Received: 9 August 2021 | Revised: 24 January 2022 | Accepted: 28 February 2022 | Accepted article online: 3 March 2022

DOI: 10.1002/prca.202100068

Proteomics Clinical Applications

RESEARCH ARTICLE

Robust subtyping of non-small cell lung cancer whole sections through MALDI mass spectrometry imaging

Charlotte Janßen[1] | Tobias Boskamp[2,1] | Lena Hauberg-Lotte[1] | Jens Behrmann[1] | Sören-Oliver Deininger[2] | Mark Kriegsmann[3,4] | Katharina Kriegsmann[5] | Georg Steinbuß[5] | Hauke Winter[4,6] | Thomas Muley[4,7] | Rita Casadonte[8] | Jörg Kriegsmann[8,9] | Peter Maaß[1]

[1]Center for Industrial Mathematics (ZeTeM), University of Bremen, Bremen, Germany

[2]Bruker Daltonics GmbH, Bremen, Germany

[3]Institute of Pathology, University Hospital Heidelberg, Heidelberg, Germany

[4]Translational Lung Research Center Heidelberg (TLRC), Member of the German Center for Lung Research (DZL), Heidelberg, Germany

[5]Department of Hematology, Oncology and Rheumatology, University Hospital Heidelberg, Heidelberg, Germany

[6]Department of Thoracic Surgery, Thoraxklinik, Heidelberg University Hospital, Heidelberg, Germany

[7]Translational Research Unit, Thoraxklinik, Heidelberg University Hospital, Heidelberg, Germany

[8]Proteopath, Trier, Germany

[9]Center for Histology, Cytology and Molecular Diagnostic, Trier, Germany

Correspondence
Dr. Charlotte Janßen and Tobias Boskamp, Center for Industrial Mathematics (ZeTeM), University of Bremen, Bibliothekstraße 5, 28359 Bremen, Germany.
Email: cjanssen@uni-bremen.de; tboskamp@uni-bremen.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Number: FKZ 031L0198A; Klaus Tschira Stiftung, Grant/Award Number: 00.010.2019

Abstract

Subtyping of the most common non-small cell lung cancer (NSCLC) tumor types adenocarcinoma (ADC) and squamous cell carcinoma (SqCC) is still a challenge in the clinical routine and a correct diagnosis is crucial for an adequate therapy selection. Matrix-assisted laser desorption/ionization (MALDI) mass spectrometry imaging (MSI) has shown potential for NSCLC subtyping but is subject to strong technical variability and has only been applied to tissue samples assembled in tissue microarrays (TMAs). To our knowledge, a successful transfer of a classifier from TMAs to whole sections, which are generated in the standard clinical routine, has not been presented in the literature as of yet.

We introduce a classification algorithm using extensive preprocessing and a classifier (either a neural network or a linear discriminant analysis (LDA)) to robustly classify whole sections of ADC and SqCC lung tissue. The classifiers were trained on TMAs and validated and tested on whole sections. Vital for a successful application on whole sections is the extensive preprocessing and the use of whole sections for hyperparameter selection.

The classification system with the neural network/LDA results in 99.0%/98.3% test accuracy on spectra level and 100.0%/100.0% test accuracy on whole section level, respectively, and, therefore, provides a powerful tool to support the pathologist's decision making process. The presented method is a step further towards a clinical application of MALDI MSI and artificial intelligence for subtyping of NSCLC tissue sections.

KEYWORDS
deep learning, lung cancer, mass spectrometry imaging, non-small cell lung cancer, whole sections

Abbreviations: ADC, adenocarcinoma; IPN, intensity profile normalization; LDA, linear discriminant analysis; MALDI MSI, matrix-assisted laser desorption/ionization mass spectrometry imaging; NSCLC, non-small cell lung cancer; SqCC, squamous cell carcinoma; TMA, tissue microarray

1 | INTRODUCTION

Adenocarcinoma (ADC) and squamous cell carcinoma (SqCC) are the most common types of non-small cell lung cancer (NSCLC) and the most common lung cancer types in general (~70%). An accurate diagnosis is essential for an effective patient treatment, but, in some cases, pathologists need to acquire several immunohistochemical (IHC) stainings of the tissue to make a robust diagnosis. IHC stainings, however, are costly and time and tissue consuming and there is not always enough tissue available for subsequent molecular analysis.

The use of matrix-assisted laser desorption and ionization (MALDI) mass spectrometry imaging (MSI) has been emerging as basis for tumor classification in recent decades [1, 2, 3]. Classical machine learning techniques have been used to classify MALDI MSI measurements of cancerous lung tissue [4, 5, 6] and other tumor types [7, 8, 9]. A first neural network for the classification of MALDI peptide imaging data was introduced by Behrman et al. [10].

To successfully train a classifier, especially a neural network with its many degrees of freedom, it is important to obtain a large dataset with enough biological variation. Tissue samples of as many patients as possible are required. Usually, tissue samples of many patients are collected in tissue microarrays (TMAs) to mimic a biopsy situation and to achieve homogeneous experimental conditions for all measured samples. Subtyping of cancerous lung tissue based on MALDI MSI measurements has been successfully performed on TMAs [11, 12]. The constructions of TMAs, however, is not part of the standard clinical process. Applying a classifier, trained on TMAs, to MALDI MSI measurements of whole sections can be difficult due to technical variations in the measurement process. While a somewhat similar approach has been recently taken to translate a classifier developed on melanoma biopsies to TMAs [13], a successful automatic classification of tissue whole sections from a model developed on TMAs has not been presented yet.

Here, for the first time, we present a classification system that robustly classifies whole sections of ADC and SqCC lung tissue based on MALDI MSI measurements. The classification system consists of an extensive preprocessing and a classification through a neural network or a linear discriminant analysis (LDA). There are two crucial points for achieving a successful classification of whole sections: an extensive preprocessing of the MALDI MSI spectra and the incorporation of whole sections into the process of choosing the classifier's hyperparameters. An independent test dataset consisting of whole sections was used for the evaluation of the algorithm.

