Bioinformatics, 36, 2020, i300–i308
doi: 10.1093/bioinformatics/btaa436
ISMB 2020

Deep multiple instance learning classifies subtissue locations in mass spectrometry images from tissue-level annotations

Dan Guo[1], Melanie Christine Föll[2,3], Veronika Volkmann[2,3], Kathrin Enderle-Ammour[2,3], Peter Bronsert[2,3,4,5], Oliver Schilling[2,3] and Olga Vitek[1,*]

[1]Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA, [2]Institute for Surgical Pathology, Medical Center – University of Freiburg, 79106 Freiburg, Germany, [3]Faculty of Medicine, University of Freiburg, 79110 Freiburg, Germany, [4]Tumorbank Comprehensive Cancer Center Freiburg, Medical Center – University of Freiburg and [5]German Cancer Consortium (DKTK) and Cancer Research Center (DKFZ), 79106 Freiburg, Germany

*To whom correspondence should be addressed.

Abstract

Motivation: Mass spectrometry imaging (MSI) characterizes the molecular composition of tissues at spatial resolution, and has a strong potential for distinguishing tissue types, or disease states. This can be achieved by supervised classification, which takes as input MSI spectra, and assigns class labels to subtissue locations. Unfortunately, developing such classifiers is hindered by the limited availability of training sets with subtissue labels as the ground truth. Subtissue labeling is prohibitively expensive, and only rough annotations of the entire tissues are typically available. Classifiers trained on data with approximate labels have sub-optimal performance.

Results: To alleviate this challenge, we contribute a semi-supervised approach *mi-CNN*. mi-CNN implements multiple instance learning with a convolutional neural network (CNN). The multiple instance aspect enables weak supervision from tissue-level annotations when classifying subtissue locations. The convolutional architecture of the CNN captures contextual dependencies between the spectral features. Evaluations on simulated and experimental datasets demonstrated that mi-CNN improved the subtissue classification as compared to traditional classifiers. We propose mi-CNN as an important step toward accurate subtissue classification in MSI, enabling rapid distinction between tissue types and disease states.

Availability and implementation: The data and code are available at https://github.com/Vitek-Lab/mi-CNN_MSI.

Contact: o.vitek@neu.edu

1 Introduction

Biochemical constitution of tissues varies with tissue types (such as epithelial and connective tissues), or disease states (such as tumor and healthy tissues). Mass spectrometry imaging (MSI) provides an untargeted characterization of the molecular composition of such tissues at spatial resolution, simultaneously quantifying hundreds of analytes without the need for chemical labels or antibodies (Spengler, 2015; Jones *et al.*, 2012). Therefore, MSI has a strong potential to become a rapid diagnostic technology in the clinic (Kriegsmann *et al.*, 2015; Vaysse *et al.*, 2017).

Although the name of the technology contains the word ‘image’, the structure of MSI data is very different from other bioimaging technologies (Fig. 1). In MSI, mass spectra are acquired at thousands of different spatial *locations* in a raster pattern throughout the tissue. MSI techniques fall into two major categories: matrix-assisted laser desorption/ionization (MALDI) MSI (Aichler and Walch, 2015) and desorption electrospray ionization (DESI) MSI (Wu *et al.*, 2013). With each technique, the mass spectrum obtained at each location is a collection of *features*, corresponding to the ions of biochemical analytes such as metabolites, lipids, peptides and proteins. The features do not contain direct information regarding the identity of the underlying analyte, except for their ratios of mass over charge *m*/*z*. For one tissue location, a typical MSI experiment reports hundreds to thousands of *m*/*z* in ascending order. The intensities of the *m*/*z* correlate with the abundance of the analyte. A plot of the abundance of one *m*/*z* across all locations is referred to as an *ion image*.

A reliable diagnostics can be achieved by supervised classification models that take as input the observed mass spectra, and predict labels such as tumor, healthy or tumor subtypes. Beyond *tissue-level classification* (classifying the entire tissues), *subtissue-level classification* (classifying the disease status of individual locations within the tissues) is of most interest. Ranking *m*/*z* features by their predictive ability is also important. Currently, training subtissue-level classifiers providing this information requires training sets of tissues with reliable *subtissue labels*.

Unfortunately, accessing a training set with reliable subtissue labels is challenging in practice. In a typical workflow, pathologists

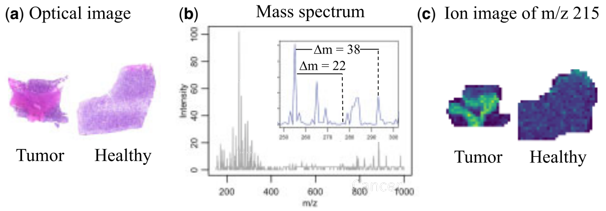

Fig. 1. MSI data. (**a**) H&E-stained optical images of a pair of tumor and healthy tissues from the human renal cell carcinoma (RCC) experiment. (**b**) Mass spectrum from one location in the tumor tissue. The inset zooms into $m/z \in (250, 300)$. Two features with *m*/*z* shift $\Delta m = 22$ can correspond to molecular ions and sodium adducts. Two features with *m*/*z* shift $\Delta m = 38$ can correspond to molecular ions and potassium adducts. (**c**) Ion images of *m*/*z* 215 of the tissues in (a)

examine the hematoxylin and eosin (H&E)-stained optical images such as in Figure 1a. To obtain subtissue labels, the pathologist must manually examine and annotate the distinct regions of each tissue (Lou *et al.*, 2017). The cost of manual work is one of the reasons to the relatively small number of biological replicates in MSI. The procedure is particularly costly for heterogeneous tissues that require labeling of multiple small sub-regions, or for tissues with challenging histology. To be transferrable to MSI, the subtissue labeling must use specialized software that takes time to learn. As the result, pathologists often avoid labeling the individual locations, and only roughly annotate the entire tissues. Figure 1c shows that, although the tissue on the left is annotated as tumor, the ion image indicates tissue heterogeneity, and the tissue likely contains both cancerous tissue and healthy kidney parenchyma. Such imprecise labeling of tissue locations compromises the accuracy of the resulting classifiers.

In addition to the labeling, high correlations between many *m*/*z* limit our ability to train accurate classifiers on MSI data. For example, in peptide MSI proteins are digested to give rise to multiple peptide ions of a same protein, and therefore have similar spatial distributions of abundance. An analyte can also produce multiple *m*/*z* ions for other reasons, that include sodium adducts, neutral loss ions, fragment ions or multiply charged ions. For example, Figure 1b illustrates the potential sodium and potassium adducts that give rise to correlated features. The high correlation in the high-dimensional vector of *m*/*z* features undermines the stability of the classifiers, and leads to overfitting (Kriegsmann *et al.*, 2015; Vaysse *et al.*, 2017).

To improve our ability to accurately classify subtissue locations in MSI from approximate tissue-level annotations, we propose a semi-supervised approach *mi-convolutional neural network (CNN)*. mi-CNN implements multiple instance learning (MIL) with a CNN. The multiple instance aspect of the approach enables weak supervision from tissue-level annotations when classifying subtissue labels. The convolutional architecture of the CNN captures potential contextual dependencies between *m*/*z*, such as sodium adducts and dehydrated ions. Evaluations on simulated and experimental datasets demonstrate that mi-CNN improved the subtissue classification as compared with traditional classifiers such as support vector machine (SVM) and CNN, and successfully reflected the truly predictive spectral features. We propose mi-CNN as an important step toward accurate subtissue classification in basic biology and clinical applications of MSI.

