Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 18(1): 495–517.
DOI: 10.3934/mbe.2021027
Received: 01 October 2020
Accepted: 03 December 2020
Published: 10 December 2020

Research article

Machine learning based classification of normal, slow and fast walking by extracting multimodal features from stride interval time series

Wajid Aziz[1,]*, Lal Hussain[2,3], Ishtiaq Rasool Khan[1], Jalal S. Alowibdi[1] and Monagi H. Alkinani[1]

[1] Department of Computer & AI, College of Computer Science and Engineering (CCSE), University of Jeddah, P.O. Box 80327, Jeddah 21589, Saudi Arabia
[2] Department of Computer Science & IT, University of Azad Jammu and Kashmir, King Abdullah Campus, Muzaffarabad 13100, Pakistan
[3] Department of Computer Science & IT, University of Azad Jammu and Kashmir, Neelum Campus, Athmuqam 13230, Pakistan

* **Correspondence:** Email: wloun@uj.edu.sa.

Abstract: The gait speed affects the gait patterns (biomechanical and spatiotemporal parameters) of distinct age populations. Classification of normal, slow and fast walking is fundamental for understanding the effects of gait speed on the gait patterns and for proper evaluation of alternations associated with it. In this study, we extracted multimodal features such as time domain and entropy-based complexity measures from stride interval signals of healthy subjects moving with normal, slow and fast speeds. The classification between different gait speeds was performed using machine learning classifiers such as classification and regression tree (CART), support vector machine linear (SVM-L), Naïve Bayes, neural network, and ensemble classifiers (random forest (RF), XG boost, averaged neural network (AVNET)). The performance was evaluated in term of accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), p-value, area under the receiver operating characteristic curve (AUC). To distinguish the slow and normal gait walking, the highest performance was yielded in terms of accuracy (100%), p-value (0.004), and AUC (1.00) using RF, XGB-L followed by XGB-Tree with accuracy (88%), p-value (0.04) and AUC (1.00). To classify the fast and normal walking, the highest performance was obtained with accuracy (88%), p-value (0.04) using XGB-L, XGB-Tree and AVNET. The highest AUC (0.94) was obtained using NB. To discriminate the fast and slow gait walking, the highest performance was obtained using SVM-R, NNET, RF, AVNET with accuracy (88%), p-value (0.04) and AUC (0.94) using RF and AUC (0.96) using XGB-L.

Keywords: averaged neural network (AVNET); classification and regression tree (CART); gait walking; support vector machine linear (SVM-L)

Abbreviations: ANN: Artificial neural network; AUC: Area under the curve; AVNET: Averaged neural network; CART: Classification and regression tree; CHF: Congestive heart failure; CNN: Central nervous system; DFA: Detrended fluctuation analysis; EEG: Electroencephalographic; GMM-UBM: Gaussian mixture models-universal background models; GPI: Gait-based person identification; MFCCs: Mel frequency cepstral coefficients; MSE: Multiscale sample entropy; NB: Naïve Bayes; NPV: Negative predictive value; NSR: Normal sinus rhythm; PD: Parkinson's disease; PLP: Perceptual lineal prediction; PPV: Positive predictive value; RF: Random forest; ROC: Receiver operating characteristic; SI: Stride interval; SVM-L: Support vector machine-linear

1. Introduction

Walking independently require a firm control to remain flexible and stable while navigating through unpredictable and complex environment [1]. The human gait patterns of a healthy individual exhibits regular fluctuations, which are due to different complexities of nonlinear motor controlling system made from cognitive, neural and mechanical components. The motor controlling system improves the gait to maintain equilibrium, continue progression and remain adaptable [2]. The different gait speeds are outcomes of the complex musculoskeletal system controlled by the central nervous system (CNS) [3]. The quantification of the biomechanical physiognomies of an individual's gait is an imperative clinical tool for evaluating normal and pathological locomotion patterns with effects of walking speed on gait biomechanics [3], and gender differences in gait kinematics [4]. These are used for devising therapeutic interventions [5] as well as to evaluate the intervention outcomes [6].

The locomotor behavior is engendered using the legs by propelling the body over the ground. The joints in the body have more degrees of freedom that is necessary for propelling the body, and the muscles have comparatively more degrees of freedom than the joints due to multiarticular muscles and antagonistic pairs of muscles [7]. The humans use plentiful and redundant degrees of freedom for different gait speeds that plays a vital role for adaptive locomotor behavior. It is however unclear that how the CNS manipulates a huge number of degrees of freedom. The complicated and redundant nature of the musculoskeletal system and differences in the motor outcomes reveal that the motor control in the CNS is extremely complex and control strategies differ for different gaits [7].

During last three decades, it has been clearly demonstrated that the human gait can be adequately analyzed using stride interval time series. The stride interval or gait cycle is the time interval between two heel-strikes of the same foot [2]. Continuous gait cycles during walking are not exactly same. Stride to stride fluctuations are created by small changes from one step to next step over a period i.e. known as gait variability. Gait variability not only represents the magnitude of fluctuations, by measuring the dispersion around the central tendency i.e. standard deviation, but also shows the serial correlation between continuous strides i.e. temporal ordering fluctuations by computing long range power law correlation [8], an empirical examination of detrended fluctuation analysis (DFA) [9] and re-interpreting DFA [10]. Hausdorff et al. employed detrended fluctuation analysis to study human gait variability with aging [11], neurodegenerative diseases [12], and even different walking conditions [13]. They observed that gait dynamics are more random or less correlated in elderly people, in neurodegenerative disease subjects and in subjects walking under constrained protocols. Costa et al. [14] proposed multiscale sample entropy (MSE) to quantitatively measure of complexity of human gait that is large for healthy subjects and correlated stochastic processes. Aziz and Arif [15] applied symbolic entropy (SyEn) to characterize gait dynamics of control and neurodegenerative signals. Hausdorff et al. [11–13] highlighted that gait dynamics has meaning and may be useful for assessing the neural control of locomotion and improving functional assessment of aging, chronic disease, and their effect on the human gait. Goshvarpour [16] used Poincare plots, Hurst exponents, and the Lyapunov exponents to quantify the dynamics of gait signals in healthy subjects who walked at their normal, slow, fast speed. Abbasi and Aziz [17] applied symbolic time series analysis to study the gait dynamics of healthy moving under constrained and unconstrained conditions. Yu et al. [18] proposed multivariate multiscale symbolic entropy to quantitatively measure complexity of normal, slow and fast walking under unconstrained and metronomic walking by considering both within- and cross-channel dependencies as well as coupling in multiple channels complex signals over a range of scales. Recently, Vasquez-Correa et al. [19] proposed Gaussian mixture models - universal background models (GMM-UBM) and i-vectors to evaluated different neurological states of Parkinson's disease (PD) using information from gait, speech and handwriting. San-Segundo et al. [20] applied frequency features and GMM-UBM approach to identify a gait-based person identification (GPI) system that uses inertial signals from a smartphone. They also integrated new feature extraction approach such as mel frequency cepstral coefficients (MFCCs) and perceptual lineal prediction (PLP) coefficients to further improve the results Moreover, San-Segundo et al. [21] recently employed i-vector approach and compared the results with GMM-UBM system. Li et al. [22] applied machine learning based classification methods such as a sensory-motor fusion-based manipulation and grasping control strategy has been developed for a robotic hand-eye system. MFCCs features were extracted by [23] to increase robustness in the detection of freezing of gait in PD, tremor detection in PD [24] and smartphone inertial signals for human activity segmentation [25].

In the past, researchers developed different techniques such as long power law correlation [8], detrended fluctuation analysis (DFA) [9], DFA to study human gait variability with aging [11], neurodegenerative diseases [12], and even different walking conditions [13] to study the complex dynamics of human gait. However, these measures have limited capability to classify patterns of gait during different walking conditions due to the involvement of different cognitive, neural, and mechanical components of motor controlling system. Recently, Hussain et al. proposed diagnosis framework for the diagnosis of epileptic seizures [20], arrhythmia detection [21] and classification of normal sinus rhythm (NSR) and congestive heart failure (CHF) subjects [22] by extracting multimodal features. The evaluation metrics revealed improved classification ability of different classifiers using multimodal features for distinguishing healthy and pathological subjects. Despite of the fundamental dissimilarities in the regulation of heart rate and the regulation of human gait, the success of research in the heart rate variability analysis has open window to similarly explore gait variability.

