Performance analysis of 10 machine learning models in lung cancer prediction

Joselyn Zapata-Paulini[1], Michael Cabanillas-Carbonell[2]
[1]Graduate School, Universidad Continental, Lima, Peru
[2]Faculty of Engineering, Universidad Privada del Norte, Lima, Peru

Article Info

Article history:

Received Mar 1, 2024
Revised Sep 25, 2024
Accepted Sep 30, 2024

Keywords:

Lung cancer
Machine learning
Models
Performance
Predicting

ABSTRACT

Lung cancer is one of the diseases with the highest incidence and mortality in the world. Machine learning (ML) models can play an important role in the early detection of this disease. This study aims to identify the ML algorithm that has the best performance in predicting lung cancer. The algorithms that were contrasted were logistic regression (LR), decision tree (DT), k-nearest neighbors (KNN), gaussian Naive Bayes (GNB), multinomial Naive Bayes (MNB), support vector classifier (SVC), random forest (RF), extreme gradient boosting (XGBoost), multilayer perceptron (MLP) and gradient boosting (GB). The dataset used was provided by Kaggle, with a total of 309 records and 16 attributes. The study was developed in several phases, such as the description of the ML models and the analysis of the dataset. In addition, the contrast of the models was performed under the metrics of specificity, sensitivity, F1 count, accuracy, and precision. The results showed that the SVC, RF, MLP, and GB models obtained the best performance metrics, achieving 98% accuracy, 98% precision, and 98% sensitivity.

Corresponding Author:

Joselyn Zapata-Paulini
Graduate School, Universidad Continental
Alfredo Mendiola 5210, Los Olivos 15311, Lima, Perú
Email: 70994337@continental.edu.pe

1. INTRODUCTION

Lung cancer is one of the diseases with the highest mortality and incidence in the world [1]. It is estimated that about 1.8 million new cases and 1.6 million deaths occur each year [2]. Lung cancer is classified into molecularly and histologically heterogeneous categories [3], [4]. The most common types include large or small cell carcinoma, adenocarcinoma, squamous cell, and carcinoid [5]. Some of these cancers originate from poorly differentiated neuroendocrine cells, resulting in more rapid metastasis, consequently, poor prognosis [6]. Among the main risk factors for lung cancer is smoking, which accounts for 80% to 90% of diagnosed cases [1], [7]. People who die from this disease are usually diagnosed late, which makes it difficult to administer effective treatment and reduces the probability of survival [8].

Globally, approximately slightly more than 20% of patients diagnosed with lung cancer live longer than 5 years [9]. In countries such as the United States, lung cancer ranks second in terms of incidence, accounting for approximately 25% of all cancer deaths in the country [10]. On the other hand, in countries with a medium-high development index, an increase in the mortality rate has been observed [11]. Lung cancer is the leading cause of cancer death in the U.S., accounting for approximately 20.8% of all cancer deaths in 2023 [12], [13]. In 2020, Japan was the second country in the Organisation for Economic Co-operation and Development (OECD) with the most lung cancer deaths [14]. According to the most recent WHO data from 2020, lung cancer deaths in several countries show alarming figures. Hungary tops the list with 8,377 deaths (7.29% of all deaths), followed by Serbia, Turkey, North Korea and China, highlighting the urgency of preventive and treatment measures [15], [16]. In countries such as Greenland, Italy, and Slovenia, a decrease in lung cancer incidence has been observed. On the other hand, in Slovakia, Poland, and the Netherlands, an increase in the incidence of this disease has been recorded [17].

In recent decades, the field of machine learning (ML) has experienced great advances in the development of sophisticated algorithms and data preprocessing [18]. Emphasizing the importance of researchers taking advantage of ML’s predictive capabilities to address the diagnosis and treatment of diseases, using mathematical models to identify patterns in the data [19]-[21]. There are 4 types of ML techniques, which are supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [22]. In general, ML models provide systems with the ability to learn and improve through training, without the need to be explicitly programmed [23], [24]. Similarly, ML algorithms seek to automate the development of analytical models to perform tasks related to the detection or prediction of objects, and diseases, among others [25], [26]. However, to achieve greater accuracy in predictions, a large amount of data related to the subject of study is required [27].

This study aims to identify the ML algorithm with the best performance in predicting lung cancer. The algorithms that were contrasted were logistic regression (LR), decision tree (DT), k-nearest neighbors (KNN), gaussian Naive Bayes (GNB), multinomial Naive Bayes (MNB), support vector classifier (SVC), random forest (RF), extreme gradient boosting (XGBoost), multilayer perceptron (MLP) and gradient boosting (GB). In addition, the article is structured in 6 sections. In section 1 introduction, the problem of the case study is detailed. Section 2 bibliographic review describes the studies related to this research. Section 3 methodology is devoted to the research methodology, which is divided into two parts: 3.1. description of the ML models and 3.2. case study. Section 4 results, presents the model training results. The last sections are 5 discussions and 6 conclusions, where the obtained results are discussed and concluded.

