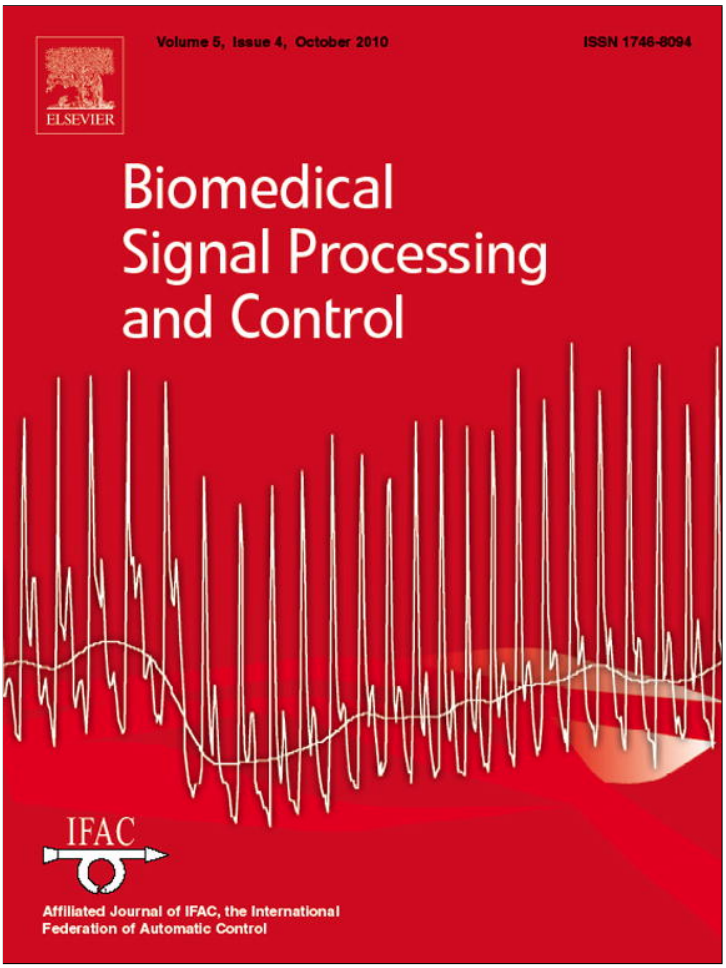

Biomedical Signal Processing and Control 5 (2010) 311–317

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Detecting fraudulent whiplash claims by support vector machines

Steinn Gudmundsson[a,*], Gudny Lilja Oddsdottir[b], Thomas Philip Runarsson[a], Sven Sigurdsson[a], Eythor Kristjansson[c]

[a] Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, Hjardarhagi 2-6, 101 Reykjavik, Iceland
[b] Faculty of Medicine, University of Iceland, Reykjavik, Iceland
[c] NeckCare, Skipholti 50c, 105 Reykjavik, Iceland

ARTICLE INFO

Article history:
Received 17 August 2009
Received in revised form 11 May 2010
Accepted 12 May 2010
Available online 12 June 2010

Keywords:
Whiplash
Insurance fraud
Time series
Classification
Support vector machines

ABSTRACT

A new method is proposed for detecting fraudulent whiplash claims based on measurements of movement control of the neck. The method is noninvasive and inexpensive. The subjects track a slowly moving object on a computer screen with their head. The deviation between the measured and actual trajectory is quantified and used as input to an ensemble of support vector machine classifiers. The ensemble was trained on a group of 34 subjects with chronic whiplash disorder together with a group of 31 healthy subjects instructed to feign whiplash injury. The sensitivity of the proposed method was 86%, the specificity 84% and the area under curve (AUC) was 0.86. This suggests that the method can be of practical use for evaluating the validity of whiplash claims.

1. Introduction

The term whiplash associated disorders (WAD) refers to a variety of clinical manifestations due to bony or soft tissue neck injuries following an acceleration-deceleration energy transfer such as a motor vehicle accident [1]. The diagnosis of WAD is an archetype of the diagnosis of a soft tissue injury where the diagnosis is mainly based on the exclusion of visible trauma on standard imaging modalities [2]. Recent figures suggest that more than 300 persons (per 100 000 in the population) with neck pain after traffic collisions are seen in emergency departments every year [3,4]. Although the majority of those diagnosed with WAD recover within the first 3 months after injury [5], a significant proportion, 20–40% remain symptomatic at 6 months and develop chronic WAD [5,6].

Patients with WAD, who are still symptomatic despite numerous physical treatments and medical care, take the natural course of action of seeking compensation for their chronic symptoms. Not surprisingly, patients with chronic WAD, have a poor reputation due to these compensation claims, especially as no demonstrable patho-anatomical signs can usually be detected to justify them. Some, who have not been hurt, may even exploit the inability of the health care system to detect a justifiable cause for the complaints to make a compensation claim. The present situation is therefore very unsatisfactory for genuine patients, the health care system and the third party payers. Insurance claims for personal injury after whiplash injury cost the United Kingdom more than £3 billion per year [7] and data from the United States reveal costs reaching $29 billion per year [8,9]. Fraudulent whiplash claims based on staged accidents are estimated to cost the UK insurance industry between £ 75 and £ 110 million every year, representing 5% of all whiplash claims [10]. More accurate diagnosis could therefore help insurers fight fraudulent claims.

