Classification of Rheumatoid Joint Inflammation Based on Laser Imaging

Anton Schwaighofer, Volker Tresp, Peter Mayer, Andreas Krause, Ingolf Mesecke-von Rhein
Helmut Rost, Georg Metzger, Gerhard A. Müller, Alexander K. Scheel

Abstract

We describe a novel system for the examination of patients suffering from rheumatoid arthritis. Basis of this system is a laser imaging technique which is sensitive to the optical characteristics of finger joint tissue. From the laser images acquired at baseline and followup, finger joints can automatically be classified according to whether the inflammatory status has improved or worsened. To perform the classification task, various linear and kernel-based systems were implemented and their performances were compared. Based on the results presented in this paper, one can conclude that the laser-based imaging permits a reliable classification of pathological finger joints, making it a sensitive method for detecting arthritic changes.

I. Introduction

Rheumatoid arthritis (RA) is an inflammatory arthropathy with a prevalence rate of about 1-2% of the population. Many patients develop severe disability early in the course of the disease because of progressive joint destruction. Hence, effective therapy with disease-modifying antirheumatic drugs (DMARDs), which retard or even prevent joint destruction, should be initiated as early as possible [1]. Therefore, early diagnosis is essential and would lead to a considerable improvement in the overall prognosis of RA patients, particularly as new effective therapeutic approaches are now widely available.

Objective quantification of joint inflammation is a major challenge not only in the clinical diagnosis of RA but also in the development of new drugs, especially biologicals. Thus, novel methods to rapidly, objectively, and reproducibly assess joint swelling are needed to supplement and objectify clinical assessment [2].

In an *in vitro* study, Prapavat et. al. [3], [4] showed that joint tissues such as bone, cartilage and synovia have distinct absorption and scattering coefficients when analyzed with laser light of a certain wavelength. They showed that there were significant differences in the optical characteristics of normal and pathological tissue.

Based on these studies, a laser-based imaging technique was developed specifically for proximal interphalangeal finger joints [2]. A laser transilluminates the finger joint from above, and a CCD sensor[1] below the finger joint captures the distribution of transmitted and deflected laser light. Fig. 1 shows a schematic drawing of the system, together with two example images taken

[1] Charge Coupled Device, a semiconductor camera element as it can also be found in normal video cameras

from healthy and inflamed joints. The goal of the work presented in this paper is to analyze if the inflammatory status of a finger joint can reliably be classified on the basis of the laser image, in combination with state-of-the-art machine learning techniques. Provided that the accuracy of the overall system is sufficiently high, the imaging technique with the automatic inflammation classification can be combined to realize a novel device that allows an inexpensive and reproducible assessment of inflammatory joint changes.

The paper is organized as follows. In Sec. II we describe the laser imaging system in more detail, as well as the process of data acquisition. Sec. III briefly lists the linear and kernel-based classifiers used in the experiments. In Sec. IV we describe how the methods were evaluated and compared. We present experimental results in Sec. V, and give conclusions and an outlook to future developments in Sec. VI.

II. Laser-based Imaging for Detecting Arthritic Changes

A. Background

Rheumatoid arthritis (RA) is a chronic inflammatory disease that often leads to progressive joint damage. Pathologically, the early stage of RA is characterized by congestion, edema and cellular infiltration of the synovial membrane whereas the typical cartilage and bone erosions usually occur at a later disease stage [5]. Synovial fluid becomes cloudy and opaque of varying degree, depending on the concentration protein, leukocytes, and debris present.

Recent studies have convincingly shown that early treatment and therefore an early diagnosis and sensitive follow-up is mandatory to prevent or at least delay joint destruction [6], [7]. Therapy with DMARDs has a proven ability to retard or prevent joint damage. Since the responsiveness to the treatment with individual DMARDs is patient specific, it is important to objectively quantify the inflammatory changes and therefore the effectiveness of these drugs, especially biologicals, during follow-up. Novel methods to sensitively assess joint swelling and inflammatory soft tissue changes in early disease stages are needed to supplement clinical assessment. These methods should be non-invasive, of low cost, examiner independent and readily available in daily practice [2].

In addition to clinical and laboratory findings, imaging techniques play an important role in the diagnosis and monitoring of RA. Until now, conventional radiography has been the main imaging tool in the assessment of RA. However, this method detects only late joint destructions (erosions) and is not sensitive to early inflammatory changes (synovitis). In contrast, magnetic resonance imaging (MRI) and ultrasound not only provide information about osseous but also about soft tissue changes such as synovitis, effusions and tendon abnormalities. Both imaging methods, MRI and ultrasound, have been evaluated as possible alternatives for the diagnosis of early arthritic changes [8], [9]. However, ultrasound and MRI also have their limitations. Ultrasound is laborious and requires a trained examiner while MRI is both time-consuming and costly.

B. The Laser-based Imaging Technique

Using an *in vitro* joint model, Prapavat et. al. [3], [4] recently performed studies on optical characteristics including absorption and scattering coefficients of synovial fluid and tissue. They found a detectable change between specimen from inflamed RA joints and that of healthy controls.

