Paweł PENDZIAŁEK*[1], Joanna POLAŃSKA[1]

Chapter 11. MALDI-MSI MOLECULAR IMAGING DATA CLASSIFICATION SYSTEM

11.1. Introduction

The aim of the research was to create an intelligent system that finds the signature of head and neck cancer based on the protein profile of mass spectrometry. It is currently the sixth most common cancer worldwide [1]. Despite the work of many teams, there is still no marker for use in biopsy and blood testing for this type of cancer. To find that marker, one of the stages are machine learning tools. In this stage, preparations properties are compared with their description from a specialist. The system was created through the fusion of histopathological information and molecular imaging data. Using both information, a mathematical model can be created that allows to distinguish diseased tissue from a healthy tissue. That model can be a part of the system used for diagnostic purposes. Mass spectrometry data from samples taken from patients can be loaded into the system. Then, using the previously created model, the system could assess whether the patient has a suspicion of head and neck cancer.

11.2. Materials & Methods

The data used to find head and neck cancer signature comes from research carried out at the Maria Sklodowska-Curie Institute – Oncology Centre (MSCI), in Gliwice (Poland). These are 5 preparations taken from patients with diagnosed head and neck cancer. Each of the samples contains diseased and healthy parts. Preparation is

* Corresponding author: pawel.pendzialek@polsl.pl, Łanowa 10/25, 41-800 Zabrze, PL.

[1] Department of Engineering and Data Exploratory Analysis, Silesian University of Technology.

a fragment of tissue with a very small thickness, therefore it is analyzed only in two-dimensional space. The pathologist places a grid over the sample (the grid consists of vertical and horizontal equidistant parallel lines). Then the pathologist examines the sample cell by cell, using a microscope. For each cell, the pathologist determines whether this fragment looks more like diseased or healthy tissue. These cells will be called mass spectra. Each of the preparations was also examined in the spectrometer using MALDI-MSI technique. This machine measured values of all 40 160 mass spectra for 109 000 mass channels with high accuracy. Gaussian Mixture Model technique was applied to reduce the number of mass channels to 3714. Mass channels are peptides (proteins) and they differ by the mz value. They will be called features. The feature values are numbers in the range from 1.814e-32 to 2.254e8. Table 11.1 shows statistics of the preparations. The number of diseased and healthy tissue spectra is calculated on the basis of the pathologist's description. In parentheses there is information what part of the preparation is the given type of tissue. In the Figure 11.1 there are shown mentioned preparations.

Table 11.1

Preparations statistics

	Prep. 1	Prep. 2	Prep. 3	Prep. 4	Prep. 5
Number of spectra	8005	11869	11823	4505	3958
Number of diseased tissue spectra	844 (10.54%)	4885 (41.16%)	5631 (47.63%)	1962 (43.55%)	1963 (49.60%)
Number of healthy tissue spectra	7161 (89.46%)	6984 (58.84%)	6192 (52.37%)	2543 (56.45%)	1995 (50.40%)

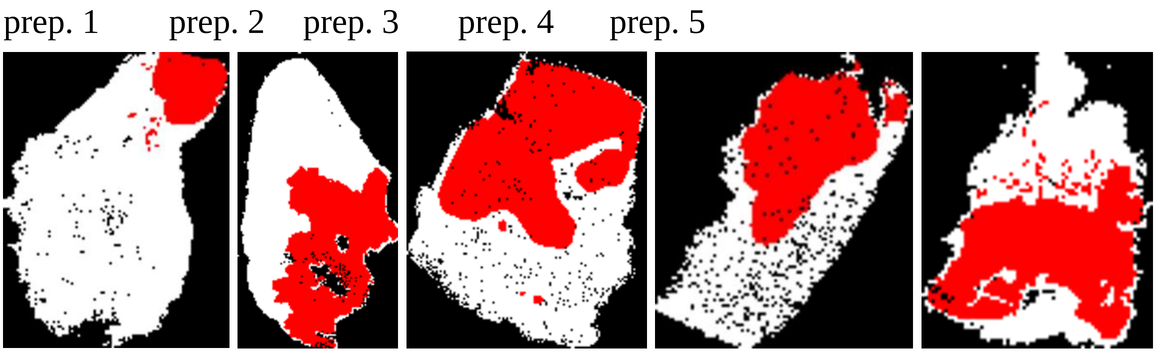

Fig. 11.1. 5 preparations used to create the system. Colors: healthy tissue spectra – white, diseased tissue spectra – red, background – black