Clinical Relevance

Correct subtyping of non-small cell lung cancer (NSCLC) is crucial for therapeutic decisions but may require several immunohistochemical stainings, which can be challenging due to limited available tissue. Previous studies presented successful classifications of cancerous lung tissue based on matrix-assisted laser desorption/ionization (MALDI) mass spectrometry imaging (MSI) measurements collected on tissue microarrays (TMAs). Technical variations in MALDI MSI measurements, however, often hamper the transfer to whole sections, thus preventing this method from being translated into clinical routine. We present a classification algorithm that provides a 100% accuracy on the 16 NSCLC whole sections in our test dataset. Our approach, therefore, provides a significant step towards an application of MALDI MSI and artificial intelligence in a clinical routine for the support of pathologists.

2 | MATERIALS AND METHODS

2.1 | MALDI MSI dataset

2.1.1 | Tissue samples

Tissue cores of NSCLC were assembled on six TMAs at the Tissue Biobank from the National Center of Tumor Diseases (NCT). Allocation of cores to TMAs was randomized, thus all TMAs represent similar mixtures of tumor types and clinical characteristics. Tissue sections of the TMAs (stained with hematoxylin and eosin [H&E], CK5/6, TTF1, Napsin and p40) were scanned at 40x magnification with a slide scanner (Aperio AT2) and provided by the Institute of Pathology, Heidelberg University, in accordance with the local ethics committee (ethics committee number S315/2020). Additionally, a diagnosis for each core was provided. We excluded all tissue cores classified as neither ADC nor SqCC and all tissue cores with an unclear diagnosis indicated by ambiguous IHC staining results (CK5/6- and p40-negativ SqCC cores and TTF1-negative ADC cores). The remaining data comprised $N = 179$ cores of ADC and $N = 223$ cores of SqCC from a total of 201 patients (2 cores per patient). Additionally, 30 H&E stained whole sections of lung tissue (15 of ADC and 15 of SqCC) from 30 patients were provided by the Institute of Pathology, Heidelberg University. Overview images of the datasets can be found in the Supporting Information (Figures S1 and S2).

Areas with high tumor cell content, low amount of necrosis and high scan quality were annotated by a thoracic pathologist (MK). Each spectrum originates from a specific point of the tissue (50 μm resolution). Spectra outside of the annotated regions were excluded from the subsequent analysis, resulting in 66,183 spectra (31,942 SqCC, 34,241 ADC) from the six TMAs, and 161,969 spectra (87,293 ADC, 74,676 SqCC) from the whole sections.

2.1.2 | Tissue preparation and MALDI MSI measurements

The preparation of the tissue and MALDI MSI measurements are described in detail in a previous study [12]. In brief, the tissue sections

were mounted on indium tin oxide-coated glass slides. After dewaxing and heat-induced antigen retrieval a trypsin digestion step (0.025 μg/μl final concentration) with an automatic reagent sprayer (TM-sprayer, HTX Technologies, Chapel Hill, NC, USA) was applied, followed by matrix application (10 mg/ml alpha-cyano-4-hydroxycinnamic acid) with the same spraying device. MALDI MSI was performed using a rapiflex MALDI tissuetyper (Bruker Daltonics) operated in positive reflector mode. MSI measurements were acquired with flexImaging (version 5.0, Bruker Daltonics) and the spectra were generated with flexControl (version 4.0, Bruker Daltonics). After MSI, the matrix was removed from the tissue sections and the tissue was H&E stained using standard protocols.

The data was imported into the SCiLS Lab software (version 2018b, Bruker Daltonics) and a baseline correction was performed using a convolutional algorithm. The MALDI MSI data was accessed via the SCiLS Lab API (version 1.0.554, Bruker Daltonics) and imported into Python for further preprocessing and the subsequent classification.

2.2 | Classification process

The MALDI MSI data were analyzed by two different classification algorithms: a state-of-the-art neural network designed for an application to MALDI peptide imaging data, and a classical machine learning technique, which consists of a simple feature selection through a receiver operating characteristic (ROC) analysis in combination with an LDA [14]. Previous to both classifications, a preprocessing pipeline is applied, which is explained in more detail below.

2.2.1 | Training, validation and test dataset

The spectra from the six TMAs were used for training the classifiers and are referred to as training dataset in the following. We randomly chose seven ADC and seven SqCC whole sections as validation dataset and kept the remaining eight ADC and eight SqCC sections as independent test dataset. Note that the test dataset does not contain data from the same patients as the validation data as the division is done section-wise. The distinction into validation and test dataset is important as the validation dataset is used in the model selection process, that is, it is used for the tuning of the neural network's hyperparameters and choice of epoch, and for the choice of number of features in the LDA classification. A test dataset is therefore required to validate the algorithm on independent data.

2.2.2 | Preprocessing pipeline

MALDI data is subject to extensive technical variability that can harm the classification process [15]. Here, we used a preprocessing pipeline aiming at reducing technical artefacts and other noise. The pipeline has shown great improvement for the classification of MALDI MSI data with classical machine learning techniques [16]. It consists of the following steps:

- reduction of the m/z range to 700–2700 Da,
- an intensity profile normalization (IPN), which aligns the distribution of spectral intensities with a reference distribution (ipn) [17],
- a statistical recalibration to reduce mass shifts (cal) [18],
- a resampling that is specifically designed for MALDI peptide imaging data reducing the mass resolution to approximately 1 Da (peptide mass resampling - pmr) [17],
- a gaussian (spatial) smoothing with a radius of 200 μm (smooth),
- a second intensity profile normalization [17],
- a log transformation of the intensities (log).

The abbreviations of the individual preprocessing steps are later used to indicate different versions of the pipeline. The full pipeline corresponds to ipn-cal-pmr-smooth-ipn-log. The preprocessing pipeline was first applied to the training data and its settings and the reference profiles computed for the IPN were saved. The exact same preprocessing was, subsequently, applied to the spectra from the whole sections. In future applications of our algorithm to other MALDI MSI measurements of lung tissue it is important to use the same preprocessing settings as the classifiers were trained to classify spectra distributions resulting from this specific preprocessing.