Based on these observations, a new imaging technique has been developed [2] which allows the *in vivo* transillumination of finger joints with laser light in the near infrared wavelength range. The scattered light distribution is detected by a camera and is used to assess the inflammatory status of the finger joint. A schematic drawing of the imaging system, together with two example images, is shown in Fig. 1. The current prototype is depicted in Fig. 2.

A description of the scattered light distribution by nine numerical features is in turn used to automatically assess the inflammatory status of the finger joint. In this second part, machine learning techniques are used to classify the inflammatory joint changes. A description of the numerical features, derived from the images, will be given in Sec. II-D and Appendix A.

[Figure 1 about here.]

[Figure 2 about here.]

C. Data Acquisition

In the clinical diagnosis of RA it is necessary to assess the synovial joint inflammation during follow-up, yet an objective quantification of inflammatory changes may be difficult. The laser imaging system was designed to support and objectify this diagnosis. In the classification step of the imaging system, the goal is to decide—based on features extracted from the optical data—if there was an improvement of joint activity or if the joints remained unchanged or worsened.

The data we consider in this article is based on a study with 22 patients with rheumatoid arthritis [2]. All of the patients had been or were being treated with disease-modifying antirheumatic drugs. The main steps in the study were as follows:

1) All proximal interphalangeal (PIP) joints II, III, IV and V of the 22 patients underwent a precise clinical examination at baseline and followup.
2) Laser images of all 176 relevant PIP joints were taken at baseline and followup, using the above described imaging technique.
3) Out of the total 176 PIP joints, only 72 showed clear clinical signs of a change in their inflammatory status. The laser images of these 72 PIP joints at baseline and followup, together with the clinically assessed change of inflammatory status, make up the raw data used in this work.

We will now give brief descriptions of steps 1 and 3, further details can be found in [2].

1) Clinical Examination: The joints considered in the present study were proximal interphalangeal (PIP) joints II to V of 22 patients. All PIP joints were examined at baseline and during a followup visit after a mean duration of 42 days.

On both visits, all patients were examined by one investigator (AKS) to assess the clinical degree of inflammation. The diagnostic criteria used were

- The clinical arthritis activity, scored according to the degree of synovitis (swelling, tenderness or warmth) on a scale ranging from 0 (inactive) to 3 (very active).
- The degree of pain for each joint, as indicated by the patients on a scale ranging from 0 (no pain) to 10 (unbearable pain)
- Joint circumference, measured by a conventional measuring tape.

For every joint under consideration, data for each of the 3 above criteria from baseline and followup were compared and rated as improvement (+), worsening (-) or unchanged (0).

2) Selecting a Clear Reference: We will later on train a classification system to best approximate the rheumatologists findings, based solely on information taken from the laser images. It is thus particularly important to have precise clinical references. Therefore, out of the total of 176 PIP joints we selected a subset of 72 joints with clinically clear signs of a change of inflammatory status between baseline and followup. Detailed criteria are given in [2].

3) Reproducibility: To enable an exact inter-individual comparison of images, fingers have to be positioned on the laser imaging device in exactly the same way during baseline and followup. This was accomplished by comparing, at followup, the live image with the image taken at baseline. This procedure achieves a repositioning accuracy of $< 1\,\mathrm{mm}$ [2]. Investigations on a group of healthy controls revealed that an accuracy of 85% can be reached in achieving reproducible laser images [2].

D. Pre-processing

Once the medical study was complete, pre-processing of the acquired data lead to a representation that was suitable for further use in developing a system to assess arthritic changes. Major steps were image pre-processing and an intra-individual comparison of images acquired at baseline and followup.

1) Feature Extraction: In the image pre-processing step, we seek to describe relevant parts of the laser image by a set of numerical features. For our aims, the most relevant part is the light intensity near the center of the finger joint. Thus, to calculate the numerical features, a horizontal line along the vertical center of the laser image is selected. The distribution of light intensity along this line has the shape of a bell curve, similar to a Gaussian density function. For this curve, we consider figures such as the maximum light intensity, the curvature of the light intensity at the maximum, and several others. A detailed description of all features is given in Appendix A.

2) Comparison of Baseline and Followup: The laser images showed high inter-individual variations in optical joint characteristics, resulting from individual differences in joint anatomy [2]. These differences can significantly overlie arthritic effects, making it impossible to tell the inflammatory status of a joint from one single image. Instead, special emphasis was put on the intra-individual comparison of baseline and followup data.

For every joint examined, data from baseline and followup visit were compared and changes in joint activity were rated as improvement, unchanged or deterioration. Similarly, the features derived from the laser images (see the previous section) at baseline and followup were subtracted. In the classification step, the task will be to predict the change of joint activity from the change of laser image features.

3) Final Data: From a machine learning point of view, the data for developing the classification system built on top of the laser imaging device is as follows: We have a set of 72 vectors, each holding the difference of ten feature values between baseline and followup. Nine of these features are derived from the laser images, the tenth feature is the difference of joint circumference between baseline and followup. For each of the 72 vectors we have a label $+1$ (indicating joints where a clinically clear improvement of joint activity had been observed, with a total of 46 joints) or label -1, indicating joints with unchanged or worsened activity, a total of 26 joints.

III. Classification Methods

In this section, we describe the employed linear and kernel-based methods we have been using to classify the inflammatory status of a finger joint from the laser image. Kernel-based methods have shown excellent performance on many challenging classification and regression tasks and thus represent the current state of the art in machine learning. In this section, we will mainly focus on design issues and give references to introductory and in-depth material on the respective methods.