Rys. 11.1. 5 preparatów użytych do utworzenia systemu. Kolory: widma zdrowe – biały, widma chore – czerwony, tło – czarny

The aim of the study was to create a system that classifies the data which was not "seen" before. Therefore, the data was split into three sets:

- training set (80% healthy and diseased tissue spectra from each of the preparations 1,2,3,4)
- test set (20% healthy and diseased tissue spectra from each of the preparations 1,2,3,4)
- validation set (all spectra from the preparation 5)

The training set was used to select the most important features and create a model. Next, the test set was used to find an optimal threshold value for classification task. This way the final model was obtained, and its quality was checked using the validation set. In the Figure 11.2 there is a block diagram which presents a scheme of classifier construction and validation. The system was implemented entirely in Python language.

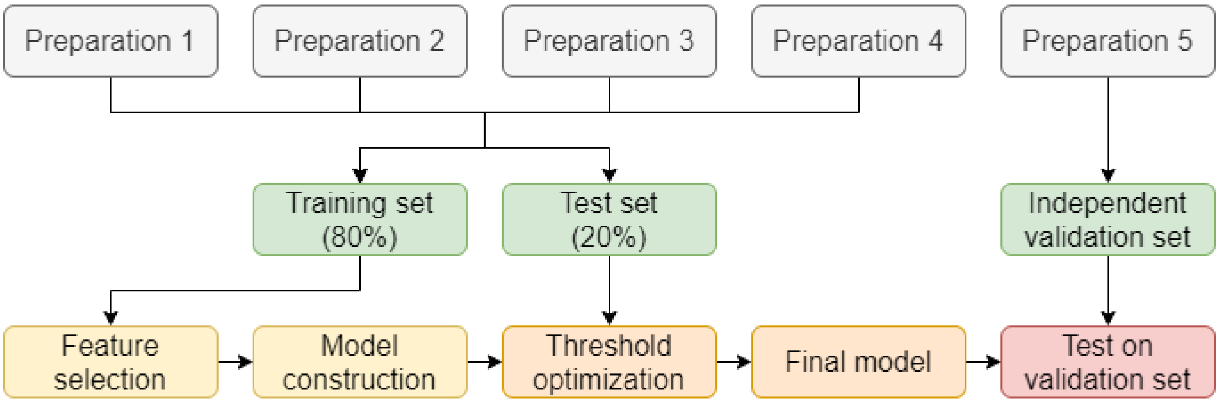

Fig. 11.2. Block diagram illustrating a classifier construction and validation scheme
Rys. 11.2. Schemat blokowy ilustrujący sposób konstrukcji i walidacji klasyfikatora

The feature values differ by up to 40 orders of magnitude (range from 1.814e-32 to 2.254e8). If such data is loaded to the classifier, it extends the computation time, and secondly, the classifier builds a model using a small number of features with the highest values. To solve this problem, the data was logarithmic. Every *x* in the data was converted to *ln(1+x)*. The number 1 was added so that the data in the range (0,1) after logarithm is positive and has small values. This method reduced the data range from 1.814e-32 to 1.923e1.

The number of features is really big (there is 3714 of them). Reduction of feature number can speed up calculations and help to find only features which are related to the head and neck cancer signature. Probably there exist features which values are very small. Also there may be a group of features that have high values, but their variety is small. In each preparation there is healthy and diseased tissue so it can be assumed that both of mentioned groups of features do not affect whether tissue part is diseased or not. The goal is to select features which values are large and highly variable.