2.2.3 | LDA

We used linear discriminant analysis (LDA) as a state-of-the-art classifier, serving as a baseline to which to compare the classification results obtained through the neural network. Being one of the best understood and most traditional supervised machine learning techniques, LDA was chosen for its structural simplicity and the absence of algorithmic hyperparameters that would need to be tuned.

As a first step, a feature selection through a ROC analysis is performed [19]. The ROC analysis examines the discriminative relevance of each single m/z-value, that is, it provides a measure for how different the intensities of a certain m/z-value are in spectra of one class compared to spectra from the other class. To this end, the area under the ROC curve (AUC) is computed for each m/z-value. Based on the absolute ROC score $r = |\text{AUC} - 0.5|$, the 200 most discriminative m/z-values (features) are chosen for the subsequent classification. LDA classifiers are trained on the first n features for all $n = 1, \ldots, 200$. The final classifier, given by the final number of features n, is chosen according to the balanced accuracy of the classifiers on the validation dataset.

2.2.4 | Neural network

The neural network used in this study maps the preprocessed spectra onto a probability for each class. The class with the maximum probability is chosen as prediction. It is termed IsotopeNet and is an adapted version of that presented by Behrman et al. [10]. A more detailed

TABLE 1 Evaluation of classification with neural network and LDA on spectra level

	Bal. accuracy	Sensitivity (SqCC)	Sensitivity (ADC)
	Neural network		
Training	0.982	0.984	0.981
Validation	0.986	0.992	0.981
Test	0.990	0.989	0.991
	LDA		
Training	0.944	0.926	0.970
Validation	0.954	0.969	0.938
Test	0.983	0.980	0.986

Balanced accuracy (Bal. accuracy) is the arithmetic mean of the sensitivities of both classes. Note that in a two-class classification problem, the sensitivity of one class corresponds to the specificity of the other.

TABLE 2 Confusion matrices of classification with neural network and LDA on core/ whole section level (Pred. = Prediction)

True\ Pred.	Training (cores)		Validation (whole sections)		Test (whole sections)	
	SqCC	ADC	SqCC	ADC	SqCC	ADC
	Neural network					
SqCC	217	6	7	0	8	0
ADC	4	175	0	7	0	8
	LDA					
SqCC	202	21	7	0	8	0
ADC	8	171	1	6	0	8

Diagonal elements of the matrices show number of correctly classified cores/sections, off-diagonal elements show number of misclassified cores/whole sections.

introduction to neural networks, the training process and the network's architecture is given in the Supporting Information S.2. Neural networks, due to their many degrees of freedom, are prone to overfitting where the network learns features specific to the training data and does not generalize well to unseen data. Overfitting is usually indicated by a high training accuracy (i.e., accuracy on the training dataset) compared to a lower validation accuracy (i.e., accuracy on the validation dataset). To avoid overfitting, we implemented several standard regularization methods and performed a hyperparameter tuning (see Supplementary Information S.2). As the final setting of the hyperparameters, we chose values that resulted in the highest balanced accuracy on the validation dataset. Each training comprises 30 epochs (in one epoch the training dataset is shown to the network once). The final network is chosen from the epoch with the highest validation accuracy.

We repeated the training with the final neural network five times as the training process involves the generation of random numbers for the initialization of the network and the batch sampling of the dataset. The resulting best network in terms of balanced accuracy on the validation dataset was chosen for further analysis. A measure of the influence of random numbers in terms of standard deviation can be found in the Supporting Information (Table S3).

3 | RESULTS AND DISCUSSION

Both classifiers (neural network and LDA) achieved high accuracies on the training and validation dataset, as well as on the independent test dataset (Table 1). The final goal in a clinical application is the classification of whole sections and not of single spectra. We, therefore, define the section-level classification result as the class with the larger number of spectra in one section (majority voting). Both classifiers were able to robustly subtype whole sections of ADC and SqCC (Table 2). The neural network classified all 14 validation sections and 16 test sections correctly. The LDA classifier misidentified only one section of the validation dataset.

To ensure a reliable classification of the whole sections, we propose to use a simple quality control, as commonly used in similar studies. If the percentage of spectra in the majority class for one section falls below a certain threshold, the prediction is considered less reliable or inconclusive. In this case, a warning to the pathologist could be issued, suggesting a particularly careful review of this section and its prediction result. For this purpose, the classification results can be visualized using a heatmap (Figure 1 and Figure S2 in the Supporting Information).

The choice of an appropriate threshold depends on the respective context, for example prior probabilities or cost of misclassification. The canonical choice would be 66.6%, resulting in equally sized percentage intervals for the three possible outcomes correct, incorrect, and inconclusive. In our study, a 90% threshold was used, thus further raising the impact of the quality control.

With the neural network classifier, three sections did not pass the quality control (two in validation, one in test set), although by majority voting all were classified correctly. In case of the LDA classifier, five sections did not pass (three in validation, two in test set), one of which was the one section classified incorrectly by majority voting (Table 3).

Interestingly, all sections caught by the quality control following the neural network classification, are also caught in the quality control following the LDA classification. This suggests that the ambiguous classification result of these sections is not due to classifier-specific deficits but that spectra from these sections have unusual patterns resulting from unusual technical or biological variability. The spectral features that usually lead to a distinct classification are apparently not prevalent in these sections. There may be other features in these sections that are indicative of the underlying tumor type but they were not sufficiently learned by the classifiers from the training data, either because the specific patterns were not available in the training dataset or because of non-optimal learning.

In total, the neural network provides slightly more accurate results than the LDA classifier. This is most visible in the accuracies per section (Table 3). The validation and test dataset include data from only 30 patients, whereas the training data contains data from 201 patients.

