An Automatic Method for Identifying Huntington's Disease using Gait Dynamics

Juliana Paula Felix*[1], Flávio Henrique Teles Vieira*[2], Gabriel da Silva Vieira†[3],
Ricardo Augusto Pereira Franco*[4], Ronaldo Martins da Costa*[5], Rogerio Lopes Salvini*[6]
*Universidade Federal de Goiás, Brazil
† Instituto Federal Goiano, Brazil
{[1]julianafelix, [5]ronaldocosta, [6]rogeriosalvini}@inf.ufg.br, [2]flavio_vieira@ufg.br,
[3]gabriel.vieira@ifgoiano.edu.br, [4]ricardofranco3@gmail.com

***Abstract*—Huntington's Disease (HD) is a genetic disorder that causes the progressive breakdown of nerve cells in the brain, reducing an individual's ability to reason, walk, and speak. Due to its severity, new approaches are important for the development of methods that contribute to the correct classification of this disease. In this paper, we propose an automatic method for diagnosing Huntington's Disease using gait dynamics information. Our approach is divided into a four-stage pipeline: preprocessing, feature extraction, classification, and diagnosis output. We evaluate the performance of our proposed method through well-known classifiers that are commonly used in machine learning problems. A publicly available database on Gait Dynamics in Neuro-Degenerative Disease is used, and the experimental results show that both Support Vector Machines (SVM) and Decision Tree (DT) were able to achieve an average accuracy of 100:0%, representing an improvement in the field.**

***Index Terms*—automatic diagnosis, Huntington's disease, machine learning, gait dynamics.**

I. Introduction

Genetic diseases result from mutations or abnormalities on chromosomes or genes, resulting in a clustering of different symptoms, from psychiatric disorders to several brain damage. Huntington's Disease (HD) is one of those genetic pathologies that causes the progressive breakdown of nerve cells in the brain, reduces the individual's ability to reason, walk and speak, and it is referred to as a fatal genetic disorder which manifests as a triad of motor, cognitive, and psychiatric symptoms which begin insidiously and progress over many years, until the death of the individual [1].

Since HD greatly affects the behavioral control system, which includes the body movement, the use of this information can help the task of identifying and also classifying individuals who suffer from this disease. In fact, it has been previously shown that some neurological diseases present increased fluctuation magnitude as well as an altered fluctuation dynamics [2]. Thus, the design of temporal parameters provides assistance in this task, supporting a way to deal with an automated classification problem as in [3], [4]. An unsteady gait, for instance, shows a pattern that differentiates a person who has a genetic disorder from a person who does not. In previous studies, this observation has been used satisfactorily to classify neurodegenerative diseases [5], or even to distinguish them from healthy people [6].

Well-known machine learning algorithms as Support Vector Machines (SVM), Decision Trees (DT) and Naive Bayes (NB) classifier are often used in pattern genetic disorder recognition in order to extract knowledge from data and to identify common behavior settings. In such proposals, a crucial point is related to the feature extraction which may influence the "learning" stage and, consequently, the assertiveness of the classifiers. Thus, a suitable feature extraction enforces the selection of appropriated descriptors and models can be better designed based on them.

Besides, noisy data disturbs the capability of built models to perform a correct classification, which requires inspection and prior analysis of the data. In general, it is treated by outlier detection strategies in order to identify extreme values that deviate from other observations or even guide in removing anomalies which raise suspicions, i.e., which differ significantly from a given set of data.

In our proposal, we identified noise observations by applying a median filter to replace the detected noise data points that were 3 standard deviations greater than or less than the median value. Furthermore, we investigate the effectiveness of two features, Coefficient of Variation (CV) and Fractal Scaling Index (α), in temporal series that were recorded from gait signals using force-sensitive resistors. The first one is obtained from the mean (μ) and standard deviation (σ) of each time series, while the second one reveals the extent of long-range correlations in time series based on statistical analysis fluctuation. The proposed outlier detection and the feature extraction technique are used to perform a binary classification between a subject with Huntington's Disease and a control subject.

In this paper, we propose an automatic method for the diagnosis of Huntington's Disease using the information from gait dynamics, and we evaluate the performance of five well-known classifiers [7], [8] (Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes (NB), Linear Discriminant Analysis (LDA) and Decision Tree(DT)) in accurately classifying gait signals from an unknown subject under consideration as being from a subject suffering with Huntington's Disease or being from a healthy subject.