There are several applications of machine learning techniques in medical applications [26], Data-driven decision-support system [27], Big data management and analytics in scientific Programming [28]. The existing techniques have limitations by not considering the multivariate dynamics of gait stride interval time series during normal, slow fast walking. We proposed and extracted multimodal features from gait stride interval time series to capture the temporal dynamics (short, medium and long term variations), complex dynamics based on entropy-based information theoretic approaches, wavelet based methods and statistical measures in order to provide the enhanced detection performance to distinguish the different walking conditions. The specific aim of the present study was to explore the applicability of multimodal features feature extraction strategy (which includes both linear and nonlinear measures) for classification of normal, slow and fast gait speed using stride interval signals. In this study, we extracted multimodal features from stride interval signals of healthy subjects walking at different speeds and then used the robust machine learning classifiers such as CART, SVM-L, NB, ensemble classifiers (RF, XGB, AVNET) for classification. The performance of different classifiers was evaluated using accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), p-value, area under receiver operating characteristic (AUC) curve.

2. Materials and methods

2.1. Dataset

Human gait is under voluntary control of the central nervous system (CNS). The spatiotemporal variations in different gait speeds are outcome of complex musculoskeletal system, which is controlled by the CNS [3]. The variations in different gait speeds may contain are very useful information about controlling mechanism of human gait, which can be extracted for clinical decision making and devising therapeutic interventions. Based on these characteristics, we proposed to use time domain statistical and entropy-based measures to detect the different gait dynamics during normal, slow and fast walking protocols. The data used in the study was taken from Physionet, which is publicly available "the research resource for complex physiologic signals" [30]. The data comprises of long-term recordings of stride interval (SI) time series of ten young healthy men, who had no history of any respiratory, neuromuscular or cardiovascular problems and were not using any medication. The mean age of the subjects was 21.7 years (range: 18–29 years), height 1.77 ± 0.08 meters (mean ± standard deviation) and weight was 71.8 ± 10.7 kg. The subjects under three different walking protocols, one-hour usual normal walk, one-hour slow pace and one hour at a fast pace. The SI was measured by using force-sensitive switches taped inside one shoe. We split the dataset into training and testing data with a 70% and 30% ratio by using a stratified sampling method. The recordings are not from the same subjects in both sets.

The Figure 1 reflect the stride interval time during one hour of walking from one of the healthy subjects during slow, normal and fast walking rates. We computed the time domain, statistical, entropy and wavelet based 16 features from these stride interval time series data. These features are then passed as input to the robust machine learning classifiers to classify slow vs normal, fast vs normal and fast vs slow stride intervals.

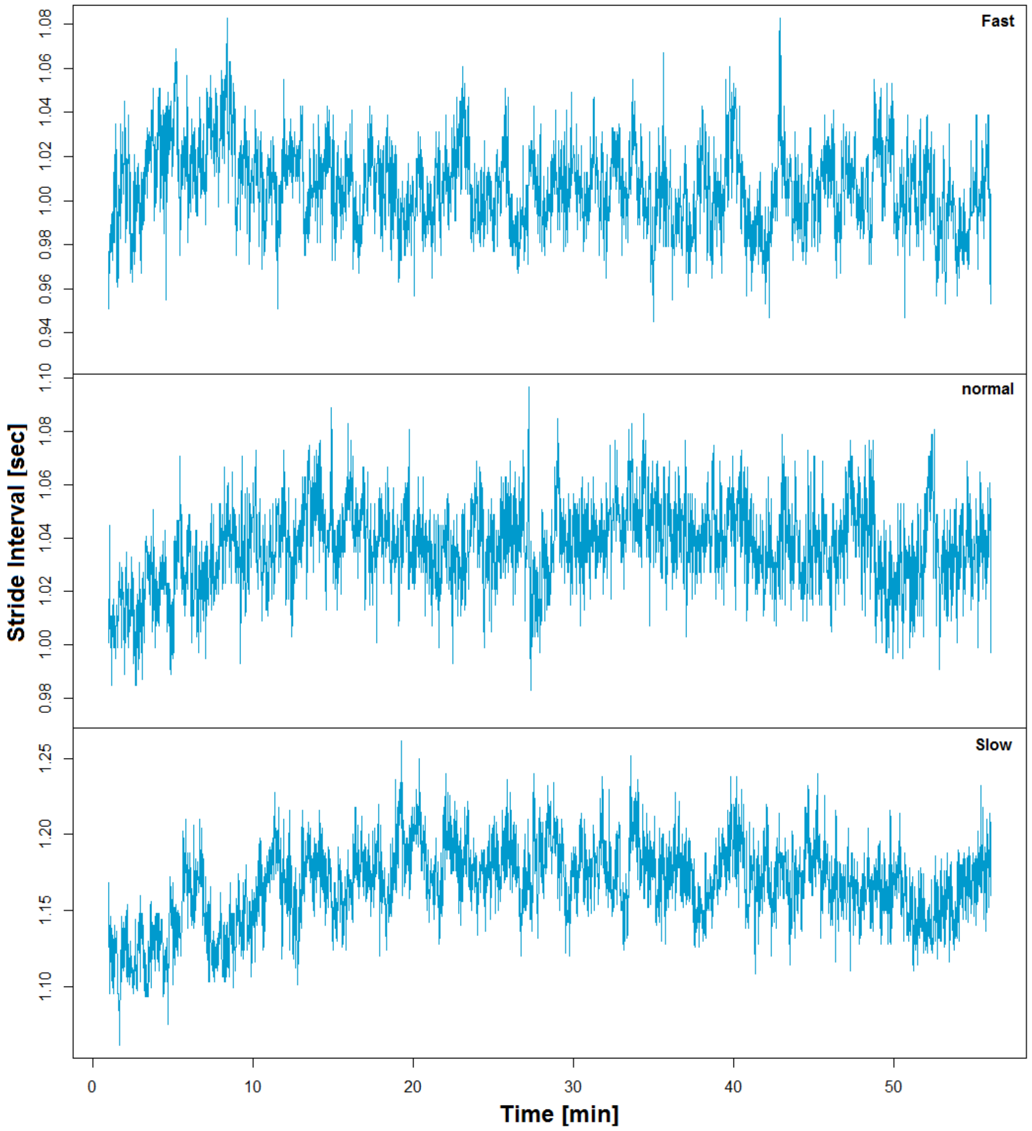


Figure 1. Example of time series of stride time during one hour of walking in a healthy young adult at slow, normal and fast walking rates.

2.2. Feature extraction

Features extraction is one of the most important steps before applying the machine learning and neural networks classification techniques for detection and prediction purposes. It requires an optimum feature set that should effectively discriminate the subjects. Features extraction is solely specific to the problem. We extracted the following features from gait stride interval time series.

2.2.1. Time domain features

Given the stride interval time series $SI = \{SI_j\}, 1 \leq j < N$. Different time domain statistical parameters SDSD, SDSI, RMSSD and SDASI were computed.

SDSD: standard deviation of differences between successive stride intervals (SI).

$$SDSD = Standard\ deviation\ (SI_{j+1} - SI_j)$$

SDSI: standard deviation of the SI time series data.

$$\text{SDSI} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N} (SI_j - \overline{SI})^2} \tag{2.1}$$

RMSSD: Is the square root of the mean squared differences of N successive stride intervals

$$\text{RMSSD} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N-1} (SI_{j+1} - SI_j)^2} \tag{2.2}$$

SDSI: Standard deviation of the average SI-intervals calculated over short periods (5 minutes) segments of the entire Signal. If $\{\mu_1, \mu_2, \mu_3, \ldots, \mu_n\}$ are average values of the segments $\{s_1, s_2, s_3, \ldots, s_n\}$, then

$$SDASI = \frac{1}{N}\sum_{i=1}^{n} \mu_i \tag{2.3}$$

2.2.2. Entropy and wavelet-based features

Biological systems are composed of multiple interacting components exhibiting the complex patterns. This pattern of change contains hidden but very useful information to understand the underlying dynamics of these systems. To compute the dynamics, researchers in the past employed different methods from information theoretic approaches such as seizure detection based on multimodal feature extraction approach [29], symbolic time series analysis to detect seizure [30], complex dynamics of electroencephalographic (EEG) motor movement signals [31], dynamics of alcoholism using sample entropy based on KD tree algorithm [32], lung cancer dynamics based on refined fuzzy entropy [33].

