Machine Learning Algorithms for Classification Patients with Parkinson's Disease and Hereditary Ataxias

Osiris Escamilla-Luna, Miguel A. Wister, and Jose Hernandez-Torruco

***Abstract*—Neurodegenerative diseases are a group of neurological conditions characterized by the loss or destruction of neurons in the central nervous system, resulting in severe impairments and death. Researchers commonly used a two-group classification (Patients with a Neurodegenerative disease vs. healthy subjects of control). Thus, the principal purpose of this article is to distinguish between Parkinson's patients and subjects with Hereditary Ataxias using machine learning techniques. We conducted experiments using a real dataset comprising Gait characteristics derived from the inertial motion sensors of a smartphone (iPhone 5S). This investigation had 67 participants, 53 of who had Parkinson's disease and 14 of whom had Hereditary Ataxias. Methods of feature selection were applied to reduce dimensionality. In addition, five classification algorithms were constructed and assessed based on their accuracy, precision, sensitivity, and specificity. The Support Vector Machine algorithm achieved an accuracy of 92.7%, a precision of 91.1%, a sensitivity of 96.2%, and a specificity of 89.1%. These results show that the suggested technique might inspire new research issues and have a direct therapeutic impact.**

***Keywords*—Neurodegenerative Disease, Gait, Machine Learning, Smartphone.**

I. Introduction

Neurodegenerative disorders impair motor abilities, resulting in an imbalanced gait pattern due to a lack of coordination and balance. Nevertheless, each neurodegenerative illness affects the patient's motions uniquely. Among neurological disorders, Parkinson's disease affects about six million individuals worldwide. In addition, it has the fastest growth rate. Therefore, it is anticipated that 13 million individuals will have PD by 2040 [1]. Parkinson's disease (PD) is a neurodegenerative disorder brought on by a substantial loss of dopamine in the forebrain. Variable indications and symptoms of Parkinson's disease include tremors, slower movement, tight muscles, decreased posture and balance, loss of natural motions, and speech and writing abnormalities [2].

Manuscript received December 9, 2022; revised December 22, 2022. Date of publication February 2, 2023. Date of current version February 2, 2023. The associate editor prof. Dinko Begušić has been coordinating the review of this manuscript and approved it for publication.

The paper was presented in part at the International Conference on Software, Telecommunications and Computer Networks (*SoftCOM*) 2022.

Authors are with the Juarez Autonomous University of Tabasco (UJAT), Mexico (e-mails: 201h18001@alumno.ujat.mx, miguel.wister@ujat.mx, and jose.hernandezt@ujat.mx).

Digital Object Identifier (DOI): 10.24138/jcomss-2022-0157

Neurologists commonly use clinical scales to diagnose PD, including the Movement Disorder Society-sponsored version of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [3] and the Hoehn & Yahr (HY) [4]. However, due to the subjectivity of neurologists, diagnosis of PD can be challenging for physicians due to overlapping symptoms of other neurodegenerative diseases. A past study reveals that about 25% of diagnoses are incorrect [5]; therefore, objective motor assessment tools are crucial for the future of PD diagnostic procedures.

Another neurological disease is Hereditary Ataxias (HA). HA is a form of neurological illness that encompasses a broad set of conditions defined by ataxic gait, lack of coordination between the hands and eyes, and cerebellar atrophy. However, the primary symptom of the category of motor diseases is progressive walking ataxia caused by neurodegeneration of components in the cerebellar cortex, brainstem, and spinal cord [6].

Recent research has incorporated Machine Learning (ML) approaches for PD classification, notably for speech analysis, comparing ML algorithms such as Naive Bayes (NB), J48, Random Forest (RF), and Support Vector Machine (SVM) [7] [8]. Generally, a two-group classification (Patients with a neurodegenerative disorder vs. healthy subjects of control (HC)) was utilized, with beneficial results, as we can see from the summary of recent work in Table I.

Some researchers utilized ML to classify PD patients and HC examining motion abilities. However, the majority of investigations [17] [18] focused on the lower limbs, identifying evidence of changes in gait patterns and comparing them to medical scales to predict the course of neurodegenerative illness. Again, accelerometers and gyroscopes were the most common technology [19] [20].

Research indicates that the iPhone's sensors are dependable and accurate enough to analyze and detect kinematic gait patterns [21], [22]. In addition, research related to gait evaluation and healthcare has shown that iPhone sensors can obtain quantified gait characteristics with adequate precision and consistency, notably in ankle position and in a manner that is easy, portable, and wearable [23].

This article should be viewed as expanding the conference paper published in [20]. The prior study used machine learning to categorize PD patients and HC based on their motor capabilities. In this project, we utilized ML to classify PD patients and individuals with another neurological condition, in

TABLE I
STUDIES THAT APPLIED MACHINE LEARNING MODELS TO MOVEMENT DATA TO DIAGNOSE PD

Ref.	*Machine learning methods*	*Number of subjects*	*Outcomes (Best result)*
[8]	Support vector machine Artificial neural networks	14 healthy subjects(HC) + 16 Parkinson's Disease patients(PD)	ANN: Accuracy = 0.894
[9]	Support vector machine Random forest Naïve Bayes	30 PD + 30 HC + 30 subjects with idiopathic hypomia(IH)	Random forest: HC vs. PD: Accuracy = 0.950 HC + IH vs. PD: Accuracy = 0.917 HC vs. IH vs. PD: Accuracy = 0.789
[10]	Support vector machine Decision tree Random forest	22 HC + 19 PD	SVM: Accuracy = 0.890
[11]	Logistic regression Naïve Bayes Random forest Decision tree	30 HC + 30 PD	Random forest: Accuracy = 0.820
[12]	Naïve Bayes K-Nearest neighbors Support vector machine	30 HC + 30 PD	SVM: Accuracy = 0.950
[13]	Random forest	26 HC + 14 PD	Accuracy = 0.946
[14]	Support vector machine	10 Early-onset cerebellar ataxia + 20 HC	Accuracy = 0.729
[15]	Artificial neural networks	1 Friedreich’s ataxia + 1 HC	Accuracy = 0.740
[16]	K-Nearest neighbors	30 Cerebellar ataxia + 30 HC	Accuracy = 0.733

this instance, HA, as opposed to healthy control participants. Our primary objective is to identify the optimal ML approach for identifying which disease a patient has.

