Journal of Biomedical Optics

SPIEDigitalLibrary.org/jbo

Computer-aided diagnosis of rheumatoid arthritis with optical tomography, Part 2: image classification

Ludguier D. Montejo
Jingfei Jia
Hyun K. Kim
Uwe J. Netz
Sabine Blaschke
Gerhard A. Müller
Andreas H. Hielscher

Computer-aided diagnosis of rheumatoid arthritis with optical tomography, Part 2: image classification

Ludguier D. Montejo,[a] Jingfei Jia,[a] Hyun K. Kim,[b] Uwe J. Netz,[c,d] Sabine Blaschke,[e] Gerhard A. Müller,[e] and Andreas H. Hielscher[a,b,f]
[a]Columbia University, Department of Biomedical Engineering, New York, New York 10025
[b]Columbia University Medical Center, Department of Radiology, New York, New York 10032
[c]Laser- und Medizin-Technologie GmbH Berlin, Berlin, Dahlem 14195, Germany
[d]Charité-Universitätsmedizin Berlin, Department of Medical Physics and Laser Medicine, Berlin 10117, Germany
[e]University Medical Center Göttingen, Department of Nephrology and Rheumatology, Göttingen 37075, Germany
[f]Columbia University, Department of Electrical Engineering, New York, New York 10025

Abstract. This is the second part of a two-part paper on the application of computer-aided diagnosis to diffuse optical tomography (DOT) for diagnosing rheumatoid arthritis (RA). A comprehensive analysis of techniques for the classification of DOT images of proximal interphalangeal joints of subjects with and without RA is presented. A method for extracting heuristic features from DOT images was presented in Part 1. The ability of five classification algorithms to accurately label each DOT image as belonging to a subject with or without RA is analyzed here. The algorithms of interest are the *k*-nearest-neighbors, linear and quadratic discriminant analysis, self-organizing maps, and support vector machines (SVM). With a polynomial SVM classifier, we achieve 100.0% sensitivity and 97.8% specificity. Lower bounds for these results (at 95.0% confidence level) are 96.4% and 93.8%, respectively. Image features most predictive of RA are from the spatial variation of optical properties and the absolute range in feature values. The optimal classifiers are low-dimensional combinations (<7 features). These results underscore the high potential for DOT to become a clinically useful diagnostic tool and warrant larger prospective clinical trials to conclusively demonstrate the ultimate clinical utility of this approach. [DOI: 10.1117/1.JBO.18.7.076002]

Keywords: optical tomography; rheumatoid arthritis; computer-aided diagnosis; image classification; light propagation in tissue; medical imaging.

Paper 12648BPRR received Feb. 7, 2013; revised manuscript received May 28, 2013; accepted for publication May 30, 2013; published online Jul. 15, 2013.

1 Introduction

A general framework for the application of computer-aided diagnosis (CAD) to the field of diffuse optical tomography (DOT) is presented in this two-part paper. We apply the framework to the diagnosis of rheumatoid arthritis (RA) from frequency-domain DOT (FD-DOT) images of human proximal interphalangeal (PIP) joints. The data set is from a recent clinical study and consists of 219 DOT images of PIP joints (99 joints from 33 subjects with RA and 120 joints from 20 healthy subjects).[1] Absorption (μ_a) and scattering (μ_s') images are available for each joint.

In Part 1,[2] we presented a framework for processing DOT images and extracting heuristic features. The classification strength of each feature is evaluated with Kruskal–Wallis analysis of variance, Dunn's test, and receiver operating characteristic (ROC) curve analysis. Three important observations are made. First, we observe that features of subjects with RA differ from features of healthy subjects ($p < 0.05$). This implies that physiological differences between subjects with RA and healthy subjects can be captured by DOT images. Our second major finding pertains to subjects with RA who do not exhibit effusions, synovitis, or erosion on magnetic resonance imaging (MRI) and ultrasound (US) scans. The DOT images of these subjects are statistically the same as the images of subjects with RA who do exhibit effusions, synovitis, or erosion, suggesting that DOT can potentially detect the onset of RA in these joints before MRI and US could show macroscopic anatomical changes. Our third major finding shows that features from μ_s' images allow for more accurate classification of each joint as affected or not affected by RA compared to features from μ_a images.

In this part (Part 2), we present a general framework for classifying DOT images of PIP joints as affected or not affected by RA using machine learning techniques. This approach allows for the use of multiple features in the analysis, thus going beyond the ROC curve analysis of Part 1, which was limited to the evaluation of one feature at a time. Classification of each PIP joint is performed with the best 30 features from Part 1 and five distinct classification algorithms—linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), *k*-nearest-neighbors (KNN), self-organizing maps (SOMs), and support vector machines (SVMs). We report the performance of each algorithm in terms of sensitivity (Se) and specificity (Sp).

The work in this paper goes beyond previously published material[1,3,4] in several ways. First, unlike in previous studies, we combine μ_a and μ_s' features in the classification analysis. Second, we are substantially increasing the number of features considered from 4 to 594. In addition to basic features, such as smallest or largest μ_a and μ_s' values, we now consider more advanced features, such as Fourier coefficients of two-dimensional (2-D) and

Address all correspondence to: Ludguier D. Montejo and Andreas H. Hielscher, Columbia University, Department of Biomedical Engineering, 500 West 120th Street, ET 351 Mudd Building, MC8904, New York, New York 10027. Ludguier D. Montejo, Tel: +212-854-2320; Fax: +212-854-8725; E-mail: ldm2106@columbia.edu; Andreas H. Hielscher, Tel: 212-854-5020; Fax: 212-854-8725; E-mail: ahh2004@columbia.edu

0091-3286/2013/$25.00

3-D spatial distribution of optical properties. Third, we compare the performance of five different classification algorithms to determine which scheme is most suitable for DOT imaging data. Fourth, we employ a feature-selection algorithm to determine the subset of image features that achieves highest Se and Sp. This step is essential given the large number of permutations possible when 594 features are considered together with five different classification algorithms. Here we employ an evolution strategy that identifies five to ten optimal features out of 594. Finally, we use intracluster correlation coefficients (ICCs) to compute the effective sample size (ESS) of each data group, which is used to adjust the Se and Sp values. This approach helps to account for bias that may arise as a result of treating each imaged finger as an independent sample and allows us to compute confidence intervals (CIs) for Se and Sp; a necessary treatment as our data consists of multiple fingers per subject and these images may not be statistically independent.

In the remainder of this paper we address the general format of multidimensional classification by presenting details on the five classification methods of interest. We review the mathematical and theoretical foundations of the feature-selection algorithm and present results from classification of DOT images of PIP joints as affected or not affected with RA. All classification results are validated through extensive cross-validation. The paper concludes with a discussion on the potential impact of CAD in the diagnosis of RA with DOT.

2 Methods

Reducing the number of features from the original 594 features to a smaller subset is motivated by various factors. First, features that poorly differentiate between the two diagnostic groups in ROC analysis are unlikely to offer substantial contributions in multidimensional analysis and should be discarded. Second, to improve the generalizability of classification results, it is generally desirable for the ratio between the number of features (l) and data samples (N) to be small (generally $l/N = 0.1$ to 0.2 is acceptable).[5] Third, the complexity of classification algorithms can increase with l, sometimes exponentially. Ultimately, the final number of features should strike a balance between these motivating factors and the desire to include as many important features as possible.