To evaluate features importance, the variance was calculated for each of them (all spectra from the training set were used to calculate it). Variance is an arithmetic mean of the squares of deviations of numbers from their arithmetic mean. Looking at this definition, both features with low values and features with low variety have relatively small variance values. The question is, how many features with the highest variance should be taken to the classifier. The threshold variance value was selected using the GMM (Gaussian Mixture Model) method. This method is to present a data distribution as a sum of a finite number of Gaussian distributions with unknown parameters [2]. For distribution of feature variances, 3 Gaussian distributions were fit (Figure 11.3). The points of intersection of "neighboring" distributions divide the features into 3 groups. The group with highest variance values was selected – only features with variance higher than 0.364 were taken from the data. This way 217 features (less than 6%) remained.

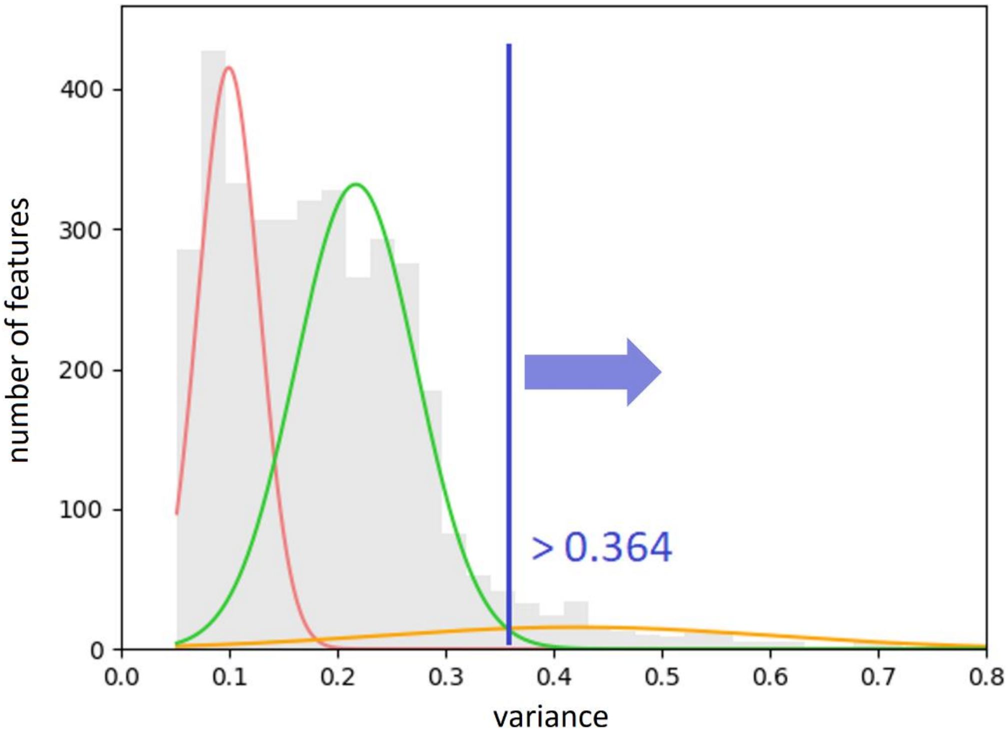

Fig. 11.3. Distribution of feature variances approximated by GMM
Rys. 11.3. Dystrybucja wariancji cech przybliżona przez GMM

The next step was the model building. The tool used for classification was SVM (Support Vector Machine) classifier with the RBF (Radial Basis Function) kernel. This is an algorithm which uses support vectors to construct the hyperplane that divides classes. In analyzed data there are two classes: the "healthy" class and the "cancer" class. The model was built on the test set.

To evaluate the model’s quality, it is necessary to know the following numbers:

- TP (true positive) – the number of data correctly classified to a positive class
- TN (true negative) – the number of data correctly classified to a negative class
- FP (false positive) – the number of data incorrectly classified to a positive class
- FN (false negative) – the number of data incorrectly classified to a negative class.

