Silesian University of Technology
Faculty of Automatic Control, Electronics and Computer Science
Institute of Informatics

Doctor of Biomedical Engineering Dissertation

Machine learning-based workflow for the analysis of MALDI-TOF mass spectrometry cancer data

mgr inż. Wojciech Sikora

Supervisor: Prof dr hab. inż. Joanna Polańska

mgr inż. Wojciech Sikora

Machine learning-based workflow for the analysis of MALDI-TOF mass spectrometry cancer data

Abstract

The subject of the dissertation is the analysis of data obtained by mass spectrometry imaging of samples obtained from patients with head and neck cancer. The aim of the work is to propose a complete workflow for the analysis of mass spectrometry imaging data, starting with raw mass spectra and ending with a model capable of classifying new observations into one of several classes. Three hypotheses were made at the beginning of this work. The first hypothesis concerns the method of peak identification in mass spectra. The hypothesis states that peak identification can be successfully performed by dividing the spectrum into parts and modeling them with Gaussian mixture models. The second hypothesis concerns the method of redundancy removal and dimensionality reduction in the data. The second hypothesis states that the information about spatial distribution available thanks to imaging, helps to remove redundancy and reduce the dimensionality of the data. The third hypothesis concerns the evaluation of the importance of features in classifiers. The hypothesis states that identifying the most important features for heterogeneous data is possible by aggregating the information from many unit models.

The first chapters of the thesis deal with the basic issues of proteomics and mass spectrometry imaging of biological samples. The first chapter briefly explains why the analysis of imaging data of biological samples is important and introduces the main concepts related to this topic. The first chapter also discusses the challenges a data analysist must face during the work with mass spectrometry data and state of the art methods for dealing with them.

In the second chapter, the techniques of data acquisition using mass spectrometry and mass spectrometry imaging are presented in more detail. First, the general operating scheme of mass spectrometers is described, along with the typical parameters that distinguish the various types of mass spectrometers and affect their applications as well

as an explanation of what a mass spectrum exactly is. The chapter also includes a description for mass spectrometry imaging process, and a visualization of the result of this process (see Figure 1).

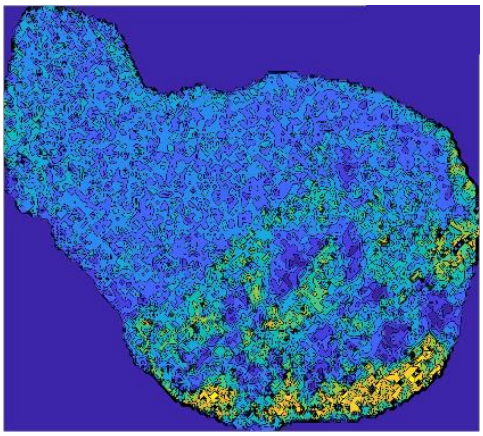

Figure 1: An example of image acquired by mass spectrometry imaging.

This is followed by the description of the main ionization methods for mass spectrometry imaging. Three methods that are most commonly used for imaging of biological samples are presented. These methods are desorption electrospray ionization (DESI), matrix-assisted laser desorption ionization (MALDI), and secondary ion mass spectrometry (SIMS). The chapter includes information about their operating principles with visualizations and comparison of working parameters. Reference is also made to other methods, which are in general modifications of the mentioned methods but are not yet widely used. Then, a description of the next steps in mass spectrometry, i.e., ion separation and detection is provided. In particular, the time-of-flight ion separation method has been described in detail, since this method was used to obtain the data analyzed in this thesis.

The third chapter focuses on presenting the data processed during the experiments. At first, the process of sample acquisition and preparation by a specialist is presented, step by step, along with all the important parameters and detailed information about the mass spectrometer used. The acquired samples with marked regions of cancer, epithelium and normal tissue are presented in the Figure 2.