2. BIBLIOGRAPHIC REVIEW

In this section, related work related to the case study is described. Radhika *et al.* [28], evaluate the performance of NB, support vector machine (SVM), DT, and LR algorithms for early diagnosis of lung cancer. Two datasets were used for training the algorithms. The results of the study showed that, with the first dataset, LR obtained the highest accuracy with a value of 0.969. With the second data set, SVM achieved the highest accuracy with a value of 0.992, followed by DT with 0.9 and NB with 0.8787. On the other hand, Dritsas and Trigka [29] compared the performance of NB, Bayesian network (BayesNet), stochastic gradient descent (SGD), SVM, LR, artificial neural network (ANN), KNN, J48, logistic model tree (LMT), RF, random tree (RT), reduced error pruning tree (RepTree), rotation forest (RotF) and adaptive boosting (AdaBoostM1) algorithms. In addition, their methodology employed the SMOTE technique and the cross-validation method for data processing. The study positioned RotF as the most efficient algorithm, achieving an accuracy and precision of 0.971, and an area under of curve (AUC) value of 0.993. On the other hand, LR achieved an accuracy and precision of 0.963, while KNN and RF obtained an accuracy and precision of 0.952. Similarly, Singh and Gupta [30] an efficient approach for lung cancer detection and classification based on images related to this disease is presented, and KNN, SVM, DT, MNB, SGD, RF, and MLP algorithms were analyzed. The results of the study positioned MLP with the best metrics, as it scored 0.8855 in accuracy and 0.8695 in precision. In contrast, the RF, SGD, MNB, DT, and SVM algorithms achieved 0.8481, 0.5771, 0.5140, and 0.5724, respectively. Patra [31], analyzed the radial basis function network (RBFN), SVC, LR, RF, J48, NB, and KNN algorithms for lung cancer prediction. The study concluded that RBFN obtained the highest accuracy with 0.8125, followed by NB and J48 with 0.7812, and KNN with 0.75. In contrast, Faisal *et al.* [32] evaluated MLP, NB, SVM, DT, gradient boosted tree (GBT), ANN, and RF algorithms for early-stage lung cancer prediction. The results position GBT as the best predictor with 0.9 in accuracy, followed by NB with 0.85, SVM and RF with 0.7917, and MLP with 0.7833. Similarly, Xie *et al.* [33] different ML algorithms are compared for the detection of biomarkers that aid in the early detection of lung cancer, SVM, RF, ANN, NB, and AdaBoost algorithms were analyzed. The study positioned NB with the best metrics with 0.1 in accuracy. In turn, Mishra and Gangwar [34] evaluate KNN, NB, DT, RF, and SVM algorithms for early detection of lung cancer. The study concluded that DT achieved the best performance with 0.1 in precision and accuracy, followed by RF with 0.98 and 0.984, KNN with 0.96 and 0.949, and NB and SVM with 0.91 in accuracy. Similarly, Gupta *et al.* [35] compare ML algorithms for lung cancer prediction, KNN, RF, and SVM algorithms were contrasted. The results of the study showed that RF performed the best with 0.842 in accuracy and 0.85 in precision, followed by SVM with 0.821 in accuracy and 0.828 in precision. Meanwhile, Ingle *et al.* [36] compared different ML algorithms for the detection of different types of lung cancer. The results position AdaBoost as the best algorithm, since it achieved 0.9074 in accuracy, 0.8180 in sensitivity, and 0.9399 in specificity. Celik *et al.* [37] perform a comparative study of different ML algorithms for lung cancer prediction. The study concluded that RF obtained the best performance with 0.9608 in accuracy. On the other hand, Mokoatle *et al.* [38] contrast XGBoost, light gradient boosting (LightGBM), RF, and convolutional neural network (CNN) algorithms for lung, breast, and prostate cancer detection. In their methodology, they employed the SMOTE technique, sentence-BERT (SBERT) and simple and efficient contrasting of sentence embeddings (SimCSE) transforms for data processing. The results of the study showed that XGBoost is the most efficient algorithm with an accuracy of 0.73. Göltepe [39] compare the performance of RF, KNN, NB, LR, DT, and SVM algorithms for lung cancer prediction. The study concluded that KNN, NB, and DT algorithms obtained the best performance with 0.71 accuracy. Similarly, Bharathy *et al.* [40] train SVM, KNN, DT, LR, NB, and RF algorithms to determine their performance in lung cancer detection. The results revealed that RF obtained an accuracy of 0.885, being the best algorithm in the lung cancer detection task. Meanwhile, Khan *et al.* [41] they evaluated multiple ML algorithms for lung cancer prediction. In their methodology, they employed the Kruskal-Walli test to select gene expression data. The results position RF as the most accurate, as it reached 0.84375. Finally, Banerjee and Das [42] they analyze SVM, RF, and ANN algorithms for lung cancer prediction. The results of the study show that ANN achieved the highest accuracy with 0.96, followed by SVM with 0.80 and RF with 0.70.

3. METHOD

In this section, the case study in which different ML models were developed and trained is presented. In the first part, the models (LR, DT, KNN, GNB, MNB, SVC, RF, XGBoost, MLP, and GB) are described. In the second part, the case study is described.

3.1. Description of the MLs models

3.1.1. Logistic regression

The LR model is one of the most widely used algorithms in medicine because of its usefulness in multivariable modeling [43], [44]. One of the most obvious advantages is its ability to convert coefficients into proportional odds [45]. In addition, LR provides us with a technique that guarantees that the training result is expressed in binary form, with values of 0 and 1 [46]. The mathematical equation of the LR model is expressed in (1).

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

3.1.2. Decision tree

The DT model is a simple tool that can separate data into categories with the use of classification rules [47]. DT is based on the divide-and-conquer strategy and is composed of leaf nodes that are connected, forming a hierarchical structure [48]. Since it is a classification model, it can be applied in various fields, including data mining and classification [49], [50]. DT can be represented in (2). Where *E* denotes the entropy, s is the sample, *Py* is the probability of occurrence of the SI event and *Pn* is the probability of occurrence of the NO event.

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

3.1.3. K-nearest neighbor

The KNN model is widely recognized for its effectiveness in data separation and can be useful when study data present ambiguities [51]. Furthermore, KNN groups data into coherent subsets and labels new data according to their similarity to the training results [52]. The model is a nonparametric algorithm, i.e., there is no fixed number of parameters independent of the data size [53]. The Euclidean equation in this model is show in (3).

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p}(x_{ri} - x_{rj})^2}, \tag{3}$$

3.1.4. Gaussian Naive Bayes

The GNB model is a probabilistic classification algorithm that has multiple applications, such as vehicle control and medical diagnosis [54]. GNB uses Bayes' rules and is based on assuming independence between features provided by the class, the model allows us to perform classifications efficiently [55]. In (4) describes the density function used in the model, where P(X|Y) is evaluated, "X" represents a class, and "Y" represents a particular object [56].

$$P(X|Y=c)=\frac{1}{\sqrt{2c\sigma_c^2}}e^{\frac{-(x-\mu_c)^2}{2\sigma_c^2}}, \tag{4}$$

3.1.5. Multinomial Naive Bayes

The MNB model is an adaptation of the Naive Bayes algorithm, which is mainly focused on text processing [57]. MNB counts the frequency with which words appear, completely ignoring binary occurrence [58]. Therefore, the model is suitable for categorizing documents [59]. The mathematical equation of the LR model is expressed in (5). Where $|V|$ corresponds to the vocabulary size and *(Ci)* indicates the total number of words.

$$P\left(d_j|C_i\right)=\prod_{t=1}^{|V|}P(W_t|C_i)^{x_t}, \tag{5}$$

3.1.6. Support vector classifier

The SVC model is based on sound principles derived from statistical learning theory, these fundamentals are used to develop models that can optimally apply classification or regression [60]. SVC is a generic classifier that can be applied in multiple fields since it can process numerical data and text [61]. The core of the model focuses on optimization, as it minimizes the common problems of ML algorithms. The objective function can be defined in (6) and (7). Where W is the vector of weights, b is the bias term, x is the feature vector, *yi* is the sample class label, and n is the number of samples.

$$\min 1/2w^2, \tag{6}$$

which is subject to:

$$y_i(wx+b)-1\geq 0, i=1\ldots.n, \tag{7}$$

3.1.7. Random forest

The RF model is recognized for obtaining excellent results in ML [62], [63]. RF is mainly used for classification and regression but can be applied to other tasks [64]. Likewise, it is used in multiple scientific fields, since it can reduce multi-source and multi-dimensional data [65].