The remaining of this paper is as follows. Section II presents some studies that are related to this paper. In Section III,

2375-0197/19/$31.00 ©2019 IEEE
DOI 10.1109/ICTAI.2019.00243

we describe the proposed method, which includes the data description and preprocessing, feature extraction, and details of the classifications performed. Results are showed in Section IV, followed by our conclusions in Section V.

II. Related Work

The analysis of gait variables of neurodegenerative diseases has been widely performed. Here, we review some of the studies that are related to this paper.

Hausdorff et al. [9] hypothesized that the stride-interval correlations would be altered by changes in neurological function associated with aging and certain disease states, and they tested their hypothesis with control subjects and subjects with Huntington's Disease. In their findings, they state that the scaling exponent, obtained after performing a detrended fluctuation analysis, was smaller in the subjects with Huntington's disease compared with disease-free controls.

In a later study, Hausorff et al. [2] analysed subjects with Amyotrophic Lateral Sclerosis (ALS) and found that the gait of patients with ALS is less steady and more temporally disorganized compared with that of healthy subjects. Moreover, they have also found that ALS, as well as Parkinson's Disease and Huntington's Disease, presented an increased stride-to-stride variability compared to healthy control subjects. Other authors, such as Tafazzoli et al. [10] and Joshi et al. [11] explored the problem of diagnosing other neurodegenerative diseases using gait information.

When it comes to automatic methods for the classification of Huntington's Disease using gait signals, some progress has been made. Zeng and Wang [3] used deterministic learning theory to perform a binary classification that distinguishes data from subjects with Huntington's disease and control subjects. Their approach consists of using all data points from the left and right swing interval, and left and right stance interval from each subject as input to the Radial Basis Function (RBF) classifier, leading to an average accuracy of 83.3%.

Baratin et al. [12] used Discrete Wavelet Transform, extracting two features from each of the seven levels of decomposition performed, and then feeding it to a Support Vector Machine (SVM), resulting on the improvement of accuracy to 86.1%. Later on, Gupta et al. [13] used the mutual information criterion to select the most essential features, in a total of 500, from a given dataset, which were used to construct a decision tree classifier, leading to an accuracy of 88.5%. Genetic algorithm to perform feature selection was done by Daliri [6], which was left with 12 features that were then fed to a support vector machine classifier, achieving 90.3% of accuracy.

In our approach, we used data signal obtained from force-sensitive resistors placed in each subject's shoe and collected as they walked. We perform a feature extraction based on metrics of fluctuation magnitude and fluctuation dynamics, leading to a simple feature vector of size two as input for the classifiers. Our results showed that the method proposed is effective, being able to achieve an average accuracy of 100.0% for both support vector machine and decision tree classifiers.

III. Proposed Method

This paper proposes an automated method for identifying Huntington's Disease using information obtained from gait dynamics. In our method, given a time series signal extracted from the gait of a subject, it goes through preprocessing and feature extraction, leading to a feature vector that serves as input to a classifier that outputs the diagnosis, that is, whether the signal belongs to a subject that has Huntington's Disease or not. The general structure of the proposed method is shown in Figure 1, and an explanation of each step follows, starting with the description of the input data.

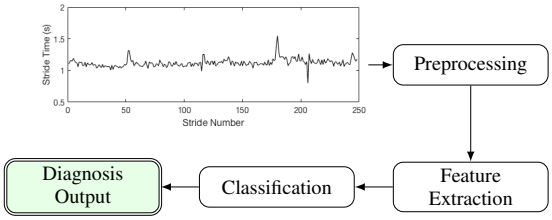


Figure 1: Outline of the proposed method.