In this work we select 30 features from the original 594 features, as it is a good compromise, resulting in $l/N = 0.13$. These 30 features are selected based on Youden indices ($Y = \mathrm{Se} + \mathrm{Sp} - 1$) from ROC curve analysis. The features with the 30 largest Y values are selected; of these, 4 are from μ_a and 26 are from μ_s' images. Throughout this paper we refer to features by the numbering in Table 1.

The original clinical data are divided into five groups (labels A, B, C, D, or E) of subjects with RA (segmented based on symptoms) and one group without RA (label H) (refer to Part 1 for the details of the clinical trial data). We established in Part 1 that the subgroups of affected subjects are not statistically different from each other based on the features we currently consider. However, each of the affected subgroups is statistically different from the cohort of healthy subjects. As a result, in the following analysis, we treat all subjects with RA as one group (affected with RA) and attempt to accurately classify an unseen data point as affected or not affected with RA.

The nomenclature used in Table 1 was established in Part 1 and follows the pattern *Feature #:Projection Name:Optical Parameter*. Shorthand notation is necessary because of the large number of features we considered in Part 1. For example, feature 19 is denoted by F03:GT:s, which translates to the mean (F03 or feature number 3) of the geometrically dominant transverse (GT) slice of the reduced scattering (s) reconstruction. Indices "a" and "s" denote μ_a and μ_s' derived features, respectively. Feature numbers F01 to F05 are basic statistical features,

Table 1 List of features with top Youden indices from ROC curve analysis (presented in Part 1).

#	Descriptive Notation	Analysis Type	Optical Variable
1	F05:UV:a	Basic	Absorption
2	F05:SV:a	Basic	Absorption
3	F05:ST:a	Basic	Absorption
4	F05:GT:a	Basic	Absorption
5	F02:UV:s	Basic	Scattering
6	F04:UV:s	Basic	Scattering
7	F04:SV:s	Basic	Scattering
8	F05:SV:s	Basic	Scattering
9	F04:SS:s	Basic	Scattering
10	F05:SC:s	Basic	Scattering
11	F05:ST:s	Basic	Scattering
12	F03:VS:s	Basic	Scattering
13	F01:VC:s	Basic	Scattering
14	F03:VC:s	Basic	Scattering
15	F04:VC:s	Basic	Scattering
16	F01:VT:s	Basic	Scattering
17	F03:VT:s	Basic	Scattering
18	F04:VT:s	Basic	Scattering
19	F03:GT:s	Basic	Scattering
20	F04:GT:s	Basic	Scattering
21	F013:SC:s	2-D-FFT	Scattering
22	F014:VS:s	2-D-FFT	Scattering
23	F015:VS:s	2-D-FFT	Scattering
24	F014:VC:s	2-D-FFT	Scattering
25	F015:VC:s	2-D-FFT	Scattering
26	F016:VC:s	2-D-FFT	Scattering
27	F014:VT:s	2-D-FFT	Scattering
28	F015:VT:s	2-D-FFT	Scattering
29	F018:VT:s	2-D-FFT	Scattering
30	F019:GS:s	2-D-FFT	Scattering

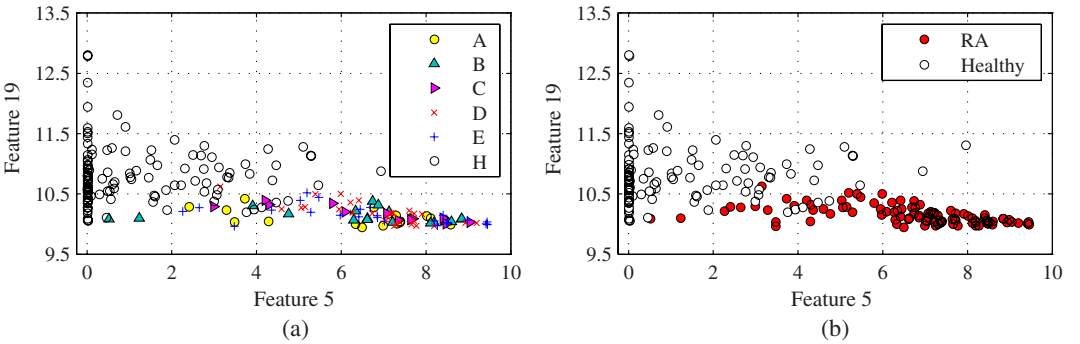

Fig. 1 Distribution of sample features. In (a) the six distinct diagnosis groups (A, B, C, D, E, H) are identified, while in (b) the five cohorts diagnosed with RA (A, B, C, D, E) are grouped into one group (RA).

F06 to F12 are Gaussian mixture model features, and labeling of FFT features starts from F13 for the first FFT coefficient. For 2-D images, the last FFT coefficient is F26, while for 3-D images it is F76.

As an example and to visualize the reduction of subgroups A to E into a single group, consider features 5 and 19 from Table 1. The allocation of data to subgroups A to E and H is presented in Fig. 1(a), where all six groups are visualized. The same data are presented in Fig. 1(b); however, in this plot, a single group (label RA) replaces subgroups A to E, resulting in only two groups of data. For the purpose of consistency, we use features 5 and 19 throughout this text when it is necessary to show 2-D plots.

In this work we study the classification performance of various multidimensional combinations of these 30 features, starting with 2-D combinations. In Secs. 2.1 to 2.4 we briefly review the five classification algorithms. The cross-validation methodology is presented in Sec. 2.5. The feature-selection algorithm, which we use to find optimal feature combination, is presented in Sec. 2.6.

2.1 *Nearest-Neighbor Classification*

The KNN algorithm is among the most basic classification algorithms because it does not require much, if any, prior knowledge of the distribution of the training or testing data. Each unseen feature vector, x, is classified according to the density of affected or healthy data points within a spatial sphere of radius r (covering k neighboring data points and two distinct data classes M).[5,6] The rules governing the assignment of a label to each testing vector x are as follows:

1. From the training data, identify the k nearest neighbors to each vector x using the Euclidean distance measure.
2. Count the number of training data vectors belonging to each class (healthy or RA).
3. Assign test vector x to the class with the maximum number of k_i samples (healthy or RA).

The choice of k affects classification: generally, larger values reduce the effect of noise, but make boundaries between classes less distinct [Fig. 2(a)]. The simplest version of the algorithm is when $k = 1$, known as the nearest-neighbor (NN) rule. In other words, a feature vector x is assigned to the class of its nearest neighbor.

2.2 *Linear and Quadratic Discriminant Analysis*

Classification with discriminant analysis (DA) is a popular parametric method based on Bayesian statistics, primarily used when the training and testing data are believed to be normally distributed. Even in cases where the data are not normally distributed, DA is generally an effective classifier as it gives rise to linear and quadratic hyperplanes that are reasonably accurate at separating the two classes. In general, for classification with DA, the posterior probability $p(\omega_i|\mathrm{x})$ of feature vector x originating from class ω_i is defined by *Bayes theorem*.[5–7]

$$p(\omega_i|\mathrm{x}) = \frac{p(\mathrm{x}|\omega_i)P(\omega_i)}{P(\mathrm{x})}, \tag{1}$$

where $P(\omega_i)$ and $P(\mathrm{x})$ are prior probabilities for class ω_i and feature vector x, respectively.[5] Classification of x is done using the maximum *a posteriori* estimate, $\max_{\omega_i} \hat{p}(\omega_i|\mathrm{x})$, and setting the prior probability for each feature vector equal to $\hat{p}(\omega_i|\mathrm{x}) \propto \hat{p}(\mathrm{x}|\omega_i)P(\omega_i)$.

