This is the accepted manuscript of the article, which has been published in Intelligent data analysis, 2019, 23(1), 215-226. The final publication is available at IOS Press through https://doi.org/10.3233/IDA-173704

Classification of patients and controls based on stabilogram signal data

Henry Joutsijoki*[1], Jyrki Rasku†[1], Ilmari Pyykkö‡[2], and Martti Juhola§[1]

[1]University of Tampere, Faculty of Natural Sciences, FI-33014 University of Tampere, Finland
[2]University of Tampere, Faculty of Medicine and Life Sciences, PB 100, FI-33014 University of Tampere, Finland

Abstract

Inner ear balance problems are common worldwide and are often difficult to diagnose. In this study we examine the classification of patients with inner ear balance problems and controls (people not suffering from inner ear balance problems) based on data derived from the stabilogram signals and using machine learning algorithms. This paper is a continuation for our earlier paper where the same dataset was used and the focus was medically oriented. Our collected dataset consists of stabilogram (a force platform response) data from 30 patients suffering from Ménière's disease and 30 students called controls. We select a wide variety of machine learning algorithms from traditional baseline methods to state-of-the-art methods such as Least-Squares Support Vector Machines and Random Forests. We perform extensive and carefully made parameter value searches and we are able to achieve 88.3% accuracy using k-nearest neighbor classifier. Our results show that machine learning algorithms are well capable of separating patients and controls from each other.

Keywords: Ménière's disease, Stabilogram signal, Machine learning, Otoneurological diseases, Classification

*Corresponding author, Tel.:+358503185860; fax:+35832191001, henry.joutsijoki@uta.fi
†jyrki.rasku@uta.fi
‡ilmari.pyykko@uta.fi
§martti.juhola@uta.fi

1 Introduction

The capability of learning new things is amazing for humans. Often in the beginning difficulties are encountered when learning new skills, but after some time we adjust in such a way that we master the new skill as we would have done it always. One of the first skills what we learn is to walk. As a baby the first steps are the most difficult ones since we do not know how to control our body balance. We may take one step and then fall. This situation is repeated for several times until we learn to control our balance.

After learning to control our body balance, unfortunately, we easily forget how complicated and fine-tuned skill it actually is and think it as self-granted. Overall, a good body balance is taken almost for granted among the majority of people. Nevertheless, we notice the importance of body balance when accidents such as falling at the slippery ice or stumbling to an object happen. These kinds of accidents can happen to everyone and cause a lot of harm to everyday life.

There are many people who have significant difficulties in maintaining balance even in good environmental conditions. Persons who have problems with their balance are usually elderly or young with some inner ear disorder. These disorders may have huge effect on the everyday life and can make even the simplest tasks impossible to perform (see, for example, [37]). If a person has some inner ear disorder, it may also increase his/her probability to get into an accident. A number of accidents happen worldwide due to inner ear disorders which may lead to severe injuries likewise leg and hip fractures, head injuries or even death [24]. Treatments for these injuries again can be expensive and require long-term hospitalization or sick leave. The total costs of accidents (including treatments and recovering) can be very high both for the patient himself/herself and for the society.

Despite a person being an elder one or young, the severity of balance problem should be evaluated, if an inner ear balance problem is suspected. Among the young, objective recording of certain types of balance deficits may help physicians to make diagnoses whereas for the elderly similar measurements could further be used in planning and proposing suitable balance training programs. However, the first step is to recognize whether or not a person has a balance problem, which is caused by an inner ear disease before other more specific diagnoses are made.

An objective evaluation of human ability to maintain balance can be done using a force platform which produces a stabilogram signal. There is, however, no commonly agreed way to interpret such results. In addition, there is a lack for commonly accepted baseline values for measures for different age groups and gender (see, for example [36], for discussion on this subject). A commonly used

method in examination of human swaying is to apply diffusion analysis introduced by Collins and De Luca [9]. Diffusion analysis has been used since then widely and some examples of it are the studies of Peterka [28] and Forsman et al. [16]. Peterka [28] studied human swaying during quiet stance without stimulation and Forsman et al. [16] applied the idea of Peterka in their own study.

In our earlier research [30] we focused on studying a measure that can efficiently tell apart the measurements from young students and older people who suffered from Ménière's disease. Compared to [16, 28] we did not use diffusion analysis in our previous study in order to have a new perspective for the analysis of human swaying process. The purpose of this article is to examine more closely machine learning methods that could improve our classification results presented in [30] using the same dataset and same features than before. In contrast to [30] we have done significant changes:

1. The number of classification methods is larger.
2. Extended parameter value search
3. Leave-one-out and nested-leave-one-out methods in computational training and tests
4. New performance measures

Hence, all the experiments presented in this paper are new.

The rest of the paper is organized as follows. Section 2 presents related works. Section 3 describes in detail the technical issues of the design of experiments including description of dataset, feature extraction, measurements, classification methods, parameter settings, and performance measures used. In Section 4 classification results are presented. Finally, Section 5 is left for discussion and conclusion.