2 Background

2.1 Subtissue-level classification in MSI

Classifying tissue locations using MSI spectra has already received a lot of attention (Kriegsmann *et al.*, 2015; Vaysse *et al.*, 2017). Various classifiers have been proposed for these task, including linear discriminant analysis (Dill *et al.*, 2010, 2011), regularized logistic regression (Eberlin *et al.*, 2014; Sans *et al.*, 2017), SVM (Calligaris *et al.*, 2015), and many others. Variations of these approaches such as nearest shrunken centroids (Bemis *et al.*, 2016) incorporate spatial smoothing to enhance the spatial stability of the results. The classifiers take as input *m*/*z* features at each location, classify the label of each location, and classify the tissues according to the majority of its location labels.

Recently, neural networks became of a great interest for MSI. Rauser *et al.* (2010) used fully connected neural networks for tumor classification, and Inglese *et al.* (2017) used unsupervised neural networks to cluster tumor tissues. CNNs, a class of deep neural networks originally designed for image classification, were also introduced. CNN convolutes the image using a small-sized kernel to capture the local connectivity within an image (Rawat and Wang, 2017). A novel application of CNN to MSI proposed to view mass spectra as 1D images. Behrmann *et al.* (2018) used a modified Residual Net with 13 935 parameters and kernel size of 3 to capture isotopic patterns in mass spectra. van Kersbergen *et al.* (2019) replaced convolutional layers in Behrmann's network with dilated convolutional layers to increase receptive size, and capture globally distributed patterns in the spectra.

Although the approaches above are quite diverse, they all rely on quality subtissue labels for training. As the result, they are undermined by training sets with approximate annotations, such as in Figure 1.

2.2 Multiple instance learning (MIL)

Multiple instance learning is a semi-supervised framework commonly used in a variety of applications such as image and video analysis (Cheplygina *et al.*, 2019) and computer-aided diagnosis (Fu *et al.*, 2010; Kandemir and Hamprecht, 2015), but so far not utilized for MSI. In contrast to the classifiers above, MIL allows weak supervision of the training data. The approach considers groups of observations, called *bags*, where ground-truth labels are only available at the bag level. The labels of the observations in a bag, called *instances*, are unknown. In a binary classification problem MIL assumes that a positive bag contains at least one positive instance, but the negative bags contain only negative instances. The homogeneity of the data in the negative bags is the key feature of the approach that enables efficient learning.

Existing MIL algorithms can be classified into two groups: bag space algorithms and instance space algorithm. Bag space algorithms, such as mi-Graph (Zhou *et al.*, 2009) and MIL with instance (Fu *et al.*, 2010), do not predict labels of individual instances. They classify the bags directly by considering similarities of input features between the bags. Instance space algorithms, such as mi-SVM (Andrews *et al.*, 2003) and MILboost (Zhang *et al.*, 2006), take features of the instances as input and predict labels of both instances and bags. For instance-level prediction, mi-SVM is one of the most accurate methods (Kandemir and Hamprecht, 2015). The method treats labels of instances in positive bags as latent variables, and estimates them from the data. Parameters of SVM are optimized by iteratively training SVM on the current instance-level labels, and updating the instance-level labels from their predictions by the current SVM.

2.3 Interpretation of black-box machine learning models

Many of the classification approaches above function like a 'black box' and lack interpretability. Post-processing of these models (Molnar, 2019) helps characterize the relative importance of each predictive feature after the model is fit. One such approach is Local Interpretable Model-agnostic Explanation (LIME; Ribeiro *et al.*, 2016), which ranks features by their importance in predicting the label of a particular observation of interest. LIME generates new observations by permuting the values of the predictive features in the dataset, and obtains the black box predictions for these new observations. Next, LIME weights the new observations by their proximity to the observation of interest, and trains a weighted interpretable model (such as linear regression with subset selection or regularization) on the new observations and their predictions. Finally, LIME repeats this procedure multiple times, and ranks the features by their frequency of being selected as predictive.

3 Multiple instance learning with convolutional neural network (mi-CNN)

3.1 Overview

For the purposes of subtissue classification in MSI, we propose to view a *tissue* as a bag, and a tissue *location* as an instance. We assume that tissues annotated as non-tumor do not have tumor locations, but tissues annotated as tumor can have both tumor and non-tumor locations. MIL allows us to train classifiers of subtissue locations on training sets with such rough tissue-level annotations. Instance space algorithms are of a particular interest for this task. Our proposed approach takes as the baseline mi-SVM, which reported high classification accuracy on similar tasks in the past, but substitutes the SVM classifier with a CNN (Fig. 2). Although CNN are frequently used for image analyses in computer vision domains, the proposed approach uses CNN is a different way. We do not apply spatial convolution on a tissue, as we expect high heterogeneity of the microenvironment within a tumor, and an insufficient spatial smoothness of the location labels. Instead, the CNN incorporates convolutional filters to *m*/*z* in individual locations to capture potential correlations between *m*/*z* of a same location. The CNN has a lightweight structure to avoid overfitting. Finally, post-processing with LIME identifies highly predictive *m*/*z* for downstream biological and clinical interpretation.

3.2 Notation

Consider tissue j and its locations i. The tissue is characterized by a collection of mass spectra $\mathbf{X}_j = \{\mathbf{X}_{ij}\}$, $i = \{1, \ldots, I_j\}$, $j = \{1, \ldots, J\}$, and each mass spectrum $\mathbf{X}_{ij}$ is a vector of M intensities of *m*/*z* features $\mathbf{X}_{ij} = \{X_{ij}^{(1)}, \ldots, X_{ij}^{(M)}\}$. Let $Y_j \in \{0, 1\}$ denote the annotation of the tissue j, and y_{ij} the subtissue label at the ith location. Note that Y_j is known, and y_{ij} is unknown. Denote π_j the probability that tissue j belongs to Class 1, and π_{ij} is the corresponding probability for the location i in that tissue. Given a mass spectrum $\mathbf{X}_{ij}$, our goal is to predict the label y_{ij} of this location, and the label Y_j of the entire tissue.

3.3 Subtissue-level classification

Using cross-entropy as the loss function, the objective of MIL is defined as

$$\max_{\Theta} \sum_{i,j} \{y_{ij} \log(\pi_{ij}) + (1 - y_{ij}) \log(1 - \pi_{ij})\}$$

$$\textbf{such that } \max_{i}(y_{ij}) = Y_j, \qquad (1)$$

where $\pi_{ij} = f(\Theta, \mathbf{X}_{ij})$ is the prediction of a classifier (a CNN) with parameters Θ.