In analyzed data, "cancer" is the positive class, and "healthy" is the negative class. After counting above values, the classifier quality indicators can be calculated:

- Sensitivity (percentage of correctly classified positive data points)
- Specificity (percentage of correctly classified negative data points)
- PPV (percentage of correct “positive” assignments)
- NPV (percentage of correct “negative” assignments)
- Accuracy (percentage of correctly classified data points)
- Balanced accuracy (arithmetic mean of sensitivity and specificity)
- F1 score (harmonic mean of sensitivity and PPV).

To calculate the probability of spectra belonging to the cancer class, an algorithm that uses decision function was implemented. The classifier creates a decision function that returns a value for each of the data point in the test set. In case of 2 classes, on the one side of the hyperplane there are data points with negative values, and on the other side there are data points with positive values. The further from the hyperplane, the higher the absolute value of the data point is and the more it fits to the class it is inside (according to the classifier). A method for calculating probabilities is to scale all of the returned values linearly to values in the range [0,1]. Probabilities were calculated to find the threshold value for which classification results are the best. To find optimal threshold the Youden’s *J* statistic was used, which is the sum of sensitivity and specificity minus 1. The best threshold value was calculated on the test set. For this threshold value, a test was performed on the independent validation set.

To check the quality of the system, all steps were repeated for the remaining splits into 4 training preparations and one preparation for independent validation. Training and test sets were created in the same way as before.

11.3. Results

In the Table 11.2 there are presented results for all 5 data splits. In the Figure 11.4 there are presented bitmaps showing classification results for independent validation preparations. Balanced accuracy for test sets is 90-92% (91%±1% on average). This results are very good. For validation preparations, a great result (almost 93%) is for preparation 1. In other cases, the result for the independent validation set is clearly worse than for the test set. The biggest difference is for preparation 5 (probably it contains a different cancer subtype) and for preparation 2 (the pathologist incorrectly assigned the lymphatic infiltration to the cancer class). However, balanced accuracy of 69-93% (81%±9% on average) obtained for validation sets is satisfactory.

Table 11.2

Results summary

		Independent validation preparation:				
		1	2	3	4	5
Number of features		204	201	206	202	217
Model building time [s]		63.01	43.94	43.22	63.54	68.43
Test set	PPV [%]	86.60	80.72	81.62	84.16	85.26
	NPV [%]	93.24	95.73	97.28	95.81	96.39
	Balanced accuracy [%]	90.26	90.18	92.05	91.47	92.35
Validation set	PPV [%]	46.58	77.34	82.69	76.27	63.22
	NPV [%]	99.89	76.29	87.49	90.21	79.38
	Balanced accuracy [%]	92.88	74.30	85.17	83.79	68.54