A. Dataset Description

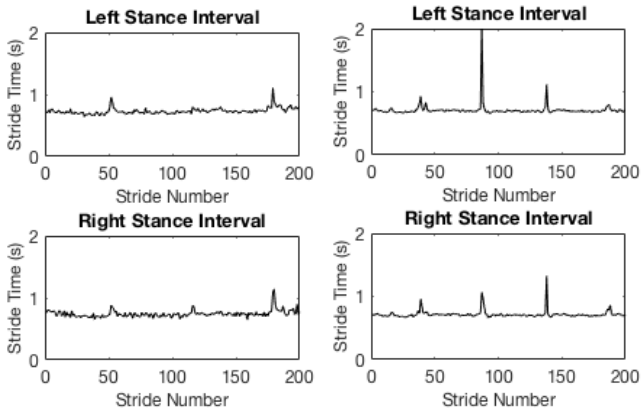
We used the publicly available database on Gait Dynamics in Neuro-Degenerative Disease provided by Hausdorff et al. [2], [9] and available online on the web page of *Physionet* [14]. It consists of a collection of different gait cycle parameters collected using force-sensitive resistors placed in the subject's shoe. Each subject was requested to walk for 5 minutes at their usual pace along a 77m hallway. In total, there are 20 (6 males / 11 females) records from patients with Huntington's disease (HD) and 16 (2 males / 14 females) records from healthy control subjects (CO).

From these force-sensitive resistors, three time series could be derived from each foot (left/right), corresponding to different phases of the gait: stride interval (the time elapsed between the first contact of two consecutive footsteps of the same foot), swing interval (time during which the foot is in the air), and stance interval (the phase during which the foot remains in contact with the ground). Moreover, another time series could be obtained based on information taken from both sensors: the double support interval (time during which both feet are in contact with the ground). An example of the left and right stance interval time series from a subject with HD and the corresponding ones derived from a healthy subject is show in Figure 2.

From each subject that participated in the study, the age, height (m), weight (kg), gender, and the mean gait speed (m/s) have been recorded. This information are displayed in Table I, where μ stands for the mean and SE stands for the standard error of the mean, calculated as $\sigma/\sqrt{n}$, where σ is the standard deviation and n is the total number of observations.

B. Data preprocessing

In order to remove some start-up effects associated with the moment the participants of the experiment began to walk in



(a) Subject with Huntington's Disease. (b) Healthy subject.

Figure 2: An example of time series of stance intervals derived from the left and right foot from participants with Huntington's Disease (Figure 2a) and from a control subject (Figure 2b).

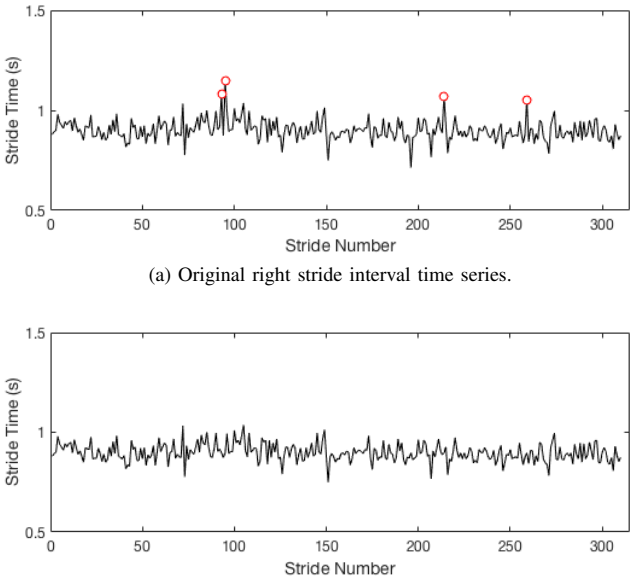
Table I: Statistical data from Huntington's Disease (HD) and Control (CO) subjects.

Group	Age (years)		Height (m)		Weight (kg)		Gait Speed (m/s)	
	μ	SE	μ	SE	μ	SE	μ	SE
HD	47	3	1.83	0.02	72.1	3.7	1.15	0.08
CO	39	4	1.83	0.02	66.8	2.7	1.35	0.04

the hallway, the first 20s of the recorded data from each one of the subjects were removed. Noise observations, mainly due to turnarounds that were occasionally needed, were identified by applying a median filter. To do so, for each time series, we calculated its median value. Then, all data points that were three standard deviations greater than or less than the median value were replaced by the median value of the corresponding time series. Figure 3 shows before and after the application of this process on the right stride interval series of a subject with HD. The red dots in Figure 3a are the points that matched the criteria of the median filter and were identified as outliers. After their values were replaced for the median value of the series, the corresponding time series presented in Figure 3b was generated.