A. Gaussian Process Classification (GPC)

In Gaussian processes [10], [11], a function

$$f(\mathbf{x}) = \sum_{j=1}^{M} v_j k(\mathbf{x}, \mathbf{x}_j, \Theta) \tag{1}$$

is described as a superposition of M kernel functions $k(\mathbf{x}, \mathbf{x}_j, \Theta)$, defined for each of the M training data points $\mathbf{x}_j$, with weight v_j. The kernel functions are parameterized by the vector Θ.

In two-class Gaussian process classification, the logistic transfer function $\sigma(f(\mathbf{x})) = (1 + e^{-f(\mathbf{x})})^{-1}$ is applied to the prediction of a Gaussian process to produce an output which can be interpreted as $\pi(\mathbf{x})$, the probability of the input $\mathbf{x}$ belonging to class 1 [12], [13].

For the experiments we chose the Gaussian kernel function (also called the squared exponential kernel)

$$k(\mathbf{x}, \mathbf{x}_j, \Theta) = \theta_0 \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x}_j)^T \mathrm{diag}(\theta_1, \ldots, \theta_d)^{-1} (\mathbf{x} - \mathbf{x}_j)\right] \tag{2}$$

with input length scales $\theta_1, \ldots, \theta_d$ where d is the dimension of the input space. $\mathrm{diag}(\theta_1, \ldots, \theta_d)$ denotes a diagonal matrix with entries $\theta_1, \ldots, \theta_d$.

For training the Gaussian process classifier (that is, determining the posterior probabilities of the parameters $v_1, \ldots, v_M, \theta_0, \ldots, \theta_d$) we used a full Bayesian approach, implemented with Readford Neal's freely available FBM software [14]. Details of the experimental setup can be found in Appendix B.

B. Gaussian Process Regression (GPR)

In Gaussian process regression [10], [15], [16] we treat the classification problem as a regression problem with target values $\{-1, +1\}$, i.e. we do not apply the logistic transfer function as in the last subsection. Any GP output < 0 is treated as indicating an example from class -1, any output $>= 0$ as an indicator for class 1.[2] The disadvantage is that the GPR prediction cannot be treated as a posterior class probability; the advantage is that the fast and non-iterative training algorithms for GPR can be applied.

The parameters $\Theta = \{\theta_0, \ldots, \theta_d\}$ of the kernel function Eq. (2) were chosen by maximizing the evidence $P(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_M, \Theta)$ with respect to Θ via a scaled conjugate gradient method [10], where $\mathbf{y}$ is a vector containing the M class labels of the training data.

Later on this method will be referred to as "GPR Bayesian"[3]. Results are also given for a simplified covariance function with $\theta_0 = 1$, $\theta_1 = \theta_2 = \ldots = \theta_d = r$, where the common length scale r was chosen by cross-validation (later on referred to as "GPR crossval").

C. Support Vector Machine (SVM)

The SVM is a maximum margin linear classifier [17], [18], [19]. As in Sec. III-B, the SVM classifies a pattern according to the sign of $f(\mathbf{x})$ in Eq. (1). The difference is that the weights $\mathbf{v} = (v_1, \ldots, v_M)^T$ in the SVM minimize the particular cost function

$$\mathbf{v}^T K \mathbf{v} + \sum_{i=1}^{M} C_i (1 - y_i f(\mathbf{x}_i))_+ \tag{3}$$

where $(\cdot)_+$ sets all negative arguments to zero. Here, $y_i \in \{+1, -1\}$ is the class label for training point $\mathbf{x}_i$. $C_i \geq 0$ is a constant that determines the weight of errors on training point $\mathbf{x}_i$, and K is an $M \times M$ matrix containing the amplitudes of the kernel functions at the training data, i.e.

[2] Similarly, neural networks are sometimes trained with sum-of-squares error function on classification problems.

[3] Though in a strict sense, it is not fully Bayesian, since we only use a point estimate for Θ instead of the full probability distribution.

$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j, \Theta)$. The motivation for this cost function comes from statistical learning theory [19]. Many authors have previously obtained excellent classification results by using the SVM. One particular feature of the SVM is the sparsity of the solution vector $\mathbf{v}$, that is, many elements v_i are zero.

In the experiments, we used both an SVM with linear kernel ("SVM linear") and an SVM with a Gaussian kernel ("SVM Gaussian"), equivalent to the Gaussian process kernel Eq. (2), with $\theta_0 = 1$, $\theta_1 = \theta_2 = \ldots = \theta_d = r$. The kernel parameter r was chosen by cross-validation.

In practical applications, for example in the medical domain [20] or in recommender systems [21], it has been noted that the standard formulation of the SVM, as given by Eq. (3) with $C_i = C$ for all i, is susceptible to unbalanced class distributions.[4] A well-known remedy is using a cost function that penalizes errors in the minority class more than errors on examples of the majority class. For SVMs, different cost functions for minority and majority class can be simply introduced in Eq. (3) by using $C_i = C_{\text{maj}}$ for the majority class and $C_i = C_{\text{min}} > C_{\text{maj}}$ for the minority class. Theoretical work on this method, specifically aimed at support vector machines, is presented in [23]. In our experiments, we found empirically that $C_{\text{min}} = 1$ and $C_{\text{maj}} = 0.8$ gives the best balance of sensitivity and specificity.