2 Related Works

Evaluation of human ability to maintain the upright stance is part of physician's daily work. First ways to make the evaluation was the use of Romberg's test. This test is applied even nowadays, if no other testing method is available. The result of this test was based on physician's subjective evaluation whether a person sway more with eyes closed than eyes open or not. Later on Collins and de Luca [9] proposed their diffusion analysis as a new tool in the balance evaluation. This

idea was based on PID (propotional, integral, derivative) control theory. Within a short swaying period postural controlling can be seen as open loop system, where no control is needed. Longer swaying, on the other hand, requires strategy or closed loop control where feedback from environment is needed (visual and force feedback). This approach is mostly used in human balance research. Later on, other methods have been used in modeling the swaying process such as fractals [13] and wavelets [27].

Fractals and wavelets approaches, however, try to model the swaying process, not predicting the differences in the model parameters. This task is difficult, because the model parameters do usually not vary a lot among measured people. On the other hand, this finding supports the idea that models describing the phenomenon are robust. Classification of human swaying has been considered earlier in some extent. For instance, works of Audiffren et al. [1] and Hewson et al. [19] are closely related to the swaying process and measured physical dimensions into previous theory.

Machine learning approach is quite novel in this field of medical expertise. In this work we outline how machine learning could be used in simulating medical diagnosing in a context of human balance disorders. Machine learning methods have been used earlier in inner ear diagnosis. For example, in [20, 40, 41, 42] a set of otoneurological diseases (including Ménière's disease) were classified using machine learning methods successfully. More specifically, in [20] half-against-half architecture was used and applied with k-NN, Support Vector Machine (SVM) and naïve Bayes classifiers. The best accuracy, 76.9%, was obtained by the SVM. In [40] k-NN and naïve Bayes classifier were used and the highest accuracy was 75.0%. Weighted k-NN and genetic algorithms were applied to classification of otoneurological diseases in [41] and around 80.0% accuracy was gained. In [42] k-NN and SVM classifiers were used with one-vs-all and one-vs-one multi-class approaches and 82.4% accuracy was obtained.

In [43] machine learning methods were used for assessing the imbalance and vestibular dysfunction. Yeh et al. [43] used SVM-based solution and within six different test setups, the highest accuracy was around 95.0%. Finally, Dastgheib et al. [11] applied machine learning methods to the diagnosis of Ménière's disease. They tested five different approaches and obtained 84% accuracy on test set. For data derived from stabilogram signals, machine learning methods have been used in [29, 30, 31, 32]. In these studies k-NN, Hidden Markov Models, SVM were used and the accuracies were over 70%.

3 Design of Experiments

3.1 Dataset, feature extraction and measurements

Our dataset consists of stabilogram signals measured from 60 participants likewise in [35]. The dataset is collected by the authors. Half of the participants were patients suffering from Ménière's disease and the other half were students from the University of Tampere and they are called controls and they were not suffering from an inner ear balance problems. Measurements were made by the permission of patient organization and for the patient organization. Hence, the collected data is register data where people are not identified and the register data does not require permission from the ethics committee. Both controls and patients voluntarily participated in the test.

Detailed information related to groupwise average ages (25 for controls and 60 for patients) and distributions of sexes within the patients (43.3%/56.7% for M/F) and controls (93.3%/6.7% for M/F) can be found from Table 1. The same dataset was used in [30] together with the same features. However, the focus in [30] was more medically oriented whereas this paper concentrates on the machine learning perspective by classifying patients and controls.

The detailed measuring procedure for the data used in this work can be found from our earlier study [30]. In [30] we conducted a two phase test where our goal was to extract such measurements that were closely related to the human ability to maintain the balance. In the first phase we measured how well a person can follow a moving object only by altering his/her center point of mass on a force platform. In the second phase of our test we measured an average time delay that occurs when a subject is following a "jumping" light source only using eyes.

The force platform contains three analog pressure sensitive voltage sensors that form a plane under the subject's feet. Using the information about the location of the sensors in relation to others, we can use a momentum equation that results in the location of resulting force acting on the platform. Analog signal is converted into digital form using DT9800 A/D converter at the frequency of 50Hz. The device used in measuring the visual time delay was a product of Micromedical Technologies[1]. This device is usually used in thorough evaluation of inner ear functionality.

Human postural control system combines information from vision, and force feedback from the ground. Our selected features closely connect these two infor-

[1] `http://www.micromedical.com/Products/VisualEyes_VNG/VisualEyes-Spectrum`

mation channels. Tracking the target signal require visual feedback while feeling the ground tells how far it is possible to lean without losing balance. One can easily verify the need for force feedback from the ground. It is somewhat difficult to walk on a soft mattress of a trampoline, because the feel of supporting force has been suppressed.

These two tests yielded us a five dimensional feature vector (t, n, d, F_x, F_y) from each subject. More specifically, t is the total time when a subject was able to follow the object, n is the number of approaches to reach the target when losing it, d is the mean eye movement delay in a saccade test. Furthermore, F_x and F_y are form factor signals in x (medio-lateral) and y (anterior-posterior) directions and they contain the information about the nature of error when following the target.

