Chemometrics and Intelligent Laboratory Systems 217 (2021) 104407

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Spatial sampling effect on data structure and random forest classification of tissue types in High Definition and Standard Definition FT-IR imaging

Danuta Liberda [a], Karolina Kosowska [a], Paulina Koziol [a,b], Tomasz P. Wrobel [a,*]

[a] Solaris National Synchrotron Radiation Centre, Jagiellonian University, Czerwone Maki 98, 30-392, Krakow, Poland
[b] Institute of Physics, Jagiellonian University, Łojasiewicza 11, 30-348, Kraków, Poland

ARTICLE INFO

Keywords:
FT-IR Imaging
Histology
Classification
Spatial resolution

ABSTRACT

Infrared radiation imaging (IR) combined with machine learning algorithms is one of the most promising techniques in tissue structures recognition and cancer diagnosis. Pancreatic tissue is heterogeneous and provides variable biochemical composition, as well as exhibits a variety of shapes and sizes having strong impact on physical aspects of light absorption. This research explores the potential of High and Standard Definitions in pancreatic tissue type prediction with random forest classification based on different spectral ranges. IR images spatial resolution depends on wavenumber and objective's numerical aperture. Therefore, data measured in Standard Definition (SD) are affected by under-sampling in the high wavenumbers while High Definition (HD) is free of this limitation for the whole analyzed spectral region. In order to investigate this effect a Random Forest model was created using a set of 56 biopsies measured in both definitions. In terms of signal variance the HD was found to be more spread out (however, with smaller scattering). SD models obtained a higher True Positive Rate than HD, however, spatial detail was better in HD models. Models trained on one definition and predicted on the other suggest that HD model can be used to successfully predict SD data. Finally, models based on only fingerprint and only high wavenumber were created to assess information content in each region. Models based on fingerprint region had very good prediction results across all classes, while for high wavenumber region results were class and definition dependent.

1. Introduction

Histopathology based on patterns of spatial tissue structures is the most commonly used technique in clinical cancer diagnosis. The tissue microenvironment is difficult to study even with the use of modern instrumental methods and the interpretation of the results is often ambiguous due to the complex chemical and spatial structure of tissues. Although cancer develops in the epithelium, it is common knowledge that the impact of the stroma on its progression is of significance [1–4]. In recent years, more and more groups of scientists have focused on research of this aspect, including groups developing modern diagnostic methods such as tissue classification using Fourier-Transform Infrared Spectroscopy (FT-IR) and machine learning [5–7]. FT-IR microscopy based methods connect tissue visualization at the same scale as traditional microscopy with analysis of the chemical composition. This allows for the isolation of individual tissue structures.

In our work, we focused on pancreatic tissues as a model for studying classification optimization. The pancreas is very heterogeneous, containing both cellular and amorphous structures. Structures such as fibroblasts, myofibroblasts, pancreatic stellate cells, immune cells, blood vessels, extracellular matrix, proteins, and growth factors can be distinguished in the stroma. The composition of the tissue is not static, it is constantly changing, especially during the progression of cancer or inflammation. Fibrous collagens, abundant in stroma, have a hierarchical molecular structure. The collagen molecule consists of three polypeptide alpha chains (~1.6 nm wide, ~300 nm long) that form the triple helical domain. The combination of the three alpha chains determines the type of collagen. Collagen molecules join both longitudinally and crosswise to form fibrils (width ~100 nm, length ~1 μm) thanks to covalent cross-linkages. Fibrils can then aggregate to form microfibers (~1 μm wide, ~10 μm long) [8,9]. The micrometric fibers are separable by high resolution IR microscopy. The above-mentioned red blood cells and lymphocytes are also found on a similar scale. Human red blood cells have 7.5–8.7 μm in diameter and 1.7–2.2 μm in thickness [10]. Lymphocytes include natural killer cells, T cells and B cells. The dense nucleus of a lymphocyte is approximately the size of a red blood cell [11]. As reported

* Corresponding author.
E-mail address: tomek.wrobel@uj.edu.pl (T.P. Wrobel).

https://doi.org/10.1016/j.chemolab.2021.104407
Received 2 April 2021; Received in revised form 14 June 2021; Accepted 25 August 2021
Available online 30 August 2021