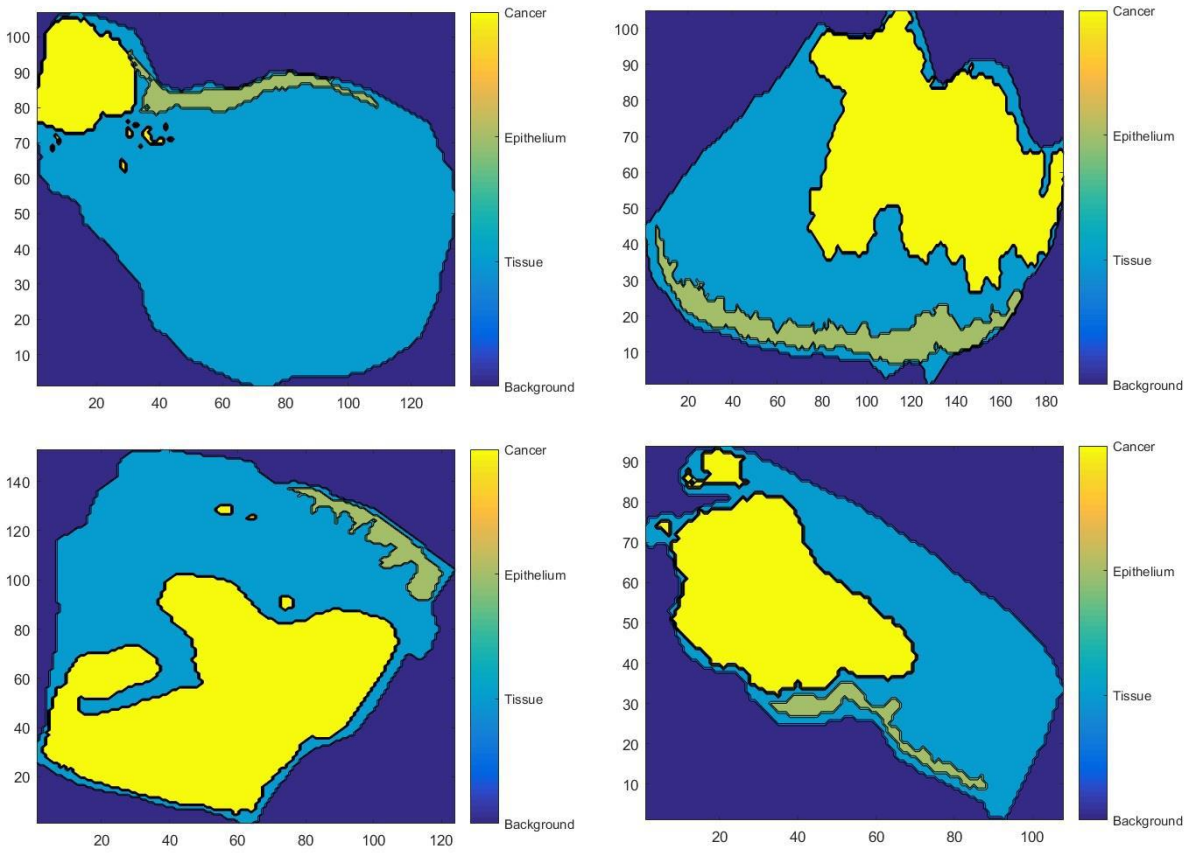

Figure 2: Acquired samples with marked regions of cancer, epithelium and normal tissue.

This chapter also describes the initial steps taken to prepare the raw data for the main part of the processing, that is, peak identification. At first the proposed method of baseline correction is described and other commonly used methods are listed like wavelet and splines-based methods. The spectra are then normalized with standard TIC (total ion count/current) method, where the sum of all intensities is calculated and each intensity is divided by this number. The last step is spectrum alignment with fast Fourier transform. The chapter also contains references to the most useful openly available tools to perform these steps.

The fourth chapter is the introduction to the main topic of mass spectrometry data analysis, peak identification. In this chapter, the current state of the art about peak identification, particularly about peak identification for proteomics related data. First, three methods for aggregating mass spectra were described and compared. Peak identification is performed on the aggregated mass spectra. The first method for peak identification was a simple peak picking method that uses a threshold for the signal-to-noise ratio for the peak intensity. The noise was defined as the mean absolute deviation of the intensity in the local neighborhood. It is a simple and commonly used method that can detect only peaks with intensities above the defined noise level. The results of peak

identification using this method on the real data were shown with different values for the window width of the local neighborhood and the threshold value for the signal-to-noise ratio. The results were unsatisfactory. The next method used and described is based on the continuous wavelet transform. This method is categorized to in the dissertation as peak modeling method. Peak modeling methods use the shape of the peak as well as the intensity of the peak. The identification with this method is done by first transforming the spectrum into the wavelet space and then identifying the ridge lines correlated with the peaks in the spectrum. In the paper, the mathematical description of the wavelet transform is given with an example as well as a detailed description of how the entire peak modeling process works with visualizations of each step and the final results.

