Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 126 (2018) 1977–1984

Procedia Computer Science

www.elsevier.com/locate/procedia

22nd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

A comparative study of machine learning algorithms for physiological signal classification

Giorgio Biagetti[a], Paolo Crippa[a,*], Laura Falaschetti[a], Giulia Tanoni[a], Claudio Turchetti[a]

[a] *DII — Department of Information Engineering, Università Politecnica delle Marche, Via Brecce Bianche, 12, I-60131 Ancona, Italy*

Abstract

The present work aims at the evaluation of the effectiveness of different machine learning algorithms on a variety of clinical data, derived from small, medium, and large publicly available databases. To this end, several algorithms were tested, and their performance, both in terms of accuracy and time required for the training and testing phases, are here reported. Sometimes a data preprocessing phase was also deemed necessary to improve the performance of the machine learning procedures, in order to reduce the problem size. In such cases a detailed analysis of the compression strategy and results is also presented.

Keywords: Machine learning, classification, biosignal analysis

1. Introduction

In the recent years machine learning algorithms are becoming of paramount importance in computer science research. Recent advancements in wearable sensors for collecting biological data, such as electrocardiography (ECG), electroencephalography (EEG), surface electromyography (sEMG), photopletysmography (PPG) and speech signals, and inertial data such as accelerometer and gyroscopic signals have lead to a complex, large and heterogeneous data processing [7]. The human activity detection as well as the diagnosis and prognosis of patients based on manual investigation of data collected from these sensors are difficult and time consuming. Therefore, the implementation of knowledge-based decision-making systems is becoming more and more important in order to exploit the advantages of these new sensors. For these systems, the machine learning algorithms play a key role because they are capable of performing the analysis of such complex data in a very efficient way [17, 11]. In the field of fitness and healthcare data are often complex, context-dependent, and heterogeneous. As a consequence, obtaining insightful information from the raw data is a challenging task. Sports trainers, clinicians and researchers often make use of statistical data analysis

* Corresponding author. Tel.: +39-071-220-4541 ; fax: +39-071-220-4464.
E-mail address: p.crippa@univpm.it

10.1016/j.procs.2018.07.255

for examining physiological signals. In this context, machine learning algorithms can be used to identify patterns in data and the advantage of such algorithms is that the recognition systems can learn from the identified data providing more efficient and accurate decisions [12]. An example of decision within healthcare informatics is the recognition whether a hospital patient suffers of a particular pathology or not (healthy patient) [10, 4, 5, 15].

Because the electrical signals collectable from patients are different and so heterogeneous, it is very important to perform experimental investigation to compare the performance of different machine learning algorithms to help select the best one for a specific set of such signals.

A large number of machine learning algorithms have been developed in order to perform classification, regression, and in general pattern recognition. Among these the most common algorithms are the support vector machine (SVM) [13], the neural networks, the k-nearest neighbor (KNN) [14, 16], the logistic regression, the decision tree [9], the random forests, the linear discriminant analysis (LDA) [26], the case-based reasoning, the naïve Bayes, and the fuzzy logic. This study compares the performance of some of the above popular machine learning methods applied to publicly available data set, in order to investigate algorithms that provide the best classification accuracy using physiological signals such as speech, EEG and gait dynamics signals.

The rest of this paper is organized as follows: Section 2 is divided into two subsections that present the used data sets and machine learning algorithms, respectively. Section 3 discusses the experimental results and related findings. Finally, Section 4 gives some conclusions.

2. Material and Methods

2.1. Data sets

In this work the following databases related to physiological signals have been considered in order to test the classification performances of the chosen machine learning algorithms.