In this study, we extracted following entropy based computational methods to detect the different gait dynamics during slow, fast and normal walking.

(1) Approximate entropy

Approximate entropy (ApEn) developed by [34] is a statistical measure used to quantify the regularities in data. It shows the probability that similar observation patterns do not repeat.

$$\text{ApEn (m, r. N)} = \emptyset^{m}(r) - \emptyset^{m+1}(r) \tag{2.4}$$

The $C^m(r)$ and $C^{m+1}(r)$ are being computed as detailed in [32]. Two parameters are set to measure the average entropy, i.e. m, which is the length of the window, and r, the criterion of similarity. We selected m = 3 and r = 0.15 times the standard deviation of data in this analysis as given in [34].

(2) Fast sample entropy with KD tree algorithmic approach

Sample entropy (SampEn) employed by [35] is a modified form of approximate entropy. It is used to assess the physiological time series signal. Sample entropy when comparing with approximate entropy shows good features like independent data length and trouble-free implementation. It can easily be implemented in many programming languages.

Thus, sample entropy can be more precisely computed using following formula:

$$\mathrm{SampEn(m, r)} = \mathrm{lim_{N\to\infty}} - \mathrm{ln}\frac{P^m(r)}{Q^m(r)} \tag{2.5}$$

Where $P^m(r)$ denotes the probability that two sequences will still match for m+1 points and $Q^m(r)$ is the probability that two sequences will matches for m points (with tolerance of τ); where self matches are excluded. In this regard Eq (2.5) can be expressed as:

$$\mathrm{SampEn(m, r, N)} = -\mathrm{ln}\frac{P^m(r)}{Q^m(r)} \tag{2.6}$$

By setting $Q = \left\{\frac{[(N-m-1)(N-m)]}{2}\right\} Q^m(r)$ and $P = \left\{\frac{[(N-m-1)(N-m)]}{2}\right\} P^m(r)$

We have $\frac{P}{Q} = \frac{P^m(r)}{Q^m(r)}$ and thus sample entropy [36] can be expressed as:

$$\mathrm{SampEn(m, r, N)} = -\ln\left(\frac{P}{Q}\right) \tag{2.7}$$

Where P is the total number of forward matches of length m+1 and Q is the total number of templates matches of length m. Here we used sample entropy with KD tree algorithmic base approached as implemented by [37] which provide improved performance and is more effective with respective to time and space complexity.

(3) Wavelet entropy

Wavelet methods are used in many applications for their nonlinear analysis, commonly used wavelet packet methods [38] are Shannon, log energy, threshold, sure and norm etc. Shannon entropy [38] was employed to measure the complexity of signal to wavelet coefficients generated by WPT where larger values show high uncertainty process and therefore higher complexity. Wavelet entropy used by [39] which provided the useful information to measure the underlying dynamical process associated with the signal. The entropy 'E' must be an additive information cost function such that E (0) = 0 and $\mathrm{E(S)} = \sum_\mathrm{i} \mathrm{E(S_i)}$.

(4) Shannon entropy

The Shannon entropy was proposed by Claude Shannon in 1948 [40]. In addition, it is the measurement of the vulnerability associated with a randomness of the data space. Shannon entropy precisely estimate the predicted value of the results found in a packet. We can describe the Shannon entropy of a random variable S as follows:

$$E(S) = -\sum_{i=1}^{n} S_i^2 log_2(S_i^2) \tag{2.8}$$

Where Si represents coefficients of signal S in an orthonormal basis. If the entropy value is greater than one, the component has a potential to reveal more information about the signal and it needs to be decomposed further in order to obtain simple frequency component of the signal [41]. By using the entropy, it gave a useful criterion for comparing and selection the best basis.

(5) Wavelet entropy

This entropy measure was proposed by [44] can mathematically defined such as:

$$\mathrm{E(S)} = \frac{\sum_i |S_i|^p}{N} \tag{2.9}$$

Where p is the power, the terminal node signal must be $1 \ll p < 2$ and S_i is the terminal waveform signal.

(6) Threshold entropy

E(Si) = 1 if $|Si| > p$ and 0 elsewhere so $E(s) = \#\ \{I$ such that $|Si| > p\}$ is the number of time instants when the signal is greater than a threshold p.

The threshold entropy value was determined using a value of 0.2.

(7) Sure entropy

The threshold of the parameter P and the values of $P \geq 0$ are used.

$$E(s) = n - \#\{i\ such\ that\ |si| \leq \mathrm{p}\} + \textstyle\sum_i min_{(si^2, p^2)} \tag{2.10}$$

Where, the discrete wavelet entropy E is a real number, s is the terminal node signal and (si) i the waveform of terminal node signals. In Sure entropy, p is a positive threshold value and must be $p \geqslant 2$ [43].

The entropy of Sure was measured at threshold 3.

(8) Norm entropy

The P is used in Normal Entropy as the power and value of $P \geq 1$. The intensity in l^p norm entropy is:

$$E(si) = |s_i|^p \tag{2.11}$$

$$so\ E(s) = \sum_i |s_i|^p{}_i = ||S||_p^p$$

The entropy of the norm was estimated at 1.1 with power.

The wavelet norm entropy represents the ordering of nonstationarity of time series fluctuation.

(9) Log energy

$$H_{logEn}(B) = -\textstyle\sum_{i=0}^{N-1} (log_2(\,Pi(B)))^2 \tag{2.12}$$

Where $Pi(B)$ denotes the function of probability distribution and is a logarithmic amount of the distribution square of these probabilities.

2.3. Classification methods

To classify different gait stride interval time series data, following robust machine learning algorithms were used.

2.3.1. Support vector machine (SVM)

Based on the empirical error minimization, the traditional methods due to small sample cases are prone to generate the overfitting problems, whereas, SVM has a good generalization ability due to the structural risk minimization principle [44]. Moreover, SVM is also appropriate [45] and provide good generalization even if the training set has some bias. SVM is also appropriate when the dataset has many features [46]. It is successfully used in many applications such as machine learning machine learning [47], pattern recognition problems [48], and medical diagnosis area [49,50] etc.

2.3.2. Naïve Bayes (NB)

The NB [51] algorithm is based on Bayesian theorem [52] and it is suitable for higher dimensionality problems. NB relate with a family of probabilistic classifier and established on Bayes theorem containing compact hypothesis of independence among several features. NB techniques were greatly biased because its probability computation errors are large. Due to the better performance [53], NB is presently used in variety of applications in present advance developments [54–57].

2.3.3. Decision tree (DT)

DT is proposed by Breiman in 1984 [58], are decision support tools of machine learning and data mining for the large size of input data, which predict the target value or class label based on several input variables. In DT, the classifier compares and check the similarities in the dataset and ranked it into distinct classes. Wang et al. [59] used DT algorithm for classifying the data based on choice of an attribute which maximizes and fix the data division. Until the conclusion criteria and condition are met, the attributes of datasets are split into several classes.

2.3.4. Ensemble classifiers

The ensemble classifiers contain a set of individually trained classifiers, their estimates are then combined when classifying the different instances using different methods [60]. These classifiers are constructed by several learning algorithms and then predict new data points by adding the weight of their predictions. Following ensemble classifiers were used in this study:

(1) Random forest (RF)

Random forest (RF) is another type of machine learning classifier which is operated by constructing an assembly of decision trees. The result is achieved by averaging the output founded from all DTs. [61]. Breiman in 2001 developed RF model by taking an extra layer with bagging strategies. It has important applications in regression, classification and in multi selections. [62]. It is a best classifier for categorization, prediction and regression purposes [63].