Since the subtissue labels y_{ij} are not observed, they are estimated by an expectation–maximization-like algorithm (Algorithm 1, similar to mi-SVM in Andrews *et al.*, 2003) minimizing the entropy loss [Equation (1)]. First, the labels of all subtissue locations are initialized with the annotations of the corresponding tissues. Next, the algorithm iterates between training CNN on the current location labels, estimating the probability π_{ij} that location i in tissue j belong to Class 1, and imputing the location labels y_{ij} from these probabilities until convergence. The constraint in Equation (1) ensures that the labels of non-tumor locations in non-tumor tissues are always classified as non-tumor. On the other hand, if no locations on a tumor tissue are classified as tumor, the location with the highest π_{ij} in this tissue will be labeled as tumor (Lines 7–10, Algorithm 1). The algorithm stops when the number of updated labels is below a threshold, or when the maximum number of iterations is reached.The architecture of CNN must be adapted to the specifics of the MSI. In these experiments the number of *m*/*z* features can be very large (up to one hundred thousand), while the number of biological replicates is relatively small (typically < 50). Therefore, the CNN should be relatively lightweight, and minimize the number of parameters to avoid overfitting. The convolution filter should be large enough to incorporate neighboring *m*/*z*, but small enough to benefit from weight sharing and computation reduction.

We propose a 1D CNN, consisting of three basic components, namely convolutional layers, pooling layers and fully connected layers. Three convolutional layers hierarchically learn the potential patterns in a mass spectrum. For each layer, the filter size is set according to the contextual dependencies between *m*/*z* of interest, such as mass shifts corresponding to sodium adducts and molecular ions. After each convolutional layer, maxpooling reduces the resolution of the previous layer by focusing on large intensities of *m*/*z* features and reducing the impact of spectral noise. The CNN includes only one fully connected layer that captures globally distributed patterns (Fig. 2). *Softmax* activation

Algorithm 1 *mi-CNN*

```
1: procedure mi-CNN($\mathbf{X}_1, \ldots, \mathbf{X}_J, Y_1, \ldots, Y_J$, threshold)
2:     Initialize: $y_{ij} = Y_j$ for $j \in 1, \ldots, J$
3:     while the number of updated labels < threshold do
4:         Compute CNN parameters $\Theta$ for current labels $y_{ij}$
5:         Compute $\pi_{ij} = f(\Theta, \mathbf{X}_{ij})$
6:         For each $j$ where $Y_j = 1$, set $y_{ij} = \text{ifelse}(\pi_{ij} > 0.5, 1, 0)$
7:         for each $j$ where $Y_j = 1$ do
8:             if $\sum y_{ij} = 0$ then
9:                 Compute $i' = arg\ \max_i \pi_{ij}$
10:                Set $y_{i'j} = 1$
11:            end if
12:        end for
13:    end while
14:    OUTPUT $(\Theta, y_{ij})$
15: end procedure
```

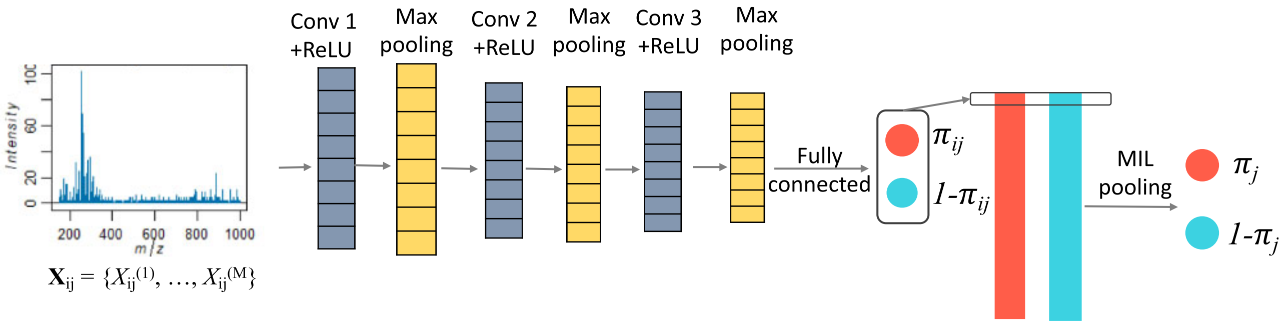

Fig. 2. Architecture of mi-CNN. π_j the probability that tissue j belongs to Class 1, and π_{ij} is the corresponding probability for the location i in that tissue. (**a**) Training set and (**b**) validation set

function is used in the output layer to generate probability of each class.

The CNN is trained using stochastic gradient descent. It calculates the partial derivative of the loss function in Equation (1) with respect to the learnable parameters in Θ by backpropagation, and iteratively updates Θ and values in each layer until convergence.

Tissue-level classification. The proposed tissue-level classification does not count the proportion of predicted location labels in a tissue. Instead, it treats each tissue as one observation, and uses the collection of mass spectra $\{\mathbf{X}_{1j}, \dots, \mathbf{X}_{I_j j}\}$ from all the locations in the tissue as its predictive features. The CNN architecture for this task is the same as the architecture for subtissue-level classification, with the exception of combining the probabilities of the individual locations into a pooling layer that estimates the probability of a tissue-level label. The pooling can be a simple max or mean pooling, or a generalized mean pooling

$$\pi_j(\pi_{1j}, \dots, \pi_{I_j j}) = \left(I_j \sum_i \pi_{ij}^r \right)^{-\frac{1}{r}}, \tag{2}$$

where I_j is the number of locations on tissue j, and r is an integer tuning parameter. The loss function is the cross-entropy of tissue-level predictions and tissue-level labels

$$\max_{\Theta} \sum_j \{ Y_j \log(\pi_j) + (1 - Y_j) \log(1 - \pi_j) \}, \tag{3}$$

where π_j is pooled probability of π_{ij}, and $\pi_{ij} = f(\Theta, \mathbf{X}_j)$ are the predicted probabilities by CNN.

3.4 Evaluation and interpretation

We evaluate the accuracy of subtissue classification by calculating the accuracy and the balanced accuracy of label predictions at individual locations. We evaluate the accuracy of tissue-level classification by calculating the accuracy and the balanced accuracy of label predictions at the entire tissues. The metrics are defined as

$$\text{Accuracy} = \frac{TP + TN}{P + N} \text{ and} \tag{4}$$

$$\text{Balanced accuracy} = \frac{1}{2}\frac{\text{TP}}{\text{P}} + \frac{1}{2}\frac{\text{TN}}{\text{N}}, \tag{5}$$

where for subtissue-level classification, TP is the number of correctly classified positive (i.e. tumor) locations across all the tissues, TN is the number of correctly classified negative (i.e. non-tumor) locations across all the tissues. P and N are the total numbers of locations across the tissues classified as tumor or non-tumor respectively. For tissue-level classification, TP, TN, P and N have the same interpretation, but for the entire tissues. Accuracy quantifies the overall proportion of correct predictions by model. When the number of observations in each class is not balanced, and the prediction of a minority class is under-represented, overall accuracy may inaccurately characterize the performance. In this case, balanced accuracy, quantifying the average of individual proportions of correct predictions in each class may provide more insights.

Even when we can report the accuracy of classification, the classifier remains a black box. Therefore, we use LIME to assist with the interpretation, and identify m/z features that play a particularly important role in classifying the labels of individual locations. We randomly select a subset of locations in the validation sets in our experiments, use LIME to select top five influential features for each location, and rank the selected features by frequency of being selected in multiple locations.

