Accuracy comparison of dimensionality reduction techniques to determine significant features from IMU sensor-based data to diagnose vestibular system disorders

Serhat İkizoğlu [a], Saddam Heydarov [b],*

[a] Control and Automation Engineering Dept., Istanbul Technical University, Istanbul, Turkey
[b] Electronics Technologies, Istanbul Gelisim University, Istanbul, Turkey

A R T I C L E   I N F O

A B S T R A C T

This study is a significant step gone to develop Machine Learning (ML) algorithm to apply to gait sensory information collected from people to identify Vestibular System (VS) disorders. Although ML is widely used as diagnostic tool in medical decision-making, there is not much research done on application of ML methods to identify VS imperfections. In this paper, we compared the accuracies of two dimensionality-reduction techniques to use with SVM with Gaussian Kernel: Feature Selection (FS) and Feature Transformation (FT) methods. T-test and Sequential Backward Selection (SBS) were used for FS and Principal Component Analysis (PCA) and Kernel Principal Component Analysis (KPCA) with polynomial and Gaussian kernels were used as FT method. Both methods were applied to the dataset formed by 22 features collected from 37 people, of whom 21 were healthy and 16 subjects had VS-disorders. The highest accuracy among FT methods was 89.2%, while it was 81.1% for FS method. SVM with Gaussian Kernel, trained with the dataset of reduced dimensionality, had computation time of few hundreds of milliseconds, which makes real-time data processing possible. The importance of this work will obviously increase with the increase in the number of initial features. As a next step, we aim to increase dataset and use additional features extracted from pressure sensors placed under the feet. We also aim to use time domain characteristics of the features to increase overall accuracy as a next step.

1. Introduction

THIS work is the continuation of the study we have done before to compare the accuracy of three ML methods applied to the sensory information taken from people with Vestibular system (VS) disorders and healthy individuals [1].

VS includes the parts of inner ear and brain and processes information taken from human sensors. It is a very important part of human balance system. To diagnose disorders related to human balance systems, clinicians use mobile balance equipment and analyse recorded body sway [2]. The current system has its drawbacks in terms of consumed time and feasibility. Automated disease classification can, on the other hand, reduce time dramatically and result in fast and reliable diagnosis of VS disorders.

ML is widely being used in the field of disease identification where faster and reliable results are required. Jaymin P. et al. [3]

used ML for heart disease predictions and applied Decision Tree method with 10-fold cross validation using WEKA tool. They also made comparison between different Decision Tree algorithms and reported their accuracies. In [4], Srivatsa S. and Parthiban G. applied ML to data set formed by information taken from diabetic people in order to predict heart disease. They first used Naive Bayes method for data mining purpose to find out those who suffer from heart related problems. They further used Support Vector Machines (SVM) with Gaussian Kernel with 10-fold cross validation as classification model. Mehmet K. and Tolga E. [5] used and Multilayer Perceptron (MLP) for diabetes identification. They applied SVM linear, SVM with polynomial kernel and SVM with radial basis function and MLP to the features extracted from the data obtained from UCI web page. They compared accuracy of these algorithms and reported SVM linear to be the most accurate one among the compared algorithms. Subha R. et al. [6] gave brief review about ML techniques applied to cardiovascular disease identification and stated the importance of the classifier and the features chosen for training model in order to reach accurate identification of the disease. Lalaantika et al. [7] developed and tested ML method for prediction of global hypokinesia, heart related disease, through MRI

* Corresponding author.
   E-mail addresses: ikizoglus@itu.edu.tr (S. İkizoğlu), sheydarov@gelisim.edu.tr (S. Heydarov).

images. They extracted all of their features to train their model from MRI images and reported that their model performed with highest accuracy on independent test set. Trambaiolli, L. R et al. [8] used ML algorithm for Alzheimer-Disease (AD) identification. They made use of SVM algorithm and extracted features from large data set formed by EEG signals. They also compared the accuracy of SVM algorithm for different feature combinations. Khan A. and Muhammed U [9] gave a review about ML approaches for diagnosis of AD. In their review, they compared different approaches and stated the 4-step model - as pre-processing, feature selection, classification and class threshold - to be used as reference for AD diagnosis.

Machine learning has also an application area in the field of gait analysis. Müller H. et al. [10] applied ML to gait analysis data taken from hospitals participating in MD-PAEDIGREE to classify data to three categories: healthy, people with Neurological and Neuromuscular Diseases (NND) or people with Juvenile Idiopathic Arthritis. They compared the accuracies of Random Forest, Boosting, Multilayer Perceptron and SVM classifiers and reported the accuracy of Random Forest, SVM and Multilayer Perceptron to be 100% and 96.4% for Boosting classifier. They also stated that the training and testing time of all models were of milliseconds providing opportunities for real-time application. In [11], Nuttaki C. investigated the use of ML for distinctly identifying gender-specific characteristics. They used mobile smartphones with built-in accelerometers as wearable sensors worn on subject's shoulder, elbow, wrist, hip, knee and ankle to collect gait data. They extracted stride length and width, step length and width, and torque of each joint as features and applied Naïve Bayes, J48 tree and SVM as ML classifiers to the dataset formed by these features. The reported accuracies for Naïve Bayes, J48 and SVM were 99%, 58% and 99.2%, respectively. They also pointed out that the study can be extended to identify key biomarkers for normal and diseased conditions. Wu J. [12] used the manifold learning algorithm applied to gait data to improve gait classification performance. He used Isometric Feature Mapping algorithm as nonlinear feature extraction method to decrease the size of the feature matrix formed by gait data that was collected from the strain gauge force platform embedded in 10 m long walkway path. After that, he applied SVM as ML model to test the accuracy. He analysed gait data of young and elderly people and reported that SVM combined with nonlinear feature extraction model performed better when compared to linear SVM. He stated that manifold learning algorithm can be used to find low-dimensional gait data inside high dimensional feature matrix that can improve model accuracy and it has potential for future clinical applications. Another research done by LeMoyne. R, et al. [13], uses ML to classify legs affected by hemiplegic disparity. They applied ML model to the gait data that was extracted through force plate measurements. The ground reaction force parameters from force plate measurements were used to derive features and Logistic regression was applied to the feature matrix to identify affected and unaffected leg. The reported accuracy was 100%.

