Active Learning for Convenient Annotation and Classification of Secondary Ion Mass Spectrometry Images

Michael Hanselmann[1], Jens Röder[1,2], Ullrich Köthe[1], Bernhard Y. Renard[3], Ron M. A. Heeren[4], Fred A. Hamprecht[1,⋆]

[1] Heidelberg Collaboratory for Image Processing (HCI), Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Speyerer Straße 6, Heidelberg, Germany. [2] Robert Bosch GmbH, CR/AEM5, Robert-Bosch-Straße 200, 31139 Hildesheim, Germany. [3] Research Group Bioinformatics (NG 4), Robert Koch Institute, Nordufer 20, Berlin, Germany. [4] FOM-AMOLF, Science park 104, 1098 XG Amsterdam, The Netherlands.

⋆ corresponding author. Email: fred.hamprecht@iwr.uni-heidelberg.de, Fax: +49 6221 54 5276.

Abstract: Digital staining for the automated annotation of Mass Spectrometry Imaging (MSI) data has previously been achieved using state-of-the-art classifiers such as random forests or support vector machines (SVMs). However, the training of such classifiers requires an expert to label exemplary data in advance. This process is time-consuming and hence costly, especially if the tissue is heterogeneous. In theory, it may be sufficient to only label few highly representative pixels of an MS image, but it is not known a priori which pixels to select. This motivates *active learning* **strategies in which the algorithm itself queries the expert by automatically suggesting promising candidate pixels of an MS image for labeling. Given a suitable querying strategy, the number of required training labels can be significantly reduced while maintaining classification accuracy. In this work, we propose active learning for convenient annotation of MSI data. We generalize a recently proposed active learning method to the multi-class case and combine it with the random forest classifier. Its superior performance over random sampling is demonstrated on Secondary Ion Mass Spectrometry data, making it an interesting approach for the classification of mass spectrometry images.**

Mass Spectrometry Imaging (MSI) (Caprioli *et al.*, 1997; McDonnell and Heeren, 2007) allows a detailed analysis of the spatial distribution of proteins, peptides, lipids or metabolites (Seeley and Caprioli, 2008b; Chaurand *et al.*, 2002). With recent efforts to standardize proteomics experiments (Taylor *et al.*, 2007; Slany *et al.*, 2009; Franck *et al.*, 2009; Green *et al.*, 2010), MSI continuously moves closer to clinical application (Fournier *et al.*, 2008; Seeley and Caprioli, 2008a,b; Walch *et al.*, 2008). In many of these recent studies, the MS image is spatially partitioned into coherent regions associated with cancer or healthy tissue, or regions corresponding to different cell types. Manual analysis requires the expert to inspect multiple m/z channel images. Moreover, analyzing the channel images independently may not even be sufficient for discriminating tissue types with similar molecular signatures. For these reasons and with data sizes of up to several gigabytes (Eijkel *et al.*, 2009) direct manual analysis becomes tedious or infeasible, emphasizing the need for automated methods.

Previous studies have shown that unsupervised methods such as hierarchical clustering (Deininger *et al.*, 2008), principal component analysis (PCA) (van de Plas *et al.*, 2007) or probabilistic latent semantic analysis (pLSA) (Hanselmann *et al.*, 2008) are useful for segmenting MS images into spectrally coherent regions based on their molecular signatures only. At the same time they are intrinsically limited by their inability to learn from expert annotations. One consequence is the lack of clear criteria for model optimization (Cord and Cunningham, 2008). If the underlying mathematical assumptions are inept for the data at hand, the user has very limited influence on the segmentation outcome.

Many recent studies have thus considered supervised approaches and demonstrated that, given a set of spatially resolved annotations or (immunohistochemical) expert labels, supervised classifiers can be used for automated discrimination of tissue types (Yanagisawa *et al.*, 2003; Schwartz *et al.*, 2005; Schwamborn *et al.*, 2007; Gerhard *et al.*, 2007; Hanselmann *et al.*, 2009b). Even so, technical and biological variability between experiments often remains significant (Meyer and Stühler, 2007). Depending on the precise application, this limits the classification accuracies that can be achieved, especially in studies where the size of the training set is small. In such scenarios, where training of classifiers that generalize well to new MSI data is difficult, more robust and reliable results might be obtained by training the classifier anew for each separate MSI set. However, labeling of MSI data is time-consuming, and, consequently, very expensive. It is thus desirable to reduce the number of required labels (i.e. labeling time for the expert) without jeopardizing classification accuracy. This motivates the application of semi-supervised learning techniques (SSL) (Zhu, 2005;

Chapelle *et al.*, 2006; Bruand *et al.*, 2011), active learning (AL) strategies (see Settles (Settles, 2009) for a review) or hybrid approaches (Rajan *et al.*, 2008).

SSL methods typically base their classification output on two sources of information: the labels given by the user and the underlying structure of the unlabeled data points. An interesting and highly interactive method for MALDI MSI analysis was recently published by Bruand et al. (Bruand *et al.*, 2011). While SSL approaches can exploit the information hidden in the unlabeled observations, they lack a concept for guiding the labeling expert. In contrast, in active learning, the algorithm iteratively queries the expert to label that observation for which additional knowledge may be most beneficial for improving the classifier's performance. By labeling the samples (observations) of a data set in a smart order, a high performance level can often be obtained with fewer training samples. Although active learning methods have shown excellent performance in many fields such as speech recognition (Riccardi and Hakkani-Tür, 2006), image classification (Joshi *et al.*, 2009), remote sensing (Li *et al.*, 2010; Mitra *et al.*, 2004; Tuia *et al.*, 2009), and biomedical imaging (Doyle and Madabhush, 2010; Oh *et al.*, 2011), only few researchers have applied them to mass spectrometry data (Zomer *et al.*, 2004; Iyuke, 2011; Shi *et al.*, 2010). None of these publications is on mass spectrometry imaging.

In this paper, we generalize a recently proposed active learning strategy (Röder *et al.*, 2012) to the multi-class setting, and combine it with the random forest classifier (Breiman, 2001), which has previously been used for efficient classification of MSI data (Hanselmann *et al.*, 2009b). We show on real world MS images that our approach results in high classification accuracies after only a few learning steps and is thus suitable for efficient annotation of MSI data sets. We further demonstrate that the algorithm has an inbuilt capacity for novelty detection, alerting the expert to previously unlabeled but distinct classes rather than blindly making a prediction. Given the same number of labels, our querying strategy outperforms traditional non-active learning by up to 10% in sensitivity and 2–4% in positive predictive value. In our experiments, random sampling requires more than twice as many labels to achieve the same performance level. Finally, our strategy does not suffer from the high variability between runs that are characteristic for the random sampling approach.

Methods

Active Learning

Active learning aims at achieving steep learning curves, i.e. high classification accuracies after seeing as few labeled training examples as possible. It is motivated by the observation that a classifier can benefit more from judiciously chosen and informative training examples than from large numbers of redundant and hence less informative examples (Schohn and Cohn, 2000). Typically, active learning approaches are iterative and "guide" the labeler in the sense that the algorithm chooses observations for which it needs labels (Settles, 2009). In each round, the algorithm requests a label for that observation (pixel) x of an (MS) image that has the maximum *training utility value (TUV)* in the set U of all unlabeled observations, and is thus expected to contribute most to improving the classifier's performance. After label assignment, the classifier is trained with the augmented label set, all unlabeled observations are reclassified, and the algorithm continues by presenting its next query. These steps are repeated until either the human expert is satisfied with the classification result or a predefined stopping criterion is met.

A meaningful TUV function balances two strategies: *exploration* of the feature space and *refinement* of the current decision boundary. The aim of exploration is to sample from those regions of feature space from which so far only few training examples are available. The rationale is that a test sample can only be classified well if enough (local) evidence is available. Whereas exploration thus seeks good sample coverage of the whole feature space, the refinement strategy tries to improve the classifier by sampling points that are close to the decision boundary, i.e. for which approximately equal probability for two or more classes is present. Fig. 1 illustrates these strategies for the binary case.

The proposed active querying strategy can be illustrated with the following thought experiment: consider three different points in a feature space, and do not assume that the true decision boundary comes from a simple parametric class, such as a hyperplane. Points 1 and 2 lie on the currently estimated decision boundary, point 3 lies far away from it. There are many labels available in the vicinity of point 1, few in the neighborhood of point 2 and none surrounding point 3. Let $x^{(i)}$ with $i \in \{1, 2, 3\}$ denote the three points.

A pure refinement strategy would favor points $\{1, 2\}$ over 3. An exploratory strategy would take more interest in 3 than in $\{1, 2\}$. We use a strategy that prefers $\{2, 3\}$ over 1, for the following reasons: Point 3 is interesting, because we know nothing about its true class (remember that

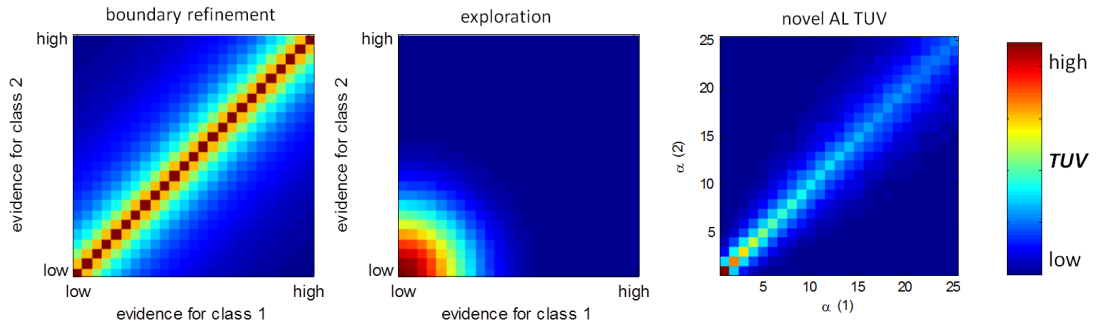

Figure 1: The figure illustrates the training utility value (TUV) of a candidate point in a binary classification setting. In (pure) decision boundary refinement, or uncertainty learning, candidate points with equal amounts of evidence for either class are preferred, regardless of how much evidence there is. In (pure) exploration, the candidate points receive a high score if the (local) evidence for both classes is low. Only the absolute "amount" of evidence is considered, its consistency is neglected. On the right, the newly proposed TUV function is shown for different parameter settings, where the evidence for classes 1 and 2 is measured by $\alpha_1 \in \{1, \ldots, 25\}$ and $\alpha_2 \in \{1, \ldots, 25\}$. We observe that our TUV function reconciles exploration and decision boundary refinement (also see also Supplementary Material B).

we do not assume a simple parametric model for the decision boundary). Point 2 is interesting because the location of the decision boundary is based on an estimate of $\hat{p}(Y|x^{(2)})$ which – being a random variable of itself – is of necessity imprecise when based on only few labeled points. There is thus some potential to be informed, or surprised, by an additional label at point 2. Point 1 is uninteresting because its estimate $\hat{p}(Y|x^{(1)})$ is based on a large number of nearby training examples, and we do not expect the decision boundary to change substantially in response to yet another label at that point. Finally, we factor the marginal density $\hat{p}(x)$ of all labeled and unlabeled points into the proposed training utility value. The reason is that estimating the decision boundary well is only relevant in populated regions of feature space. The vehicle used to capture the above intuition is a *second order distribution*, that is, the distribution of the probabilistic point estimate $\hat{p}(Y|x)$. This distribution and its use in a training utility value are defined next.