3.5 Implementation

We implemented *mi-CNN* using Tensorflow (Allaire and Tang, 2019) in the RStudio environment. We constructed a CNN architecture of three convolutional layers with Rectified Linear Unit (ReLU) activation, and a fully connected layer. The filter sizes of each convolutional layer were set as 38, 18 and 16. The network had 1774 trainable parameters in total for an input length of 850. CNN were trained using batch stochastic gradient descent optimization. Training one epoch of the renal cell carcinoma (RCC) dataset with 5350 spectra took ~10 s, and training the entire model took ~1.5 h on a computer with 64 RAM and 3.6 GHz CPU. Baseline model mi-SVM was implemented in R following (Andrews *et al.*, 2003). The maximum number of iterations of mi-SVM was set as 200. The kernel function used was radial basis function with gamma as 0.0012 in simulation datasets and human RCC data, and sigmoid function with gamma as 0.00125 in human bladder cancer data. LIME was implemented using R package *lime* (Pedersen and Benesty, 2019). The number of bins for continuous variable was set as 4 and the kernel width was set as 0.1 in *lime*.

4 Data

We evaluated the performance of mi-CNN on five datasets. Two experimental datasets represent two human cancers, and two different MSI acquisition strategies (DESI, characterizing metabolites and lipids and MALDI, characterizing peptides). We further simulated three datasets with known ground truth, inspired from one of the experimental datasets.

4.1 Human RCC experiment

The experiment aimed to classify locations in human renal tissues as tumor versus healthy. Pairs of tumor and healthy tissue sections were collected from eight human donors with RCC. The tissues were subjected to serial H&E staining. Pathology examination of the H&E-stained tissues was unable to classify the tissues at the subtissue resolution, and only annotated each entire tissue section as tumor or healthy (Fig. 3).

Data from the tissues were acquired using DESI ionization source on a Thermo Finnigan LTQ ion trap mass spectrometer in negative mode. The mass range covered 150–1000 Da. In total, 7567 mass spectra were collected from on average 472 locations per tissue. Prior to classification, the spectra were normalized by total ion current (TIC) and resampled to unit mass resolution, which produced 850 m/z features per mass spectrum. The data are available in R package *CardinalWorkflow* (Bemis, 2019). The pairs of tissues were randomly split into a training set (six pairs) and validation set (two pairs).

4.2 Human bladder cancer experiment

The experiment aimed to classify human bladder cancer tissues as tumor versus stroma. Two tissue microarrays (TMAs) containing core needle biopsies from resected formalin-fixed and paraffin-embedded bladder tissues of 49 patients were built, and each TMA was mounted onto a separate glass slide (Fig. 4). A pathologist annotated 42 tissue cores by carefully examining sub-areas of each tissue

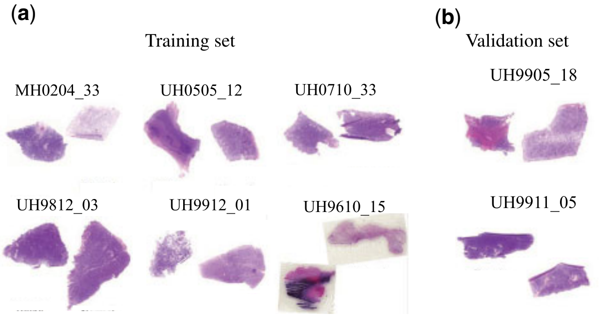

Fig. 3. Human RCC experiment. Pairs of tumor and healthy tissues from eight donors were H&E stained, and examined by a pathologist. For each pair, the tissue on the left has the pathology annotation of tumor, and the tissue on the right has the pathology annotation of healthy. The subtissue-level annotations were not available for this experiment. (a) Training set and (b) Validation set

and color-coded subregions presenting tumor and subregions presenting stroma (Fig. 4). The annotations are viewed as ground truth in this article. The label tumor was assigned to tissue cores containing tumor subregions, and the label stroma to cores containing only stroma.

The proteins in the tissues were digested with trypsin and the peptides were covered with alpha-cyano-4-hydroxycinnamic acid matrix and analyzed with an AB SCIEX 4800 MALDI Time-of-Flight (TOF)/TOF mass spectrometer in positive mode. The mass range was 800–2300 Da. Subregion annotations containing 3152 mass spectra in total and 77 spectra per tissue were extracted via an affine transformation strategy (Föll *et al.*, 2019). The two datasets were resampled, combined and pre-processed using Cardinal and MALDIquant algorithms on https://usegalaxy.eu (Bemis *et al.*, 2015; Föll *et al.*, 2019; Gibb and Strimmer, 2012). The major pre-processing steps comprised peak picking, re-calibration, removal of contaminants and TIC normalization. The pre-processed file contained 593 *m/z* features. Annotated tissues from one slide were used as training set (21 tissues), and on the second slide as the validation set (21 tissues). The split aims to test the robustness of the classifier to experimental batch effects.

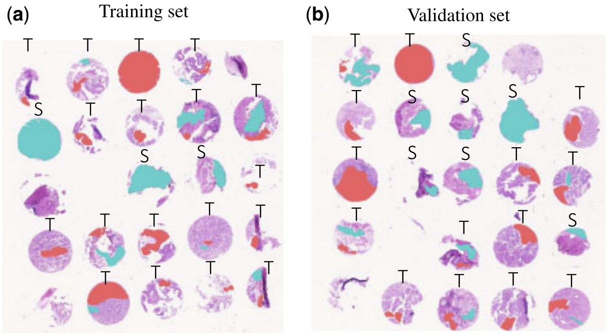

Fig. 4. Human bladder cancer experiment. H&E-stained optical images of human bladder cancer tissues after data acquisition. Letters above each tissue are tissue-level annotations (T, tumor; S, stroma). The colors inside each core indicate subtissue-level pathology (red, tumor; blue, stroma), viewed as the ground truth. (**a**) Training set: 3 purely stroma tissues and 18 tissues with both tumor and stroma locations. (**b**) Validation set: 7 purely stroma tissues and 14 tissues with both tumor and stroma

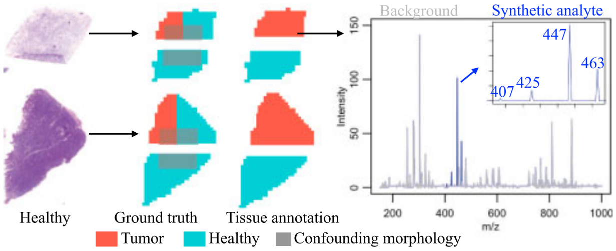

Fig. 5. Simulated Dataset 1. Healthy tissues from the RCC experiment were split into halves. Locations on the left half of the upper newly created tissues were labeled as tumor, and the remaining locations as healthy. The labels were viewed as the ground truth. To mimic pathology annotations, the entire upper tissues were annotated as tumor, and the lower tissues as healthy. A synthetic analyte with four features, differentially abundant between tumor and healthy, was added to the experimental spectra. Its intensity was confounded by a morphology structure spanning both tissue types

Simulated Dataset 1: one differentially abundant analyte with four features, and a complex background. The simulation is based on the mass spectra from eight healthy tissues in RCC dataset. It mimicked real-life variation in feature intensities, while providing the ground truth regarding both the labels of the tissue locations and the predictive features.

First, the eight healthy tissues in the RCC dataset were split into two halves, as shown in Figure 5. Since the mass spectra from these tissues have real-life biological and technological variation, but no systematic variation between the tissue types, they are viewed as a complex background.