While identification of the precise nature of the physical impairment in patients with WAD has proven to be difficult, some advances have been made in recent years in the development of objective assessment methods. These include tests that address sensorimotor control deficits like disturbed head–neck awareness [11], disturbed neck movement control [12] and oculomotor disturbances [13,14]. A recent study [15] used the total cervical range of motion to discriminate between whiplash and healthy subjects with good results. Grip et al. [16] report promising results on classifying WAD subjects from controls using three-dimensional neck movement data in conjunction with a neural network classifier. Dvir et al. [17] studied maximal versus feigned active cervical motion in healthy subjects and were able to differentiate effectively between the two levels of effort using the coefficient of variation of the range of motion.

"The Fly" test [12] is a computerized method to measure the accuracy of head and neck movements. It measures the ability of the patient to correct cervical spine movements on a moment-to-

* Corresponding author.
E-mail addresses: steinng@hi.is (S. Gudmundsson), glo@hi.is (G.L. Oddsdottir), tpr@hi.is (T.P. Runarsson), sven@hi.is (S. Sigurdsson), eythork@simnet.is (E. Kristjansson).

1746-8094/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bspc.2010.05.004

Fig. 1. The three tasks: easy (left), medium (middle) and difficult (right). The gray curve is the Fly trajectory and the black curve is a trajectory recorded from an asymptomatic subject.

moment basis. This is an important proprioceptive function for the regulation of movements, i.e. detection and correction of errors, via feedback and reflex mechanisms, when performing active movements. This method has demonstrated impaired movement control in patients with a history of whiplash injury when compared to healthy controls [12]. The prior Fly test has recently been reformed both as a measurement method (test) and a treatment method. This was accomplished by creating incremental difficult classes of unpredictable movement tasks according to specific criteria (see Section 2.2).

The new classification method proposed here for detecting fraudulent whiplash claims is noninvasive, inexpensive and risk free. It combines a revised version of the Fly test for cervicocephalic kinesthetic sensibility [12], with a state of the art pattern recognition algorithm. A set of features is extracted from the neck movement measurements and used as input to a statistical pattern classifier. The features include the deviation of the measured trajectory from the actual trajectory which was found useful for discriminating between asymptomatic and chronic whiplash subjects in [12]. Another feature is the rate of change of acceleration (jerk) which was found to be significantly higher in a group patients with insidious neck pain and WAD when compared to healthy subjects [18]. The features also include entropy measures which are frequently employed in the analysis of physiological time series. To the best of our knowledge, this is the first time machine learning has been applied to the problem of differentiating between subjects with genuine WAD symptoms and subjects who try to fake results for personal gain. The initial hypothesis was that neck movement in asymptomatic subjects faking neck injury was significantly different than in subjects with chronic WAD and that this difference could be reliably detected by a combination of the Fly test and a statistical pattern classifier.

2. Materials and methods

2.1. Subjects

Sixty-five individuals participated in the study, divided into two study groups. A group of 31 healthy individuals (16 women and 15 men) ages 16–67 years (mean 37.9, SD 16.7) was recruited from staff and students of various local businesses (including insurance companies) and schools. A group of 34 patients (28 women and 6 men) with chronic WAD, after being injured in motor vehicle accidents, ages 21–56 years (mean 41.3, SD 9.2) was recruited through contacts with physical therapists in Reykjavik. Samples of convenience were used in both groups. The asymptomatic group had no history of musculoskeletal pain or injury in the neck, upper back or upper arms. To be included in the WAD group, the subjects should have a history of one or more whiplash injuries and have had symptoms for more than 6 months and less than 15 years, from the neck, cervicalgi, cervicogenic headache (above 4 on the visual analog scale) and restricted movement in the upper cervical spine. Individuals were excluded if their symptoms corresponded to grades III or IV, as classified by the Quebec Task Force on whiplash-associated disorders [1] or if they had rheumatic or neurological disorders of any kind. All participants completed questionnaires recording descriptive data and general health. The WAD group completed additionally the Whiplash Disability Questionnaire [19], the Tampa Scale of Kinesiophobia [20] and the SF-36 Health Survey [21]. Ethical clearance for the study was obtained from the National Bioethics Committee and informed consent was obtained from all participants.

2.2. Measurements

The method for measuring movement control in the neck is briefly summarized here, the reader is referred to [12] for details. An electromagnetic tracking system, *3space Fastrak system* (Polhemus Inc, Colchester, VT), with a sampling rate of 120 Hz, was used in this study. This system has been used to assess position sense in the neck [22,23] and the range of motion in the neck [24] and shoulder [25]. The system computes the position and orientation of two sensors at discrete time intervals as they move through space. One sensor is placed on the forehead and the other at the back of the head. The horizontal and vertical differences between the sensor positions are used to determine the position of a cursor on a computer screen situated 1 m in front of the subject. This cursor indicates movements of the head. Another cursor on the screen (the *Fly*) traces out predetermined movement tasks. The subjects are asked to use the cursor, derived from the sensors on the head, to follow the cursor of the Fly as accurately as possible. Only the cursors are visible, not their trajectories, which makes prediction of movement difficult. A recently developed software package is used to carry out the recordings.

Three movement tasks of varying difficulty (easy, medium and difficult) were used in the study. The difficulty level was determined by the geometry of the movement tasks (Fig. 1), the velocity of the target (the Fly) and the length of the trajectories. The movement tasks differ from the ones described in [12] which all had similar geometry and trajectory lengths, and were therefore all of similar difficulty level.