We organized the rest of the paper as follows; section II discusses the design implementation of this experiment, including data collection, data preprocessing, feature extraction, and selection. We also described in this section the machine learning algorithms used in this paper. Next, section III deals with the performance evaluation of each algorithm used in every dataset, and section IV consists of the conclusion.

II. METHODS

A. Data Collection

We conducted this project in Mexico City with the Manuel Velasco Suarez National Institute of Neurology and Neurosurgery (INNN-MVS). This institute investigates and treats patients with walking disorders, including Parkinson's disease and other neurodegenerative diseases like Hereditary Ataxias. INNN-MVS enabled the creation of a Gait laboratory in a large enough space (20m long by 3m wide) to capture gait characteristics and ensure that patients with walking disorders can move around comfortably.

In this study, two groups were comprised of 53 PD patients (27 males, 26 females, mean age standard deviation: 65.6±14.2 years) and 14 HA patients (HA) (seven males, seven females, mean age standard deviation: 43.2.8±23.0 years). The specialists diagnosed the patients with PD, and HA. The local Ethical Committee adhered to the most recent version of the Declaration of Helsinki, which approved the study protocol after receiving written informed consent from each participant. We used a protocol that accounted for the patient's motor disturbances and gait deterioration, such as loss of balance, precision, and movement speed, to acquire the data. This protocol states that INNN-MVS medical staff supported Gait data collection to prevent accidents, avoid contact to allow patients to walk freely during the trajectory, and respect the illness-induced swings that prevented patients from walking in a straight line. During ten days, data were collected from patients with varying disease severity during their medical visits. With the patient's comfort and freedom of movement in mind, movement sensors were placed on the subjects' ankles at each step. This arrangement of sensors was optimal for reducing walking gait disturbances, see Fig. 1. The protocol specifies that each participant covers 20 meters of the track at an average walking speed. Therefore, we simultaneously placed a smartphone on both the right and left ankles for each participant (R_ANKLE and L_ANKLE, respectively). See Fig. 2.

We utilized the three-axis accelerometer of an iPhone 5S to collect data from each subject's ankle. Figure 3 depicts the raw tri-accelerometer axis value for the walking activity.

B. Data Preprocessing

Preprocessing begins with linear interpolation to calibrate the data sensor and compensate for timing differences. Then,

accelerometer data is recorded at a variable sampling rate amidst setting a recording frequency.

As accelerometers measure the rate of change of the velocity of an object (acceleration), the average acceleration of a motionless accelerometer should be equal to the acceleration of gravity; however, due to their high sensitivity, the Gait data captured by these devices is consistently fluctuating as the acceleration values change. Therefore, we employed zero normalization to eliminate the constant signal effects on each ax [23].

Generally, data is smoothed to eliminate unwanted noise and identify outliers (data points significantly differ from the remainder). A common technique for smoothing data is the moving average, which calculates the mean of the points within a sliding window. This method eliminates insignificant differences between data points. This procedure is analogous to a low-pass filter with a response that is smoothed by its difference.

The values measured along the sensor axis depend on the device's position and method of attachment to the subject's body. However, when attached rigidly to the ankle, the relative motion between the sensor and the ankle can be disregarded. Sensor inertial signals are recorded relative to the smartphone in a coordinate system. For our purposes, the Gait dynamics must be represented in a coordinated manner relative to the subject. It is only sometimes feasible to manipulate or determine the subject-related device orientation. The device will likely maintain a different orientation when collecting data from multiple subjects. To account for this change in orientation, we calculate the orientation-invariant signal acceleration magnitude as the square root of the sum of the squares of each vector value in the time series.

The acceleration signal was used to identify strides by identifying changes in the signal corresponding to heel contact with the ground. Acceleration increases at the beginning of the stride and decreases at its conclusion.

Fig. 1. Arrangement of sensors in both the right and left ankles

We implemented an algorithm to determine the maximum and minimum local value of acceleration magnitude from signal acceleration magnitude. We only considered peaks that were more significant than the subject's standard deviation. A find peaks function identifies all peaks that are more notable than the minimum value and returns the value and index of each peak. The minimum values are calculated by the same algorithm using the same data with the signals inverted as input. To determine the beginning of each stride, we considered minimum peaks below the standard deviation of all values. Each peak chosen as the beginning of a stride should be preceded and followed by a maximum peak higher and lower than its standard deviation, respectively [24]. Fig. 4 depicts the filtered accelerometer value of the walking activity.

Fig. 2. Subject walking 20 meters at normal speed

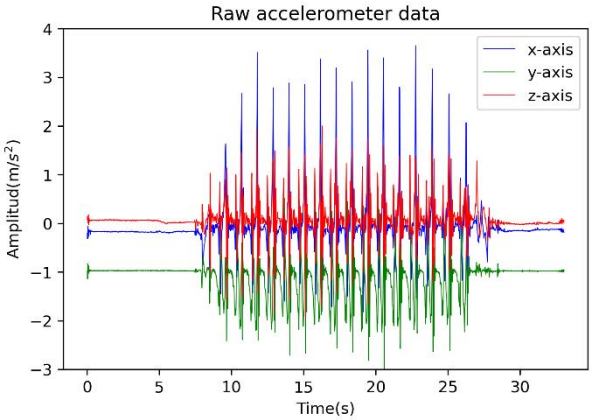

Fig. 3. Raw accelerometer data collected from an iPhone 5S

C. Feature Extraction

We used a computer-assisted Gait assessment tool, "iGAIT" [25], based on MATLAB to help extract gait features. The iGait tool extracted 56 gait characteristics from data taken from each patient's right and left ankle accelerometers: six Spatio-temporal, fifteen Frequency domain, and seven Regularity and Symmetry of step characteristics (28 for each accelerometer). Table II contains exhaustive lists of each type of variable. Given the position of the capture instrument, the movement axes of sensors correspond to gait movements as follows:

• The Y-axis corresponds to vertical movement (VER).

• The X-axis corresponds to the anterior-posterior movement (AP).

• The Z-axis corresponds to the middle-lateral movement (ML).