In the fifth chapter, the method of peak identification is presented, which is used as one of the steps of the proposed workflow for the analysis of mass spectrometry imaging data. The goal of this method is to model the entire mass spectrum using Gaussian mixture models. The method has been presented in the cited publications. The first step of the method is to divide the mass spectrum into smaller fragments and model each fragment with a Gaussian mixture model, essentially modeling the peaks as normal distributions. At the beginning of the chapter, the motivation and arguments for this approach to peak identification are discussed. The arguments are related to the nature of the peaks in mass spectra acquired by time-of-flight mass spectrometry and to the nature of the MALDI ionization method. In further parts, the experiments aimed at improving the originally proposed spectrum splitting method are described. First, the original method is presented, which, in simplified terms, uses a peak identification method to search for "true" peaks and uses the peaks found to split the signal.

Consideration was then given for how the evaluation of spectral fragments should be done. Two rules were established that should be followed when splitting the spectrum. The first rule states that a portion of the spectrum should be as small as possible, i.e., it should contain a single peak if possible. The second rule states that a grouping of peaks, i.e., a group of peaks that overlap, cannot be divided into different parts and should be combined into a single part. Next, experiments were conducted using two new approaches to spectrum splitting. The first method was very similar to the original method. The method looks for peaks using the continuous wavelet transform and splits the signal at the lowest points between the peaks. The second method focuses on finding optimal division points without prior peak detection. The division points are found by local minimum value search. In the local minimum search, an additional constraint is applied based on the local neighborhood for the minimum value of the local

minimum. If the value of the local minimum is below the threshold, the spectrum is not split at that point. The two methods were compared together with manual division (see Figure 3) on real data and evaluated with defined rules. In the end, the last method proved to be the best. The paper also contains the pseudocode for the implementation of this method.

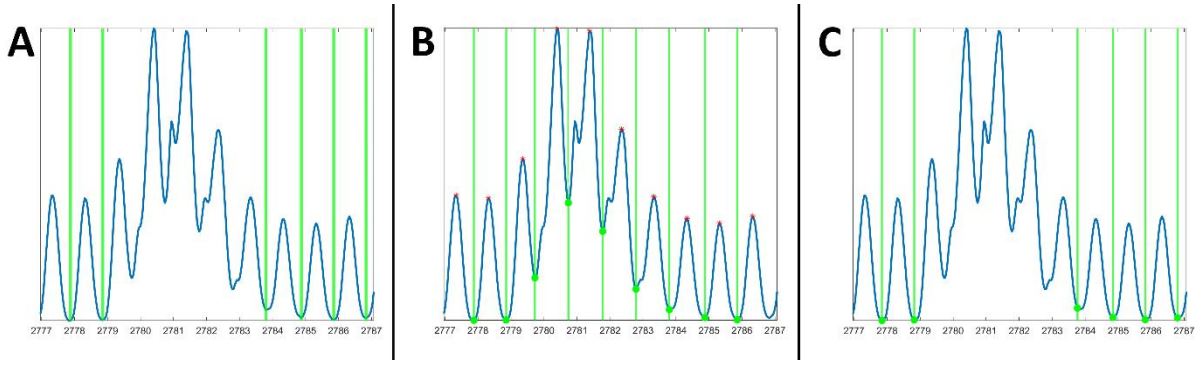

Figure 3: Comparison of methods for search of spectrum division points. Manual division (A). Division with continuous wavelet transform-based peak identification (B). Division by search for local minima (C).