Training Utility Value Function

Above, we have informally discussed favorable properties of the TUV function. This section approaches the problem from a more theoretical perspective and may be skipped by the less mathematically interested reader.

Let $(\mathcal{X}, \mathcal{Y}) = \{(x^{(1)}, y^{(1)}), ..., (x^{(N)}, y^{(N)})\}$ be the set of N training samples, i.e. mass spectra $x^{(k)}$ with M channels and corresponding class labels $y^{(k)} \in \{1, \ldots, d\}$. Let further L be a loss function, i.e. a function that quantifies the penalty associated with an incorrect classification. The

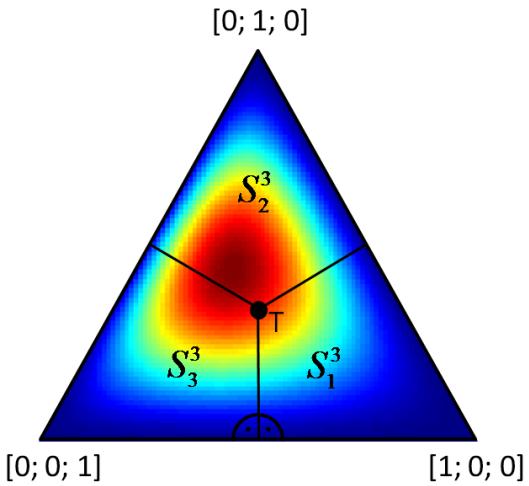

Figure 2: Each vertex of the simplex $\mathbb{S}^3$ corresponds to one of the $d = 3$ classes of interest. The mapping function θ (cf. eq. (3)) maps each point on the simplex to one of these classes. In the canonical case, each point is assigned to the closest vertex and hence to the class associated with that vertex. Figuratively, threshold point T (which lies in the center of the simplex) is used to partition $\mathbb{S}^3$ into three parts $\mathbb{S}^3_j, j = 1, \ldots, 3$. $\mathbb{S}^3_j$ is the Voronoi region associated with the j-th vertex. The posterior estimate for a test point can now be interpreted as a point on this simplex. In the TUV for Random Forests section, we further describe how a Dirichlet-distribution can be employed to describe the second-order distribution of the posterior. Color-coding is used to show an example for such a second order distribution, where blue indicates low and red indicateds high probability. A uniformly-colored simplex would correspond to an uninformative prediction. In contrast, in the example the plotted Dirichlet distribution is concentrated in part $\mathbb{S}^3_2$ of the simplex, indicating a preference for class two.

lowest achievable classification error, for a given loss function L, data distribution $p(x, y)$ and classification rule θ, is given by the overall *expected risk* $\int_{\mathcal{X}} R\big(\pi(x)\big)p(x)\,dx$. The conditional risk for misclassifying a point at position x is given by

$$R\big(\pi(x)\big) := \mathbb{E}_{Y|x}\Big(L\Big(Y = y, \theta\big(\pi(x)\big)\Big)\Big) \tag{1}$$

$$\pi(x) := \Big[p(Y = 1|x), \ldots, p(Y = d|x)\Big]^T \tag{2}$$

Here, $L(y, z)$ is the loss associated with a prediction z if the true class label is y; $\pi(x) \in \mathbb{S}^d$ is the vector of class conditional probabilities for each of the d classes which, thanks to the normalization constraint $\sum_{y\in\mathcal{Y}} p(y|x) = 1$, lies in the unit simplex $\mathbb{S}^d$ with d vertices (see Fig. 2 for an example with $d = 3$ vertices). Finally, θ is a classification rule $\mathbb{S}^d \mapsto \{1, \ldots, d\}$ which maps any point in the simplex to one of the d classes. The canonical mapping function θ employs the winner-takes-all strategy, i.e. maps each point from the simplex to its closest vertex (and hence to the class associated with that vertex).

In practice, the true class conditional probabilities $\pi(x)$ are not known, but need to be estimated from training data (Hastie *et al.*, 2009). Classifiers such as logistic regression or polychotomous logistic regression offer point estimates $q_y^0(x) := \hat{p}(Y = y|x),\ y \in \mathcal{Y}$ which can be compiled in a d-dimensional vector $q^0(x) = [\hat{p}(Y = 1|x), \ldots, \hat{p}(Y = d|x)] \in \mathbb{S}^d$. Plugging this point estimate into the conditional risk gives

$$R\big(q^0(x)\big) := \sum_{y \in \mathcal{Y}} L\Big(y, \theta\big(q^0(x)\big)\Big) \cdot q_y^0(x) \tag{3}$$

This quantity is the key ingredient of uncertainty sampling, which has been presented in many variants (Baum, 1991; Tong and Koller, 2000; Scheffer *et al.*, 2001). This class of active learning algorithms seeks to reduce the estimated expected risk by querying additional labels near the decision boundary, where the conditional risk is greatest. The implicit hope is that additional labels may drive the updated class conditional probability towards one of the simplex vertices, that is, to obtain unequivocal evidence for the dominance of one class. Uncertainty sampling is very simple to implement and widely used, but it is a pure exploitation / refinement strategy: it will never explore uncharted regions of feature space. Indeed, it will spend all of its queries around the current decision boundary. In addition, uncertainty sampling only relies on a point estimate of the posterior distribution and does not consider the uncertainties of the class conditional probability estimates themselves. This "second-order" uncertainty is implicitly taken into account in schemes such as error reduction sampling (Roy and McCallum, 2001; Zhu *et al.*, 2003). However, such look-ahead schemes require a (rank-one) update of the current classification boundary and turn out to be relatively expensive.

The novelty in Ref. (Röder *et al.*, 2012) is that it makes explicit, and capitalizes on, the uncertainty of the class-conditional probability itself. The latter, like any estimate that is obtained from finite training data, is subject to uncertainty. The prerequisite for their procedure is that the classifier must provide not merely a point estimate $q^0(x)$ for the class conditional probability, but a full *second-order distribution* over $q(x)$ as expressed by a probability density function $g\big(q(x)\big)$. More specifically, an estimated second-order distribution over the class-conditional probability can be written as

$$g\big(q(x)\big) := \frac{\partial G\big(q(x)\big)}{\partial q(x)} \tag{4}$$

$$G\big(q(x)\big) := Pr\Big(\hat{p}(Y = 1|x) \leq q_1(x) \wedge \ldots \wedge \hat{p}(Y = d|x) \leq q_d(x)\Big) \tag{5}$$

with density g and cumulative distribution function G.

If such a second-order distribution is available, the point estimate $q^0(x)$ can be identified with $q^0(x) \equiv \mathbb{E}_q\big(q(x)\big)$ and $R\big(q^0(x)\big)$ from Eq. (3) can be rewritten as $R\Big(\mathbb{E}_q\big(q(x)\big)\Big)$.

Now, in Ref. (Röder *et al.*, 2012) we argue that this estimate is overly conservative and tends to overrate the utility of samples whose intrinsic (i.e. Bayesian) uncertainty is high. We contrast it with the following distributional estimate, which measures the risk at location x arising from intrinsic uncertainty and insufficient training combined,

$$\mathbb{E}_q\Big(R\big(q(x)\big)\Big) := \sum_{y \in \mathcal{Y}} \int L\Big(y, \theta\big(q(x)\big)\Big) \cdot q_y(x) \cdot g\big(q(x)\big)\, dq(x) \tag{6}$$

We further argue that the extent by which these estimates differ, when weighted with the estimated marginal density $\hat{p}(x)$ (to take into account the importance of location x), is a good training utility value (TUV), or measure of interestingness, for yet unlabeled observations. Specifically, we posit

$$TUV(x) = \hat{p}(x)\bigg(R\Big(\mathbb{E}_q\big(q(x)\big)\Big) - \mathbb{E}_q\Big(R\big(q(x)\big)\Big)\bigg) \tag{7}$$

and show superior active learning curves when averaging over a large number of datasets.

This TUV function can be seen to naturally balance both exploration and refinement, see Fig. 1. In particular, unlike uncertainty sampling strategies, this criterion eventually desists from querying further labels near the decision boundary in areas where multiple labels are already available: These areas exhibit high intrinsic uncertainty that cannot be removed by additional label queries. Also, the proposed criterion does not have additional parameters as required by heuristic strategies that alternate between exploration and exploitation phases (Brinker, 2003).

To summarize this discussion: in areas with few labels, and in the absence of a parametric model that is known to govern the true posterior probability of a class, the estimate of the class conditional probability is of necessity imprecise. This uncertainty is reflected in a broad second-order distribution, which leads to lower values of $\mathbb{E}_q\Big(R\big(q(x)\big)\Big)$ as compared to the more conservative $R\Big(\mathbb{E}_q\big(q(x)\big)\Big)$. If, on the other hand, the local evidence is high, the second-order distribution is narrow, yielding similar values for both terms. An example is given in Supplementary Material A.

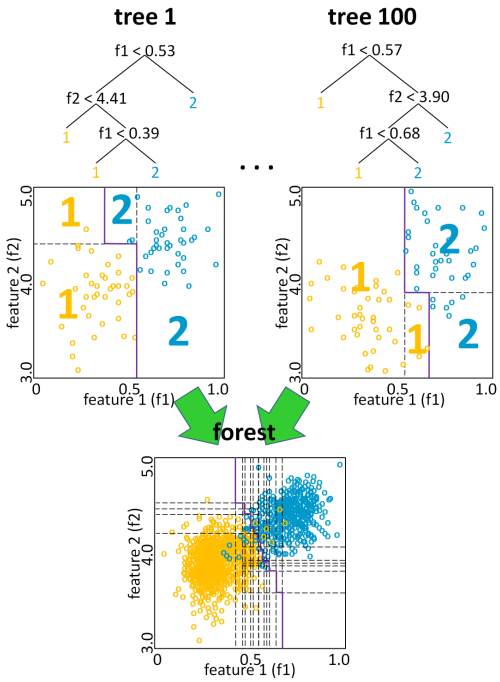

Figure 3: The random forest classifier is an ensemble of decision trees where the single trees are constructed from bootstrap samples. At each node of a tree, the feature that allows for the best class separation is chosen (with respect to the subset of features selected for that node). The corresponding partitioning of the feature space is shown with the decision boundary plotted in purple. The collection of trees forms the random forest whose classification is based on the majority votes of the individual trees.

Random Forests

The random forest (Breiman, 2001) (cf. Fig. 3) is a state-of-the-art ensemble classifier which comprises n_{tree} decision trees. Each individual tree constitutes a crisp classifier and is constructed from a bootstrap sample of size N of all available training samples. Tree construction starts at the root node and proceeds down toward the leaf nodes. In each node, a subset of the M features (i.e. mass channels) is chosen at random (a typical subset size being $\sqrt{M}$), and the feature that allows for the best class separation of the samples in the node is selected. After splitting the node, the algorithm continues on the next level until all nodes are pure, i.e. contain samples with consistent class labels. All samples which are not part of the bootstrap sample, the so-called out-of-bag samples, can be used to obtain a performance estimate for the classifier. A query sample is classified by putting it down each of the trees in the ensemble until it reaches the leaf nodes. The distribution over classes obtained for a single query sample cannot strictly be interpreted as a posterior probability, but does give an indication of how certain the classifier is in its prediction. Many studies have shown that the random forest classifier is robust to overfitting and label noise (Breiman, 2001; Saf-

fari *et al.*, 2009), delivers state-of-the-art prediction accuracy (Caruana *et al.*, 2008; Ulintz *et al.*, 2006), can handle a large number of input variables (Lin and Jeon, 2006; Breiman, 2004), allows for fast training, and is robust with respect to the exact choice of the two hyperparameters: number of trees, and size of the random feature subset evaluated at a node (Pardo and Sberveglieri, 2008).