D. Generalized Linear Model (GLM)

Generalized linear models are well established statistical techniques to solve regression and classification problems [24], [25], [26]. A GLM is built up from a linear model for the input data, together with a link function that relates the linear predictor to the mean of the outcome variables. If we choose the canonical link function for Bernoulli distributions, the output of the linear model $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ is in turn input to the logistic transfer function $\sigma(\cdot)$

$$\sigma(f(\mathbf{x})) = (1 + e^{-f(\mathbf{x})})^{-1} = (1 + e^{-\mathbf{w}^T\mathbf{x}})^{-1} \tag{4}$$

that computes $\pi(\mathbf{x})$, the probability of the input $\mathbf{x}$ belonging to class 1. Training of the GLM was done by iteratively re-weighted least squares (IRLS), implemented with the Netlab toolbox by Chris Bishop and Ian Nabney [27].

[4] Yet this property is also shared by other classification algorithms, such as neural networks [22].

IV. TRAINING AND EVALUATION

One of the challenges in developing the classification system for the laser imaging system is the low number of available training examples. Data was collected through an extensive medical study, but only data from 72 fingers were found to be suitable for further use. As with many medical applications, data acquisition is rather costly and needs to be conducted carefully. Based on the positive results presented in this paper, it is planned to acquire further data in future studies.

A. Training

From the currently available 72 training examples, classifiers need to be trained and evaluated reliably. Part of the standard methodology for small data sets is N-fold cross-validation, where the data are partitioned into N equally sized sets. The system is trained on $N-1$ of those sets and tested on the Nth data set left out, this is repeated N times.

Since we wish to make use of as much training data as possible, $N = 36$ seemed the appropriate choice, giving test sets with two examples in each iteration. For some of the methods model parameters needed to be tuned (for example, choosing SVM kernel width), where again cross-validation is employed. This leads to the following procedure for training and evaluation:

```
Run 36-fold CV
  For Bayesian methods or methods without tunable
      parameters (SVM linear, GPC, GPR Bayesian, GLM):
      Use full training set to train classifier
  For Non-Bayesian methods
      (SVM Gaussian, GPR crossval):
      Run 35-fold CV on the training set, choose
      parameter to minimise CV error
  evaluate the classifier on the 2 example test set
```

This nested loop ensures that in no case any of the test examples is used for training or parameter tuning.

B. ROC Curves

In medical diagnosis, biometrics and other areas, the common means of assessing a classification method is the receiver operating characteristics (ROC) curve. An ROC curve plots sensitivity versus 1-specificity[5] for different thresholds of the classifier output. ROC curves can be compared at a coarse level by calculating the area under the ROC curve by integration. Based on the ROC curve it can be decided how many false positives resp. false negatives one is willing to tolerate, thus helping to tune the classifier threshold to best suit a certain application. A random assignment of classes to data would result in an ROC curve in form of a diagonal line from (0,0) to (1,1) with an ROC area of 0.5.

Acquiring the ROC curve typically requires the classifier output on an independent test set. We instead use the union of all test set outputs in the cross-validation routine. This means that the ROC curve is based on outputs of slightly different models, yet this still seems to be the most suitable solution for such few data. For all classifiers we assess the area of the ROC curve and the cross-validation error rate. Here the above mentioned threshold on the classifier output is chosen such that sensitivity equals specificity.

V. Results

[Table 1 about here.]

[Figure 3 about here.]

Table I lists error rates for all methods listed in Sec. III. Gaussian process regression (GPR Bayesian) with an error rate of $\approx 14\%$ clearly outperforms all other methods, which all achieve comparable error rates in the range of $20 \dots 24\%$. We attribute the good performance of GPR to its inherent feature relevance detection, which is done by adapting the length scales θ_i in the covariance function Eq. (2). A large θ_i means that the i-th feature is essentially ignored.

In an additional experiment we wanted to find out if classification results could be improved by using only a subset of input features[6]. We found that only the performance of the two linear classifiers (GLM and SVM linear) could be improved by input feature selection. Both now achieve an error rate of 16.67%, which is only slightly worse than GPR on the full feature set (see Table I). In absolute numbers, GLM and SVM linear with reduced feature set differ from

[5] Sensitivity = $\frac{\text{true positives}}{\text{true positives}+\text{false negatives}}$
Specificity = $\frac{\text{true negatives}}{\text{true negatives}+\text{false positives}}$

[6] This was done with the input relevance detection algorithm of the neural network tool SENN, a variant of sequential backward elimination where the feature that least affects the neural network output is removed. The feature set was reduced to the three most relevant ones.

GPR Bayesian only by 2 misclassifications on the test set. We can thus not identify a “clear winner” out of this set of three methods, which is also confirmed by statistical hypothesis testing (see the following paragraph).

A. *Significance Tests*

Using a statistical hypothesis test, we compared all classification methods pairwise. A description of the test is given in Sec. C. It turned out the three best methods (GPR Bayesian, and GLM and SVM linear with reduced feature set) perform better than all other methods at a confidence level of 90% or more. Amongst the three best methods, no significant difference could be observed.