The remainder of the chapter deals with fitting Gaussian mixture models to the split spectrum. Since it is impossible to calculate the optimal parameters for the Gaussian mixture model analytically, the fitting is performed using the expectation-maximization (EM) algorithm. The paper contains a general and mathematical description of the algorithm as well as a step-by-step visualization of the process of fitting a Gaussian mixture model to a randomly selected part of the split spectrum. A custom implementation of the EM algorithm was created for experiments and the pseudocode was included in the paper. The last subsections deal with the selection of the optimal value for the parameter k, i.e., the number of mixture elements to be fitted to a part of the spectrum. The quality of the models is described by the likelihood, based on which the Bayes information criterion is calculated to introduce a penalty for model complexity. With this approach, an optimal number of parameters is chosen. Due to the stochastic nature of the EM algorithm, the process is repeated several times. Ultimately, the value of k that occurs most frequently is chosen. The boxplots included in the dissertation (see Figure 4) clearly show how, for very low values of k, the variance of the model likelihood is high and gradually decreases. In the end, the original number of over one hundred thousand data points in each spectrum was reduced to 9454 elements describing the averaged mass spectrum model.

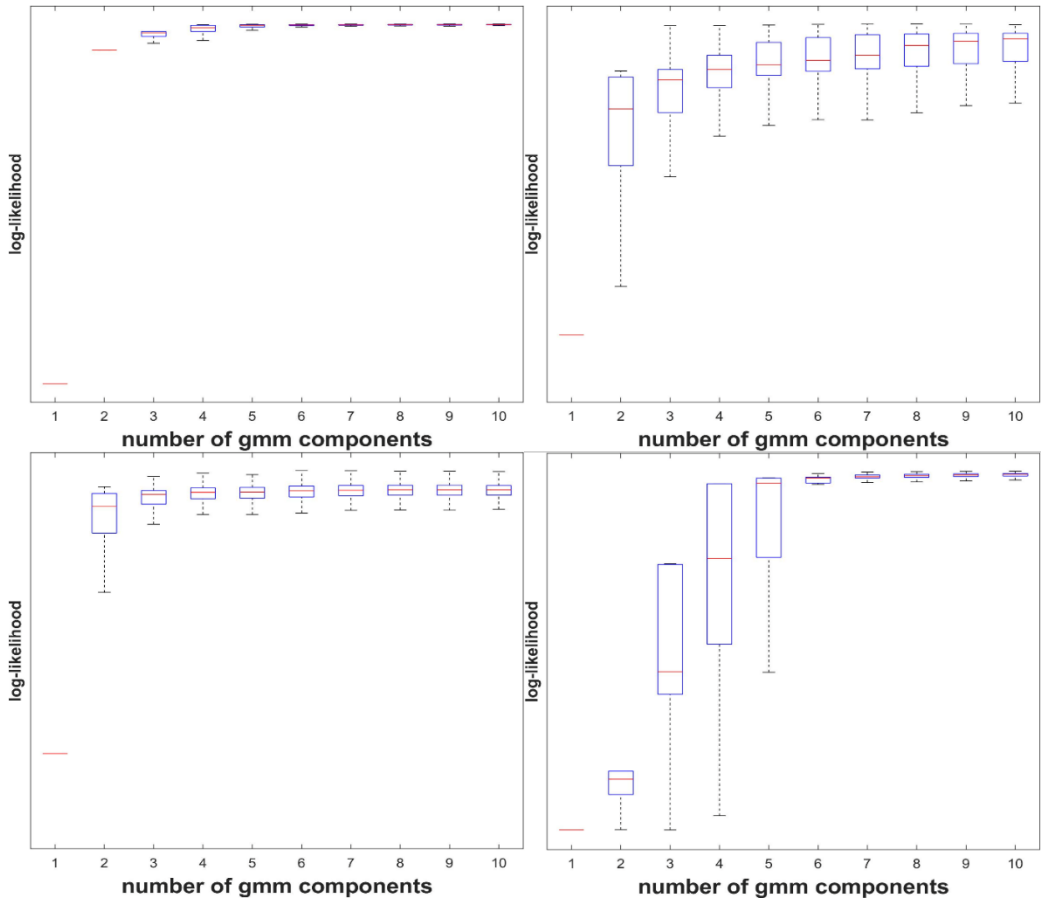

Figure 4: Boxplots for 4 randomly chosen spectrum parts.