2.3. Procedure

All the subjects were measured during the day. The participants were seated on a wooden chair and instructed to assume a comfortable position facing forward. The examiner explained the intention and nature of the task required of the participants. To familiarize them with the task, all the participants executed the same movement task twice. This task was different from the ones used during the measurements and was only used for instructional purposes. The participants were required to repeat each of the three movement tasks in Fig. 1 three times, with a 10 s interval between each task. The test was performed in random order across tasks and trials. The duration of the easy, medium and difficult tasks was 25, 40 and 50 s, respectively. The participants had no knowledge about the different difficulty grades of the movement tasks beforehand. After a break of 10 min, the asymptomatic subjects were asked to feign a neck disorder using the protocol from [17] which involves reading the following paragraph to the subjects: "*Imagine that one year ago you were involved in a motor vehicle collision. As a result you have suffered from various symptoms, like headache and neck pain. Today, although symptom free, you claim damages for cervical impairment after this car collision. In the next set of measurements, try to convince me that your claim is well founded and that you still suffer from those symptoms.*" The measurements were then repeated in the same way as before.

As a result, two sets of measurements exist for the asymptomatic group, sincere and feigned. In the analysis below, the feigned performance of the asymptomatic group together with the measurements from the whiplash group are used for classification.

2.4. Data analysis

The analysis was carried out using the Matlab program (Mathworks, Natick, MA) after exporting the raw data from the recording software. Each trial results in a bi-variate time series, $P(t)$, containing the on-screen x and y coordinates derived from the Fastrak system. The corresponding actual (Fly) trajectory is denoted by $Q(t)$.

In the following, the time series measurements are replaced by *features*, succinct vectorial representations of the data. Since the recording protocol used has only recently been developed, the choice of features is not obvious. Several features which were thought to be relevant were extracted from the data. A scatter plot matrix of the features was used to determine which features to use in the classification.

2.4.1. Feature extraction

The feature CDEV represents the deviation of the measured trajectory from the actual trajectory. It is defined as the average distance between $P(t)$ and $Q(t)$ over the whole time series. Closely related features are XDEV which measures the average horizontal deviation between $P(t)$ and $Q(t)$ and YDEV which measures the corresponding vertical deviation. For all three features, the largest 10% of the values are discarded prior to averaging in order to prevent minor "mishaps" having a large effect.

The INSIDE feature is another measure of closeness between the measured and actual trajectories. It is defined as the number of points which satisfy $||P(t) - Q(t)||_2 < R$ divided by the length of the series. A separate value of R was used for each task, $R = 3$ for the easy task, $R = 3.3$ for the medium difficult task and $R = 3.75$ for the difficult task. The values were taken from a previous study (unpublished.)

The fifth feature is CURVLEN, the length of the observed trajectory, normalized by the length of the actual trajectory, the lengths being computed by summing up the distances between consecutive points of the time series. A feature JERK was computed using the procedure from [26]. It is obtained by differentiating the time series of $P(t)$ three times, employing a convolution-based boxcar smoothing filter at each step to reduce noise (width 100 samples). The resulting instantaneous acceleration values are then squared and averaged to obtain a single value. As a byproduct of this computation, the feature MEANVEL, computed as the mean velocity of the user controlled cursor, is obtained. The final two features considered here are based on the spectral entropy of amplitude deviations. The entropy of the x-axis deviations is defined as $H_x = -\sum p_x \log p_x$ where p_x is the normalized power spectrum of the series $P_x(t) - Q_x(t)$. The entropy of the vertical deviations, H_y, is defined analogously. The spectral entropy quantifies how sinusoidal the signal is. A pure sinusoid has entropy zero and uncorrelated white noise has entropy one. The corresponding features are referred to as HX and HY.

2.4.2. Support vector machine ensemble

The support vector machine has over the last decade become a popular approach to pattern classification since it can deliver state-of-the-art performance on a wide variety of real-world classification problems.

Let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ denote a set with m labeled examples, where $\mathbf{x}_i \in \mathbb{R}^d$ represent the training examples and $y_i \in \{-1, 1\}$ are the labels. In the case that each subject is represented by a single trial from a single task, the training set consists of $m = 65$ examples, i.e. the total number of participants in the study. The feature vector has $d = 9$ elements corresponding to the nine different features described in Section 2.4.1 and the label -1 denotes the whiplash group and $+1$ the feign group.

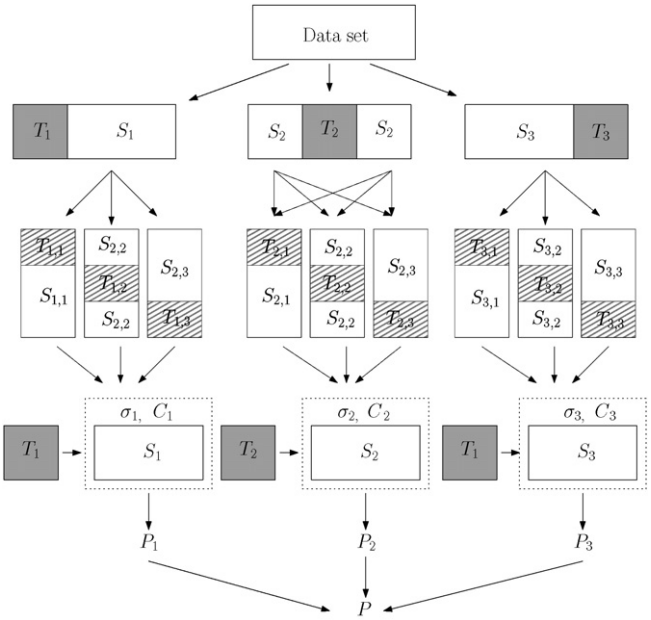

Fig. 2. The nested cross-validation procedure used to obtain an estimate of the performance measure P of the ensemble classifier. Threefold cross-validation is used in this example while 10-fold cross-validation is used in the study. The data set is split into three equally sized parts. A test partition T_i is set aside while the remaining data S_i is used for training. Each training set S_i is then partitioned further into three smaller training sets $S_{i,j}$ and three test set $T_{i,j}$ partitions. A search of "optimal" values of SVM parameters C, γ is carried out by training ensembles on each sub-partition $S_{i,j}$ and evaluating their accuracy (misclassification rate) on $T_{i,j}$. The (C, γ) pair with the highest accuracy is then used to train an ensemble on the whole S_i data set. A performance measure P_i is obtained by classifying the test set T_i. The individual performance measures P_i are finally combined in a single value P which is then reported.