An individual signal that is used in describing the form factor is a signal $\mathbf{e} = e_1, \ldots, e_T$ where e_1 is the difference of target and control squares in sampling instant 1. Similarly, e_T is the respective difference in sampling instant T. Form factors F_x and F_y are calculated as a quotient of kurtosis and variance for respective signals e_x and e_y. High F presents the situation where a subject manages to stay on a target well. However, small amounts of large extrusions occur. Low F, on the other hand, is characteristic for a process where a person does not stay properly on target and there is a great number of sufficiently large error incidents.

3.2 Classification

3.2.1 Methods

In this research we examined several classification methods ranging from traditional baseline methods to current state-of-the-art methods. More specifically, we tested the following methods: Classification tree (CT) (CART algorithm [4, 5, 14, 25, 45]), k-nearest neighbor classifier (k-NN) [10, 15, 45], Linear discriminant analysis (LDA) [2, 8], Logistic regression (LR) [12, 22], Least-Squares Support Vector Machine (LS-SVM) [38, 39], Mahalanobis discriminant analysis (MDA) [6], naïve Bayes (NB) variants [33, 34, 45], Quadratic discriminant analysis (QDA) [8, 21] and Random Forests (RF) [7, 23]. These algorithms were selected because they have shown great performance in many applications [20, 40, 42, 45] and they extend our earlied study [30] significantly. Naïve Bayes method was tested with and without kernel density estimation (KDE) [18]. When KDE was used with NB, we tested four different kernels: Box, Epanechnikov, Gaussian and triangle.

Overall, classification methods tested can be divided into two groups:

- Parameter free algorithms
- Parameter-dependent algorithms

Classification was conducted using the leave-one-out (LOO) method for the parameter free algorithms and for the parameter-dependent algorithms nested leave-one-out (NLOO) was applied. LOO and NLOO approaches were chosen because they have the benefit of maximizing the size of training set which is important when the dataset is relatively small as in our case it is. However, the difference between aforementioned categories become evident since parameter free methods (LDA, LR, MDA, NB and its variants and QDA) require the classification procedure to be performed only once whereas for the rest of the methods (CT, k-NN, LS-SVM, and RF) the classification procedure must be repeated with different parameter settings.

Logistic regression differs from the other parameter free classification methods. In LR the output for the test instance is not a class label directly but it is a real number close to 0 or 1. A final class label for each instance was derived by rounding the output to the closest integer (0/1). Parameter settings are explained in detail in Section 3.2.2. After finishing the LOO or NLOO procedure, performance measures (see Section 3.3) were evaluated. We used Matlab 2017a together with Statistics and Machine Learning Toolbox and Parallel Computing Toolbox in our tests. All the tests were run using a desktop computer having Intel i7-3960X 3.5GHz processor, 32GB memory and Win7 operating system.

3.2.2 Parameter settings

The first parameter-dependent algorithm is Classification tree (CART). In the implementation of CART algorithm we varied the number observations required to be in each splitting node in a tree. We tested values of 1, 5 and 10. Moreover, we used 1 as a minimum number of observations what each leaf must include. In addition, we used Gini's diversity index as a splitting criterion. Classification was performed using LOO since we present the results of each parameter settings in result table.

In the cases of k-NN, LS-SVM and RF (other parameter-dependent algorithms) NLOO was used due to parameter optimization and to prevent possible overfitting. NLOO includes two loops where the outer loop is for model assessment whereas the inner loop is for model selection including parameter optimization. In other words, in each round of outer loop, we repeat LOO procedure to a training set equal number of times as we have parameter values to be tested.

Hence, we find optimized parameter values for each training set separately. When the optimal parameter setting is found, a classifier is trained again using the whole training set and tested with the test instance extracted from the dataset in outer loop. Optimal parameter setting was selected using accuracy which is specified in detail in Section 3.3.

For Random Forests classifier the number of trees ($\#trees$) is the most crucial parameter. We varied the number of trees in a forest from 1 to 50 with the step size of 1. In the case of k-NN classifier there are three main parameters included:

1. Distance measure
2. Possible distance weighting
3. k value.

In this study we examined eight distance measures: Chebychev metric, correlation measure, cosine measure, Euclidean metric, Mahalanobis metric, Manhattan metric, standardized Euclidean metric and Spearman measure. In addition, three typical distance weighting schemes were tested: No weighting that is every instance has weight of 1, inverse weighting ($\frac{1}{dist}$) and squared inverse weighting ($\frac{1}{dist^2}$) with respect to distance $dist$. Each one of the distance measure and weight combination (24 altogether) was tested with k values $\{1, 3, \ldots, 57\}$. Only odd k values were tested in order to exclude the possibility of ties. Moreover, k value is limited with the size of a training set. Due to NLOO procedure the maximum value for k was 57. Our classification task is a two-class problem and, hence, ties do not occur when odd k values are used. The NLOO classification procedure was repeated with all aforementioned k values

The kernel selection is an important issue when LS-SVM is used in practice. The purpose of the kernel is to map data from the input space to a high-dimensional space where the two classes can be separated using a maximum margin hyperplane. Only kernels satisfying the conditions of Mercer's theorem [38] can be used. The notation of $K(\cdot, \cdot)$ refers to a kernel function which is needed for constructing the LS-SVM classifier and when a new instance is classified. For our research we selected eight kernel functions which are represented in Table 2.