Feature extraction method is also widely used to determine relevant features and to remove redundant ones, so that overall accuracy is improved and wrong classification rate is reduced. Rueda A. and Krishnan S. [14] used machine learning to diagnose Parkinson's disease in early stage. They identified prominent set of parameters that can best represent the disease. By using sustained vowel recordings, they found Mel-Frequency Cepstral Coefficients and used them as discriminative features. Aich S. et al. [15] compared accuracy of machine learning algorithm applied to different set of features. They compared Random Forest classification algorithm with original feature set extracted from voice recordings and with transformed set of features by using Principal Component Analysis (PCA). They found that Random Forest with PCA based feature reduction method showed better accuracy with 96.83%

compared to original data set. They furthermore stated that Intrinsic Mode Functions were discriminative features. In [16], Vipani R. et al. employed ML algorithm to develop an automatic classifier for Parkinson disease (PD) and Huntington's (HD) diseases. They used gait data extracted from 63 subjects and applied Hilbert transform for feature extraction. Using Logistic Regression, they classified the subjects. They tested the classifier accuracy using MATLAB and found the accuracy of the model to identify PD and HD as 85.22%. The accuracy of the algorithm to detect healthy subjects was found to be 87.79%. Baby M. et al. [17] proposed a method to distinguish healthy people from the ones diagnosed with PD based on statistical features extracted from pre-processed gait analysis data using Wavelet Transform. After that, Artificial Neural Network (ANN) was used as classification algorithm and its performance was compared with SVM and Naive Bayes classifier.

ML together with feature transformation (FT) and feature selection (FS) techniques is also used as classification method for diabetes, kidney disease and heart diseases. In [18], Vaishali, R. et al. made use of ML to increase the accuracy of models used to predict Type-2 Diabetes. By using Goldberg's Genetic algorithm applied to pre-processed features, they decreased the number of features by 50% and used Multi Objective Evolutionary Fuzzy Classifier on the new data set. They found the accuracy around 83% by separating the data set as 70% training and 30% test sets. Wibawa M. et al. [19] developed machine learning based disease identification tool for early diagnosing of chronic kidney disease (CKD). They used correlation based feature selection method to identify discriminative features among 24 parameters including signs, symptoms and risk factors that might be related to CKD. They used UCI ML repository containing 400 instances. They compared three different ML models, namely, SVM, k-Nearest Neighbour (kNN) and Naive Bayes and found that kNN with feature selection method showed the highest accuracy with 98.1%. Zhao T.T. et al. [20] applied machine learning to classify heart disease, which is one of the most important research areas in clinical decision support systems. They proposed discriminant minimum class locality preserving canonical correlation analysis (DMPCCA) to extract discriminative features from the dataset formed by 1579 patients. They compared three different feature extraction techniques, namely, PCA, DMPCCA and Canonical Correlation Analysis (CCA). They applied SVM to processed feature sets and found DMPCCA to give the best result. Yekkala I. et al. [21] also used ML for heart disease prediction. They analysed three different models: Bagged Tree, Random Forest and AdaBoost with Particle Swarm Optimization (PSO) as feature subset-selection method to predict with high accuracy. They found that Bagged Tree, when used with PSO, gave the best experimental result. Subanya B. and Rajalaxmi R. R. [22] studied metaheuristic algorithm to determine the optimal feature subset that improves classification accuracy and removes redundant parameters causing wrong classifications. They used Swarm Intelligence based Artificial Bee Colony (ABC) to find discriminative features and found that SVM, used together with ABC, showed high accuracy with just only seven features.

Although there is lots of work done on disease identification using ML, almost no study has been carried out on ML application to VS disorders. One of the very few studies is the one from Yeh S.C. et al. [23], where the authors used SVM as ML tool to analyse quantified balance indices of the patients who had gone through designed interactive virtual reality rehabilitation program. The focus of the study was to see the effect of virtual reality as rehabilitation program on quantified balance indices parameter. Machine learning was not used as a disease identification tool in that study.

Table 1 shows the relevance of the research items used in the Literature Review part to our current research.

In the whole of our project, we are developing a ML algorithm to use for identification of VS imperfections. In our previous study,
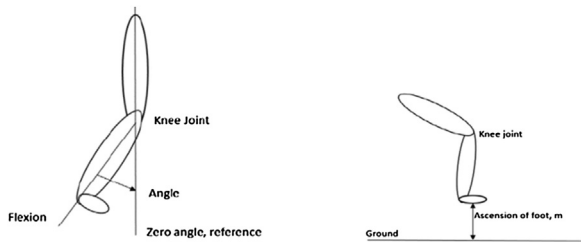
Fig. 1. A visual for ascension made by foot during walking.

we searched for the method to give the highest accuracy where we examined three ML methods [1]. SVM with Gaussian Kernel was found to perform best with an accuracy of 83.3%. When applying the learning model, one of the most important steps to improve the dataset quality is the dimensionality reduction process. In this paper, we compare the accuracies of two different dimensionality-reduction techniques, the Feature Selection (FS) and the Feature Transformation (FT) methods and search for the one to present the highest accuracy as dimensionality reduction tool. In this context, the rest of the paper is arranged as follows: In Section 2, we describe the dataset formation and features used to train the models. In Section 3, we give brief background information about the feature reduction techniques. This section is followed by introduction of experimental results and the set-up description. Finally, we submit conclusion and state the future work we plan to do.

2. Background informatıon

2.1. Dataset formation

Dataset was formed by using sensory information taken from 37 people, out of which 21 were healthy and 16 people were diagnosed to have VS imperfections. Among all the subjects, 11 people from VS group and 10 people from healthy group were female. The mean age of healthy participants and people with VS imperfections was 48.4 and 55.6, respectively. People with VS imperfections were suffering from different types of balance disorders such as Benign Paroxysmal Positional Vertigo (BPPV), Multiple Sclerosis (MS), Vestibular Neuritis (VN) etc. The patients were primarily diagnosed using Dynamic Posturography. Computerized Dynamic Posturography is a clinical method to determine the problem behind the balance disorder of the patient that differentiates between sensory, motor and central adaptive functional impairments [24]. It makes use of a movable platform on which the patient tries to control the balance using information from the vestibular, visual and proprioceptive systems. Sensors were placed on subjects' feet, knee and waist, consistent with literature about motion tracking and gait analysis [25,26]. We asked people to walk an 11.5 m long straight path and during this period, sensory data were collected for further data processing and feature extraction. The 22 features we used to train our models were highly consistent with the ones widely used in literature about gait analysis [27–30]. As dataset dimension was not too big and there were not enough samples for each VS-disorder group, only binary classification was performed, where subjects were classified as belonging either to 'healthy' or to 'VS-disorder' group. The 22 features used to train the learning models and definitions for some of them are given in Tables 2 and 3, respectively. Figs. 1–3 visualize some of the features used.