SVMs seek a *maximum margin* hyperplane in feature space which separates the two classes so that the distance from the hyperplane to the closest examples, known as *support vectors*, is maximal. Maximizing the margin can be related to optimizing bounds on the expected misclassification rate of the classifier [27].

The *1-norm soft margin SVM* is obtained by solving the following quadratic optimization problem over $\mathbf{w}$, $\boldsymbol{\xi}$ and b:

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \qquad i = 1, \ldots, m \end{aligned} \tag{1}$$

where $\mathbf{w}$ is a d-dimensional vector, $\mathbf{x}i$ is an m-dimensional vector and b is a scalar. The parameter C controls the trade-off between maximizing the margin and allowing misclassified training examples. Classification of example $\mathbf{x}$ is performed by computing

$$f(\mathbf{x}) = \operatorname{sgn}\left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + b\right) \tag{2}$$

One of the most important properties of SVMs is that the mapping $\boldsymbol{\phi}$ is not explicitly needed, only the inner product (kernel) $k(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle$ between vectors $\boldsymbol{\phi}(\mathbf{x})$ and $\boldsymbol{\phi}(\mathbf{z})$ is needed. A commonly used kernel is the radial basis function which will be used in the following

$$k(\mathbf{x}, \mathbf{z}) = \exp(-\gamma||\mathbf{x} - \mathbf{z}||^2) \tag{3}$$

where $\gamma > 0$ is a user-specified scale parameter and $\boldsymbol{\phi}(\boldsymbol{x})$ is a nonlinear mapping. Solving the dual of the above optimization problem

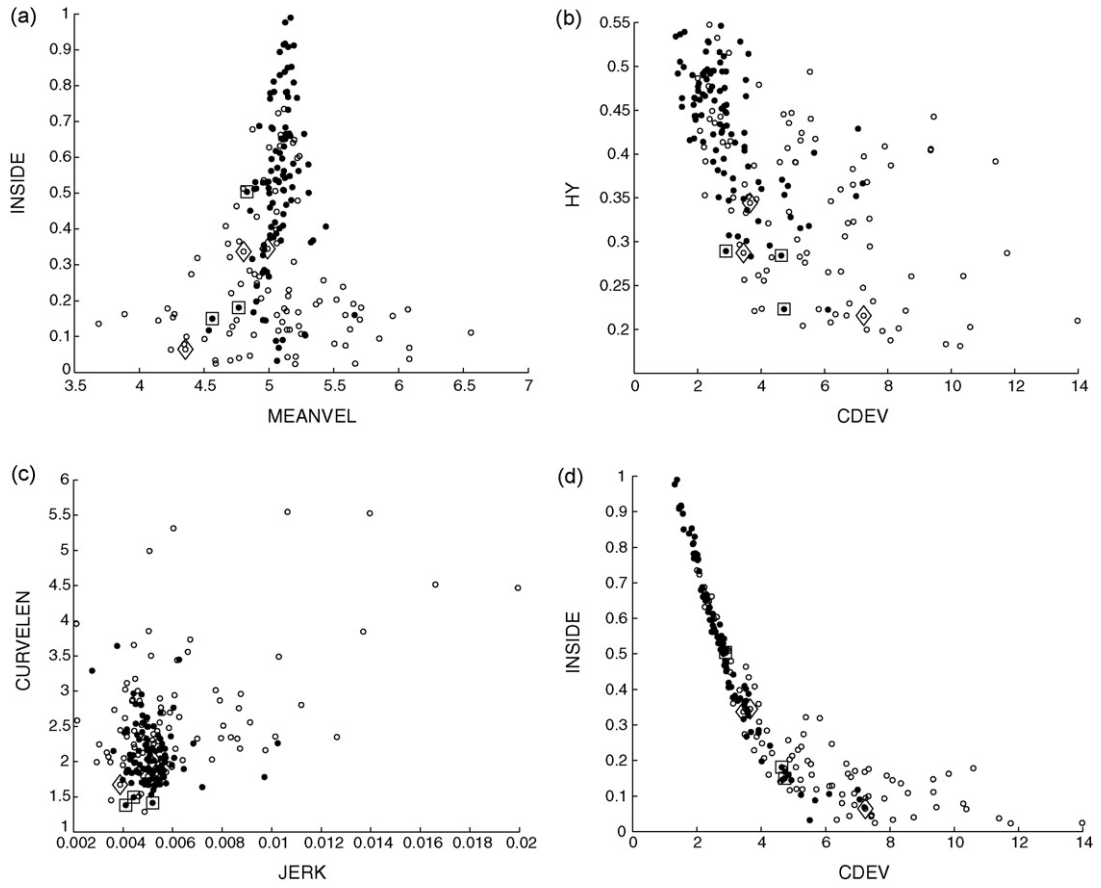

Fig. 3. Scatter plots for selected features (easy task, three trials per subject). Whiplash (black dots) and feign (open circles). The squares identify a single whiplash subject and the diamonds identify a single feign subject. Both subjects are arbitrarily chosen.

and rewriting the classification rule in terms of the dual solution avoids the need for computing vectors $\boldsymbol{\phi}(\mathbf{x})$ explicitly, only the kernel values are required. The LIBSVM package [28] was used to train the SVM classifiers and perform the subsequent classification. Each feature was linearly scaled into the interval $[0, 1]$ by subtracting the minimum of the feature values over all the training examples and dividing by the range of feature values over all the examples. The test data was scaled accordingly, using the min and max values of the training data.