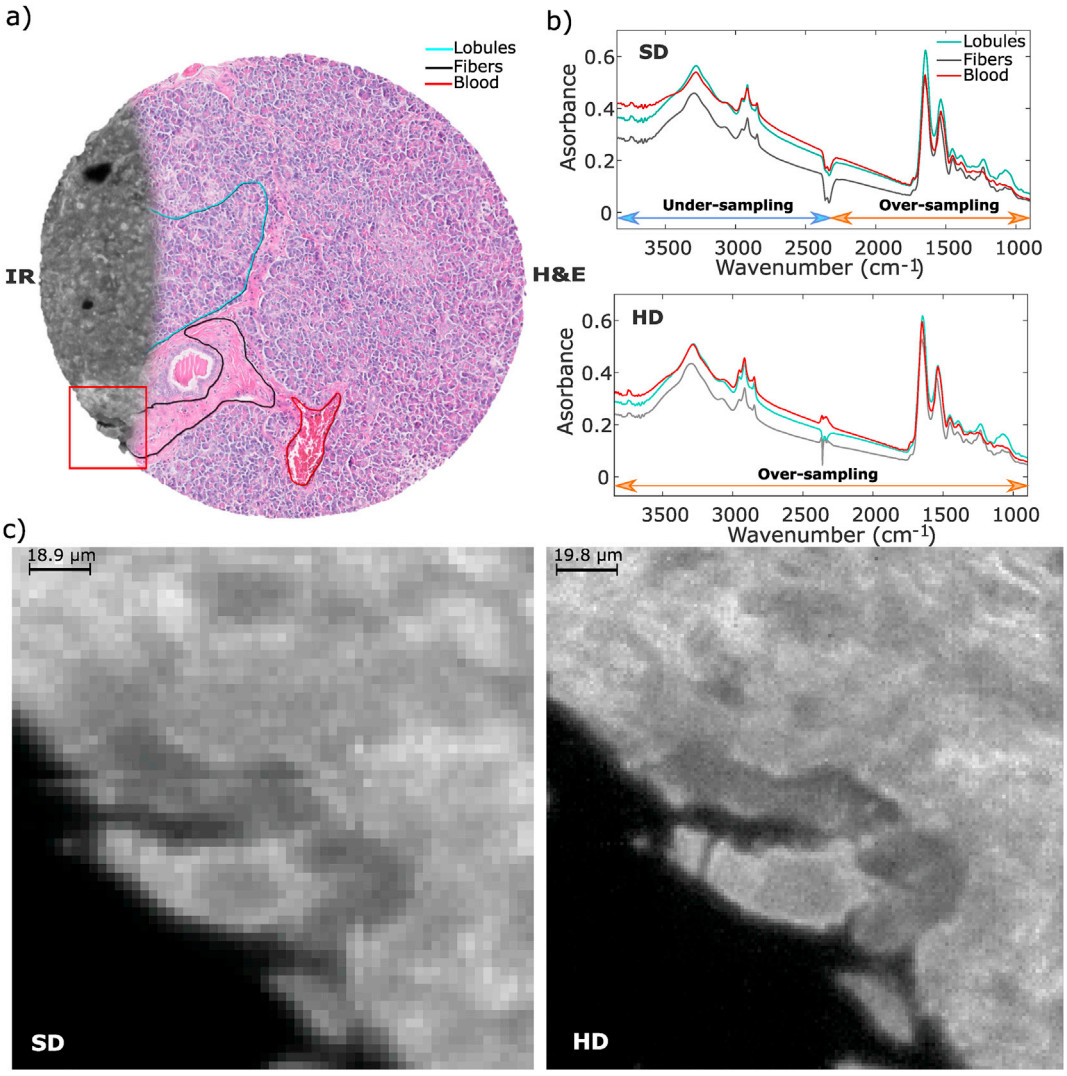

Fig. 1. Comparison of pancreatic tissue FT-IR imaged with different spatial resolution: a) H&E image of analyzed biopsy with overlayed part imaged in HD FT-IR, b) mean spectra of three classes: Lobules, Fibers, and Blood for SD and HD data set, c) SD and HD images for Amide A band (3282 cm^{-1}) of tissue part marked with red frame in subsection a). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

before [12,13], Random Forest classifier (RF) is a great tool for FT-IR based tissue classification and has been chosen to compare the different spatial resolution datasets. There are two aspects of this algorithm that make it robust for overfitting and irrelevant features, moreover, RF is scale-invariant. In the case of using a single decision tree the model can overfit when it optimizes to single set of observations and loses generalizability to new objects. Therefore, in RF high number of decision trees are used with different observation subsets, chosen from analyzed data through bagging procedure [14], as an input. When a new sample comes into the model, the number of votes (one from each tree) is calculated and indicates to which class the new sample belongs. Class with the majority of votes is a classification output. The second aspect is a selection of random subspaces of features (equal to the square root of total variables number) for each decision tree, which prevents the model against decision trees correlation. Considering above advantages of RF and also the fact that it retains information about the importance of individual features in model creation, we found it to be sufficient for assessment of classification potential of data measured in HD and SD definition.