2.2. Dimensionality reduction

Dimensionality reduction is a pre-processing tool for high dimensional data analysis, especially for the dataset with small number of samples and large number of features. In the case where

Table 1
Summary of the Literature Review part.

Literature review	The relevance to the current research
[1]	Here we presented our previous study. The result of this study (SVM with Gaussian kernel) was used to compare the effectiveness of feature extraction methods.
[2]	General overview on how VS disorders are normally diagnosed is presented.
[3–9]	These papers show some fields (heart disease, PD, AD, diabetes) where ML is widely used as disease identification tool. We showed what methods they used and what were the results. Our aim was to give some overview to the reader about ML application in medical decision making.
[10–13]	Here we presented ML application in the gait analysis field. We investigated which sensory information are used (IMU, force plates, mobile phones as wearable sensors etc.), how the model accuracies are increased. In our research, we analysed gait parameters to diagnose VS imperfection, so we aimed to make the reader familiar with current gait analysis methods.
[14–17]	Aim of this paper is to find feature extraction method that can increase accuracy and decrease computation time which will be even more important as data set dimension increases. Therefore, in these reviews, we showed why feature extraction is used, what were the ML model accuracies applied to original dataset and to the dataset formed by feature extraction methods.
[18–22]	Here we presented general pipeline on how FT and FS are used to increase accuracy of the model. The pipeline can be summarized as; • apply different FT and FS methods to form new feature set • apply the ML model to the feature set • if the accuracy is increased, use that method to extract discriminative features. We used similar pipeline in our research.
[24]	Here we showed one of the few researches that we were able to find where ML was applied in VS field. Our aim was to show that ML is rarely used in VS field, and also, in these studies ML is not used as disease identification tool, rather it is used for different purposes.

Table 2
Features used to traın ml models.

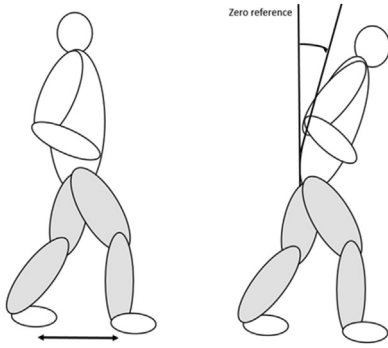| Feature no	Feature definition
1	Right foot average step length
2	Left foot average step length
3	Average speed
4	Total travelled distance - left foot
5	Total travelled distance - right foot
6	Step symmetry 1 (total travelled distance - right foot / total travelled distance - left foot)
7	Step symmetry 2 (total travelling time - right foot / total travelling time - left foot)
8	Left knee bending angle
9	Average bending angle - left knee
10	Right knee bending angle
11	Average bending angle - right knee
12	Maximum lateral waist swing to left
13	Maximum lateral waist swing to right
14	Left knee maximum swing
15	Right knee maximum swing
16	Average swing - right knee
17	Average swing - left knee
18	Average ascension - right foot
19	Average ascension - left foot
20	Waist - maximum posterior swing
21	Waist - maximum anterior swing
22	Waist inclination angle during walking

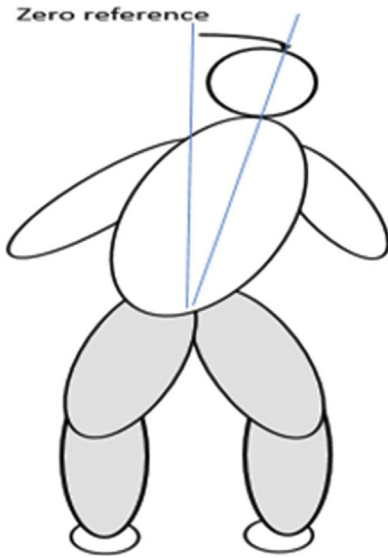**Fig. 2.** A visual for step length.

Fig. 3. A sample illustration for waist inclination angle during walking.

data set is small, using many features may lead to the learning algorithm to over-fit the data; also, it may cause high algorithm complexity. There are two major kinds of dimensionality reduction methods, FS and FT, where the former chooses smallest feature subset from original features leading to highest learning algorithm accuracy; the latter transforms features to a new reduced space by preserving most of the relevant information related to previous subset [31]. Feature selection method selects the subset from the initial feature set to have least number of features by removing redundant, irrelevant or noisy data and thus improving the data set quality [32]. On the other hand, FT performs some transformation depending on the learning model and transforms correlated features to a new subspace [33].

2.2.1. Feature selection method

Generally, high dimensional data can include features that may be irrelevant, thus increase feature-set dimension, which results in increasing the search space and algorithm computation time. Feature selection process is a method used to select subset of features from original space that contains least number of features giving highest accuracy [34]. Advantages of feature selection method can be summarized as follows [32]:

- Reduces feature space dimension and decreases computation time.
- Increases data set quality by removing irrelevant features.
- The resulting model has higher accuracy.

- Makes data visualization easier due to reduced number of features.

Feature selection can be categorized into three classes as filter, wrapper and hybrid method. Filter method is the simplest one and does not consider the classification algorithm. It just performs statistical analysis on data set to choose feature subset yielding highest accuracy [35]. Wrapper method, on the other hand, performs selection by considering the learning algorithm, which slows down the feature selection process [36]. Filter methods have less computation time compared to wrapper methods as they do not involve any learning model, but it results in less reliability [34]. Hybrid method uses advantages of both the filter and wrapper methods. It uses independent method to choose feature subset and evaluates its performance on model accuracy by applying it to the learning model used [37].

Widely used filter method in bioinformatics is to apply univariate criterion on each feature while assuming there is not any interaction between them. We used the *t*-test and compared p values, that is, absolute value of t statistics to see how effective it is in distinguishing people with VS disorder from healthy individuals. The *t*-test, also known as Student's test, is a widely used statistical method in comparing two groups. It is a method of testing hypothesis about the mean of samples drawn from normally distributed population when the standard deviation of the population is unknown. It has two types, namely, the independent and the paired *t*-test. In independent *t*-test, one sample from two independent normally distributed populations are compared, while in paired *t*-test the samples are not randomly selected; the second sample is obtained by doing certain changes on the first sample [38]. For the given binary classification problem, t value of sample can be calculated as [39];

$$t(x^i) = \frac{|\mu_{i1} - \mu_{i2}|\sqrt{n_1 n_2}}{\sqrt{n_2 \sigma_{i1}^2 + n_1 \sigma_{i2}^2}} \tag{1}$$

where μ_{i1} and σ_{i1} denotes the mean and standard deviation of the i-th feature of the first class, while μ_{i2} and σ_{i2} are the mean and standard deviation of the second class. n_1 and n_2 are number of samples of the first and second class, respectively.