Three trials are carried out for each of the three different tasks, resulting in 9 measurements per subject available during training and classification.

A single SVM classifier can be trained using data from all the trials by combining feature values for all three tasks and three trials in a single feature vector. This approach has at least two drawbacks. First, it is quite common that even healthy subjects perform badly on some of the trials, e.g. due to lapses in attention, and the corresponding elements of the feature vector will have values that are "off the charts". This problem can be mitigated somewhat by averaging feature values over trials. Secondly, multiple trials run the risk of some of them being unusable or even missing. Since SVMs are sensitive to outliers [29] and do not handle missing values properly an alternative approach based on an ensemble of classifiers was used instead. Three SVM classifiers were trained, each specializing in a single task. Each trial forms a separate training example and each subject is therefore represented by three examples (or less in case of missing data) in the training set for a given task. Classification of (new) subjects is performed by sending each available trial through the corresponding SVM (depending on the task) which subsequently casts a vote for either the whiplash or the feign class. The final classification is based on majority vote, with ties broken arbitrarily. The resulting ensemble is relatively robust towards missing data and outliers, both in the training stage and later during classification.

2.4.3. Model selection and validation

A nested cross-validation procedure was used to assess the performance of the classifier ensemble and to carry out model selection (tuning of parameters C and γ) for individual classifiers. The outer cross-validation loop (10-folds) was used to estimate a ROC curve for the classifier and was done on a per subject basis to prevent the same subject being simultaneously included in both the training and test sets. For each training set partition, a separate cross-validation procedure was used to select the best model by performing a grid search over a range of values of the C and γ parameters and selecting the pair giving the lowest cross-validation error. The procedure is further illustrated in Fig. 2.

The ensemble has three pairs of (C, γ) parameters, i.e. six hyper-parameters in all. The amount of training data available for tuning the hyper-parameters is approximately $9/10 \cdot 65 \cdot 3 \cdot 3 \approx 527$ examples which corresponds to around 88 examples per parameter.

3. Results

3.1. Feature selection

Fig. 3(a)–(d) shows selected 2D projections of the feature matrix. Each subject is represented by three dots, where a single dot corresponds to one trial. The squares and diamonds illustrate how the

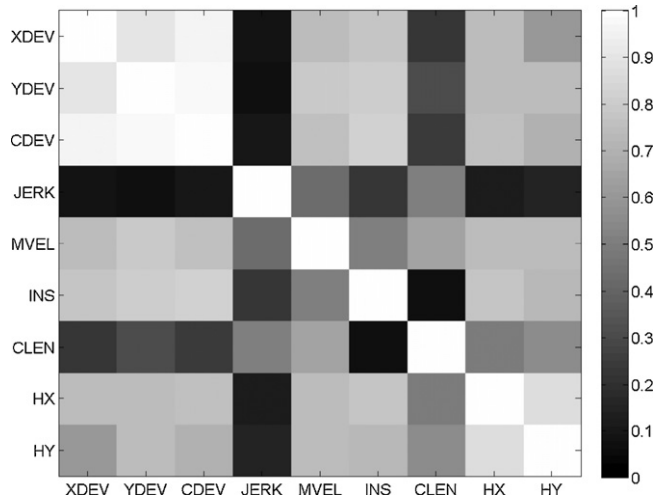

Fig. 4. Linear correlation between all feature-pairs (absolute values.) White color denotes strong correlation and black indicates no correlation. The plot was obtained by aggregating feature values from both groups for the difficult task.

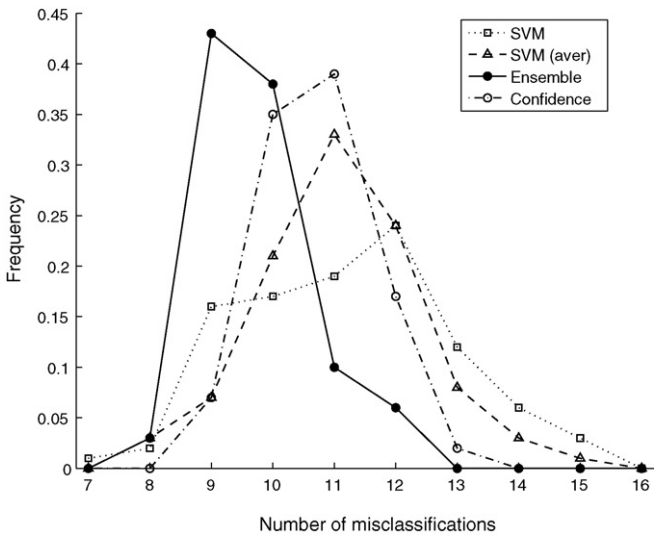

Fig. 5. Comparison of a single SVM classifier which aggregates the measurements from all tasks and trials (dotted curve), an SVM which averages over trials (dashed curve), the ensemble classifier (solid curve) and an ensemble classifier employing a confidence measure (dash-dotted curve). The graph shows how frequently a given misclassification count was obtained in repeated cross-validation runs.