The prior probabilities for each class are defined to be equal, $P(\omega_i) = P(\omega_j)\ \forall\ i, j$, so classification depends only on the likelihood function estimate, $\hat{p}(\mathrm{x}|\omega_i)$. The likelihood functions are assumed to follow the general multivariate normal distribution,

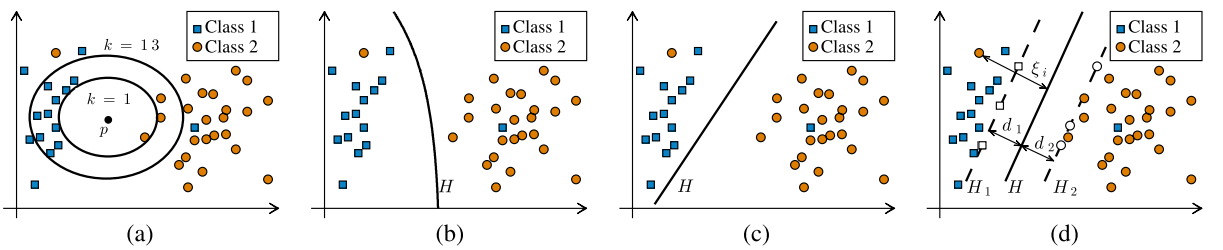

Fig. 2 KNN (a), QDA (b), LDA (c), and SVM (d) boundaries for a nonlinearly separable two-class problem. (a), Examples of $k = 1$ and $k = 13$ for KNN are shown for a new data point (▲). (d), Support vector data points are denoted by white squares and circles.

$$p(\mathrm{x}|\omega_i)=\frac{1}{(2\pi)^{l/2}|\Sigma_i|^{1/2}}\exp\left[-\frac{1}{2}(\mathrm{x}-\vec{\mu}_i)^T\Sigma_i^{-1}(\mathrm{x}-\vec{\mu}_i)\right], \quad i=1,\ldots,M, \tag{2}$$

where l is the dimensionality of x, M the number of classes (here $M=2$), $\vec{\mu}_i$ the mean value, and Σ_i the covariance matrix of class ω_i. For classification purposes, estimates for μ_i and Σ_i are computed from the training data, using the maximum likelihood estimation. Following Bayes, classification is performed using the discriminant function

$$g_i(x)=\ln[p(x|\omega_i)P(\omega_i)], \tag{3}$$

where $P(\omega_i)$ is the *a priori* probability of class ω_i. Here, the assumption that $P(\omega_i)=P(\omega_j)\ \forall\ i,j$ is used. The decision surfaces resulting from the discriminant functions are $g_i(x)-g_j(x)=0$.

Two distinct classification methodologies arise from this theory. In the first and more general case, the covariance matrix for each class is estimated independently, resulting in hyperquadratic decision curves [Fig. 2(b)]. This method is generally referred to as QDA. In the second case, the covariance matrices of the two groups are assumed to be identical ($\Sigma_i=\Sigma_j\ \forall\ i,j$). The resulting decision curves are hyperplanes and the method is called LDA [Fig. 2(c)]. In both cases, the individual features are assumed to be statistically independent, resulting in diagonal covariance matrix estimates.[5]

2.3 *Self-Organizing Maps*

The SOM algorithm is a type of constrained clustering algorithm, where an initial set of randomly distributed and discrete points (neurons) are allowed to self-organize into a smooth manifold. The self-organizing process is achieved through training, a type of competitive learning process, and is typically referred to as vector quantization. After clustering is complete, each neuron is assigned a class label (healthy or RA) based on the number of training vectors from each training class (in this way, similar to the KNN algorithm). Finally, the testing data are input and each feature vector x in the testing set is assigned to its topologically corresponding neuron. The test vector x therefore inherits the class label of its assigned neuron.

Our team of researchers previously presented the theoretical developments necessary for the application of SOMs to CAD of DOT images.[3,8] In this work we use SOMs to perform image classification in multidimensional feature space, varying the number of neurons (n) and learning rate (l) used for pattern learning. The total neurons are varied between 9, 16, and 25. The learning rate is varied between 0.01, 0.1, and 1.0.

2.4 *Support Vector Machines*

SVM is an increasingly popular algorithm in machine learning because of its utility in classification and pattern recognition. We use the SVM algorithm for the general two-class problem, where the classes are not necessarily linearly separable [Fig. 2(d)].[5,6,9] We review the well-established SVM theory for completeness.

The optimal separating line (hyperplane in n-dimensional space), H, is denoted as $H\rightarrow\omega^T x+b=0$. The optimal separating hyperplane is obtained by maximizing the margin, $m=d_1+d_2$, which can be rewritten as a function of the separating plane as $m=2/\|\omega\|$. Then, the primal SVM is a quadratic program where maximizing the margin m corresponds to minimizing $\|\omega\|$ and can be written as

$$\max_{\alpha\geq 0,\beta\geq 0}\min_{\omega,b,\xi}\left\{\frac{1}{2}\|\omega\|^2+C\sum_i\xi_i-\sum_i\alpha_i[y_i(w^Tx_i+w_0)-1+\xi_i]-\sum_i\beta_i\xi_i\right\}. \tag{4}$$

Slack variables ξ_i allow handling of nonseparable data points, C represents a penalty for misclassified data points [Fig. 2(d)], and α and β are Lagrange multipliers. The dual SVM is obtained by finding the gradient of the quadratic program with respect to all variables (ω, b, and ξ_i) and combining the results with the nonnegativity constraints on the Lagrange multipliers α and β. The dual SVM for the nonseparable two-class case can then be written as

$$L_D:\ \max\left[\sum_i\alpha_i-\frac{1}{2}\sum_i\sum_j\alpha_i\alpha_j y_i y_j k(x_i,x_j)\right] \quad \text{subject to } \sum_i\alpha_i y_i=0,\quad \alpha_i\in[0,C], \tag{5}$$

where the kernel $k(x_i,x_j)$ is a function that takes two input vectors and computes a scalar product, $k(x,y)=\phi(x)^T\phi(y)$. This is popularly known as the kernel trick and defines the type of separating hyperplane used for classification. This quadratic program is solved for α_i. Subsequently, the optimal separating hyperplane can be recovered from one of the constraints, $\omega=\sum_i\alpha_i y_i x_i$, and classification is done using the scheme $f(x)=\mathrm{sign}[\sum_i\alpha_i y_i k(x_i,x_j)+b]$. The value of b is found from only those data vectors that are on the margin [i.e., the support vectors, Fig. 2(d)]. The support vectors correspond to vectors with nonzero and not-C α's and by assuming $\xi_i=0$. Then, $b=\tilde{b}$, where b is computed from $y_i(\omega^Tx_i-\tilde{b}_i)-1+\xi_i=0$.

