# Evaluation of machine learning algorithms in the early detection of Parkinson's disease: a comparative study

**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 Jan 17, 2024
Revised Feb 16, 2024
Accepted Mar 16, 2024

***Keywords:***

Algorithm
Comparative
Early detection
Machine learning
Parkinson

**ABSTRACT**

Parkinson's is a neurodegenerative disease that generally affects people over 60 years of age. The disease destroys neurons and increases the accumulation of α-synuclein in many parts of the brain stem, although at present its causes remain unknown. It is therefore a priority to identify a method that can detect the disease, and this is where machine learning models become important. This study aims to perform a comparative analysis of machine learning models focused on the early detection of Parkinson's disease. Logistic regression (LR), support vector machines (SVM), decision trees (DT), extra trees classifiers (ETC), K-nearest neighbors (KNN), random forests (RF), adaptive boosting (AdaBoost) and gradient boosting (GB) algorithms are described and developed to identify the one that offers the best performance. In the training stage, we used the Oxford University dataset for Parkinson's disease detection, which has a total of 23 attributes and 195 records on patient voice recordings. The article is structured into six sections, such as introduction, related work, methodology, results, discussions, and conclusions. The metrics of accuracy, sensitivity, F1 count, and precision were used to measure the models' performance. The results position the KNN model as the best predictor with 95% accuracy, precision, sensitivity, and F1 score.

***Corresponding Author:***

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

## 1. INTRODUCTION

Parkinson's disease is a disorder that impacted approximately 6 million people globally in 2016 [1]. Over the past two decades, a significant increase in the incidence of this disease has been observed, although the reasons behind this increase are not fully elucidated [2], [3]. This condition is characterized by the progressive loss of neurons of the substantia nigra pars compacta [4] and the accumulation of α-synuclein in different areas of the brainstem, but its origin remains an unsolved enigma for the scientific community [5], [6]. With approximately 3% to 5% of cases linked to genetic bases and 16% to 36% related to hereditary factors, Parkinson's is now classified as a heterogeneous disease affecting various regions of the nervous system [7], [8].

In this context, the prevalence of Parkinson's disease in developed countries is about 3% in the general population and about 1% in people over 60 years of age [9], [10]. People of different ethnic backgrounds may be affected, although men are slightly more predisposed to the disease [11], [12]. The age of onset, previously estimated at 50.8 years, is now around 60 years, and the first symptoms may appear between 21 and 40 years of age, sometimes extending into the 50 years [13], [14]. Definitive diagnosis of Parkinson's disease is made by autopsy; however, clinical diagnosis is made by diagnostic certainty: clinically possible, clinically probable, and clinically definite Parkinson's disease [15].

The standardized incidence rate worldwide is 8 to 18 cases per 100,000 inhabitants [16]. In Spain, an annual incidence rate of 168 per 100,000 inhabitants was calculated, affecting mostly elderly men, but decreasing in women over 79 years of age [17]. On the other hand, in Rotterdam, the Netherlands, there was a slight increase in the incidence of Parkinson's disease from 0.3 per 1,000 inhabitants aged 55 to 65 years to 4.4 per 1,000 inhabitants over 85 years, with men being more likely to develop the disease [18]. Similarly, in the northern area of Manhattan, United States, an incidence of 107 cases per 100,000 inhabitants was identified, where women had the lowest prevalence as opposed to men, as well as white and Hispanic people [19]. In the case of Taiwan, an annual incidence of 10.4 cases per 100,000 population and a mortality rate of 40.4% was recorded [20]. In Australia, a prevalence of 85 cases per 100,000 population was recorded [21]. Similarly, China has an annual incidence of 170 cases per 100,000 inhabitants over 50 years of age [22]. It is therefore necessary to find an effective method for early detection of the disease, and it is in this context that machine learning (ML) comes into play.

ML is a subfield of artificial intelligence (AI) that allows computers to learn and improve from training with data related to a specific topic and is characterized by not being explicitly programmed [23]. It involves the application of algorithms and statistical models to examine data and subsequently use this information to predict outcomes or make decisions [24], [25]. Disease prediction using ML involves analyzing large volumes of medical data to train algorithms to predict the likelihood of a patient developing a particular disease [26]. During training, the models seek to identify patterns and relationships between the data that may be determinant in the detection or prediction of a disease [27].

The aim of this study is to address the urgent need for effective methods for the early detection of Parkinson's disease. To this end, a comparative analysis of several ML models is performed in this context. Specifically, logistic regression (LR), support vector machines (SVM), decision trees (DT), extra trees classifier (ETC), K-nearest neighbors (KNN), random forests (RF), AdaBoost and gradient boosting (GB) algorithms are described and developed to identify the model offering the best performance in this task.

This work contributes to research by addressing unsolved problems and areas requiring improvement in early Parkinson's detection. The contributions focus on the comprehensive comparison and evaluation of multiple ML models, considering their specific performance on the proposed task. The article is structured in six main parts. Section 1 introduction, we contextualize the problems of the case study. Section 2 review of the literature, the relevant literature is reviewed to provide an overview of previous approaches and highlight gaps in knowledge. Section 3 methodology, we detail the methodology, which is divided into two sections. In section 3.1, we discuss the ML models, and in section 3.2, we develop the case study. Then, section 4 results, we present the results. In the last two parts section 5 discussion, and section 6 conclusion, we discuss the results and present the conclusions of the study, summarizing the contributions and outlining possible future directions for research.

## 2. REVIEW OF THE LITERATURE

In this section, we detail some works related to the case study. The authors of the study [28], analyzed and developed different ML models for the early detection of Parkinson's disease, for the training of the models they used a dataset on the non-motor and olfactory characteristics of the patients. The results of the study positioned the LR model as the best disease detector with 0.97159 in accuracy. Similar to the study [29], where they analyze SVM, RF, DT, KNN, and multi-layer perceptron (MLP) models to detect Parkinson's disease, differentiating healthy from diseased people, in the training stage they used the University of California at Irvine (UCI) dataset with 195 voice recordings. The study concluded that MLP is the best detector with 0.9831 accuracy, followed by SVM with 0.95 accuracy. Similarly, Sharma and Mishra [30] analyzed seven ML models focused on predicting Parkinson's disease, they used a dataset from Oxford University with voice biometric recordings from 31 patients. The results position the AdaBoost model as the best predictor with a precision and accuracy of 0.87 and 0.864 respectively, followed by RF and LR which achieved a precision of 0.85 and 0.8305 respectively.

Swaroopa and Saritha [31], analyzed multiple ML algorithms for Parkinson's disease classification and detection, for training the models they used a dataset storing voice recordings of people with and without Parkinson's. The results of the study positioned the extreme gradient boosting (XGBoost) model as the best with 0.95 accuracy. On the other hand, [32] conducted an evaluation of multiple ML algorithms for the identification of patients with Parkinson's disease, the models were trained on a dataset of 195 vowel phonation records. The results of the study showed that the SVM model achieved 0.914 in accuracy. Similarly, Sakar *et al.* [33] explore the SVM model in predicting Parkinson's disease by recording speech, isolated words, and short sentences. The study concluded that SVM achieved an accuracy of 0.775. In [34], the authors analyze DT, LR, and artificial neural networks (ANN) models to identify the model with the best accuracy in diagnosing Parkinson's disease. The training results showed that ANN achieved better performance with 0.929 in accuracy.

Nahar *et al.* [35] conducted a comparative study of different ML models focused on early detection of Parkinson's disease, in the methodological part they employed feature selection methods such as Boruta, recursive feature elimination (RFE), and RF classifier. The results show that the GB model achieved the best performance with 0.79 in precision and 0.7916 in accuracy, followed by ETC with 0.73 in precision and 0.75 in accuracy. Saeed *et al.* [36], perform a comparative analysis of different ML models for early prediction of Parkinson's disease, for training the models they used a dataset storing 240 patient voice recordings. The training results positioned the KNN model as the best predictor with 0.8833 in accuracy.

Likewise, Kundu *et al.* [37] seek to diagnose Parkinson's with the use of ML models, for this, they used the recorded voice dataset from the ICU machine learning repository. The results of the study positioned the AdaBoost model as the best predictor with 0.974 accuracy. Tiwari [38] analyzed eight models focused on Parkinson's disease prediction, as part of their methodology they employed feature selection of 5, 10, and 20 random selections. The training results positioned RF as the best predictor with 0.905 in precision and 0.975 in accuracy, followed by DT with 0.905 in precision and 0.905 in accuracy. Mandal and Sairam [39] optimize the accuracy of some models focused on early diagnosis of Parkinson's disease. The results showed that LR achieved the best performance with a sensitivity and specificity of 0.983 and 0.996, respectively. On the other hand, [40] performed a comparative study of ML algorithms for the early detection of Parkinson's disease. The results position the DT model as the best predictor with 0.9477 in accuracy. The study [41], focused on the early detection of Parkinson's disease through the analysis of voice signals. Recordings of patients pronouncing vowels, processed using wavelet transforms, are used. The use of the genetic algorithm (GA) for feature selection and its integration with the KNN classifier is highlighted. The results reveal that the vowel "a" and the KNN offer an accuracy of 0.9118, positioning them as key elements in the effective detection of the disease. Similarly, in [42] with the use of voice recordings and body movement information of patients, they seek to predict the probability of developing Parkinson's disease. The results show that the KNN model achieved a better performance with 0.88 in accuracy. Finally, the authors of the study [43] make a thorough review of ML algorithms that can help in the prediction of Parkinson's disease, as part of their methodology they employed the SMOTE oversampling technique to generate synthetic values and in training feature selection to optimize the models. The results of the study ranked RF as the best with 0.974 in accuracy.