3.1.8. Extreme gradient boosting

The XGB model is an ML algorithm that employs multiple weak learners to achieve a greater effect, it is used in multiple sectors, such as medicine [66]. The fundamental basis of XGB is the injection of numerous DT in each interaction to improve its performance, since these trees focus on the most difficult points to predict [67]. Also, to avoid over-fitting the data, the model uses a combination of the GB algorithm and regularization techniques [68], [69]. The (8) used to calculate the predictions of an XGB tree is detailed below. Where y is the final prediction of the model and *f(x)* is the prediction of the i-th DT.

$$\hat{y}_\iota=\sum_{t=1}^{m}f_t(x_i), \tag{8}$$

3.1.9. Multi-layer perceptron classifier

MLP model is a powerful tool for supervised training using multiple data output examples known to the algorithm [70]. MLP is a kind of ANN that generally consists of three layers, which are input, hidden, and output [71]. Similarly, the model is feedforward type, since it uses a backpropagation technique to learn [72]. Each node of the model is governed by (9). Where h_{1j} is the output of node *j*, w_{ij} represents the input gate of node *j* in the hidden layer h_1, x_i is the input corresponding to node *j* and b_j is the bias associated to node *j*.

$$h_{1j}=f(\sum_{i=0}^{n}w_{ij}x_i+b_j), \tag{9}$$

3.1.10. Gradient boosting

The GB model belongs to one of the most powerful classes of ML algorithms, due to its proven efficiency in various applications and areas of study [73]. Moreover, GB is one of the algorithms that mainly focuses on accuracy and speed for data processing [74]. The model is based on the statistical concept developed by Friedman, where the algorithm is optimized by using the gradient in the function space. The algorithm can be expressed as (10). Where $\hat{y}$ is the final model accuracy, *f(x)* is the prediction function, γ is the learning coefficient and *h(x)* is the prediction of the i-th weakest model.

$$\hat{y}=f(x)=\sum\gamma*h(x), \tag{10}$$

3.2. Case study

3.2.1. Understanding the dataset

The data used for this study were obtained from the Kaggle platform. The dataset has a total of 309 patient records collected from the online lung cancer prediction system website, with patients aged 21-87 years, balancing the records between males and females. The data contains 14 attributes, such as gender (M=male, F=emale), patient age, smoking (yes=2, no=1), yellow fingers (yes=2, no=1), anxiety (yes=2, no=1), pressure (yes=2, no=1), chronic disease (yes=2, no=1), fatigue (yes=2, no=1), allergy (yes=2, no=1), frostbite (yes=2, no=1), alcohol consumption (yes=2, no=1), cough (yes=2, no=1), shortness of breath (yes=2, no=1), difficulty swallowing (yes=2, no=1), chest pain (yes=2, no=1), lung pain (yes=2, no=1), lung cancer (yes=positive, no=negative). The development process of this investigation is detailed in Figure 1.

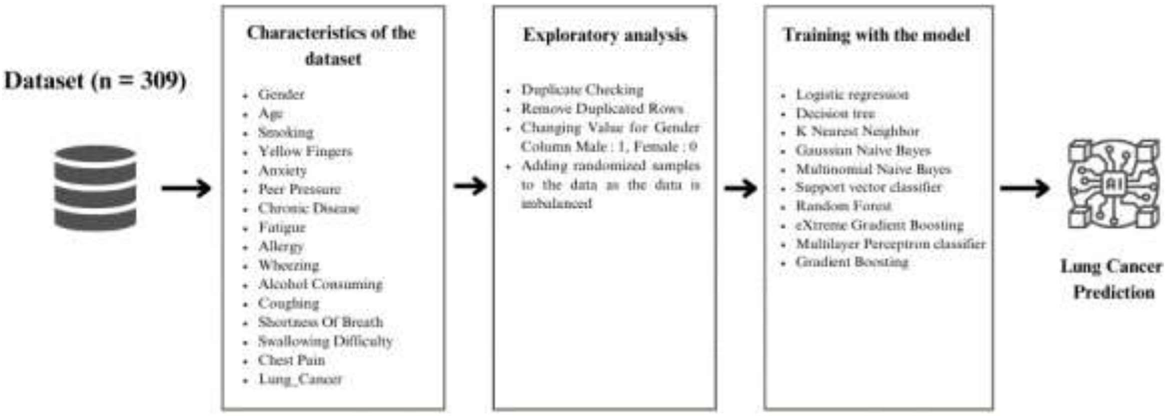


Figure 1. Case study development process

3.2.2. Preparation of the dataset

In this phase, a thorough analysis of the data in each column was performed as a starting point for processing. First, the necessary libraries were imported to perform a general exploration of the data. During the analysis, the types of variables found in each column were examined, as shown in supplementary Table 1. In addition, it was verified that there were no null data, after which the existence of duplicate data was verified, and once identified, the presence of duplicate data was eliminated. It should be noted that a column called “sex” was identified, which stored a character representative of the patient’s sex. In this regard, it was changed to binary values, where “1” was male and “0” female. Similarly, the column “lung cancer” was changed from string to number, where “1” was positive and “0” negative. In addition, the rest of the variables where “yes” was “2” and “no” was “1” were transformed to binary, where “1” was yes and “0” was no. Finally, the type of variable storing the column “sex” and “lung cancer” was changed, since it was originally identified as an object and when transforming the data it was changed to integer. The results of the data processing are shown supplementary Table 2.

Table 1. Information about the dataset

#	Column	Non-null	Dtype
0	gender	309 non-null	int64
1	age	309 non-null	int64
2	smoking	309 non-null	int64
3	yellow_fingers	309 non-null	int64
4	anxiety	309 non-null	int64
5	peer_pressure	309 non-null	int64
6	chronic disease	309 non-null	int64
7	fatigue	309 non-null	int64
8	allergy	309 non-null	int64
9	wheezing	309 non-null	int64
10	alcohol consuming	309 non-null	int64
11	coughing	309 non-null	int64
12	shortness of breath	309 non-null	int64
13	swallowing difficulty	309 non-null	int64
14	chest pain	309 non-null	int64
15	lung_cancer	309 non-null	int64

Table 2. Analysis of dataset variables

	0	1	2	3	4	...	279	280	281	282	283
Gender	1	1	0	1	0	...	0	0	1	1	1
Age	69	74	59	63	63	...	59	59	55	46	60
Smoking	0	1	0	1	0	...	0	1	1	0	0
Yellow_fingers	1	0	0	1	1	...	1	0	0	1	1
Anxiety	1	0	0	1	0	...	1	0	0	1	1
Peer_pressure	0	0	1	0	0	...	1	0	0	0	0
Chronic disease	0	1	0	0	0	...	0	1	0	0	0
Fatigue	1	1	1	0	0	...	0	1	1	0	1
Allergy	0	1	0	0	0	...	1	1	1	0	0
Wheezing	1	0	1	0	1	...	1	0	0	0	1
Alcohol consuming	1	0	0	1	0	...	0	0	0	0	1
Coughing	1	0	1	0	1	...	1	0	0	0	1
Shortness of breath	1	1	1	0	1	...	0	1	1	0	1
Swallowing difficulty	1	1	0	1	0	...	1	0	0	1	1
Chest pain	1	1	1	1	0	...	0	0	1	1	1
dfa	1	1	0	0	0	...	1	0	0	0	1

3.2.3. Exploratory analysis of the data

In the analysis of the “lung cancer” column, an imbalance was observed in the diagnosed cases, as there is a higher number of records with positive diagnoses, as shown in Figure 2. Therefore, this imbalance should be taken into account for the development of the models. Figure 3 shows the visualization of the variables according to their distribution by risk factors. According to Figures 3(a) and 3(b) people who consume alcohol and smoke have a higher probability of developing lung cancer (76 vs. 3). Similarly, people who smoke have a high probability of developing lung cancer (55 vs. 16)u, as do those who consume alcohol (69 vs. 4). On the other hand, people who do not consume alcohol and do not smoke can also develop lung cancer (38 cases).