TUV for Random Forests

We now combine the TUV with the random forest classifier in a multi-class setting. As discussed above, given a test sample the random forest classifier provides a distribution over tree votes. To obtain both a density estimate and a meaningful measure for the uncertainty (pure leaves suggest perfect certainty and are hence misleading), we train the random forest with all labeled examples from previous learning rounds plus a predefined fraction of samples from a uniformly distributed auxiliary class "0". After training, all hitherto unlabeled MSI samples are classified. Among these points the next query candidate is selected.

The number of trees $v_i(x)$ voting for the $d+1$ classes ($i = 0, 1, \ldots, d$) can now be interpreted as an indicator for how certain the classifier's assessment for x is. Simply put: The more trees vote for the auxiliary class, the weaker the local evidence for the other classes and thus the higher the uncertainty of the classifier. At the same time, the relative number of votes for the remaining classes is an indicator how far x lies from the decision boundary. Generalizing the Beta distribution from Ref. (Röder *et al.*, 2012) to multiple classes, we model the probability density function $g(q)$ (cf. eq. (5)) with a Dirichlet distribution, which is parameterized by the number of trees voting for classes 1 to d. This yields $g(q) = Dir(q|\alpha)$ where $\alpha \in \mathbb{N}_+^d, \alpha_y = 1 + v_y(x)$, and $\sum_{i=1}^{d} \alpha_i = d + n_{tree}$ (see Fig. 2). The complete mathematical derivation is detailed in Supplementary Material B.

Fig. 1 and Supplementary Material C show that this choice yields a TUV function that obeys both exploration and refinement principles. Computation of the TUV requires Monte Carlo integration over parts of the simplex. An efficient implementation is discussed in Supplementary Material D-E; MATLAB code is available from *http://hci.iwr.uni-heidelberg.de/MIP/Software*. An overview of the active learning method is given in algorithm 1.

Experiments

Data. We used secondary ion mass spectrometry (SIMS) data acquired from orthotopic human breast cancer xenografts (MCF-7) grown in mice. For data acquisition, a Physical Electronics

Algorithm 1 : Overview of our active learning procedure. The user interacts with the algorithm by answering the label queries in step 5.

Query label for observation x with the largest density in feature space
for $k = 1$ to maxIterations **do**
1. Uniformly sample from the bounding box enclosing all observations in feature space and label the obtained auxiliary samples as "0" (frequency controlled by resampling parameter)
2. Combine user-labeled samples and "0"-samples to train a random forest classifier with $d+1$ classes
3. Classify all unlabeled observations $x \in U$, i.e. all observations to which the user has not yet assigned a label
4. Drop random forest votes for class "0" to obtain d-dimensional vectors α for all unlabeled observations $x \in U$ with $\alpha_i = 1 + v_i(x), i = 1, \ldots, d$ where d is the number of classes and $v_i(x)$ is the number of trees that vote for class i given observation x
5. Query user label for that observation x that has the highest *training utility value (TUV)* among all yet unlabeled observations (i.e. $\max_{x \in U} TUV(x)$, cf. Eq. (7) and Supplementary Material D)
end for

TRIFT II TOF SIMS equipped with an Au+ liquid metal ion cluster gun was used. The tumor samples were embedded in gelatin, flash-frozen, cryo-sectioned to $\approx 10\mu m$ and thaw-mounted on a cold indium tin oxide-coated glass slide. The tissues were not washed prior to SIMS analysis, which was confined to a mass range of 0–2000 Da. The spectral resolution was rebinned to 0.1 Da and the range between 0–400 Da was selected, resulting in 4009 mass channels. Due to the large amount of data processed in this study, short acquisition times of 2 seconds per spot were used. Consequently, the spatial resolution had to be rebinned to $35 \times 35\mu m$ per pixel in order to guarantee a reasonable number of ion counts in each mass spectrum.

Three out of the six slices used in a previous study (Hanselmann *et al.*, 2009b, 2008) were selected for evaluation of our active learning method - one from the bottom (entitled S4), middle (S7) and top (S11) of the stack of available parallel slices of the tumor. The spectra in the three data sets were baseline corrected by channelwise subtraction of the minimum, normalized by their total ion count, and features were extracted with a peak picker based on local maximum detection. The dimensionality of the resulting spectra varied from 64 to 69 for the three sets. Crisp gold standard labels were obtained by Hematoxylin-Eosin (HE) staining of parallel slices and five classes of interest were identified: necrotic tumor, viable tumor, tumor interface, gelatin, glass/hole (see (Hanselmann *et al.*, 2009b) and Supplementary Material F for a more detailed description). All observations (pixels) for which label information is available were used in the evaluation of the methods. The class distribution among the labels corresponding to these observations determines the (maximum) number of different regions/classes in the segmentation result. Since section S4 only contains labels for four of the five classes, S4 was segmented into four regions. In contrast,

S7 and S11 were segmented into five regions.

Evaluation Criteria. We compared our active learning approach (AL-RF) to random sampling (RS) that in each learning step randomly queries the label of a hitherto unlabeled observation. Random sampling was used for comparison as it is known to be "surprisingly effective, being competitive with more complex approaches" (Cawley, 2011) and performs reasonably well in many studies (Guo and Schuurmans, 2008; Settles and Craven, 2008). It has thus been established as the de facto baseline strategy to compare new active learning algorithms to. Prediction accuracy was measured by sensitivity (SE) and positive predictive value (PPV). Sensitivity is defined as $SE = \frac{TP}{TP+FN}$ where TP is the number of true positives and FN is the number of false negatives. The positive predictive value estimates the ratio of samples that are correctly classified as class k among all samples that are classified as k, that is $PPV = \frac{TP}{TP+FP}$ where FP is the number of false positives. We averaged the obtained SE and PPV rates over all four (slice S4), respectively five classes (slices S7, S11).

Due to the non-deterministic nature of the RS strategy and the Monte Carlo integration, we repeated the active learning method and the random sampling approach 100 times and averaged the obtained results in each learning step. To obtain reliable quality estimates, in addition, we repeated the random forest training and classification in each learning step five times. We drew 300 samples to perform the Monte Carlo integrations and employed stratified sampling to balance the labels in the training set. In both approaches, the learning was started with an empty set of labeled points (in practical applications a number of initial labels might already be given, as it is e.g. also possible in AMASS (Bruand *et al.*, 2011)), exactly one label was queried in each active learning step where the ground truth label map served as oracle, and a 0-1 loss function was assumed.

Results

Fig. 4 and Supplementary Material G-H report the obtained classification accuracies on the three MSI datasets. Results are given for both querying strategies and an increasing number of learning steps. Ideally, the learning curves are steep, such that high classification accuracies are obtained after only few learning steps. Since this is typically achieved by first querying the labels that have the highest potential of increasing the classifier's performance, it is also insightful to examine which training points the methods select within a fixed number of learning steps (here: 100). Intuitively, some of the classes are easier to distinguish than others, which is likely to manifest itself in the

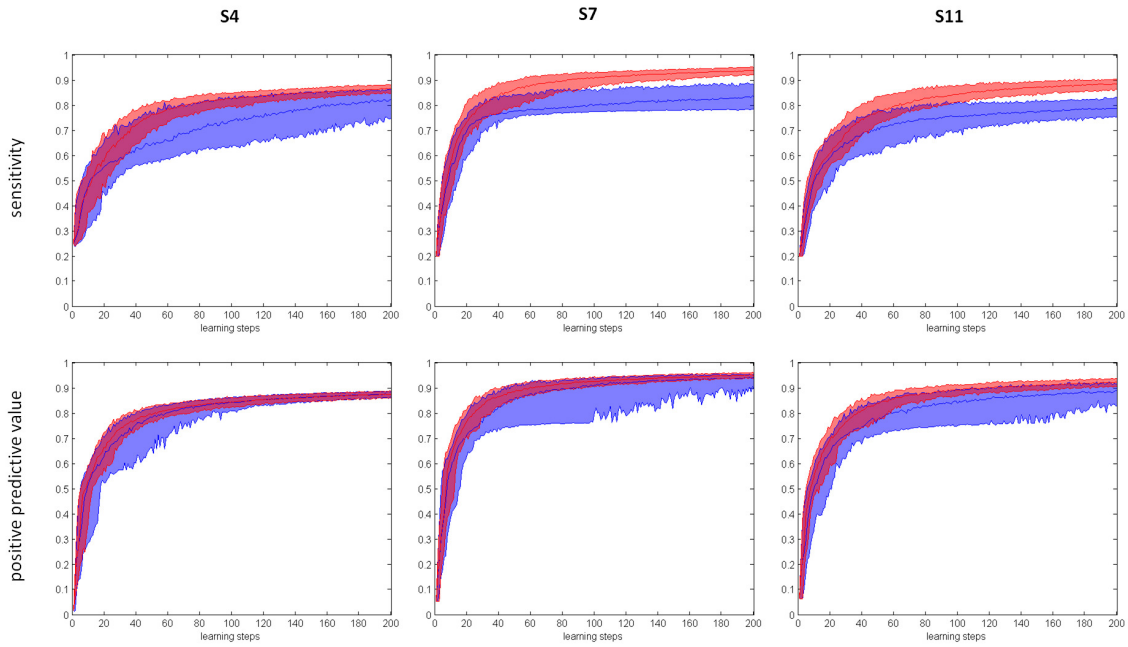

Figure 4: Learning curves obtained for random sampling (blue) and our active learning approach (red). Accuracies are measured by sensitivity (top row) and positive predictive value (bottom row). In each learning step, one additional label is queried. The plots show the median as well as the band between the 95% quantile and the 5% quantile for the 100 repeats. In contrast to random sampling, our active learning approach exhibits significantly lower variance between the different learning runs and the band around the median gets thinner over the course of iterations. At the same time, it significantly outperforms random sampling.

training point selection of the active learning strategy. Results are shown in Fig. 5, and a step-by-step example for slice S7 is given in Supplementary Material I. In detail, the following results were obtained for slices S4, S7, and S11:

Slice S4. Fig. 4 and Supplementary Material G reveal that our active learning scheme (AL-RF) performed similarly to random sampling (RS) in the first few learning steps and significantly outperformed RS as soon as more than ≈ 20 learning steps were executed. Due to the steeper learning curve, AL-RF improved on RS by about 10% in sensitivity after 100 iterations. RS needed more than 200 learning steps (i.e. twice as many labels) to achieve the same performance level. For a large number of learning steps, RS eventually collected a sufficient number of samples from all classes and hence converged towards the sensitivity rates obtained with AL-RF. However, the margin was still more than 5% after 200 iterations (cf. Supplementary Material H). Regarding positive predictive value, AL-RF slightly outperformed RS in the first ≈ 70 learning steps, that is, in the regime which is most interesting for a learning from sparse annotations.