Thus, any data point (or feature vector) x can be classified using the above scheme, where $f(x)=\pm 1$ states whether the data point x is classified into class $+1$ or -1. The kernel trick is very useful as it allows classification with nonlinear hyperplanes. It allows mapping of a d-dimensional input vector from L-space into a higher dimensional H (Hilbert) feature space using the basis function Φ, such that $x_i\rightarrow\Phi(x_i)$. There are many kernel functions that can be used, and in this work we implement the linear, quadratic, polynomial, and radial basis function (RBF) kernels. The linear kernel is trivial. The quadratic, polynomial, and RBF kernels are listed in order below.

$$k(x,y)=(x_1y_1+x_2y_2)^2, \tag{6}$$

$$k(x,y)=(x^Ty+1)^p, \tag{7}$$

$$k(x,y)=\exp\left(-\frac{1}{2\sigma^2}\|x-y\|^2\right). \tag{8}$$

2.5 *Cross-Validation and Quantification of Classification Accuracy*

In general, classification algorithms seek to determine a boundary that best separates two (or more) distinct groups of data. In this work we consider a two-class problem, where the classes are

affected or not affected with RA. There are two steps to this process: (1) training and (2) testing the algorithms. In the training phase, the algorithm determines the decision boundary that best separates the training data into its two classes. In the testing phase, the ability of the algorithm to accurately classify an unseen data point is evaluated (this is the only way to infer how well the classification algorithm will perform on future data).

We remove any bias that may be introduced from treating each imaged finger as an independent sample by employing a modified version of the leave-one-out cross-validation procedure (LOOCV) to train and test. In contrast to the standard LOOCV procedure, where one sample (finger) is used for testing while the remaining samples are used for training, we leave out all samples (fingers) belonging to one single subject (three fingers for subjects with RA and six fingers for subjects without RA). The remaining samples are used for training the algorithm. In the testing phase, each of the testing samples is classified as TP, TN, FP, or FN. This process is repeated for each of the 53 distinct subjects (each repetition is called an iteration).

The overall performance of the algorithm is computed by summing the TP, TN, FP, and FN values computed from each of the 53 LOOCV iterations. From these results, the sensitivity [$\mathrm{Se} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$] and specificity [$\mathrm{Sp} = \mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$] values are computed for each feature combination and classification algorithm. We compute CI for Se and Sp that take into account the effective sample size (as discussed in Part 1). We report the Se and Sp values and their corresponding 95% CI (lower and upper bounds), that is, the interval within which we are confident the true values of Se and Sp lie (with 95% confidence).

2.6 *Selection of Best Features*

In our work to date, our strategy for finding the combination of features that yields the best results (Se and Sp) in the classification of RA has been to evaluate the ability to diagnose with all possible feature combinations.[1,3,4] The same analysis cannot be performed for large numbers of features as the number of combinations is $2^n - 1 - n$, where n is the number of features. In this work $n = 594$, and the number of possible combinations is astronomical.

We overcome this problem in two steps. First, we reduce the dimensionality l of the feature space by considering only those 30 features with the largest Youden index ($\mathrm{Se} + \mathrm{Sp} - 1$). Second, we employ an optimization algorithm to determine a subset of these 30 features that yield optimal or near-optimal classification results. The algorithm does not test all possible feature combinations; instead, it samples only a subset of these combinations while still achieving high classification accuracy.

In particular, we employ an evolution strategy algorithm generally referred to as (1, λ)-single-parent evolution strategy (ES) or greedy feature-selection rule. This is an optimization technique based on ideas of adaptation and evolution.[5,10] The ES algorithm determines which multidimensional set of features achieves optimal (or near-optimal) Se and Sp.

Beginning with a set of parent features (p^k), the algorithm has two steps: (1) mutation and (2) selection. In the mutation step, a total of M mutants, denoted as λ^k, are generated to compete with their parents p^k. The mutation occurs in three steps: (1) M_a new feature combinations are generated by adding a new feature to p^k; (2) M_r new features are generated by replacing an existing feature in p^k with one of the remaining features; (3) M_d new feature combinations are generated by dropping an existing feature in p^k. Thus, the total number of mutants M ($M_a + M_r + M_d$) in generation k are obtained by adding, dropping, or replacing one feature from parent features p^k.

In the selection step, new parents p^{k+1} are selected from the current set of parents p^k and their λ^k mutants. The selection of the new parent features p^{k+1} is made by selecting the feature combination (with dimensionality d) that yields the largest augmented Youden index, defined as

$$Y^*(\mathrm{Se}, \mathrm{Sp}) = \mathrm{Se} + \mathrm{Sp} + \alpha L_{\mathrm{Se}} + \beta L_{\mathrm{Sp}} - \delta d - 1, \quad (9)$$

where L_{Se} and L_{Sp} are the lower bounds of the CI for Se and Sp (i.e., $\mathrm{CI}^{\alpha}_{\mathrm{Se}}$ and $\mathrm{CI}^{\alpha}_{\mathrm{Sp}}$ from Part 1), respectively. The scaling factors α, β, and δ control the contribution of the lower bound values (L_{Se} and L_{Sp}) and dimensionality (d) of the selected feature combination on Y^*, and are all set to 0.001. In this way, feature combinations with higher lower bounds and low dimensionality are preferred.

The mutation and selection operators are applied in a loop and an iteration of the loop is called a generation, denoted by k. We begin the process by specifying the initial set of parents p^0. The sequence of generations continues until we are unable to improve Y^*, that is, $Y^{*k} = Y^{*k+1}$. The algorithm finds the feature combination with the highest Se and Sp values, favoring combinations with higher L_{Se} and L_{Sp} and lower dimensionality d. The process is summarized as follows:

1. A feature combination (may be a single feature) is chosen as the parents p^k of the current generation k.
2. All M possible λ^k mutants of generation k are obtained by adding (M_a), dropping (M_d), or replacing (M_r) one feature from the parent combination p^k ($M = M_a + M_r + M_d$).
3. Of all λ^k mutants and their p^k parents, the combination with the largest Y^* becomes the parents of the next generation (p^{k+1}).
4. Set $k \leftarrow k + 1$ and repeat this process until the objective function Y^* does not improve (i.e., $Y^{*k} = Y^{*k+1}$).

This procedure is formulated as an optimization problem, where the objective function, ϕ^k_λ, is defined as the winning Y^* of generation k [Fig. 3(a)]. In this sense, we seek to maximize ϕ by selecting the feature combination that maximizes Y^* at each generation and define it as

$$\phi^k_\lambda \rightarrow \max_i (Y^*_\lambda)^k = \max_i (\mathrm{Se}^{k,i} + \mathrm{Sp}^{k,i} + \alpha L^{k,i}_{\mathrm{Se}} + \beta L^{k,i}_{\mathrm{Sp}} - \delta d^{k,i} - 1), \quad (10)$$

where the index i refers to the ith feature combination in generation k. $\mathrm{Se}^{k,i}$ and $\mathrm{Sp}^{k,i}$ are the Se and Sp from the LOOCV procedure using the ith feature combination of the kth generation. Similarly, $L^{k,i}_{\mathrm{Se}}$ and $L^{k,i}_{\mathrm{Sp}}$ are the lower bounds of the 95.0% CI for $\mathrm{Se}^{k,i}$ and $\mathrm{Sp}^{k,i}$, while $d^{k,i}$ is the dimensionality of the corresponding feature combination. The algorithm guarantees that $\phi^{k+1}_\lambda > \phi^k_\lambda$ until convergence, where the solution converges to a near-optimal combination (a local maximum) that maximizes ϕ_λ [Fig. 3(b)]. Se and Sp are not required to increase at each iteration.