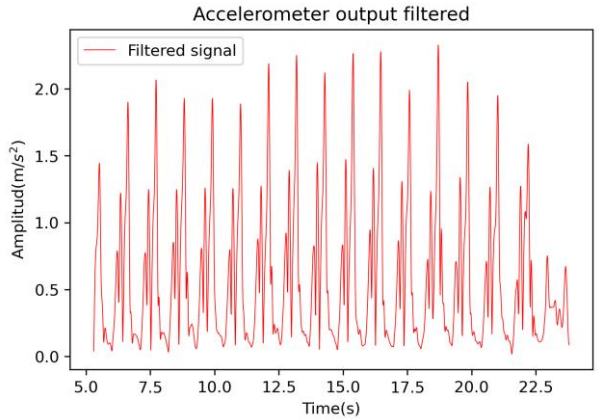

Fig. 4. Filter and segmented accelerometer data

After detecting heel contacts, the number of steps is determined by counting the identified contacts. Next, cadence is calculated by dividing the number of steps by the time spent walking while walking velocity is determined by dividing the distance by the time spent walking. Next, the average step length is calculated by dividing the walking distance by the number of steps. Finally, the square root of the mean acceleration (RMS) indicates the intensity of motion, and the RMS values of the three acceleration directions (VT, AP, and ML) are calculated using (1):

$$RMS_d = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{di} - \bar{x}_d)^2} \quad (1)$$

where x_{di} is the acceleration along either the VT, AP, or ML axis, N is the size of the acceleration signal, and $\bar{x}_d$ is the average acceleration along any axis. iGAIT extracts six Spatio-temporal characteristics from acceleration data, including cadence, mean step length, velocity, RMS_{VT}, RMS_{AP}, and RMS_{ML}.

iGAIT also extracts six frequency domain features from each direction of the acceleration data. These features measure the amplitude and frequency of body movements and are taken from each direction of the acceleration data. The integral of the power spectral density (IPSD) was estimated by a periodogram using (2) and (3). Where $0 \leq \omega \leq \pi$ represents the angular frequency, x_i represents the acceleration along the VT, AP, or ML axis, and N represents the total number of acceleration samples. The frequency with the highest PSD value is known as the primary frequency. The remaining four frequency features ($Fr1$, $Fr2$, $Fr3$, and $Fr4$) correspond to the frequencies when the PSD is cumulated (CPSD), which is calculated using Equation (4) for 50%, 75%, 90%, and 99% of IPSD, respectively. iGAIT extracts eighteen frequency characteristics from the acceleration data, including IPSD, primary frequency, $Fr1$, $Fr2$, $Fr3$, and $Fr4$ in each of the three directions (VT, ML, and AP).

$$PSD(e^{j\omega}) = \frac{1}{2\pi N}\left|\sum_{i=1}^{N} x_i e^{-j\omega i}\right|^2 \quad (2)$$

$$IPSD = \int_0^{\pi} PSD(\omega)d\omega \quad (3)$$

$$CPSD(\omega) = \int_0^{\omega} PSD(\omega)d\omega \quad (4)$$

The autocorrelation coefficient measures the relationship between a time series' past and future values. To quantify the regularity and symmetry of gait, iGAIT employs autocorrelation coefficients of acceleration data that are independent of their direction. Autocorrelation coefficients of acceleration data can be estimated without bias using (5), where x_i represents acceleration data, and $f_c(t)$ represents autocorrelation coefficients, t represents the time. When the time lag t is equal to the periodicity of the acceleration x_i, the $f_c(t)$ series will contain a peak.

$$f_c(t) = \frac{1}{N+|t|}\sum_{i=1}^{N-|t|} x_i x_{i+t}. \quad (5)$$

TABLE II
FEATURES EXTRACTED FROM IGAIT

Spatio-Temporal Features	*Frequency Features*	*Regularity and Symmetry Features*
Cadence (step/min)	IPSD VER	Symmetry in VER
Mean step length (m)	Frequency at 50% energy (Hz) VER	Symmetry in AP
Velocity (m/s)	Frequency at 75% energy (Hz) VER	Stride Regularity in VER
RMS VER	Frequency at 90% energy (Hz) VER	Stride Regularity in ML
RMS ML	Frequency at 100% energy (Hz) VER	Stride Regularity in AP
RMS AP	IPSD ML	Step Regularity in VER
	Frequency at 50% energy (Hz) ML	Step Regularity in AP
	Frequency at 75% energy (Hz) ML	
	Frequency at 90% energy (Hz) ML	
	Frequency at 100% energy (Hz) ML	
	IPSD AP	
	Frequency at 50% energy (Hz) AP	
	Frequency at 75% energy (Hz) AP	
	Frequency at 90% energy (Hz) AP	
	Frequency at 100% energy (Hz) AP	

A. Datasets

We obtained three sets of data, one for each ankle (PH_RANKLE and PH_LANKLE) with the 28 gait features described in the preceding section and the combination of the two (PH_ANKLES) with 56 Gait features. Each dataset contains 67 instances.

B. Feature Selection

Feature selection techniques are required to alleviate the effect of the curse of dimensionality, improve the performance of the designed model, shorten the time required for the learning process, and enhance data comprehension. This study employed the f-test for one-way Analysis of Variance (ANOVA). The F-test computes the variance ratio between two groups and within a group for a given characteristic. In this case, the groups are instances with the same target value. Greater values of the f-test indicate smaller distances within the groups and greater distances between the groups. In this ANOVA feature selection method using the f-test, the features are ranked according to their f-score [26].

F. SMOTE

A problem with imbalanced classification is insufficient instances of the minority type for a model to learn the decision boundary effectively. One solution to this difficulty is to oversample samples from the minority group. Before fitting a model, oversampling can be achieved by simply duplicating examples from the minority type in the training dataset. This method can balance the category distribution but provides no additional data to the model.

Synthesizing new samples from the minority category is an improvement over duplicating minority samples. Therefore, this method is an acceptable form of data enhancement for tabular data. The Synthetic Minority Oversampling Technique(SMOTE) is conceivably the most popular technique for synthesizing new examples.

Setting samples close to the feature space, marking a line between the instances in the feature space, and drawing a new sample at a point along the line is how SMOTE operates. In particular, SMOTE selects a sample at random from the minority class. Then, SMOTE identifies the n closest neighbors that belong to the same minor category for any value of a minor type, draws a straight line with that neighbor, and generates random values until they exhibit a synthetic ratio [27].

G. Classifiers

This work uses five supervised classifiers among the most recently used to classify people based on Gait features: Logistic Regression, Naive Bayes, K-Nearest Neighbors, Support Vector Machine, Decision Trees, and Random Forests.