## 3. METHOD

In this section, we present the methodology of our study, which is divided into two stages. In the first stage, we detail the models (LR, SVM, DT, ETC, KNN, RF, AdaBoost, and GB) of ML selected for Parkinson's detection. In the second stage, we conduct a comprehensive analysis of the dataset, followed by preprocessing and training of the models.

### 3.1. Description of the ML models

#### 3.1.1. Logistic regression

LR models are statistical algorithms that analyze the relationship of variables with another qualitative, dichotomous dependent variable and one or several independent variables [44], [45]. This model is usually used to study the effects of predictor variables on categorical variables, so that the prediction result is usually in binary to confirm the presence or absence of a specific event [46]. It is mainly useful to study the relationship of the state of some disease to determine whether a person is sick or not, so it is widely used in the health field [47], [48]. The mathematical formula of the model can be represented in (1). *Y* represents an event's probability, denoted as *P(Y).*

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

#### 3.1.2. Support vector machines

The SVM model is an algorithm based on statistical learning that focuses mainly on minimizing structural risk; it is an easy-to-use model, reliable in predictions, and fast in data processing [49]. In the model, use is made of a hypothesis space composed of linear functions within a high-dimensional feature space, and these functions are trained using optimization algorithms that apply a learning bias derived from the fundamentals of statistical learning theory [50]. SVM significantly decreases the most common ML problems such as optimization problems, and it is in this aspect that the model excels relative to others [51]. The main objective of the model is to predict the objective values with the use of test data and certain variables [52], [53]. The model can be expressed in (2) and (3). In equation $yi$, which represents the sample class label, the vector of weights W, the feature vector x, the bias b, and the sample size n are used.

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

Subject to:

$$y_i(wx+b)-1 \geq 0, i = 1 \ldots .n, \tag{3}$$

### 3.1.3. Decision tree

The DT model is a commonly used classification and prediction algorithm employed in early statistical algorithms [54]. DT is widely used to develop classification models since this model has similarities with human reasoning and is very easy to understand [55]. Moreover, it is not only known for its multiple fields of applications but also for its robustness and interpretability [56]. In data classification, DT employs the "divide and conquer" technique, whereby the model is dedicated to identifying features and establishing patterns in large datasets, to facilitate the selection and predictive modeling process [57]. The highlight of decision trees is their ability to separate data sets into recursive subsets according to the values of related input fields or predictors, these separations result in descendant nodes, which are known as leaves or end nodes [58], [59]. In (4) the mathematical equation of the model is expressed. Within the equation, *s* is used to represent the sample, *E* is interpreted as the entropy, *Pn* is used to express the probability of not occurring, and *Py* is used to express the probability of occurrence.

$$E(s) = \sum_{k=0}^{n} \binom{n}{k} - Py * \log 2Pn, \tag{4}$$

### 3.1.4. Extra trees classifier

The ETC model is an algorithm that constructs multiple DTs in its training stage, to produce the class that is most frequent among the classes (classification) or the average prediction value (regression) of the individual trees as a result of prediction [60]. The model is similar to RF, but they differ in the way of splitting the tree nodes, since ETC selects features from the dataset randomly, and the best split of a random subset is chosen, while RF, employs all features without distinction [61], [62]. The ETC algorithm has found applications in a variety of areas, including medical diagnosis, wind speed forecasting, and diabetes detection [63]. The architecture of the ETC model is presented in Figure 1.

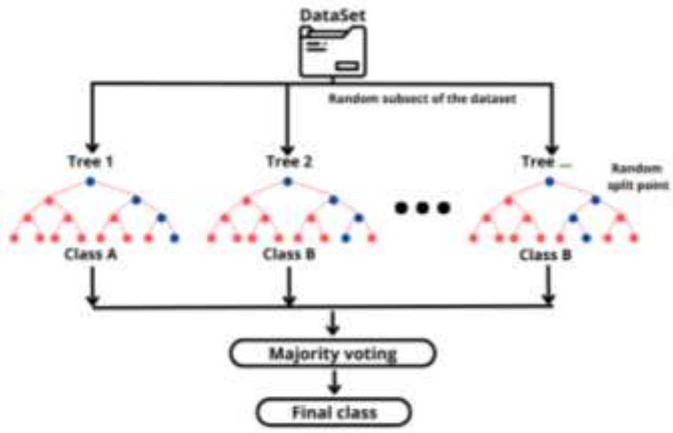


Figure 1. Architecture of the ETC model

### 3.1.5. K-nearest neighbors

The KNN model is a simple algorithm and is widely used in many scientific fields [64]. It is a non-parametric model that does not make any assumptions about the distribution of the data, so it is widely used in regression analysis and data classification [65]. Moreover, it is based on identifying the k nearest k data points within the training set, after that, it performs the prediction or classification of the data based on the majority class of the nearest neighbors [66]. To measure the distance between its neighbors, the model uses the Euclidean equation for continuous variables, as shown in (5), and for discrete variables, it uses the overlap metric [67], [68] The model is widely used in the prediction of some diseases such as type 2 diabetes or for movie recommendations [69], [70].

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ri} - x_{rj})^2}, \quad (5)$$

### 3.1.6. Random forest

The RF model consists of using regression trees that make use of bootstrap resampling and randomization of predictors, leading to high prediction accuracy [71]. RF is a more comprehensive version of the DT model, as it uses multiple classifiers to achieve better accuracy instead of a single classifier [72]. The model is one of the most popular within ML, as it can account for correlation and interactions between features [73]. In regression, the RF model averages the predictions of each tree, but in classification, it accumulates the majority class votes using the class information provided by the individual trees [74]. As shown in (6) shows the formula that the model uses to estimate the predictions of each tree [75]. Where $E_\theta$ refers to the expectation with respect to the random parameter, conditional on $X$ and the data set $D_n$.

$$\bar{r_n}(X, D_n) = E_\theta[r_n(X, \theta, D_n)], \quad (6)$$

### 3.1.7. Adaptive boosting

The AdaBoost model is an algorithm that uses multiple weak learners to create a strong learner, which is commonly used for data classification and regression [76]. The model works by interactively training the weak learners and adjusting the misclassified weights, so that the final prediction is made by weighting the results of the learners' predictions [77]. On the other hand, AdaBoost has been integrated with other ML models to improve their performance, such as, for example, boosting the accuracy of DT [78]. The model equation is detailed in (7). Where $F_T(x)$ expresses the final prediction of $x$, $f_t(x)$ describes the low-power model prediction, $\hat{y}$ is used to represent the number of low-power models and $\alpha_t$ refers to the weight coefficient.

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

### 3.1.8. Gradient boosting

The GB model belongs to one of the most powerful ML families due to its wide range of practical applications [79]. In addition, it focuses on speed and accuracy [80]. GB is used for both regression and classification and like AdaBoost, it uses weak learners to create a stronger one [81]. The model is optimized in function space with the use of gradient, as it is based on Friedman's statistical development [82]. The model equation can be expressed in (8). Where *f(x)* represents the prediction function, $\hat{y}$ denotes the final model accuracy, γ is the learning coefficient and *h(x)* corresponds to the prediction of the i-th least robust model.

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

## 3.2. Case study

### 3.2.1. Understanding the dataset

The Oxford University dataset for Parkinson's disease detection was used to train the models. The dataset stores 195 records with 23 attributes recording various biomedical measurements of the voice, all presented in ASCII format. The variables stored name (patient name and record), mean fundamental vocal frequency (MDVP: Fo(Hz)), maximum fundamental vocal frequency (MDVP: Fhi(Hz)), minimum fundamental vocal frequency (MDVP: Flo(Hz)), a measure of the tonal components of speech (NHR), measure assessing the relationship between harmonics and noise in speech (HNR), the fractal dimension of the speech signal (DFA), two measures of dynamic complexity in the speech signal (RPDE and D2), three measures assessing the variation in frequency (Spread1, Spread2, and PPE), measures quantifying the variation of the fundamental frequency in speech (Jitter, Jitter(Abs), RAP, PPQ and Jitter: DDP), measures assessing amplitude variation in the voice (Shimmer, Shimmer(dB), Shimmer: APQ3, Shimmer: APQ5, MDVP: APQ, Shimmer: DDA) and status ("1" represents Parkinson's and "0" indicates that the subject is healthy). The development process of the study is detailed in Figure 2.

### 3.2.2. Data preparation

As an initial step before proceeding with the training of our models, we performed an analysis of the characteristics of the dataset. First, we imported the necessary libraries to perform an exploratory analysis, where we verified the attribute names. We also examined the data types contained in each attribute, as presented in Table 1. We observed that the dataset includes one object-type attribute, one integer-type attribute, and 22 float-type attributes. This analysis allowed us to confirm the absence of missing values. Finally, we inspected both the unique values and the information stored in the dataset, as shown in Table 2.