Second, the newly created tissues were assigned tissue- and subtissue-level labels. The left half of the upper newly created tissues was labeled as tumor, and the remaining locations as healthy. These labels were viewed as the ground truth. To mimic pathology annotations at the tissue level, the entire upper tissues were annotated as tumor, and the lower tissues as healthy.

Next, one synthetic differentially abundant analyte between the tumor and the healthy locations was added to the experimental spectra. The simulation incorporated a morphology (grey area in Fig. 5) that confounded the intensity of the differentially abundant analyte and spanned both tumor and the healthy tissue locations. The intensity X'_{ij} of this analyte at location i in tissue j was simulated as follows

$$\begin{gathered} X'_{ij} = \mu + S_j + \delta_{ij} + \varepsilon_{ij} \\ S_j \overset{iid}{\sim} \mathcal{N}(0, \sigma_S^2), \varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2),\ \delta_{ij} \overset{iid}{\sim} \mathcal{N}((-\mathrm{I}_{\mathrm{out}} + \mathrm{I}_{\mathrm{in}})\Delta\mu, \sigma_\delta^2)\ ' \end{gathered} \tag{6}$$

where μ is the mean intensity of the analyte for tumor or stroma, S_j is the biological between-tissue variation, δ_{ij} is the variation between the morphological region and background, and ε_{ij} is the biological and technological variation between locations of a same tissue. All the random variables are independent. I_{in} and I_{out} are indicators of whether a tissue location is inside or outside a morphology region, and $\Delta\mu$ is the mean intensity shift of locations inside or outside the morphological region. Here $\mu = 50$ for tumor and $\mu_2 = 150$ for healthy, $\sigma_S = 0.15\mu$, $\Delta\mu = 5$, $\sigma_\delta = 0.1\Delta\mu$ and $\sigma_\varepsilon = 0.1\mu$.

Finally, we simulated four individual *m/z* features generated by this analyte. The features correspond to dehydrated ions $[M-H-H_2O]^-$ (*m/z* 407), molecular molecules $[M-H]^-$ (m/z 425), sodium adducts $[M-2H+N_a]^-$ (*m/z* 447) and potassium adducts $[M-2H+K]^-$ (*m/z* 463). Each feature was simulated as $X_{ij}^{(m/z)} \sim Dirchlet(1,1,1,1) \cdot X'_{ij}$.

Similarly to the RCC dataset, the tissues were split into a training set of six tissue pairs, and a validation set of two tissue pairs.

Simulated Dataset 2: one analyte with differential relative intensity of two of the four features, and a complex background. We mimicked a situation where tumor locations affect the relative intensities of features of a same analyte. We assumed that the synthetic analyte produced more potassium adducts in tumor locations, but more sodium adducts in healthy tissues. The simulation repeated the procedure above, while setting the mean intensity of the analyte to $\mu = 50$ for both tumor and healthy locations, and setting the total intensity of molecular ions and dehydrated ions to $X_{ij}^{(407,425)}$. The total intensity

Table 1. Simulated Dataset 1: classification accuracy

	Compare with	SVM	CNN	mi-SVM	mi-CNN
Training	Tissue annotations	0.895 (0.895)	0.948 (0.948)	0.885 (0.882)	0.751 (0.742)
	Subtissue labels	0.747 (0.833)	0.747 (0.833)	0.809 (0.862)	0.979 (0.981)
Validation	Subtissue labels	0.778 (0.671)	0.752 (0.831)	0.833 (0.693)	0.975 (0.964)

Note: Accuracy [Equation (4)] and balanced accuracy [in parentheses, Equation (5)]. The first two rows evaluate the accuracy with respect to tissue-level annotations. The last four rows evaluate the accuracy with respect to labels of within-tissue locations.

of adducts $X_{ij}^{(447,463)}$ was simulated from $Dirichlet(\alpha = c(1,1)) \cdot X'_{ij}$. Next, in the tumor locations we set the intensity of sodium adducts $X_{ij}^{(447)} = 0.2 \cdot X_{ij}^{(447,463)}$, the intensity of potassium adducts $X_{ij}^{(463)} = 0.8 \cdot X_{ij}^{(447,463)}$. In the healthy locations we set the intensity of sodium adducts $X_{ij}^{(447)} = 0.8 \cdot X_{ij}^{(447,463)}$, and the intensity of potassium adducts $X_{ij}^{(463)} = 0.2 \cdot X_{ij}^{(447,463)}$ in healthy.

Simulated Dataset 3: impact of biological variation, technological variation and sample size. The simulation evaluated the effect of biological and technological variation, and of the number of tissues in the training set, on the performance of mi-CNN. We simulated training sets with between 13 and 130 tissues, half of which annotated at the tissue level as tumor, and the other half as healthy. Each simulated tissue was characterized by 25 locations, with spectra randomly selected from the healthy tissues in the RCC experiment to represent complex background. As in Datasets 1 and 2, only half of the locations in the tumor-annotated tissues had tumor locations as the ground truth. The synthetic analyte was simulated as in Equation (6), with $\mu = 50$ for the tumor locations and $\mu = 150$ for the healthy locations. σ_S varied from 0.1μ to 0.3μ, σ_ε varied between 0.05μ and 0.15μ.

5 Results

5.1 Results for the simulated datasets

Taking as input tissue-level annotations, mi-CNN accurately classified subtissue labels. We compared the ability of mi-CNN and mi-SVM, and that of the classical CNN and SVM, to classify subtissue labels on Simulated Dataset 1. Table 1 shows that SVM and CNN had high accuracy when comparing the classified locations to tissue-level annotations in the training set. This is expected, as the methods were trained to minimize the classification loss with respect to tissue-level annotations. However, these predictive patterns were undermined by the mislabeled healthy locations in the tumor-annotated tissues of the training set. When comparing the classifications to the ground truth at the location level, the methods had worse accuracy (and worse balanced accuracy, that accounts for differences in the number of tumor and healthy locations) in both the training and the validation dataset. Figure 6 details the predictions on the validation set. It illustrates that SVM had poor predictions for both tumor and healthy locations, while CNN had poor predictions for healthy locations.

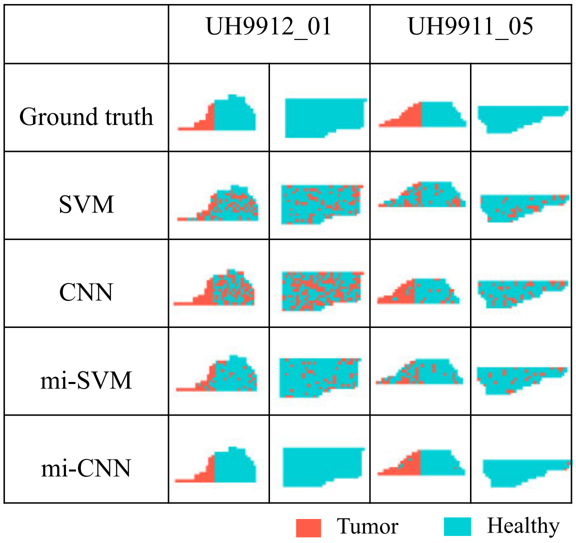

Fig. 6. Simulation Dataset 1: subtissue-level classification on the validation set

Although the accuracy of mi-SVM and mi-CNN classification compared with tissue-level annotations was lower than that of SVM and CNN (Table 1, Row 1 and 2), their results were closer to the ground truth location labels, both on the training (Table 1, Rows 3 and 4) and the validation sets (Table 1, Rows 5 and 6). Figure 6 illustrates that mi-SVM, and in particular mi-CNN, classified the labels of the individual locations more correctly.