(2) XGBoost classifier

Chen and Guestrin proposed XGBoost a gradable machine learning system in 2016 [64]. This system was the most popular and standard system when it was employed in the field of machine learning Kaggle in 2015 however it was employed in 17 solutions out of a total of 29 solutions. Based on the good performance, the XGBoost machine learning classifier was used to train and detect the wind turbine fault. The Gradient boosting is original model of XGBoost, which combines and relate the weak base with stronger learning models by an iterative manner [65].

(3) Averaged neural network (AVNET) model

The AVNET contain multiple neural network models applying the same dataset and predicts using the average of the predictions outcome from all of the constituent model [66]. Due to models initialization on either by fitting the models or different random number seeds on bootstrap data samples of the original training set (i.e. bagging [67] the neural network) the models may be dissimilar to each other. When multiple neural networks are fit at different random number seeds, all the resultant models are used for prediction, by averaging model scores are first Ripley 1996. In classification problems, the average class probabilities or prior probabilities are produced the final class prediction as opposed by voting from the individual class predictions [68].

2.4. Performance evaluation measures

There is variety of measures that are generally employed to compute the proposed system performance. For the detection of breast cancer, using machine learning classifiers performance can be measured by computing PPV, NPV, specificity, sensitivity, and total accuracy.

TN: correct classification of normal.

FN: incorrect classification of normal.

TP: correct classification of abnormal.

FP: incorrect classification of abnormal.

2.4.1. Accuracy

Measure of usefulness or effectiveness of the classification scheme is called Accuracy. Following equation can be employed for the computation of accuracy reflected below:

$$\text{Total Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{2.13}$$

2.4.2. Sensitivity

Measure of the classifier's ability for identifying the positive class patterns is called sensitivity. It is the probability of positive test given that patient suffering from disease. Also known as True Positive Rate (TPR). The following equation is used to compute the sensitivity represented by:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \tag{2.14}$$

2.4.3. Specificity

Specificity is used to calculate the classifier's ability for identifying negative class patterns. It measures the proportion of negatives that are identified correctly. Also known as False Positive Rate (FPR). It can be obtained by using the following equation represented by:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{2.15}$$

2.4.4. Positive predictive value (PPV)

Mathematically PPV can be expressed as below:

$$PPV = \frac{TP}{TP+FP} \tag{2.16}$$

where TP denote that the test makes a positive prediction and subject has positive result under gold standard while FP is the event that test make a positive prediction and subject make negative result.

2.4.5. Negative predictive value (NPV)

Mathematically NPV can be expressed using below equation:

$$NPV = \frac{TN}{TN+FN} \tag{2.17}$$

where TN represents that test make negative prediction and subject has also negative results, while FN represents that test make negative prediction and subject has positive result.

2.5. Area under the receiver operating characteristic (ROC) curve

The ROC is graphed against the true positive rate (TPR), i.e. sensitivity and false positive rate (FPR), i.e. the slow & fast, slow & normal and fast & normal subjects' specificity values. The mean values of features for slow & fast subjects are graded as 1 and 0 respectively, the same was repeated for other combinations. The ROC function is then transferred to this vector, which plots each sample value against the values of specificity and sensitivity. ROC is one of the popular methods of calculating success in order to diagnose and interpret the efficacy of a classifier [69]. The sensitivity is graphed against the y-axis, and the x-axis is graphed against the 1-Specificity. The portion of a square unit is represented by the area under the receiver operating characteristic (AUC) curve. Its value varies from 0 to 1. The distinction is shown by AUC > 0.5. The superior diagnostic tool is shown by the greater AUC. Sensitivity represents right positive cases calculated by dividing the total positive cases, while Specificity represents negative cases expected as positive, calculated by dividing the total number of negative cases.

3. Results

In this study, we extracted the multimodal features extracting approach to compute the dynamics of gait data with normal, slow and fast walking. Based on the spatiotemporal and complex dynamics, we computed time domain, statistical and entropy-based complexity methods and wavelet packet based on gait dynamics with normal, slow and fast walking. We applied machine learning classifiers such as CART, SVM linear, Naïve Bayes, Neural Network and ensemble classifiers such as random forest, XGBoost, AVNET.

The performance to distinguish the fast and slow gait walking based on different classifiers was computed in terms of accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), p-value and area under the receiver operator characteristic (AUC) curve. The classifiers ability to identify the positive class patterns is measured using sensitivity while negative class patterns was measured using specificity. In most of the classifiers, either sensitivity and PPV are higher or specificity and NPV are higher, so merely relying on these measures is not appropriate to judge the overall classifiers performance. So, AUC which is computed in combination of sensitivity and specificity was measured – a more reliable measure to measure the classification performance alongwith P-value and accuracy. The sensitivity, specificity, PPV and NPV values of the ensemble classifiers are higher as reflected in Table 1, which resulted in higher AUC. The highest performance in terms of accuracy and p-value to classify fast vs slow gait walking was obtained using SVM-R, NNET, RF and AVNET accuracy (88.0%), p-value (0.04) followed by SVM-L, NV, XGB-L and XGB-Tree with accuracy (75%) and p-value (0.14). Moreover, the highest performance in terms of AUC was obtained using RF with AUC (0.94) followed by XGB AUC (0.94), XGB-Tree and AVNET AUC (0.91), NNET with AUC (0.88), NB with AUC (0.81) as reflected in Table 1. A highest sensitivity and NPV of 100% was obtained using SVM-R, NB, NNET, RF and AVNET. Moreover, the highest PPV (100%) was obtained using XGB-L followed by PPV (80%) using SVM-R, NNET, RF and AVNET.

Table 1. Classification accuracy of fast and slow gait walking based on different classifiers.

Classifier	accuracy	sensitivity	specificity	PPV	NPV	P-Value	AUC
CART	0.50	1.00	0.00	0.50	0.00	0.64	0.50
SVM-L	0.75	0.75	0.75	0.75	0.75	0.14	0.68
SVM-R	0.88	1.00	0.75	0.80	1.00	0.04	0.81
NB	0.75	1.00	0.50	0.67	1.00	0.14	0.81
NNET	0.88	1.00	0.75	0.80	1.00	0.04	0.88
RF	0.88	1.00	0.75	0.80	1.00	0.04	0.94
XGB-L	0.75	0.50	1.00	1.00	0.67	0.14	0.94
XGB-Tree	0.75	0.75	0.75	0.75	0.75	0.14	0.91
AVNET	0.88	1.00	0.75	0.80	1.00	0.04	0.91

Table 2. Classification accuracy of fast and normal gait walking based on different classifiers.

Classifier	accuracy	sensitivity	specificity	PPV	NPV	P-Value	AUC
CART	0.50	1.00	0.00	0.50	NA	0.64	0.50
SVM-L	0.63	0.50	0.75	0.67	0.60	0.36	0.87
SVM-R	0.88	0.75	1.00	1.00	0.80	0.04	0.93
NB	0.75	0.75	0.75	0.75	0.75	0.14	0.94
NNET	0.88	0.75	1.00	1.00	0.80	0.04	0.88
RF	0.88	1.00	0.75	0.80	1.00	0.04	0.94
XGB-L	0.88	1.00	0.75	0.80	1.00	0.04	0.88
XGB-Tree	0.88	1.00	0.75	0.80	1.00	0.04	0.88
AVNET	0.88	0.75	1.00	1.00	0.80	0.04	0.84

To distinguish the fast and normal gait dynamics, the highest performance in terms of accuracy (88.0%), p-value (0.04) were obtained using NNET, RF, XGB-L, XGB-Tree and AVNET. The highest performance in terms of AUC was obtained NB & RF with AUC (0.94) followed by XGB-L & XGB-Tree with AUC (0.88), SVM-L with AUC (0.87), NNET with AUC (0.63), AVNET with