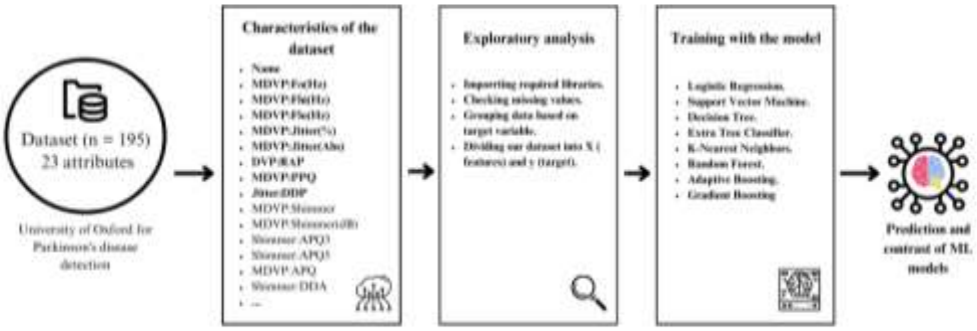


Figure 2. Case study development process

Table 1. Summary information of the data set

| # | Column | Non-null | Count | Dtype |
|---|---|---|---|---|
| 0 | name | non-null | 195 | object |
| 1 | MDVP:Fo(Hz) | non-null | 195 | float64 |
| 2 | MDVP:Fhi(Hz) | non-null | 195 | float64 |
| 3 | MDVP:Flo(Hz) | non-null | 195 | float64 |
| 4 | MDVP:Jitter(%) | non-null | 195 | float64 |
| 5 | MDVP:Jitter(Abs) | non-null | 195 | float64 |
| 6 | MDVP: RAP | non-null | 195 | float64 |
| 7 | MDVP: PPQ | non-null | 195 | float64 |
| 8 | Jitter: DDP | non-null | 195 | float64 |
| 9 | MDVP: Shimmer | non-null | 195 | float64 |
| 10 | MDVP:Shimmer(dB) | non-null | 195 | float64 |
| 11 | Shimmer: APQ3 | non-null | 195 | float64 |
| 12 | Shimmer: APQ5 | non-null | 195 | float64 |
| 13 | MDVP: APQ | non-null | 195 | float64 |
| 14 | Shimmer: DDA | non-null | 195 | float64 |
| 15 | NHR | non-null | 195 | float64 |
| 16 | HNR | non-null | 195 | float64 |
| 17 | Status | non-null | 195 | int64 |
| 18 | RPDE | non-null | 195 | float64 |
| 19 | DFA | non-null | 195 | float64 |
| 20 | Spread1 | non-null | 195 | float64 |
| 21 | Spread2 | non-null | 195 | float64 |
| 22 | D2 | non-null | 195 | float64 |
| 23 | PPE | non-null | 195 | float64 |
| Dtypes: float64(22), int64(1), object(1) | | | | |
| Memory usage: 36.7+ KB | | | | |

Table 2. Content of the data set

| | 0 | 1 | 2 | ... | 192 | 193 | 194 |
|---|---|---|---|---|---|---|---|
| Name | phon_R01_S01_1 | phon_R01_S01_2 | phon_R01_S01_3 | ... | phon_R01_S50_4 | phon_R01_S50_5 | phon_R01_S50_6 |
| MDVP:Fo(Hz) | 119.992 | 122.4 | 116.682 | ... | 174.688 | 198.764 | 214.289 |
| MDVP:Fhi(Hz) | 157.302 | 148.65 | 131.111 | ... | 240.005 | 396.961 | 260.277 |
| MDVP:Flo(Hz) | 74.997 | 113.819 | 111.555 | ... | 74.287 | 74.904 | 77.973 |
| MDVP:Jitter(%) | 0.00784 | 0.00968 | 0.0105 | ... | 0.0136 | 0.0074 | 0.00567 |
| MDVP:Jitter(Abs) | 0.00007 | 0.00008 | 0.00009 | ... | 0.00008 | 0.00004 | 0.00003 |
| MDVP:RAP | 0.0037 | 0.00465 | 0.00544 | ... | 0.00624 | 0.0037 | 0.00295 |
| MDVP:PPQ | 0.00554 | 0.00696 | 0.00781 | ... | 0.00564 | 0.0039 | 0.00317 |
| Jitter: DDP | 0.01109 | 0.01394 | 0.01633 | ... | 0.01873 | 0.01109 | 0.00885 |
| MDVP: Shimmer | 0.04374 | 0.06134 | 0.05233 | ... | 0.02308 | 0.02296 | 0.01884 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| Shimmer: DDA | 0.06545 | 0.09403 | 0.0827 | ... | 0.03804 | 0.03794 | 0.03078 |
| NHR | 0.02211 | 0.01929 | 0.01309 | ... | 0.10715 | 0.07223 | 0.04398 |
| HNR | 21.033 | 19.085 | 20.651 | ... | 17.883 | 19.02 | 21.209 |
| Status | 1 | 1 | 1 | ... | 0 | 0 | 0 |
| RPDE | 0.414783 | 0.458359 | 0.429895 | ... | 0.407567 | 0.451221 | 0.462803 |
| DFA | 0.815285 | 0.819521 | 0.825288 | ... | 0.655683 | 0.643956 | 0.664357 |
| Spread1 | -4.813031 | -4.075192 | -4.443179 | ... | -6.787197 | -6.744577 | -5.724056 |
| Spread2 | 0.266482 | 0.33559 | 0.311173 | ... | 0.158453 | 0.207454 | 0.190667 |
| D2 | 2.301442 | 2.486855 | 2.342259 | ... | 2.679772 | 2.138608 | 2.555477 |
| PPE | 0.284654 | 0.368674 | 0.332634 | ... | 0.131728 | 0.123306 | 0.148569 |
| PPE | 0.284654 | 0.368674 | 0.332634 | ... | 0.131728 | 0.123306 | 0.148569 |

### 3.2.3. Data preparation

In Figure 3 it is evident that the data set presents a significant imbalance in the target variable "status". It is observed that there is a predominance of records corresponding to patients diagnosed with Parkinson's, representing 75.38% of the total, compared to 24.62% of records belonging to individuals who do not suffer from the disease. This imbalance may hurt the performance of ML models, so methods to mitigate this drawback should be evaluated.

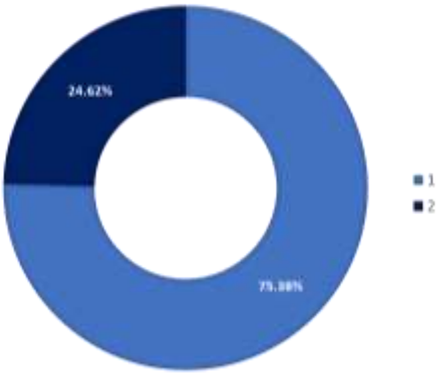


Figure 3. Target variable "status"

Figure 4 shows the correlations between the attributes and the target variable. In this graph, shorter bars indicate lower correlations, while longer bars represent higher correlations. The attributes MDVP: Fo(Hz), MDVP: Flo(Hz), HNR, and MDVP: Fhi(Hz) show the lowest correlations, with values close to or below -0.3. On the other hand, the attributes spread2, PPE, and spread1 exhibit some of the strongest correlations, with values of 4, 5, and 6, respectively, along with the target variable "status" with 10. The remaining attributes show correlations in the range of levels 2 and 3, indicating a strong relationship with the target variable.

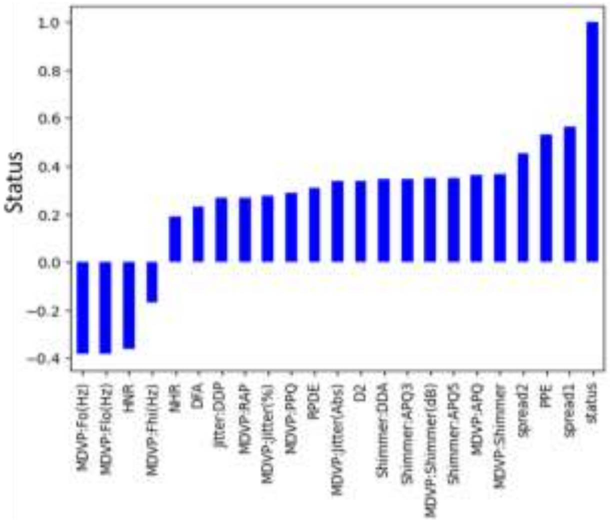
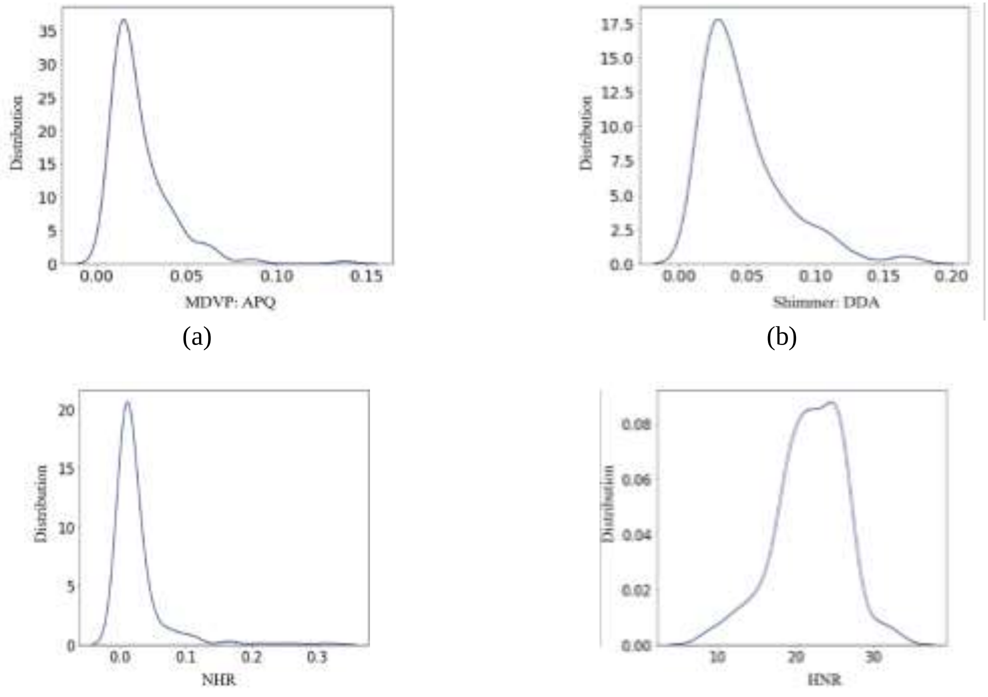