Slice S7. Over the whole range of the first 200 iterations and especially for low numbers

of learning steps, our approach outperformed RS with respect to PPV. At the same time, it significantly outperformed RS regarding sensitivity, leading to a gain of more than 10% after 100 and also after 200 learning steps. Again, RS required more than twice as many labels to reach the performance level of AL-RF after 100 steps. The sensitivity of the RS algorithm increased very slowly such that after 500 iterations the sensitivity was still at a comparably low level of 86%.

Fig. 5 reveals that RS resulted in a classifier that mostly confused the necrotic class (indicated in red) with the viable class (light green). In contrast, AL-RF yielded significantly better results. Gelatin and glass spectra did not pose a challenge for either strategy.

Slice S11. Regarding sensitivity as well as positive predictive value, the results obtained for slice S11 proved to be highly similar to the results for slice S7. AL-RF again outperformed RS with respect to both sensitivity and positive predictive value. After 100 and 200 learning steps it resulted in SE and PPV rates which were approximately 9% respectively 4-6% higher than the results yielded with RS. Fig. 5 shows that RS again failed to achieve good classification performance for the necrotic class. AL-RF performed significantly better, but still confused several necrotic samples with viable cancer and some with glass. Apparently, additional learning steps are necessary to learn to reliably discriminate necrotic and viable tumor in this data set.

Discussion

Classification Performance. Given a fixed number of learning steps AL-RF resulted in positive predictive values which were slightly higher or comparable to the ones obtained with RS. At the same time, AL-RF significantly outperformed RS with respect to sensitivity by up to 10%, as soon as more than 15–20 labels were queried. It also exhibited significantly lower variance between runs, as can be seen from Fig. 4. The main conclusion is that AL-RF has the potential to reduce labeling times without trading for classification accuracy.

Training Point Selection. Fig. 5 shows that RS largely failed to discriminate necrotic from viable tumor tissue. The necrotic area has only small spatial extent, such that random sampling only selected few corresponding training points. In comparison, AL-RF selected more than twice as many necrotic samples on slices S7 and S11. This choice seems reasonable, since discriminating viable and necrotic tumor is the most challenging task of our classification problem. In any case, AL-RF yielded a significantly better classification result with respect to these classes (cf. Fig. 5). Whereas the necrotic and viable tumor samples are rather close in feature space, the non-tissue classes gelatin and glass have little spectral overlap, which simplifies their classification. Indeed,

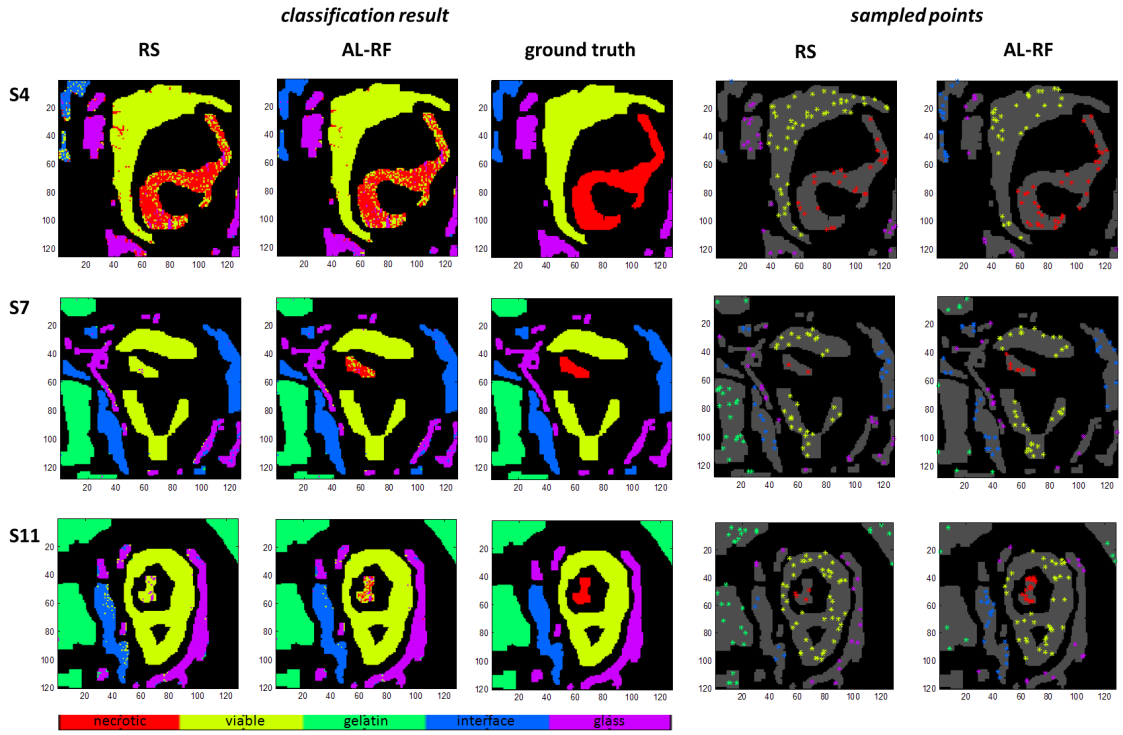

Figure 5: The classification results after 100 learning steps with our active learning method (AL-RF) and random sampling (RS). To obtain the crisp classification, we first averaged the probability maps gathered in the 100 repeats and then took the maximum likelihood estimate in each pixel. On the right, the selected training points for a representative learning run are plotted (we refrain from plotting the training points for all 100 repeats to keep the images uncluttered). Since the area of the necrotic class is comparatively small in slices S7 and S11, random sampling only selects very few training points for that class, leading to a bad classification result. In contrast, AL-RF requests more training samples for that class, yielding a superior classification. At the same time, it samples less points from the gelatin and glass classes, which have less overlap with the other classes in feature space than e.g. necrotic and viable tissue and are thus easier to learn.

AL-RF queried far fewer samples from these classes than RS, and the corresponding areas in Fig. 5 are less densely sampled. We conclude that AL-RF seems to construct training sets which are consistent with our expectations and prior knowledge about the classification task at hand.

Influence of the Number of Trees. There is some freedom in the exact choice of the second-order distribution. The Dirichlet, as a member of the exponential family with the correct support, is a canonical choice. While it allows the combination with the successful random forest classifier, using the tree votes as parameters introduces a certain shortcoming: When increasing the overall number of trees in the ensemble, the parameters specifying the Dirichlet distributions grow larger, which results in a narrower distribution. Thus, ultimately the uncertainty estimate is dependent on the number of trees. However, the number of trees in a random forest is fixed, typically between 100 and 200. Our experiments demonstrate that for this choice our criterion works well in practice.

Method's Assumptions. Supervised learning can only be as good as the labels provided, and it is thus important for the expert to ensure that the assigned labels are correct. This requires a certain level of interaction between the active learning approach and the microscopy software.

Unsupervised Segmentation can Assist the Labeling Process. Alternatively, PCA or pLSA scores may be used as overlays when assigning labels. These low-dimensional summaries of the MSI data often reveal structures that are not apparent from individual channel images but are often visible in the stained images (see Supplementary Material J for details).

Computation Time. Training of the random forest and subsequent classification took less than 1s on a standard desktop PC (2 GHz dual core processor with 2 GBytes of RAM). Computing the risk estimates for all unlabeled observations (cf. Supplementary Material C) required another 1.5-2s. Performance improvements may be achieved by employing an online version of the random forest classifier (Saffari *et al.*, 2009; Fuchs and Buhmann, 2009) or by querying multiple labels in each iteration (Cebron and Berthold, 2009), but this is beyond the scope of this paper. While a speed-up is always desirable, the measured computation times are clearly below the time that an expert typically needs for labeling the query point.

Future Work. Since AL-RF is based on the random forest classifier, which was shown to work well on complex MALDI signatures by several studies (see e.g. (Wu *et al.*, 2003)), and since the results for discriminating similar tissue classes such as viable and necrotic tissue are encouraging, we expect that AL-RF may also become an interesting tool for MALDI MSI analysis. Confirming or refuting this belief is an interesting avenue of future research. Also, the analyzed xenograft tumors

are rather homogeneous in nature. Thus, it will be interesting to analyze tissue types which are characterized by spectrally more overlapping signatures. Due to the reasons given above, we believe that AL-RF is suitable for this task.

Conclusions

Due to the enormous amount of data produced by modern-day instruments, routine clinical application of mass spectrometry imaging will not be possible without computational analysis (Eidhammer *et al.*, 2007). Robust training of supervised classifiers requires a set of expert labels that reflects the variability between patients and instrument settings. The high variability encountered in practice jeopardizes reproducibility and motivates the collection of expert labels for each newly acquired MSI data set. However, labeling is time-consuming and thus expensive. Consequently, novel algorithms are needed that yield the highest possible classification accuracies and at the same time require as little user-interaction as possible. We have demonstrated, how active learning can be used for the efficient annotation and classification of SIMS data. We have further demonstrated that it outperforms random sampling by a large margin if only a small number of labels are made available for training. Harvesting this potential is worthwhile as mass spectrometry imaging is moving closer to clinical application.

Acknowledgments

We thank Kristine Glunde (Johns Hopkins University School of Medicine, Baltimore, USA) and Erika R. Amstalden (FOM-AMOLF, Amsterdam, The Netherlands) for providing the tissue sections and MSI data, as well as Boaz Nadler (Weizman Institute of Science, Rehovot, Israel), Anna Kreshuk (University of Heidelberg, Germany), Xinghua Lou (Memorial Sloan-Kettering Cancer Center, New York, USA), and Marc Kirchner (Childrens Hospital Boston, USA) for fruitful discussions. We furthermore gratefully acknowledge financial support by the DFG under grant no. HA4364/6-1 (MH, BYR, FAH), and the Robert Bosch GmbH (JR, FAH). RMAH gratefully acknowledges financial support from the programme P24 of the Dutch national program COMMIT. Finally, we thank our reviewers for helpful comments and suggestions.

Supplementary Material

A: Working Example

Consider a binary classification example with class labels $\{-1; +1\}$. Assume that the unknown true posterior distribution $q_{+1}(x)$ is given by two point masses of 0.5 at 0.2 and 0.9.

Then, a point estimate for the posterior class probability can be obtained from

$$\hat{p}(+1|x) = \int_0^1 q g_x(q_{+1}(x)) dq \tag{8}$$

which for 0-1 loss yields

$$\hat{p}(Y = +1|x) = 0.2 \cdot 0.5 + 0.9 \cdot 0.5 = 0.55. \tag{9}$$

It follows that $q^0(x) = [0.45, 0.55]$. The classifier assigns classes -1 and +1 according to

$$\theta\big(\hat{p}(Y = 2|x)\big) = sgn\big(\hat{p}(Y = +1|x) - 0.5\big). \tag{10}$$

Thus, with Eq. (3) from the manuscript and Supplementary Material D (see below) we obtain

$$R\big(q^0(x)\big) = \sum_{y \in \mathcal{Y}} L\Big(y, \theta\big(q^0(x)\big)\Big) \cdot q_y^0(x) \tag{11}$$

$$= \min_{\tilde{y} \in \{-1;+1\}} \sum_{y \in \{-1;+1\}} L(y, \tilde{y}) \cdot q_y^0(x) \tag{12}$$

$$= \min_{\tilde{y} \in \{-1;+1\}} \{1 \cdot 0.55, 1 \cdot 0.45\} = 0.45. \tag{13}$$

This is the best we can do if only a point estimate of the true posterior distribution is available. Additional knowledge can be obtained by querying labels at x or its neighborhood.