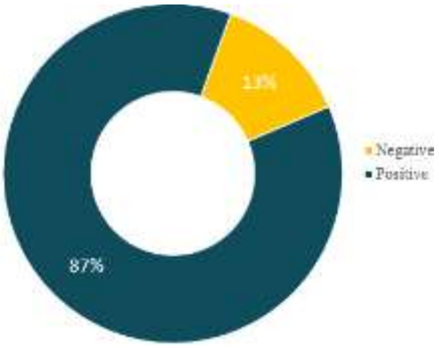
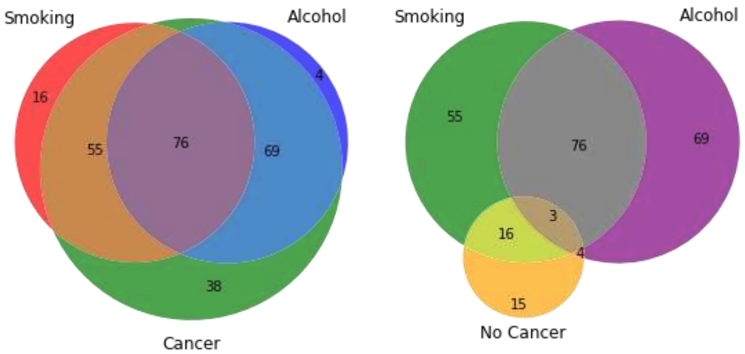


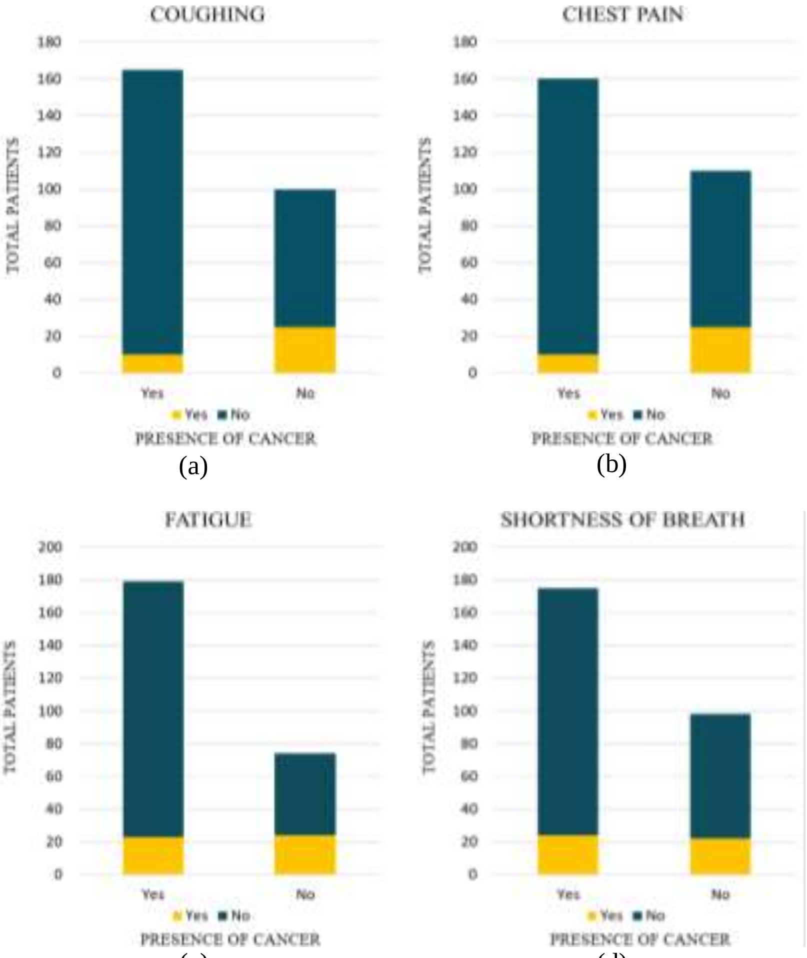
Figure 2. Analysis of lung cancer diagnoses



(a) (b)

Figure 3. Visualization of variables, (a) risk factors present in the positive diagnosis of cancer and (b) risk factors present in the negative diagnosis of cancer

Figure 4 presents the statistical graphs that analyze the symptoms of lung cancer. It is observed that cough and chest pain are more frequent in patients diagnosed with this disease and less common in those without it. Therefore, identifying these symptoms can be vital for the diagnosis of lung cancer, according to Figures 4(a) and 4(b). On the other hand, Figures 4(c) and 4(d) show that fatigue and shortness of breath are manifested in patients with and without lung cancer, but it can be evidenced that their presence is twice as high in patients diagnosed with cancer. Consequently, these symptoms can be considered as indicators for the diagnosis of lung cancer.



(a)

(b)

(c)

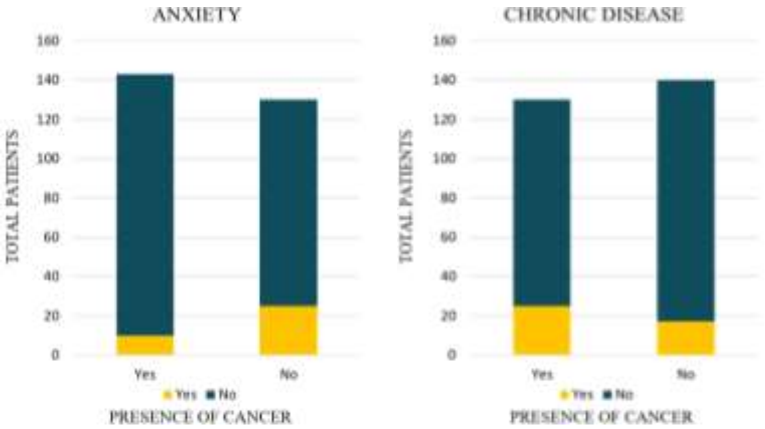
(d)

Figure 4. Symptoms of lung cancer; (a) analysis of cough and diagnosis of lung cancer, (b) analysis of chest pain and diagnosis of lung cancer, (c) analysis of fatigue and the diagnosis of lung cancer, and (d) analysis of shortness of breath and diagnosis of lung cancer

Figure 5 presents the statistical graphs that analyze the secondary symptoms of lung cancer. Figure 5(a) shows that anxiety appears to have a slight relationship with this condition. Similarly, chronic disease has a subtle correlation with lung cancer, according to Figure 5(b). Therefore, since chronic disease is a physiological comorbidity, it could be taken as a secondary symptom to predict lung cancer.

3.2.4. Data processing and modeling

Before the modeling and training of the algorithms, the RandomOverSampler technique of the Imblearn library was used to balance the data of the “lung cancer” column, which showed a significant imbalance in the number of positive and negative diagnoses. Subsequently, the dataset was split into two parts, one part for training and one part for testing. In addition, scaling was applied to both parts and then the models were trained.



(a) (b)

Figure 5. Secondary symptoms, (a) analysis of anxiety and the diagnosis of lung cancer and (b) analysis of chronic disease and lung cancer diagnosis

4. RESULTS

For this study, LR, DT, KNN, GNB, MNB, SVC, RF, XGBoost, MLP, and GB algorithms were trained to find the best-performing model for lung cancer prediction. The dataset was provided by Kaggle, these data were processed and cleaned for algorithm training. The training results are detailed in Table 3.