Support Vector Machines: It is a discriminative classifier defined mathematically by separating hyperplanes. It is very efficient in binary classification with small data sizes. Since the data size used in the experiment is comparatively more minor, SVM is expected to show good performance. The algorithm is used in solving regression problems also [28].

Naive Bayes: Based on Bayes' theorem, the Naive Bayes Classifier is a straightforward probabilistic classification technique. In Bayes' theorem, if two independent occurrences occur randomly and in succession, the chance of the second event occurring is the likelihood that one of these events occurs. For example, the product rule in (6) can be expressed using two distinct expressions by utilizing the change property.

$$P(X \cap Y) = P(X|Y)P(Y) = P(Y|X)P(X) \tag{6}$$

As shown in (7), Bayes' theorem describes the relationship between a random event occurring in a random process and the conditional probability of another unexpected event.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{7}$$

In (7), Bayes calculates the probability of the dependent states that are likely to occur in any scenario. In this equation, P (X) represents the problem's input probability, P (X) indicates the likelihood of possible exit status, and P (Y |X) reflects the chance of an output of Y given input X. The NB classification method analyzes the link between dependent and independent attributes to generate a conditional probability for each relationship. the effects of independent variables on the dependent variable are combined To classify a new sample [29].

K-Nearest Neighbors: It is one of the supervised machine learning algorithms most used and one of the simplest classification algorithms. It is a method of data classification that determines what cluster a data point is in by examining the data points around it. For example, a point is categorized based on the majority vote of its neighbors, with the point being assigned to the category most common amongst its K nearest neighbors estimated by a distance function [30].

Decision Tree: It is one of the most straightforward machine learning algorithms that can be used to solve classification and regression problems. A decision tree is a simple flowchart-like tree structure that facilitates binary classification by making iterative decisions for different attributes based on the outcomes of random events [28].

Random Forest: A classification algorithm developed by Breiman and Cutler uses an ensemble of tree predictors. In Random Forest, each tree is constructed by bootstrapping the training data and for each split randomly selected subset of features is used. Splitting is made based on purity measures [31].

H. Performance Evaluation

All the classifiers used above are evaluated based on Accuracy, Precision, Sensitivity(recall), Specificity, F1 score, and AUC, respectively. The formulas for (8), (9), (10), (11), and (12) are based on a confusion matrix, as shown in Table III. AUC is the two-dimensional area under the Receiver Operating Characteristic (ROC) curve.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{6}$$

$$Precision = \frac{TP}{(TP+FP)} \tag{7}$$

$$Sensitivity = \frac{TP}{(TP+FN)} \tag{8}$$

$$Specificity = \frac{TN}{(TN+FP)} \tag{9}$$

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{10}$$

TABLE III
CONFUSION MATRIX

	Classified as PD Patients	*Classified as HA Patients*
Real Patients with PD	True Positives	False Negatives
Real Patients with HA	False Positives	True Negatives

where TP, TN, FP, and FN are True Positives, True Negatives, False Positives, and False Negatives, respectively.

I. Software

We performed our data pre-processing and analysis in Python version 3.6. We used software packages: pandas (version 1.3.4), Numpy (version 1.18) [32], Scipy (version 1.7.1) [33], imbalanced-learn (version 0.9.0), and scikit-learn (version 1.0.1) [34].

III. RESULTS AND DISCUSSION

A. Feature Selection

After applying the method of ANOVA with the f-test to each dataset, we selected the first ten features. The selected features are shown in Table IV and were included in the subsequent analysis.

TABLE IV
SELECTED FEATURES

Feature number	*Both ankles*[a]	*Right ankle*	*Left ankle*
	Feature	*Feature*	*Feature*
1	Freq_90_VER_L	Freq_100_AP	Freq_90_VER
2	RMS_ML_L	Velocity	RMS_ML
3	Freq_75_VER_L	Sym_AP	Freq_75_VER
4	Freq_50_VER_L	Freq_100_ML	Freq_50_VER
5	Freq_100_VER_L	Freq_90_AP	Freq_100_VER
6	Freq_100_AP_R	RMS_AP	IPSD_ML
7	Velocity_R	Stride_Reg_ML	Velocity
8	Sym_AP_R	Freq_90_ML	RMS_AP
9	IPSD_ML_L	Cadence	Cadence
10	Velocity_L	Freq_50_VER	IPSD_AP

[a] L=Left ankle, R=Right ankle

B. Oversampling

In the three datasets analyzed for this study, the proportion of individuals with PD was 79.1%, while the proportion of patients with HA was 20.9%. (imbalance rate: 3.78). As a result, we discovered a moderate imbalance in the class variable. Because they attempt to predict classes with greater weight, classifiers trained on these skewed data are more likely to produce biased outcomes. Therefore, it may increase precision. However, the precision of a variable with low frequency is likely to decrease. This study utilized SMOTE (sampling strategy: minority and k neighbors: 5) to circumvent the imbalance problem of these datasets. This method produced the following new balanced datasets:

- PH_ANKLES: 53 PD vs. 53 HA.
- PH_RANKLE: 53 PD vs. 53 HA.
- PH_LANKLE: 53 PD vs. 53 HA.

C. Hyperparameter Tunning and Cross Validation

We conducted an Exhaustive Grid Search to determine the optimal hyperparameters for every model. Table V displays the hyper-parameter values utilized by the algorithms. Following that, we utilized a 10-FCV evaluation scheme:

- Cross-validation randomly divides the original data into k subsets of equal size.
- It performs k loops in which k-1 of the original dataset is used for training and the last k for testing.
- The results of each iteration are averaged to obtain a single result.

TABLE V
MODEL HYPERPARAMETERS

Model	*Parameter*	*Value*
Logistic Regression	C	0.1
Support Vector Machines	kernel	rbf
	C	1
	gamma	1
Naive Bayes	var_smoothing	$1x10^{-9}$
K-Nearest Neighbors	n_neighbors	3
	metric	minkowski
Decision Tree	criterion	entropy
	max_depth	4
	max_leaf_nodes	5
	min_samples_split	10
	min_samples_leaf	1
Random Forest	n_estimators	50
	criterion	entropy
	max_depth	4
	max_leaf_nodes	5
	min_samples_split	10
	min_samples_leaf	1

For each dataset, we performed 50 runs using each classifier described previously. In each sequence, a different seed is set. Using a different seed ensures that the train and test sets will be divided differently. In addition, for each run, Accuracy, Sensitivity, and Specificity were calculated. This procedure was executed on all datasets. We then calculated the average of these measurements across 50 runs.