Figure 4. Correlations of different attributes with the target variable "status"

On the other hand, Figure 5 plots the distribution of several continuous variables relevant to our analysis. In particular, Figure 5(a) shows the distribution of the MDVP: APQ attribute, which quantifies the mean of the amplitude differences between successive peaks in the speech signal. In the plot, a leftward skew in the distribution of this attribute is visible, indicating a trend toward lower values. Similarly, in Figure 5(b), we plot the distribution of the Shimmer: DDA attribute, which measures the variation in the amplitude of the speech signal based on the average distance between local maxima and minima in the signal. Again, we can observe a leftward skew in this distribution, suggesting a prevalence of lower values in this attribute. Continuing with the analysis of Figure 5(c), we note that the distribution of the NHR attribute also shows a leftward bias, in line with the trends observed in the previous attributes. In contrast, in Figure 5(d), we present the distribution of the HNR attribute, which shows a distribution that roughly resembles a normal distribution, with a symmetry around its mean. These distribution patterns are essential for understanding the variability of these variables in our data set and can provide valuable information.



(a) (b)

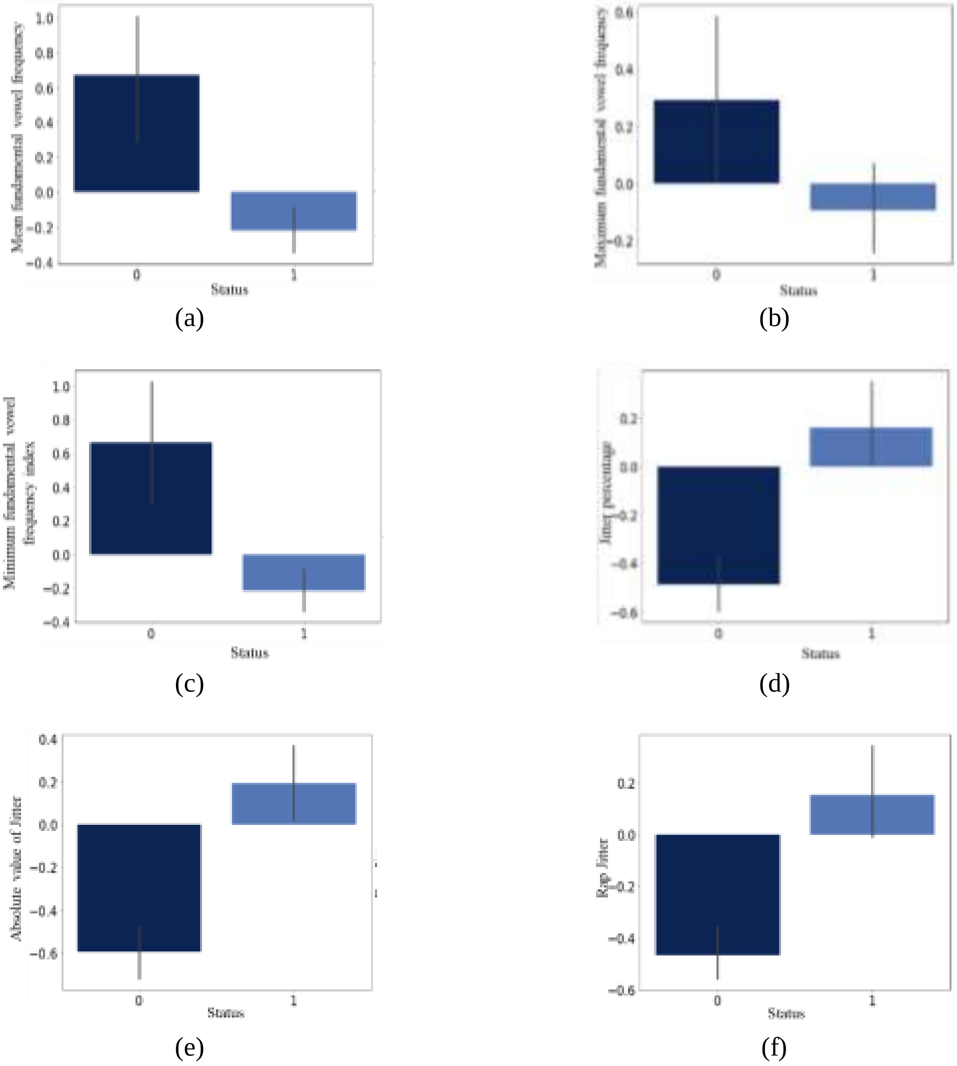
(c) (d)

Figure 5. Distribution of the continuous functions: (a) distribution of the "MDVP: APQ" attribute, (b) distribution of the "Shimmer: DDA" attribute, (c) distribution of the "NHR" attribute, and (d) distribution of the "HNR" attribute

In Figure 6, we illustrate the relationship between the continuous attributes and the target variable to identify the factors that influence the probability of developing Parkinson's disease. In Figure 6(a), we observe that as the mean vocal fundamental frequency (MDVP: Fo(Hz)) increases, the probability of developing Parkinson's decreases, as fewer positive "1" cases related to this attribute are recorded. Similarly, in Figure 6(b), we can appreciate that as the maximum fundamental vowel frequency (MDVP: Fhi(Hz)) increases, the probability of suffering from Parkinson's is practically zero, since the positive cases related to this attribute are few less than "0". Similarly, according to Figure 6(c), an increase in the minimum fundamental vocal frequency index (MDVP: Flo(Hz)) is associated with a minimal probability of developing Parkinson's in patients. On the other hand, in Figure 6(d), it is observed that an increase in the Jitter percentage (MDVP: Jitter(%)), which measures the variation in the period between consecutive cycles of the speech signal, correlates with an increase in the likelihood of developing Parkinson's disease. This is because the data set contains a higher number of cases related to this attribute and positive cases (greater than "0"). Furthermore, according to Figure 6(e), an increase in the absolute value of Jitter (MDVP: Jitter(Abs)) translates into a higher probability of developing Parkinson's disease, as more positive cases (0.2) related to this attribute are recorded. Likewise, according to Figure 6(f), an increase in fast Jitter (MDVP: RAP) is associated with an increased likelihood of developing the disease (0.1), making it a relevant factor in clinical diagnosis.

Likewise, in Figure 7 we analyze the influence of two attributes with the target variable "status" to see the probabilities that a patient may develop Parkinson's disease. In Figure 7(a) we check the influence of the attributes MDVP: Fo(Hz) and PPE on the variable "status". In this plot, we note that 8 Parkinson's cases with positive diagnoses were found to be associated with low MDVP: Fo(Hz) values, ranging up to 100 Hz while presenting PPE in the range of 0.19 to 0.31. Furthermore, multiple instances of Parkinson's disease are observed to have relationships with medium and high MDVP: Fo(Hz) levels and a PPE in the range of 0.1 to

0.45. On the other hand, Figure 7(b) identified 8 instances of low MDVP: Fo(Hz) up to 100 Hz with spread1 (from -6 to -4.5 approximately) are related to the probability of Parkinson's disease. Also, several cases of Parkinson's are visible for mean with spread1 (from -7 to -3 approximately). In the case of very high MDVP: Fo(Hz), some Parkinson's cases are also observed together with spread1 (from -6.5 to -5, approximately). Similarly, according to Figure 7(c), no case of low PPE with spread1 is related to Parkinson's disease. In addition, numerous Parkinson's cases with positive diagnosis can be seen in the high PPE category, with spread1 varying between -6.5 and -3.5. This area shows a considerable density of Parkinson's disease cases.