B. *Comparison with Baseline Methods*

A further set of experiments was conducted to compare the performance of the methods described in Sec. III with the performance of baseline methods. We chose the nearest neighbour classifier[7] as the baseline method, with the evalutation framework described in Sec. IV. Using the full set of ten features, the nearest neighbour classifier achieved an error rate of 22.2%, and 27.8% with the reduced feature set. Thus only the three best performing classifiers have achieved significant advantages over baseline.

It is also interesting to consider results that use only one single feature. Using only finger circumference in a GPR Bayesian classifier, we achieved an error rate of 22.2%. While finger circumference can thus already provide some initial information on the inflammatory status, the laser image features are essential to achieve a prediction accuracy that is suitable for medical use.

C. *ROC Curves*

For the three best classification methods (GPR Bayesian, and GLM and SVM linear with reduced feature set), we have plotted the receiver operating characteristics (ROC) curve in Fig. 3. According to the ROC curve a sensitivity of $\approx 80\%$ can be achieved with a specificity at around 90%. GPR Bayesian seems to give best results, both in terms of error rate and shape of the ROC curve.

[7] For each test point, assign the label of the training point that has minimum Euclidean distance to the test point.

D. Choice of Methods

To summarize, when the full set of features was used, Bayesian kernel-based classifiers (Gaussian process regression) appear to have advantages over the other approaches due to their inherent input relevance detection. Comparable yet slightly worse results could be achieved by performing feature selection *a priori* and reducing the number of input features to the three most significant ones. In particular, the error rates of linear classifiers (GLM and linear SVM) improved by this feature selection, whereas more complex classifiers did not benefit. The best overall classification method for the laser imaging system seems to be Gaussian process regression (GPR), operating on the full set of features, although—as stated before—the differences in performance to the two linear methods GLM and SVM with reduced feature set are not statistically significant. Still we can draw the important conclusion that a sensitivity of 80% can be reached at a specificity of approximately 90%.

Further developments of the classification step in the laser imaging system will incorporate information from established medical imaging systems such as magnetic resonance imaging (MRI). MRI provides information about soft tissue changes in the finger joint and thus can be used to assess objectively the inflammatory status of the joint. In contrast, in the present study the inflammatory status is assessed by a rheumatologist and the patients subjective degree of pain, thus we may expect a certain degree of label noise in the data on which the classification system has been trained.

VI. CONCLUSIONS

In this paper we have reported results from the analysis of a new system to classify joint inflammation (synovitis in patients with rheumatoid arthritis, RA) based on images captured by a novel laser imaging technique. Out of a set of linear and kernel-based classification methods, Gaussian process regression performed best, followed closely by generalized linear models and the linear support vector machine, the latter two operating on a reduced feature set. The promising results achieved with the best methods showed that a further development of the RA classification system is desirable. We achieved a sensitivity of 80% at a specificity of approximately 90%. Further studies to improve on these results are currently being prepared.

In summary, the new laser imaging device with the RA classification system might allow an early and reliable monitoring of inflammatory processes. Laser imaging is of only limited help for an individual early arthritis diagnosis due to anatomical inter-individual differences of the joint structures. Thus, the new technique may be especially useful for a sensitive followup analysis of joint inflammation, and therefore may provide important information about both the response to medication and for the objective quantification of the effectiveness of antirheumatic medication. Although the current study has only evaluated the laser imaging technique on patients with rheumatoid arthritis, it may as well be suitable to examine arthritic changes of other genesis.

The new laser imaging technique is easy to handle, non-invasive and inexpensive. It therefore has many advantages over conventional imaging techniques and provides information about inflammatory processes in early disease stages. Laser imaging may supplement our imaging armamentarium and help to better assess arthritis patients. However, additional studies with more patients and a comparison to other, established imaging techniques have to be performed before the overall usefulness of this new technique can conclusively be evaluated.

A. Future Work

Further medical and technical studies on the presented laser-based imaging technique are currently being prepared. In particular, we plan to carefully compare the laser technique with laboratory findings and other established imaging techniques (MRI and ultrasound) which have shown a high sensitivity in detecting early arthritic changes. Using this information, it is hoped that sensitivity and specificity of the RA classification system can significantly be enhanced. Ultimately, this should lead to an accurate and fully automatic classification of the inflammatory joint status based on the laser imaging technique.

Acknowledgments

Anton Schwaighofer gratefully acknowledges support through an Ernst-von-Siemens scholarship. The authors thank the two anonymous reviewers for their helpful suggestions to improve the presentation of the paper. Part of this work has been implemented with Radford Neal's FBM software, and with the Netlab toolbox by Ian Nabney and Chris Bishop. Thanks go to all of them for making their software available to the public.

APPENDIX

A. Computing Features from Laser Images

Starting with the laser image, as captured by the CCD camera, smoothing and resizing are the first pre-processing steps.

- Crop the original image to a size of 55×415 pixels. Here, only the lines near the vertical center of the image are retained, that is, the lines around the center of the finger joint.
- Partition the image into disjoint patches of size 5×5 pixels and average the pixel values over each patch. This leads to a matrix M of size 11×83, containing a smoothed and resized version of the original laser image that concentrates on the center point of the finger joint.