1. *The Parkinson data set (DS1).* This data set is composed of several voice samples that have been recorded from 23 subjects with Parkinson's disease (PD) and from 8 healthy subjects. From each of the 31 subjects, an average of six voice samples have been obtained, for a total of 195 voice recordings. From all these speech signals a total of 22 features have been calculated using standard and and nonstandard measurement techniques that generated a single number for each of the 195 voice recordings [25].
2. *Parkinson speech dataset with multiple types of sound recordings (DS2).* This data set is composed of several voice samples that have been recorded from 40 subjects (20 subjects with PD and 20 healthy subjects) for the training set, and from 28 subjects with PD for the testing set. For the training set, 26 voice samples including sustained vowels, numbers, words, and short sentences in Turkish language have been recorded from each subject for a total of 1040 recordings. For the testing set three samples of the sustained vowels "a" and "o" have been recorded from each subject for a total of 168 recordings. Finally a set of 26 linear and time-frequency based features have been extracted from each voice sample [29].
3. *Epileptic seizure recognition data set (DS3).* This data set is composed of data that have been recorded from 500 subjects. For each subject 23.6 seconds of brain activity have been recorded. The corresponding time-series have been sampled into 4097 data points representing the values of the EEG signal as a function of the discrete time. Every 4097 data points have been divided and shuffled into 23 segments, each segment contains 178 data points for 1 second. This results in $23 \times 500 = 11500$ pieces of information (rows), each of which contains 178 data points for 1 second (columns) [3].
4. *EEG eye state data set (DS4).* This data set is composed of data obtained from one continuous EEG signal whose the duration was 117 seconds. This EEG signal has been measured by using the Emotiv EEG Neuroheadset. During the EEG measurement, the eye state has been detected via a camera and added manually to the file after analysing the video frames. Two eye states have been considered: the "eye-closed" state and the "eye-open" state. The continuous EEG signal over time for 14 electrodes was sampled and 14980 samples were obtained as observations. Of these, 6722 and 8255 observations are related to the "eye-closed" and "eye-open" states, respectively [30, 28].

5. *Gait dynamics in neuro-degenerative disease data base (DS5).* This data set is composed of data obtained from gait dynamics measured from 63 subjects: 15 subjects are patients with PD, 20 subjects are patients with Huntington's disease (HD), and 13 subjects are patients with amyotrophic lateral sclerosis (ALD). 16 healthy subjects are also included as a control. The raw data were measured using force-sensitive resistors, with the output roughly proportional to the force under the foot. The following stride-to-stride measures of footfall contact times were derived from these signals as features, i.e. left stride interval, right stride interval, left swing interval, right swing interval, left swing interval (% of stride), right swing interval (% of stride), left stance interval, right stance interval, left stance interval (% of stride), right stance interval (% of stride), double support interval, and double support interval (% of stride) [24, 22, 23].
6. *Gait in Parkinson's disease - vertical ground reaction force (VGRF) (DS6).* This data set is composed of data obtained from gait measurements of 93 subjects with PD, and 73 healthy subjects as a control. The data set includes the vertical ground reaction force records of subjects as they walked at their usual, self-selected pace for approximately 2 minutes on level ground. Underneath each foot were 8 sensors (Ultraflex Computer Dyno Graphy, Infotronic Inc.) that measure force (in newtons) as a function of time [2, 1]. The output of each of these 16 sensors has been digitized and recorded at 100 samples per second, and the records also include two signals that reflect the sum of the 8 sensor outputs for each foot [18].

2.2. *Algorithms*

To objectively identify the diseases from the gait data, we used the following different types of classifiers.