Table 3. Training results

	Precision (%)	Recall (%)	F1-score (%)	Support		Precision (%)	Recall (%)	F1-score (%)	Support
	Logistic regression					Decision tree			
0	0.96	1	0.98	64	…	0.93	0.97	0.95	64
1	1	0.95	0.97	56	…	0.96	0.91	0.94	56
Accuracy	-	-	0.97	120	…	-	-	0.94	120
Macro avg	0.98	0.97	0.97	120	…	0.94	0.94	0.94	120
Weighted avg	0.98	0.97	0.97	120	…	0.94	0.94	0.94	120
	KNN					Gaussian Naive Bayes			
0	0.93	1	0.96	64	…	0.95	0.89	0.92	64
1	1	0.91	0.95	56	…	0.88	0.95	0.91	56
Accuracy	-	-	0.96	120	…	-	-	0.92	120
Macro avg	0.96	0.96	0.96	120	…	0.92	0.92	0.92	120
Weighted avg	0.96	0.96	0.96	120		0.92	0.92	0.92	120
	Multinomial Naive Bayes					SVM			
0	0.89	0.73	0.8	64	…	0.98	0.98	0.98	64
1	0.75	0.89	0.81	56	…	0.98	0.98	0.98	56
Accuracy	-	-	0.81	120	…	-	-	0.98	120
Macro avg	0.82	0.81	0.81	120	…	0.98	0.98	0.98	120
Weighted avg	0.82	0.81	0.81	120		0.98	0.98	0.98	120
	Random forest					XGBoost			
0	0.98	0.98	0.98	64	…	0.98	0.97	0.98	64
1	0.98	0.98	0.98	56	…	0.96	0.98	0.97	56
Accuracy	-	-	0.98	120	…	-	-	0.97	120
Macro avg	0.98	0.98	0.98	120	…	0.97	0.98	0.97	120
Weighted avg	0.98	0.98	0.98	120		0.98	0.97	0.98	120
	MLP classifier					Gradient boosting			
0	0.98	0.98	0.98	64	…	0.98	0.98	0.98	64
1	0.98	0.98	0.98	56	…	0.98	0.98	0.98	56
Accuracy	-	-	0.98	120	…	-	-	0.98	120
Macro avg	0.98	0.98	0.98	120	…	0.98	0.98	0.98	120
Weighted avg	0.98	0.98	0.98	120		0.98	0.98	0.98	120

After training, the LR, DT, KNN, GNB, MNB, SVC, RF, XGBoost, MLP, and GB algorithms achieved 97%, 94%, 96%, 92%, 81%, 98%, 98%, 97%, 98%, and 98% accuracy, respectively. The complete information is shown in Table 3. In addition, Figure 6 shows the precision percentage of the algorithms visually, which helps to easily compare the performance of each model.

According to the results in Table 3, the SVC, RF, MLP, and GB models obtained the best performance metrics, since they achieved 98% in accuracy, 98% in precision, and 98% in sensitivity.

In the second place, the LR and XGBoost models achieved 97% accuracy, 98% precision, and 97% sensitivity. In third place is the KNN model that achieved 96% accuracy, precision and sensitivity. In the last places are the DT, GNB, and MNB models, which achieved the lowest metrics, 94%, 92%, and 81% accuracy.

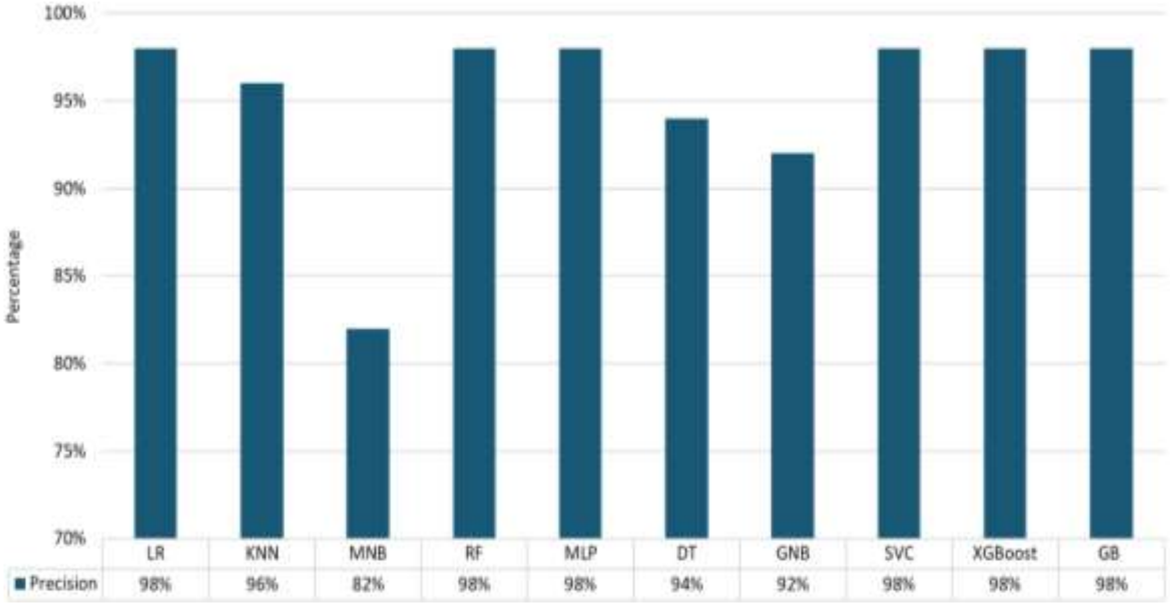


Figure 6. Precision of algorithms for predicting lung cancer

5. DISCUSSION

Lung cancer remains one of the cancers with the highest incidence and mortality worldwide, claiming thousands of lives each year. Early detection is crucial to improving survival rates, as it allows for timely treatment interventions. This study investigated the effects of ML models on lung cancer prediction. While previous studies have explored the impact of various ML algorithms on disease prediction, the literature lacks comprehensive comparisons between a wide range of ML models (such as LR, DT, KNN, GNB, MNB, SVC, RF, XGBoost, MLP, and GB). This study aims to fill this gap by identifying the model with the best performance in predicting lung cancer.

After training, it was found that the SVC, RF, MLP, and GB models exhibited the highest performance metrics, each achieving 98% accuracy, precision, and sensitivity in predicting lung cancer. These results are similar to those obtained in [34], where the RF model achieved a performance of 98%. Similarly, Xie *et al.* [33] RF, it achieved 96.68% accuracy. In contrast, in studies [35], [40], [41], [42] the RF model achieved performance metrics by 84.2%, 88.3%, 84.37%, and 70%, respectively, which are lower than those obtained in this study. Similarly, the MLP model in our study achieved 98% accuracy, outperforming studies [30], [32], which reported 88.55% and 78.33%, respectively. On the other hand, in this study, the LR and XGBoost models achieved 97% accuracy, 98% accuracy, and 97% sensitivity, results similar to those presented in [28], [29] as these studies achieved a performance of 96.9% and 96.3% in the LR and XGBoost models. In the opposite position, [38] it determined that the XGBoost model achieved an accuracy of 73%, which is a lower result than those obtained in this study. Furthermore, in [29], [31], [34], [39] the KNN model scored 95.2%, 75%, 96%, and 71%, respectively. These results are mostly similar to those obtained in this study, as the KNN model achieved 96% in precision, accuracy, and sensitivity. This indicates that our dataset and preprocessing methods may have contributed significantly to the higher throughput.