To describe the optical characteristics of the joint, the following nine numerical features are computed from matrix M:

1) Average light intensity along line 6, that is, $\frac{1}{83}\sum_{j=1}^{83} M_{6,j}$
2) Maximum light intensity $\max_{i,j}\{M_{i,j}\}$
3) Maximum of average light intensities along each line $\max_i\{\frac{1}{83}\sum_{j=1}^{83} M_{i,j}\}$.
4) Let a be the line containing this maximum average light intensity, $a = \arg\max_i\{\sum_{j=1}^{83} M_{i,j}\}$. Let $b = \arg\max_j\{M_{a,j}\}$ be the point of maximum light intensity in line a. Feature 4 is computed as the average light intensity around the maximum $\frac{1}{7}\sum_{j=b-3}^{b+3} M_{a,j}$.
5) The light intensity along line a is bell shaped. We thus fit a Gaussian density function to the light intensity in line a, by approximating $\log M_{a,\cdot}$ by a polynomial of degree two. Features 5 through 7 are the coefficients of this polynomial.
6) Curvature of the light intensity curve around its maximum. We use the simple heuristic[8] $M_{a,b} - M_{a,b-5}$ for the curvature left of the maximum (feature 8) and $M_{a,b} - M_{a,b+5}$ for the curvature right of the maximum (feature 9).

B. Setup for Neal's FBM Software

We used Radford Neal's FBM software (availabe from `http://www.cs.toronto.edu/~radford/`) for implementing Gaussian process classification. As a prior distribution for kernel parameter θ_0 we chose a Gamma distribution. $\theta_1 \ldots \theta_d$ are samples of a hierarchical Gamma

[8] We might of course compute the curvature based on the Gaussian density approximation, yet this would duplicate errors caused by poorly fitting polynomials.

distribution. In FBM syntax, the prior is `0.05:0.5 x0.2:0.5:1`. Sampling from the posterior distribution was done by persistent hybrid Monte Carlo, following the example of a 3-class problem in [14].

C. Performance Comparison

In order to compare the performance of two given classification methods, one usually employs statistical tests. An overview of tests commonly used in the machine learning community can be found in [28]. When using such tests to compare the classification methods for data from the laser imaging system, we noticed two points: For some tests (such as the 5×2 paired t test proposed by Dietterich) we need to further subdivide the data, which is unacceptable for our few training data. Furthermore, a test for our needs should exploit the fact that we perform a pairwise comparison in the cross-validation procedure, where two classification methods make their predictions on a test point.

We thus propose a simple hypothesis test that allows a performance comparison of two given classification methods A and B. A similar test has been used in [29] to compare text categorization methods.

We start out by looking at the test points one by one, computing the predictions of both methods A and B on the test point, and checking which of the methods predicts correctly. Basis of the proposed test is a matrix of counts, similar to the data used by McNemar's test (see [30], chapter 8)

Number of points which	A classifies correctly	A misclassifies
B classifies correctly	a	b
B misclassifies	c	d

where $a + b + c + d = N$ gives the total number of test points. Similarly to McNemar's test, we assume that both counts a and d (the number of examples that are correctly classified resp. misclassified by both methods) do not contribute to the performance difference of methods A and B.

We assume the remaining counts b and c to be the sufficient statistics of a Binomial random variable, where the parameter θ is the proportion of cases where method A performs better than method B.

The null hypothesis H_0 is that the parameter $\theta = 0.5$, that is, both methods A and B have the same performance. Hypothesis H_1 is that $\theta > 0.5$. The test statistics under the null hypothesis is the Binomial distribution $\mathrm{Bi}(i|b+c,\theta)$ with parameter $\theta = 0.5$. We reject the null hypothesis if the probability of observing a count $k \geq c$ under the null hypothesis

$$P(k \geq c) = \sum_{i=c}^{b+c} \mathrm{Bi}(i|b+c, \theta = 0.5) \tag{5}$$

is sufficiently small.

References

[1] T. W. Huizinga, K. P. Machold, F. C. Breedveld, P. E. Lipsky, and J. S. Smolen. "Criteria for early rheumatoid arthritis: From bayes' law revisited to new thoughts on pathogenesis." *Arthritis and Rheumatism*, vol. 46, no. 5, pp. 1155–1159, May 2002.

[2] A. K. Scheel, A. Krause, I. Mesecke-von Rheinbaben, G. Metzger, H. Rost, V. Tresp, P. Mayer, M. Reuss-Borst, and G. A. Müller. "Assessment of proximal finger joint inflammation in patients with rheumatoid arthritis, using a novel laser-based imaging technique." *Arthritis and Rheumatism*, vol. 46, no. 5, pp. 1177–1184, 2002.

[3] V. Prapavat, W. Runge, A. Krause, J. Beuthan, and G. A. Müller. "Bestimmung von gewebeoptischen Eigenschaften eines Gelenksystems im Frühstadium der rheumatoiden Arthritis (in vitro)." *Minimal Invasive Medizin*, vol. 8, pp. 7–16, 1997.

[4] V. Prapavat, W. Runge, J. Mans, A. Krause, J. Beuthan, and G. A. Müller. "The development of a finger joint phantom for the optical simulation of early inflammatory rheumatic changes." *Biomed Tech (Berl)*, vol. 42, no. 11, pp. 319–326, 1997.