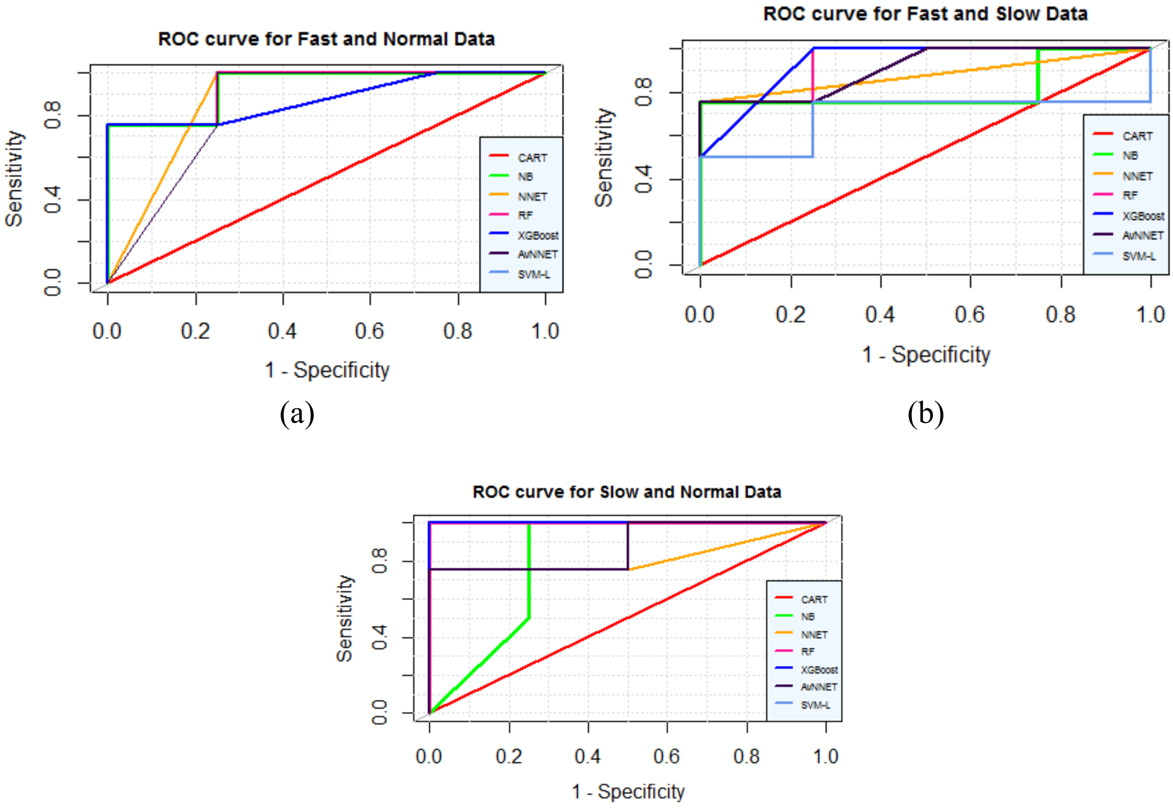
AUC (0.84) CI as reflected in Table 2 alongwith other performance metrics. The other performance metrics computed are reflected in Table 2.

Table 3. Classification accuracy of slow and normal gait walking based on different classifiers.

Classifier	accuracy	sensitivity	specificity	PPV	NPV	P-Value	AUC
CART	0.50	1.00	0.00	0.50	NA	0.64	0.50
SVM-L	0.88	1.00	0.75	0.80	1.00	0.04	0.75
SVM-R	0.75	0.50	0.75	0.75	0.75	0.14	0.81
NB	0.75	0.50	1.00	1.00	0.67	0.14	0.81
NNET	0.75	0.75	0.75	0.75	0.75	0.14	0.81
RF	1.00	1.00	1.00	1.00	1.00	0.004	1.00
XGB-L	1.00	1.00	1.00	1.00	1.00	0.004	1.00
XGB-Tree	0.88	1.00	0.75	0.80	1.00	0.04	1.00
AVNET	0.75	0.75	0.75	0.75	0.75	0.14	0.88

To classify the slow vs normal gait walking subjects, the highest performance was obtained using RF and XGB-L with accuracy (100%), p-value (0.004) followed by SVM-R, XGB-Tree with accuracy (88%), p-value (0.04), NB, NNET, AVNET with accuracy (75%), p-value (0.14). The highest performance in terms of AUC was obtained using RF, XGB-L, XGB-Tree with AUC (1.00) followed by AVNET with AUC (0.88), NB & NNET, SVM-R with AUC (0.81) as depicted in Table 3 alongwith other performance measures. The other performance metrics computed are reflected in Table 3.

The Figure 2(a–c) shows the AUC to classify different gait walking speeds using strides interval signals by extracting multimodal features and employing robust machine learning techniques. To classify fast vs slow gait walking, the highest separation was obtained using RF, XGB-L with AUC (0.94) followed by XGB-tree, AVNET with AUC (0.91), NNET AUC (0.88), NB, SVM-R AUC (0.81), SVM-L AUC (0.68) and CART with AUC (0.50) as depicted in Figure 2(a). To distinguish the fast and normal walking, the highest separation was obtained using NB, RF with AUC (0.94) followed by SVM-R with AUC (0.93); NNET, XGB-L, XGB-tree with AUC (0.88), SVM-L with AUC (0.87), AVNET AUC (0.84) and CART with AUC (0.50) as reflected in Figure 2(b). Moreover, to distinguish the slow and normal gait walking, the highest separation was obtained using RF, XGB-L, XGB-tree with AUC (1.00) followed by AVNET with AUC (0.88), SVM-R, NB, NNET with AUC (0.81), SVM-L with AUC (0.75) and CART (0.50) as shown in Figure 2(c). The results reveal that the highest separation was obtained to distinguish slow vs normal followed by slow vs fast and fast vs normal showing that high differences among the gait stride interval walking between these groups accordingly.



(a)

(b)

(c)

Figure 2. The area under the receiver operating characteristic (AUC) curve to classify different gait walking speeds using stride interval signals a) fast vs normal, b) fast vs slow, c) slow vs normal.

Feature ranking algorithms are mostly used for ranking features independently without using any supervised or unsupervised learning algorithm. A specific method is used for feature ranking in which each feature is assigned a scoring value, then selection of features will be made purely on the basis of these scoring values [70]. The selection of Chi-square features is that they rank the features based on the statistical significance test and only take into account those features that depend on the class label. As the use of training data improves the ability to distinguish between classes with similar characteristics, supervised classifiers are highly consistent and produce precise results. We applied Chi-Squared method to rank the features in order to determine the overall feature important from all four datasets. So, we selected top three important features out of total sixteen features and then we reduced number of features in train set by including only the data of top three selected features i.e. wavelet threshold (WTh), wavelet log energy (WLogEn), wavelet shannon (WShannon) as reflected in Figure 3 followed by root mean square (RMS), wavelet sure (WSure), RMSSD, and SDSSD.

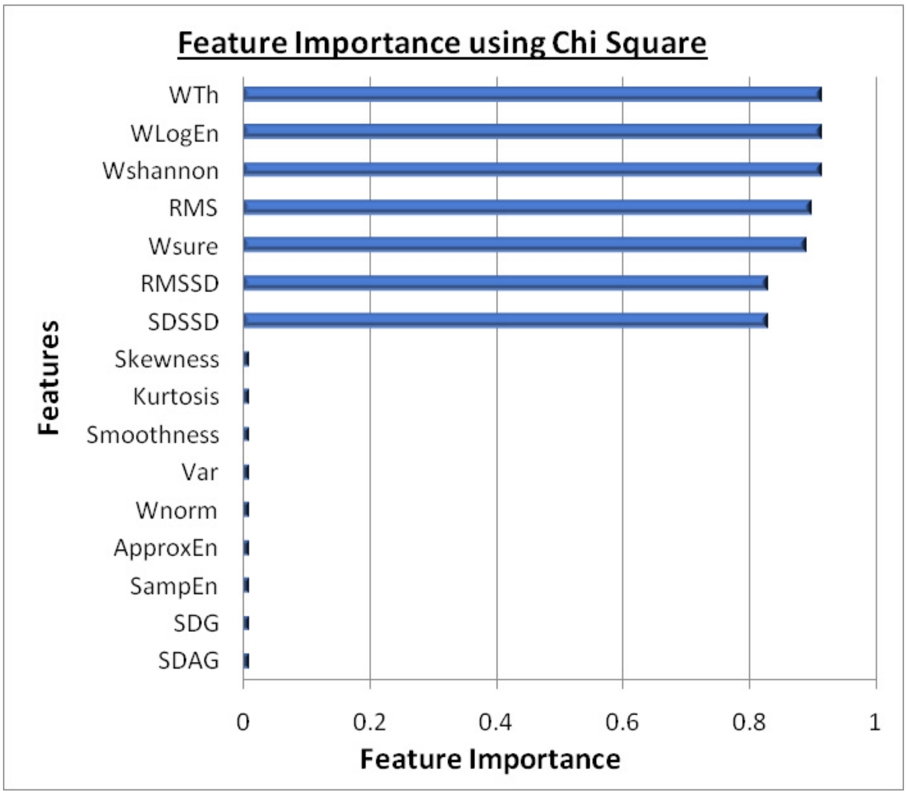


Figure 3. Feature importance with Chi squared feature selection method

Table 4. Comparing means with independent sample t-test to distinguish fast from slow gait walking.