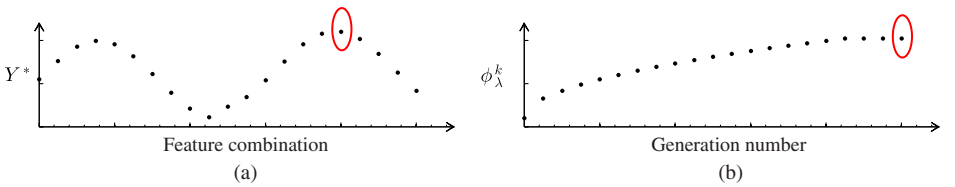

Fig. 3 (a), Sample within-generation values of the augmented Youden index Y^* for all possible feature combinations (mutants). (b), Sample evolution of objective function over multiple generations.

To identify the appropriate parents for the first generation (p^0), we first perform LOOCV using all possible 2-D feature combinations and compute Y^* for each combination. Feature pairs that yield the five highest Y^* values are selected as the first-generation parents p^0. Thus, five distinct optimization runs are executed for each combination of algorithm type (KNN, DA, SOM, SVM) and algorithm parameters (number of neighbors, discriminant type, SVM kernel, etc.), where the initial set of features p^0 is different for each run.

3 Results

We start by showing three examples of typical decision boundaries. Figure 4 shows these boundaries for three different classification algorithms applied to a data set consisting of two features. The two features are the minimum and mean value of μ_s' in images of all healthy subjects (blue dots) and subjects with RA (red dots). The classification algorithms that produce the decision boundaries are LDA, QDA, and SVM. In the case of LDA and QDA, all the data points influence placement of the boundary, while the support vectors determine the SVM boundary only (markers with white face). Here, 34 support vectors (<16% of the data) are identified, 20 from healthy subjects and 14 from subjects with RA. In these three examples, the entire data are used to train (i.e., no cross-validation).

From these plots we determine TN, TP, FN, and FP values. In these example cases we find that LDA achieves 91% Se and

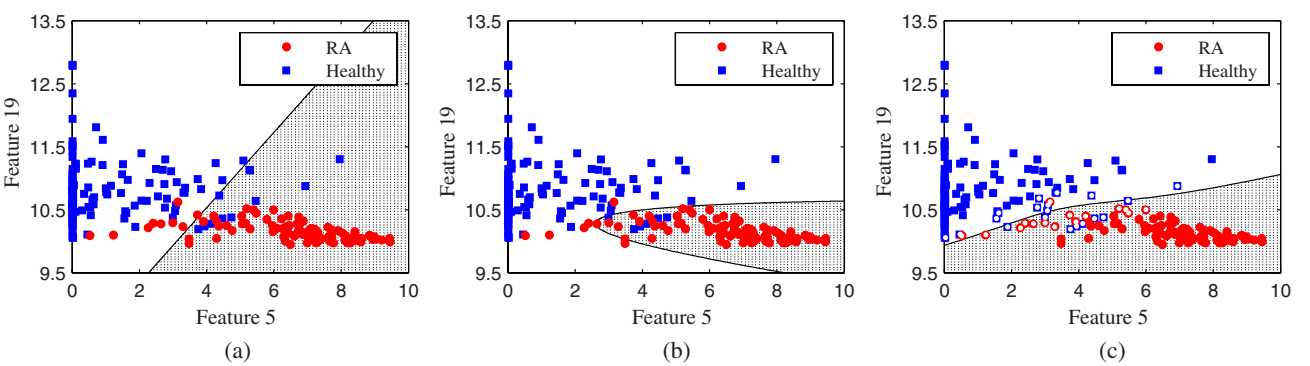

Fig. 4 LDA (a), QDA (b), and SVM (c) decision boundaries separating affected from healthy data using image features 5 and 19. Support vector data points are identified by circles and squares with white face.

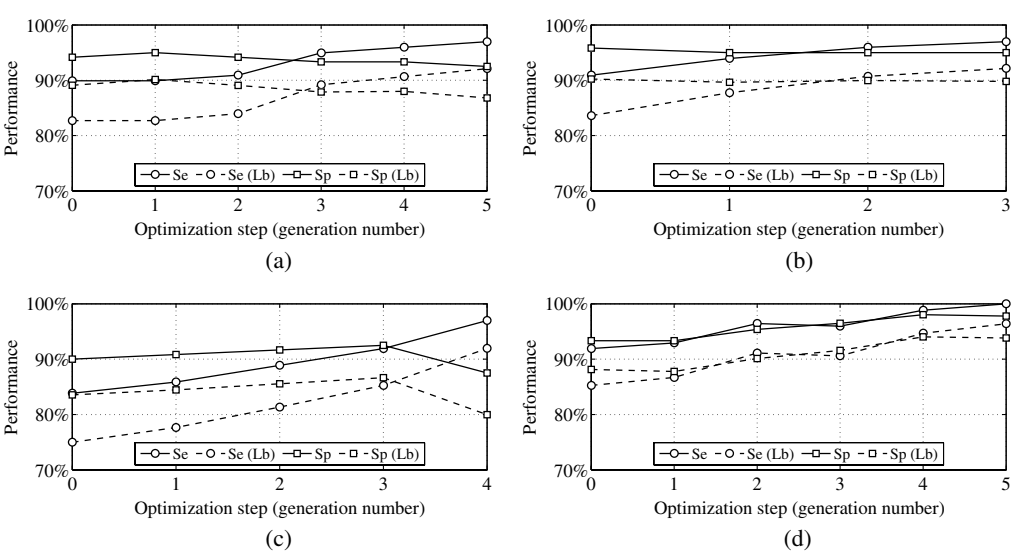

Fig. 5 Sample Se and Sp evolution paths obtained with (a) KNN, (b) DA, (c) SOM, and (d) SVM classifiers. The lower bounds (Lb) for the 95% confidence interval are shown for Se and Sp, respectively. The upper bound in these cases is 100.0%.

91% Sp (FP = 11, FN = 9), QDA results in 93% Se and 95% Sp (FP = 6, FN = 7), and SVM achieves 98% Se and 93% Sp (FP = 9, FN = 2). A similar analysis is performed for all other feature combinations identified by the feature-selection algorithm; however, in those cases, we do perform extensive cross-validation as described in Sec. 2.5.

Results from the evolution algorithm are summarized for each classification method in Secs. 3.1 to 3.4. For brevity, we report only the results obtained with three distinct seed parents used in the optimization algorithms (out of a possible five). The results of nonreported optimization runs are well within the trend presented by the reported cases.

In general, the optimization algorithm always converges to a combination of two to six features in one to six generations and always improves on the initial Se and Sp by 5% to 15% for all classification methods (Fig. 5). The optimal set of features typically includes basic statistical features derived from the raw reconstruction data, basic features from projections of the 3-D data set, as well as coefficients from the Fourier transform of the data (refer to Table 1 for definitions). The algorithm fails to improve on the original set of features only twice (KNN, 1 neighbor, runs 1 and 3).