Sequential backward selection (SBS) was used as wrapper method to find the feature subset that increases model accuracy. Unlike the filter method, in the wrapper type feature selection, the learning model is also taken into account. SBS algorithm starts training by using all features and continues by removing features until there is no change in the model accuracy.

2.2.2. Feature transformation method

PCA is a statistical method that reduces dimensionality of the raw matrix. It uses an orthogonal transformation to convert set of observations of variables that may have correlation between them to new uncorrelated variables, where the new subspace has a dimension less than or equal to previous feature space. These new variables are called principle components and the problem formulation is based on choosing the proper number of principle components for the dimensionality reduction.

Before running the PCA algorithm, raw feature data set is normalized with zero-mean and unit-variance [40]. Thus, each normalized feature vector $x_N^{(i)}$ will be of the form

$$x_N^{(i)} = \frac{x^{(i)} - \mu}{\sigma} \tag{2}$$

where μ is the mean value of the *i*-th feature vector $x^{(i)}$ and σ is the standard deviation of $x^{(i)}$.

The problem definition to find the optimal hyper-plane can be given as determining the hyper-plane, on which sum of squares of
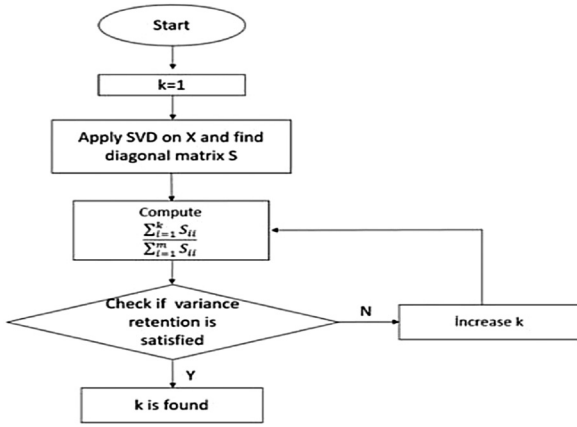
Fig. 4. PCA algorithm to find the right value of k.

projections of samples are minimum. PCA algorithm, as explained above, reduces the dimension of feature matrix from n to k and the main design parameter for this method involves choosing the right k value. The procedure to find the proper k value can be summarized as follows [40].

a) Calculate covariance matrix, $\sum$, using normalized features
b) Calculate S matrix (diagonal matrix) using singular value decomposing applied on $\sum$
c) Set k = 1
d) Using 99% variance retention as rule of thumb, calculate the ratio using expression (3);

$$\frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{m} S_{ii}} \geq 0.99 \tag{3}$$

where S is diagonal matrix found by SVD and m is number of features

e) If inequality is violated, stop
f) Else, increase k & go to step (d)

Fig. 4 shows the flow diagram for the PCA method based on S matrix.

In standard PCA method, the data is assumed to be in low dimensional linear subspace. When the data is low dimensional nonlinear subspace kernel modified PCA (KPCA) is used [41]. The main idea of KPCA is to increase the original dataset to higher dimensional space, where linear separation is possible. The steps to follow to apply KPCA method can be given as follows [42];

• Transform original dataset to new high dimensional space using chosen nonlinear kernel function
• Compute the covariance of the new dataset and apply SVD to find matrix U, which contains principal components
• Chose first k principal components from matrix U
• Use principal components to find new transformed features
• Use new features and apply the learning model to find the accuracy

Polynomial and Gaussian kernel functions were used to transform the original dataset to new high dimensional space.
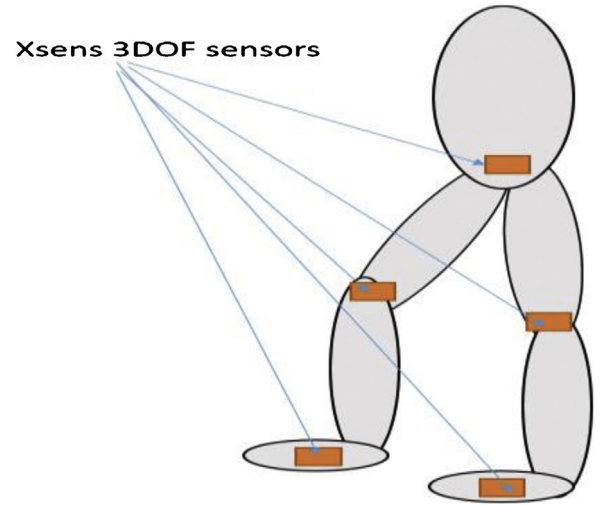

Fig. 5. Illustration for sensor placement on the body.

Table 3
Some important features and their definitions.

Parameter	Definition
Waist posterior oscillation [deg]	Waist angle as a result of oscillation done during walking
Lateral oscillation [deg]	Oscillation done to the right or left wrt stationary position of human body
Ascension made by foot [m]	Distance between foot and ground during toe off
Speed [cm/s]	Ratio of length of walking path to time spent on walking

3. Experimental setup and results

3.1. Data collection

Sensory information from human subjects was collected using MATLAB and MTW2 Wireless 3DOF Motion Tracker IMU sensors from Xsens, which include 3D accelerometer, gyroscope and magnetometer [43]. IMU sensors provide acceleration data for gait analysis. Nevertheless, these sensors suffer from bias errors which result in high position errors if not corrected. Thus, we have developed algorithms to correct position data for both the feet and the knees [44]. During the data acquisition process, attention was paid that the Xsens sensors were within the communication range of Bluetooth to avoid packet loss due to the communication interruption. Also, Xsens uses proprietary radio protocol called Awinda, based on low-cost 2.4 GHz ISM chipsets, which detects and handles occasional packet loss in real-time processing. When the communication is lost, the Xsens sensors store data in the buffer and retransmit it when the communication is back. After retransmission, the data is removed from buffer. These buffers can store up to 1000 data packets. Taking into account that the sampling rate of our sensors was 100 Hz, 10 s of missed data can be retransmitted when wireless communication is back [45]. The sensors are placed on waist, knees and feet of the subjects as shown in Figs. 5 and 6 [25,26]. Technical specifications of the sensor are listed in Table 4. By examining some of the specifications like roll, pitch and static accuracy given in Table 4, it can be seen that the sensors are accurate enough to use for gait analysis. The data was collected in Cerrahpasa Medical School-Istanbul and attention was given to the fact that IMU sensors would not be affected by environmental conditions such as magnetic fields created by nearby devices [44]. Therefore, we collected data mainly on weekends with electronic devices in offices turned off in order to minimize the effect of magnetic field
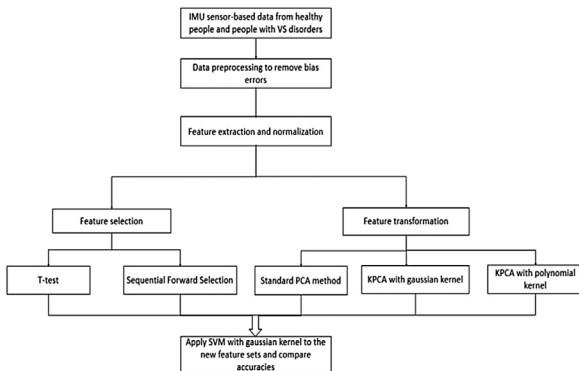
Fig. 6. Human subject with wearable motion sensors.