1. *Linear, cubic and Gaussian support vector machine (SVM).* The support-vector machine or network is a supervised learning technique for two-group classification problems. The machine conceptually implements the following idea: input vectors are non-linearly mapped to a very high-dimension feature space. In this feature space a linear decision surface is constructed. Special properties of the decision surface ensures high generalization ability of the learning machine [13]. Given a training set, whose elements are marked as belonging to one of two categories, the SVM builds a model that assigns the elements of the testing set to one category or the other, making it a non-probabilistic binary linear classifier. The SVM model represents the samples as points in space, mapped so that the samples of the separate categories are divided by a clear gap that is as wide as possible. New samples are then mapped into that same space and predicted as belonging to a category based on which side of the gap they fall in. The SVM algorithm can efficiently perform both a linear and a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. Using a cubic or a Gaussian function for the kernel we obtain the so-called cubic SVM and Gaussian SVM, respectively.
2. *K-nearest neighbor (KNN).* This classifier is one of the most popular neighborhood classifiers in pattern recognition and machine learning because of its simplicity and efficiency. It categorizes each unlabelled test example using the label of the majority of examples among its k-nearest (most similar) neighbors in the training data set. The similarity depends on a specific distance metric, therefore, the performance of the classifier strictly depends on the distance metric used. However, it suffers of memory requirements and time complexity, because it is fully dependent on every example in the training set [31].
3. *Decision tree.* This classifier partitions the input space into small segments, and labels these small segments with one of the various output categories. However, conventional decision tree only does the partitioning to the coordinate axes. With the growth of the tree, the input space can be partitioned into very small segments so as to recognize subtle patterns [27]. The main drawback is that overgrown trees could lead to overfitting.
4. *Linear discriminant analysis (LDA).* This classifier consists in finding the projection hyperplane that minimizes the interclass variance and maximizes the distance between the projected means of the classes. Similarly to Karhunen-Loève transform (KLT) [21, 20, 6, 8], these two goals can be achieved by solving an eigenvalue problem with the corresponding eigenvectors defining the hyperplane of interest to be used for the classification [19].
5. *Logistic regression.* This algorithm is a variant of the linear regression, specialized for the case when the dependent variable is binary. Instead of directly fitting the dependent variable, its probability of occurrence is used, and

then a decision threshold is employed to make the final judgement. Since there are no closed-form solutions to this problem, typically an iterative approach such as maximum likelihood is used to fit the model.

3. Experimental Results

The previously mentioned classification algorithms have been tested on the datasets DS1 through DS6. The first four datasets were used "as is", passing all the recorded features and/or signals to the classifiers, while the last two were also manipulated to try and find better features for the classification task.

3.1. Classification results on original data sets

Results for the unmodified datasets are shown in Tables 1–6.

Table 1. Parkinson data set (DS1) – Binary classification: Healthy, Parkinson disease (matrix dimensions: 195×23).

Algorithm	Accuracy [%]	Train time [s]	Test time [s]
KNN	97.44	0.394	0.099
LDA	84.62	0.548	0.047
SVM Linear	79.49	0.572	0.037
SVM Cubic	92.31	0.575	0.037
SVM Gaussian	84.62	0.538	0.040
Decision Tree	84.62	0.435	0.071
Logistic Regression	25.64	< 0.001	< 0.001

Table 2. Parkinson speech data set with multiple types of sound recordings (DS2) – Binary classification: Healthy, Parkinson disease (matrix dimensions: 1208×27).

Algorithm	Accuracy [%]	Train time [s]	Test time [s]
KNN	70.12	0.411	0.105
LDA	65.98	0.533	0.048
SVM Linear	65.98	0.720	0.025
SVM Cubic	73.86	1.265	0.020
SVM Gaussian	67.22	0.607	0.056
Decision Tree	66.39	0.452	0.071
Logistic Regression	66.80	0.940	0.009

Table 3. Epileptic seizure recognition data set (DS3) – Binary classification: Healthy, Epilepsy (matrix dimensions: 11500×179).

Algorithm	Accuracy [%]	Train time [s]	Test time [s]
KNN	94.57	0.457	0.727
LDA	82.30	0.645	0.036
SVM Linear	80.91	2.831	0.193
SVM Cubic	96.61	1.621	0.087
SVM Gaussian	80.00	5.496	0.684
Decision Tree	93.70	0.711	0.070
Logistic Regression	82.13	1.502	0.011

3.2. Classification results on compressed data sets

DS5 and DS6 were further investigated to see if a compression of the features helped the recognition.

Table 4. EEG eye state data set (DS4) – Binary classification: eye-open state, eye-closed state (matrix dimensions: 14980×15).