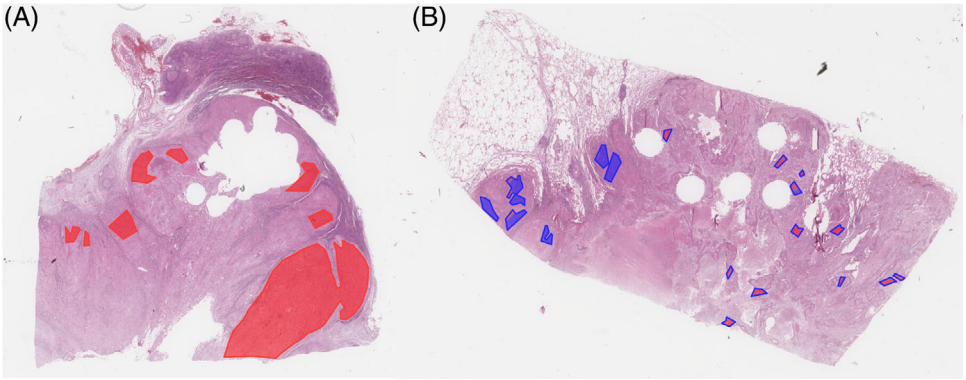

FIGURE 1 Optical image of sections (A) ADC2 and (B) SqCC7 with overlaid annotations of tumor areas (lines) and heatmaps (filling) of the neural network classification. Heatmaps show the location of spectra in the respective color they are classified as. The colors red/blue correspond to ADC/SqCC, respectively. Heatmaps of all test sections can be found in the Supporting Information (Figure S2).

TABLE 3 Accuracies per section

Validation dataset			Test dataset		
	NN	LDA		NN	LDA
ADC3	99.97%	98.41%	ADC1	99.79%	98.64%
ADC6	100.00%	99.96%	ADC2	100.00%	100.00%
ADC7	99.73%	100.00%	ADC4	100.00%	99.92%
ADC8	100.00%	100.00%	ADC5	100.00%	99.22%
ADC9	**66.09%**	***32.58%***	ADC10	100.00%	99.14%
ADC13	98.98%	**83.83%**	ADC11	91.35%	**84.53%**
ADC15	100.00%	98.96%	ADC12	99.20%	99.35%
SqCC1	100.00%	100.00%	ADC14	99.42%	100.00%
SqCC4	**87.41%**	**84.82%**	SqCC2	100.00%	92.27%
SqCC5	100.00%	100.00%	SqCC3	100.00%	100.00%
SqCC8	100.00%	100.00%	SqCC6	100.00%	97.45%
SqCC10	100.00%	94.17%	SqCC7	**62.60%**	**72.87%**
SqCC13	99.91%	95.61%	SqCC9	100.00%	99.76%
SqCC14	100.00%	99.58%	SqCC11	100.00%	100.00%
			SqCC12	100.00%	100.00%
			SqCC15	100.00%	99.60%

Bold letters highlight sections that did not pass the quality control (i.e., less than 90% of spectra of majority tumor type). Italic letters highlight the only section that is classified incorrectly (ADC9 - LDA classification). Note that this section would be caught by the quality control as designated.

The slightly higher accuracies on the training dataset of the neural network compared to the LDA (Table 1 and 2) might indicate that the biological variations in the training dataset are better captured by the neural network, or that the neural network is able to better handle noise and other artefacts in the training data. The training of a neural network and choice of its hyperparameters, however, are far more challenging and time consuming than the training of a LDA. While this is not relevant for the clinical application itself (as the trained classifier would be provided), this is an important factor for future studies in this area.

Additionally, the influence of random numbers on the training process of the neural network and especially on the classification success on the test dataset is high (see Table S3). Even though the network we chose, based on the best validation accuracy, provides a successful classifier in our case, it cannot be guaranteed that a reproduction of the experiments will result in a similarly successful classifier with high test accuracies. The LDA classifier on the other hand is not influenced by random effects and its results can be reproduced accurately.

On our system, the evaluation of 5400 spectra (approximate average number of spectra within annotated tumor areas in one whole section in our dataset) with a neural network on 4 GPUs takes about 3 s, with the LDA classifier it takes below one second on the CPU. The preprocessing takes about 233 s. The computations were done on a 20-core processor (Intel Xeon Silver 4210), and 4 NVIDIA GeForce GTX 2080 Ti. Both evaluation times are only a small addition to the MALDI data preparation and measurement, which take several hours. For comparison, IHC staining including all necessary preparation steps takes about 1–2 h per section as well.

3.1 | Using whole sections for validation

There are two points that were crucial to achieve a robust and accurate classification of the whole sections: first, to use whole sections in the validation dataset and, second, the extensive preprocessing.

Including whole sections in the validation dataset means that they were used in the model selection process. They served as an indicator of overfitting, for the choice of epoch and for the choice of hyperparameters. The hyperparameters of the neural network include the learning rate, batch size, weight decay parameter, and the amount of dropout. For the LDA classifier, the validation dataset was used to find the best number of features used for the classification model. Note that the final accuracy of the classification algorithm is determined on the independent test dataset.

Figure 2A shows that during the training of the neural network high accuracies on training, validation, and test dataset were achieved already after the first epoch. Note that one epoch consists of many