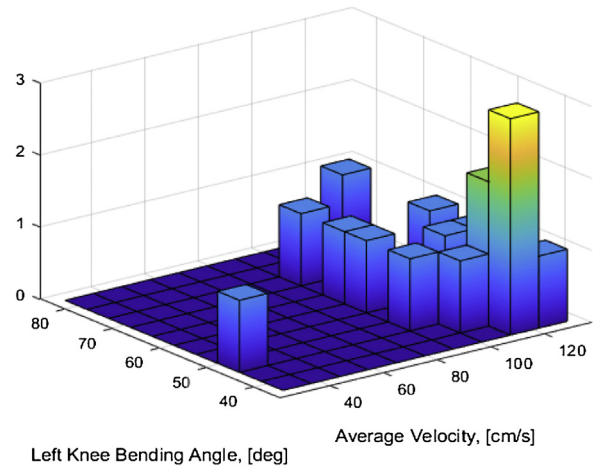Fig. 7. The block diagram of the overall methodology.
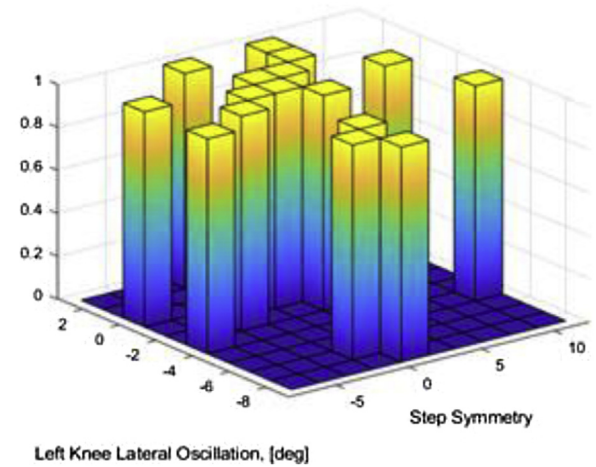

Fig. 8. 3D histogram plot for left knee bending angle vs average velocity.

Fig. 9. 3D histogram plot for left knee lateral oscillations vs step symmetry.
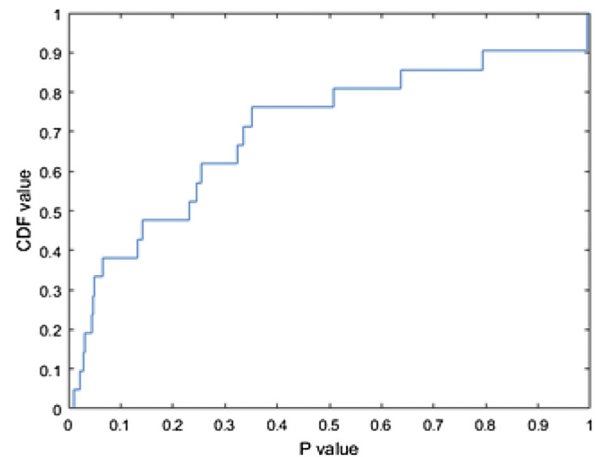

Fig. 10. CDF plot for p-values.

interference. The subjects were asked to move on a flat path to be sure having zero change in z-axis data (Fig. 6). Furthermore, Ethical Committee approval was taken to conduct the experiments. Experimental procedure was explained to subjects before the tests and they were asked to sign consent form.

3.2. Results

The block diagram of the overall methodology is given in Fig. 7. The initial 22 features were supposed to correlate more or less within themselves, so that the number of discriminative features would be significantly less than this number. Figs. 8 and 9 show 3D histogram plots of some important features for VS group. Observing these figures provides the pre-understanding that 'left knee bending angle' is somewhat correlated with 'average velocity', whereas 'left knee lateral oscillations' and 'step symmetry' are independent features.

We used MATLAB Classification App Learner toolbox for data processing, model training and testing performances of feature selection and transformation algorithms.

T-test filter method and SBS as wrapper method were used for feature selection, while standard PCA and KPCA were applied as feature transformation method together with SVM with Gaussian Kernel to the training dataset. Both methods were applied to the feature set consisting of 37 people. In order to get better visualization about how effectively the two groups were separated by
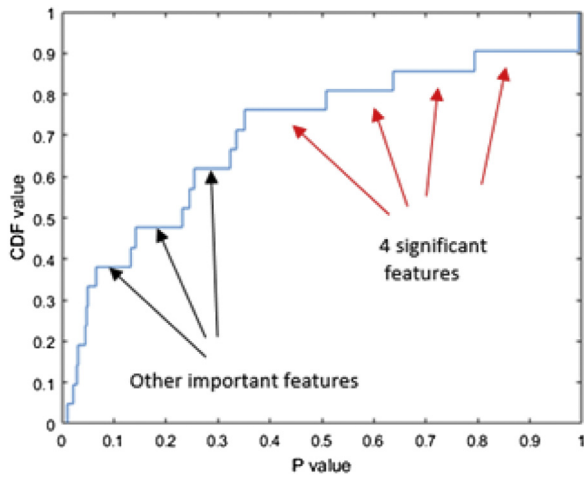
Fig. 11. Recognizing the number of most significant features that lead to maximum training accuracy.

Fig. 12. ROC curve for the learning model with SBS method.

Table 4
Technical specifications of the sensors used.

Parameter	Value	Unit
Roll, pitch, accuracy	<0.5	degree
Static accuracy	<1	degree
Dynamic accuracy	2	degree RMS
Angular resolution	0.05	degree
Internal sampling rate	1800	Hz
Maximum acceleration	16	g
Maximum update rate	75	Hz

Table 5
Some key model parameters before and after t-test.

Parameters	Before t-test	After t-test
Training time	6.18 seconds	0.75 seconds
Prediction speed	610 observations/second	1200 observation/second

Table 6
Some key model parameters before and after SBS.