[5] D. McGonagle, W. Gibbon, P. O'Connor, D. Blythe, R. Wakefield, M. Green, D. Veale, and P. Emery. "A preliminary study of ultrasound aspiration of bone erosion in early rheumatoid arthritis." *Rheumatology*, vol. 38, pp. 329–331, 1999.

[6] J. Kim and M. Weisman. "When does rheumatoid arthritis begin and why do we need to know?" *Arthritis and Rheumatism*, vol. 43, pp. 473–482, 2000.

[7] P. Conaghan and P. Brooks. "Disease-modifying antirheumatic drugs, including methotrexate, gold, antimalarials, and d-penicillamine." *Curr Opin Rheumatol*, vol. 7, pp. 167–73, 1995.

[8] P. Scutellari and C. Orzincolo. "Rheumatoid arthritis: Sequences." *European Journal of Radiology*, vol. 27, no. 31–38, 1998.

[9] M. Backhaus, T. Kamradt, D. Sandrock, D. Loreck, J. Fritz, K. Wolf, H. Raber, B. Hamm, G. Burmester, and M. Bollow. "Arthritis of the finger joints." *Arthritis and Rheumatism*, vol. 42, pp. 1232–1245, 1999.

[10] D. J. MacKay. "Introduction to Gaussian processes." In C. M. Bishop, ed., "Neural Networks and Machine Learning," vol. 168 of *NATO Asi Series. Series F, Computer and Systems Sciences*. Springer Verlag, 1998.

[11] C. K. Williams. "Computation with infinite neural networks." *Neural Computation*, vol. 10, no. 5, pp. 1203–1216, 1998.

[12] C. K. Williams and D. Barber. "Bayesian classification with gaussian processes." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1342–1351, 1998.

[13] D. Barber and C. K. Williams. "Gaussian processes for Bayesian classification via hybrid Monte Carlo." In M. C. Mozer, M. I. Jordan, and T. Petsche, eds., "Advances in Neural Information Processing Systems 9," MIT Press, 1997.

[14] R. M. Neal. "Monte carlo implementation of gaussian process models for bayesian regression and classification." Technical Report 9702, Department of Statistics, University of Toronto, 1997.

[15] C. E. Rasmussen. *Evaluation of Gaussian Processes and other methods for non-linear regression*. Ph.D. thesis, University of Toronto, 1996.

[16] C. K. Williams and C. E. Rasmussen. “Gaussian processes for regression.” In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, eds., “Advances in Neural Information Processing Systems 8,” MIT Press, 1996.

[17] C. J. Burges. “A tutorial on support vector machines for pattern recognition.” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, June 1998.

[18] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

[19] V. N. Vapnik. *The nature of statistical learning theory*. Springer Verlag, 1995.

[20] K. Veropoulos, C. Campbell, and N. Cristianini. “Controlling the sensitivity of support vector machines.” In “Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99), SVM Workshop,” 1999.

[21] T. Zhang. “On the dual formulation of regularized linear systems with convex risks.” *Machine Learning*, vol. 46, no. 1, pp. 91–129, 2002.

[22] D. Lowe and A. R. Webb. “Optimized feature extraction and the bayes decision in feed-forward classifier networks.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 355–364, April 1991.

[23] Y. Lin, Y. Lee, and G. Wahba. “Support vector machines for classification in nonstandard situations.” Technical Report 1016, Department of Statistics, University of Wisconsin, Madison, WI, USA, March 2000.

[24] L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Verlag, second edition, 2001.

[25] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, second edition, 1989.

[26] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman & Hall, 1995. First CRC Press reprint 2000.

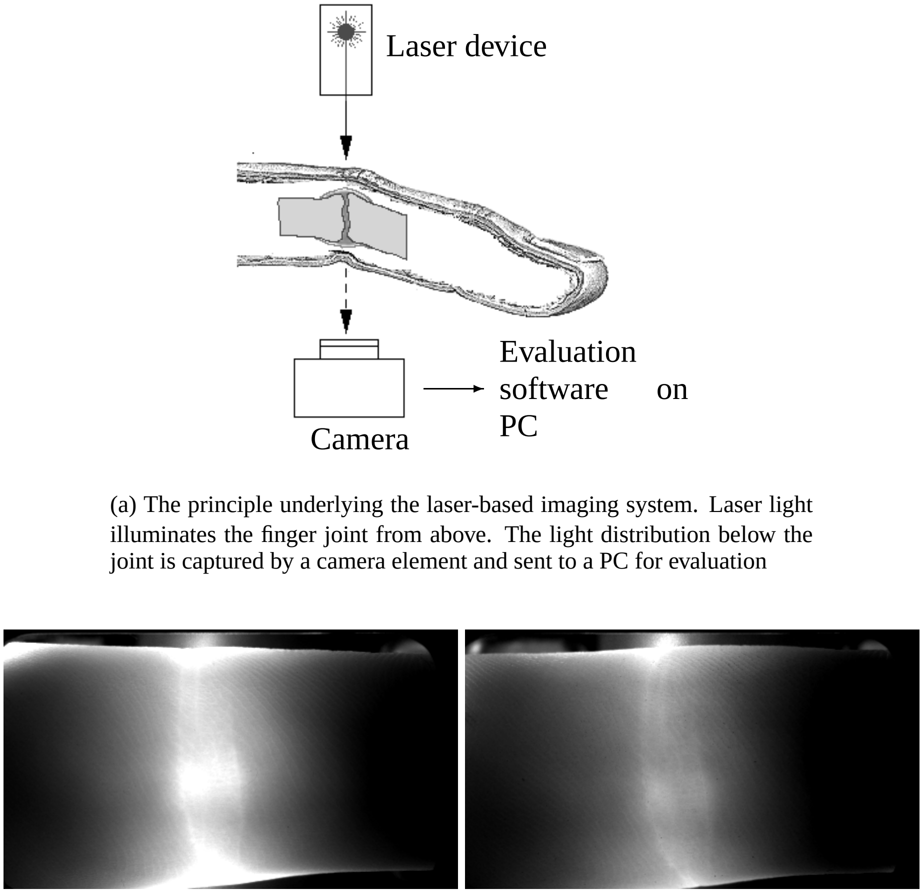
[27] I. T. Nabney. *Netlab. Algorithms for Pattern Recognition*. Springer Verlag, 2002.