Table 2 shows that the results are not limited to situations when the predictive analyte is differentially abundant. Qualitatively similar results are obtained with the predictive pattern in Simulated Dataset 2.

mi-CNN improved subtissue classification by leveraging changes in relative abundances of features from a same analyte. Tables 1 and 2 show that mi-CNN and CNN had higher classification accuracy with respect to the location labels as compared to mi-SVM and SVM. To evaluate whether the improved accuracy was due to the CNN's ability to capture the contextual relationships between related *m/z*, we ranked the predictive features by their importance in these methods using LIME. Figure 7 compares the relative

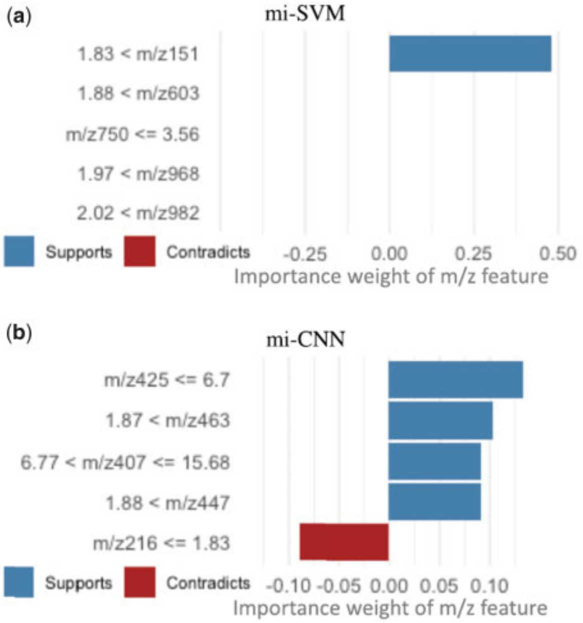

Fig. 7. Simulated Dataset 1: LIME-based importance of *m/z* features when classifying a tumor location in the validation set. A location in the simulated tissue UH9912_01 was classified correctly by both mi-SVM and mi-CNN. However, only mi-CNN captured the four *m/z* features (407, 425, 447 and 463) from the synthetic differentially abundant analyte. (**a**) mi-SVM and (**b**) mi-CNN

Table 2. Simulated Dataset 2: classification accuracy

	Compare with	SVM	CNN	mi-SVM	mi-CNN
Training	Tissue annotations	0.532 (0.530)	0.778 (0.777)	0.860 (0.856)	0.700 (0.690)
	Subtissue labels	0.565 (0.538)	0.734 (0.810)	0.758 (0.776)	0.877 (0.800)
Validation	Subtissue labels	0.530 (0.449)	0.869 (0.896)	0.771 (0.500)	0.912 (0.701)

Note: As Table 1, for Simulated Dataset 2.

importance of the top five features, when classifying a tumor location in one tissue with mi-SVM and mi-CNN. Both methods classified this location correctly. However, while in mi-SVM the most predictive feature is part of the background, mi-CNN ranked the *m*/*z* features (407, 425, 447 and 463) of the synthetic differentially abundant analyte among the top five most predictive.

Out of 200 randomly selected locations, mi-CNN consistently ranked all these features among the top five most predictive in 32.3% of the locations, and at least one of these features among the top five most predictive in 99.3% of the locations. The respective numbers for mi-SVM were very low, 0% and 6%. This illustrates the utility of incorporating the domain-specific information in the size of the convolution filter in the neural network.

In presence of larger variation, accurate subtissue-level classification with mi-CNN required a larger sample size. We evaluated the accuracy of mi-CNN with respect to subtissue labels on Simulated Dataset 3. Figure 8a shows that, in situations where both between-tissue and within-tissue variation is relatively small, mi-CNN can have a high classification accuracy on the validation set, even when trained on a relatively small number of 12 biological replicates. Figure 8c and d illustrates that the between-tissue variation dominates the classification accuracy, and the within-tissue variation has a relatively small impact. Including more biological replicates is beneficial when variation is large.

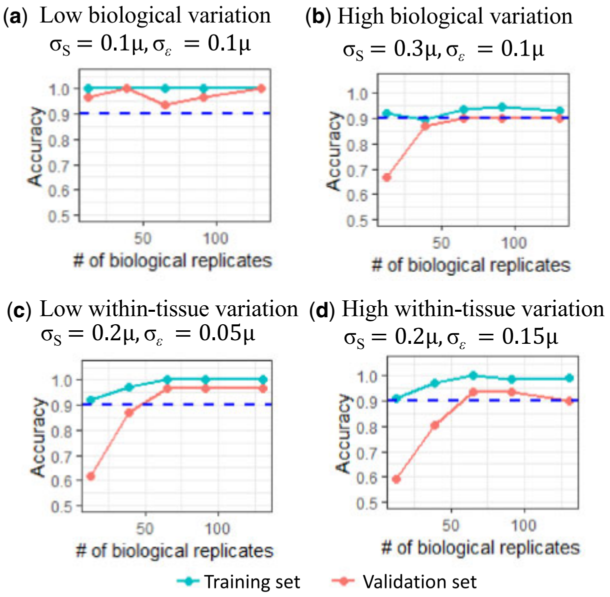

Fig. 8. Simulated Dataset 3: impact of biological (a,b) and technological variation (c,d) of the synthetic analyte, and of the number of training set tissues, on the accuracy of mi-CNN with respect to subtissue labels. When biological variation is relatively small, mi-CNN correctly classified subtissue locations, even with a small number of biological replicates in the training set. Including more biological replicates is beneficial when variation is large

5.2 Results for the experimental datasets

RCC experiment. Although subtissue-level ground truth was not available for the RCC experiment, we used the fact that the tissue sections annotated as healthy were expected to be free from tumor. Therefore, we evaluated the classifications with respect to the homogeneity of subtissue classification of the healthy sections. Figure 9 illustrates that, on the training set, mi-SVM and mi-CNN both had homogeneous predictions of healthy on healthy tissue. On the validation set, mi-CNN had slightly more homogeneous predictions of healthy on healthy tissues than mi-SVM. The predictions of SVM and mi-SVM had no substantial difference in this dataset. CNN has less homogeneous predictions of healthy on healthy tissues than mi-CNN in both training and validation set. This indicates that mi-CNN can improve prediction on healthy locations by considering healthy locations in the tumor tissues.

LIME-based interpretation of mi-SVM and mi-CNN highlighted different features as highly predictive. For mi-SVM, *m*/*z* 181, 215, 760, 865 and 898 were ranked as the top 5 most important. For mi-CNN, these were *m*/*z* 217, 751, 773, 885 and 886. These results indicate that the choice of the classifier plays an important role in both predictive accuracy and the choice of predictive features in this dataset.