We report the Se and Sp values to which the algorithm converges, the initial and final feature combinations, and the lower and upper bounds of the 95.0% CI for Se and Sp. The following convention is used: Se (L_{Se},U_{Se}), where L_{Se} and U_{Se} are the lower (L) and upper (U) bounds around the computed Se value within which we have 95.0% confidence that the true value of Se lies. This interval is computed using the ESS and ICC as presented in Part 1 and corrects for correlation between fingers from the same subject. The ESS and ICC values used throughout the analysis are presented in Table 2.

Table 2 ICC and ESS values for each of the 30 features listed in Table 1.

	Affected		Healthy	
Feature number	ICC	ESS	ICC	ESS
1	0.40	54.8	0.31	47.0
2	0.46	51.5	0.33	45.4
3	0.41	54.2	0.20	60.6
4	0.44	52.6	0.31	47.4
5	0.34	58.6	0.18	62.5
6	0.39	55.6	0.14	70.1
7	0.38	56.2	0.14	70.3
8	0.40	55.2	0.17	64.6
9	0.38	56.2	0.12	74.8
10	0.38	56.4	0.18	63.2
11	0.40	54.8	0.16	67.0
12	0.39	55.7	0.14	70.7
13	0.45	52.3	0.14	70.5
14	0.38	56.1	0.14	70.1
15	0.35	58.1	0.14	69.6
16	0.41	54.5	0.16	66.1
17	0.38	56.2	0.15	68.4
18	0.31	61.0	0.16	67.1
19	0.58	46.0	0.28	49.8
20	0.38	55.9	0.14	70.7
21	0.37	56.7	0.18	62.7
22	0.40	55.1	0.12	75.5
23	0.40	55.0	0.12	76.0
24	0.39	55.5	0.12	74.2
25	0.37	57.1	0.15	67.9
26	0.42	53.9	0.17	65.1
27	0.40	54.8	0.11	78.0
28	0.40	54.7	0.11	77.7
29	0.42	53.6	0.20	60.7
30	0.54	47.5	0.29	49.2

3.1 *k-Nearest-Neighbors*

Classification with KNN is performed using various numbers of neighboring points ($k = 1, 3, 5, 7, 11, 15, 21$). For each k, five optimization runs, each with distinct original parent features, are executed, resulting in $7 \times 5 = 35$ distinct executions of the optimization algorithm. The best optimization runs for $k = 1, 5, 11$ are summarized in Table 3. Results from the remaining optimization runs are well within the trend represented by these sample results. The largest Y^* value is achieved using $k = 5$, with Se = 96.0%(90.7%, 100.0%) and Sp = 94.2%(89.0%, 100.0%). All five distinct runs with $k = 5$ converge to the same final set of features, {3,8,15,19}, independent of the initial set of features. Using too few (i.e., $k = 1$) or too many ($k > 11$) neighbors results in lower Se and Sp.

3.2 *Discriminant Analysis*

The three best classification results obtained with LDA and QDA algorithms are shown in Table 4. Classification with QDA is marginally better than classification with LDA. Classification with QDA converges to optimal features {5,15,19} with Se = 97.0%(92.1%, 100.0%) and Sp = 95.0%(89.8%, 100.0%). Classification with LDA converges to features {5,9,15,19} with Se = 97.0%(92.1%, 100.0%) and Sp = 93.3%(87.8%, 100.0%). QDA achieves higher Se and Sp, and converges to a lower-dimensional optimal feature combination compared to LDA. Both methods consistently select the same set of optimal features independent of the initial feature set, although LDA selects one additional feature compared to QDA (feature 9).

Table 3 Classification results with the KNN algorithm using 1, 5, and 11 nearest neighbors.

Neighbors (k)	Sensitivity [% (95% CI)]	Specificity [% (95% CI)]	Initial combination	Final combination	Optimization steps
1	93.9 (88.0, 100.0)	93.3 (88.0, 100.0)	6, 15	6, 15	1
	93.9 (87.9, 100.0)	92.5 (86.8, 100.0)	8, 19	8, 14, 19	2
	93.9 (88.1, 100.0)	91.7 (85.3, 100.0)	5, 19	5, 19	1
5	96.0 (90.7, 100.0)	94.2 (89.0, 100.0)	15, 19	3, 8, 15, 19	4
	96.0 (90.7, 100.0)	94.2 (89.0, 100.0)	6, 18	3, 8, 15, 19	5
	96.0 (90.7, 100.0)	94.2 (89.0, 100.0)	5, 19	3, 8, 15, 19	4
11	97.0 (92.1, 100.0)	92.5 (86.8, 100.0)	3, 10	7, 8, 19, 21	6
	97.0 (92.1, 100.0)	92.5 (86.8, 100.0)	3, 12	7, 8, 19, 21	6
	97.0 (92.1, 100.0)	92.5 (86.8, 100.0)	3, 6	7, 8, 19, 21	5

Table 4 Classification results with the DA algorithm using linear and quadratic functions.

Discriminant type	Sensitivity [% (95% CI)]	Specificity [% (95% CI)]	Initial combination	Final combination	Optimization steps
Linear	97.0 (92.1, 100.0)	93.3 (87.8, 100.0)	5, 12	5, 9, 15, 19	4
	97.0 (92.1, 100.0)	93.3 (87.8, 100.0)	5, 19	5, 9, 15, 19	3
	97.0 (92.1, 100.0)	93.3 (87.8, 100.0)	8, 19	5, 9, 15, 19	4
Quadratic	97.0 (92.1, 100.0)	95.0 (89.8, 100.0)	5, 30	5, 15, 19	4
	97.0 (92.1, 100.0)	95.0 (89.8, 100.0)	5, 6	5, 15, 19	3
	97.0 (92.1, 100.0)	95.0 (89.8, 100.0)	8, 19	5, 15, 19	3

3.3 *Self-Organizing Maps*

For each set of first-generation parents, the optimization algorithm is executed for each possible combination of neurons ($n = 9, 16, 25$) and learning rates ($l = 0.01, 0.1, 1.0$), for a total of $9 \times 5 = 45$ optimization runs. Results from classification with SOMs are summarized for $n = 9, 16$ and $l = 1.0$ in Table 5. While the number of neurons n do have an impact on the classification accuracy, we find that the learning rate l does not make a significant difference; for this reason only results that demonstrate the dependence on n are shown.

Classification with SOMs leads to different optimal features for each run, meaning that the optimal feature combination is dependent on the initial set of features. The best classification results are $\mathrm{Se} = 97.0\%(92.1\%, 100.0\%)$ and $\mathrm{Sp} = 91.7\%(85.7\%, 100.0\%)$, with $n = 9$ and $l = 1.0$. The performance of SOMs with $n = 16$ is similar; however, performance

Table 5 Classification results with the SOM algorithm using neurons $n = 9, 16$, respectively ($l = 1.0$).

Neurons (n)	Sensitivity [% (95% CI)]	Specificity [% (95% CI)]	Initial combination	Final combination	Optimization steps
9	89.9 (82.7, 100.0)	92.5 (86.9, 100.0)	6, 10	10, 11	2
	93.9 (88.1, 100.0)	90.8 (84.7, 100.0)	15, 19	15, 19	1
	97.0 (92.1, 100.0)	91.7 (85.7, 100.0)	19, 24	8, 17, 19	5
16	90.9 (84.0, 100.0)	94.2 (89.1, 100.0)	3, 7	3, 12	3
	94.9 (89.3, 100.0)	89.2 (82.7, 100.0)	6, 14	6, 11	2
	90.9 (84.0, 100.0)	94.2 (89.1, 100.0)	6, 18	6, 14	2

significantly decreases with $n = 25$, suggesting that using 25 neurons results in overfitting the data.