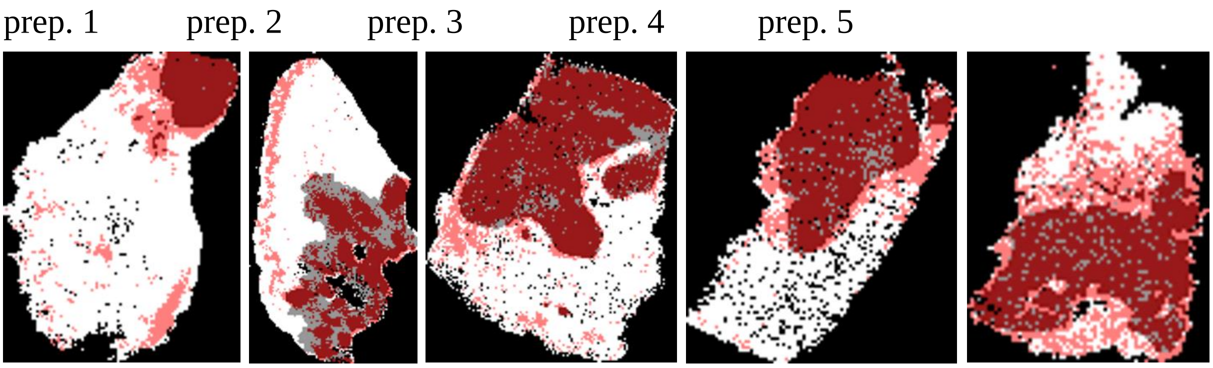

Fig. 11.4. Classification results for validation sets. Colors: TN – white, FP – light red, FN – gray, TP – dark red, background – black

Rys. 11.4. Wyniki klasyfikacji dla zbiorów walidacyjnych. Kolory: TN – biały, FP – jasny czerwony, FN – szary, TP – ciemny czerwony, tło – czarny

11.4. Discussion

11.4.1. Interpretation of the results

The goal was to find the signature of head and neck cancer. There were done 5 different splits, and from 3714 features only 277 were selected at least once to the classifier. In the Figure 11.5 there is a Venn diagram that shows how many features were common only for specific classifiers. For example, 15 features were selected only in the case when preparation 2 was the validation set (blue color on the very top). 44.4% of these 277 features were always selected, so the selection was quite repetitive and these 123 features are a representative part of this set.

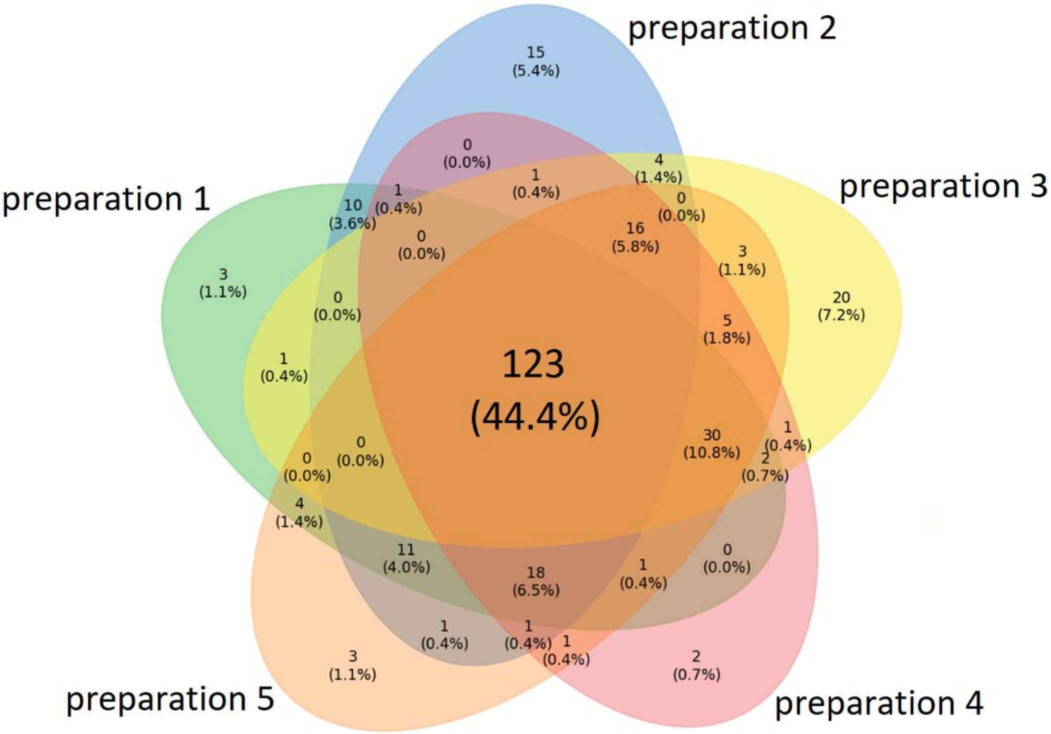

Fig. 11.5. A Venn diagram illustrating how many features were common only for specific classifiers
Rys. 11.5. Diagram Venna ilustrujący ile cech było wspólnych tylko dla konkretnych klasyfikatorów

In the Figure 11.6 there is an axis with marked mz values for 277 selected features. Most of them have small mz values. Probably mass channels with low mz values have bigger impact on healthy and diseased spectra distinction.