In the present study, a comprehensive dataset was used to evaluate the performance of various ML models in predicting lung cancer. However, the quality of the dataset, including its size, diversity, and feature representation, can significantly affect the effectiveness of models. While our study achieved high-throughput metrics, more research is needed to validate these findings in different populations and clinical settings. In addition, the generalizability of the results may be limited by the specific characteristics and preprocessing techniques used in our study.

Future studies should consider diverse datasets and alternative feature sets to ensure broader applicability. In addition, it is essential to continue refining these models and exploring their integration into clinical practice to maximize their impact on patient care. The incorporation of real-time data and the development of hybrid models could provide significant improvements in the accuracy and clinical utility of ML models in lung cancer prediction.

Our findings provide conclusive evidence of the potential of ML in medical diagnostics and the need for continuous improvement and validation of these models with diverse datasets. Effective implementation of ML models in clinical settings has the potential to revolutionize the early diagnosis of lung cancer, thereby improving patient outcomes and optimizing healthcare resources.

6. CONCLUSION

After presenting the training results of the LR, DT, KNN, GNB, MNB, SVC, RF, XGBoost, MLP, and GB models for the prediction of lung cancer, the following conclusions were reached. The models obtained outstanding results, mainly SVC, RF, MLP, and GB which achieved the best metrics in precision, accuracy, and sensitivity for lung cancer prediction and may be vital for early detection to help improve patient prognosis. Similarly, the LR, XGBoost, KNN, DT, and GNB models achieved exceptional metrics for cancer prediction. Except for MNB, which achieved less than 90% performance. Additionally, bad habits related to alcohol consumption and smoking are factors that have a higher presence in patients diagnosed with lung cancer. On the other hand, symptoms such as cough, chest pain, and chronic disease are indicators that, although present in patients without cancer, should also be considered for an early diagnosis of lung cancer. Finally, all of the algorithms trained and analyzed in this study proved to be useful tools for lung cancer prediction. Although the models achieved outstanding metrics, it is recommended for future research to explore models such as CNNs or recurrent neural networks (RNNs) to see if they offer improvements in prediction accuracy. In addition, it would be beneficial to evaluate additional, larger, more diverse, or collected datasets from different geographic regions, to determine the generalizability and performance of models in different contexts, allowing them to be tested for effectiveness.

REFERENCES

[1] World Health Organization, “Lung cancer,” *WHO*, 2023. https://www.who.int/news-room/fact-sheets/detail/lung-cancer (accessed Dec. 03, 2023).

[2] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, “Global cancer statistics, 2012,” *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, Mar. 2015, doi: 10.3322/caac.21262.

[3] N. Howlader *et al.*, “The effect of advances in lung-cancer treatment on population mortality,” *New England Journal of Medicine*, vol. 383, no. 7, pp. 640–649, Aug. 2020, doi: 10.1056/nejmoa1916623.

[4] W. D. Travis, E. Brambilla, A. P. Burke, A. Marx, and A. G. Nicholson, “Introduction to the 2015 World Health Organization classification of tumors of the lung, pleura, thymus, and heart,” *Journal of Thoracic Oncology*, vol. 10, no. 9, pp. 1240–1242, Sep. 2015, doi: 10.1097/JTO.0000000000000663.

[5] W. D. Travis *et al.*, “The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification,” *Journal of Thoracic Oncology*, vol. 10, no. 9, pp. 1243–1260, Sep. 2015, doi: 10.1097/JTO.0000000000000630.

[6] K. C. Thandra, A. Barsouk, K. Saginala, J. S. Aluru, and A. Barsouk, “Epidemiology of lung cancer,” *Wspolczesna Onkologia*, vol. 25, no. 1, pp. 45–52, 2021, doi: 10.5114/wo.2021.103829.

[7] M. B. Schabath and M. L. Cote, “Cancer progress and priorities: lung cancer,” *Cancer Epidemiology Biomarkers and Prevention*, vol. 28, no. 10, pp. 1563–1579, Oct. 2019, doi: 10.1158/1055-9965.EPI-19-0221.

[8] R. Nooreldeen and H. Bach, “Current and future development in lung cancer diagnosis,” *International Journal of Molecular Sciences*, vol. 22, no. 16, p. 8661, Aug. 2021, doi: 10.3390/ijms22168661.

[9] A. H. Krist *et al.*, “Screening for lung cancer: US preventive services task force recommendation statement,” *JAMA - Journal of the American Medical Association*, vol. 325, no. 10, pp. 962–970, Mar. 2021, doi: 10.1001/jama.2021.1117.

[10] D. Yang, Y. Liu, C. Bai, X. Wang, and C. A. Powell, “Epidemiology of lung cancer and lung cancer screening programs in China and the United States,” *Cancer Letters*, vol. 468, pp. 82–87, Jan. 2020, doi: 10.1016/j.canlet.2019.10.009.

[11] World Health Organization, “The top 10 causes of death,” *World Health Organization*, vol. 12, 2020, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed Dec. 03, 2023).

[12] National Cancer Institute, “Cancer stat facts: lung and bronchus cancer. secondary cancer stat facts: lung and bronchus cancer,” *Surveillance, Epidemiology, and End Results Program*, 2021, [Online]. Available: https://seer.cancer.gov/statfacts/html/lungb.html (accessed Dec. 03, 2023).

[13] “American Cancer Society,” “Lung cancer statistics | how common is lung cancer,” *Facts and Figures 2020*, 2020, [Online]. Available: https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html (accessed Dec. 03, 2023).

[14] “Cáncer de pulmón, traquea y bronquios: fallecimientos por país OCDE,” *Statista*, 2023, [Online]. Available: https://es.statista.com/estadisticas/588401/numero-de-muertes-por-neoplasia-en-determinados-paises-de-la-ocde/ (accessed Dec. 03, 2023).

[15] World Health Organization, “World Health Organization. world health rankings live longer live better 2016,” *World Health Organization*, 2016, [Online]. Available: www.worldlifeexpectancy.com/world-health-rankings (accessed Dec. 03, 2023).

[16] B. Zhou *et al.*, “Worldwide burden and epidemiological trends of tracheal, bronchus, and lung cancer: a population-based study,” *eBioMedicine*, vol. 78, p. 103951, Apr. 2022, doi: 10.1016/j.ebiom.2022.103951.

[17] J. Huang *et al.*, "Distribution, risk factors, and temporal trends for lung cancer incidence and mortality: a global analysis," *Chest*, vol. 161, no. 4, pp. 1101–1111, Apr. 2022, doi: 10.1016/j.chest.2021.12.655.

[18] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/s12525-021-00475-2.