[28] T. G. Dietterich. “Approximate statistical tests for comparing supervised classification learning algorithms.” *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[29] Y. Yang and X. Liu. “A re-examination of text categorization methods.” In “Proceedings of ACM SIGIR 1999,” ACM Press, 1999.

[30] J. L. Fleiss. *Statistical Methods for Rates and Proportions*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, second edition, 1981.

List of Figures



(a) The principle underlying the laser-based imaging system. Laser light illuminates the finger joint from above. The light distribution below the joint is captured by a camera element and sent to a PC for evaluation

(b) Laser image of a proximal interphalangeal (PIP) joint of a healthy control

(c) Laser image of an actively inflamed PIP joint

Fig. 1. Schematic drawing of the laser-based imaging method and two example images, as they are captured by the camera element. On both images, the palm is on the left, the finger tip on the right side. For actively inflamed joints, the extended area in the middle of the image typically appears darker due to changed optical characteristics of the joint

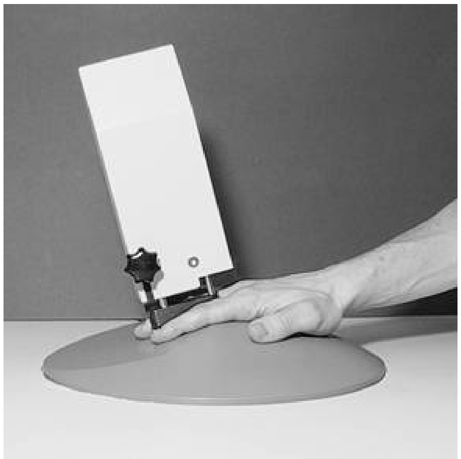
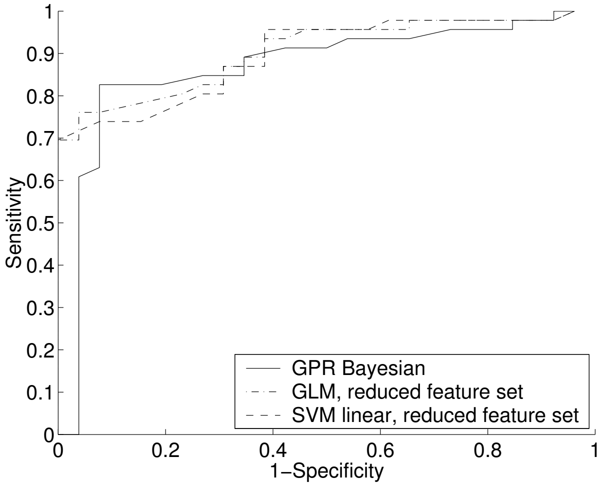


Fig. 2. The prototype laser imaging system. A laser light source (in the upper part of the apparatus) transilluminates the patient's proximal interphalangeal joint. The finger is placed in a specially designed holder that eases a reproducable positioning



Method	ROC area
GPR Bayesian	0.84
SVM linear, reduced feature set	0.86
GLM, reduced feature set	0.86

Fig. 3. ROC curves of the best classification methods, both on the full data set and on a reduced data set where *a priori* feature selection was used to retain only the three most relevant features

LIST OF TABLES

Method	Error rate
GLM	20.83%
GLM, reduced feature set	16.67%
GPR Bayesian	$\mathbf{13.89}\%$
GPR crossval	22.22%
GPC	23.61%
SVM linear	22.22%
SVM linear, reduced feature set	16.67%
SVM Gaussian	20.83%

TABLE I

ERROR RATES OF DIFFERENT CLASSIFICATION METHODS ON THE RHEUMATOID ARTHRITIS PREDICTION PROBLEM. ALL ERROR RATES HAVE BEEN COMPUTED BY 36-FOLD CROSS-VALIDATION, WITH THE CLASSIFIER THRESHOLD SET SUCH THAT SENSITIVITY EQUALS SPECIFICITY. "REDUCED FEATURE SET" INDICATES EXPERIMENTS WHERE *a priori* FEATURE SELECTION HAS BEEN USED