Algorithm	Accuracy [%]	Train time [s]	Test time [s]
KNN	82.38	0.421	0.230
LDA	62.82	0.564	0.047
SVM Linear	62.58	16.538	0.143
SVM Cubic	57.18	128.033	0.020
SVM Gaussian	75.47	3.913	0.538
Decision Tree	78.20	0.482	0.074
Logistic Regression	63.15	0.998	0.009

Table 5. Gait dynamics in neuro-degenerative disease data set (DS5) – Multi-class classification: Healthy, Parkinson disease, ALS disease (matrix dimensions: 10314×12).

Algorithm	Accuracy [%]	Train time [s]	Test time [s]
KNN	72.65	0.415	0.165
LDA	61.83	0.564	0.046
SVM Linear	66.93	29.941	0.057
SVM Cubic	24.64	164.562	0.060
SVM Gaussian	79.63	2.955	0.326
Decision Tree	77.01	0.476	0.073

Table 6. Gait in Parkinson's disease - vertical ground reaction force (DS6) – Binary classification: Healthy, Parkinson disease (matrix dimensions: 4512×2000).

Algorithm	Accuracy [%]	Train time [s]	Test time [s]
KNN	99.64	0.939	380.169
LDA	75.51	1.644	0.078
SVM Linear	75.34	5651.882	313.824
SVM Cubic	54.97	22237.372	129.825
SVM Gaussian	99.25	8625.951	144.639
Decision Tree	80.93	2.831	0.098
Logistic Regression	75.52	5.421	0.025

DS5 features were reduced by discarding data for the right foot and then computing *1)* the step-to-step interval average, *2)* the swing interval average, *3)* the maximum step-to-step interval, *4)* the minimum step-to-step interval, *5)* their standard deviation, *6)* and variance, *7)* the stance interval duration, *8)* the stance percentage of the step-to-step interval, *9)* the average speed. Classification results are shown in Table 7. As can be seen, the SVM-based classifiers clearly benefitted from the reduction in the feature space dimension.

Table 7. Gait dynamics in neuro-degenerative disease data set (DS5) – Multi-class classification: Healthy, Parkinson disease, ALS disease (matrix dimensions: 42×9).

Algorithm	Accuracy [%]	Train time [s]	Test time [s]
KNN	62.5	0.413	0.082
LDA	75.0	0.546	0.047
SVM Linear	87.5	0.788	0.054
SVM Cubic	87.5	0.788	0.056
SVM Gaussian	50.0	0.746	0.058
Decision Tree	75.0	0.432	0.072

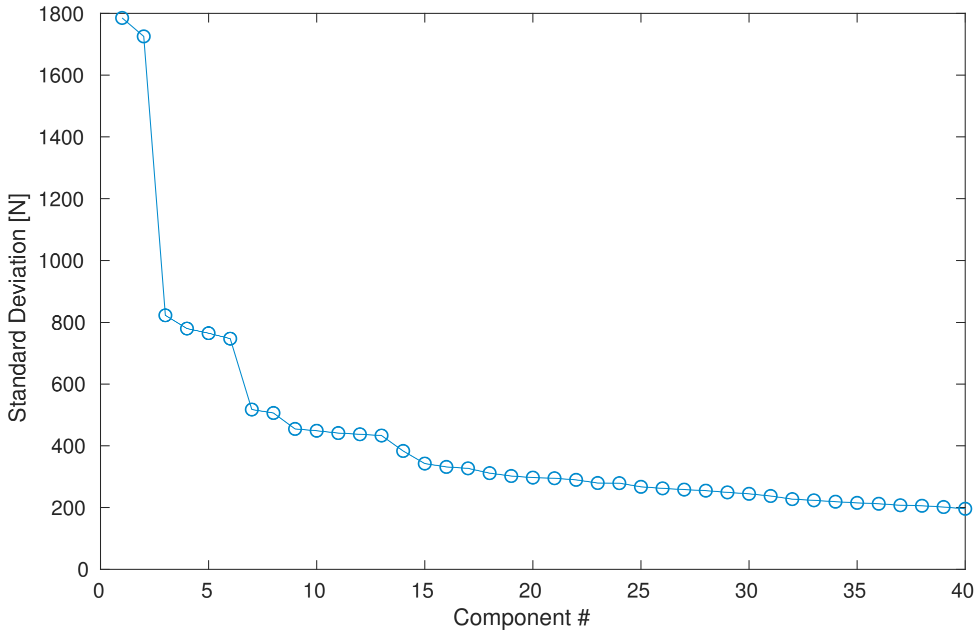

Fig. 1. Eigenvalues from the DS6.