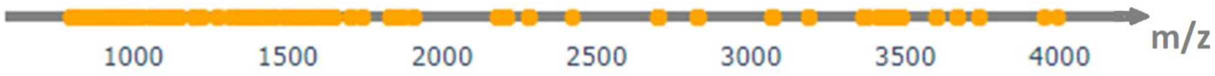

Fig. 11.6. mz values of 277 selected features
Rys. 11.6. Wartości mz dla 277 wybranych cech

In the left part of Figure 11.7 there is a correlation matrix between all features. The features can be divided into 2 main groups (the border is about feature number 2000) which have stronger correlation than with features from another group. In the right part of Figure 11.7 there is a correlation matrix between 277 features which are in at least one classifier. Most of them are from the group with smaller mz values. Selected features does not have much stronger correlation. Despite this, the classification results are satisfactory. If the number of preparations was greater, and if all of them were certainly the same cancer subtype, the results would probably be better.

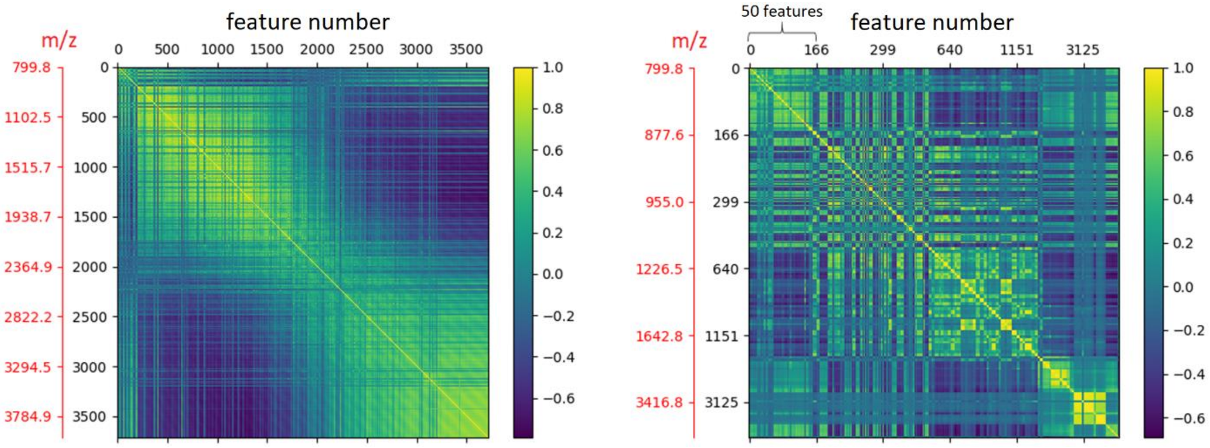

Fig. 11.7. Correlation matrices: between all 3714 features (left) and between 277 selected features (right)

Rys. 11.7. Macierze korelacji: pomiędzy wszystkimi 3714 cechami (z lewej) i pomiędzy 277 wybranymi cechami (z prawej)

11.4.2. Comparison with standard approaches

At the beginning, the data was logarithmic to decrease its numerical range. The second advantage of this solution is that the natural logarithm is an increasing monotone function. It means that the features that had very small values before logarithm, still have smaller values comparing to the other features. Another popular solution for data processing is normalization, for example Z standardization. It is subtracting the arithmetic mean of the data from each variable and dividing this difference by the standard deviation. This can be performed separately for each feature and then they become more "comparable" with each other. Normalization is a good method if the features are independent of each other, but in analyzed data features are protein values. If the feature has very small values, probably this protein is not essential to distinguish classes in the preparation. If normalization is used, unfortunately such features can have a huge impact on the model. Moreover, the numbers can still have big numerical range.