3.4 *Support Vector Machines*

Classification with SVMs is performed using linear, quadratic, polynomial, and RBF kernels. Classification with the polynomial kernel is explored with polynomials of degrees 3, 4, 5, 6, and 7. The RBF kernel is explored by varying σ (0.1, 0.5, 1.0, 2, 5, 10.0). A total of five distinct classification runs (each with distinct features as seeds) are performed for each combination of kernel method and kernel parameter for a total of $5 + 5 + 5 \times 5 + 5 \times 6 = 65$ distinct runs.

Results from classification with linear, quadratic, and polynomial kernels are summarized in Table 6. Linear, quadratic, and low-order polynomial kernels can separate the data well. The classification accuracy of polynomials of higher order ($p \geq 6$) is lower compared to low-order polynomials. This is expected as it is well known that higher-order polynomials can severely overfit the data, resulting in poor cross-validation results. Marginally less accuracy is obtained with the RBF kernel, $\sigma = 3$ providing the best results (omitted for brevity).

The optimal combination always converges to the same set of features, independent of initial features, depending only on the kernel type. The largest Y^* value is achieved using a polynomial of order 3, with Se $= 100\%(96.4\%, 100.0\%)$ and Sp $= 97.8\%(93.8\%, 100.0\%)$, and converges to optimal features {4,5,6,12,15,30}. Feature 5 is selected as an optimal classifier by all kernels, while features 15 and 19 are selected as optimal features by three kernels.

3.5 *Best Feature Selection*

The frequency with which each of the original 30 features appears as an optimal classifier is presented in Fig. 6. Features 8, 15, and 19 are chosen as optimal features 73.0%, 40.0%, and 73.0% of the time across all KNN optimization runs [Fig. 6(a)]. Features 5 (100.0%), 15 (80.0%), and 19 (80.0%) are chosen as optimal features most often across all DA iterations [Fig. 6(b)]. Similarly, features 3 (20.0%), 6 (20.0%), 8 (33.0%), and 19 (33.0%) are chosen as optimal features most often across all SOM iterations [Fig. 6(c)]. Finally, features 5 (88.0%), 6 (60.0%), 19 (48.0%), and 30 (48.0%) are chosen as optimal features most often across all SVM iterations [Fig. 6(d)].

Feature 5 is chosen most often by DA (100.0%) and SVM (88.0%), while feature 19 is chosen most often by KNN (73.0%), and SOM ((33.0%). Feature 19 occurs as an optimal feature at least 20.0% of the time in all classification methods. Feature 15 appears as an optimal feature >20.0% of the time for three classification methods. Features 3, 5, 6, and 8 appear as an optimal feature >20.0% of the time for two classification methods. Finally, features 7, 11, 21, and 30 are chosen at least 20.0% of the time by one classification method.

4 Discussion and Conclusions

In this two-part paper, we present a general framework for the application of CAD techniques to DOT. In Part 1 we focus on feature-extraction methods, while in Part 2, image classification with machine learning techniques takes center stage. As a specific example, the framework is applied to the classification of FD-DOT images of 219 PIP joints ($N = 53$ subjects). The goal is to classify each PIP joint as affected or not affected with RA based on the analysis of single image features (Part 1) and

Table 6 Classification results with the SVM algorithm using linear, quadratic, and polynomial kernels.

SVM kernel	Sensitivity [% (95% CI)]	Specificity [% (95% CI)]	Initial combination	Final combination	Optimization steps
Linear	97.0 (92.1, 100.0)	93.3 (87.8, 100.0)	5, 12	5, 9, 15, 19	4
	97.0 (92.1, 100.0)	93.3 (87.8, 100.0)	5, 19	5, 9, 15, 19	3
	97.0 (92.1, 100.0)	93.3 (87.8, 100.0)	8, 19	5, 9, 15, 19	4
Quadratic	97.0 (92.1, 100.0)	95.0 (89.8, 100.0)	5, 30	5, 15, 19	4
	97.0 (92.1, 100.0)	95.0 (89.8, 100.0)	5, 6	5, 15, 19	3
	97.0 (92.1, 100.0)	95.0 (89.8, 100.0)	8, 19	5, 15, 19	3
Polynomial of order 3	100.0 (96.4, 100.0)	97.8 (93.8, 100.0)	5, 30	4, 5, 6, 12, 15, 30	5
	100.0 (96.4, 100.0)	97.8 (93.8, 100.0)	5, 6	4, 5, 6, 12, 15, 30	5
	100.0 (96.4, 100.0)	97.8 (93.8, 100.0)	6, 27	4, 5, 6, 12, 15, 30	6
Polynomial of order 4	97.0 (91.7, 100.0)	94.2 (88.0, 100.0)	5, 6	5, 30	2
	97.0 (91.7, 100.0)	94.2 (88.0, 100.0)	6, 29	5, 30	3
	97.0 (91.7, 100.0)	94.2 (88.0, 100.0)	8, 19	5, 30	3
Polynomial of order 5	97.0 (92.2, 100.0)	93.3 (87.5, 100.0)	5, 30	5, 19	2
	97.0 (92.2, 100.0)	93.3 (87.5, 100.0)	8, 19	5, 19	2
	97.0 (92.2, 100.0)	93.3 (87.5, 100.0)	5, 19	5, 19	1

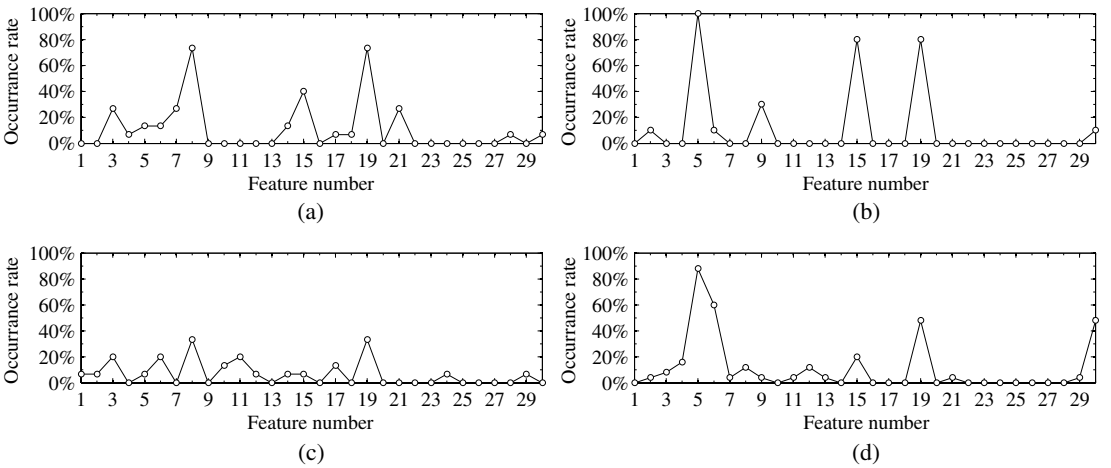

Fig. 6 Frequency with which all features appear as optimal classifiers using (a) KNN, (b) DA, (c) SOM, and (d) SVM.

combinations of multiple image features (Part 2). For image classification with multiple features, we compare the performances of five different algorithms, including KNN, LDA, QDA, SOM, and SVM. Given the large number of possible permutations of the 594 features we extract, it is necessary to implement a feature-selection algorithm to determine the subset of image features that achieves the highest Se and Sp in combination with each of the five classification algorithms. Results are validated through extensive cross-validation, where the ground truth is the clinical diagnosis made according to the American College of Rheumatology (ACR) 2010 criteria.[11]

The procedure for training and testing the algorithms is a two-step process using a modified version of the LOOCV procedure. The first phase uses data from $N-1$ subjects to train the classifier and then retrospective classification is performed on the data from the remaining one subject (testing phase). This procedure is repeated so that data from each subject (three images for subjects with RA and six images for subjects without RA) is left out one time each.