For DS6, a principal component analysis (PCA) was employed to reduce the number of features, as follows. First, only 47 subjects were used, each of which providing a 2-minute recording sampled at 100 Hz. Each signal track was splitted into 1.25 s long windows, without overlapping. Each time window is thus composed of 125 samples from each of the 16 force sensors, for a total of 2000 features. Since there are a total of 96 recorded time windows per subject, the total number of observations was 4512.

The ensuing data matrix, 4512×2000, was reduced by means of a PCA to evaluate the effect of the size of the parameter space on the various classification algorithms.

Figure 1 reports the square root of the eigenvalues of the covariance matrix of the data. This is the standard deviation of each principal component coefficient. Figure 2 reports the cumulative sum of said eigenvalues, i.e., the total explained variance of the first principal components. As can be seen, most of the variance is explained by the first few components, so it is reasonable to try very strong compressions down to the order of half a dozen components.

Table 8 reports the results of the classification experiments on such data set for different numbers of components. As expected, the best accuracies were obtained with just 5 or 7 components. The dimensionality reduction actually helps the classification, as shown by contrasting with Table 6, which reports the results obtained without PCA.

4. Conclusion

In this paper we presented the results of several commonly employed machine learning algorithms applied to the automatic classification of clinical data to help in the diagnosis of different diseases. From the experiments, it is apparent that no single algorithm can be deemed to be the best in all cases. Among all those tested, KNN was the one to give more consistent results, though it is known to require lots of storage for the models if the data set is large. The SVM family also suffers from big data sets, and benefits the most from data compression. Optimal feature selection, either automated by means such as PCA or manually performed, obviously still plays a fundamental role in improving machine learning performance.

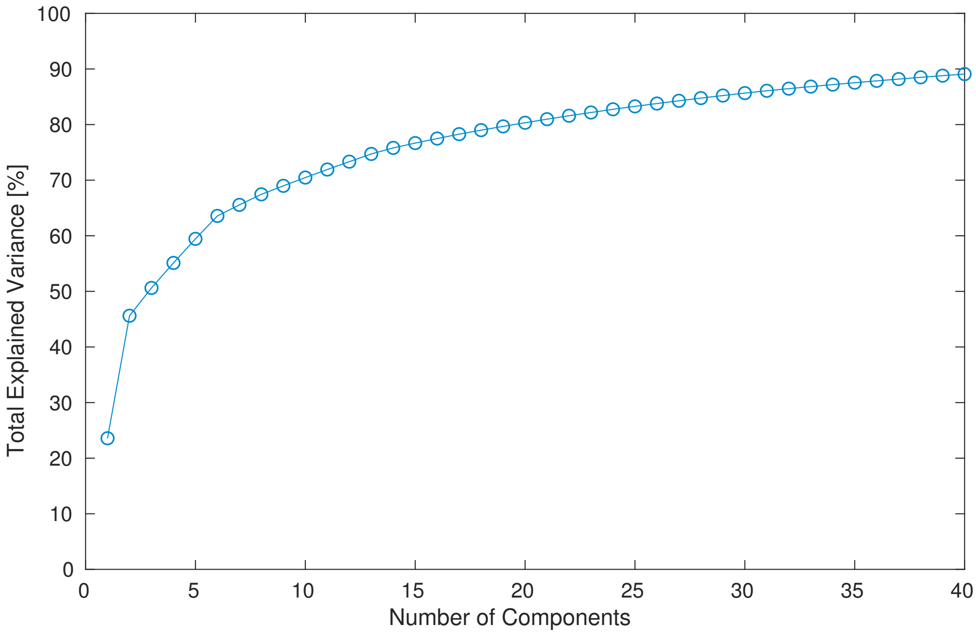

Fig. 2. Cumulative sum of the eigenvalues from DS6.