The reduction of feature number was big (more than 94% of them were rejected). Without feature selection, 5 hours after starting the system the model was still not obtained. This operation was necessary not only to speed up calculations, but also to find only the features that create the signature of head and neck cancer.

In the standard approach, the classifier does not calculate probabilities of data points belonging to each class – it only creates a hyperplane and assignments to the classes are made according to it. In used classifier there is implemented a default method for calculating probabilities. This method internally uses 5-fold cross-validation to calculate them [3]. For this method the best Youden's index was obtained for the threshold of 30% for validation preparation 5. The optimal threshold for the implemented method (probabilities calculated using the values returned by decision function) is 54%. In the Figure 11.8 there are heatmaps of cancer probabilities for both default and implemented method.

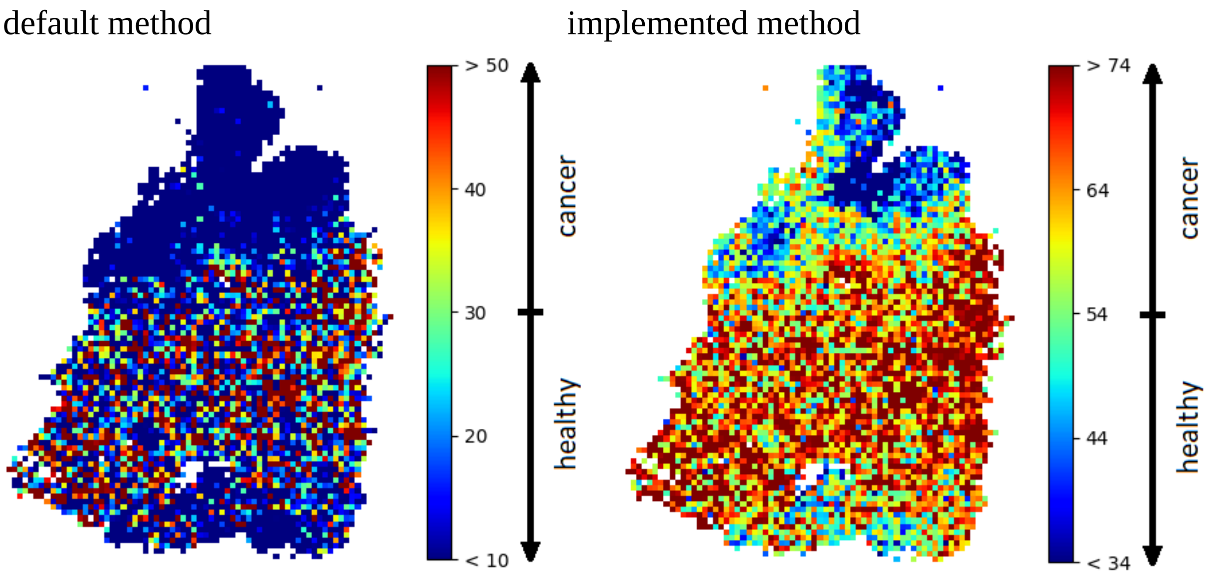

Fig. 11.8. Heatmaps of cancer probability [%] for validation preparation 5
Rys. 11.8. Heatmapy prawdopodobieństwa nowotworu [%] dla preparatu walidacyjnego 5

Comparison of the results for validation preparation 5 for 3 mentioned methods is shown in the Table 11.3 and in the Figure 11.9. The results without threshold detection are the worst and for method based on decision function they are the best. Without detection, the system tended to assign too many spectra to the healthy class (low sensitivity). For implemented method, balanced accuracy is more than 4 p.p. (percentage points) better than for the default method and more than 8 p.p. better than without calculating probabilities. Sensitivity is over 86% and F1 score is much higher that in other cases.

Model building time is much lower than in the default method because this method uses additional cross-validation.