Now assume that the true posterior distribution $g\big(q(x)\big)$ is known. Using Eq. (6) of the manuscript we obtain

$$\mathbb{E}_q\Big(R\big(q(x)\big)\Big) = \sum_{y \in \{-1;+1\}} \int L\Big(y, \theta\big(q(x)\big)\Big) \cdot q_y(x) \cdot g\big(q(x)\big)\, dq(x) \tag{14}$$

$$= 1 \cdot (1 - 0.9) \cdot 0.5 + 1 \cdot 0.2 \cdot 0.5 = 0.15, \tag{15}$$

where we have used that i.e. $q_{-1}(x) = 1 - q_{+1}(x)$. Thus, here the classic risk estimate is over-

pessimistic. Whereas $R\big(q(x)\big)$ is large, $\mathbb{E}_q\Big(R\big(q(x)\big)\Big)$ is much smaller, rendering x attractive for selection. From this example, it also becomes clear that the TUV is non-negative.

B: Derivation of the Second-order Distribution Estimate

Dirichlet Distribution is Conjugate to Multinomial Distribution

Let $\tilde{\pi}(x) = [p(Y = 0|x), \ldots, p(Y = d|x)]^T$ be the vector of true class conditional probabilities for each of the $d + 1$ classes and let $v_i(x)$ be the number of trees voting for class i where $i = 0$ is the auxiliary class. $v(x)$ can be modeled as a realization of a multinomially distributed random variable with density

$$Mult\big(v_0(x), \ldots, v_d(x) \mid \tilde{\pi}; n_{tree}\big) = \binom{n_{tree}}{v_0(x), \ldots, v_d(x)} \prod_{i=0}^{d} \tilde{\pi}_i^{v_i(x)} \tag{16}$$

where $n_{tree} = \sum_i v_i(x)$ is the total number of trees. The Dirichlet distribution, given by

$$Dir(\tilde{\pi} \mid \alpha) = \frac{\Gamma\left(\sum_{i=0}^{d} \alpha_i\right)}{\prod_{i=0}^{d} \Gamma\left(\alpha_i\right)} \prod_{i=0}^{d} \tilde{\pi}_i^{\alpha_i - 1} \tag{17}$$

for $\alpha = [\alpha_0, \ldots, \alpha_d]^T$ is conjugate to the multinomial distribution (Casella and Berger, 2002), and uniform on the simplex for $\alpha_i = 1, i = 0, \ldots, d$. Thus, applying Bayesian inference and multiplying the multinomial with the uniform prior yields the posterior distribution estimate (Bishop, 2007)

$$g\big(\tilde{\pi} \mid v_0(x), \ldots, v_d(x)\big) \propto Mult\big(v_0(x), \ldots, v_d(x) \mid \tilde{\pi}; n_{tree}\big) \cdot Dir(\tilde{\pi} \mid 1, ..., 1) \tag{18}$$

with

$$g\big(\tilde{\pi} \mid v_0(x), \ldots, v_d(x)\big) = Dir\big(\tilde{\pi} \mid 1 + v_0(x), \ldots, 1 + v_d(x)\big). \tag{19}$$

In this work we treat the output of the random forest as a realization of the true posterior which is unknown. We note, however, that RF is not known to be consistent (Biau *et al.*, 2008).

Dropping the Votes for the Reference Class

We are now interested in the distribution estimate for the scenario where the reference class (class 0) is dropped. It is known (Casella and Berger, 2002; Abramowitz and Stegun, 1965) that for stochastically independent and Gamma-distributed $Y^{(k)} \sim Gamma(\alpha_k, 1), k = 0, \ldots, d$ it holds

that

$$\left(\frac{Y^{(0)}}{\sum_{k=0}^{d} Y^{(k)}}, \frac{Y^{(1)}}{\sum_{k=0}^{d} Y^{(k)}}, \ldots, \frac{Y^{(d)}}{\sum_{k=0}^{d} Y^{(k)}}\right) \sim Dir(\alpha_0, \ldots, \alpha_d). \tag{20}$$

Thus, defining

$$W^{(i)} := \frac{Y^{(i)}}{\sum_{k=0}^{d} Y^{(k)}}, i = 0, \ldots, d \tag{21}$$

yields that

$$(W^{(0)}, \ldots, W^{(d)}) \sim Dir(\alpha_0, \ldots, \alpha_d). \tag{22}$$

We show that it follows that $(\tilde{W}^{(1)}, \ldots, \tilde{W}^{(d)}) \sim Dir(\alpha_1, \ldots, \alpha_d)$ where $\tilde{W}^{(i)} = \frac{W^{(i)}}{\sum_{l=1}^{d} W^{(l)}}$.

Proof: Using equation (21) we obtain:

$$\left(\frac{W^{(1)}}{\sum_{l=1}^{d} W^{(l)}}, \ldots, \frac{W^{(d)}}{\sum_{l=1}^{d} W^{(l)}}\right) = \left(\frac{\frac{Y^{(1)}}{\sum_{k=0}^{d} Y^{(k)}}}{\sum_{l=1}^{d} \frac{Y^{(l)}}{\sum_{k=0}^{d} Y^{(k)}}}, \ldots, \frac{\frac{Y^{(d)}}{\sum_{k=0}^{d} Y^{(k)}}}{\sum_{l=1}^{d} \frac{Y^{(l)}}{\sum_{k=0}^{d} Y^{(k)}}}\right) \tag{23}$$

$$= \left(\frac{Y^{(1)}}{\sum_{l=1}^{d} Y^{(l)}}, \ldots, \frac{Y^{(d)}}{\sum_{l=1}^{d} Y^{(l)}}\right) \tag{24}$$

which is again Dirichlet-distributed with parameters $\alpha_1, \ldots, \alpha_d$ (Casella and Berger, 2002).

C: Risk Reduction

Fig. 6 shows the expected risk reduction for adding a test point x to the set of training data, depending on its prediction result by the random forest. It demonstrates that our TUV criterion considers both exploration and boundary refinement.

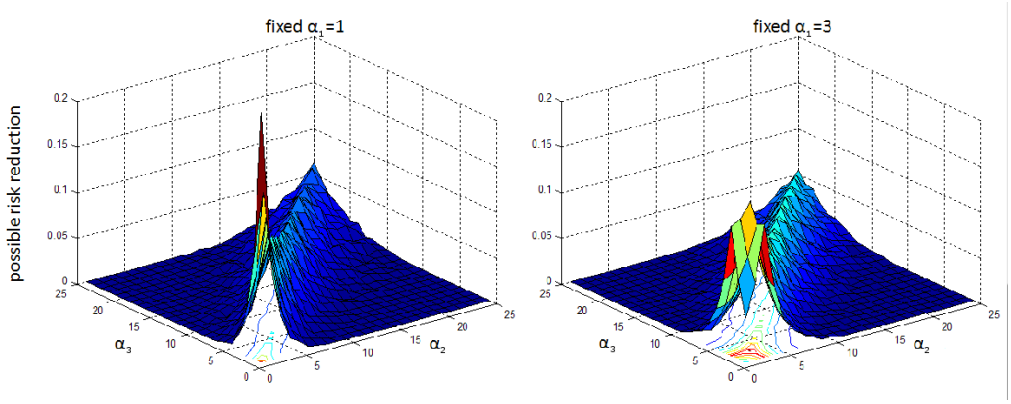

Figure 6: The figure shows the expected risk reduction $TUV(x) = (R(\mathbb{E}_q(q(x))) - \mathbb{E}_q(R(q(x))))$ for the 3-class case. The density $p(x)$ was fixed to 1 in both examples. We further fixed α_1 to 1 and 3 respectively (from left to right). Note that the TUV is symmetric in the parameters α_i such that fixing α_2 or α_3 instead leads to the same results. The highest TUV scores are obtained for α_2 and α_3 equal to 1 respectively 3, i.e. for uniform parameters (see contour plot in bottom plane). The TUV thus obeys the principle of exploitation/boundary refinement. Since the point $\alpha_i = 3\ \forall i$ corresponds to a lower level of uncertainty than the point $\alpha_i = 1\ \forall i$ (note that less local evidence is available), the maximum on the right is lower than the one on the left. It thus also obeys the concept of exploration. Also note that the TUV is symmetric with respect to the two varying parameters. Equal values for α_2 and α_3 lead to higher $TUVs$ than differing values.

D: Implementing the TUV Criterion

Computing Risk Estimates

Evaluation of the risk estimate $R\big(q^0(x)\big)$ (see eq. (3) of the manuscript) is straightforward. If the canonical mapping function θ is used that maps each point in the simplex to its closest vertex and under the assumption of 0-1 loss, i.e. $L(y,\tilde{y}) = 1 \forall y \neq \tilde{y}$ and $L(y,y) = 0$, we simply need to compute

$$R\big(q^0(x)\big) = \sum_{y\in\mathcal{Y}} L\Big(y, \theta\big(q^0(x)\big)\Big) \cdot q^0_y(x) = \min_{\tilde{y}\in\mathcal{Y}} \sum_{y\in\mathcal{Y}} L(y,\tilde{y}) \cdot q^0_y(x). \tag{25}$$

Computation of the distributional risk estimate $\mathbb{E}_q\left(R(q(x))\right)$ is more involved. Using a Dirichlet second-order distribution $g(q)$, the formula for the distributional estimate for the conditional risk at x can be rewritten as follows:

Let $B(\alpha) = \prod_{l=1}^d \Gamma(\alpha_l)/\Gamma(\sum_{l=1}^d \alpha_l)$ be the multinomial Beta function. From the manuscript we have

$$\mathbb{E}_q\Big(R\big(q(x)\big)\Big) := \sum_{y\in\mathcal{Y}} \int L\Big(y, \theta\big(q(x)\big)\Big) \cdot q_y(x) \cdot g\big(q(x)\big)\, dq(x). \tag{26}$$

Plugging in $g(q) = Dir(q|\alpha)$ with $\alpha \in \mathbb{N}_+^d, \alpha_y = 1 + v_y(x)$, and $\alpha^T\mathbf{1} = d + n_{tree}$, where $v_y(x)$ is the number of trees voting for class y given x and $\mathbf{1}$ is a unit vector, we obtain

$$= \sum_{y\in\mathcal{Y}} \int L\Big(y, \theta\big(q(x)\big)\Big) \cdot q_y(x) \cdot \Big[\frac{1}{B(\alpha)} \prod_{l=1}^d q_l(x)^{\alpha_l - 1}\Big] dq(x). \tag{27}$$

Let $e_y \in \mathbb{N}^d$ be a unit-length vector where only the yth entry is equal to one. Then the above equation can be rewritten as

$$= \sum_{y\in\mathcal{Y}} \int L\Big(y, \theta\big(q(x)\big)\Big) \cdot \frac{B(\alpha + e_y)}{B(\alpha)} \cdot \underbrace{\frac{1}{B(\alpha + e_y)} \prod_{l=1}^d q_l(x)^{(\alpha+e_y)_l - 1}}_{=: Dir\left(q(x)|\alpha+e_y\right)} dq(x). \tag{28}$$

which – using theorem 1 (cf. Supplementary Material E) – is equivalent to

$$= \sum_{y\in\mathcal{Y}} \int L\Big(y, \theta\big(q(x)\big)\Big) \cdot \frac{\alpha_y}{\sum\limits_k \alpha_k} \cdot Dir\big(q(x)|\alpha + e_y\big) dq(x). \tag{29}$$

Under the assumption that $L(y, y) = 0$, a final rewrite of eq. (29) yields

$$= \sum_{\tilde{y} \in \mathcal{Y}} \sum_{\substack{y \in \mathcal{Y} \\ y \neq \tilde{y}}} L(y, \tilde{y}) \frac{\alpha_y}{\sum_k \alpha_k} \underbrace{\int_{\mathbb{S}^d_{\tilde{y}}} Dir(q(x)|\alpha + e_y) dq(x)}_{*}, \tag{30}$$

where $\mathbb{S}^d_{\tilde{y}}$ denotes the yth part of the simplex, i.e. the part that belongs to class $\tilde{y}$ (see fig. 3 of the manuscript).