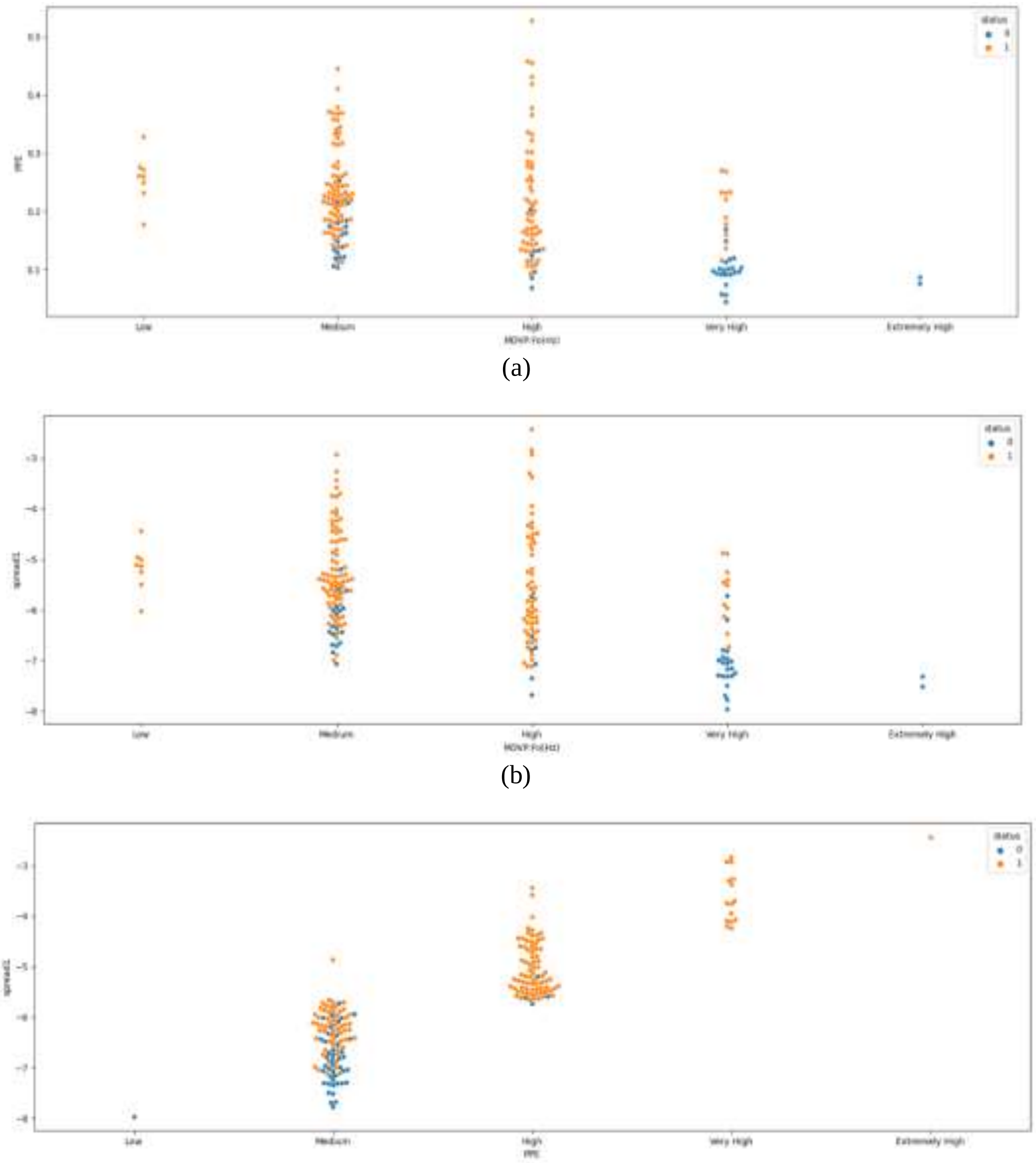
(a) (b)

(c) (d)

(e) (f)

Figure 6. Relationship of continuous attributes with the target variable: (a) mean fundamental vowel frequency with the target variable, (b) maximum fundamental vowel frequency with the target variable, (c) minimum fundamental vowel frequency index with the target variable, d) Jitter percentage with the target variable, (e) absolute value of Jitter with the target variable, and (f) rap Jitter with the target variable

In Figure 8, a comprehensive analysis of the correlation of the target variable "status" concerning the other attributes was carried out. Through this analysis, we sought to identify significant relationships and to identify the influence of various attributes on the target variable. In the graph generated, it is observed that several attributes present a significant correlation with the "status" variable. Among the attributes that show a strong correlation are PPE, D2, spread1, spread2, RPDE, MDVP: Shimmer, MDVP: Shimmer(dB), Shimmer: APQ3, Shimmer: APQ5, MDVP: APQ, Shimmer: DDA, MDVP: Jitter(%) and MDVP: Jitter(Abs). Furthermore, it is important to note that strong multicollinearity was observed among several of the independent variables analyzed. Therefore, it is likely that we will have to discard some of them.



(a)

(b)

(c)

Figure 7. Influence of two attributes with the target variable to determine the probability of having Parkinson's: (a) testing the influence of "MDVP: Fo(Hz)" and "PPE" with "status", (b) testing the influence of "MDVP: Fo(Hz)" and "spread1" with "status", and (c) testing the influence of "PPE" and "spread1" with "status"

### 3.2.4. Data processing and modeling

Before training the models, we have to process and clean the dataset so that the algorithms can achieve better performance. First, we divide our dataset into two main groups: one called "Y," which stores the target variable, and another called "X," which contains all the other attributes. The "X" group consists of 195 records and 22 different attributes, while the "Y" group includes 195 records with the corresponding target variable. We then proceeded to divide these groups into training and test sets in a proportion of 80% and 20%, respectively. This division allows us to use 80% of the data to train our models and the remaining 20% to evaluate their performance. In addition to data partitioning, we implemented a feature scaling process. This technique is fundamental to standardizing the attributes, meaning that we transform their values so that they have a mean of zero and a standard deviation of one. This is important to ensure that differences in attribute scaling do not negatively affect the performance of our models. We then proceeded to train ML models focused on the early detection of Parkinson's disease.

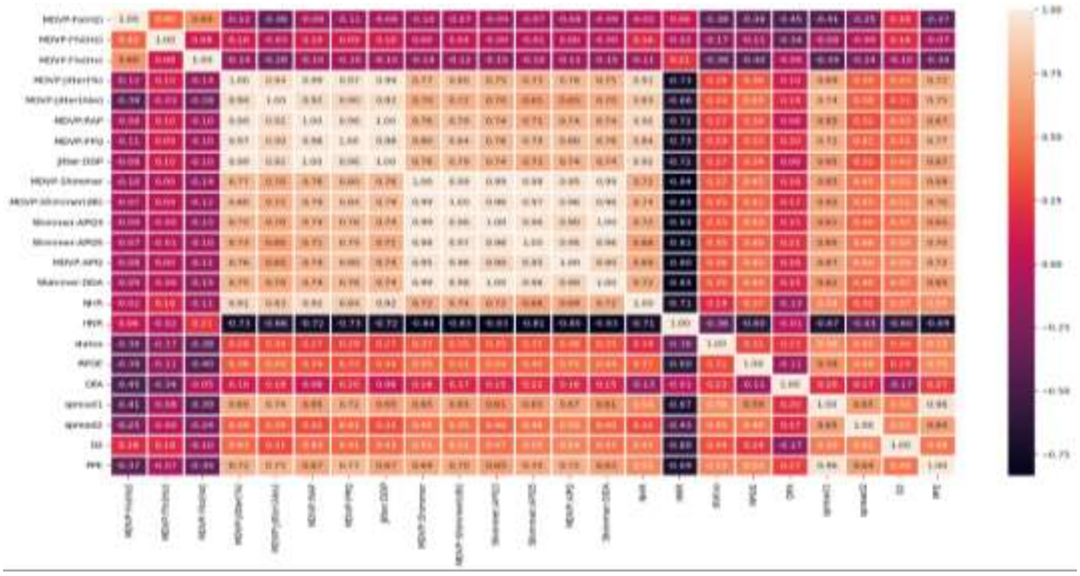


Figure 8. Correlation of characteristics

## 4. RESULTS

After performing a thorough preprocessing process of the dataset provided by the University of Oxford for Parkinson's disease detection, we proceeded to train several ML models to identify the most effective one based on the metrics of precision, accuracy, sensitivity, and F1 score. The models we trained included LR, SVM, DT, ETC, KNN, RF, and AdaBoost, as well as GB. The results of this training are detailed in Table 3, which provides a comprehensive view of each model's performance relative to the aforementioned metrics.

Table 3. Model training results

| Logistic regression | | | | | K neighbor classifier | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | F1-score (%) | Recall (%) | Precision (%) | Support | | F1-score (%) | Recall (%) | Precision (%) | Support |
| 0 | 0.67 | 0.57 | 0.80 | 7 | … | 0.86 | 0.86 | 0.86 | 7 |
| 1 | 0.94 | 0.97 | 0.91 | 32 | … | 0.97 | 0.97 | 0.97 | 32 |
| Macro avg | 0.80 | 0.77 | 0.86 | 39 | … | 0.91 | 0.91 | 0.91 | 39 |
| Weighted avg | 0.89 | 0.90 | 0.89 | 39 | … | 0.95 | 0.95 | 0.95 | 39 |
| Accuracy | 0.90 | | | 39 | … | 0.95 | | | 39 |
| ROC AUC score: 0.7700892857142857 | | | | | ROC AUC score: 0.9129464285714287 | | | | |
| Support vector classifier | | | | | Random forest | | | | |
| | F1-score (%) | Recall (%) | Precision (%) | Support | … | F1-score (%) | Recall (%) | Precision (%) | Support |
| 0 | 0.62 | 0.57 | 0.67 | 7 | … | 0.71 | 0.71 | 0.71 | 7 |
| 1 | 0.92 | 0.94 | 0.91 | 32 | … | 0.94 | 0.94 | 0.94 | 32 |
| Macro avg | 0.77 | 0.75 | 0.79 | 39 | … | 0.83 | 0.83 | 0.83 | 39 |
| Weighted avg | 0.87 | 0.87 | 0.87 | 39 | … | 0.90 | 0.90 | 0.90 | 39 |
| Accuracy | 0.87 | | | 39 | | 0.90 | | | 39 |
| ROC AUC score: 0.7544642857142856 | | | | | ROC AUC score: 0.8258928571428571 | | | | |
| Decision tree | | | | | AdaBoost | | | | |
| | F1-score (%) | Recall (%) | Precision (%) | Support | … | F1-score (%) | Recall (%) | Precision (%) | Support |
| 0 | 0.53 | 0.71 | 0.42 | 7 | … | 0.50 | 0.57 | 0.44 | 7 |
| 1 | 0.85 | 0.78 | 0.93 | 32 | … | 0.87 | 0.84 | 0.90 | 32 |
| Macro avg | 0.69 | 0.75 | 0.67 | 39 | … | 0.69 | 0.71 | 0.67 | 39 |
| Weighted avg | 0.79 | 0.77 | 0.83 | 39 | … | 0.80 | 0.79 | 0.82 | 39 |
| Accuracy | 0.77 | | | 39 | | 0.79 | | | 39 |
| ROC AUC score: 0.7477678571428572 | | | | | ROC AUC score: 0.7075892857142857 | | | | |
| Extra tree classifier | | | | | Gradient boosting | | | | |
| | F1-score (%) | Recall (%) | Precision (%) | Support | … | F1-score (%) | Recall (%) | Precision (%) | Support |
| 0 | 0.31 | 0.29 | 0.33 | 7 | … | 0.71 | 0.71 | 0.71 | 7 |
| 1 | 0.86 | 0.88 | 0.85 | 32 | … | 0.94 | 0.94 | 0.94 | 32 |
| Macro avg | 0.58 | 0.58 | 0.59 | 39 | … | 0.83 | 0.83 | 0.83 | 39 |
| Weighted avg | 0.76 | 0.77 | 0.76 | 39 | … | 0.90 | 0.90 | 0.90 | 39 |
| Accuracy | 0.77 | | | 39 | | 0.90 | | | 39 |
| ROC AUC score: 0.5803571428571428 | | | | | ROC AUC score: 0.8258928571428571 | | | | |