D. Results and Discussion

The classification results for the right and left ankles, as well as the ankles as a whole, are displayed in Table VI. The Cross-Validation results are shown in Fig. 5, 6, and 7. SVM produced the best results within the PH ANKLES data set. With a mean accuracy of 0.927 ± 0.012, the average diagnostic performance was exemplary. Other models performed better than indifferent metrics in this dataset; for example, K-Nearest Neighbors had a precision of 0.963 ± 0.028 and a specificity of 0.971 ± 0.018. SVM also had the best performance for PH RANKLE, with a mean accuracy of 0.894 ± 0.01. Nonetheless, K-Nearest Neighbors had the highest average specificity in that dataset, with 0.877 ± 0.031. In addition, SVM performed the best for PH LANKLE, with an accuracy of 0.914 ± 0.009. However, K-Nearest Neighbors had the highest average specificity in that dataset, with 0.877 ± 0.031. The SVM model had the

TABLE VI
CLASSIFIER PERFORMANCE

Datasets	*Model*	*Accuracy*	*Precision*	*Sensitivity*	*Specificity*	*F1*	*AUC*
Both Ankles	Logistic Regression	0.787 ± 0.017	0.796 ± 0.022	0.803 ± 0.019	0.771 ± 0.026	0.784 ± 0.018	0.787 ± 0.017
	Support Vector Machines	**0.927 ± 0.012**	0.911 ± 0.018	**0.962 ± 0.002**	0.891 ± 0.023	**0.931 ± 0.011**	**0.927 ± 0.012**
	Naive Bayes	0.761 ± 0.011	0.754 ± 0.016	0.813 ± 0.020	0.709 ± 0.012	0.770 ± 0.014	0.761 ± 0.011
	K-Nearest Neighbors	0.836 ± 0.017	**0.963 ± 0.028**	0.701 ± 0.029	**0.971 ± 0.018**	0.794 ± 0.024	0.836 ± 0.018
	Decision Tree	0.712 ± 0.025	0.820 ± 0.053	0.586 ± 0.048	0.837 ± 0.055	0.654 ± 0.036	0.711 ± 0.026
	Random Forest	0.807 ± 0.024	0.847 ± 0.031	0.774 ± 0.034	0.839 ± 0.034	0.792 ± 0.028	0.807 ± 0.024
Right Ankle	Logistic Regression	0.795 ± 0.018	0.816 ± 0.028	0.797 ± 0.022	0.794 ± 0.032	0.792 ± 0.019	0.796 ± 0.018
	Support Vector Machines	**0.894 ± 0.019**	**0.865 ± 0.025**	**0.960 ± 0.010**	0.829 ± 0.034	**0.903 ± 0.016**	**0.894 ± 0.019**
	Naive Bayes	0.675 ± 0.012	0.665 ± 0.016	0.742 ± 0.017	0.606 ± 0.023	0.690 ± 0.012	0.674 ± 0.012
	K-Nearest Neighbors	0.745 ± 0.018	0.855 ± 0.039	0.613 ± 0.024	**0.877 ± 0.031**	0.692 ± 0.022	0.745 ± 0.019
	Decision Tree	0.744 ± 0.040	0.805 ± 0.057	0.672 ± 0.056	0.816 ± 0.056	0.709 ± 0.046	0.744 ± 0.040
	Random Forest	0.785 ± 0.023	0.810 ± 0.029	0.777 ± 0.036	0.792 ± 0.031	0.776 ± 0.027	0.784 ± 0.023
Left Ankle	Logistic Regression	0.740 ± 0.014	0.727 ± 0.019	0.807 ± 0.018	0.674 ± 0.022	0.751 ± 0.015	0.741 ± 0.014
	Support Vector Machines	**0.914 ± 0.009**	**0.872 ± 0.014**	**0.994 ± 0.009**	0.834 ± 0.018	**0.925 ± 0.008**	**0.914 ± 0.009**
	Naive Bayes	0.716 ± 0.016	0.711 ± 0.022	0.770 ± 0.022	0.662 ± 0.024	0.726 ± 0.018	0.716 ± 0.017
	K-Nearest Neighbors	0.698 ± 0.013	0.829 ± 0.046	0.510 ± 0.021	**0.885 ± 0.019**	0.606 ± 0.025	0.698 ± 0.014
	Decision Tree	0.697 ± 0.027	0.724 ± 0.045	0.710 ± 0.052	0.683 ± 0.062	0.682 ± 0.033	0.697 ± 0.026
	Random Forest	0.728 ± 0.018	0.756 ± 0.034	0.709 ± 0.031	0.747 ± 0.036	0.712 ± 0.022	0.728 ± 0.018

highest average AUC in the PH Ankles dataset, at 0.927 ± 0.012. The ROC Curves are depicted in Fig. 8, 9, and 10.

We conducted a Mann-Whitney U test with a significance level of 0.05 to compare accuracy differences between the model with the highest performance and the others. In the PH_ANKLES dataset, the results revealed a highly significant distinction. As the p-value obtained from the Mann-Whitney U test (as shown in Table VII) is significant, we conclude that the SMV model has a significantly higher yield than the other four models. Next, in the PH_RANKLE dataset, we compare accuracy differences between the SVM model and the other models, and the results revealed a significant distinction. We conclude that the SVM model has a significantly higher yield than the other four models, as the p-value from the Mann-Whitney U test (as shown in Table VIII) is significant. Finally, in the PH_LANKLE dataset, we compare accuracy differences between the SVM model and the other models (as shown in Table IX). Again, the results revealed a highly significant distinction. However, the SVM model has a significantly higher yield than the other models.

Using the dataset containing both ankles, SVM was more accurate than Logistic regression, KNN, Decision Tree, and Random Forest. Most models performed better when utilizing the dataset containing both ankles as opposed to the same model utilizing datasets containing only one ankle.