Parameter settings are kernel and dataset dependent on LS-SVM. In LS-SVM a common parameter for all kernels is regularization parameter C which controls a trade-off between a maximum margin and minimum classification error. In addition, hyperparameters are encountered with different kernels and the number of hyperparameters vary between each kernel from 0 to 2. For the RBF kernel we

need a hyperparameter σ describing the width of Gaussian function. For the hyperbolic tangent kernel there are two hyperparameters (κ and δ). Hyperparameters must satisfy constraints $C, \sigma, \kappa > 0$ and $\delta < 0$ or, otherwise, kernels do not fulfill the conditions of Mercer's theorem.

Table 3 shows the detailed information related to parameter value spaces of each kernel used. Based on the information given in Table 3 we performed extensive parameter value search for all kernels. Since the main objective of this paper is not to present a novel parameter optimization process (e.g. applying some new evolutionary algorithm), we performed traditional one, two or three dimensional grid search depending on the kernel.

3.3 Performance measures

Table 4 shows the confusion matrix for our study in general form. There are numerous performance measures developed for two-class classification problems which can be derived from the confusion matrix. We selected seven of them to be used. Firstly, we examined true positive (TP) and true negative (TN) values which describe the number of correctly classified controls and patients respectively. Secondly, we chose true positive rate (TPR) also known as sensitivity or recall and true negativity rate (TNR) also called specificity. These performance measures explain the proportion of correctly classified controls and patients within the selected group and supplement the TP and TN information. In other words, we have

$$TPR = \frac{TP}{TP + FN} 100\%$$

and

$$TNR = \frac{TN}{TN + FP} 100\%.$$

Thirdly, we use accuracy as one performance measure. Accuracy is commonly applied and it explains the overall performance (i.e. how many patients and controls were classified correctly with respect to the size of the dataset). It is one of the most used performance measure in machine learning community and, thus, used also in this paper. Accuracy can be defined in different ways but we use the following definition for it

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} 100\%.$$

Although accuracy is widely used, it does not always give the right perspective from the results, if it is used alone. For example, if we have a dataset where the

one class is very small and the other is very large, we can obtain high accuracy even all the instances from the smaller class would be misclassified. In order to avoid this kind of problem, we use other performance measures together with accuracy. Fourthly, we use F_1-score which is the weighted average of precision (PRE)

$$PRE = \frac{TP}{TP + FP}$$

and TPR [17]. F_1-score can be interpreted in terms of TP, FP and FN as follows

$$F_1 = \frac{2TP}{2TP + FN + FP}.$$

Since F_1-score does not take into account true negative value, we decided to use also Matthews correlation coefficient (MCC) [3, 26] as one performance measure. MCC gives as an output a number from the interval $[-1, 1]$ where 1 indicates perfect classification and -1 total misclassification. MCC is commonly used and it describes the classification results well even the class sizes would be highly skewed. More specifically, MCC can be defined as follows

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

4 Results

Tables 5 and 6 present the detailed classification results. In Table 5 k-NN algorithm was used with various distance measures and weightings. A separate table for the k-NN results was given due to a large number of distance measures and weighting combinations. Thus, the analysis of results would be easier for a reader. The optimal parameter value has not been given in Table 5 since NLOO approach may give different optimal k value for each training set.

When examining the results of Table 5, we notice that four of the total 24 distance and weighting combinations achieved 88.3% accuracy. These were cosine measure with (squared) inverse weightings and (standardized) Euclidean metric with squared inverse weighting. Overall, Table 5 includes relatively large differences between distance measures which can be seen in ACC, F_1 and MCC performance measures. The worst result in terms of accuracy was obtained by the Spearman measure having maximum of 75.0% accuracy. However, good results yielded by the k-NN classifier may be explained with the fact that k-NN is a local classifier which can perform well with rather small datasets.

The classification of the controls (students) in Table 5 was performed slightly better than that of the patients since in many cases 28 controls out of 30 were recognized correctly. For the patients the respective number was 25. One possible reason why the controls were recognized better than the patients might be that the controls are more a homogeneous group compared to the patients. Moreover, the mean age of controls compared to patients may have influence on this kind of result. The patients may have symptoms in a wide scale from mild to very difficult. This increases the heterogeneousity of the patients and may affect on the classification.

Table 6 shows the classification results when other than k-NN algorithm were used in classification. Similarly, as in Table 5 we do not present the optimal parameter values in Table 6 for LS-SVM and RF classifiers due to NLOO procedure applied. Compared to Table 5 results Table 6 shows similar results with respect to accuracies. Now, naïve Bayes classifier with and without KDE yielded the highest accuracy (86.7%) which is an interesting result. The success of simple NB classifier may be explained with the size of the datasets (60 instances and five features). Probability based NB classifier works with small datasets whereas LS-SVM and RF classifiers may work better with larger datasets.

Discriminant analysis based methods, LR and LS-SVMs had good accuracies, F_1 score and MCC values. Accuracy, for example, was between 83.3% and 85.0% which is comparable with the k-NN results given in Table 5. For CART algorithm 78.3% and 80.0% accuracies were obtained depending on the value of "minparent" parameter. Random Forests, which can be considered as an extension for the CART algorithm, obtained 80.0% accuracy. The result is good but it did not outperform other classification methods.