results may vary between trials in two arbitrarily chosen subjects, one from the whiplash group (squares) and the other from the feign group (diamonds.) From Fig. 3(a) it can be seen that values of the INSIDE feature are considerably larger for the whiplash group indicating that the feign group is exaggerating neck movement. This is further confirmed by the feature MEANVEL, since its values for the whiplash group lie in a narrow range compared to the feign group. This behavior is also reflected in the CDEV feature in Fig. 3(b) which shows that the feign group deviates more from the actual trajectory. The same figure also shows that the entropy of the amplitude deviations is lower for the feign group, suggesting more regularity in the tracking process. Fig. 3(c) reveals that both the jerk-index and curve length features have limited ability to discriminate the two groups, at least when considered alone. The failure of the jerk index to provide useful information may be due to the fact, that even the simplest movement task is an order of magnitude more complex than the type of movement considered in [18]. It is possible that repeated differentiation amplifies noise to such an extent that any useful signal present gets lost. Finally, Fig. 3(d) shows that CDEV and INSIDE are strongly correlated. Similar examination of HX versus HY revealed significant correlation. The correlation between all pairs of features is summarized in Fig. 4. From their respective definitions it is obvious that CDEV is also strongly correlated with both XDEV and YDEV. Based on the above considerations, the set of candidate features was narrowed down to CDEV, *HY* and MEANVEL. Only two features, CDEV and HY, were included in the final feature set since the inclusion of MEANVEL was found to hurt performance.

3.2. A single SVM versus an ensemble

The ensemble classifier was first compared to a single SVM classifier where the feature values corresponding to all the tasks and trials for CDEV and H_y were combined in a single vector ($m = 65$ and $d = 3 \cdot 3 \cdot 2 = 18$). This comparison helps establish whether the added complexity of the ensemble is justified. Since the data set is relatively small, the performance estimates obtained with cross-validation are affected by the order of the training examples. The cross-validation procedure was therefore repeated 100 times, randomly shuffling the training examples each time. Fig. 5 shows the (cross-validated) estimate of the misclassification count. Using trial averages instead ($d = 3 \cdot 2 = 6$) gave almost identical results. Keeping in mind that the simulations are not independent, the ensemble appears to be more robust since it seldom misclassifies more than 10 examples. Averaging the results, the misclassification rate for a single classifier was estimated as 0.17 ± 0.02 and 0.15 ± 0.01 for the ensemble classifier. The fraction of support vectors for the ensemble classifier was 0.57 for the easy task, 0.44 for the medium difficult task and 0.46 for the difficult task (on average).

The SVM ensemble was also compared to an identical ensemble classifier which employed linear discriminant classifiers (LDA) instead of SVMs[1]. The purpose of this experiment was to investigate whether the complexity of a nonlinear SVM was justified. The misclassification rate for the LDA ensemble was 0.21 with a sensitivity of 74% and a specificity of 83%. Similar results were obtained with an ensemble of linear SVMs (maximum margin LDAs.)

3.3. Confidence measures

The proposed method does not provide a measure of confidence with the classification results. When the voting is done over multiple tasks and trials, the vote count (or their ratio) can provide such a measure. If, say, "feign" receives four votes and "whiplash" five, the clinician may simply declare the test as inconclusive. When the voting is done over few individual tasks and trials, e.g. because of missing data, this strategy is not very useful. An alternative is to use an SVM variant with posterior probability estimates and use the average of the probabilities for a single class as a measure of confidence. A classification is obtained by applying a threshold, e.g. 0.5, to the average. This strategy was tested by using the LIBSVM package to obtain individual SVMs with probability estimates. The misclassification count of the resulting ensemble is shown in Fig. 5. The confidence estimates come at the cost of a decrease in performance. This is most likely due to the fact that the probability SVM variant does not use the training data as efficiently as the standard SVM. See [28] and references therein for details of the estimation procedure.

[1] The LDAs were obtained with Matlab's `classify` function.

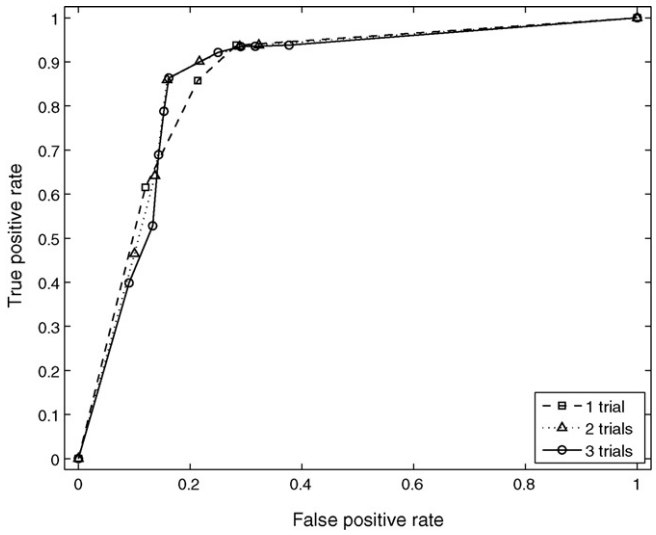

Fig. 6. ROC curve for the ensemble classifier (solid) which shows the true positive rate (sensitivity) versus the false positive rate (1−specificity). Also shown are ROC curves for a classifier which use one (dashed) or two (dotted) trials for each task.

3.4. Accuracy and ROC analysis of the ensemble classifier

An ensemble classifier with sensitivity 86% and specificity 84% was obtained (estimate based on cross-validation). This corresponds to an misclassification rate of 15%. Fig. 6 shows the ROC curve for the ensemble classifier (solid curve) obtaining by thresholding the number of votes for the whiplash group and averaging the results from 100 runs of the cross-validation procedure. The area under curve (AUC) was 0.86. Two additional curves are shown, the dashed curve corresponds to using only a single trial during classification, while the dotted curve corresponds to using two trials. In all cases, multiple trials were used in training and voting was done over all three tasks. The results suggest that using multiple trials during the prediction stage instead of just one improves performance. An improvement is obtained by going from one to two trials. Adding a third trial does not have a noticeable effect.