Integration over Part $\mathbb{S}^d_{\tilde{y}}$ of the Simplex

We use Monte Carlo integration (Hammersley, 1960) to approximate term (*) in eq. (30) which has to be calculated for each observation x (see Active Learning section of the manuscript). For each x we sample n_{sample} times from the corresponding Dirichlet distribution. For the experiments in the manuscript, $n_{sample} = 300$ was used.

Sampling from a Dirichlet distribution can efficiently be performed using Minka's Fastfit toolbox (Minka, 2004). Sampling from a d-dimensional Dirichlet distribution with parameters α_x boils down to drawing one sample from each of the d Gamma distributions $Gamma(\alpha_i, 1)$ with subsequent normalization by division with the sum. Since Minka's code is fast, it can in theory be used to sequentially draw samples from Dirichlet distributions with different parameterizations α. However, performing these calculations independently for all observations x is still very time-consuming. We speed up the procedure by exploiting the fact that in our scenario the parameterizations of the individual Dirichlet distributions are highly similar. Since $\alpha_i = 1 + v_i(x), i = 1, \ldots, d$, where $v_i(x)$ represents the number of trees voting for the classes i, it follows that $\alpha_i \in \{1, 2, \ldots, n_{tree} + 1\}$, that is the range of parameters is limited.

We can thus reuse the Gamma samples as follows: Assume, that for each observation x we want to draw n_{sample} d-dimensional samples $s^{(k)}, k = 1, \ldots, n_{sample}$ from a Dirichlet distribution parameterized by vector α to perform the Monte Carlo integration. Therefore, we first draw n_{sample} samples from $\Gamma(k), k = 1, \ldots, (n_{tree} + 1)$ and store the results in a $(n_{tree} + 1) \times n_{sample}$ matrix C which serves as a lookup table. Then, the n_{sample} requested samples are "constructed" from C by first selecting the d rows from the lookup table that correspond to the parameters $\alpha(i), i = 1, \ldots, d$ and storing them in a $d \times n_{sample}$ matrix S. Next, we randomly permute each row of S to avoid bias in case of non-unique α_i and do a column-wise normalization of S such that each column contains one Dirichlet sample $s^{(k)}, k = 1, \ldots, n_{sample}$[1].

[1] In our experiments, the procedure described above led to a significant speed-up factor of ≈ 100. We propose to

Given a threshold point T, we then determine for each of the n_{sample} samples $s^{(k)}, k = 1, \ldots, n_{sample}$ to which part $\mathbb{S}_j^d$ of the simplex $\mathbb{S}^d$ it belongs (see fig. 3 in the manuscript). The set Z_j of samples $\tilde{s}$ that fall in part $\mathbb{S}_j^d$ can be expressed by (Hanselmann *et al.*, 2009a)

$$Z_j = \left\{ \tilde{s} \in \left\{ s^{(1)}, \ldots, s^{(n_{sample})} \right\} \middle| 1 - \frac{\tilde{s}_j}{\tilde{s}_j + \tilde{s}_k} < 1 - \frac{T_j}{T_j + T_k} \forall k \in \{1, ..., d\} \setminus \{j\} \right\}. \tag{31}$$

Note that the assignment of a point to a corner of the simplex can efficiently be found in $d - 1$ pairwise comparisons. In each step, we compare the ratios in (31) with respect to two dimensions, e.g. $k = 1$ and $j = 2$. The dimension for which the inequality holds (here 1 or 2) is declared the "winner" which in the next step is compared to the next dimension ($k = winner(\{1, 2\}), j = 3$) and so on.

In case of 0-1 loss the threshold point resides in the center of the simplex such that the formula simplifies and we only have to determine the column-wise maxima of U in order to calculate the assignments.

E: Theorem 1

Proposition: Let $B(\alpha) = \frac{\prod_{l=1}^{d} \Gamma(\alpha_l)}{\Gamma(\sum_{l=1}^{d} \alpha_l)}$ be the multinomial Beta function. Then it holds that

$$\frac{B(\alpha + e_y)}{B(\alpha)} = \frac{\alpha_y}{\sum_k \alpha_k}. \tag{32}$$

Proof:

$$\frac{B(\alpha + e_y)}{B(\alpha)} = \frac{\Gamma(\alpha_1) \cdot \ldots \cdot \Gamma(\alpha_y + 1) \cdot \ldots \cdot \Gamma(\alpha_d)}{\Gamma(\alpha_1 + \ldots + (\alpha_y + 1) + \ldots + \alpha_d)} \tag{33}$$

$$\cdot \frac{\Gamma(\alpha_1 + \ldots + \alpha_y + \ldots + \alpha_d)}{\Gamma(\alpha_1) \cdot \ldots \cdot \Gamma(\alpha_y) \cdot \ldots \cdot \Gamma(\alpha_d)} \tag{34}$$

$$= \frac{\Gamma(\alpha_y + 1)}{\Gamma(\alpha_y)} \cdot \frac{\Gamma(\sum_k \alpha_k)}{\sum_k \alpha_k \Gamma(\sum_k \alpha_k)} = \frac{\alpha_y \Gamma(\alpha_y)}{\Gamma(\alpha_y)} \cdot \frac{1}{\sum_k \alpha_k} \tag{35}$$

$$= \frac{\alpha_i}{\sum_k \alpha_k} \tag{36}$$

where we use the iterative definition of the Gamma function (Bishop, 2007), that is $\Gamma(n + 1) = n\Gamma(n)$.

construct C (and thus S) anew in each turn of the active learning procedure.

F: Labels

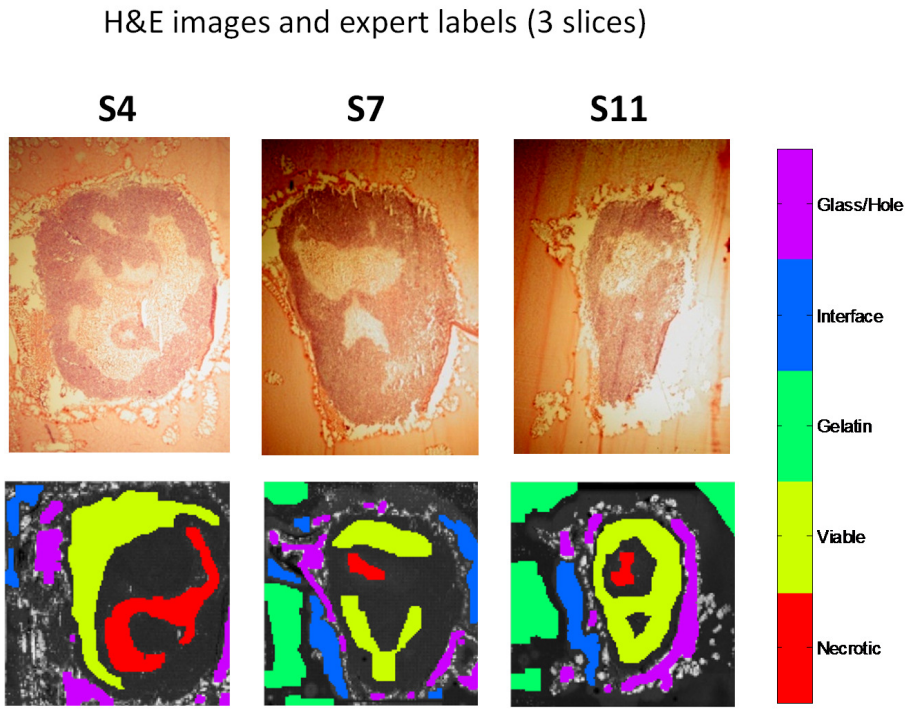

Figure 7: The gold standard labels for the three MSI data sets (bottom row) are obtained from Hematoxylin-Eosin (H&E) staining of parallel slices (top row). The labeling is only partial: labels for the five tissue classes of interest are color-coded whereas black and white indicates that no label information is available.

G: Mean Sensitivities and Positive Predictive Values

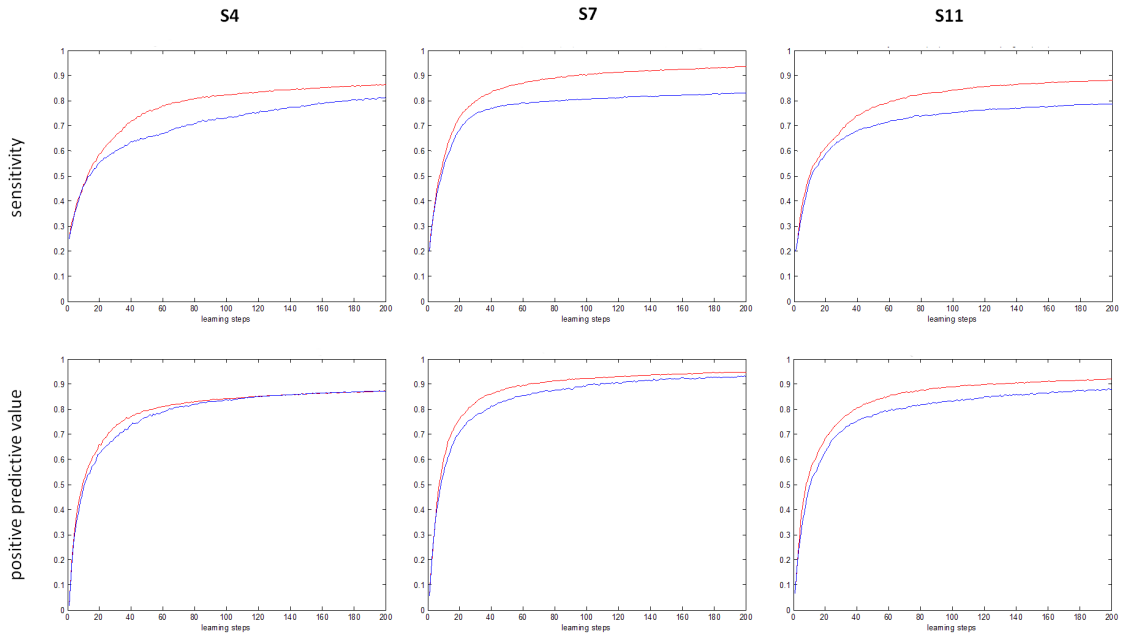

Figure 8: Comparison of the sensitivities and positive predictive values obtained with random sampling (blue) and our active learning approach (red) (averaged over 100 repeats). On all sets, our method outperforms random sampling with respect to sensitivity and positive predictive value.