The same trend with respect to the classification of controls versus patients was seen in Table 6 as in Table 5. The controls were, generally speaking, classified better than the patients which supports the hypothesis of the homogeneousity of the controls compared to the patients. Moreover, the age of the controls can also reflect to the classification results.

5 Discussion and conclusion

This paper dealt with classification of patients and controls using stabilogram derived data. Our study is a natural continuation for our study [30] where the same dataset was first used and examined. The focal idea was to find a suitable machine learning method which would separate patients and controls as well as possible

and the classification results would be improved from the results given in [30]. The stabilogram signal data originates from 30 controls and 30 patients suffering from Ménière's disease which is one of the most common inner ear disorders encountered. From the stabilogram signal data five features were extracted and used in classification.

Overall, we performed 45 different test setups with extensive parameter value search. We managed to achieve 88.3% accuracy using k-NN classifier. From the other classification methods naïve Bayes and LS-SVMs showed very good results having 85% or above accuracies. The results showed that patients were more difficult to classify than controls which indicates that controls may be more homogeneous group compared to patients. This again shows that there can be locality behavior into some extent in our dataset.

Stabilogram signal data measured with a force platform are difficult. Finding correct features from the signal data can be challenging and the differences between signals of different subjects can be very small. However, this study shows that separating patients from controls can be done with a high performance measures when the right features and machine learning algorithm are found.

Ménière's disease is the most commonly diagnosed inner ear disease and it is encountered worldwide. The variety in severity of Ménière's disease and inner ear diseases, generally speaking, is large. Some people may have mild symptoms and live relatively normal life whereas in the worst case scenario an inner ear disease can be disabling. A very important issue is to diagnose the disease as early as possible and to start a proper treatment so that the quality of life would be as good as possible. At the same time we can save funds and resources both from the individual point of view and from the society perspective.

Our study concentrated on the first step of diagnose, classification of a control and a patient. When we have a machine learning method which tells to a physician whether or not a person is probably a patient, a physician can use this information in his/her diagnose. Machine learning can be a helpful tool for a physician when diagnosing an inner ear disease which again enables physician to give more efficiently proper instructions for a patient how to improve his/her condition.

Although our paper focused on separating patients and controls especially in the context of Ménière's disease, our research can be extended into other domains as well. Balance is an important issue as already in the introduction was stated and there are domains where good balance is a key point. In sports and sports medicine balance measurements can be performed to improve possible disorders and skills. Moreover, another domain where this research can be applied is game industry. Nowadays, various accessories can be attached to computers and consoles. A

few years ago Nintendo designed a balance board called Wii Balance Board [44] which could be used for muscle training for example. Hence, this research has possibilities to extend even to a larger consumer markets.

From the methodological point of view there are possibilities how to improve our research further. In future, when we have a larger dataset, machine learning techniques from the field of artificial neural networks (ANNs) can be considered as possible classification method. ANNs were left out from this study due to the small number of instances. Artificial neural networks have proven good results in many applications, but they may require a large training set (especially deep learning techniques) in order to work properly.

Acknowledgements

The first author is thankful for Finnish Cultural Foundation Pirkanmaa Regional Fund for the support.

References

[1] J. Audiffren, I. Bargiotas, N. Vayatis, P. Vidal, and D. Ricard, A nonlinear scoring approach for evaluating balance: Classification of elderly as fallers and non-fallers, *Plos One* 11(12)(2016), e0167456.

[2] S. Balakrishnama and A. Ganapathiraju, Linear discriminant analysis - A brief tutorial, Technical Report, Institute for signal and information processing, 1998, pp. 1-8.

[3] P. Baldi, S. Brunak, Y.Chauvin, C.A.F. Andersen, and H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, *Bioinformatics* 16(5)(2000), 412-424.

[4] L. Bel, D. Allard, J.M. Laurent, R. Cheddadic, and A. Bar-Hend, CART algorithm for spatial data: Application to environmental and ecological data, *Computational Statistics & Data Analysis* **53**(8)(2009), 3082-3093.

[5] H.R. Bittencourt and R.T. Clarke, Use of classification and regression trees (CART) to classify remotely-sensed digital images, in: *Proceedings of the 2003 IEEE International Geoscience and Remote Sensing Symposium*, 2003, **6**, pp. 3751-3753.

[6] G. Bohling, Classical normal-based discriminant analysis, Technical report, Kansas Geological Survey, 2006, pp. 1-24. http://people.ku.edu/gbohling/EECS833

[7] L. Breiman, Random Forests, *Machine Learning* **45**(1)(2001), 5-32.

[8] K.J. Cios, W. Pedrycz, R.W. Swiniarski, and L.A. Kurgan, *Data Mining: A Knowledge Discovery Approach, Springer*, New York, NY, USA, 2007.

[9] J.J. Collins and C.J. De Luca, Open-loop and closed-loop control of posture: A random-walk analysis of center-of-pressure trajectories, *Experimental Brain Research* **95**(2)(1993), 308-318.

[10] T. Cover and P. Hart, Nearest neighbor pattern classification, *IEEE Transactions on Information Theory* **13**(1)(1967), 21-27.