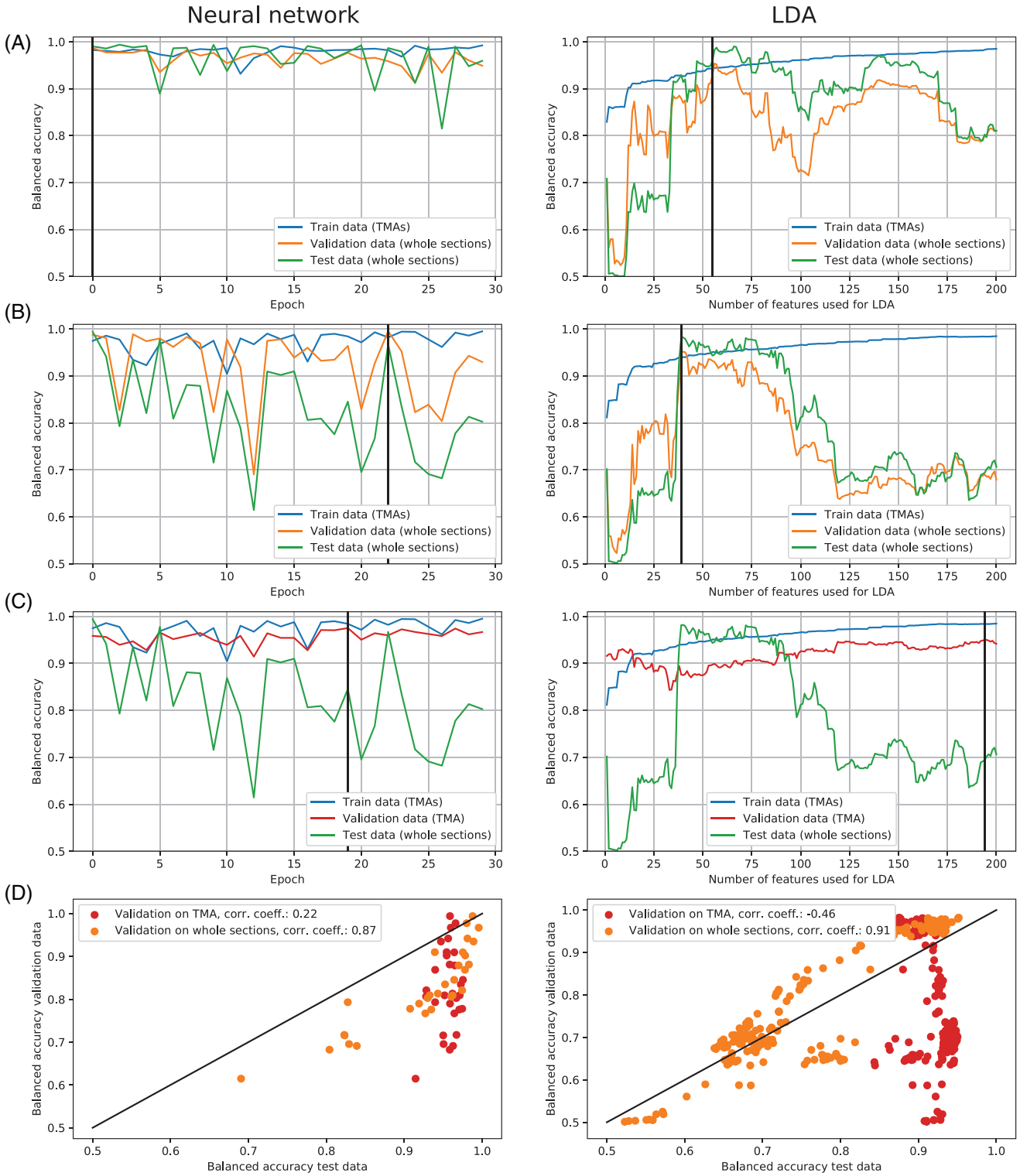

FIGURE 2 Balanced accuracies on training (blue), validation (orange/red) and test dataset (green) of the neural network and the LDA with the full pipeline over epochs/number of features (A) for validation on whole sections, (B) trained on only 5 TMAs and validated on whole sections and (C) the identical training as in (B) but with validation on the remaining TMA instead of whole sections, and (D) correlation between validation and test accuracies for validation on one TMA and validation on whole sections (corr. coeff. = correlation coefficient). The vertical lines in A-C indicate the respective model with the highest validation accuracy, which is selected for further analysis.

steps in the gradient descent algorithm as we use a batch-wise stochastic gradient descent. The validation and test accuracies vary from epoch to epoch indicating a smaller or larger amount of overfitting. Both are correlated, suggesting that a high validation accuracy is a good criterion for achieving a high test accuracy. The training accuracy, however, remains more or less stable across epochs, and hence is not a good indicator of epochs that provide classifiers that achieve high test accuracies.

To highlight the need for including whole sections in the validation dataset, we additionally re-trained the network on TMAs 1–5 and validated the training on the remaining TMA 6 (Figure 2C). The identical run was additionally validated on the whole sections usually used for validation (Figure 2B). Figures 2C and 2D show that, when using TMA spectra for validation, validation and test accuracies are no longer correlated and choosing the epoch based on the validation accuracy would lead to a low accuracy on the test dataset. Note that the choice of TMA 6 for validation was purely arbitrary, similar effects are observed when using any of the other TMAs for validation.

The same is true for the LDA classifier. While the accuracy on the training dataset is increasing with the amount of features used for the LDA, the validation and test accuracies are not (Figure 2A). When using a validation dataset with spectra from TMA cores for the choice of the number of features, only low accuracies on the test dataset would be reached (Figure 2C).

The LDA classifier would require a higher number of features for a high training accuracy, that in turn would lead to lower accuracies on the whole sections. This indicates that using a higher number of features for the LDA would not lead to learning more biologically relevant features, but rather to an overfitting, meaning that non-biologically-meaningful features specific to the training data would be learned.

In general, a validation dataset is important as an indicator for overfitting, as the classifier's training process is only influenced by the training data and the classifier might learn features that are specific to the training data and non-biologically meaningful. Using data from a TMA as validation dataset is in this case not sufficient. A possible reason might be that in the training process the classifiers learn features that are only specific to TMAs and that are not present in whole sections. These differences in spectra from TMAs vs. spectra from whole sections might be due to the different preparation processes applied to TMAs and whole sections, respectively, resulting in different technical characteristics of the acquired MSI data. What exactly these differences are, how they can be observed and, more importantly, avoided, is still not clear and needs to be investigated further. Nevertheless, including data from whole sections in the model selection process helps to choose an epoch and hyperparameters/number of features that result in a classifier that learned biological features that are also prevalent in whole sections.

3.2 | Influence of preprocessing pipeline

The second prerequisite to achieve very high classification accuracies is the extensive preprocessing. Theoretically, a sufficiently complex and appropriately designed neural network could be able to learn all preprocessing steps and provide very accurate classification results on the raw spectra. In practice, however, this is not possible when only a limited and insufficient amount of training data samples is available. The preprocessing used in this study made it possible to achieve high accuracies using only the available limited training dataset.