C. Feature Extraction

Previous studies have shown that, in healthy adults, the fluctuation magnitude is relatively small [2], [9]. In adults with certain types of neurological diseases, however, the fluctuation magnitude and the fluctuation dynamics are, in general, altered [9], [15]. Therefore, measures of fluctuation of magnitude (stride-to-stride variability), such as the coefficient of variation (CV), as well as metrics of of fluctuation dynamics (how the stride time changes from one stride to the next), such as the fractal scaling index α, have great potential of differentiating healthy subjects from those suffering from some types of



(a) Original right stride interval time series.

(b) Right stride interval time series after preprocessing.

Figure 3: Before (a) and after (b) applying the median filter on the right stride time series from a subject suffering with Huntington's Disease.

neurodegenerative disorder, including Huntington's Disease [2], [9], [16].

Having that in mind, we were interested in verifying the effectiveness of these two features alone – CV and α – in performing a binary classification between a subject with Huntington's Disease and a control subject. The first feature, the coefficient of variation (CV), is a measure of the magnitude of stride-to-stride variability and gait unsteadiness. It can be determined by calculating $100 \times (\sigma/\mu)$, where μ stands for the mean and σ stands for the standard deviation of each time series from each subject.

Using Detrended Fluctuation Analysis (DFA), we computed the second feature, the fractal scaling index α, which is a measure of the degree to which one stride interval is correlated with previous and subsequent intervals of different time scales. The DFA method [17] is performed as follows. First, the time series to be analyze (with N samples) is integrated. Then, the integrated time series is divided into windows of equal length n – that is, each window considers n strides. In each window, a least squares line is fit to the data, representing the trend in that window. Considering that the y coordinate of the straight line segments is denoted by $y_n(k)$, then, for each window, the integrated time series $y(k)$ is detrended by subtracting the local trend, $y_n(k)$. Next, the root-mean-square fluctuation of this integrated and detrended time series is then calculated by

$$F(n) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k) - y_n(k)]^2}.$$

The detrending process, followed by fluctuation measurement, is repeated over a range of different window sizes n. Then, a log-log graph of $F(n)$ against n is constructed, and the slope of least-square regression line fit to this graph is calculated to be the scaling exponent α. It has been shown that the region of $10 \leq n \leq 20$ provides a statistically robust estimate of stride time correlation properties despite the length of data [9], [18], and thereupon this same range was used here.

Therefore, each time series was finally represented by the coefficient of variation (CV) and its fractal scale index α, making up a feature set of size two, consisted of two numerical values, later used to feed each of the five classifiers.

D. Classification

In machine learning problems, some classifiers such as Support Vector Machine (SVM) [19], K-Nearest Neighbors (KNN) [20], Naive Bayes (NB) [21], Linear Discriminant Analysis (LDA) [22] and Decision Trees (DT) [23] are often employed. In this paper, these well-known and aforementioned classifiers are used and a comparison of their accuracy performance is made, showing which ones get the best assertiveness. For the SVM classifier, we used a linear kernel; the Euclidean distance was used as metric for KNN. Performance validation was carried out using the leave-one-out cross-validation (LOOCV) method [24], a fairly common technique applied in machine learning experiments to estimate generalization error. In every run, one sample was removed from the training data at a time for testing. Our reported classification accuracies represent the average accuracies on the test sets. All experimentation code to perform classification was developed using MATLAB R2017a using Statistics and Machine Learning Toolbox.

IV. Results

Table II and Table III present the average classification accuracies of our method, along with its standard error of the mean (SE), of the five classifiers (SVM, KNN, NB, LDA and DT). In Table II, the results consider features extracted from the left and right foot gait parameters. The results show that, when using parameters derived from the left stride, left swing and left stance interval, overall classification accuracies of 94.4%, 83.3%, and 94.4% were obtained, respectively. On the other hand, the highest average accuracy when right stride and right swing parameters were used turned out to be 97.2%, 91.7%, respectively, while right stance parameters produced an average of 100.0% accuracy.

When double support interval parameters were considered (see Table III), the best overall accuracy achieved was 94.4% when either KNN, NB or DT was used as a classifier. When observing all gait variables presented in Table II and Table III, the right stance interval is the one that stands out, for it produced the highest average accuracy for both SVM, KNN and DT (100.0%, 97.2%, and 100.0%, respectively). Linear Discriminant Analysis' highest accuracy was 83.3%, and it was obtained when left stride or left stance signals were used, while NB's highest accuracy obtained was 94.4% when features extracted from the double support signal was taken as input to the classifier.