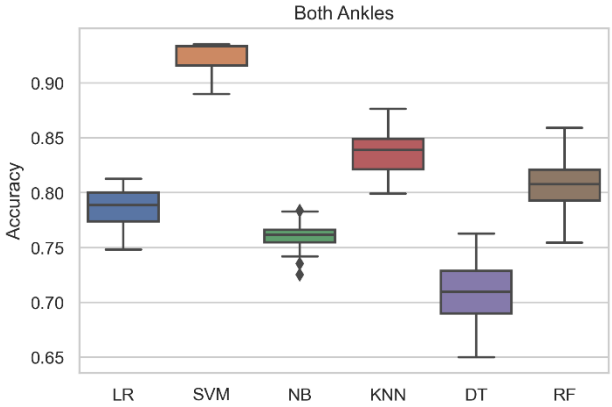

Fig. 5. PH_ANKLES dataset results from the five models

TABLE VII
SIGNIFICANCE TEST IN THE PH_ANKLES DATASET

	p-value
SVM vs Logistic Regresion	3.242x10-18
SVM vs Naive Bayes	3.229x10-18
SVM vs KNN	3.242x10-18
SVM vs Decision Tree	3.242x10-18
SVM vs Random Forest	3.231x10-18

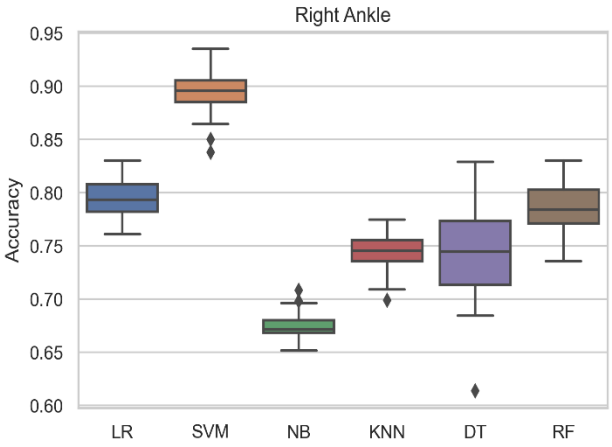

Fig. 6. PH_RANKLE dataset results from the five models

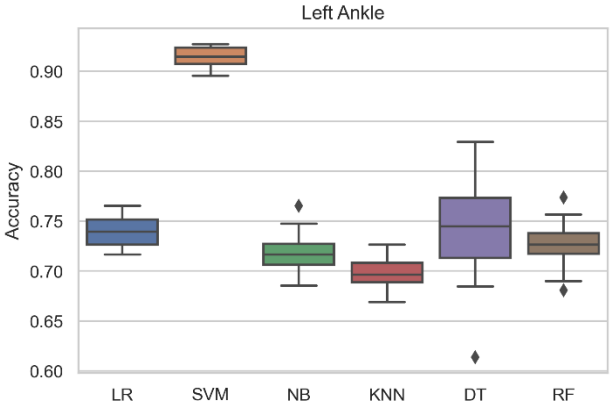

Fig. 7. PH_LANKLE dataset results from the five models

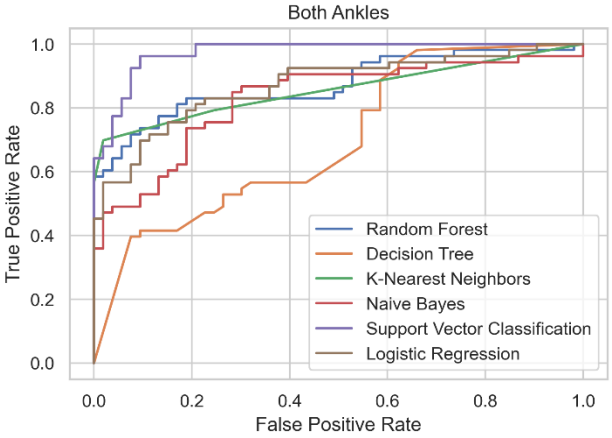

Fig. 8. PH_ANKLES dataset roc curves from the five models

TABLE VIII
SIGNIFICANCE TEST IN THE PH_RANKLE DATASET

	p-value
SVM vs Logistic Regresion	3.472x10-18
SVM vs Naive Bayes	3.459x10-18
SVM vs KNN	3.462x10-18
SVM vs Decision Tree	3.487x10-18
SVM vs Random Forest	3.468x10-18

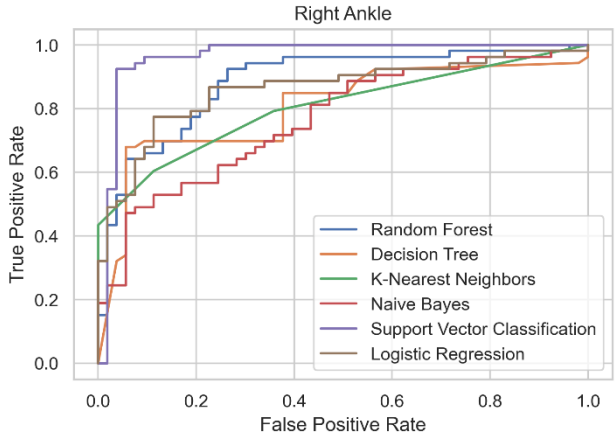

Fig. 9. PH_RANKLE dataset roc curves from the five models

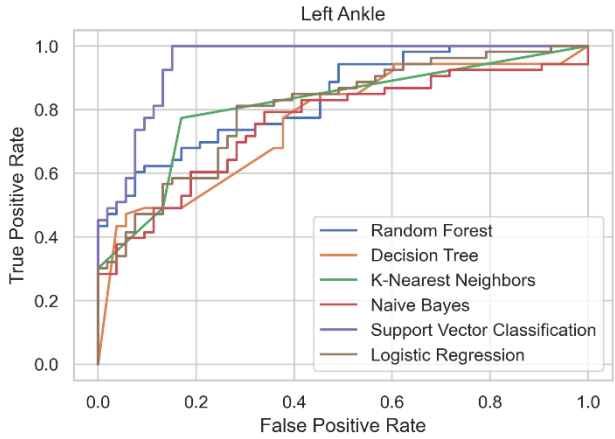

Fig. 10. PH_LANKLE dataset roc curves from the five models

TABLE IX
SIGNIFICANCE TEST IN THE PH_LANKLE DATASET

	p-value
SVM vs Logistic Regresion	3.455x10-18
SVM vs Naive Bayes	3.458x10-18
SVM vs KNN	3.457x10-18
SVM vs Decision Tree	3.465x10-18
SVM vs Random Forest	3.455x10-18