Fig. 7 illustrates the effect of using different tasks on classifier performance. In all cases there are three trials per task. The easy and difficult tasks give classifiers which perform worse in general than the classifier trained on the medium difficult task. The latter gives comparable results to the classifier using all three tasks. Incorporating the easy and difficult tasks does not degrade overall performance, which demonstrates the robustness of ensemble predictors in general, but then again, there is no apparent gain in performance. Visual inspection of the difficult task indicated that the subjects had some trouble completing it. In retrospect, this task may simply have been too difficult.

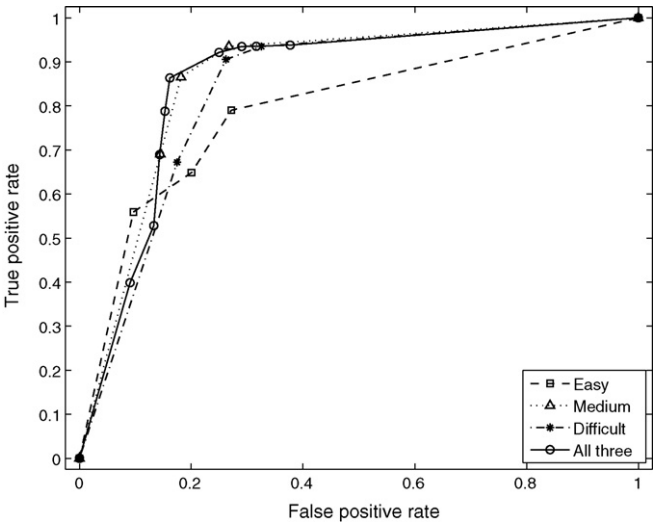

Fig. 7. ROC curve for the classifiers trained on individual tasks: easy (dashed), medium (dotted) and difficult (dash-dotted). The solid curve corresponds to the ensemble classifier trained on all three tasks.

4. Discussion

In general, irrelevant features harm classifier performance and should therefore be excluded prior to training. In clinical applications there is often preference for simple models over complex ones. Model interpretability was therefore deemed important and the feature selection was biased towards using a small number of features.

While the healthy group consisted of males and females in equal numbers, there was considerable imbalance in the WAD group where the females outnumbered males five to one. Since samples of convenience were used, this imbalance may be related to the observation that women are more prone to persistent symptoms after whiplash injuries [4,5]. If there is a significant difference in performance between males and females with WAD on the Fly test, the group mismatch will affect the results presented here. This issue warrants investigation in future studies.

Although the discrimination studies cited in Section 1 are not directly comparable to this study because of differences in instrumentation, evaluation and different target groups, it is still worthwhile to compare the accuracies reported in each of the studies. Grip et al. [16] report a sensitivity of 90% and specificity of 88% for a classifier trained on 59 WAD subjects and 56 controls. Using a measure derived from the total cervical range of motion, Prushansky et al. [15] report a sensitivity of 78% and specificity of 76% for discriminating between 101 chronic whiplash patients and 75 healthy subjects. The classification was obtained by thresholding the derived measure.

5. Conclusions

A new method is proposed for evaluating the validity of whiplash claims. The measurement process is simple, risk free and inexpensive. The accuracy in terms of sensitivity and specificity indicate that the method is of practical significance. With respect to medicolegal issues an independent study with a much larger number of subjects and a less biased gender distribution between the WAD and feign groups would be needed to confirm the results presented here.

The ensemble classifier was found to have a slight advantage in terms of accuracy over an SVM classifier which aggregates multiple measurements in a single feature vector, as well as being more robust. Furthermore, utilizing multiple trials in the ensemble during classification gives a slight performance gain over a single trial. The task of medium difficulty appears to be more useful than either the easy or the difficult tasks for discriminating between the two groups. An ensemble which included all three tasks had comparable performance to a classifier trained on the medium difficult task only. More data is now being collected in order to resolve this issue.

The inclusion of additional data may further improve the ensemble accuracy. Age and gender could be relevant parameters since the rate of neck injuries due to whiplash tend to be higher for females than males and also for young adults in the age group 18–24 years [4,5]. Clinical data such as the results of the SF-36 health survey and the Whiplash Disability Questionnaire could also be included.

In the near future, we plan to develop a severity index for neck injuries by a simple extension of the new method. A severity index can be obtained by training a classifier with continuous outputs (e.g. an SVM with probability estimates) on groups of asymptomatic subjects and patients with neck injuries. Prior to computing the severity index, feigned performance will be investigated using the method presented here.

Disclosure of commercial interests

NeckCare is a start-up innovation company developing the Fly method for diagnostic and treatment purposes. The tasks created in the new Fly method are patent pending.

Acknowledgments

The project was partly supported by The Icelandic Center for Research (RANNIS). We thank Magnús K. Gíslason for providing the code to calculate the jerk-index and Ólafur Gíslason for his assistance with the Fly program. The authors wish to thank the anonymous reviewers for their comments, which helped to improve the paper considerably.