To investigate the influence of the preprocessing, we successively removed certain preprocessing steps from the pipeline (log transformation and spatial denoising) and replaced the IPN with a standard total ion count (TIC) normalization. The results can be found in Figure 3 and Table S4 in the Supporting Information. As for the full pipeline, the training process was repeated five times and the resulting best network in terms of balanced validation accuracy was used for further analysis.

For the neural network, the pipeline without log transformation (ipn-cal-smooth-pmr-ipn) achieves the best result on the validation and test dataset in terms of balanced accuracy. Compared to the full pipeline, slightly higher values on validation and test dataset are reached but slightly lower values on the training dataset (Table S4). Note that even though the stability of the training is decreased when removing the log transformation, this does not influence the classification success of the final network, which is chosen from the epoch with the highest validation accuracy. The hyperparameter tuning was only performed for the full pipeline and another set of hyperparameter values might achieve a more stable training process for the reduced pipeline as well. Similar results are found for the LDA classifier. The pipeline ipn-cal-pmr-smooth-ipn performs comparably well as the full pipeline (ipn-cal-smooth-pmr-ipn-log). It follows, that the log transformation of the peaks does not make a big difference for the classification.

The additional removal of the spatial smoothing, however, has a bigger effect on the training accuracy of neural network and LDA. For the neural network, the training accuracy reaches only values below approximately 94% (pipelines ipn-cal-pmr-ipn and cal-pmr-tic) (Figure 3). The same can be observed for the LDA where significantly lower accuracies on the training data are achieved, even when using a high number of features. Additionally, the spatial smoothing seems to have a big influence on the variation of validation and test accuracies with the number of features (Figure 3). Without the spatial smoothing high accuracies are achieved when using at least 50 features and stay high when adding more features, with the spatial smoothing accuracies drop when adding more than 50 but less than 100 features. This might hint at a negative influence of the spatial smoothing making the classification of the LDA less stable. However, decisive is only the final classifier chosen from one fixed number of features. Additionally, note that the selection of features via ROC analysis is also computed from the preprocessed spectra. It follows that, for example, features 50–100 for the full preprocessing are not necessarily the same as features 50–100 for a different preprocessing pipeline, which could explain the different behavior of validation and test accuracies for the different pipelines. In general, adding more features can decrease the accuracies, because the additional features may be highly correlated and introduce only little new information.

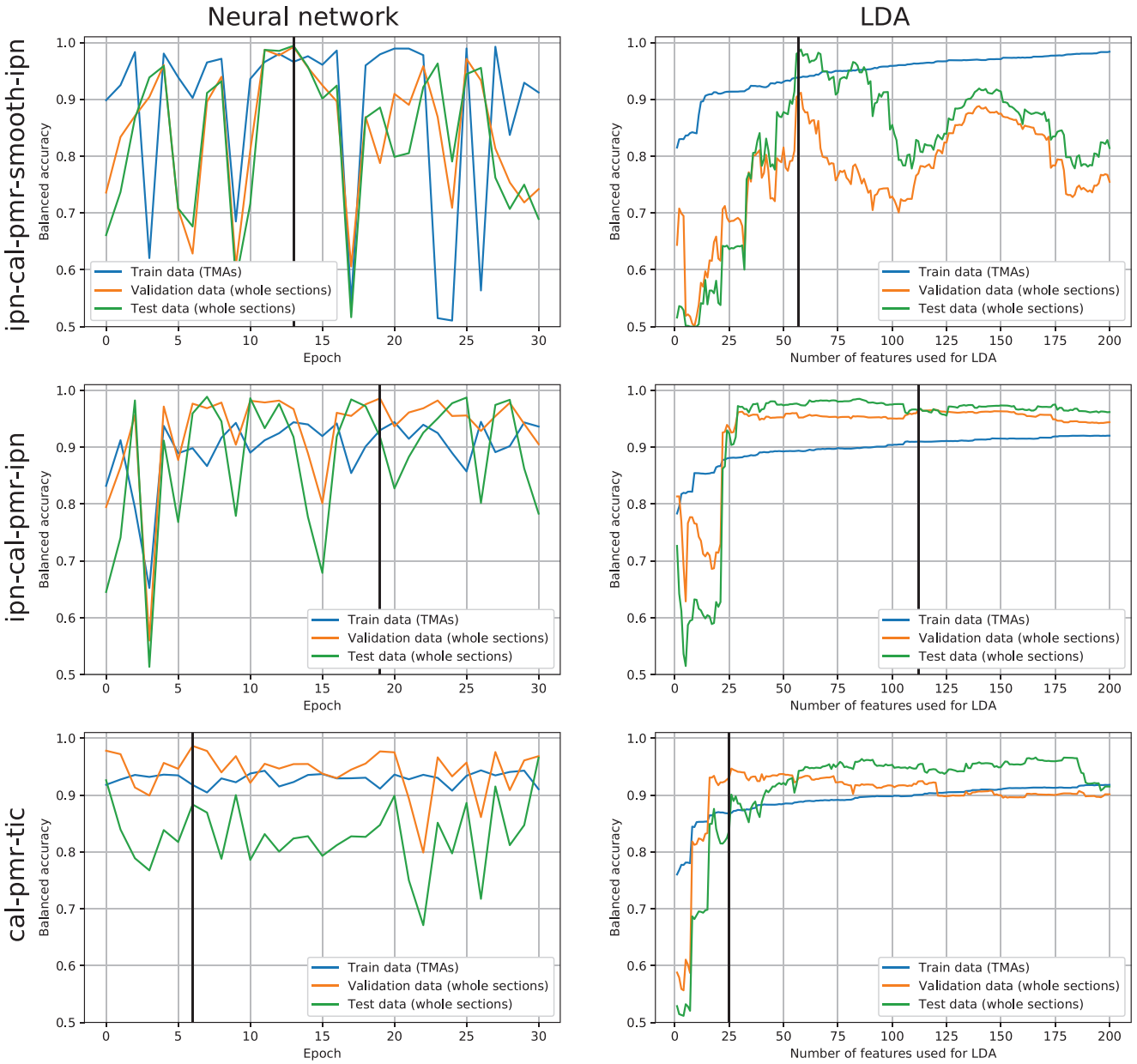

FIGURE 3 Balanced accuracies on training (blue), validation (orange) and test dataset (green) of the neural network and the LDA for different preprocessing pipelines. The vertical lines indicate the respective models with the highest validation accuracy, which are selected for further analysis. Accuracies for the respective best model for different pipelines are given in the Supporting Information in Table S4.

