4765–4800, 2022. https://doi.org/10.1007/s10462-022-10275-5

[44] J. Zapata-Paulini and M. Cabanillas-Carbonell, "Evaluation of machine learning algorithms in the early detection of Parkinson's disease: A comparative study," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 35, no. 1, pp. 222–237, 2024. http://doi.org/10.11591/ijeecs.v35.i1.pp222-237

[45] X. Wei, "A method of enterprise financial risk analysis and early warning based on decision tree model," *Security and Communication Networks*, vol. 2021, no. 1, pp. 1–9, 2021. https://doi.org/10.1155/2021/6950711

[46] F. Ton, O. Jiang, and V. Chang, "Development of an accurate operational definition for asthma using decision tree model," *Respirology*, vol. 24, no. S2, pp. 166–167, 2019. https://doi.org/10.1111/resp.13700_216

[47] S. Han, B. D. Williamson, and Y. Fong, "Improving random forest predictions in small datasets from two-phase sampling designs," *BMC Med. Inform. Decis. Mak.*, vol. 21, 2021. https://doi.org/10.1186/s12911-021-01688-3

[48] S. Abdullah and G. V. Prasetyo, "Easy ensemmble with random forest to handle imbalanced data in classification," *Journal of Fundamental Mathematics and Applications (JFMA)*, vol. 3, no. 1, pp. 39–46, 2020. https://doi.org/10.14710/jfma.v3i1.7415

[49] X. Yang, "Prediction of credit risk based on logistic regression and random forest technique," in *ACM International Conference Proceeding Series*, 2022, pp. 531–535. https://doi.org/10.1145/3558819.3565138

[50] P. A. A. Resende and A. C. Drummond, "A survey of random forest based methods for intrusion detection systems," *ACM Computing Surveys (CSUR)*, vol. 51, no. 3, pp. 1–36, 2018. https://doi.org/10.1145/3178582

[51] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, pp. 3–42, 2006. https://doi.org/10.1007/s10994-006-6226-1

[52] L. Peng, R. Yuan, L. Shen, P. Gao, and L. Zhou, "LPI-EnEDT: An ensemble framework with extra tree and decision tree classifiers for imbalanced lncRNA-protein interaction data classification," *BioData Mining*, vol. 14, no. 1, 2021. https://doi.org/10.1186/s13040-021-00277-4

[53] S. M. Mastelini, F. K. Nakano, C. Vens, and A. C. P. de L. F. de Carvalho, "Online extra trees regressor," *IEEE Transaction on Neural Networks and Learning Systems*, vol. 34, no. 10, pp. 6755–6767, 2022. https://doi.org/10.1109/TNNLS.2022.3212859

[54] B. Dhananjay, N. P. Venkatesh, A. Bhardwaj, and J. Sivaraman, "Cardiac signals classification based on Extra Trees model," in *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, 2021, pp. 402–406. https://doi.org/10.1109/SPIN52536.2021.9565992

[55] M. Ntahobari, L. Kuhlmann, M. Boley, and Z. R. Hesabi, "Enhanced Extra Trees classifier for epileptic seizure prediction," in *2022 5th International Conference on Signal Processing and Information Security (ICSPIS)*, 2023, pp. 175–179. https://doi.org/10.1109/ICSPIS57063.2022.10002677

[56] M. M. Hameed, M. K. Alomar, F. Khaleel, and N. Al-Ansari, "An extra tree regression model for discharge coefficient prediction: Novel, practical applications in the hydraulic sector and future research directions," *Mathematical Problem Engineering*, vol. 2021, no. 1, 2021. https://doi.org/10.1155/2021/7001710

[57] Y. P. Raykov, A. Boukouvalas, F. Baig, and M. A. Little, "What to do when k-means clustering fails: A simple yet principled alternative algorithm," *PLoS One*, vol. 11, no. 9, pp. 1–28, 2016. https://doi.org/10.1371/journal.pone.0162259

[58] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, p. 1295, 2020. https://doi.org/10.3390/electronics9081295

[59] N. S. Sagheer and S. A. Yousif, "A parallel clustering analysis based on Hadoop multinode and Apache Mahout," *Iraqi Journal of Science*, vol. 62, no. 7, pp. 2431–2444, 2021. https://doi.org/10.24996/ijs.2021.62.7.32

[60] R. W. Sembiring Brahmana, F. A. Mohammed, and K. Chairuang, "Customer segmentation based on RFM model using k-means, k-medoids, and DBSCAN methods," *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, vol. 11, no. 1, pp. 32–43, 2020. https://doi.org/10.24843/LKJITI.2020.v11.i01.p04

[61] N. R. Panda, "A review on logistic regression in medical research," *National Journal of Community Medicine*, vol. 13, no. 4, pp. 265–270, 2022. https://doi.org/10.55489/njcm.134202222

[62] C. Wallisch *et al.*, "Review of guidance papers on regression modeling in statistical series of medical journals," *PLoS One*, vol. 17, no. 1, pp. 1–20, 2022. https://doi.org/10.1371/journal.pone.0262918