The LR, SVM, DT, ETC, KNN, RF, AdaBoost, and GB models achieved an accuracy of 89%, 87%, 83%, 76%, 95%, 90%, 82%, and 90% respectively. Similarly, they achieved an accuracy of 90%, 87%, 77%, 77%, 77%, 95%, 90%, 79% and 90% respectively. According to Table 3, most of the trained models have achieved exceptional performance, such as the KNN model, which achieved the best metrics in precision, accuracy, sensitivity and F1 count with 95%, 95%, 95%, 95%, and 95% respectively, making it the best model for early Parkinson's detection. In the second place, we have the RF and GB models with 90% accuracy, 90% in precision, 90% in sensitivity, and 90% in F1 count. In third place is the LR model with 89% in precision, 90% in accuracy, 90% in sensitivity, and 89% in F1 count. Followed by the SVM model with 87% accuracy, 87% precision, 87% accuracy, 87% sensitivity, and 87% F1 count. The DT model achieved fifth place with 83% precision, 77% accuracy, 77% sensitivity and 79% F1 count. The AdaBoost models achieved 82% precision and 79% accuracy. Finally, the ETC model recorded the worst metrics in Parkinson's detection with 76% in precision and 77% in accuracy.

## 5. DISCUSSION

Parkinson's disease, as a neurodegenerative condition, affects large numbers of people worldwide, having a significant impact on quality of life. Its diagnosis remains a challenge, with causes still largely unknown and definitive confirmation required by autopsy. The primary purpose of this study was to conduct a comparative analysis of ML models aimed at early detection of this disease.

The training of the models was performed with the Oxford University dataset for Parkinson's disease detection, which stores 195 records with 23 attributes on voice, presented in ASCII format. After applying optimization methods and techniques such as feature selection and scaling, the ML models were trained. The LR, SVM, DT, ETC, KNN, RF, AdaBoost, and GB models achieved an accuracy of 89%, 87%, 83%, 76%, 95%, 90%, 82%, and 90% respectively. They also achieved an accuracy of 90%, 87%, 77%, 77%, 77%, 95%, 90%, 90%, 79% and 90% respectively. The model that achieved the best performance was KNN with 95% in the metrics of precision, accuracy, sensitivity, and F1 count. On the contrary, studies [36], [41], [42] obtained lower results than this study with the same model, achieving 88.33%, 91.18% and 88% accuracy, respectively. Differentiating mainly with the studies [36], [41], where they used a different data set than the one used in this study.

On the other hand, the RF and GB models achieved second place with precision, sensitivity, accuracy, and F1 count metrics of 90%. Similar to [38], where they applied feature selection for data processing, the RF model achieved 90.5% precision and 97.3% accuracy. In contrast, [30] used the same dataset from Oxford University, but achieved 85% in precision and 87% in accuracy with the RF model, which are lower than those obtained in this study. One of the most marked differences is the use of optimization techniques, since in [30] they did not employ feature selection. In contrast, [43] applied the SMOTE oversampling technique to generate synthetic data and achieved 97.4% accuracy with the RF model, which is a higher result than that obtained in this study.

In the case of the GB model, in the study [35] the model managed to obtain precision and accuracy of 79% with the use of optimization techniques such as Boruta, recursive feature elimination (RFE), and RF classifier, but achieving lower results than those of this study about the GB model. Likewise, the LR model achieved third place with 89% accuracy and 90% precision. Similar to those obtained in [30], where LR obtained 83.05% in precision and 83% in accuracy, using the same data set, but with different optimization techniques. In contrast to [28], with the use of a dataset on the non-motor and olfactory characteristics of the patients, the model achieved 97.159% accuracy, being superior results to those obtained in this study. In the case of the SVM model, a performance of 87% in precision and accuracy was achieved in this study, similar to that achieved in [33], where the model achieved a result of 77.5% in accuracy with the use of voice recordings with short and long sentences. In contrast, in [29], [32] the model achieved a better performance with 95% and 91.4%, with the main difference in the dataset used for training the models. For its part, the DT model achieved 83% accuracy and 77% precision, lower results than those achieved in studies [38], [40] where the model achieved 90.5% and 94.77% accuracy, respectively.

The AdaBoost model obtained a precision of 82% and an accuracy of 79%. Similar to the study [30], where the model achieved 87% precision and 86.4% accuracy. Different from those achieved in [37], where with a different data set than the one used in this study, the AdaBoost model achieved 97.4% accuracy. Finally, the ETC model achieved 76% in precision and 77% in accuracy, similar to [35], since the model achieved 73% in precision and 75% in accuracy. As we have been able to observe the results of this study agree with most of the related studies, in some cases even surpassing them.

Based on the findings, it is clear that ML models, in particular KNN, RF and GB, demonstrate a remarkable ability for early detection of Parkinson's disease, and would help to improve patients' long-term quality of life. However, many times the performance of the models is conditioned with the dataset used and the optimization techniques and methods applied to the data. These results support the future utility of ML models in the early detection of neurodegenerative diseases, but also emphasize the need for further research and the development of representative data sets. This study contributes to the understanding of the capabilities and limitations of ML models in the early detection of Parkinson's disease. Continued improvement of these methodologies, along with critical consideration of dataset characteristics, will guide future research toward more accurate and effective approaches to address this challenging health problem.

## 6. CONCLUSION

Parkinson's disease has had a rapid increase in incidence and mortality rates in the last two decades, so finding an early detection method could improve patients' lives. This study aimed to perform a comparative analysis of ML models focused on early detection of Parkinson's disease. Therefore, LR, SVM, DT, ETC, KNN, RF, AdaBoost, and GB algorithms were developed and analyzed to identify the one that offers the best performance. The Oxford University dataset for Parkinson's disease detection was used to train these models. This dataset, as well as related work, had patient voice information, including single words, short sentences, and audio frequencies. The results rank the KNN model as the best with 95% on the metrics of precision, accuracy, sensitivity, and F1 count.

In addition, during the correlation analysis of the data, we noticed that, depending on some particular situations, the probability of suffering from Parkinson's disease increases or decreases. As in the case of Jitter values, when there is an increase in the frequencies recorded by this attribute, the probability of suffering from the disease increases dramatically, since this factor is present in most cases with a positive diagnosis of Parkinson's disease. On the other hand, the MDVP frequencies are also determinant, since when the frequencies increase the probability of Parkinson's decreases. Therefore, these attributes, as well as others that were analyzed, could be determinant in the early identification of Parkinson's disease.

Finally, the models have proven to be a reliable method for Parkinson's detection, so they could be used in clinical trials. However, their effectiveness is conditioned by the data set used and the optimization techniques employed. Therefore, in the future, it would be a priority to contrast the models of this study with other datasets and other optimization techniques to determine the best model for early detection of Parkinson's disease.

## REFERENCES

[1] V. L. Feigin *et al.*, “Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016,” *The Lancet Neurology*, vol. 18, no. 5, pp. 459–480, May 2019, doi: 10.1016/S1474-4422(18)30499-X.

[2] E. R. Dorsey, T. Sherer, M. S. Okun, and B. R. Bloem, “The emerging evidence of the Parkinson pandemic,” *Journal of Parkinson’s Disease*, vol. 8, no. s1, pp. S3–S8, Dec. 2018, doi: 10.3233/JPD-181474.

[3] G. Deuschl *et al.*, “The burden of neurological diseases in Europe: an analysis for the Global Burden of Disease Study 2017,” *The Lancet Public Health*, vol. 5, no. 10, pp. e551–e567, Oct. 2020, doi: 10.1016/S2468-2667(20)30190-0.

[4] E. Benmalek, J. Elmhamdi, and A. Jilbab, “Voice assessments for detecting patients with Parkinson’s diseases in different stages,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 6, pp. 4265–4271, Dec. 2018, doi: 10.11591/ijece.v8i6.pp4265-4271.

[5] A. Samii, J. G. Nutt, and B. R. Ransom, “Parkinson’s disease,” *The Lancet*, vol. 363, no. 9423, pp. 1783–1793, May 2004, doi: 10.1016/S0140-6736(04)16305-8.