[19] G. Sahoo, A. K. Nayak, P. K. Tripathy, and J. Tripathy, "A novel machine learning based hybrid approach for breast cancer relapse prediction," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 32, no. 3, pp. 1655–1663, Dec. 2023, doi: 10.11591/ijeecs.v32.i3.pp1655-1663.

[20] O. Babu P., C. C. Sobin, and J. Ali, "Building machine learning-based prediction system for critical diseases," in *Deep Learning for Cognitive Computing Systems: Technological Advancements and Applications*, De Gruyter, 2022, pp. 75–96.

[21] T. Suresh, T. A. Assegie, S. Ganesan, R. L. Tulasi, R. Mothukuri, and A. O. Salau, "Explainable extreme boosting model for breast cancer diagnosis," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, pp. 5764–5769, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5764-5769.

[22] I. H. Sarker, "Machine learning: algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.

[23] I. H. Sarker, M. M. Hoque, M. K. Uddin, and T. Alsanoosy, "Mobile data science and intelligent apps: concepts, AI-based modeling and research directions," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 285–303, Feb. 2021, doi: 10.1007/s11036-020-01650-z.

[24] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *Journal of Big Data*, vol. 7, no. 1, p. 41, Dec. 2020, doi: 10.1186/s40537-020-00318-5.

[25] M. I. Jordan and T. M. Mitchell, "Machine learning: trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.

[26] T. A. Assegie, R. L. Tulasi, V. Elanangai, and N. K. Kumar, "Exploring the performance of feature selection method using breast cancer dataset," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 25, no. 1, pp. 232–237, Jan. 2022, doi: 10.11591/ijeecs.v25.i1.pp232-237.

[27] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. S. Rellermeyer, "A survey on distributed machine learning," *ACM Computing Surveys*, vol. 53, no. 2, pp. 1–33, Mar. 2020, doi: 10.1145/3377454.

[28] P. R. Radhika, R. A. S. Nair, and G. Veena, "A comparative study of lung cancer detection using machine learning algorithms," in *Proceedings of 2019 3rd IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT 2019*, Feb. 2019, pp. 1–4, doi: 10.1109/ICECCT.2019.8869001.

[29] E. Dritsas and M. Trigka, "Lung cancer risk prediction with machine learning models," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 139, Nov. 2022, doi: 10.3390/bdcc6040139.

[30] G. A. P. Singh and P. K. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans," *Neural Computing and Applications*, vol. 31, no. 10, pp. 6863–6877, Oct. 2019, doi: 10.1007/s00521-018-3518-x.

[31] R. Patra, "Prediction of lung cancer using machine learning classifier," in *Communications in Computer and Information Science*, vol. 1235 CCIS, 2020, pp. 132–142.

[32] M. I. Faisal, S. Bashir, Z. S. Khan, and F. Hassan Khan, "An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer," in *2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology, ICEEST 2018*, Dec. 2018, pp. 1–4, doi: 10.1109/ICEEST.2018.8643311.

[33] Y. Xie *et al.*, "Early lung cancer diagnostic biomarker discovery by machine learning methods," *Translational Oncology*, vol. 14, no. 1, p. 100907, Jan. 2021, doi: 10.1016/j.tranon.2020.100907.

[34] A. Mishra and S. Gangwar, "Lung cancer detection and classification using machine learning algorithms," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 6 S, pp. 277–282, Jun. 2023, doi: 10.17762/ijritcc.v11i6s.6920.

[35] A. Gupta *et al.*, "A study on prediction of lung cancer using machine learning algorithms," *Research Square*, pp. 1–13, 2022, doi: 10.21203/rs.3.rs-1912967/v1.

[36] K. Ingle, U. Chaskar, and S. Rathod, "Lung cancer types prediction using machine learning approach," in *Proceedings of CONECCT 2021: 7th IEEE International Conference on Electronics, Computing and Communication Technologies*, Jul. 2021, pp. 01–06, doi: 10.1109/CONECCT52877.2021.9622568.

[37] A. E. Celik, J. Rasheed, and A. Yahyaoui, "Machine learning approaches for lung cancer prediction," in *Proceedings - International Conference on Advanced Computer Information Technologies, ACIT*, Sep. 2022, pp. 540–543, doi: 10.1109/ACIT54803.2022.9913114.

[38] M. Mokoatle, V. Marivate, D. Mapiye, R. Bornman, and V. M. Hayes, "A review and comparative study of cancer detection using machine learning: SBERT and SimCSE application," *BMC Bioinformatics*, vol. 24, no. 1, p. 112, Mar. 2023, doi: 10.1186/s12859-023-05235-x.

[39] Y. Göltepe, "Performance of lung cancer prediction methods using different classification algorithms," *Computers, Materials and Continua*, vol. 67, no. 2, pp. 2015–2028, 2021, doi: 10.32604/cmc.2021.014631.

[40] S. Bharathy, R. Pavithra, and B. Akshaya, "Lung cancer detection using machine learning," in *Proceedings - International Conference on Applied Artificial Intelligence and Computing, ICAAIC 2022*, May 2022, pp. 539–543, doi: 10.1109/ICAAIC53929.2022.9793061.

[41] F. Khan, K. Pradhan, and D. Sinha, "A model for lung cancer prediction," in *Proceedings - 2021 3rd International Conference on Advances in Computing, Communication Control and Networking, ICAC3N 2021*, Dec. 2021, pp. 251–255, doi: 10.1109/ICAC3N53548.2021.9725462.

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

[43] A. Hudon, M. Beaudoin, K. Phraxayavong, S. Potvin, and A. Dumais, "Enhancing predictive power: integrating a linear support vector classifier with logistic regression for patient outcome prognosis in virtual reality therapy for treatment-resistant schizophrenia," *Journal of Personalized Medicine*, vol. 13, no. 12, p. 1660, Nov. 2023, doi: 10.3390/jpm13121660.

[44] M. M. Saim and H. Ammor, "Comparative study of machine learning algorithms (SVM, logistic regression and KNN) to predict cardiovascular diseases," *E3S Web of Conferences*, vol. 351, p. 01037, May 2022, doi: 10.1051/e3sconf/202235101037.

[45] J. Jeppesen, J. Christensen, P. Johansen, and S. Beniczky, "Personalized seizure detection using logistic regression machine learning based on wearable ECG-monitoring device.," *Seizure*, vol. 107, pp. 155–161, Apr. 2023, doi: 10.1016/j.seizure.2023.04.012.

[46] E. Rave Gómez, “Dependence between income and savings rates in professionals in the south of the Aburrá Valley (in Spanish: Dependencia entre ingresos y tasas de ahorro en profesionales del sur del valle de Aburrá),” *Ecos de Economía*, vol. 17, no. 37, pp. 161–175, Nov. 2013, doi: 10.17230/ecos.2013.37.7.

[47] C. S. Lee, P. Y. S. Cheang, and M. Moslehpour, “Predictive analytics in business analytics: decision tree,” *Advances in Decision Sciences*, vol. 26, no. 1, pp. 1–29, 2022, doi: 10.47654/V26Y2022I1P1-30.

[48] A. L. López-Lobato, H. G. Acosta-Mesa, and E. Mezura-Montes, “Blood cell image segmentation using convolutional decision trees and differential evolution,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14502 LNAI, 2024, pp. 315–325.