The sixth chapter is devoted to feature engineering, with the goal of further reducing the dimensionality of the data and removing redundancy. The key aspect of this part of the thesis is the comparison of the spatial distribution of the elements on the samples. At the beginning, we discuss the goals and threats associated with dimensionality reduction. Then, our strategy for noise removal is described. The strategy consists of filtering the noise using parameters of the normal distributions describing the elements of the spectrum model, in particular the σ parameter describing the shape of the peak and the λ parameter describing the intensity of the peak. The distribution of these parameters was examined for the entire population of spectrum elements. The distributions were modeled using the Gaussian mixture model and the elements modeling the noise were removed. As a result, the number of elements was reduced from 9454 to 2884.

Further dimensionality reduction was achieved by comparing the spatial distribution of nearby elements (features) on the physical samples. The logic behind this is that when fitting Gaussian mixtures, a peak in the spectrum can be modeled with more than one normal distribution. This occurs because the bass line correction is not ideal and because

a 'true" peak is not exactly a normal distribution. Due to uncertainties in peak detection caused by slight variations during ion motion in the magnetic field, the true peaks in the mass spectrum are slightly skewed. This is characterized by a slight flattening of the right slope of the peak. If the neighboring elements of the spectrum model statistically have the same spatial distribution, they are merged into a single feature.

The next sections of the paper explain the statistical tests used to compare the spatial distributions of the elements. First, the Kolmogorov-Smirnov statistical test is described with an example, along with a visualization of the test statistics when comparing two artificially generated probability distributions. It is a test that can be used for comparing two empirical one-dimensional probability distributions. For the comparison of spatial distributions, the extension of this test proposed by Peacock was used. Peacock's test is explained by comparing two randomly selected elements of the spectrum model. A detailed description and visualizations of the process are provided. Since there are no significance levels for Peacock's test, the critical values were calculated using numerical simulations. Since each sample has a different size, the critical value was calculated separately for each sample. To obtain a single *p*-value for the test, the one-sided *p*-values of all unit tests were combined using Fisher's method. As a result, the number of features decreased from 2884 to 2392.

The last subsection refers to the detection of isotopic envelopes. The isotopic envelope is an expression of a specific molecule that contains different isotopes of atoms in its chemical composition, causing differences in mass and therefore differences in mass-to-charge ratio (m/z). Isotopic envelopes hinder the analysis of mass spectrum and it is beneficial to represent them as a single feature at the place of the dominant peak. Usually the difference in atomic mass between consecutive peaks is 1 Da. Peaks in an isotopic envelope should have similar shape and their spatial distribution should be the same. Using this information an algorithm for isotope envelope search was created. The algorithm first checks the distance between the compared normal distributions (the difference in μ values), and if the distance is within a numerically acquired range, the shape of the normal distributions is checked (the ratio of σ values). If both conditions are met, the spatial distribution is compared. Normal distributions (features) that satisfy all three conditions are then merged at the position of the dominant feature, and the value of the new feature is the sum of all elements in the envelope.

At the end of the chapter, the results of the feature engineering are presented. Finally, the spectrum model, consisting of 9454 elements, each described by a single normal distribution, was reduced to a set of 888 features, confirming the second hypothesis

stated in the thesis. The final number of features is in accordance with the expectations derived from the field-specific knowledge.

The seventh chapter is devoted to the application of statistical and machine learning methods to train classifiers capable of making a prediction for a new observation based on the processed data. First, methods widely used for this purpose and corresponding publications are listed. Then, the sampling of data into training, testing and validation sets is addressed. The validation set was set aside with a size of 10% of the total data set. Then, the training and test sets were selected for cross-validation by stratified sampling to maintain the class balance in each set. A new split is made for each unit model. Then, the equations and explanations for all model performance measures are presented. The basic model evaluation measures calculated using the confusion matrix are accuracy, precision, sensitivity, specificity, positive predictive value, and F1 score. More complex methods for model evaluation and comparison are then described. A detailed description of the ROC curves, precision-sensitivity curves, and NPV-PPV curves is provided with a detailed description of each plot and visualizations for trained models.