In order to visualize fine biological structures and investigate the influence of data quality and resolution on classifier performance, FT-IR Imaging was used in this study with two objectives (36x, $NA = 0.5$ and 15x, $NA = 0.4$) giving different projected pixel sizes (1.1 μm and 2.7 μm) and spatial resolutions. However, pixel size should not be confused with image resolution - as it often happens for this modality. Due to the diffraction limit, image resolution is wavelength dependent and defined by the Rayleigh criterion as follows:

$$d = \frac{0.61 \cdot \lambda}{NA},$$

where d describes limit of resolution (minimal distance between two resolvable points), λ is a wavelength and NA describes numerical aperture of an objective. As easily deducible, better resolution is achieved for shorter wavelengths and higher NA. However, this is a general rule derived for simple circular aperture. Fortunately, a model for light propagation through FT-IR system with Cassegrain objectives and Focal Plane Array was thoroughly derived by means of Fourier optics [15]. One of the conclusions defined optimal sampling rate x (pixel size in other words) to be:

$$x = \frac{0.25 \cdot \lambda}{NA},$$

ensuring proper collection of signal's spatial frequencies. This states that

Table 1
List of top 20 metrics for SD and HD data sets sort with descending importance in RF classification with their spectral ranges. Metrics with numbers from 1:20 range correspond to maximum band intensity, 21:40 correspond to band center of gravity, and 41:60 band integration. Metrics which were important in classification of both SD and HD datasets are marked with an asterix.

SD		HD	
Metric no.	Spectral range (cm^{-1})	Metric no.	Spectral range (cm^{-1})
44*	1010–1155	**11**	1360–1390
4*	1010–1155	**44***	1010–1155
6	1190–1210	**48***	1190–1290
14	1490–1590	**32**	1360–1425
27*	1210–1280	**51**	1360–1390
46	1190–1210	**9***	1295–1320
49*	1295–1320	**49***	1295–1320
45	1160–1180	**27***	1210–1280
8	1190–1290	**4***	1010–1155
19	3025–3100	**8***	1190–1290
34	1490–1590	**35**	1710–1760
54	1490–1590	**1***	940–980
7*	1210–1280	**25**	1160–1180
28*	1190–1290	**38**	3000–3025
5	1160–1180	**12**	1360–1425
1*	940–980	**28***	1190–1290
59	3025–3100	**29**	1295–1320
48*	1190–1290	**7***	1210–1280
9*	1295–1320	**21**	940–980
41	940–980	**33**	1425–1480

the optimal sampling rate is 2.44 times the limiting spatial frequency of the signal, in contrary to the general Nyquist criterion where the factor is 2. Application of this analysis to the $NA = 0.5$ objective with projected pixel size of 1.1 μm (specific pixel sizes x and used NAs), gives a limit for wavelengths (images) around $\lambda = 2.2$ μm (4545 cm^{-1}). This means that for this objective, the sampling rate is optimal throughout the full mid-IR spectral range. Meeting this criterion allows to define it as High Definition (HD) objective since all spatial frequencies are over-sampled. For the second objective ($NA = 0.4$ and projected pixel size of 2.7 μm) optimal sampling is met for $\lambda = 4.32$ μm (2314 cm^{-1}). Which means it is only optimal for fingerprint region while signal loss caused by under-sampling may occur in high spectral region. Therefore this objective will be termed as Standard Definition (SD), even though parts of the spectrum are sampled at HD. This situation is schematically put in Fig. 1.

Large pixels biochemically average the heterogeneous tissue environment. The HD imaging allows obtaining smaller pixels (1.1 μm), although of course it is associated with longer analysis time. This technology is still being developed and improved. However, there is no evidence yet that HD quality IR imaging gives an analytical advantage beyond image quality, which is of value in itself. It has been shown that HD gives comparable models as SD for breast tissue [4] and that different spectral regions offer varying levels of classifier performance [16]. However, these studies compared only maximal performance of each of the modalities internally, without exploring cross-definition information content. Therefore, in this study, we compared SD and HD imaging using the same Pancreatic Tissue Micro Array (TMA) to explore whether HD will create a more robust model, as it encompasses all the information that SD provides with additional features resolved.