IV. CONCLUSIONS

In this paper, we implemented various classification methods for Neurodegenerative Disease, and they could differentiate between PD and HA. This analysis was conducted with a small number of Hereditary Ataxia patients because it is difficult to access a more significant number of them due to the disease's rareness in the overall population and limited availability in specialized medical centers. The data were preprocessed and segmented to extract Gait characteristics. We implemented a feature selection technique to extract the essential features to optimize the algorithm's dimensionality. In addition, we utilized an oversampling technique to rectify the imbalanced nature of these datasets. We implemented the Logistic Regression, KNN, SVM, Decision Tree, and Random Forest classification methods. The SVM with C equal to 1, gamma equal to 1, and an RBF kernel achieved 92.7%

accuracy, making it the most accurate classification method. In a previous investigation [35], the results using an enhanced weight voting ensemble (EWVE) method showed a performance in differentiation between PD vs. ataxia patients (AUC 0.974 ± 0.036, sensitivity 0.829 ± 0.217, specificity 0.969 ± 0.038). However, our results obtained in the same metrics using a support vector machine and a k-nearest neighbor model showed a better effect in sensitivity and specificity. Results inspire future work in constant monitoring because binary classification may be helpful, e.g., to better monitor the condition's progression by identifying an aggravation reflected in a change in the manner in which the study participants walk. Also we will use different walk lengths to analyze new performances in future investigations. This field of study necessitates a monitoring system with real-time data transfer.

ACKNOWLEDGMENT

We would like to thank Dr. Catherine Boll for her support in carrying out the measurements for this project. We are very grateful to the staff and employees of Manuel Velasco Suarez National Institute of Neurology and Neurosurgery for their collaboration. We are also grateful to all patients who participated and the accompanying family members. Finally, we would like to express our gratitude to Juarez Autonomous University of Tabasco for supporting the academic resources needed for this research.

REFERENCES

[1] E. R. Dorsey and B. R. Bloem, "The Parkinson Pandemic—A Call to Action," JAMA Neurol, vol. 75, no. 1, p. 9, Jan. 2018, doi: 10.1001/jamaneurol.2017.3299.

[2] D. Twelves, K. S. M. Perkins, and C. Counsell, "Systematic review of incidence studies of Parkinson's disease," Mov Disord., vol. 18, no. 1, pp. 19–31, Jan. 2003, doi: 10.1002/mds.10305.

[3] C. G. Goetz et al., "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan," Mov Disord., vol. 22, no. 1, pp. 41–47, Jan. 2007, doi: 10.1002/mds.21198.

[4] M. M. Hoehn, "Parkinsonism: onset, progression, and mortality," p. 17.

[5] E. Tolosa, G. Wenning, and W. Poewe, "The diagnosis of Parkinson's disease," p. 12.

[6] P. Durga, V. S. Jebakumari, and D. Shanthi, "Diagnosis and Classification of Parkinsons Disease Using Data Mining Techniques," vol. 3, no. 14, p. 5, 2016.

[7] J. Sujatha, "Performance Evaluation of Machine Learning Algorithms in the Classification of Parkinson Disease Using Voice Attributes," vol. 12, no. 21, p. 7, 2017.

[8] D. Buongiorno, I. Bortone, G. D. Cascarano, G. F. Trotta, A. Brunetti, and V. Bevilacqua, "A low-cost vision system based on the analysis of motor features for recognition and severity rating of Parkinson's Disease," BMC Med Inform Decis Mak, vol. 19, no. S9, p. 243, Dec. 2019, doi: 10.1186/s12911-019-0987-5.

[9] F. Cavallo, A. Moschetti, D. Esposito, C. Maremmani, and E. Rovini, "Upper limb motor pre-clinical assessment in Parkinson's disease using machine learning," Parkinsonism & Related Disorders, vol. 63, pp. 111–116, Jun. 2019, doi: 10.1016/j.parkreldis.2019.02.028.

[10] Y. Klein, R. Djaldetti, Y. Keller, and I. Bachelet, "Motor dysfunction and touch-slang in user interface data," Sci Rep, vol. 7, no. 1, p. 4702, Dec. 2017, doi: 10.1038/s41598-017-04893-1.

[11] C. Urcuqui et al., "Exploring Machine Learning to Analyze Parkinson's Disease Patients," in 2018 14th International Conference on Semantics, Knowledge and Grids (SKG), Guangzhou, China, Sep. 2018, pp. 160–166. doi: 10.1109/SKG.2018.00029.

[12] M. Ricci, G. Di Lazzaro, A. Pisani, N. B. Mercuri, F. Giannini, and G. Saggio, "Assessment of Motor Impairments in Early Untreated Parkinson's Disease Patients: The Wearable Electronics Impact," IEEE J. Biomed. Health Inform., vol. 24, no. 1, pp. 120–130, Jan. 2020, doi: 10.1109/JBHI.2019.2903627.

[13] A. Kuhner et al., "Correlations between Motor Symptoms across Different Motor Tasks, Quantified via Random Forest Feature Classification in Parkinson's Disease," Front. Neurol., vol. 8, p. 607, Nov. 2017, doi: 10.3389/fneur.2017.00607.

[14] A. Mannini et al., «Automatic classification of gait in children with early-onset ataxia or developmental coordination disorder and controls using inertial sensors», Gait & Posture, vol. 52, pp. 287-292, feb. 2017, doi: 10.1016/j.gaitpost.2016.12.002.

[15] R. LeMoyne, F. Heerinckx, T. Aranca, R. De Jager, T. Zesiewicz, y H. J. Saal, «Wearable body and wireless inertial sensors for machine learning classification of gait for people with Friedreich's ataxia», en 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN), San Francisco, CA, USA, jun. 2016, pp. 147-151. doi: 10.1109/BSN.2016.7516249.

[16] C. Pradhan et al., «Automated classification of neurological disorders of gait using spatio-temporal gait parameters», Journal of Electromyography and Kinesiology, vol. 25, n.º 2, pp. 413-422, abr. 2015, doi: 10.1016/j.jelekin.2015.01.004.