Table 11.3

Results for validation preparation 5

	Without detection	Default method	Implemented method
Sensitivity [%]	25.73	38.61	86.70
Specificity [%]	94.34	90.03	50.38
Accuracy [%]	60.31	64.53	68.39
Balanced accuracy [%]	60.03	64.32	68.54
F1 score [%]	39.13	51.92	73.13
PPV [%]	81.72	79.21	63.22
NPV [%]	56.35	59.85	79.38
Model building time [s]	68.34	440.5	68.34

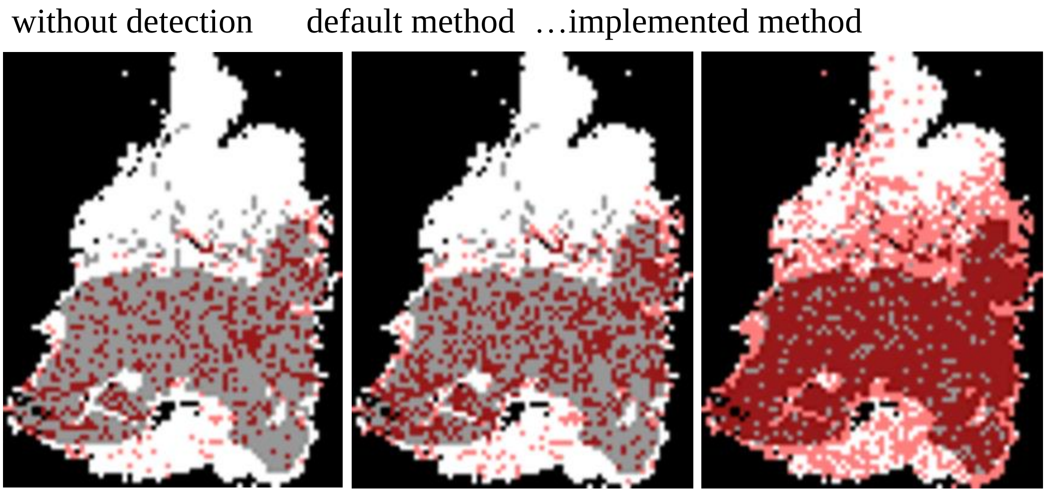

Fig. 11.9. Classification results for validation preparation 5. Colors: TN – white, FP – light red, FN – gray, TP – dark red, background – black

Rys. 11.9. Wyniki klasyfikacji dla preparatu walidacyjnego 5. Kolory: TN – biały, FP – jasny czerwony, FN – szary, TP – ciemny czerwony, tło – czarny

Acknowledgements

Work was partially financially supported by the Polish National Science Centre, project BiTIMS grant 2015/19/B/ST6/01736.

Bibliography

1. F. Bray, et al, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA Cancer J. Clin.* (2018) **68**:394-424.
2. Description of GMM method: https://scikit-learn.org/stable/modules/mixture.html, accessed: 16 March, 2021.
3. Description of the default method used by classifier to calculate probabilities: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html, accessed: 21 March, 2021.

MALDI-MSI MOLECULAR IMAGING DATA CLASSIFICATION SYSTEM

Abstract

The aim of the research was to create an intelligent classification system for head and neck cancer biopsy results based on the mass spectrometry protein profile. The material was molecular imaging data and information about the tissue type of samples taken from 5 patients (40 160 mass spectra in total, measured for 109 000 mass channels). Gaussian Mixture Model technique was applied to reduce the number of mass channels to 3714. Further reduction was obtained by selection of most variable features based on the variance distribution of peptides abundance in preparation. The SVM classifier was used and the optimal threshold detection method was implemented. The learning was performed in the process of cross-validation per patient, and the proportion of training and test set was 80:20. Results of the classification are satisfactory. Balanced accuracy for test set was 90-92% (91%±1% on average), for validation set was 69-93% (81%±9% on average).

Keywords: oncology, mass spectrometry, machine learning, classification.