References

[1] W.O. Spitzer, M.L. Skovron, L.R. Salmi, J.D. Cassidy, J. Duranceau, S. Suissa, E. Zeiss, Scientific monograph of the quebec task force on whiplash-associated disorders: redefining "whiplash" and its management, Spine 20 (8 Suppl) (1995) 1S–73S.
[2] N. Bogduk, N. Yoganandan, Biomechanics of the cervical spine. Part 3. Minor injuries, Clin. Biomech. 16 (4) (2001) 267–275.
[3] L.W. Holm, L.J. Carroll, J.D. Cassidy, S. Hogg-Johnson, P. Côté, J. Guzman, P. Peloso, M. Nordin, E. Hurwitz, G. van der Velde, E. Carragee, S. Haldeman, The burden and determinants of neck pain in whiplash associated disorders after traffic collisions, Spine 33 (4S) (2008) S52–S59.
[4] K.P. Quinlan, J.L. Annest, B. Myers, G. Ryan, H. Hill, Neck strains and sprains among motor vehicle occupants-United states, Accid. Anal. Prev. 36 (1) (2004) 21–27.
[5] R.J. Brison, L. Hartling, W. Pickett, A prospective study of acceleration-extension injuries following rear-end motor vehicle collisions, J. Musculoskelet. Pain 8 (1) (2000) 97–113.
[6] L. Barnsley, S. Lord, N. Bogduk, Whiplash injury, Pain 58 (3) (1994) 283–307.
[7] C.C. Joslin, S.N. Khan, G.C. Bannister, Long-term disability after neck injury, J. Bone Joint Surg. Br. 86 (2004) 1032–1034.
[8] M.D. Freeman, A.C. Croft, A.M. Rossignol, Whiplash associated disorders: redefining whiplash and its management, Spine 23 (9) (1998) 1043–1049.
[9] L.J. Blincoe, A.G. Seay, E. Zaloshnja, T.R. Miller, E.O. Romano, S. Luchter, R.S. Spicer, The economic impact of motor vehicle crashes, Tech. Rep. DOT HS 809 446, U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, DC, 2002.
[10] Tackling Whiplash: Prevention, Care, Compensation, Association of British Insurers (ABI), London, 2008.
[11] J. Treleaven, G. Jull, M. Sterling, Dizziness and unsteadiness following whiplash injury: characteristic features and relationship with cervical joint position error, J. Rehabil. Med. 35 (1) (2003) 36–43.
[12] E. Kristjansson, L. Hardardottir, M. Asmundardottir, A new clinical test for cervicocephalic kinesthetic sensibility, Arch. Phys. Med. Rehabil. 85 (3) (2004) 490–495.
[13] C. Tjell, A. Tenenbaum, S. Sandström, Smooth pursuit neck torsion test—a specific test for whiplash associated disorders? J. Whiplash Relat. Disord. 1 (2) (2002) 9–24.
[14] J. Treleaven, G. Jull, N. LowChoy, Smooth pursuit neck torsion test in whiplash-associated disorders: relationship to self-reports of neck pain and disability, dizziness and anxiety, J. Rehabil. Med. 37 (4) (2005) 219–223.
[15] T. Prushansky, E. Pevzner, C. Gordon, Z. Dvir, Performance of cervical motion in chronic whiplash patients and healthy subjects, Spine 31 (1) (2001) 37–43.
[16] H. Grip, F. Ohberg, U. Wiklund, Y. Sterner, J. Karlsson, B. Gerdle, Classification of neck movement patterns related to whiplash-associated disorders using neural networks, IEEE Trans. Inform. Technol. Biomed. 7 (4) (2003) 412–418.
[17] Z. Dvir, T. Prushansky, C. Peretz, Maximum versus feigned active cervical motion in healthy patients, Spine 26 (15) (2001) 1680–1688.
[18] P. Sjölander, P. Michaelson, S. Jaric, M. Djupsjöbacka, Sensorimotor disturbances in chronic neck pain—range of motion, peak velocity, smoothness of movement and repositioning acuity, Man Ther. 13 (2) (2008) 122–131.
[19] M. Pinfold, K.R. Niere, E.F. O'Leary, J.L. Hoving, S. Green, R. Buchbinder, Validity and internal consistency of a whiplash-specific disability measure, Spine 29 (3) (2004) 263–268.
[20] S. Kori, R. Miller, D. Todd, Kinesophobia: a new view of chronic pain behaviour, Pain Manage. 3 (1990) 35–43.
[21] J. Ware, K. Snow, M. Kosinski, SF-36 Health Survey: Manual and Interpretation Guide, The Health Institute, Boston, MA, 1993.
[22] E. Kristjansson, P. Dall'alba, G. Jull, Cervicocephalic kinaesthesia: reliability of a new test approach, Physiother. Res. Int. 6 (4) (2001) 224–235.
[23] J. Treleaven, G. Jull, N. LowChoy, The relationship of cervical joint position error to balance and eye movement disturbances in persistent whiplash, Man Ther. 11 (2) (2006) 96–106.
[24] M. Sterling, G. Jull, B. Vicenzino, J. Kenardy, R. Darnell, Development of motor system dysfunction following whiplash injury, Pain 103 (2003) 65–73.
[25] J.-L. Yang, S.-Y. Chen, C.-W. Chang, J.-J. Lin, Quantification of shoulder tightness and associated shoulder kinematics and functional deficits in patients with stiff shoulder, Man Ther. 14 (1) (2009) 81–87.
[26] S. Kitazawa, T. Goto, T. Urushihara, Quantitative evaluation of reaching movements in cats with and without cerebellar lesions using normalized integral of jerk, in: N. Mano, I. Hamada, M. DeLong (Eds.), Role of the Cerebellum and Basal Ganglia in Voluntary Movement, Elsevier, 1993, pp. 11–19.
[27] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA, 2001.
[28] C. Chang, C. Lin, LIBSVM: A Library for Support Vector Machines, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.
[29] L. Xu, K. Crammer, D. Schuurmans, Robust support vector machine training via convex outlier ablation, in: The 21st National Conference on Artificial Intelligence, AAAI, 2006.