H: Performance After Fixed Numbers of Learning Steps

set	criterion	method	50	100	150	200
S4	SE	RS	65.3	73.0	77.7	81.4
		AL	**75.6**	**82.4**	**84.8**	**86.4**
	PPV	RS	76.8	83.7	**86.1**	**87.3**
		AL	**79.4**	**84.2**	**86.1**	87.2
S7	SE	RS	78.3	80.6	82.0	83.1
		AL	**85.5**	**90.3**	**92.2**	**93.6**
	PPV	RS	83.5	89.5	92.0	93.1
		AL	**88.4**	**92.2**	**93.7**	**94.8**
S11	SE	RS	70.1	75.2	77.3	78.6
		AL	**77.3**	**84.3**	**87.0**	**88.1**
	PPV	RS	77.7	83.1	86.3	87.9
		AL	**83.4**	**89.0**	**90.8**	**92.1**

Table 1: Average sensitivities and positive predictive values for the three datasets after 50, 100, 150 and 200 learning steps. Our active learning approach significantly outperforms random sampling on all sets.

I: Intermediate Steps of the Active Learning Algorithm

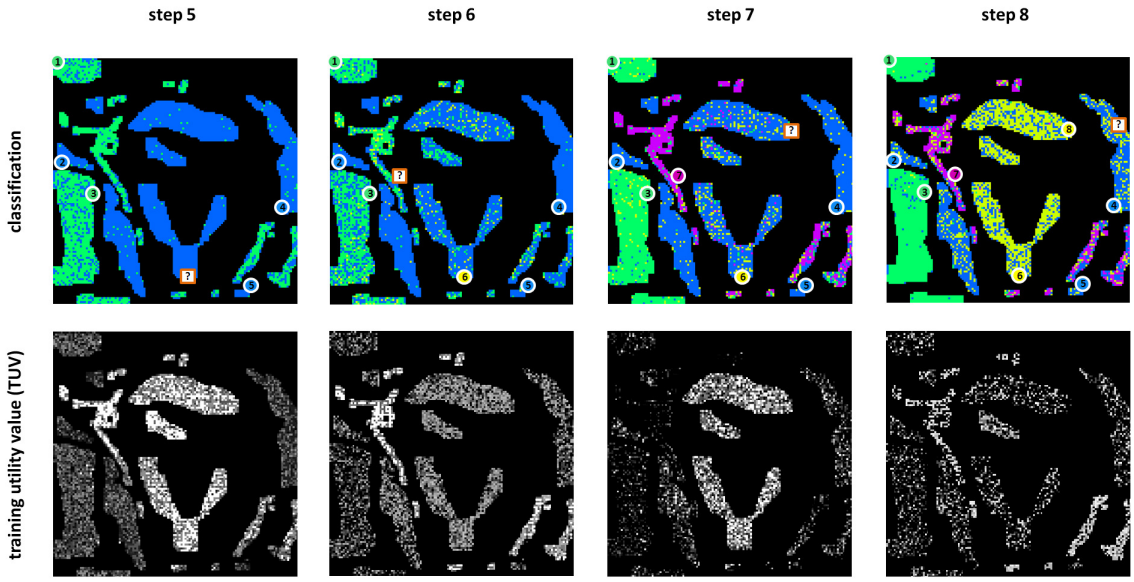

Figure 9: Here, we present intermediate steps for a single run of our active learning algorithm. The top row displays the classification results obtained after learning steps 5-8 (color coding as in Fig. 7). The bottom row shows the corresponding training utility value (TUV) maps where light areas correspond to high TUVs, i.e. points with high possible risk reduction. For instance, in step 5 the TUV is high for the observations corresponding to the necrotic, viable and glass classes, and a sample from the viable area is selected (indicated by the question mark). Consequently, in the next step the classification result for that class *slightly* improves and the respective TUV values decrease. Whereas the TUV values for the necrotic class also decrease since the viable and necrotic class are close in feature space, the TUV values for the glass area stay high, such that in the next step, a glass sample is picked, leading to significant improvement in the classification accuracy of that class.

J: Unsupervised Segmentation can Assist the Labeling Process

Probabilistic latent semantic analysis (pLSA) is an unsupervised learning technique which has successfully been employed for the segmentation of mass spectrometry images (Hanselmann *et al.*, 2008). It has also been added to the latest release of Bruker's ClinProTools software (Deininger *et al.*, 2012). Like all unsupervised learning techniques, pLSA does not make use of label information, but performs the analysis on the observed data only. It thus requires less user interaction since no labels have to be provided. On the other hand, it disregards the extra information which may be available and improve the segmentation result.

Unsupervised learning methods can assist the labeling process: PCA or pLSA scores may be used as overlays when assigning labels. These low-dimensional summaries of the MSI data often reveal structures that are not apparent from individual channel images but are often visible in the stained images.

Choosing the optimal number of components/segments k^* to decompose the MS image into is a challenging task which is specific to unsupervised learning methods. Usually, k^* has to be set manually or needs to be estimated from the data. However, in our case labeling information is available and we know that the expected number of classes is four (S4) respectively five (S7, S11). Not surprisingly, setting the number of components accordingly yielded the best segmentation results. In our experiments, we have thus set $k^* = 4$ for S4 respectively $k^* = 5$ for S7 and S11 and in each case have used all spectra of the set for which label information was available.

Fig. 10 shows that pLSA delivers surprisingly good results on all data sets, given that no label information was used in the decomposition. Many of the segments from the ground truth label maps in Fig. 7 are correctly reconstructed (e.g. the glass regions in S4, S7, S11 or the gelatin region in S11). At the same time, we notice that the pLSA algorithm has severe problems in discriminating the two tissue classes (necrotic and viable tumor). Whereas one might argue that in case of S4 the discrimination is possible from components 2 and 3, none of the five components created from S7 and S11 is discriminative for necrotic and/or viable tumor. Comparing the pLSA results with the segmentations given in Fig. 5 of the manuscript confirms that the supervised active learning algorithm is significantly more accurate in classifying necrotic and viable tumor. pLSA, as any other unsupervised learning method, is based on mathematical assumptions about the properties of the "clusters" within a data set. Internally, it relies on a distance metric that - pLSA being an unsupervised method - cannot be tweaked by the provided label information. Thus, good results can only been obtained as long as its general assumptions fit the characteristics

of the data at hand. This is a principle limitation of unsupervised methods. In our examples, pLSA seems to work well for classes which are distant in feature space but seem to work less well if (some of) these classes are close or even overlap.

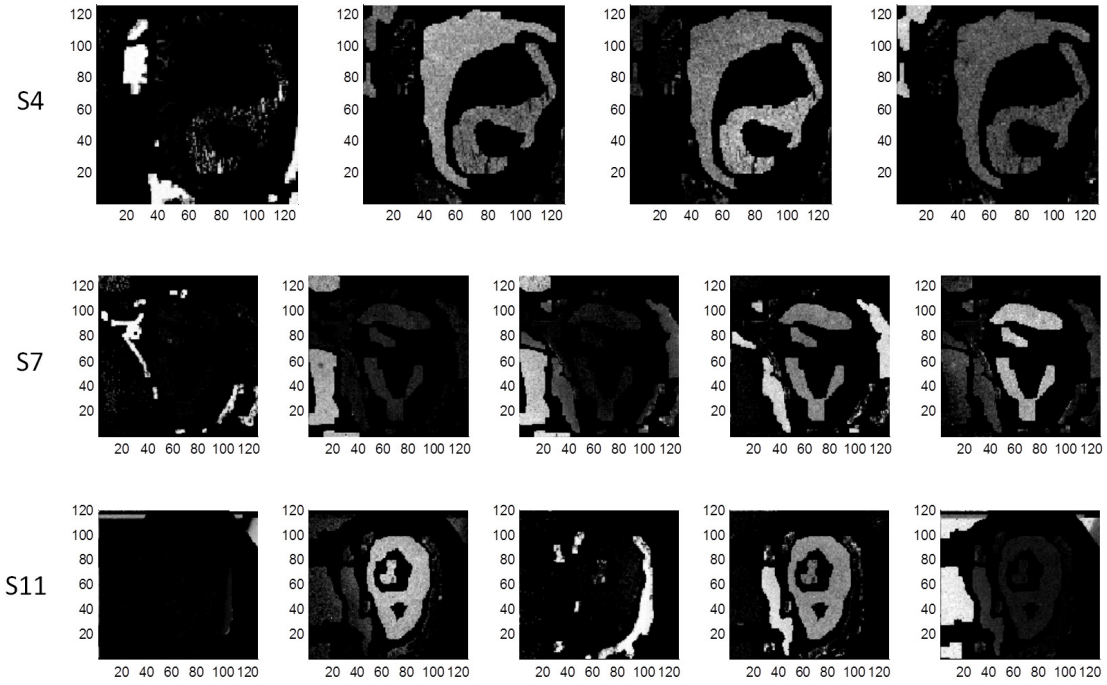

Figure 10: Segmentation results obtained from probabilistic latent semantic analysis (pLSA). Comparison with the ground truth labels given in Fig. 7 reveals that overall the pLSA results are very good, given that no label information was used. The components found on S4 seem to correspond to the classes glass, tissue (most likely viable), tissue (most likely necrotic), and interface. The complementarity of components 2-4 is not very well expressed, which may be due to the proximity of these classes in feature space. Whereas this result is encouraging, we also observed shortcomings of the method. For slice S5 we obtained five components which seem to match to glass, gelatine (1), gelatine (2), interface, and tissue (necrotic and viable tumor) regions. Whereas glass and gelatin are well identified, we were not able to better separate necrotic from viable tissue, even for different numbers of pLSA components. The same holds true for slice S11. Here component 1 picks up some signal variation. At the same time, pLSA again misses to separate necrotic from viable tumor.

References

Abramowitz, M. and Stegun, I. (1965). *Handbook of Mathematical Functions*. Dover.

Baum, E. (1991). Neural net algorithms that learn in polynomial time from examples and queries. *IEEE Trans. on Neural Networks*, **2**, 5–19.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, **9**, 2015–2033.

Bishop, C. (2007). *Pattern Recognition and Machine Learning*. Springer.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

Breiman, L. (2004). Consistency of a simple model of random forests. Technical report, University of California, Berkeley.

Brinker, K. (2003). Incorporating diversity in active learning with support vector machines. *Proc. of the 20th Int. Conf. on Machine Learning*, pages 59–66.

Bruand, J., Alexandrov, T., Sistla, S., Wisztorski, M., Meriaux, C., Becker, M., Salzet, M., Fournier, I., Macagno, E., and Bafna, V. (2011). AMASS: Algorithm for MSI analysis by semi-supervised segmentation. *Journal of Proteome Research*, **10**, 4734–4743.

Caprioli, R., Farmer, T., and Gile, J. (1997). Molecular imaging of biological samples: Localization of peptides and proteins using MALDI-TOF MS. *Analytical Chemistry*, **69**, 4751–4760.

Caruana, R., Karampatziakis, N., and Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. *Proc. of the 25th Int. Conf. on Machine Learning*, pages 96–103.

Casella, G. and Berger, R. (2002). *Statistical Inference*. Duxbury Advanced Series.

Cawley, G. (2011). Baseline methods for active learning. *JMLR Workshop and Conf. Proc.*, **16**, 47–57.

Cebron, N. and Berthold, M. (2009). Active learning for object classification: from exploration to exploitation. *Data Mining and Knowledge Discovery*, **18**(2), 283–299.