Feature	Sign. (2-tailed)	Mean diff	Std. Error Diff.	95 % CI Lower	95% CI Upper
SDASI	0.09842	−0.0053	0.0031	−0.0117	0.0011
SDSI	0.03168	−0.0285	0.0122	−0.0543	−0.0028
SDSD	0.00874	−0.0114	0.0039	−0.0196	−0.0033
RMSSD	0.00532	−0.0176	0.0055	−0.0292	−0.0059
SampEn	0.39450	0.0641	0.0735	−0.0903	0.2186
Approx.En	0.33056	−0.0009	0.0009	−0.0027	0.0009
WShannon	0.00139	2654.8444	703.4682	1176.9126	4132.7761
WlogEn	0.00012	−1270.6982	259.9173	−1816.7643	−724.6322
Wth	0.00024	715.3000	156.8173	385.8390	1044.7610
WSure	0.00037	−1862.8891	426.3469	−2758.6106	−967.1676
WNorm	0.00986	−138.5148	48.0112	−239.3825	−37.6471
RMS	0.00066	−0.3021	0.0735	−0.4565	−0.1477
Var	0.14395	−0.0032	0.0021	−0.0076	0.0012
Smoothness	0.26438	0.0000	0.000003	0.0000	0.0000
Kurtosis	0.42277	−0.1911	0.2330	−0.6806	0.2983
Skewness	0.25048	−0.1708	0.1438	−0.4729	0.1314

Table 3 reflect the statistical analysis of feature extracted from stride interval time series to distinguish the fast and slow gait walking to compare means using independent sample t-test. We have extracted 16 features from time domain (SDASI, SDSI, SDSD, RMSSD), entropy and wavelet based features (SampEn, Approx.En, WShannon, WlogEn, Wth, WSure, WNorm) and statistical features (RMS, Var, Smoothness, Kurtosis, Skewness) from stride interval time series data. Table 4 reflects the significance values, mean difference, standard error difference and 95% confidence interval of difference with lower and upper bound. The highest significance was obtained with wavelet features such as Wavelet log energy with P-value (0.00012) followed by wavelet threshold with P-value (0.00024), wavelet sure P-value (0.00037), RMS P-value (0.00066), wavelet Shannon P-value (0.00139), RMSSD P-value (0.00532), SDSD P-value (0.00874), wavelet norm P-value (0.00986) etc.

The results reveal that these features are accordingly important to distinguish these stride interval time series conditions.

4. Discussions

Biological systems work in a coordinated manner and generate information in form of biological signals. Due to the dynamical behavior of these systems, the biological signals exhibit stochastic and non-stationary behavior including Stride-to-stride variations of gait cycle timing in parkinson’s disease (PD) [13], human gait dynamics based on multiscale entropy [14], stride intervals using threshold dependent symbolic entropy [15].

Like other biological signals, human also exhibits complex dynamical behavior due to the controlling mechanism exercised by CNS through musculoskeletal system [3]. The involvement of large number of muscles which contribute to human movement, and they show distinct and complex activation patterns during different walking speed. Numerous studies demonstrated that stride interval time series of healthy individual and walking with normal speed exhibit higher complexity, which decreases with aging, diseases and under constrained walking conditions.

Based on the nonstationary and complex dynamics of gait walking patterns, we extracted multimodal linear and nonlinear features of stride interval time series data of health subject walking with normal, slow and fast speeds. In this study, after extraction the multimodal features, we used robust machine learning classifiers such as CART, SVM, RF, XBG, NNET and AVNET. The performance of classifiers evaluated using specificity, sensitivity, PPV, NPV, accuracy and AUC revealed classification ability proposed framework. The highest performance to classify fast vs slow walking was obtained using SVM-R, NNET, RF, AVNET with accuracy (88%), sensitivity & NPV (100%), p-value (0.04), while the highest separation was obtained using RF and XGB-L with AUC (0.94) followed by XGB-tree, AVNET with AUC (0.91). Likewise, to classify the normal and fast walking, the highest performance was obtained using SVM-R, NNET, XGB-L, XGB-tree with accuracy (88%), p-value (0.04), while the highest separation was obtained using NB, RF with AUC (0.94) followed by SVM-L with AUC (0.93). Moreover, to classify the slow and normal walking, the highest performance was obtained using RF, XGB-L with accuracy (100%), p-value (0.004) and AUC (1.00).

5. Conclusions

The aim of this study is to predict the changes in gait dynamics when the participant walk at normal, fast and slow speeds. The gait rhythmical changes during fast, slow and normal walling are of nonlinear and non-stationary and temporal based variations. To quantify these dynamics, we extracted time domain and entropy-based complexity features. To predict the rhythmical changes between different gait walking, we applied robust machine learning techniques such as support vector machine, classification and regression tree (CART), Naïve Bayes, ensemble classifiers such as random forest, XGBoost and averaged neural network (AVNET). The prediction performance was computed in terms of different performance measuring metrics such as specificity, sensitivity, positive predictive value, negative predictive value, accuracy, area under the receiving operating curve and p-value. The results reveal that the proposed multimodal features using robust machine learning algorithms can be very useful to predict the changes in gait rhythms which can be very helpful in predicting the dynamics between the subsystems. In this study, we classified the healthy subject walking with different speeds. The proposed framework can be used to classify control and neuro-degenerative subjects as well as risk of falls in elderly subjects.

6. Limitations of study and future recommendations

Currently, we extracted multimodal features and employed robust machine learning techniques on a small dataset with a smaller number of examples. In future we will apply the proposed methods on larger dataset with other pathologies and clinical profiles of the patients. We will compute the classification performance based on the ranked features. We will also compute the association among features in order to determine the strength of association, which will further assist the clinicians for improving the diagnosis. We will also consider other important aspects for feature extracting strategies in order to further improve the classification performance. This approach will be used to detect the human physical activity based on different inputs on human inertial signals acquired through smartphones in order to improve the healthy lifestyle and comfort of the people. The multimodal feature extracting approach will also help to recognize a user identity based on their gait. Further multimodal feature extracting approach from gait stride intervals will include in improvement of biometric recognition, user-friendliness and security etc.

Acknowledgments

The authors extend their appreciation to the research is funded by the Deanship of Scientific Research, University of Jeddah, Saudi Arabia vide grant number UJ-02-013-ICGR.

Conflict of interest

The authors declare no conflict of interest in this paper.

References

1. J. B. Kiriella, V. E. Di Bacco, K. L. Hollands, W. H. Gage, Evaluation of the effects of prescribing gait complexity using several fluctuating timing imperatives, *J. Mot. Behav.*, **52** (2020), 570–577.