2. Materials and methods

2.1. Sample preparation and measurements

Paraffin-embedded PA2081b pancreatic Tissue Micro Array (TMA) was purchased from Biomax Inc. One 5 μm thick TMA slice was mounted on a regular glass, and stained with Eosin and Hematoxylin (H&E) for histopathological annotation. Second slice used for FT-IR measurements was placed on a BaF_2 salt plate and deparaffinized with 24 h hexane bath.

56 biopsies coming from 28 patients were measured in transmission mode with Bruker Vertex70v Spectrometer coupled with Hyperion 3000 microscope with a 64x64 FPA detector, 15x objective (2.7 μm projected pixel size, NA = 0.4) and 36x objective (1.1 μm projected pixel size, NA = 0.5). Spectra acquisition was done in 3850 cm^{-1} - 900 cm^{-1} spectral range, with 8 cm^{-1} spectral resolution and zero filling factor of 1. Sample and background signals were co-averaged 4 and 64 times, respectively.

2.2. Data preprocessing

To increase spectral and spatial quality, data were denoised with the Minimum Noise Fraction method (MNF), with 15 bands used for reconstructions [17,18]. Before Principal Component Analysis (PCA), Rubberband baseline correction was applied where each spectrum was divided into segments with points: 3598, 3097, 3000, 2808, 1758, 1712, 1601, 1477, 1427, 1358, 1288, 1716, 1149, 980, 941 cm^{-1} determining for linear baseline subtraction. Following, CO_2 region (2804-1763 cm^{-1}) was removed, and spectra were normalized to Amide I band 1650 cm^{-1} frequency. Data was mean centered.

Whole spectral region based classification was done with metrics (based on previously well-defined bands – listed in Table 1 in Supplementary Materials) determined as: band maximum value, band integration and center of band gravity. Each of this metrics was normalized to corresponding type of Amide I band metric. Metrics for spectral data in the range 4000–2700 cm^{-1} (high wavenumbers) and 1900-800 cm^{-1} (fingerprint region) were created separately. The same number of metrics (112) was generated for both regions, using 8 spectral ranges each (listed in Table 1 in Supplementary Materials). Metrics were calculated as ratios of band maximum absorbance, band integration, and center of band gravity, between an individual spectral range and all remaining spectral ranges. Furthermore, additional type of metric was introduced, calculated as ratio of specific band maximum absorbance and remaining bands integrations. The same process of metrics creation was applied for HD and SD data. Images of the TMA were created based on selected wavelength of Amide I after denoising. Figures were prepared to show the pixel distribution of classes, and mark biological structures based on histological annotations on H&E stained TMA image.

2.3. Classification

Random Forest Classification at the pixel level was done with fivefold cross-validation at the patient level. Following number of patients: 5, 5, 6, 6, 6 were randomly chosen to test set in each fold. The same patients were chosen in SD and HD test sets. Metrics were used as Random Forest classification input. Classifications were done for the whole, fingerprint and high wavenumber spectral region. Additionally cross-definition models:

- model based on HD model set with prediction on SD test set,
- model based on SD model set with prediction on HD test set,

were created, to check classification ability of HD and SD data sets to predicted other definition. The degree of overlap of these models will have an impact on the measurement strategy for larger scale measurements and clinical translation.

3. Results and discussion

3.1. Data exploration

Pancreatic healthy tissue is dominated by acini cells arranged in lobules, the fibrous connective tissue surrounding lobules, excretory ducts, and interweaving blood vessels [19]. Those structures are relatively easy to distinguish by a histopathologist using hematoxylin and eosin (H&E) stained tissue microscopic image. In Fig. 1a a healthy

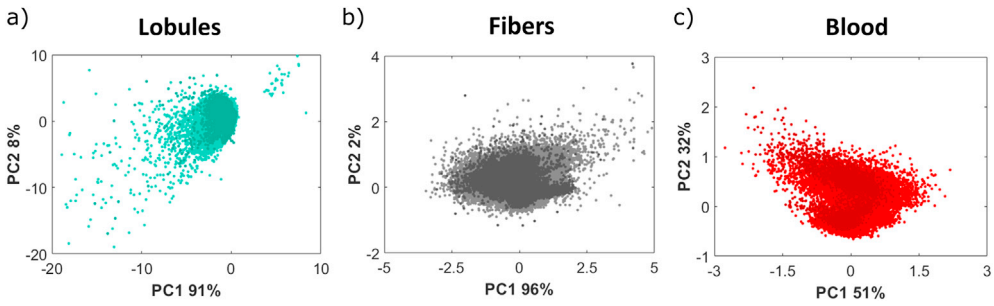

Fig. 2. Score plots of Principal Component Analysis for HD (brighter color palette) and SD (darker color palette) data sets for: a) Lobules, b) Fibers and c) Blood spectra. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