Table II: Average classification accuracies when considering left and right foot parameters.

Classifier	Left Foot						Rigth Foot					
	Stride		Swing		Stance		Stride		Swing		Stance	
	μ (%)	SE (%)	μ (%)	SE (%)	μ (%)	SE (%)	μ (%)	SE (%)	μ (%)	SE (%)	μ (%)	SE (%)
SVM	91.7	4.7	83.3	6.3	91.7	4.7	94.4	3.9	88.9	5.3	100.0	0.0
KNN	94.4	3.9	77.8	7.0	80.6	6.7	94.4	3.9	88.9	5.3	97.2	2.8
NB	86.1	5.8	83.3	6.3	86.1	5.8	86.1	5.8	91.7	4.7	91.7	4.7
LDA	83.3	6.3	80.6	6.7	83.3	6.3	80.6	6.7	83.3	6.3	80.6	6.7
DT	88.9	5.3	77.8	7.0	94.4	3.9	97.2	2.8	86.1	5.8	100.0	0.0

Table III: Average classification accuracy when considering double support interval parameters.

Classifier	Double Support	
	μ (%)	SE (%)
SVM	80.6	6.7
KNN	94.4	3.9
NB	94.4	3.9
LDA	58.3	8.3
DT	94.4	3.9

To better analyse each classifier's best performance, we constructed a confusion matrix, presented in Table IV, considering the case where each classifier achieves its highest average accuracy. We can observe that SVM and DT were able to correctly classify all data from both Huntington's Disease and Control classes. KNN has mistakenly classified one subject as being healthy when, in fact, it should be classified as HD. NB confused exactly one HD subject as being from CO, and vice-versa. On the other hand, LDA is the classifier that presented greater confusion, wrongly classifying 5 HD subjects as being healthy and 1 CO subject as having the disease. A summary of our results is presented in Table V. If compared to other results that classifies HD subjects from healthy subjects, such as the ones that mentioned in Section II, our method presented greater assertiveness for some of the classifiers, and was even able to achieve no error at all for SVM and DT, reaching the accuracy of 100%.

V. Conclusions

In this paper, an automatic method for identifying Huntington's disease using gait dynamics was presented. Our proposal considered the fluctuations caused by neurological disorders in order to identify repetitive patterns that conduct to the correct disease classification. Divided into four major steps, the proposed method deals with noisy data through outlier detection and rejection at a preprocessing stage. It presents a feature extraction approach based on the Coefficient of Variation and the Fractal Scaling Index. Moreover, it uses well-known

Table IV: Classifiers confusion matrices.

		Predicted									
		SVM		KNN		NB		LDA		DT	
		HD	CO	HD	CO	HD	CO	HD	CO	HD	CO
Expected	HD	20	0	19	1	19	1	15	5	20	0
	CO	0	16	0	16	1	15	1	15	0	16

Table V: Summary of classification accuracies achieved by each of the five classifiers analyzed in this paper.

Classifier	LDA	NB	KNN	SVM	DT
Accuracy	83.3%	94.4%	97.2%	100.0%	100.0%

machine learning classifiers to perform a binary classification that indicates whether a person has the investigated genetic disorder or not. Then, in the last stage, the results that were obtained are presented and evaluated.

In the experimental design, three parameters of the gait were considered individually from the left and right foot: stride, swing and stance intervals. The results showed a high assertiveness for these variables as well as for the five classifiers used (SVM, KNN, NB, LDA, DT), where each of them achievied more than 80% of accuracy. KNN worked well with the left and right foot stride interval, achieving 94.4% of accuracy. Moreover, SVM and DT reached an impressive 100% accuracy when considering the right foot stance interval.

The method proposed here represents an improvement on previous results that can be found in the literature, and the reduced feature vector size used in our approach, of size two, adds to the novelty of our work, which provides a simple but efficient method for identifying Huntington's Disease using gait dynamics.

Acknowledgment

The first author would like to thank CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil) for the financial support.

References

[1] H. D. S. America. Overview of huntington's disease. Last accessed in July, 2nd, 2019. [Online]. Available: https://hdsa.org/what-is-hd/overview-of-huntingtons-disease/

[2] J. M. Hausdorff, A. Lertratanakul, M. E. Cudkowicz, A. L. Peterson, D. Kaliton, and A. L. Goldberger, "Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis," *Journal of applied physiology*, vol. 88, no. 6, pp. 2045–2053, 2000.