2. J. M. Hausdorff, C. K. Peng, Z. Ladin, J. Y. Wei, A. L. Goldberger, Is walking a random walk? Evidence for long-range correlations in stride interval of human gait, *J. Appl. Physiol.*, **78** (1995), 349–358.
3. C. A. Fukuchi, R. K. Fukuchi, M. Duarte, Effects of walking speed on gait biomechanics in healthy participants: a systematic review and meta-analysis, *Syst. Rev.*, **8** (2019), 153.
4. A. Phinyomark, S. T. Osis, B. A. Hettinga, D. Kobsar, R. Ferber, Gender differences in gait kinematics for patients with knee osteoarthritis, *BMC Musculoskelet. Disord.*, **17** (2016), 157.
5. M. P. Kadaba, H. K. Ramakrishnan, M. E. Wootten, Measurement of lower extremity kinematics during level walking, *J. Orthop. Res.*, **8** (1990), 383–392.
6. S. R. Simon, Quantification of human motion: gait analysis—benefits and limitations to its application to clinical problems, *J. Biomech.*, **37** (2004), 1869–1880.
7. Z. Chen, P. C. Ivanov, K. Hu, H. E. Stanley, Effect of nonstationarities on detrended fluctuation analysis, *Phys. Rev. E.*, **65** (2002), 041107.
8. M. R. Pierrynowski, A. Gross, M. Miles, V. Galea, L. McLaughlin, C. McPhee, Reliability of the long-range power-law correlations obtained from the bilateral stride intervals in asymptomatic volunteers whilst treadmill walking, *Gait Posture*, **22** (2005), 46–50.
9. S. Damouras, M. D. Chang, E. Sejdić, T. Chau, An empirical examination of detrended fluctuation analysis for gait data, *Gait Posture*, **31** (2010), 336–340.
10. J. B. Dingwell, J. P. Cusumano, Re-interpreting detrended fluctuation analyses of stride-to-stride variability in human walking, *Gait Posture*, **32** (2010), 348–353.
11. J. M. Hausdorff, P. L. Purdon, C. K. Peng, Z. Ladin, J. Y. Wei, A. L. Goldberger, Fractal dynamics of human gait: stability of long-range correlations in stride interval fluctuations, *J. Appl. Physiol.*, **80** (1996), 1448–1457.
12. J. M. Hausdorff, S. L. Mitchell, R. Firtion, C. K. Peng, M. E. Cudkowicz, J. Y. Wei, et al., Altered fractal dynamics of gait: reduced stride-interval correlations with aging and Huntington's disease, *J. Appl. Physiol.*, **82** (1997), 262–269.
13. J. M. Hausdorff, M. E. Cudkowicz, R. Firtion, J. Y. Wei, 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, *Mov. Disord.*, **13** (1998), 428–437.
14. M. Costa, C. K. Peng, A. L. Goldberger, J. M. Hausdorff, Multiscale entropy analysis of human gait dynamics, *Phys. A Stat. Mech. Its Appl.*, **330** (2003), 53–60.
15. W. Aziz, M. Arif, Complexity analysis of stride interval time series by threshold dependent symbolic entropy, *Eur. J. Appl. Physiol.*, **98** (2006), 30–40.
16. A. Goshvarpour, A. Goshvarpour, Nonlinear analysis of human gait signals, *Int. J. Inf. Eng. Electron. Bus.*, **4** (2012), 15–21.
17. A. Q. Abbasi, W. A. Loun, Symbolic time series analysis of temporal gait dynamics, *J. Signal Process. Syst.*, **74** (2014), 417–422.
18. J. Yu, J. Cao, W. H. Liao, Y. Chen, J. Lin, R. Liu, Multivariate multiscale symbolic entropy analysis of human gait signals, *Entropy*, **19** (2017), 557.
19. J. C. Vasquez-Correa, T. Bocklet, J. R. Orozco-Arroyave, E. Noth, Comparison of user models based on GMM-UBM and i-vectors for speech, handwriting, and gait assessment of Parkinson's disease patients, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020. Available from: https://ieeexplore.ieee.org/abstract/document/9054348.

20. R. San-Segundo, R. Cordoba, J. Ferreiros, L. F. D’Haro-Enríquez, Frequency features and GMM-UBM approach for gait-based person identification using smartphone inertial signals, *Pattern Recognit. Lett.*, **73** (2016), 60–67.
21. R. San-Segundo, J. D. Echeverry-Correa, C. Salamea-Palacios, S. Lebai Lutfi, J. M. Pardo, I-vector analysis for Gait-based Person Identification using smartphone inertial signals, *Pervasive Mob. Comput.,* **38** (2017), 140–153.
22. Y. Hu, Z. Li, G. Li, P. Yuan, C. Yang, R. Song, Development of sensory-motor fusion-based manipulation and grasping control for a robotic hand-eye system, *IEEE Trans. Syst. Man, Cybern. Syst.*, **47** (2016), 1169–1180.
23. R. San-Segundo, R. Torres-Sánchez, J. Hodgins, F. De la Torre, Increasing robustness in the detection of freezing of gait in Parkinson’s disease, *Electronics*, **8** (2019), 119.
24. A. Zhang, R. San-Segundo, S. Panev, G. Tabor, K. Stebbins, A. S. Whitford, et al., Automated tremor detection in Parkinson’s disease using accelerometer signals, *Proceedings of the 2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies*, ACM, New York, NY, USA, 2018. Available from: https://dl.acm.org/doi/abs/10.1145/3278576.3278582.
25. R. San-Segundo, J. M. Montero, R. Barra-Chicote, F. Fernández, J. M. Pardo, Feature extraction from smartphone inertial signals for human activity segmentation, *Signal Process.*, **120** (2016), 359–372.
26. H. Zhou, J. Tang, H. Zheng, Machine learning for medical applications, *Sci. World J.*, **2015** (2015), 1–1.
27. H. Ma, Y. Zuo, T. Li, C. L. P. Chen, Data-driven decision-support system for speaker identification using E-Vector system, *Sci. Program.*, **2020** (2020), 1–13.
28. H. Wu, M. Liu, S. Zhang, Z. Wang, S. Cheng, Big data management and analytics in scientific programming: A deep learning-based method for aspect category classification of question-answering-style reviews, *Sci. Program.*, **2020** (2020), 1–10.
29. L. Hussain, Detecting epileptic seizure with different feature extracting strategies using robust machine learning classification techniques by applying advance parameter optimization approach, *Cogn. Neurodyn.*, **12** (2018), 271–294.
30. L. Hussain, W. Aziz, J. S. Alowibdi, N. Habib, M. Rafique, S. Saeed, et al., Symbolic time series analysis of electroencephalographic (EEG) epileptic seizure and brain dynamics with eye-open and eye-closed subjects during resting states, *J. Physiol. Anthropol.*, **36** (2017), 21.
31. L. Hussain, W. Aziz, S. Saeed, S. A. Shah, M. S. A. Nadeem, A. Awan, et al., Complexity analysis of EEG motor movement with eye open and close subjects using multiscale permutation entropy (MPE) technique, *Biomed. Res.*, **28** (2017), 7104–7111.
32. L. Hussain, W. Aziz, S. Saeed, S. A. Shah, M. S. A. Nadeem, I. A. Awan, et al., Quantifying the dynamics of electroencephalographic (EEG) signals to distinguish alcoholic and non-alcoholic subjects using an MSE based K-d tree algorithm, *Biomed. Eng.-Biomed. Tech.*, **63** (2018), 481–490.
33. L. Hussain, W. Aziz, A. A. Alshdadi, M. S. A. Nadeem, I. R. Khan, Q. A. Chaudhry, Analyzing the dynamics of lung cancer imaging data using refined fuzzy entropy methods by extracting different features, *IEEE Access*, **7** (2019), 64704–64721.
34. S. M. Pincus, Approximate entropy as a measure of system complexity, *Proc. Natl. Acad. Sci.*, **88** (1991), 2297–2301.