[6] A. J. Lees, J. Hardy, and T. Revesz, “Parkinson’s disease,” *The Lancet*, vol. 373, no. 9680, pp. 2055–2066, Jun. 2009, doi: 10.1016/S0140-6736(09)60492-X.

[7] B. R. Bloem, M. S. Okun, and C. Klein, “Parkinson’s disease,” *The Lancet*, vol. 397, no. 10291, pp. 2284–2303, Jun. 2021, doi: 10.1016/S0140-6736(21)00218-X.

[8] L. V Kalia and A. E. Lang, “Parkinson’s disease,” *The Lancet*, vol. 386, no. 9996, pp. 896–912, Aug. 2015, doi: 10.1016/S0140-6736(14)61393-3.

[9] A. H. Rajput, “Frequency and cause of Parkinson’s disease,” *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques*, vol. 19, no. S1, pp. 103–107, Feb. 1992, doi: 10.1017/S0317167100041457.

[10] M. C. de Rijk *et al.*, “Prevalence of Parkinson’s disease in Europe: a collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group,” *Neurology*, vol. 54, no. 11, pp. 21–23, Jun. 2000.

[11] M. Baldereschi *et al.*, “Parkinson’s disease and parkinsonism in a longitudinal study,” *Neurology*, vol. 55, no. 9, pp. 1358–1363, Nov. 2000, doi: 10.1212/WNL.55.9.1358.

[12] B. C. L. Lai, M. Schulzer, S. Marion, K. Teschke, and J. K. C. Tsui, “The prevalence of Parkinson’s disease in British Columbia, Canada, estimated by using drug tracer methodology,” *Parkinsonism & Related Disorders*, vol. 9, no. 4, pp. 233–238, Mar. 2003, doi: 10.1016/S1353-8020(02)00093-7.

[13] R. Inzelberg, E. Schechtman, and D. Paleacu, “Onset age of Parkinson disease,” *American Journal of Medical Genetics*, vol. 111, no. 4, pp. 459–460, Sep. 2002, doi: 10.1002/ajmg.10586.

[14] U. B. Muthane, H. S. Swamy, P. Satishchandra, M. N. Subhash, S. Rao, and D. Subbakrishna, “Early onset Parkinson’s disease: Are juvenile- and young-onset different?,” *Movement Disorders*, vol. 9, no. 5, pp. 539–544, Jan. 1994, doi: 10.1002/mds.870090506.

[15] D. B. Calne, B. J. Snow, and C. Lee, “Criteria for diagnosing Parkinson’s disease,” *Annals of Neurology*, vol. 32, no. S1, pp. S125–S127, 1992, doi: 10.1002/ana.410320721.

[16] L. M. de Lau and M. M. Breteler, "Epidemiology of Parkinson's disease," *The Lancet Neurology*, vol. 5, no. 6, pp. 525–535, Jun. 2006, doi: 10.1016/S1474-4422(06)70471-9.

[17] J. Benito-León *et al.*, "Incidence of Parkinson disease and parkinsonism in three elderly populations of central Spain," *Neurology*, vol. 62, no. 5, pp. 734–741, Mar. 2004, doi: 10.1212/01.WNL.0000113727.73153.68.

[18] L. M. L. de Lau, P. C. L. M. Giesbergen, M. C. de Rijk, A. Hofman, P. J. Koudstaal, and M. M. B. Breteler, "Incidence of parkinsonism and Parkinson disease in a general population," *Neurology*, vol. 63, no. 7, pp. 1240–1244, Oct. 2004, doi: 10.1212/01.WNL.0000140706.52798.BE.

[19] R. Mayeux *et al.*, "The frequency of idiopathic Parkinson's disease by age, ethnic group, and Sex in Northern Manhattan, 1988–1993," *American Journal of Epidemiology*, vol. 142, no. 8, pp. 820–827, Oct. 1995, doi: 10.1093/oxfordjournals.aje.a117721.

[20] R. C. Chen *et al.*, "Prevalence, incidence, and mortality of PD: a door-to-door survey in Ilan county," *Neurology*, vol. 57, no. 9, pp. 1679–1686, Nov. 2001, doi: 10.1212/WNL.57.9.1679.

[21] A. C. Jenkins, "Epidemiology of parkinsonism in Victoria," *Medical Journal of Australia*, vol. 2, no. 11, pp. 496–502, Sep. 1966, doi: 10.5694/j.1326-5377.1966.tb97295.x.

[22] S. J. Wang *et al.*, "Parkinson's disease in Kin-Hu, Kinmen: a community survey by neurologists," *Neuroepidemiology*, vol. 13, no. 1–2, pp. 69–74, 1994, doi: 10.1159/000110361.

[23] I. Straw and H. Wu, "Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction," *BMJ Health Care Inform*, vol. 29, no. 1, Apr. 2022, doi: 10.1136/bmjhci-2021-100457.

[24] B. Tuvshinjargal and H. Hwang, "Development of online service for brain disease prediction using machine learning," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, Oct. 2021, pp. 505–508. doi: 10.1109/ICTC52510.2021.9620880.

[25] Y. Lai *et al.*, "Identification of immune microenvironment subtypes and signature genes for Alzheimer's disease diagnosis and risk prediction based on explainable machine learning," *Frontiers in Immunology*, vol. 13, Dec. 2022, doi: 10.3389/fimmu.2022.1046410.

[26] J. Harvey *et al.*, "Machine learning-based prediction of cognitive outcomes in de novo Parkinson's disease," *npj Parkinson's Disease*, vol. 8, no. 1, Nov. 2022, doi: 10.1038/s41531-022-00409-5.

[27] J. H. Park *et al.*, "Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data," *npj Digital Medicine*, vol. 3, no. 1, Mar. 2020, doi: 10.1038/s41746-020-0256-0.

[28] K. N. R. Challa, V. S. Pagolu, G. Panda, and B. Majhi, "An improved approach for prediction of Parkinson's disease using machine learning techniques," in *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, Oct. 2016, pp. 1446–1451. doi: 10.1109/SCOPES.2016.7955679.

[29] R. Alshammri, G. Alharbi, E. Alharbi, and I. Almubark, "Machine learning approaches to identify Parkinson's disease using voice signal features," *Frontiers in Artificial Intelligence*, vol. 6, Mar. 2023, doi: 10.3389/frai.2023.1084001.

[30] K. Sharma and A. Mishra, "Prediction of Parkinson's disease using machine learning techniques," *SSRN Electronic Journal*, 2020, doi: 10.2139/ssrn.3742953.

[31] V. Swaroopa and S. J. Saritha, "Machine learning approches for the classification of Parkinson disease," *International Journal for Modern Trends in Science and Technology*, vol. 7, no. 11, pp. 93–97, 2021.

[32] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, Apr. 2009, doi: 10.1109/TBME.2008.2005954.

[33] B. E. Sakar *et al.*, "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828–834, Jul. 2013, doi: 10.1109/JBHI.2013.2245674.

[34] R. Das, "A comparison of multiple classification methods for diagnosis of Parkinson disease," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1568–1572, Mar. 2010, doi: 10.1016/j.eswa.2009.06.040.

[35] N. Nahar, F. Ara, M. A. I. Neloy, A. Biswas, M. S. Hossain, and K. Andersson, "Feature selection based machine learning to improve prediction of Parkinson disease," in *BI 2021: Brain Informatics*, 2021, pp. 496–508. doi: 10.1007/978-3-030-86993-9_44.

[36] F. Saeed *et al.*, "Enhancing Parkinson's disease prediction using machine learning and feature selection methods," *Computers, Materials & Continua*, vol. 71, no. 3, pp. 5639–5658, 2022, doi: 10.32604/cmc.2022.023124.

[37] M. Kundu, M. A. Nashiry, A. K. Dipongkor, S. Sarmin Sumi, and M. A. Hossain, "An optimized machine learning approach for predicting Parkinson's disease," *International Journal of Modern Education and Computer Science*, vol. 13, no. 4, pp. 68–74, Aug. 2021, doi: 10.5815/ijmecs.2021.04.06.

[38] A. K. Tiwari, "Machine learning based approaches for prediction of Parkinson's disease," *Machine Learning and Applications: An International Journal*, vol. 3, no. 2, pp. 33–39, Jun. 2016, doi: 10.5121/mlaij.2016.3203.

[39] I. Mandal and N. Sairam, "New machine-learning algorithms for prediction of Parkinson's disease," *International Journal of Systems Science*, vol. 45, no. 3, pp. 647–666, Mar. 2014, doi: 10.1080/00207721.2012.724114.

[40] T. Mohesh, K. Gowtham, P. Vijeesh, and S. Arun Kumar, "Parkinsons disease prediction using machine learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 6, pp. 1393–1411, Jun. 2022, doi: 10.22214/ijraset.2022.44075.

[41] N. Benayad, Z. Soumaya, B. D. Taoufiq, and A. Abdelkrim, "Features selection by genetic algorithm optimization with K-nearest neighbour and learning ensemble to predict Parkinson disease," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 2, pp. 1982–1989, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1982-1989.

[42] N. D. Bala and A. S, "Machine learning algorithms for detection of Parkinson's disease using motor symptoms: speech and tremor," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 6, pp. 47–50, Mar. 2020, doi: 10.35940/ijrte.F7129.038620.