[49] G. Karalis, “Decision trees and applications,” in *Advances in Experimental Medicine and Biology*, vol. 1194, 2020, pp. 239–242.

[50] O. Iparraguirre-Villanueva *et al.*, “Classification of tweets related to natural disasters using machine learning algorithms,” *International Journal of Interactive Mobile Technologies*, vol. 17, no. 14, pp. 144–162, Aug. 2023, doi: 10.3991/ijim.v17i14.39907.

[51] M. Bansal, A. Goyal, and A. Choudhary, “A comparative analysis of k-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning,” *Decision Analytics Journal*, vol. 3, p. 100071, Jun. 2022, doi: 10.1016/j.dajour.2022.100071.

[52] K. Taunk, S. De, S. Verma, and A. Swetapadma, “A brief review of nearest neighbor algorithm for learning and classification,” in *2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019*, May 2019, pp. 1255–1260, doi: 10.1109/ICCS45141.2019.9065747.

[53] H. A. Abu Alfeilat *et al.*, “Effects of distance measure choice on k-nearest neighbor classifier performance: a review,” *Big Data*, vol. 7, no. 4, pp. 221–248, Dec. 2019, doi: 10.1089/big.2018.0175.

[54] I. Wickramasinghe and H. Kalutarage, “Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation,” *Soft Computing*, vol. 25, no. 3, pp. 2277–2293, Feb. 2021, doi: 10.1007/s00500-020-05297-6.

[55] T. R. Gadekallu *et al.*, “Early detection of diabetic retinopathy using pca-firefly based deep learning model,” *Electronics (Switzerland)*, vol. 9, no. 2, p. 274, Feb. 2020, doi: 10.3390/electronics9020274.

[56] M. V. Anand, B. Kiranbala, S. R. Srividhya, K. C., M. Younus, and M. H. Rahman, “Gaussian Naïve Bayes algorithm: a reliable technique involved in the assortment of the segregation in cancer,” *Mobile Information Systems*, vol. 2022, pp. 1–7, Jun. 2022, doi: 10.1155/2022/2436946.

[57] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, “Multinomial naive bayes for text categorization revisited,” in *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 2004, pp. 488–499.

[58] S. Silva, A. S. Vieira, P. Celard, E. L. Iglesias, and L. Borrajo, “A query expansion method using multinomial naive bayes,” *Applied Sciences (Switzerland)*, vol. 11, no. 21, p. 10284, Nov. 2021, doi: 10.3390/app112110284.

[59] S. Raschka, “Naive Bayes and text classification i - introduction and theory,” *ArXiv*, 2014, [Online]. Available: http://arxiv.org/abs/1410.5329.

[60] G. Sajiv and G. Ramkumar, “Classification and prediction of breast cancer based on support vector classifier on histopathological images,” in *Proceedings of the 2nd IEEE International Conference on Advances in Computing, Communication and Applied Informatics, ACCAI 2023*, May 2023, pp. 1–8, doi: 10.1109/ACCAI58221.2023.10199554.

[61] P. Saigal and V. Khanna, “Multi-category news classification using support vector machine based classifiers,” *SN Applied Sciences*, vol. 2, no. 3, p. 458, Mar. 2020, doi: 10.1007/s42452-020-2266-6.

[62] M. Schonlau and R. Y. Zou, “The random forest algorithm for statistical learning,” *Stata Journal*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.

[63] Y. J. Luwe, C. P. Lee, and K. M. Lim, “Wearable sensor-based human activity recognition with ensemble learning: a comparison study,” *International Journal of Electrical and Computer Engineering*, vol. 13, no. 4, pp. 4029–4040, Aug. 2023, doi: 10.11591/ijece.v13i4.pp4029-4040.

[64] R. Couronné, P. Probst, and A. L. Boulesteix, “Random forest versus logistic regression: a large-scale benchmark experiment,” *BMC Bioinformatics*, vol. 19, no. 1, p. 270, Dec. 2018, doi: 10.1186/s12859-018-2264-5.

[65] A. Sarica, A. Cerasa, and A. Quattrone, “Random forest algorithm for the classification of neuroimaging data in Alzheimer’s disease: A systematic review,” *Frontiers in Aging Neuroscience*, vol. 9, no. OCT, Oct. 2017, doi: 10.3389/fnagi.2017.00329.

[66] O. Alghushairy, F. Ali, W. Alghamdi, M. Khalid, R. Alsini, and O. Asiry, “Machine learning-based model for accurate identification of druggable proteins using light extreme gradient boosting,” *Journal of Biomolecular Structure and Dynamics*, pp. 1–12, Oct. 2023, doi: 10.1080/07391102.2023.2269280.

[67] T. Thenmozhi and R. Helen, “Feature selection using extreme gradient boosting Bayesian optimization to upgrade the classification performance of motor imagery signals for BCI,” *Journal of Neuroscience Methods*, vol. 366, p. 109425, Jan. 2022, doi: 10.1016/j.jneumeth.2021.109425.

[68] A. Ramón, A. M. Torres, J. Milara, J. Cascón, P. Blasco, and J. Mateo, “Extreme gradient boosting-based method to classify patients with COVID-19,” *Journal of Investigative Medicine*, vol. 70, no. 7, pp. 1472–1480, Oct. 2022, doi: 10.1136/jim-2021-002278.

[69] 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*, vol. 14, no. 4, pp. 494–503, 2023, doi: 10.14569/IJACSA.2023.0140455.

[70] H. Taud and J. F. Mas, “Multilayer perceptron (MLP),” *Geomatic Approaches for Modeling Land Change Scenarios*, 2018, pp. 451–455.

[71] Y. Chen *et al.*, “Development of a machine learning classifier for brain tumors diagnosis based on DNA methylation profile,” *Frontiers in Bioinformatics*, vol. 1, Nov. 2021, doi: 10.3389/fbinf.2021.744345.

[72] M. Desai and M. Shah, “An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN),” *Clinical eHealth*, vol. 4, pp. 1–11, 2021, doi: 10.1016/j.ceh.2020.11.002.

[73] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Frontiers in Neurorobotics*, vol. 7, no. Dec, 2013, doi: 10.3389/fnbot.2013.00021.

[74] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

BIOGRAPHIES OF AUTHORS

Joselyn Zapata-Paulini bachelor in systems engineering and computer science from the Universidad de Ciencias y Humanidades, Master in Science with environmental management and sustainable development at the Universidad Continental, Peru. She has several international publications. Specialized in the areas of augmented reality, virtual reality, and the internet of things. Author of scientific articles indexed in IEEE Xplore, Scopus, and WoS. She can be contacted at email: 70994337@continental.edu.pe.

Michael Cabanillas-Carbonell engineer and master in systems engineering from the National University of Callao - Peru, Ph.D. candidate in Systems Engineering and Telecommunications at the Polytechnic University of Madrid. Ex president of the chapter of the Education Society IEEE-Peru. Conference Chair of the Engineering International Research Conference IEEE Peru EIRCON. Specialization in software development, artificial intelligence, machine learning, business intelligence, and augmented reality. Reviewer IEEE Peru and author of more than 100 scientific articles indexed in IEEE Xplore and Scopus. He can be contacted at email: mcabanillas@ieee.org.