35. M. Costa, A. L. Goldberger, C. K. Peng, Multiscale entropy analysis of complex physiologic time series, *Phys. Rev. Lett.*, **89** (2002), 068102.
36. J. S. Richman, J. R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, *Am. J. Physiol. Circ. Physiol.*, **278** (2000), H2039–H2049.
37. J. L. Bentley, Multidimensional binary search trees used for associative searching, *Commun. ACM*, **18** (1975), 509–517.
38. D. Wang, D. Miao, C. Xie, Best basis-based wavelet packet entropy feature extraction and hierarchical EEG classification for epileptic detection, *Expert Syst. Appl.*, **38** (2011), 14314–14320.
39. O. A. Rosso, S. Blanco, J. Yordanova, V. Kolev, A. Figliola, M. Schürmann,et al., Wavelet entropy: a new tool for analysis of short duration brain electrical signals, *J. Neurosci. Methods*, **105** (2001), 65–75.
40. Y. Wu, Y. Zhou, G. Saveriades, S. Agaian, J. P. Noonan, P. Natarajan, Local Shannon entropy measure with statistical tests for image randomness, *Inf. Sci.*, **222** (2013), 323–342.
41. S. Ekici, S. Yildirim, M. Poyraz, Energy and entropy-based feature extraction for locating fault on transmission lines by using neural network and wavelet packet decomposition, *Expert Syst. Appl.*, **34** (2008), 2937–2944.
42. E. Avci, D. Hanbay, A. Varol, An expert discrete wavelet adaptive network based fuzzy inference system for digital modulation recognition, *Expert Syst. Appl.*, **33** (2007), 582–589.
43. I. Turkoglu, A. Arslan, E. Ilkay, An intelligent system for diagnosis of the heart valve diseases with wavelet packet neural networks, *Comput. Biol. Med.*, **33** (2003), 319–331.
44. Y. Li, C. Y. Wee, B. Jie, Z. Peng, D. Shen, Sparse multivariate autoregressive modeling for mild cognitive impairment classification, *Neuroinformatics*, **12** (2014), 455–469.
45. C. J. C. Burges, A tutorial on support vector machines for pattern recognition, *Data Min. Knowl. Discov.*, **2** (1998), 121–167.
46. F. J. Huang, Y. LeCun, Large-scale learning with svm and convolutional for generic object categorization, *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, **1** (2006), 284–291.
47. A. Gammerman, Z. Luo, J. Vega, V. Vovk, *Conformal and Probabilistic Prediction with Applications: 5th International Symposium*, Springer, Madrid, Spain, 2016.
48. V. N. Vapnik, An overview of statistical learning theory, *IEEE Trans. Neural Networks*, **10** (1999), 988–999.
49. A. P. Dobrowolski, M. Wierzbowski, K. Tomczykiewicz, Multiresolution MUAPs decomposition and SVM-based analysis in the classification of neuromuscular disorders, *Comput. Methods Programs Biomed.*, **107** (2012), 393–403.
50. A. Subasi, Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders, *Comput. Biol. Med.*, **43** (2013), 576–586.
51. C. Gao, Q. Cheng, P. He, W. Susilo, J. Li, Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack, *Inf. Sci.*, **444** (2018), 72–88.
52. Y. Yamauchi, M. Mukaidono, Probabilistic inference and Bayesian theorem based on logical implication, *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, Springer, Berlin, Heidelberg, 1999. Available from: https://link.springer.com/chapter/10.1007/978-3-540-48061-7_40.

53. G. X. Yuan, C. H. Ho, C. Lin, Recent advances of large-scale linear classification, *Proc. IEEE*, **100** (2012), 2584–2603.
54. N. A. Zaidi, Y. Du, G. I. Webb, On the effectiveness of discretizing quantitative attributes in linear classifiers, *IEEE Access*, **8** (2020), 198856–198871.
55. J. Zhang, C. Chen, Y. Xiang, W. Zhou, Y. Xiang, Internet traffic classification by aggregating correlated naive bayes predictions, *IEEE Trans. Inf. Forensics Secur.*, **8** (2013), 5–15.
56. C. Chen, G. Zhang, J. Yang, J. C. Milton, A. D. Alcántara, An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier, *Accid. Anal. Prev.*, **90** (2016), 95–107.
57. P. Bermejo, J. A. Gámez, J. M. Puerta, Speeding up incremental wrapper feature subset selection with Naive Bayes classifier, *Knowledge-Based Syst.*, **55** (2014), 140–147.
58. F. J. Ariza-Lopez, J. Rodriguez-Avi, M. V. Alba-Fernandez, Complete control of an observed confusion matrix, *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018. Available from: https://ieeexplore.ieee.org/abstract/document/8517540.
59. L. M. Wang, X. L. Li, C. H. Cao, S. M. Yuan, Combining decision tree and Naive Bayes for classification, *Knowledge-Based Syst.*, **19** (2006), 511–515.
60. L. Hussain, W. Aziz, A. S. Khan, A. Q. Abbasi, S. Z. Hassan, Classification of electroencephlography (EEG) alcoholic and control subjects using machine learning ensemble methods, *J. Multidiscip. Eng. Sci. Technol.*, **2** (2015), 126–131.
61. A. Criminisi, J. Shotton, E. Konukoglu, Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning, *Found. Trends® Comput. Graph. Vis.*, **7** (2011), 81–227.
62. R. Genuer, J. M. Poggi, C. Tuleau-Malot, Variable selection using random forests, *Pattern Recognit. Lett.*, **31** (2010), 2225–2236.
63. L. Breiman, Bagging predictors, *Mach. Learn.*, **24** (1996), 123–140.
64. T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016. Available from: https://dl.acm.org/doi/abs/10.1145/2939672.2939785.
65. J. H. Friedman, Greedy function approximation: A gradient boosting machine, *Ann. Stat.*, **29** (2001), 1189–1232.
66. P. Leray, P. Gallinari, Feature selection with neural networks, *Behaviormetrika*, **26** (1999), 145–166.
67. K. Ha, S. Cho, D. MacLachlan, Response models based on bagging neural networks, *J. Interact. Mark.*, **19** (2005), 17–30.
68. D. Stephens, M. Diesing, A comparison of supervised classification methods for the prediction of substrate type using multibeam acoustic and legacy grain-size data, *PLoS One*, **9** (2014), e93950.
69. K. Hajian-Tilaki, Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation, *Casp. J. Intern. Med.*, **4** (2013), 627–635.
70. H. Wang, T. M. Khoshgoftaar, K. Gao, A comparative study of filter-based feature ranking techniques, *2010 IEEE International Conference on Information Reuse & Integration*, 2010, Available from: https://ieeexplore.ieee.org/abstract/document/5558966.

Appendix -A

Glossary

A

Accuracy is a measure of usefulness or effectiveness of the classification scheme is called Accuracy

Area under the receiver operating characteristic curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied

B

Big Data Management is the organization, administration and ***governance*** of ***large*** volumes of both structured and unstructured ***data***

C

Central nervous system is the part of the nervous system consisting primarily of the brain and spinal cord

Congestive heart failure is a chronic progressive condition that affects the pumping power of your ***heart*** muscles

D

Detrended fluctuation analysis is a method for determining the statistical self-affinity of a signal

Data-Driven Decision-Support System support decision making by allowing users to extract data from large databases, which are often in corporate data warehouses.

E

Epilepsy is a central nervous system (neurological) disorder in which brain activity becomes abnormal, causing seizures or periods of unusual behavior, sensations, and sometimes loss of awareness

Ensemble classifiers is a set of classifiers whose individual decisions are combined in some way (typically by weighted or unweighted voting) to classify new examples

L

Locomotion means the act or ability of an entity or person to transport or move oneself from place to place.

M

Multiscale entropy provides insights into the complexity of fluctuations over a range of time scales and is an extension of standard sample entropy measures

Mel frequency cepstral coefficients are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum")

N

Neurodegenerative diseases are a heterogeneous group of **disorders** that are characterized by the progressive degeneration of the structure and function of the central nervous system or peripheral nervous system

Negative predictive value is the probability that subjects with a negative screening test truly don't have the disease

P

Parkinson's disease is a brain disorder that leads to shaking, stiffness, and difficulty with walking, balance, and coordination

Positive predictive value is the probability that subjects with a **positive** screening test truly have the disease

S

Sensitivity is a measure of the classifier's ability for identifying the positive class patterns is called sensitivity.

Specificity is used to calculate the classifier's ability for identifying negative class patterns