Table 8. Gait in Parkinson's disease - vertical ground reaction force (DS6) – Binary classification with PCA compression.

PCA Components	KNN [%]	SVM Linear [%]	SVM Cubic [%]	SVM Gaussian [%]	Decision Tree [%]	LDA [%]	Logistic Regression [%]
2	59.1	61.6	48.9	61.0	58.4	61.6	61.6
3	92.2	61.6	92.9	94.3	91.8	61.6	61.6
4	99.3	61.6	99.6	99.7	98.2	61.6	61.6
5	99.6	61.6	**99.8**	99.6	**98.3**	61.6	61.6
6	99.7	61.6	99.8	99.7	97.2	61.6	61.6
7	**99.8**	77.7	99.7	**99.8**	97.3	76.8	75.4
8	99.7	85.3	99.7	99.6	96.1	84.0	84.6
9	99.6	84.1	99.8	99.2	97.2	83.7	83.9
10	99.4	84.8	99.7	98.8	97.1	83.7	84.4
11	99.2	84.9	99.4	98.3	97.1	83.9	83.9
12	99.1	87.4	99.3	97.7	96.9	86.1	85.9
13	99.1	87.9	99.6	96.5	97.0	87.4	87.9
14	98.3	88.4	99.6	94.8	96.9	87.9	87.9
15	98.7	88.1	99.6	93.5	96.9	88.2	88.6
16	98.7	88.2	99.4	92.7	96.1	88.1	88.0
17	98.8	88.2	99.3	89.9	96.8	88.5	88.5
18	98.9	88.9	99.6	91.4	96.6	88.5	89.2
19	99.3	89.2	99.7	88.6	96.1	89.2	89.7
20	99.1	**89.4**	99.3	87.0	96.1	**89.4**	**89.9**

References

[1] Alam, M.N., Garg, A., Munia, T.T.K., Fazel-Rezai, R., Tavakolian, K., 2017. Vertical ground reaction force marker for Parkinson's disease. PLOS ONE 12, 1–13.

[2] Alkhatib, R., Diab, M., Moslem, B., Corbier, C., El Badaoui, M., 2015. Gait-ground reaction force sensors selection based on ROC curve evaluation. Journal of Computer and Communications 3, 13–19.

[3] Andrzejak, R.G., Lehnertz, K., Mormann, F., Rieke, C., David, P., Elger, C.E., 2001. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. Phys. Rev. E 64, 061907.

[4] Begum, S., Ahmed, M.U., Funk, P., Xiong, N., Folke, M., 2011. Case-based reasoning systems in the health sciences: A survey of recent trends and developments. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 41, 421–434.

[5] Biagetti, G., Crippa, P., Curzi, A., Orcioni, S., Turchetti, C., 2014. A multi-class ECG beat classifier based on the truncated KLT representation, in: UKSim-AMSS 8th European Modelling Symposium on Computer Modelling and Simulation (EMS2014), Pisa, Italy. pp. 93–98.

[6] Biagetti, G., Crippa, P., Falaschetti, L., Orcioni, S., Turchetti, C., 2016a. Multivariate direction scoring for dimensionality reduction in classification problems. Smart Innovation, Systems and Technologies 56, 413–423.

[7] Biagetti, G., Crippa, P., Falaschetti, L., Orcioni, S., Turchetti, C., 2016b. Wireless surface electromyograph and electrocardiograph system on 802.15.4. IEEE Transactions on Consumer Electronics 62, 258–266.

[8] Biagetti, G., Crippa, P., Falaschetti, L., Orcioni, S., Turchetti, C., 2017. An investigation on the accuracy of truncated DKLT representation for speaker identification with short sequences of speech frames. IEEE Transactions on Cybernetics 47, 4235–4249.