[3] W. Zeng and C. Wang, "Classification of neurodegenerative diseases using gait dynamics via deterministic learning," *Information Sciences*, vol. 317, pp. 246–258, 2015.

[4] W. Aziz and M. Arif, "Complexity analysis of stride interval time series by threshold dependent symbolic entropy," *European journal of applied physiology*, vol. 98, no. 1, pp. 30–40, 2006.

[5] S. M. Keloth, S. P. Arjunan, and D. Kumar, "Computing the variations in the self-similar properties of the various gait intervals in parkinson disease patients," in *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. Seogwipo, South Korea: IEEE, July 2017, pp. 2434–2437.

[6] M. R. Daliri, "Automated diagnosis of alzheimer disease using the scale-invariant feature transforms in magnetic resonance images," *Journal of medical systems*, vol. 36, no. 2, pp. 995–1000, 2012.

[7] T. M. Mitchell, *Machine Learning*, 1st ed. McGraw-Hill, Inc., 1997.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

[9] J. M. Hausdorff, S. L. Mitchell, R. Firtion, C.-K. Peng, M. E. Cudkowicz, J. Y. Wei, and A. L. Goldberger, "Altered fractal dynamics of gait: reduced stride-interval correlations with aging and huntington's disease," *Journal of applied physiology*, vol. 82, no. 1, pp. 262–269, 1997.

[10] F. Tafazzoli, G. Bebis, S. Louis, and M. Hussain, "Genetic feature selection for gait recognition," *Journal of Electronic Imaging*, vol. 24, pp. 31–36, 02 2015.

[11] D. Joshi, A. Khajuria, and P. Joshi, "An automatic non-invasive method for parkinson's disease classification," *Computer methods and programs in biomedicine*, vol. 145, pp. 135–145, 2017.

[12] E. Baratin, L. Sugavaneswaran, K. Umapathy, C. Ioana, and S. Krishnan, "Wavelet-based characterization of gait signal for neurological abnormalities," *Gait & posture*, vol. 41, no. 2, pp. 634–639, 2015.

[13] K. Gupta, A. Khajuria, N. Chatterjee, P. Joshi, and D. Joshi, "Rule based classification of neurodegenerative diseases using data driven gait features," *Health and Technology*, pp. 1–14, 2018.

[14] J. M. Hausdorff, A. Lertratanakul, M. E. Cudkowicz, A. L. Peterson, D. Kaliton, and A. L. Goldberger. Gait dynamics in neuro-degenerative disease data base. Last accessed in March, 10th, 2019. [Online]. Available: https://physionet.org/physiobank/database/gaitndd/

[15] J. M. Hausdorff, M. E. Cudkowicz, R. Firtion, J. Y. Wei, and A. L. Goldberger, "Gait variability and basal ganglia disorders: stride-to-stride variations of gait cycle timing in parkinson's disease and huntington's disease," *Movement disorders*, vol. 13, no. 3, pp. 428–437, 1998.

[16] B. Goldfarb and S. Simon, "Gait patterns in patients with amyotrophic lateral sclerosis." *Archives of physical medicine and rehabilitation*, vol. 65, no. 2, pp. 61–65, 1984.

[17] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of dna nucleotides," *Physical review e*, vol. 49, no. 2, p. 1685, 1994.

[18] C.-K. Peng, S. Buldyrev, A. Goldberger, S. Havlin, M. Simons, and H. Stanley, "Finite-size effects on long-range correlations: Implications for analyzing dna sequences," *Physical Review E*, vol. 47, no. 5, p. 3730, 1993.

[19] I. Steinwart and A. Christmann, *Support vector machines*. Springer Science & Business Media, 2008.

[20] J. Kim, B.-S. Kim, and S. Savarese, "Comparing Image Classification Methods: K-Nearest Neighbor and Support Vector Machines," *Ann Arbor*, vol. 1001, pp. 48 109–2122, 2012.

[21] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[22] G. M. Foody and A. Mathur, "A Relative Evaluation of Multiclass Image Classification by Support Vector Machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 6, pp. 1335–1343, 2004.

[23] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[24] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.