[11] Z.A. Dastgheib, O.R. Pouya, B. Lithgow, and Z. Moussavi, Comparison of a new ad-hoc classification method with support vector machine and ensemble classifiers for the diagnosis of Meniere's disease using EVestG signals, in: *Proceedings of the 2016 IEEE Canadian Conference on Electrical and Computer Engineering*, 2016, pp. 1-4.

[12] S. Dreiseitl and L. Ohno-Machado, Logistic regression and artificial neural network classification models: a methodology review, *Journal of Biomedical Informatics* **35**(5-6)(2002), 352-359.

[13] M. Duarte and V. Zatsiorsky, On the fractal properties of natural human standing, *Neuroscience Letters* **283**(3)(2000), 173-176.

[14] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2001.

[15] S.A. Dudani, The distance-weighted k-nearest-neighbor rule, *IEEE Transaction on Systems, Man and Cybernetics* **6**(4)(1976), 325-327.

[16] P. Forsman, A. Tietäväinen, A. Wallin, and E. Hæggström, Modeling balance control during sustained waking allows posturographic sleepiness testing, *Journal of Biomechanics* **41**(13)(2008), 2892-2894.

[17] C. Goutte and E. Gaussier, A probabilistic interpretation of precision, recall and F-score, with implication for evaluation, in: *Proceedings of the 27th*

European Conference on Information Retrieval, LNCS 3408, 2005, pp. 345-359.

[18] T. Hastie, R. Tibshirani, and J. Friedmann, *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, Springer, New York, NY, USA, 2nd edition, 2009.

[19] D. Hewson, N. Singh, H. Snoussi, and J. Duchene, Classification of elderly as fallers and non-fallers using center of pressure velocity, in: Proceedings of the 32nd Annual Conference of the IEEE Engineering in Medicine and Biology Society, 2010, pp. 3678-3681.

[20] H. Joutsijoki, K. Varpa, K. Iltanen, and M. Juhola, Machine learning approach to an otoneurological classification problem, in: *Proceedings of the 35th Annual Conference of the IEEE Engineering in Medicine and Biology Society*, 2013, pp. 1294-1297.

[21] K.S. Kim, H.H. Choi, C.S. Moon, and C.W. Mun, Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions, *Current Applied Physics* **11**(3)(2011), 740-745.

[22] I. Kurt, M. Ture, and A.T. Kurum, Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease, *Expert Systems with Applications* **34**(1)(2008), 366-374.

[23] A. Liaw and M. Wiener, Classification and regression by randomForest, *R news* **2**(3)(2002), 18-22.

[24] J.R. Lindsay, Ménière's disease: Histopathological observations, *Archives of Otolaryngology* **39**(4)(1944), 313-318.

[25] W.-Y. Loh, Classification and regression trees, *Wiley Interdisciplinary Review: Data Mining and Knowledge Discovery* **1**(1)(2001), 14-23.

[26] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 lysozyme, *Biochimica et Biophysica Acta (BBA) - Protein Structure* **405**(2)(1975), 442-451.

[27] C. Morales and E. Kolaczyc, Wavelet-based multifractal analysis of human balance, *Annals of Biomedical Engineering* **30**(4)(2002), 588-597.

[28] R. Peterka, Postural control model interpretation of stabilogram diffusion analysis, *Biological Cybernetics* **82**(4)(2000), 335-343.

[29] J. Rasku, A method for the classification of corrective activity in context dependent postural controlling tasks, *Computers in Biology and Medicine* **39**(10)(2009), 940-945.

[30] J. Rasku, H. Joutsijoki, I. Pyykkö, and M. Juhola, Prediction of a state of a subject on the basis of a stabilogram signal and video oculography test, *Computer Methods and Programs in Biomedicine* **108**(2)(2012), 580-588.

[31] J. Rasku and M. Juhola, A detection method of body movement signals measured with magnetic tracking device for human balance investigations, *International Journal of Medical Engineering and Informatics* **2**(1)(2010), 37-51.

[32] J. Rasku, I. Pyykkö, M. Juhola, M. Garcia, T. Harris, L. Launer, G. Eiriksdottir, K. Siggeirsdottir, P. Jonsson, H.J. Hoffman, H. Petersen, C. Rasmussen, P. Caserotti, E. Toppila, S. Pajala, and V. Gudnason, Evaluation of the postural stability of elderly persons using time domain signal analysis, *Journal of Vestibular Research* **22**(5-6)(2012), 243-252.

[33] J. Ren, S.D. Lee, X. Chen, B. Kao, R. Cheng, and D. Cheung, Naive Bayes classification of uncertain data, in: *Proceedings of 2009 Ninth IEEE International Conference in Data Mining*, 2009, pp. 944-949.

[34] I. Rish, An empirical study of the naive Bayes classifier, in: *Proceedings of the International Joint Conference of Artificial Intelligence, Workshop of Empirical Methods in Artificial Intelligence*, 2001, pp. 41-46.

[35] K. Safi, S- Mohammed, F. Attal, Y. Amirat, L. Oukhellou, M. Khalil, J.-M. Gracies, and E. Hutin, Automatic segmentation of stabilometric signals using hidden markov model regression, *IEEE Transactions on Automation Science and Engineering* (in press) (2017).