[63] O. Iparraguirre Villanueva *et al.*, "Comparison of predictive machine learning models to predict the level of adaptability of students in online education," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 4, 2023. https://doi.org/10.14569/IJACSA.2023.0140455

[64] M. Zivkovic *et al.*, "Training logistic regression model by hybridized multi-verse optimizer for spam email classification," in *Lecture Notes in Networks and Systems*, vol. 552, 2023, pp. 507–520. https://doi.org/10.1007/978-981-19-6634-7_35

[65] P. Sherubha, L. J. Ahmed, K. S. Kannan, and S. P. Sasirekha, "Adaptive boosting model for breast cancer prediction," *Journal of Intelligent and Fuzzy Systems*, vol. 45, no. 2, pp. 3417–3431, 2023. https://doi.org/10.3233/JIFS-230086

[66] S. Dalal *et al.*, "Machine learning-based forecasting of potability of drinking water through adaptive boosting model," *Open Chem*, vol. 20, no. 1, pp. 816–828, 2022. https://doi.org/10.1515/chem-2022-0187

[67] J. Tang, A. Henderson, and P. Gardner, "Exploring AdaBoost and random forests machine learning approaches for infrared pathology on unbalanced data sets," *Analyst*, vol. 146, no. 19, pp. 5880–5891, 2021. https://doi.org/10.1039/D0AN02155E

[68] Y. Wang and L. Feng, "An adaptive boosting algorithm based on weighted feature selection and category classification confidence," *Appl. Intell.*, vol. 51, pp. 6837–6858, 2021. https://doi.org/10.1007/s10489-020-02184-3

[69] Z. Zheng and Y. Yang, "Adaptive boosting for domain adaptation: Towards robust predictions in scene segmentation," *IEEE Transactions on Image Processing*, vol. 31, pp. 5371–5382, 2021. https://doi.org/10.1109/TIP.2022.3195642

[70] S. A. Fayaz, S. Kaul, M. Zaman, and M. A. Butt, "An adaptive gradient boosting model for the prediction of rainfall using ID3 as a base estimator," *Revue d' Intelligence Artificielle*, vol. 36, no. 2, pp. 241–250, 2022. https://doi.org/10.18280/ria.360208

[71] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif. Intell. Rev.*, vol. 54, pp. 1937–1967, 2021. https://doi.org/10.1007/s10462-020-09896-5

[72] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front. Neurorobot.*, vol. 7, 2013. https://doi.org/10.3389/fnbot.2013.00021

[73] A. Mayr, H. Binder, O. Gefeller, and M. Schmid, "The evolution of boosting algorithms: From machine learning to statistical modelling," *Methods Inf. Med.*, vol. 53, no. 6, pp. 419–427, 2014. https://doi.org/10.3414/ME13-01-0122

[74] S. Park, S. Jung, S. Jung, S. Rho, and E. Hwang, "Sliding window-based LightGBM model for electric load forecasting using anomaly repair," *J. Supercomputing*, vol. 77, pp. 12857–12878, 2021. https://doi.org/10.1007/s11227-021-03787-4

[75] M. Gan, S. Pan, Y. Chen, C. Cheng, H. Pan, and X. Zhu, "Application of the machine learning LightGBM model to the prediction of the water levels of the lower Columbia River," *J. Mar. Sci. Eng*, vol. 9, no. 5, p. 496, 2021. https://doi.org/10.3390/jmse9050496

[76] B. Li *et al.*, "GNSS/INS integration based on machine learning LightGBM model for vehicle navigation," *Appl. Sci.*, vol. 12, no. 11, p. 5565, 2022. https://doi.org/10.3390/app12115565

[77] Y. Zheng, "A default prediction method using XGBoost and LightGBM," in *2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, 2022, pp. 210–213. https://doi.org/10.1109/ICICML57342.2022.10009823

[78] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785

[79] R. Sujatha, S. L. Aarthy, J. M. Chatterjee, A. Alaboudi, and N. Z. Jhanjhi, "A machine learning way to classify autism spectrum disorder," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 6, pp. 182–200, 2021. https://doi.org/10.3991/ijet.v16i06.19559

8 AUTHORS

Michael Cabanillas-Carbonell is an engineer with a Masters in Systems Engineering, pursuing a PhD in Systems Engineering and Telecommunications at the Polytechnic University of Madrid. Conference Chair of the Engineering International Research Conference IEEE Peru EIRCON. Research professor and international lecturer specializing in software development, artificial intelligence, machine learning, business intelligence, and augmented reality. He has authored more than 100 scientific articles indexed in IEEE Xplore, Scopus, and WoS (E-mail: mcabanillas@ieee.org).

Joselyn Zapata-Paulini is a Systems Engineering and Computer Science graduate from the Universidad de Ciencias y Humanidades, a Masters in Science with environmental management and sustainable development at the Universidad Continental, Peru. She has several international publications. She is specialized in augmented reality, virtual reality, machine learning, and the Internet of Things and is author of scientific articles indexed in IEEE Xplore, Scopus, and WoS (E-mail: 70994337@continental.edu.pe).