The next section describes the algorithm for training the classifier using multinomial logistic regression. In this method, the logistic regression is run multiple times, each time using a different feature as the independent variable. The feature with the best probability is selected as the first independent variable of the final model. Then, logistic regression is run again for each remaining feature, this time along with all previously selected features. The best feature is selected as the next independent variable of the final model. The process is repeated until a decision is made, based on the calculated Bayes factor, that the increase in model likelihood does not offset the increased model complexity caused by an additional element, preventing overfitting. The paper shows how the performance of the model changes when new features are added to the model, using both ROC (see Figure 5) and precision sensitivity curves. Several models were trained with this algorithm, each time for a new split in training and testing sets. The results are presented in the form of a table with average values for all the mentioned model performance measures and 95% confidence intervals for the average values. In addition, a procedure for further optimization of trade-offs between any pairs of opposing measures is described. In this work, optimization of the trade-off between positive and negative predictive values was performed by maximizing the Youden index value.

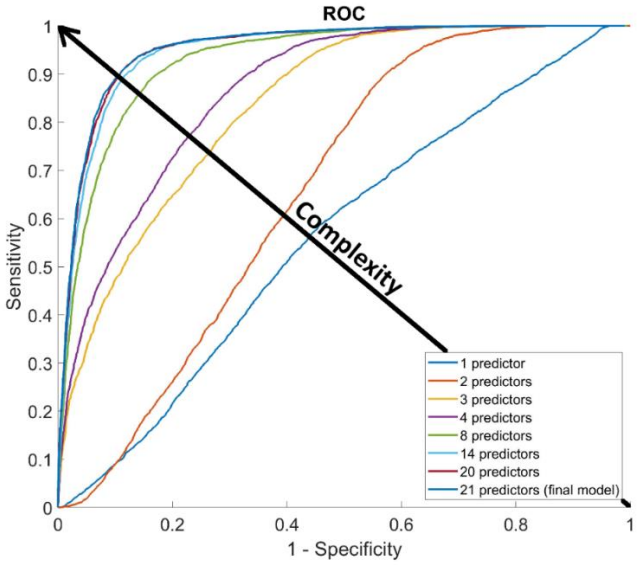

Figure 5: Influence of model complexity on the classification performance.

The second group of classifiers are simple neural networks. Using the same division into training and testing sets, classifiers were trained in the form of fully connected neural networks with two hidden layers and a number of nodes in each layer equal to the number of features. The results were presented in analogy to the logistic regression based method. Compared to the more extensive, iterative logistic regression algorithm, the neural networks performed worse but the performance of was still very good. The high performance for both methods confirms the high quality of the data processing.

The final sections of the paper focus on evaluating the importance of features and thus validating the last hypothesis made in the paper. Feature evaluation methods are proposed for both groups of classifiers. Classifiers trained with logistic regression are ordered lists of features due to the nature of the algorithm. Therefore, the score for a feature in a unit model is simple. The feature score is equal to $\frac{1}{x}$, where x is the position of the feature on the list. The total score for a feature is the average score of the feature across all unit models.

For neural networks, scores were calculated in a similar way, but first the features had to be ordered. Two methods were used to interpret the black-box model that neural network is. The first method is the locally interpretable model-agnostic explanation (LIME). This method can be used to interpret the feature importance of an observation classified by any model, since it requires only the input values of the observation and the prediction made by the classifier. This method attempts to interpret the importance of each future by perturbing the values of that future to generate new observations, e.g., by drawing the new values for the feature from a normal distribution with parameters

calculated based on the entire population of feature values in the dataset. The newly generated observations are assigned a class based on the model to be explained. Another explainable model, such as decision trees, is then trained on the resulting dataset. Feature importance is then determined by inspecting the interpretable model. The method explains a single observation in the dataset, to obtain the ordered list of feature importance for the unit model, several thousand observations were explained in this way and the results combined. After the ordered list is obtained, the final feature values are calculated in the same way as in the logistic regression based algorithm.

The second method for interpreting neural network models is Shapley values. This is an interpretation method applicable for any model based on the concept that each feature has some influence on the final classification result and this influence can be measured by removing the feature from the equation and examining the influence on the classification. Exact Shapley values are computed for each possible subset of features, so the computational complexity increases exponentially with the size of the feature set. For this reason, most implementations, including the one used in this paper, compute estimates of Shapley values by restricting the number of subsets. It also simulates the idea of removing a feature from the equation, since a feature cannot be easily removed from the explained model. To simulate this behavior, multiple calculations must be averaged for randomly selected values. The result of feature importance evaluation was presented by showing the images for top five most important feature for each method (see Figure 6).