[9] Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth & BrooksCole Advanced Books & Software., Monterey, CA.

[10] Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms, in: Proceedings of the 23rd international conference on Machine learning, pp. 161–168.

[11] Christopher, M.B., 2016. Patter Recognition and Machine Learning. Springer-Verlag New York.

[12] Clifton, D.A., Gibbons, J., Davies, J., Tarassenko, L., 2012. Machine learning and software engineering in health informatics, in: 2012 First International Workshop on Realizing AI Synergies in Software Engineering (RAISE), Zurich, Switzerland. pp. 37–41.

[13] Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297.

[14] Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13, 21–27.

[15] Crippa, P., Curzi, A., Falaschetti, L., Turchetti, C., 2015. Multi-class ECG beat classification based on a Gaussian mixture model of Karhunen-Loève transform. International Journal of Simulation: Systems, Science and Technology 16, 2.1–2.10.

[16] Dasarathy, B.V., 2002. Handbook of data mining and knowledge discovery, Oxford University Press, Inc., New York, NY, USA. chapter Data Mining Tasks and Methods: Classification: Nearest-neighbor Approaches, pp. 288–298.

[17] Duda, R.O., Hart, P.E., Stork, D.G., 2012. Pattern classification. John Wiley & Sons.

[18] Frenkel-Toledo, S., Giladi, N., Peretz, C., Herman, T., Gruendlinger, L., Hausdorff, J.M., 2005. Effect of gait speed on gait rhythmicity in Parkinson's disease: variability of stride time and swing time respond differently. Journal of NeuroEngineering and Rehabilitation 2, 23.

[19] Fukunaga, K., 2013. Introduction to statistical pattern recognition. Academic Press.

[20] Gianfelici, F., Biagetti, G., Crippa, P., Turchetti, C., 2005. A novel KLT algorithm optimized for small signal sets, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '05), pp. 405–408.

[21] Gianfelici, F., Turchetti, C., Crippa, P., 2009. A non-probabilistic recognizer of stochastic signals based on KLT. Signal Processing 89, 422 – 437.

[22] Hausdorff, J.M., Cudkowicz, M.E., Firtion, R., Wei, J.Y., Goldberger, A.L., 1998. Gait variability and basal ganglia disorders: Stride-to-stride variations of gait cycle timing in Parkinson's disease and Huntington's disease. Movement disorders 13, 428–437.

[23] Hausdorff, J.M., Lertratanakul, A., Cudkowicz, M.E., Peterson, A.L., Kaliton, D., Goldberger, A.L., 2000. Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis. Journal of applied physiology 88, 2045–2053.

[24] Hausdorff, J.M., Mitchell, S.L., Firtion, R., Peng, C.K., Cudkowicz, M.E., Wei, J.Y., Goldberger, A.L., 1997. Altered fractal dynamics of gait: reduced stride-interval correlations with aging and Huntington's disease. Journal of applied physiology 82, 262–269.

[25] Little, M.A., McSharry, P.E., Hunter, E.J., Spielman, J., Ramig, L.O., 2009. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering 56, 1015–1022.

[26] McLachlan, G., 2004. Discriminant analysis and statistical pattern recognition. volume 544. John Wiley & Sons.

[27] Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1, 81–106.

[28] Sabancı, K., Koklu, M., 2015. The classification of eye state by using kNN and MLP classification models according to the EEG signals. International Journal of Intelligent Systems and Applications in Engineering 3, 127–130.

[29] Sakar, B.E., Isenkul, M.E., Sakar, C.O., Sertbas, A., Gurgen, F., Delil, S., Apaydin, H., Kursun, O., 2013. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. IEEE Journal of Biomedical and Health Informatics 17, 828–834.

[30] Wang, T., Guan, S.U., Man, K.L., Ting, T.O., 2014. EEG eye state identification using incremental attribute learning with time-series classification. Mathematical Problems in Engineering 2014.

[31] Weinberger, K.Q., Saul, L.K., 2009. Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research 10, 207–244.