Finally, the replacement of the IPN with the TIC normalization additionally decreases the classification accuracies. The IPN aims at reducing technical variations in the data [17], which could explain its positive effect on the classification results.

4 | CONCLUDING REMARKS

We presented a fully-automated classification algorithm for robustly subtyping ADC and SqCC tumors in lung tissue whole sections based on MALDI MSI measurements. Note, that to apply our algorithm no expert knowledge, for example, to perform peak picking is required. For the first time, MALDI MSI measurements of lung tissue whole sections were classified with high accuracy. While further testing on data acquired at other labs and with different protocols is needed, our results suggest that the presented algorithm provides a step towards a clinical application of MALDI MSI and artificial intelligence and can be of great benefit for the decision making process of the pathologist.

In order to assess the significance of the neural network methodology to the performance of our tissue typing method, we compared it to a modified version in which the classification step was performed by LDA. Our results indicate that the LDA method is only slightly inferior to the neural network version, which may suggest that it is already "good enough," and that efforts for further improving the classification

accuracy should aim at other parts of the method, such as, for example, proper preprocessing to reduce technical variation. On the other hand, neural networks have only recently been applied to the analysis of MALDI MSI data, which may leave more room for improving this methodology than the classical, well understood LDA algorithm.

To apply the classification algorithm to a new whole section in a clinical routine, the workflow presented here requires the pathologist to manually annotate areas with high tumor cell content. This step could be further automated by utilizing a second neural network for detecting tumor areas based on the optical image of the whole section. Recent applications of U-Net architectures for semantic segmentation have been used very successfully to detect skin cancer in whole sections [20]. An application to lung tissue is currently investigated in a follow-up study. Additionally, the implementation of methods to make the classifier's decision explainable to the pathologist, for example by displaying the m/z-values that were the most relevant for the classifier's decision, will be the subject of future work.

ACKNOWLEDGMENTS

This study was in parts funded through the DIAMANT project (German Federal Ministry of Education and Research, BMBF, FKZ 031L0198A, Charlotte Janßen, Georg Steinbuß) and through the MALDISTAR project (Klaus Tschira Stiftung gGmbH, Project 00.010.2019, Jens Behrmann). Christiane Zgorzelski is acknowledged for excellent technical assistance.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Charlotte Janßen https://orcid.org/0000-0003-2121-9733
Tobias Boskamp https://orcid.org/0000-0002-5233-7962
Mark Kriegsmann https://orcid.org/0000-0002-7319-3646
Rita Casadonte https://orcid.org/0000-0002-9054-4786