Chapelle, O., Zien, A., and Schölkopf, B. (2006). *Semi-Supervised Learning*. MIT Press.

Chaurand, P., Schwartz, S., and Caprioli, R. (2002). Imaging mass spectrometry: a new tool to investigate the spatial organization of peptides and proteins in mammalian tissue sections. *Current Opinion in Chemical Biology*, **6**, 676–681.

Cord, M. and Cunningham, P. (2008). *Machine Learning Techniques for Multimedia*. Springer, 1st edition.

Deininger, S.-O., Ebert, M., Fütterer, A., Gerhard, M., and Röcken, C. (2008). MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *Journal of Proteome Research*, **7**(12), 5230–5236.

Deininger, S.-O., Meyer, K., and Walch, A. (2012). Application note mt-111: Concise interpretation of MALDI imaging data by probabilistic latent semantic analysis (pLSA). Technical report, Bruker Daltonics.

Doyle, S. and Madabhush, A. (2010). Consensus of ambiguity: Theory and application of active learning for biomedical image analysis. *Pattern Recognition in Bioinformatics (LNCS 6282)*, pages 313–324.

Eidhammer, I., Flikka, K., Martens, L., and Mikalsen, S.-O. (2007). *Computational Methods for Mass Spectrometry Proteomics*. John Wiley and Sons.

Eijkel, G., Kükrer-Kaletas, B., van der Wiel, I., Kros, J., Luider, T., and Heeren, R. (2009). Correlating MALDI and SIMS imaging mass spectrometric datasets of biological tissue surfaces. *Surface and Interface Analysis*, **41**, 675–685.

Fournier, I., Wisztorski, M., and Salzet, M. (2008). Tissue imaging using MALDI-MS: a new frontier of histopathology proteomics. *Expert Reviews in Proteomics*, **5**(3), 413–424.

Franck, J., Arafah, K., Elayed, M., Bonnel, D., Vergara, D., Jacquet, A., Vinatier, D., Wisztorski, M., Day, R., Fournier, I., and Salzet, M. (2009). MALDI imaging mass spectrometry - state of the art technology in clinical proteomics. *Molecular and Cellular Proteomics*, **8**, 2023–2033.

Fuchs, T. and Buhmann, J. (2009). Inter-active learning of randomized tree ensembles for object detection. *3rd IEEE ICCV Workshop on On-line Computer Vision*, pages 1370–1377.

Gerhard, M., Deininger, S.-O., and Schleif, F.-M. (2007). Statistical classification and visualization of MALDI-imaging data. *Symp. on Computer-Based Medical Systems*, **20-22**, 403–405.

Green, F., Gilmore, I., Lee, J., Spencer, S., and Seah, M. (2010). Static SIMS-VAMAS interlaboratory study for intensity repeatability, mass scale accuracy and relative quantitation. *Surface and Interface Analysis*, **42**(3), 129–138.

Guo, Y. and Schuurmans, D. (2008). Discriminative batch mode active learning. *Advances in Neural Information Processing Systems (NIPS)*, (20), 593–600.

Hammersley, J. (1960). Monte carlo methods for solving multivariable problems. *The Annals of the New York Academy of Science*, **86**, 844–874.

Hanselmann, M., Kirchner, M., Renard, B., Amstalden, E., Glunde, K., Heeren, R., and Hamprecht, F. (2008). Concise representation of mass spectrometry images by probabilistic latent semantic analysis. *Analytical Chemistry*, **80**(24), 9649–9658.

Hanselmann, M., Köthe, U., Kirchner, M., Renard, B., Heeren, R., and Hamprecht, F. (2009a). Multivariate segmentation of compositional data. *Proc. of the 15th Int. Conf. on Discrete Geometry for Computer Imagery (DGCI), Lecture Notes in Computer Science (LNCS)*, **5810**, 180–192.

Hanselmann, M., Köthe, U., Kirchner, M., Renard, B., Amstalden, E., Glunde, K., Heeren, R., and Hamprecht, F. (2009b). Toward digital staining using imaging mass spectrometry and random forests. *Journal of Proteome Research*, **8**(7), 3558–3567.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition.

Iyuke, F. (2011). *Active Learning for the Prediction of Asparagine and Aspartate Hydroxylation Sites on Human Proteins*. Master's thesis, Ottawa-Carleton Institute for Biomedical Engineering, Ottawa, Canada.

Joshi, A., Porikli, F., and Papanikolopoulos, N. (2009). Multi-class active learning for image classification. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2372–2379.

Li, J., Bioucas-Dias, J., and Plaza, A. (2010). Semisupervised hyperspectral image segmentation using multinominal logistic regression with active learning. *IEEE Trans. on Geoscience and Remote Sensing*, **48**(11), 4085–4098.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Society*, **101**, 578–590.

McDonnell, L. and Heeren, R. (2007). Imaging mass spectrometry. *Mass Spectrometry Reviews*, **26**, 606–643.

Meyer, H. and Stühler, K. (2007). High-performance proteomics as a tool in biomarker discovery. *Proteomics*, **7 Suppl 1**, 18–26.

Minka, T. (2004). Fastfit toolbox for MATLAB, version 1.2. http://research.microsoft.com/en-us/um/people/minka/software/fastfit/. accessed March 2009.

Mitra, P., Shankar, B., and Pal, S. (2004). Independent component analysis for the extraction of reliable protein signal profiles from maldi-tof mass spectra. *Pattern Recognition Letters*, **25**(9), 1067–1074.

Oh, S., Lee, M., and Zhang, B.-T. (2011). Ensemble learning with active example selection for imbalanced biomedical data classification. *5th IAPR Intern. Conf. on Pattern Recognition in Bioinformatics, Lecture Notes in Computer Science (LNCS)*, **6282**, 316–325.

Pardo, M. and Sberveglieri, G. (2008). Random forests and nearest shrunken centroids for the classification of sensor array data. *Sensor and Actuators*, **131**, 93–99.

Rajan, S., Ghosh, J., and Crawford, M. (2008). An active learning approach to hyperspectral data classification. *IEEE Trans. on Geoscience and Remote Sensing*, **46**(4), 1231–1242.

Riccardi, G. and Hakkani-Tür, D. (2006). Active learning: Theory and applications to automatic speech recognition. *IEEE Trans. on Speech and Audio Processing*, **13**(4), 1–8.

Röder, J., Kunzmann, K., Nadler, B., and Hamprecht, F. (2012). Active learning with distributional estimates. *Conf. on Uncertainty in Artificial Intelligence*.

Roy, N. and McCallum, A. (2001). Toward optimal active learning through sampling estimation of error reduction. *Proc. of the 18th Int. Conf. on Machine Learning*, pages 441–448.

Saffari, A., Leistner, C., Santner, J., Godec, M., and Bischof, H. (2009). On-line random forests. *3rd IEEE ICCV Workshop on On-line Computer Vision*, pages 1393–1400.

Scheffer, T., Decomain, C., and Wrobel, S. (2001). Active hidden markov models for information extraction. *Proc. Int. Conf on Advances in Intelligent Data Analysis*, pages 309–318.

Schohn, G. and Cohn, D. (2000). Less is more: Active learning with support vector machines. *Proc. of the 17th Int. Conf. on Machine Learning*, pages 839–846.

Schwamborn, K., Krieg, R., Reska, M., Jakse, G., Knuechel, R., and Wellmann, A. (2007). Identifying prostate carcinoma by MALDI-imaging. *International Journal of Molecular Medicine*, **20**, 155–159.

Schwartz, S., Weil, R., Thompson, R., Shyr, Y., Moore, J., Toms, S., Johnson, M., and Caprioli, R. (2005). Proteomic-based prognosis of brain tumor patients using direct-tissue matrix-assisted laser desorption ionization mass spectrometry. *Cancer Research*, **65**(17), 7674.

Seeley, E. and Caprioli, R. (2008a). Imaging mass spectrometry: Towards clinical diagnostics. *Proteomics - Clinical Applications*, **2**, 1435–1443.

Seeley, E. and Caprioli, R. (2008b). Molecular imaging of proteins in tissues by mass spectrometry. *Proceedings of the National Academy of Sciences (PNAS)*, **105**(47), 18126–18131.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Settles, B. and Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks. *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, pages 1070–1079.

Shi, J., Lin, W., and Wu, F.-X. (2010). Statistical analysis of mascot peptides identification with active logistic regression. *Int. Conf. on Bioinformatics and Biomedical Engineering*, pages 1–4.

Slany, A., Haudek, V., Gundacker, N., Griss, J., Mohr, T., Wimmer, H., Eisenbauer, M., Elbling, L., and Gerner, C. (2009). Introducing a new parameter for quality control of proteome profiles: Consideration of commonly expressed proteins. *Electrophoresis*, **30**(8), 1306–28.

Taylor, C., Paton, N., Lilley, K., Binz, P., Julian, R. J., Jones, A., Zhu, W., R., A., Aebersold, R., Deutsch, E., Dunn, M., Heck, A., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T., Patterson, S., Ping, P., Seymour, S., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T., Whitelegge, J., Wilkins, M., Xenarios, I., Yates, J., and Hermjakob, H. (2007). The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology*, **25**(8), 887–893.

Tong, S. and Koller, D. (2000). Support vector machine active learning with applications to text classification. *Proc. Int. Conf on Machine Learning*, pages 999–1006.

Tuia, D., Ratle, F., Pacifici, F., Kanevski, M., and Emery, W. (2009). Active learning methods for remote sensing image classification. *IEEE Trans. on Geoscience and Remote Sensing*, **47**(7), 2218–2232.

Ulintz, P., Zhu, J., Qin, Z., and Andrews, P. (2006). Improved classification of mass spectrometry database search results using newer machine learning approaches. *Molecular and Cellular Proteomics*, **5**, 497–509.

van de Plas, R., Ojeda, F., Dewil, M., van den Bosch, L., de Moor, B., and Waelkens, E. (2007). Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis. *Proc. of the Pacific Symp. of Biocomputing*, **12**, 458–469.

Walch, A., Rauser, S., Deininger, S.-O., and Höfler, H. (2008). MALDI imaging mass spectrometry for direct tissue analysis: a new frontier for molecular histology. *Histochemistry and Cell Biology*, **130**(3), 421–434.

Wu, B., Abbott, T., Fishman, D., McMurray, G., Mor, G., Stone, K., Ward, D., Williams, K., and Zhao, H. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, **19**(13), 1636–1643.

Yanagisawa, K., Shyr, Y., Xu, B., Massion, P., Larsen, P., White, B., Roberts, J., Edgerton, M., Gonzalez, A., Nadaf, S., Moore, J., Caprioli, R., and Carbone, D. (2003). Proteomic patterns of tumour subsets in non-small-cell lung cancer. *The Lancet*, **362**(9382), 433–439.

Zhu, X. (2005). Semi-supervised learning literature survey. Computer Sciences Technical Report 1530, University of Wisconsin–Madison.

Zhu, X., Lafferty, J., and Ghahramani, Z. (2003). Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. *Proc. of the ICML 2003 Workshop on The Continuum from Labeled to Unlabeled Data*, pages 58–65.

Zomer, S., del Nogal Sánchez, M., Brereton, R., and Pérez Pavón, J. (2004). Active learning support vector machines for optimal sample selection in classification. *Journal of Chemometrics*, **18**, 294–305.