We find that all five classification algorithms achieve clinically relevant sensitivities and specificities >90%. The best combination of sensitivity and specificity is obtained with SVM using a polynomial kernel of degree 3. The optimal features corresponding to these results are {4,5,6,12,15,30}. For this case, the sensitivity is 100% with a 95% CI of (96.4%,100.0%). The specificity is 97.8% with a 95% CI of (93.8%,100.0%).

Features 4 and 5 capture the range and mean μ_a and μ_s' values in the PIP joint, while features 6, 12, 15, and 30 all capture the variation of μ_s' across the joint. Feature 4 corresponds to the ratio of maximum μ_a and minimum μ_a values in the transverse slice across the middle of the PIP joint (F05:GT:a); feature 5 is the minimum value of the μ_s' unstructured reconstruction (F02:UV:s); feature 6 is the variance value of the unstructured reconstruction μ_s' data (F04:UV:s); feature 12 is the mean value of the variation among all μ_s' sagittal slices (F03:VS:s); feature 15 is the variance value of the variation among all μ_s' coronal slices (F04:VC:s); and feature 30 is the absolute value of the sixth coefficient of the 2-D-FFT of the central sagittal slices of μ_s' (F16:VC:s).

Features that most often achieve accurate classification are associated with global absolute values of the absorption and scattering data and their spatial variation near the PIP joint. We see evidence that features that quantify spatial variation across PIP joints are smaller for subjects with RA compared to healthy subjects. This is in concordance with our earlier findings and the images shown in Fig. 2 of Part 2. The synovial fluid as well as the surrounding tissue experience changes in optical properties in subjects with RA.[12,13] The inflammatory process starts in the synovium, leading to changes in the cell and tissue structure. Cell proliferation can be observed and the appearance of the synovial fluid changes from a clear, yellowish substance to a turbid, gray-yellowish substance. The number of leukocytes per mL increases from 100–200 in healthy conditions to 1000–100,000 during stages 1 and 2 of the disease. Leukocytes have a diameter of approximately 7–20 μm and therefore have an effect on the scattering coefficient. Furthermore, the protein content in the synovial fluid approximately triples from 10–20 g/L to 30–60 g/L.[14,15] In addition, neovascularization in the surrounding tissue has been related to synovitis,[16] which leads to an increase in the absorption coefficient.

Overall, the net effect of these changes is an increase in absorption and scattering in the finger joint affected by RA, resulting in an optical profile similar to the physiology surrounding the joint. Thus the spatial variation of optical properties decreases in subjects with RA. Although similar results have been described before, it appears that our CAD algorithms can extract these features with higher accuracy and produce sensitivities and specificities at levels not reported before. All five classification algorithms achieve clinically relevant sensitivities and specificities >90.0%, some even >97.0%. The computed 95.0% CIs for our results offer further validation that our results are robust. These observations warrant further multicenter prospective clinical studies.

One particular focus of these trials should further investigate an intriguing finding in our current study. We observe that joints of subjects with RA but without radiological evidence of effusion, erosion, or synovitis (as determined by MRI or US) are statistically identical to joints of subjects with RA that have detectable effusion, erosion, or synovitis; indeed, these joints are also statistically different from joints of healthy subjects. These joints are clear examples of cases where DOT imaging is sensitive to symptoms associated with RA before they become detectable by MRI or US. A longitudinal study is necessary to determine if these joints eventually evolve to exhibit evidence of effusion, erosion, or synovitis on MRI or US scans so as to meet the ACR criteria for RA. If, indeed, that turns out to be the case, one could prove that DOT can detect the presence of symptoms

associated with RA at an earlier stage than other imaging methods.

In addition to diagnosis of RA, we expect that the framework established in this work (CAD with DOT) can be deployed in the diagnosis and monitoring of other diseases, including breast cancer and peripheral artery disease.

Acknowledgments

The authors thank Julio D. Montejo for his contribution during the early stage of this project. This work was supported in part by a grant from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS-5R01AR046255), which is part of the National Institutes of Health. Furthermore, L.D.M. was partially supported by an NIAMS training grant on Multidisciplinary Engineering Training in Musculoskeletal Research (5 T32 AR059038 02).

References

1. A. Hielscher et al., "Frequency-domain optical tomographic imaging of arthritic finger joints," *IEEE Trans. Med. Imag.* **30**(10), 1725–1736 (2011).
2. L. D. Montejo et al., "Computer-aided diagnosis of rheumatoid arthritis with optical tomography, Part 1: feature extraction," *J. Biomed. Opt.* **52**(6), 066011 (2013)
3. C. D. Klose et al., "Computer-aided interpretation approach for optical tomographic images," *J. Biomed. Opt.* **15**(6), 066020 (2010).
4. C. D. Klose et al., "Multiparameter classifications of optical tomographic images," *J. Biomed. Opt.* **13**(5), 050503 (2008).
5. S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Elsevier/Academic Press, Amsterdam, Boston (2006).
6. C. Bishop, *Pattern Recognition and Machine Learning, Information Science and Statistics*, Springer, New York (2006).
7. G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York (1992).
8. C. D. Klose et al., "Computer-aided classification of rheumatoid arthritis in finger joints using frequency domain optical tomography," *Proc. SPIE* **7169**, 716915 (2009).
9. C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.* **20**(3), 273–297 (1995).
10. T. Bäck, *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*, Oxford University Press, New York (1996).
11. D. Aletaha et al., "2010 Rheumatoid arthritis classification criteria: An American College of Rheumatology/European League Against Rheumatism collaborative initiative," *Arthritis. Rheum.* **62**(9), 2569–2581 (2010).
12. V. Prapavat et al., "The development of a finger joint phantom for the optical simulation of early inflammatory rheumatic changes," *Biomed. Tech. (Berl)* **42**(11), 319–326 (1997).
13. V. Prapavat, "Anwendung der experimentellen Systemanalyse zur Informations—Gewinnung aus Streulicht im Frühstadium entzündlich-rheumatischer Veränderungen," Ph.D. Thesis, Technical University Berlin (1997).
14. H. Mohamed-Ali, R. W. Hauer, and H. Sorensen, "Morphology and growth behavior of synovial cells in monolayer culture," *Z Rheumatol.* **50**(2), 74–81 (1991).
15. L. Dahlberg et al., "Proteoglycan fragments in joint fluid. Influence of arthrosis and inflammation," *Acta Orthop. Scand.* **63**(4), 417–423 (1992).
16. Z. Szekanecz et al., "Angiogenesis and vasculogenesis in rheumatoid arthritis," *Curr. Opin. Rheumatol.* **22**(3), 299–306 (2010).