Human bladder cancer experiment. Figure 10 compares the classification of SVM, CNN, mi-SVM and mi-CNN with the ground truth subtissue-level labels on selected heterogeneous tumor tissue and pure stroma tissue. Similar to results on Simulated Datasets 1 and 2, SVM and CNN classified many stroma locations in the tumor tissue as tumor in the training dataset (see Fig. 10). Not surprisingly, both SVM and CNN had poor predictions in the validation set, presenting mixture predictions of tumor and stroma in the stroma tissue.

mi-SVM and mi-CNN improved the classification of SVM and CNN in terms of both accuracy and balanced accuracy (Table 3). From Figure 10, mi-CNN correctly classified more stroma locations in the tumor tissues than mi-SVM for both training and validation tissues. In addition, mi-CNN had the smallest number of false positives on the stroma tissues, showing most clean classifications in stroma tissues in Figure 10.

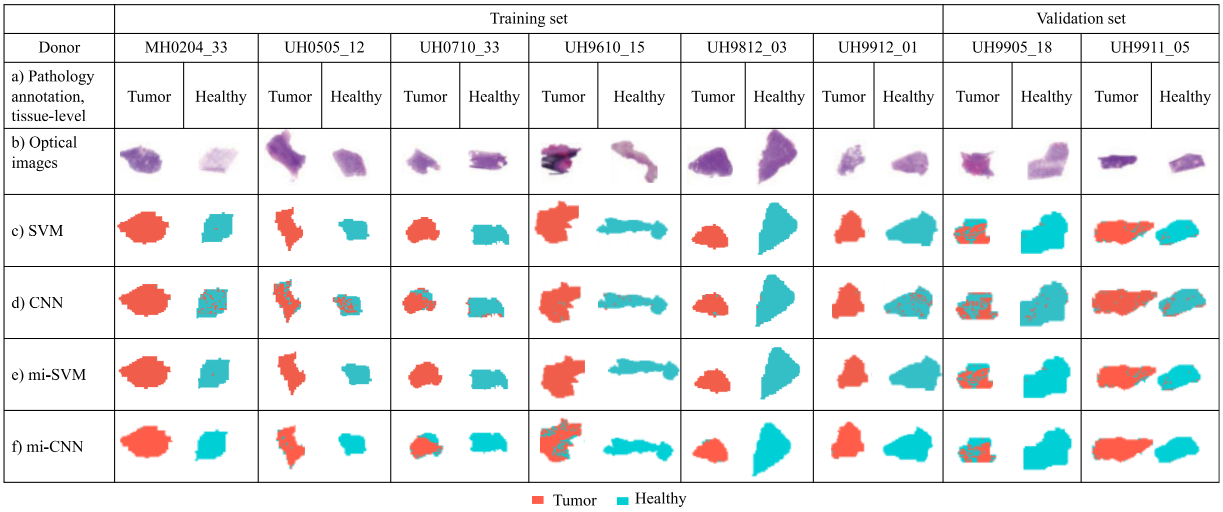

Fig. 9. Classification accuracy: the RCC experiment. (**a**) Tissue-level pathology annotations. (**b**) Optical images of H&E stained tissues. (**c-f**) Subtissue-level classifications

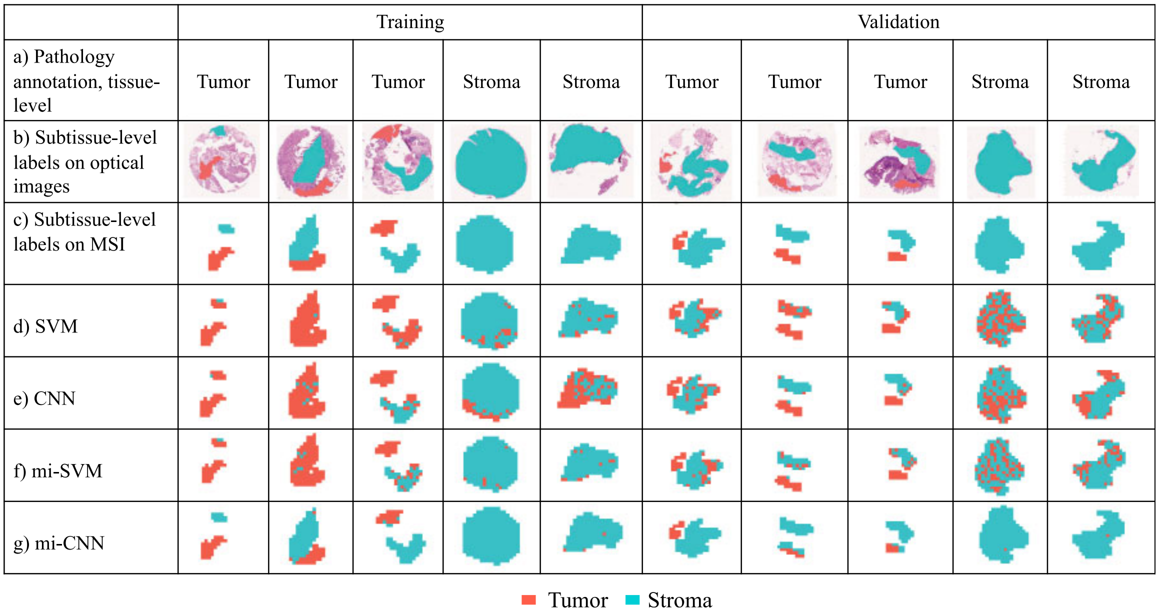

Fig. 10. Classification accuracy: the human bladder cancer experiment. (**a**) Tissue-level pathology annotations. (**b**) Subtissue-level pathology labels on optical images. (**c**) Subtissue-level labels on MSI (viewed as ground truth). (**d–g**) Subtissue-level classifications

Table 3. Classification accuracy: the human bladder cancer experiment

		SVM	CNN	mi-SVM	mi-CNN
Training	Tissue annotations	0.959 (0.946)	0.827 (0.946)	0.939 (0.946)	0.800 (0.855)
	Subtissue labels	0.801 (0.793)	0.767 (0.759)	0.847 (0.842)	0.941 (0.941)
Validation	Subtissue labels	0.755 (0.750)	0.779 (0.774)	0.827 (0.823)	0.928 (0.928)

Note: Values without the parentheses are accuracy calculated by Equation (4). Values in parenthesis are balanced accuracy calculated by Equation (5).

LIME analysis of mi-CNN classifications of a subset of 200 locations in validation set selected *m*/*z* 925.44, 944.44, 946.44, 1105.54 and 1198.69 as most predictive. Among those, *m*/*z* 944.44 is likely to be Histone 2 A, which is known to be upregulated in tumors, and *m*/*z* 1105.54 is likely to correspond to Collagen I which is known to be upregulated in stroma. LIME analysis of mi-SVM selected five different predictive features, *m*/*z* 1669.73, 1475.72, 1529.7, 963.44 and 1054.49.

6 Discussion

We introduced mi-CNN, a deep MIL approach for classifying subtissue locations in MSI experiments. The multiple instance aspect of the approach enabled training the classifier with weak supervision, using rough tissue-level annotations in the training set. The convolutional architecture of the CNN captured contextual dependencies between the spectral features. Evaluations on simulated and experimental datasets demonstrated that mi-CNN improved the subtissue classification as compared with traditional SVM and CNN.