Parameters	Before SBS	After SBS
Training time	6.18 seconds	0.85 seconds
Prediction speed	610 observations/second	800 observation/second

Table 7
Some key model parameters before and after PCA.

Parameters	Before PCA	After PCA
Training time	6.18 seconds	0.9 seconds
Prediction speed	610 observations/second	240 observation/second

Table 8
Some key model parameters before and after KPCA with polynomial kernel.

Parameters	Before KPCA With Polynomial Kernel	After KPCA With Polynomial Kernel
Training time	6.18 seconds	1.05 seconds
Prediction speed	610 observations/second	220 observation/second

each feature, Cumulative Distribution Function (CDF) of p-values are plotted in Fig. 10. Considering p < 0.05, we recognize that about 65% of the features fulfil this null hypothesis, in other words, around 14 features ($65\% * 22 \approx 14$) can be discriminative according to t-test. To find out the discriminative features, we used misclassification error (MCE) parameter, which is defined as the ratio of misclassified observations to total observations [46].

MATLAB was used to evaluate MCE. We found that four features led to minimum MCE. The result was intuitive, as by examining the graph in Fig. 10, one can recognize that there are four most significant features (Fig. 11).

In Fig. 11, the red arrows indicate the features that increase the model accuracy around 15–20% each. Though there are even some other features on the CDF plot (pointed with black arrows) that increase the accuracy around 5–10% each, we can comment that those features are effective to rise the accuracy -for example- from 75% to 80%, while they cannot increase it from 60% to 65% etc.

Using MATLAB, the most significant four features found were average velocity, left knee lateral swing, average ascent by right foot and average step length'. These four features were used to form a training set for ML and to determine the accuracy of SVM with Gaussian Kernel. Accuracy was found to be 72.2%. Together with the accuracy, training time and prediction speed are other important parameters to point to overall performance of the method.
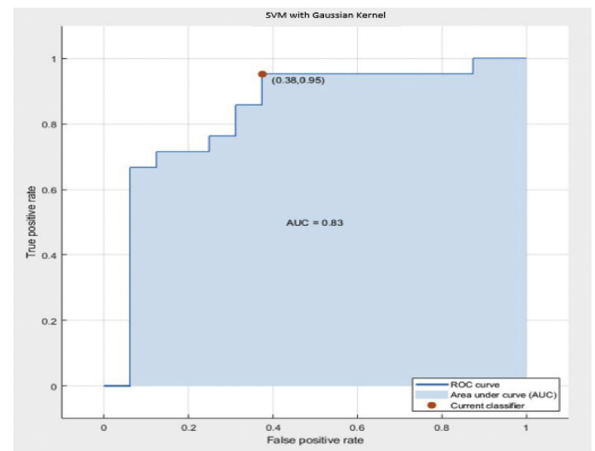
Table 5 shows numerical values for training time and prediction speed parameters before and after t-test application.

To test the SBS method, MATLAB built-in "sequentialfs" function has been used and search and performance criteria were set as "backward" and classifier performance, respectively. By applying SBS method, 16 features, were found to be discriminative. The most discriminative 16 features were used to train the learning model. The accuracy was found to be 81.1%. Receiver Operating Characteristics was used to visualize the performance of the learning model trained with features found by SBS method (Fig. 12). Table 6 shows performance metric of the model before and after SBS method application.

On the other side, we applied the PCA procedure where we first normalized the feature matrix using zero mean - unit variance method. Next, SVD was performed using MATLAB and 22 × 22 U and S matrices were found. The diagonal S matrix was used to find the proper k value with predefined 99% variance specification. The k value was determined to be 13; that is, the original feature space consisting of 22 features was transformed to a new space with 13 features. Using these 13 features, which are in fact a linear combination of the original 22 features, we trained SVM with Gaussian kernel and found the accuracy as 82.6%. Table 7 presents values for
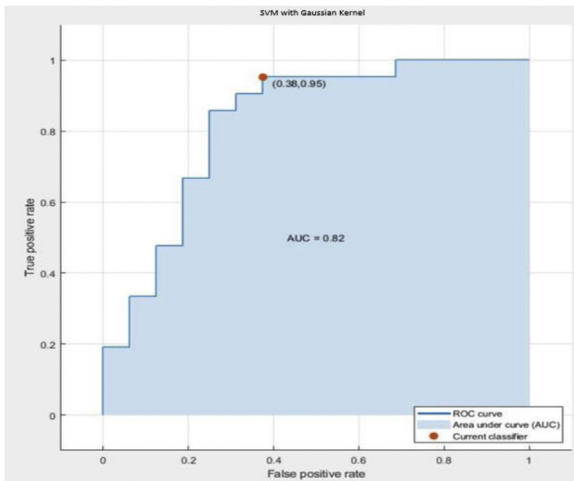
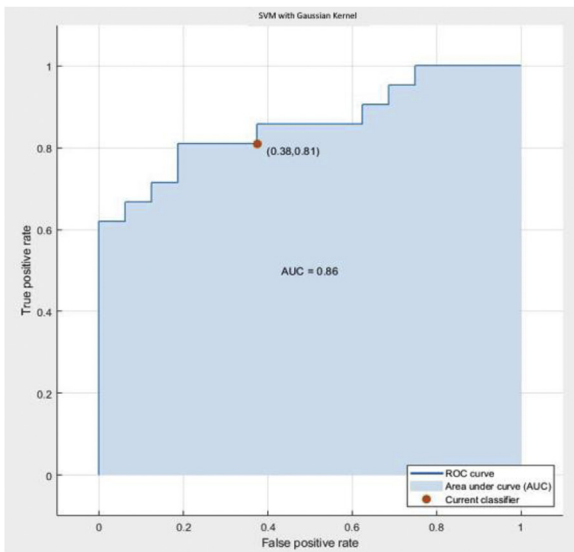**Fig. 13.** ROC curve for the learning model for KPCA with polynomial kernel.

Fig. 14. ROC curve for the learning model for KPCA with Gaussian kernel.

Table 9
Some key model parameters before and after KPCA with Gaussıan kernel.

Parameters	Before KPCA With Gaussian Kernel	After KPCA With Gaussian Kernel
Training time	6.18 seconds	1.2 seconds
Prediction speed	610 observations/second	170 observations/second

training time and prediction speed parameters before and after PCA application.

Next, we applied KPCA algorithm where polynomial and Gaussian kernel function were used for nonlinear transformation of data set to new high dimensional space. By applying standard PCA in new space and setting k = 13, we found transformed features for both kernel functions. After the feature transformation process, SVM with Gaussian kernel was used as the learning model to test the accuracy. The accuracy for KPCA with polynomial kernel function and Gaussian kernel function was found to be 81% and 89.2%, respectively. Figs. 13 and 14 shows ROC curves for KPCA with polynomial and gaussian kernel functions, respectively.