REFERENCES

1. Meding, S., Nitsche, U., Balluff, B., Elsner, M., Rauser, S., Schöne, C., Nipp, M., Maak, M., Feith, M., Ebert, M. P., Friess, H., Langer, R., Höfler, H., Zitzelsberger, H., Rosenberg, R., & Walch, A. (2012). Tumor classification of six common cancer types based on proteomic profiling by MALDI imaging. *Journal of Proteome Research*, *11*(3), 1996–2003. https://doi.org/10.1021/pr200784p
2. Kriegsmann, J., Kriegsmann, M., & Casadonte, R. (2015). MALDI TOF imaging mass spectrometry in clinical pathology: A valuable tool for cancer diagnostics. *International Journal of Oncology*, *46*(3), 893–906. https://doi.org/10.3892/ijo.2014.2788
3. Balluff, B., Hanselmann, M., & Heeren, R. (2017). Mass spectrometry imaging for the investigation of intratumor heterogeneity. In R. R. Drake & L. A. McDonnell (Eds.), *Applications of mass spectrometry imaging to cancer* (pp. 201–230). Academic Press. https://doi.org/10.1016/bs.acr.2016.11.008
4. Yanagisawa, K., Shyr, Y., Xu, B. J., Massion, P. P., Larsen, P. H., White, B. C., Roberts, J. R., Edgerton, M., Gonzalez, A., Nadaf, S., Moore, J. H., Caprioli, R. M., & Carbone, D. P. (2003). Proteomic patterns of tumour subsets in non-small-cell lung cancer. *The Lancet*, *362*(9382), 433–439. https://doi.org/10.1016/S0140-6736(03)14068-8
5. Rahman, S. J., Gonzalez, A. L., Li, M., Seeley, E. H., Zimmerman, L. J., Zhang, X. J., Manier, M. L., Olson, S. J., Shah, R. N., Miller, A. N., Putnam, J. B., Miller, Y. E., Franklin, W. A., Blot, W. J., Carbone, D. P., Shyr, Y., Caprioli, R. M., & Massion, P. P. (2011). Lung cancer diagnosis from proteomic analysis of preinvasive lesions. *Cancer Research*, *71*(8), 3009–3017. https://doi.org/10.1158/0008-5472.CAN-10-2510
6. Kriegsmann, M., Casadonte, R., Kriegsmann, J., Dienemann, H., Schirmacher, P., Hendrik Kobarg, J., Schwamborn, K., Stenzinger, A., Warth, A., & Weichert, W. (2016). Reliable entity subtyping in non-small cell lung cancer by matrix-assisted laser desorption/ionization imaging mass spectrometry on formalin-fixed paraffin-embedded tissue specimens. *Molecular & Cellular Proteomics*, *15*(10), 3081–3089. https://doi.org/10.1074/mcp.M115.057513
7. Lazova, R., Seeley, E. H., Kutzner, H., Scolyer, R. A., Scott, G., Cerroni, L., Fried, I., Kozovska, M. E., Rosenberg, A. S., Prieto, V. G., Shehata, B. M., Durham, M. M., Henry, G., Rodriguez-Peralto, J. L., Riveiro-Falkenbach, E., Schaefer, J. T., Danialan, R., Fraitag, S., Vollenweider-Roten, S., & Caprioli, R. M. (2016). Imaging mass spectrometry assists in the classification of diagnostically challenging atypical Spitzoid neoplasms. *Journal of the American Academy of Dermatology*, *75*(6), 1176–1186 e4. https://doi.org/10.1016/j.jaad.2016.07.007
8. Boskamp, T., Lachmund, D., Oetjen, J., Cordero Hernandez, Y., Trede, D., Maass, P., Casadonte, R., Kriegsmann, J., Warth, A., Dienemann, H., Weichert, W., & Kriegsmann, M. (2017). A new classification method for MALDI imaging mass spectrometry data acquired on formalin-fixed paraffin-embedded tissue samples. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1865*(7), 916–926. https://doi.org/10.1016/j.bbapap.2016.11.003
9. Klein, O., Kanter, F., Kulbe, H., Jank, P., Denkert, C., Nebrich, G., Schmitt, W. D., Wu, Z., Kunze, C. A., Sehouli, J., Darb-Esfahani, S., Braicu, I., Lellmann, J., Thiele, H., & Taube, E. T. (2019). MALDI-imaging for classification of epithelial ovarian cancer histotypes from a tissue microarray using machine learning methods. *PROTEOMICS – Clinical Applications*, *13*(1), e1700181. https://doi.org/10.1002/prca.20170018117
10. Behrmann, J., Etmann, C., Boskamp, T., Casadonte, R., Kriegsmann, J., & Maaß, P. (2017). Deep learning for tumor classification in imaging mass spectrometry. *Bioinformatics*, *34*(7), 1215–1223. https://doi.org/10.1093/bioinformatics/btx724
11. Groseclose, M. R., Massion, P. P., Chaurand, P., & Caprioli, R. M. (2008). High-throughput proteomic analysis of formalin-fixed paraffin-embedded tissue microarrays using MALDI imaging mass spectrometry. *Proteomics*, *8*(18), 3715–3724. https://doi.org/10.1002/pmic.200800495
12. Kriegsmann, M., Zgorzelski, C., Casadonte, R., Schwamborn, K., Muley, T., Winter, H., Eichhorn, M., Eichhorn, F., Warth, A., Deininger, S.-O., Christopoulos, P., Thomas, M., Longerich, T., Stenzinger, A., Weichert, W., Müller-Tidow, C., Kriegsmann, J., Schirmacher, P., & Kriegsmann, K. (2020). Mass spectrometry imaging for reliable and fast classification of non-small cell lung cancer subtypes. *Cancers*, *12*(9). https://doi.org/10.3390/cancers12092704
13. Lazova, R., Smoot, K., Anderson, H., Powell, M. J., Rosenberg, A. S., Rongioletti, F., Pilloni, L., D'Hallewin, S., Gueorguieva, R., Tantcheva-Poór, I., Obadofin, O., Camacho, C., Hsi, A., Kluger, H. H., Fadare, O., & Seeley, E. H. (2020). Histopathology-guided mass spectrometry differentiates benign nevi from malignant melanoma. *Journal of Cutaneous Pathology*, *47*(3), 226–240. https://doi.org/10.1111/cup.13610
14. Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications*, *30*(2), 169–190. https://doi.org/10.3233/AIC-170729

15. Balluff, B., Hopf, C., Porta Siegel, T., Grabsch, H. I., & Heeren, R. M. (2021). Batch effects in MALDI mass spectrometry imaging. *Journal of the American Society for Mass Spectrometry*, *32*(3), 628–635. https://doi.org/10.1021/jasms.0c00393
16. Deininger, S.-O., Bollwein, C., Casadonte, R., Wandernoth, P., Pereira Lopes Gonçalves, J., Kriegsmann, K., Kriegsmann, M., Boskamp, T., Kriegsmann, J., Weichert, W., Schirmacher, P., Ly, A., & Schwamborn, K. (2022). Multi-center evaluation of tissue classification by matrix-assisted laser desorption/ionization mass spectrometry imaging. *Analytical Chemistry*. unpublished.
17. Boskamp, T., Casadonte, R., Hauberg-Lotte, L., Deininger, S., Kriegsmann, J., & Maass, P. (2021). Cross-normalization of MALDI mass spectrometry imaging data improves site-to-site reproducibility. *Analytical Chemistry*, *93*(30), 10584–10592. https://doi.org/10.1021/acs.analchem.1c01792
18. Boskamp, T., Lachmund, D., Casadonte, R., Hauberg-Lotte, L., Kobarg, J. H., Kriegsmann, J., & Maass, P. (2019). Using the chemical noise background in MALDI mass spectrometry imaging for mass alignment and calibration. *Analytical Chemistry*, *92*(1), 1301–1308. https://doi.org/10.1021/acs.analchem.9b04473
19. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
20. Le'Clerc Arrastia, J., Heilenkötter, N., Otero Baguer, D., Hauberg-Lotte, L., Boskamp, T., Hetzer, S., Duschner, N., Schaller, J., & Maass, P. (2021). Deeply supervised UNet for semantic segmentation to assist dermatopathological assessment of basal cell carcinoma. *Journal of Imaging*, *7*(4), 71.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Janßen, C., Boskamp, T., Hauberg-Lotte, L., Behrmann, J., Deininger, S. - O., Kriegsmann, M., Kriegsmann, K., Steinbuß, G., Winter, H., Muley, T., Casadonte, R., Kriegsmann, J., & Maaß, P. (2022). Robust subtyping of non-small cell lung cancer whole sections through MALDI mass spectrometry imaging. *Prot. Clin. Appl.*, *16*, e2100068. https://doi.org/10.1002/prca.202100068