[17] E. Rovini, C. Maremmani, and F. Cavallo, "How Wearable Sensors Can Support Parkinson's Disease Diagnosis and Treatment: A Systematic Review," Front. Neurosci., vol. 11, p. 555, Oct. 2017, doi: 10.3389/fnins.2017.00555.

[18] G. Rigas et al., "Assessment of Tremor Activity in the Parkinson's Disease Using a Set of Wearable Sensors," IEEE Trans. Inform. Technol. Biomed., vol. 16, no. 3, pp. 478–487, May 2012, doi: 10.1109/TITB.2011.2182616.

[19] J. Mei, C. Desrosiers, and J. Frasnelli, "Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature," Front. Aging Neurosci., vol. 13, p. 633752, May 2021, doi: 10.3389/fnagi.2021.633752.

[20] Jayadev, S.; Bird, T.D. Hereditary ataxias: Overview. Genetics in Medicine 2013, 15, 673–683. doi:10.1038/gim.2013.28.

[21] B. Sun, Y. Wang, and J. Banda, "Gait Characteristic Analysis and Identification Based on the iPhone's Accelerometer and Gyrometer," Sensors, vol. 14, no. 9, pp. 17037–17054, Sep. 2014, doi: 10.3390/s140917037.

[22] D. A. Torres et al., «Detection of fatigue on gait using accelerometer data and supervised machine learning», IJGUC, vol. 11, n.º 4, p. 474, 2020, doi: 10.1504/IJGUC.2020.108475.

[23] F. D. Acosta-Escalante, E. Beltrán-Naturi, M. C. Boll, and J.-A. Hernández-Nolasco, "Gait Pattern Recognition in Patients with Hereditary Ataxia's Using Supervised Learning Algorithms Based on Small Subsets Smartphone Sensor Data," MATHEMATICS & COMPUTER SCIENCE, preprint, Dec. 2020. doi: 10.20944/preprints202012.0054.v1.

[24] O. Escamilla-Luna, M. A. Wister and J. Hernandez-Torruco, "Classification Algorithms for Analyzing Parkinson's Disease Patient," 2022 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), 2022, pp. 1-6, doi: 10.23919/SoftCOM55329.2022.9911391.

[25] M. Yang, H. Zheng, H. Wang, S. McClean, and D. Newell, "iGAIT: An interactive accelerometer based gait analysis system," Computer Methods and Programs in Biomedicine, vol. 108, no. 2, pp. 715–723, Nov. 2012, doi: 10.1016/j.cmpb.2012.04.004.

[26] R. Dhanya, I. R. Paul, S. Sindhu Akula, M. Sivakumar, and J. J. Nair, "A Comparative Study for Breast Cancer Prediction using Machine Learning and Feature Selection," in 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, May 2019, pp. 1049–1055. doi: 10.1109/ICCS45141.2019.9065563.

[27] H. Byeon and B. Kim, "Applying Synthetic Minority Over-sampling Technique and Support Vector Machine to Develop a Classifier for Parkinson's disease," IJACSA, vol. 12, no. 3, 2021, doi: 10.14569/IJACSA.2021.0120311.

[28] S. Thapa, S. Adhikari, A. Ghimire, and A. Aditya, "Feature Selection Based Twin-Support Vector Machine for the Diagnosis of Parkinson's Disease," in 2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC), Kuching, Malaysia, Dec. 2020, pp. 1–6. doi: 10.1109/R10-HTC49770.2020.9356984.

[29] E. Avuçlu y A. Elen, «Evaluation of train and test performance of machine learning algorithms and Parkinson diagnosis with statistical measurements», Med Biol Eng Comput, vol. 58, n.º 11, pp. 2775-2788, nov. 2020, doi: 10.1007/s11517-020-02260-3.

[30] O. Asmae, R. Abdelhadi, C. Bouchaib, S. Sara, and K. Tajeddine, "Parkinson's Disease Identification using KNN and ANN Algorithms based on Voice Disorder," in 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Meknes, Morocco, Apr. 2020, pp. 1–6. doi: 10.1109/IRASET48871.2020.9092228.

[31] E. Celik and S. I. Omurca, “Improving Parkinson’s Disease Diagnosis with Machine Learning Methods,” in 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), Istanbul, Turkey, Apr. 2019, pp. 1–4. doi: 10.1109/EBBT.2019.8742057.
[32] C. R. Harris et al., “Array programming with NumPy,” Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.
[33] P. Virtanen et al., “SciPy 1.0: fundamental algorithms for scientific computing in Python,” Nat Methods, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.
[34] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” MACHINE LEARNING IN PYTHON, p. 6.
[35] J. Song et al., «Differential diagnosis between Parkinson’s disease and atypical parkinsonism based on gait and postural instability: Artificial intelligence using an enhanced weight voting ensemble model», Parkinsonism & Related Disorders, vol. 98, pp. 32-37, may 2022, doi: 10.1016/j.parkreldis.2022.04.003.

Osiris Escamilla-Luna. His research interests are machine learning, especially its application in medicine. He received the B.Sc. degree in electrical engineering and the M.Sc. degree in engineering from the Juarez Autonomous University of Tabasco (UJAT), Mexico. He is currently pursuing the Ph.D. degree in computer science with the Juarez Autonomous University of Tabasco.

Miguel A. Wister. His research interests are in ubiquitous computing, focusing on ambient intelligence and the Internet of Things (IoT). Professor Wister received his Ph.D. in Information Technology and Communications Engineering from the University of Murcia (UM), Spain, in 2008. He earned an MSc in Informatics Technology at the Monterrey Institute of Technology and Higher Education (ITESM), Mexico, in June 1997, and a BSc degree in Informatics at the Juarez Autonomous University of Tabasco (UJAT), Mexico, in 1993. He has tutored more than 40 BSc and MSc students at the UJAT and Olmeca University. Dr. Wister has also authored and co-authored 30 scientific papers published in national and international peer-reviewed conferences and journals.

Jose Hernandez-Torruco. Doctor in Computer Science from the Juarez Autonomous University of Tabasco (UJAT). Professor of the Academic Division of Information Sciences and Technologies (DACyTI) of the UJAT since 2000. His area of interest is machine learning and data mining, especially its application in medicine. He has published several articles in English and Spanish in peer-reviewed and indexed journals and at national and international conferences. He has been responsible for several institutional projects and has participated as a collaborator in financed projects. In addition, he has directed several theses at the Bachelor's, Master's, and Doctorate levels.