Tables 8 and 9 show performance values in terms of training time and prediction speed for KPCA with polynomial and Gaussian kernels, respectively.

As a next step, we increased the number of features found as a result of feature selection method to equate to 13, number of features used in PCA method, to see whether the model accuracy would capture the feature transformation method from the perspective of classification result. Thus, we added 9 more features to equate to 13 and found the learning model accuracy to be 77.8%. The result was as expected, as the features found by FS methods are subsets of original features, while the features found FT methods are linear/nonlinear combinations of original features.

4. Conclusion

This study is an important part of a project, where our goal in the whole of the project is to develop a ML algorithm to be used to identify the vestibular system disorder for individuals who suffer from imbalance when walking. The aim of this study was to apply dimensionality reduction techniques and find out or obtain discriminative features that can increase the accuracy of the algorithm and decrease the training time. We analysed 22 initial features collected from 37 people consisting of 21 healthy ones and 16 individuals with VS disorders. We applied t-test and SBS as feature selection method and standard PCA and KPCA with polynomial and Gaussian kernels as feature transformation methods to decrease the sample matrix dimension. We determined 4 discriminative features as 'ascension of right foot, average velocity, left knee lateral swing and average step length' by using t-test method, while 16 features were found to be discriminative by applying SBS method. By using MATLAB Classification App Learner, we trained SVM with Gaussian Kernel with a matrix formed by features found with t-test and SBS method from 37 people and the resulting accuracy was 72.2% and 81.1% for t-test and SBS methods, respectively On the other hand, application of PCA and KPCA methods decreased the number of features from 22 to 13. Standard PCA method provided an accuracy of 82.6%, which was slightly below than the model accuracy (83.3%), while applying KPCA with polynomial and gaussian kernel functions resulted in an accuracy of 81% and 89.2%, respectively. The high accuracy resulted with gaussian kernel function, implied that there is nonlinear correlation between features. Besides its performance regarding the accuracy, the FT method also decreased the training time dramatically, around 6 times.

The importance of dimensionality reduction process will be even more obvious when the sample matrix size will increase. In the near future, we plan to include new features obtained from insole pressure sensors. On the other hand, in this study we did not perform sub-classification between diseases in order to prevent over-fitting of the training model, as most of the patients (11 out of 16) had the same diagnosis as BPPV. As a next step, we also aim to increase dataset under each VS disorder label that will obviously raise the overall performance of the ML algorithm and will further help to sub-classify between reasons of imbalance. We are also searching for new features to increase the overall accuracy in the diagnosis. In this context, we consider to analyse the time domain characteristics of the features using chaotic approach, complex network analysis, fractality/multi-fractality etc. It is obvious that for these types of analyses we need larger datasets, so that the statistical feature extraction will be meaningful.

CRediT authorship contribution statement

Serhat İkizoğlu: Supervision. **Saddam Heydarov:** Supervision.

Acknowledgments

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at https://doi.org/10.1016/j.bspc.2020.101963.

References

[1] S. Heydarov, S. İkizoğlu, K. Şahin, E. Kara, T. Çakar, A. Ataş, Performance comparison of ML methods applied to motion sensory information for identification of vestibular system disorders, in: 2017 10th International Conference on Electrical and Electronics Engineering (ELECO), IEEE, 2017, pp. 1112–1116.

[2] D. Basta, M. Rossi-Izquierdo, A. Soto-Varela, A. Ernst, Mobile posturography: posturographic analysis of daily-life mobility, Otol. Neurotol. 34 (2) (2013) 288–297.

[3] K.N.R. Challa, V.S. Pagolu, G. Panda, B. Majhi, An improved approach for prediction of Parkinson's disease using machine learning techniques, arXiv preprint (2016), arXiv:1610.08250.

[4] G. Parthiban, S.K. Srivatsa, Applying machine-learning methods in diagnosing heart disease for diabetic patients, Int. J. Appl. Inf. Syst. (IJAIS) 3 (2012), 2249-0868.

[5] M.S. Kurt, T. Ensarı, Diabet diagnosis with support vector machines and multi-layer perceptron, in: Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2017, IEEE, 2017, pp. 1–4.

[6] R. Subha, K. Anandakumar, A. Bharathi, Study on cardiovascular disease classification using machine learning approaches, Int. J. Appl. Eng. Res. 11 (6) (2016) 4377–4380.

[7] L. Sharma, G. Gupta, V. Jaiswal, Classification and development of tool for heart diseases (MRI images) using machine learning, in: Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference, IEEE, 2016, pp. 219–224.

[8] L.R. Trambaiolli, A.C. Lorena, F.J. Fraga, P.A. Kanda, R. Anghinah, R. Nitrini, Improving Alzheimer's disease diagnosis with machine learning techniques, Clin. EEG Neurosci. 42 (3) (2011) 160–165.

[9] A. Khan, M. Usman, Early diagnosis of alzheimer's disease using machine-learning techniques: a review paper Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference, vol. 1, IEEE, 2015, pp. 380–387.

[10] R. Joyseeree, R.A. Sabha, H. Müller, Applying Machine Learning to Gait Analysis Data for Disease Identification, MIE, 2015, pp. 850–854.

[11] C. Nutakki, H. Edakkepravan, S. Gunasekaran, L.P. Ramachandran, V. Sasi, B. Nair, S. Diwakar, Torque analysis of male-female gait and identification using machine learning, in: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2018, pp. 2103–2106.

[12] J. Wu, Automated recognition of human gait pattern using manifold learning algorithm, in: 2012 8th International Conference on Natural Computation, IEEE, 2012, pp. 199–202.

[13] R. LeMoyne, W. Kerr, T. Mastroianni, A. Hessel, Implementation of machine learning for classifying hemiplegic gait disparity through use of a force plate, in: 2014 13th International Conference on Machine Learning and Applications, IEEE, 2014, pp. 379–382.

[14] A. Rueda, S. Krishnan, Feature analysis of dysphonia speech for monitoring parkinson's disease, in: Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE, IEEE, 2017, pp. 2308–2311.

[15] S. Aich, K. Younga, K.L. Hui, A.A. Al-Absi, M. Sain, A nonlinear decision tree based classification approach to predict the Parkinson's disease using different feature sets of voice data, in: Advanced Communication Technology (ICACT), 2018 20th International Conference, IEEE, 2018, pp. 638–642.