The approach assumed that, in a binary classification problem, a tissue labeled as tumor had at least one tumor location, but the tissues labeled as non-tumor were tumor-free. This assumption is reasonable for MSI, as homogeneous healthy tissue biopsies are relatively easy to obtain, however tumor biopsies are more likely to contain a mix of tumor and non-tumor regions. In a case where both non-tumor and tumor tissues are heterogeneous, the proposed approach is no longer suitable since the reliable label of non-tumor is crucial to the method. Although we only discussed binary classification, mi-CNN can also be adapted to multi-class classification, such as different grades of tumor tissues or multiple tissue types.

In contrast to the typical applications of CNN in computer vision, the CNN architecture in this work did not include spatial convolution of tissues. This is a consequence of typically high heterogeneity of the microenvironment within a tumor, and of lack of spatial smoothness of location labels.

At the same time, the CNN architecture took advantage of the mass spectral patterns to alleviate the high dimensionality and the high correlations in the predictive feature space. In this work, the size of convolutional filters captured one of the most common sources of correlations between *m*/*z*, i.e. the presence of molecular ions and their adducts. The *m*/*z* dependencies can become more complicated and ambiguous in other cases, e.g. with larger mass ranges. The convolutional aspects can be easily adapted to such situations types by changing the size of filter and the network depth.

Although neural networks have a large parameter space and need large training datasets, we found that mi-CNN worked well on the relatively small numbers of biological tissues. This may be due to a combination of the CNN architecture, which uses locally connected neurons and weight sharing filters to reduce the parameter space and the computational cost, and a relatively large number of heterogeneous subtissue locations available for training.

Overall, we found that mi-CNN is well-suited for training subtissue-level classifiers on datasets with tissue-level annotations. This is particularly important in situations where tumor and non-tumor tissues are tightly connected, making manual labeling of the training sets difficult or even impossible at all. The approach is an important step toward taking a full advantage of MSI's capability of providing molecular information, and minimizing manual labor for tissue imaging and classification.

Acknowledgement

We thank K. A. Bemis for helpful discussions.

Funding

The work was supported in part by the NSF [award DBI-1759736 to O.V.]; and by the German Research Council [DFG, SCHI 871/11-1 to O.S.].

Conflict of Interest: none declared.

References

Aichler,M. and Walch,A. (2015) MALDI imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. *Lab. Invest.*, **95**, 422–431.

Allaire,J.J. and Tang,Y. (2019) *Tensorflow: R Interface to 'TensorFlow'*. R package version 2.0.0.

Andrews,S. *et al.* (2003) Support vector machines for multiple-instance learning. In: *Advances in Neural Information Processing Systems*, p. 577.

Behrmann,J. *et al.* (2018) Deep learning for tumor classification in imaging mass spectrometry. *Bioinformatics*, **34**, 1215–1223.

Bemis,K. *et al.* (2015) Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments. *Bioinformatics*, **31**, 2418–2420.

Bemis,K. *et al.* (2016) Probabilistic segmentation of mass spectrometry images helps select important ions and characterize confidence in the resulting segments. *Mol. Cell. Proteomics*, **15**, 1761–1772.

Bemis,K.A. (2019) *CardinalWorkflows: Datasets and Workflows for the Cardinal Mass Spectrometry Imaging Package*. R package version 1.17.0.

Calligaris,D. *et al.* (2015) MALDI mass spectrometry imaging analysis of pituitary adenomas for near-real-time tumor delineation. *Proc. Natl. Acad. Sci. USA*, **112**, 9978–9983.

Cheplygina,V. *et al.* (2019) Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.*, **54**, 280–296.

Dill,A.L. *et al.* (2010) Multivariate statistical differentiation of renal cell carcinomas based on lipidomic analysis by ambient ionization imaging mass spectrometry. *Anal. Bioanal. Chem.*, **398**, 2969–2978.

Dill,A.L. *et al.* (2011) Multivariate statistical identification of human bladder carcinomas using ambient ionization imaging mass spectrometry. *Chemistry*, **17**, 2897–2902.

Eberlin,L.S. *et al.* (2014) Molecular assessment of surgical-resection margins of gastric cancer by mass-spectrometric imaging. *Proc. Natl. Acad. Sci. USA*, **111**, 2436–2441.

Föll,M.C. *et al.* (2019) Reproducible mass spectrometry imaging data analysis in Galaxy. *GigaScience*, **8**, 628719.

Fu,Z. *et al.* (2010) MILIS: multiple instance learning with instance selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, **33**, 958.

Gibb,S. and Strimmer,K. (2012) MALDIQUANT: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*, **28**, 2270–2271.

Inglese,P. *et al.* (2017) Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer. *Chem. Sci.*, **8**, 3500–3511.

Jones,E.A. *et al.* (2012) Imaging mass spectrometry statistical analysis. *J. Proteomics*, **75**, 4962–4989.

Kandemir,M. and Hamprecht,F.A. (2015) Computer-aided diagnosis from weak supervision: a benchmarking study. *Comput. Med. Imaging Graph.*, **42**, 44–50.

Kriegsmann,J. *et al.* (2015) MALDI TOF imaging mass spectrometry in clinical pathology: a valuable tool for cancer diagnostics (review). *Int. J. Oncol.*, **46**, 893–906.

Lou,S. *et al.* (2017) An experimental guideline for the analysis of histologically heterogeneous tumors by MALDI-TOF mass spectrometry imaging. *Biochim. Biophys. Acta Proteins Proteom.*, **1865**, 957–966.

Molnar,C. (2019) *Interpretable Machine Learning*. Lulu. com.

Pedersen,T.L. and Benesty,M. (2019) *LIME: Local Interpretable Model-Agnostic Explanations*. R package version 0.5.1.

Rauser,S. *et al.* (2010) Classification of HER2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry. *J. Proteome Res.*, **9**, 1854–1863.

Rawat,W. and Wang,Z. (2017) Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.*, **29**, 2352–2449.

Ribeiro,M.T. *et al.* (2016) "Why should I trust you?" Explaining the predictions of any classifier. In: *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 1135.

Sans,M. *et al.* (2017) Metabolic markers and statistical prediction of serous ovarian cancer aggressiveness by ambient ionization mass spectrometry imaging. *Cancer Res.*, **77**, 2903–2913.

Spengler,B. (2015) Mass spectrometry imaging of biomolecular information. *Anal. Chem.*, **87**, 64–82.

van Kersbergen,J. *et al.* (2019) Cancer detection in mass spectrometry imaging data by dilated convolutional neural networks. In: *Medical Imaging 2019: Digital Pathology*, Vol. **10956**, p. 109560I, International Society for Optics and Photonics.

Vaysse,P.M. *et al.* (2017) Mass spectrometry imaging for clinical research-latest developments, applications, and current limitations. *Analyst*, **142**, 2690–2712.

Wu,C. *et al.* (2013) Mass spectrometry imaging under ambient conditions. *Mass Spectrom. Rev.*, **32**, 218–243.

Zhang,C. *et al.* (2006) Multiple instance boosting for object detection. In: *Advances in Neural Information Processing Systems*, p. 1417.

Zhou,Z.-H. *et al.* (2009) Multi-instance learning by treating instances as non-IID samples. In: *Proceedings of the 26th Annual International Conference on Machine Learning*, p. 1249.