[36] F. Scoppa, R. Capra, M. Gallamini, and R. Schiffer, Clinical stabilometry standardization: basic definitions-acquisition interval-sampling frequency, *Gait Posture* **37**(2)(2013), 290-292.

[37] D. Stephens, I. Pyykkö, K. Varpa, H. Levo, D. Poe, and E. Kentala, Self-reported effects of Ménière's disease on the individual's life: A qualitative analysis, *Otology & Neurology* **31**(2) (2010), 335-338.

[38] J.A.K. Suykens, T. van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least squares support vector machines*, World Scientific, New Jersey, USA, 2002.

[39] J.A.K. Suykens and J. Vandewalle, Least squares support vector machines, *Neural Processing Letters* **9**(3)(1999), 293-300.

[40] K. Varpa, K. Iltanen, and M. Juhola, Machine learning method for knowledge discovery experimented with otoneurological data, *Computer Methods and Programs in Biomedicine* **91**(2)(2008), 154-164.

[41] K. Varpa, K. Iltanen, and M. Juhola, Genetic algorithm based approach in attribute weighting for a medical data set, *Journal of Computational Medicine* **2014**(2014), Article ID 526801.

[42] K. Varpa, H. Joutsijoki, K. Iltanen, and M. Juhola, Applying one-vs-one and one-vs-all classifiers in k-nearest neighbour method and support vector machines to an otoneurological multi-class problem, in: *User Centred Networked Health Care - Proceedings of MIE 2011*, 2011, pp. 579-583.

[43] S.-C. Yeh, M.-C. Huang, P.-C. Wang, T.-Y. Fang, M.-C. Su, P.-Y. Tsai, and A. Rizzo, Machine learning-based assessment tool for imbalance and vestibular dysfunction with virtual reality rehabilitation system, *Computer Methods and Programs in Biomedicine* **116**(3)(2014), 311-318.

[44] Wikipedia, Wii Balance Board, `https://en.wikipedia.org/wiki/Wii_Balance_Board`. Accessed January 29, 2018.

[45] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Wu, Z.-H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, Top 10 algorithms in data mining, *Knowledge and Information Systems* **14**(1)(2008), 1-37.

Table 1: Describing information of the dataset.

Class	Sex (M/F)	Frequency	Proportion	Average age (years)
Controls	28/2	30	50.0%	25 ± 4
Patients	13/17	30	50.0%	60 ± 12
Total	41/19	60	100.0%	$-$

Table 2: LS-SVM kernels used in our study and their descriptions.

Kernel	Description
Linear	$K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$
Quadratic	$K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^T\mathbf{y})^2$
Cubic	$K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^T\mathbf{y})^3$
Polynomial kernel ($degree = 4$)	$K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^T\mathbf{y})^4$
Polynomial kernel ($degree = 5$)	$K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^T\mathbf{y})^5$
Polynomial kernel ($degree = 6$)	$K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^T\mathbf{y})^6$
Radial Basis Function	$K(\mathbf{x}, \mathbf{y}) = \exp(-\\|\mathbf{x} - \mathbf{y}\\|^2/2\sigma^2)$
Hyperbolic tangent	$K(\mathbf{x}, \mathbf{y}) = \tanh(\kappa\mathbf{x}^T\mathbf{y} + \delta)$

Table 3: LS-SVM kernels and the parameter value spaces used in this study.

Kernel	C	σ	κ	δ
Linear	$\{2^{-7}, \ldots, 2^{10}\}$	-	-	-
Quadratic	$\{2^{-7}, \ldots, 2^{10}\}$	-	-	-
Cubic	$\{2^{-7}, \ldots, 2^{10}\}$	-	-	-
Polynomial kernel ($degree = 4$)	$\{2^{-7}, \ldots, 2^{10}\}$	-	-	-
Polynomial kernel ($degree = 5$)	$\{2^{-7}, \ldots, 2^{10}\}$	-	-	-
Polynomial kernel ($degree = 6$)	$\{2^{-7}, \ldots, 2^{10}\}$	-	-	-
Radial Basis Function	$\{2^{-7}, \ldots, 2^{10}\}$	$\{2^{-7}, \ldots, 2^{10}\}$	-	-
Hyperbolic tangent	$\{2^{-7}, \ldots, 2^{10}\}$	-	$\{2^{-7}, \ldots, 2^{4}\}$	$\{-2^{3}, \ldots, -2^{-7}\}$

Table 4: Confusion matrix for two-class classification problem.

		Predicted label	
		Control	Patient
True label	Control	TP	FN
	Patient	FP	TN

Table 5: The results of k nearest neighbors method variants. Abbreviations TP, TPR, TN, TNR, ACC, F_1 and MCC stand for true positive, true positive rate, true negative, true negative rate, accuracy, F_1 score and Matthews correlation coefficient. From the performance measures TPR, TNR and ACC are presented in percentages.