The final chapter contains discussion about the conducted experiments, conclusions drawn, and plans for future research.

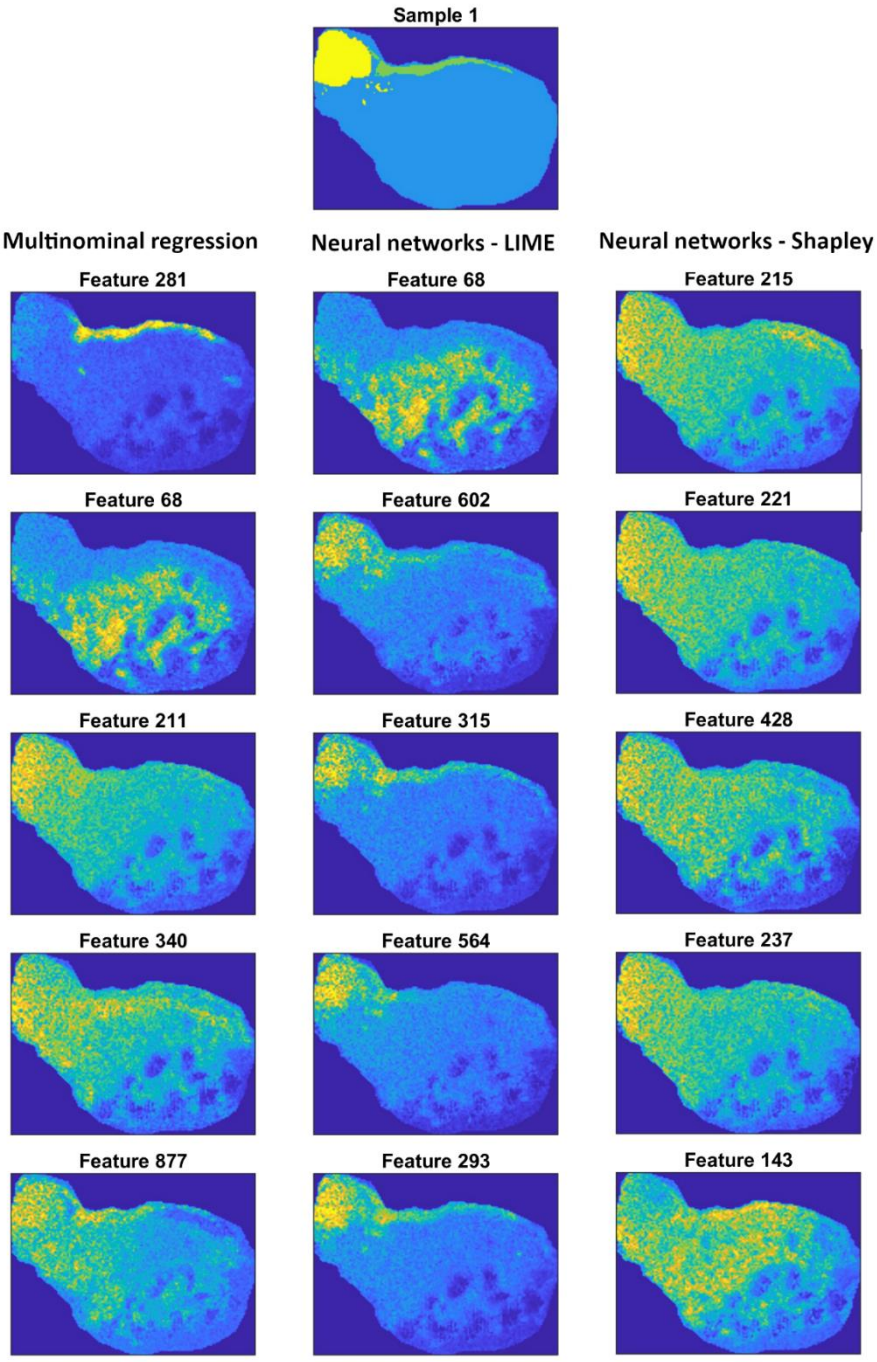

Figure 5: Images for top five features for sample number 1.