[43] N. Chintalapudi, V. R. Dhulipalla, G. Battineni, C. Rucco, and F. Amenta, "Voice biomarkers for Parkinson's disease prediction using machine learning models with improved feature reduction techniques," *Journal of Data Science and Intelligent Systems*, vol. 1, no. 2, pp. 92–98, Apr. 2023, doi: 10.47852/bonviewJDSIS3202831.

[44] S. Domínguez-Almendros, N. Benítez-Parejo, and A. R. Gonzalez-Ramirez, "Logistic regression models," *Allergologia et Immunopathologia*, vol. 39, no. 5, pp. 295–305, Sep. 2011, doi: 10.1016/j.aller.2011.05.002.

[45] Q. Q. Wang *et al.*, "[Overview of logistic regression model analysis and application]," *Zhonghua Yu Fang Yi Xue Za Zhi (Chinese Journal of Preventive Medicine)*, vol. 53, no. 9, pp. 955–960, 2019.

[46] T. G. Nick and K. M. Campbell, "Logistic regression," in *Topics in Biostatistics*, 2007, pp. 273–301. doi: 10.1007/978-1-59745-530-5_14.

[47] E. Boateng and F. Oduro, "Predicting microfinance credit default: A study of Nsoatreman Rural Bank, Ghana," *Journal of Advances in Mathematics and Computer Science*, vol. 26, no. 1, pp. 1–9, Jan. 2018, doi: 10.9734/JAMCS/2018/33569.

[48] E. Y. Boateng and D. A. Abaye, “A review of the logistic regression model with emphasis on medical research,” *Journal of Data Analysis and Information Processing*, vol. 07, no. 04, pp. 190–207, 2019, doi: 10.4236/jdaip.2019.74012.

[49] A. Zendehboudi, M. A. Baseer, and R. Saidur, “Application of support vector machine models for forecasting solar and wind energy resources: A review,” *Journal of Cleaner Production*, vol. 199, pp. 272–285, Oct. 2018, doi: 10.1016/j.jclepro.2018.07.164.

[50] S. Raghavendra. N and P. C. Deka, “Support vector machine applications in the field of hydrology: a review,” *Applied Soft Computing*, vol. 19, pp. 372–386, Jun. 2014, doi: 10.1016/j.asoc.2014.02.002.

[51] Y. Tian, Y. Shi, and X. Liu, “Recent advances on support vector machines research,” *Technological and Economic Development of Economy*, vol. 18, no. 1, pp. 5–33, Apr. 2012, doi: 10.3846/20294913.2012.661205.

[52] A. Çevik, A. E. Kurtoğlu, M. Bilgehan, M. E. Gülşan, and H. M. Albegmprli, “Support vector machines in structural engineering: A review,” *Journal of Civil Engineering and Management*, vol. 21, no. 3, pp. 261–281, Feb. 2015, doi: 10.3846/13923730.2015.1005021.

[53] 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, 2023, doi: 10.14569/IJACSA.2023.0140455.

[54] B. de Ville, “Decision trees,” *WIREs Computational Statistics*, vol. 5, no. 6, pp. 448–455, Nov. 2013, doi: 10.1002/wics.1278.

[55] S. B. Kotsiantis, “Decision trees: a recent overview,” *Artificial Intelligence Review*, vol. 39, no. 4, pp. 261–283, Apr. 2013, doi: 10.1007/s10462-011-9272-4.

[56] V. G. Costa and C. E. Pedreira, “Recent advances in decision trees: an updated survey,” *Artificial Intelligence Review*, vol. 56, no. 5, pp. 4765–4800, May 2023, doi: 10.1007/s10462-022-10275-5.

[57] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, “An introduction to decision tree modeling,” *Journal of Chemometrics*, vol. 18, no. 6, pp. 275–285, Jun. 2004, doi: 10.1002/cem.873.

[58] C. H. Flayer, C. Perner, and C. L. Sokol, “A decision tree model for neuroimmune guidance of allergic immunity,” *Immunology & Cell Biology*, vol. 99, no. 9, pp. 936–948, Oct. 2021, doi: 10.1111/imcb.12486.

[59] P. Geurts, A. Irrthum, and L. Wehenkel, “Supervised learning with decision tree-based methods in computational and systems biology,” *Molecular BioSystems*, vol. 5, no. 12, pp. 1593–1605, 2009, doi: 10.1039/b907946g.

[60] S. G and G. Ramkumar, “A robust breast cancer classification model using extra-trees classifier for histopathological image,” in *2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, May 2023, pp. 1–7. doi: 10.1109/ACCAI58221.2023.10199852.

[61] R. K. Grace and M. I. Priyadharshini, “Wind speed prediction using extra tree classifier,” in *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, Apr. 2023, pp. 1–4. doi: 10.1109/ICEEICT56924.2023.10157692.

[62] B. Baranidharan, A. Pal, and P. Muruganandam, “Cardio-vascular disease prediction based on ensemble technique enhanced using extra tree classifier for feature selection,” *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 3, pp. 3236–3242, Sep. 2019, doi: 10.35940/ijrte.C5404.098319.

[63] M. Arya, H. Sastry G, A. Motwani, S. Kumar, and A. Zaguia, “A novel extra tree ensemble optimized DL framework (ETEODL) for early detection of diabetes,” *Frontiers in Public Health*, vol. 9, Feb. 2022, doi: 10.3389/fpubh.2021.797877.

[64] Z. Zhang, “Introduction to machine learning: k-nearest neighbors,” *Annals of Translational Medicine*, vol. 4, no. 11, pp. 218–218, Jun. 2016, doi: 10.21037/atm.2016.03.37.

[65] D. Prasad, S. K. Goyal, A. Sharma, A. Bindal, and V. S. Kushwah, “System model for prediction analytics using K-nearest neighbors algorithm,” *Journal of Computational and Theoretical Nanoscience*, vol. 16, no. 10, pp. 4425–4430, Oct. 2019, doi: 10.1166/jctn.2019.8536.

[66] A. Agafonov and A. Yumaganov, “Short-term traffic flow forecasting using a distributed spatial-temporal K-nearest neighbors model,” in *2018 IEEE International Conference on Computational Science and Engineering (CSE)*, Oct. 2018, pp. 91–98. doi: 10.1109/CSE.2018.00019.

[67] F. Nigsch, A. Bender, B. van Buuren, J. Tissen, E. Nigsch, and J. B. O. Mitchell, “Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization,” *Journal of Chemical Information and Modeling*, vol. 46, no. 6, pp. 2412–2422, Nov. 2006, doi: 10.1021/ci060149f.

[68] 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*, vol. 14, no. 2, 2023, doi: 10.14569/IJACSA.2023.0140272.

[69] D. K. Behera, M. Das, S. Swetanisha, and P. K. Sethy, “Hybrid model for movie recommendation system using content K-nearest neighbors and restricted Boltzmann machine,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 1, pp. 445–452, Jul. 2021, doi: 10.11591/ijeecs.v23.i1.pp445-452.

[70] R. Garcia-Carretero, L. Vigil-Medina, I. Mora-Jimenez, C. Soguero-Ruiz, O. Barquero-Perez, and J. Ramos-Lopez, “Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population,” *Medical & Biological Engineering & Computing*, vol. 58, no. 5, pp. 991–1002, May 2020, doi: 10.1007/s11517-020-02132-w.

[71] S. J. Rigatti, “Random forest,” *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, Jan. 2017, doi: 10.17849/insm-47-01-31-39.1.

[72] A. B. Shaik and S. Srinivasan, “A brief survey on random forest ensembles in classification model,” in *International Conference on Innovative Computing and Communications*, 2019, pp. 253–260. doi: 10.1007/978-981-13-2354-6_27.

[73] X. Chen and H. Ishwaran, “Random forests for genomic data analysis,” *Genomics*, vol. 99, no. 6, pp. 323–329, Jun. 2012, doi: 10.1016/j.ygeno.2012.04.003.

[74] H. Tyralis, G. Papacharalampous, and A. Langousis, “A brief review of random forests for water scientists and practitioners and their recent history in water resources,” *Water*, vol. 11, no. 5, Apr. 2019, doi: 10.3390/w11050910.

[75] G. Biau, “Analysis of a random forests model,” *Journal of Machine Learning Research*, vol. 13, pp. 1063–1095, 2012.

[76] T. A. Assegie, R. L. Tulasi, and N. K. Kumar, “Breast cancer prediction model with decision tree and adaptive boosting,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 1, pp. 184–190, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp184-190.

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

[78] 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, Apr. 2022, doi: 10.18280/ria.360208.

[79] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Frontiers in Neurorobotics*, vol. 7, 2013, doi: 10.3389/fnbot.2013.00021.

[80] 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.

[81] A. S. Tyurin and P. V. Saraev, “Adaptation of natural gradient boosting model for production environment,” in *2022 4th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*, Nov. 2022, pp. 648–650. doi: 10.1109/SUMMA57301.2022.9973991.

[82] A. Mayr, H. Binder, O. Gefeller, and M. Schmid, “The evolution of boosting algorithms. From machine learning to statistical modelling,” *Methods of Information in Medicine*, vol. 53, no. 06, pp. 419–427, Jan. 2014, doi: 10.3414/ME13-01-0122.

## 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, machine learning 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. President of the chapter of the Education Society IEEE-Peru. Conference Chair of the Engineering International Research Conference IEEE Peru EIRCON. Advisor and Jury of Engineering Thesis in different universities in Peru. International lecturer in Spain, United Kingdom, South Africa, Romania, Argentina, Chile, China. Specialization in software development, artificial intelligence, machine learning, business intelligence, augmented reality. Reviewer IEEE Peru. He can be contacted at email: mcabanillas@ieee.org.