[16] R. Vipani, S. Hore, S. Basak, S. Dutta, Gait signal classification tool utilizing hilbert transform based feature extraction and logistic regression based classification, in: Research in Computational Intelligence and Communication Networks (ICRCICN), 2017 Third International Conference, IEEE, 2017, pp. 57–61.

[17] M.S. Baby, A.J. Saji, C.S. Kumar, Parkinson's disease classification using wavelet transform based feature extraction of gait data, in: Circuit, Power and Computing Technologies (ICCPCT), 2017 International Conference, IEEE, 2017, pp. 1–6.

[18] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, S. Nalluri, Genetic algorithm based feature selection and MOE fuzzy classification algorithm on pima indians diabetes dataset, in: Computing Networking and Informatics (ICCNI), 2017 International Conference, IEEE, 2017, pp. 1–5.

[19] M.S. Wibawa, I.M.D. Maysanjaya, I.M.A.W. Putra, Boosted classifier and features selection for enhancing chronic kidney disease diagnose, in: Cyber and IT Service Management (CITSM), 2017 5th International Conference, IEEE, 2017, pp. 1–6.

[20] T.T. Zhao, Y.B. Yuan, Y.J. Wang, J. Gao, P. He, Heart disease classification based on feature fusion Machine Learning and Cybernetics (ICMLC), 2017 International Conference, vol. 1, IEEE, 2017, pp. 111–117.

[21] I. Yekkala, S. Dixit, M.A. Jabbar, Prediction of heart disease using ensemble learning and particle swarm optimization, in: Smart Technologies For Smart Nation (SmartTechCon), 2017 International Conference, IEEE, 2017, pp. 691–698.

[22] B. Subanya, R.R. Rajalaxmi, Artificial bee colony based feature selection for effective cardiovascular disease diagnosis, Int. J. Sci. Eng. Res. 5 (5) (2014) 606–612.

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

[24] M.S.P.A. Eric Miller, M. Hall, Computerized Dynamic Posturography: Clinical Application & Contribution in Three Cases, 2001.

[25] D. Roetenberg, H. Luinge, P. Slycke, Xsens MVN: Full 6DOF Human Motion Tracking Using Miniature Inertial Sensors, Xsens Motion Technologies BV, Tech. Rep., 2009.

[26] A.R. Anwary, H. Yu, M. Vassallo, Optimal foot location for placing wearable IMU sensors and automatic feature extraction for gait analysis, IEEE Sens. J. 18 (6) (2018) 2555–2567.

[27] R. Takeda, S. Tadano, A. Natorigawa, M. Todoh, S. Yoshinari, Gait posture estimation using wearable acceleration and gyro sensors, J. Biomech. 42 (15) (2009) 2486–2494.

[28] S. Tadano, R. Takeda, K. Sasaki, T. Fujisawa, H. Tohyama, Gait characterization for osteoarthritis patients using wearable gait sensors (H-Gait systems), J. Biomech. 49 (5) (2016) 684–690.

[29] S. Qiu, Z. Wang, H. Zhao, H. Hu, Using distributed wearable sensors to measure and evaluate human lower limb motions, IEEE Trans. Instrum. Meas. 65 (4) (2016) 939–950.

[30] A. Ng, Support Vector Machines. Machine Ascension Learning, 2008.

[31] N. Chumerin, M.M. Van Hulle, Comparison of two feature extraction methods based on maximization of mutual information, in: 2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, IEEE, 2006, pp. 343–348.

[32] L. Ladha, T. Deepa, Feature selection methods and algorithms, Int. J. Comput. Sci. Eng. 3 (5) (2011) 1787–1797.

[33] S. Keller, A.C. Braun, S. Hinz, M. Weinmann, Investigation of the impact of dimensionality reduction and feature selection on the classification of hyperspectral EnMAP data, in: 2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), IEEE, 2016, pp. 1–5.

[34] S. Khalid, T. Khalil, S. Nasreen, A survey of feature selection and feature extraction techniques in machine learning, in: 2014 Science and Information Conference, IEEE, 2014, pp. 372–378.

[35] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1-2) (1997) 273–324.

[36] L. Yu, H. Liu, Feature selection for high-dimensional data: a fast correlation-based filter solution, Proceedings of the 20th International Conference on Machine Learning (ICML-03) (2003) 856–863.

[37] L. Rangarajan, Bi-level dimensionality reduction methods using feature selection and feature extraction, Int. J. Comput. Appl. 975 (2010) 8887.

[38] P. Crepel, S.E. Fienberg, J. Gani, Statisticians of the Centuries, Springer Science & Business Media, 2013.

[39] B. Chandra, M. Gupta, An efficient statistical feature selection approach for classification of gene expression data, J. Biomed. Inform. 44 (4) (2011) 529–535.

[40] A. Ng, Principal Component Analysis. Machine Learning, 2008.

[41] I.K. Timotius, I. Setyawan, A.A. Febrianto, Face recognition between two person using kernel principal component analysis and support vector machines, Int. J. Electr. Eng. Inform. 2 (1) (2010) 55–63.

[42] H.M. Ebied, Feature extraction using PCA and Kernel-PCA for face recognition, in: 2012 8th International Conference on Informatics and Systems (INFOS) (Pp. MM-72), IEEE, 2012.

[43] Xsens website: https://www.xsens.com/products/mtw-awinda/.

[44] S. Ikizoglu, K. Şahin, A. Ataş, E. Kara, T. Çakar, IMU Acceleration Drift Compensation for Position Tracking in Ambulatory Gait Analysis, ICINCO 2017-2, Madrid, 2017.

[45] M. Paulich, M. Schepers, N. Rudigkeit, G. Bellusci, Xsens MTw Awinda: Miniature Wireless Inertial-magnetic Motion Tracker for Highly Accurate 3D Kinematic Applications, Xsens, Enschede, The Netherlands, 2018.

[46] J.H. Kim, Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap, Comput. Stat. Data Anal. 53 (11) (2009) 3735–3745.

S. İkizoğlu graduated from Istanbul Technical University (ITU)- Control and Computer Engineering. He completed his doctorate at the same institute in 1992. He is currently Associate Professor at ITU-Control and Automation Engineering Dept. His area of interest mainly covers:Measurement, Instrumentation, Control and Mechatronics.

S. Heydarov graduated from Electrical and Electronics Engineering Department with Control Engineering concentration from Middle East Technical University-Turkey. He is currently Ph.D student in Control and Automation Eng. Dept. at Istanbul Technical University. He is also lecturer at Istanbul Gelisim University, Electronics Technologies department. His research area of interest includes data analysis, modeling and simulation of biomedical systems.