	Controls		Patients				
Method	TP	TPR	TN	TNR	ACC	F_1	MCC
Chebychev measure and equal weights	26	86.7	24	80.0	83.3	0.84	0.67
Chebychev measure and inverse weights	26	86.7	24	80.0	83.3	0.84	0.67
Chebychev measure and inverse squared weights	26	86.7	24	80.0	83.3	0.84	0.67
Manhattan metric and equal weighting	27	90.0	24	80.0	85.0	0.86	0.70
Manhattan metric and inverse weighting	27	90.0	25	83.3	86.7	0.87	0.74
Manhattan metric and squared inverse weighting	24	80.0	23	76.7	78.3	0.79	0.57
Correlation measure and equal weighting	25	83.3	23	76.7	80.0	0.81	0.60
Correlation measure and inverse weighting	25	83.3	24	80.0	81.7	0.82	0.63
Correlation measure and squared inverse weighting	26	86.7	24	80.0	83.3	0.84	0.67
Cosine measure and equal weighting	26	86.7	23	76.7	81.7	0.83	0.64
Cosine measure and inverse weighting	**28**	**93.3**	**25**	**83.3**	**88.3**	**0.89**	**0.77**
Cosine measure and squared inverse weighting	**28**	**93.3**	**25**	**83.3**	**88.3**	**0.89**	**0.77**
Euclidean measure and equal weighting	26	86.7	23	76.7	81.7	0.83	0.64
Euclidean measure and inverse weighting	26	86.7	24	80.0	83.3	0.84	0.67

Euclidean measure and squared inverse weighting	**28**	**93.3**	**25**	**83.3**	**88.3**	**0.89**	**0.77**
Mahalanobis measure and equal weighting	26	86.7	23	76.7	81.7	0.83	0.64
Mahalanobis measure and inverse weighting	27	90.0	23	76.7	83.3	0.84	0.67
Mahalanobis measure and squared inverse weighting	24	80.0	23	76.7	78.3	0.79	0.57
Standardized Euclidean measure and equal weighting	26	86.7	24	80.0	83.3	0.84	0.67
Standardized Euclidean measure and inverse weighting	**28**	**93.3**	**25**	**83.3**	**88.3**	**0.89**	**0.77**
Standardized Euclidean measure and squared inverse weighting	26	86.7	25	83.3	85.0	0.85	0.70
Spearman measure and equal weighting	26	86.7	19	63.3	75.0	0.78	0.51
Spearman measure and inverse weighting	26	86.7	18	60.0	73.3	0.76	0.48
Spearman measure and squared inverse weighting	26	86.7	18	60.0	73.3	0.76	0.48

Table 6: The results of CT, LDA, LR, LS-SVM, MDA, NB, QDA and RF classifiers. Abbreviations TP, TPR, TN, TNR, ACC, F_1 and MCC mean true positive, true positive rate, true negative, true negative rate, accuracy, F_1 score and Matthews correlation coefficient. From the performance measures TPR, TNR and ACC are presented in percentages.

	Controls		Patients				
Method	TP	TPR	TN	TNR	ACC	F_1	MCC
Classification tree (CART) ($minparent = 1$)	24	80.0	23	76.7	78.3	0.79	0.57
Classification tree (CART) ($minparent = 5$)	25	83.3	23	76.7	80.0	0.81	0.60
Classification tree (CART) ($minparent = 10$)	24	80.0	23	76.7	78.3	0.79	0.57
Linear discriminant analysis	26	86.7	24	80.0	83.3	0.84	0.67
Logistic regression	25	83.3	25	83.3	83.3	0.83	0.67
LS-SVM Linear kernel	26	86.7	24	80.0	83.3	0.84	0.67
LS-SVM Quadratic kernel	27	90.0	24	80.0	85.0	0.86	0.70
LS-SVM Cubic kernel	26	86.7	24	80.0	83.3	0.84	0.67
LS-SVM 4th degree polynomial kernel	27	90.0	24	80.0	85.0	0.86	0.70
LS-SVM 5th degree polynomial kernel	27	90.0	24	80.0	85.0	0.86	0.70
LS-SVM 6th degree polynomial kernel	27	90.0	24	80.0	85.0	0.86	0.70
LS-SVM RBF kernel	26	86.7	24	80.0	83.3	0.84	0.67
LS-SVM Sigmoid kernel	26	86.7	24	80.0	83.3	0.84	0.67
Mahalanobis discriminant analysis	27	90.0	24	80.0	85.0	0.86	0.70
Naïve Bayes (normal distribution assumption)	**27**	**90.0**	**25**	**83.3**	**86.7**	**0.87**	**0.74**
Naïve Bayes (kernel smoothing density estimation and triangle kernel)	**27**	**90.0**	**25**	**83.3**	**86.7**	**0.87**	**0.74**

Naïve Bayes (kernel smoothing density estimation and Epanechnikov kernel)	**27**	**90.0**	**25**	**83.3**	**86.7**	**0.87**	**0.74**
Naïve Bayes (kernel smoothing density estimation and box kernel)	26	86.7	25	83.3	85.0	0.85	0.70
Naïve Bayes (kernel smoothing density estimation and Gaussian kernel)	**27**	**90.0**	**25**	**83.3**	**86.7**	**0.87**	**0.74**
Quadratic discriminant analysis	26	86.7	24	80.0	83.3	0.84	0.67
Random Forest	25	83.3	23	76.7	80.0	0.81	0.60