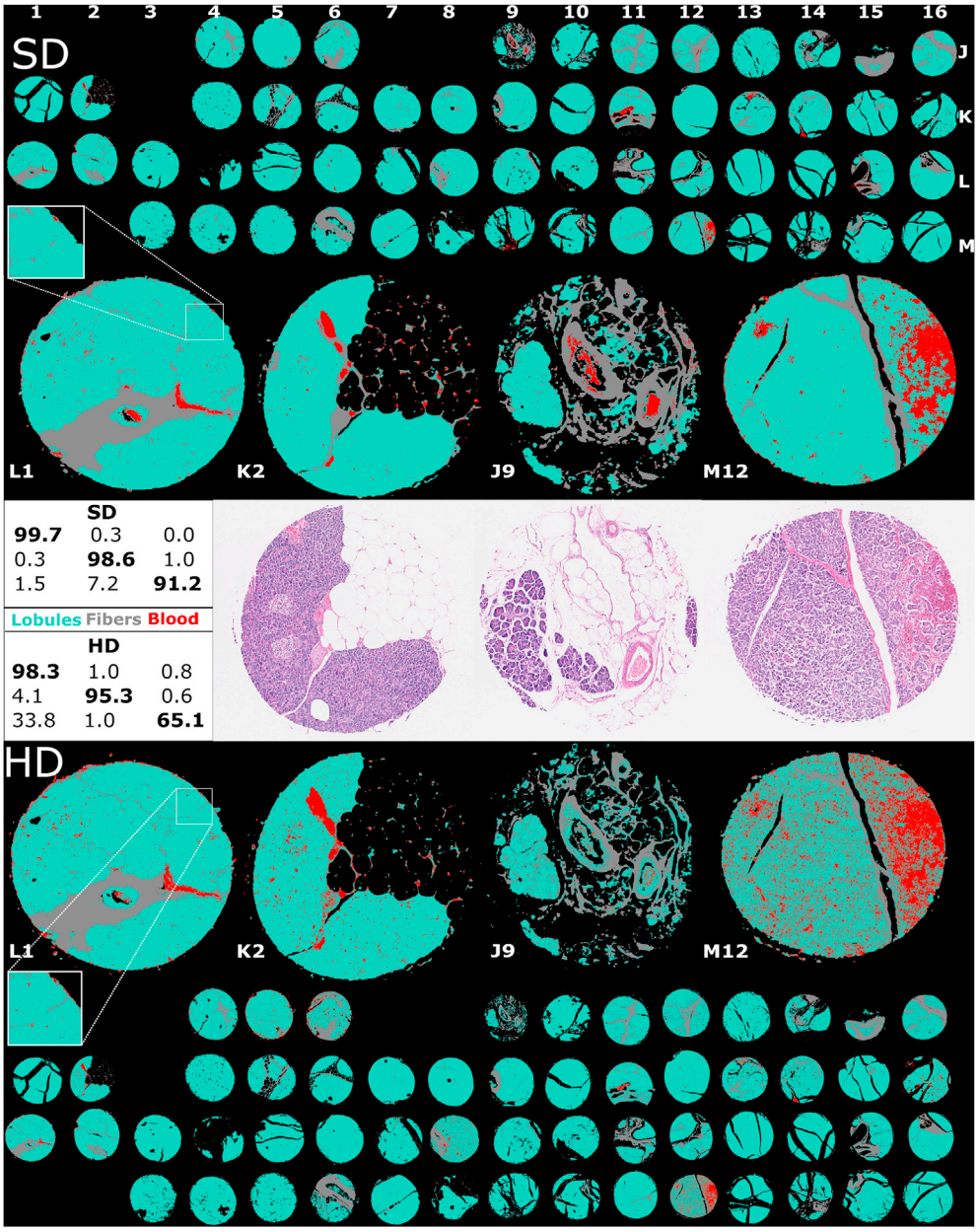

Fig. 3. Classification results of fivefold cross-validation for predicted images: SD upper figure panel and HD lower figure panel. Confusion matrices and H&E images for chosen biopsies are presented in the middle figure panel.

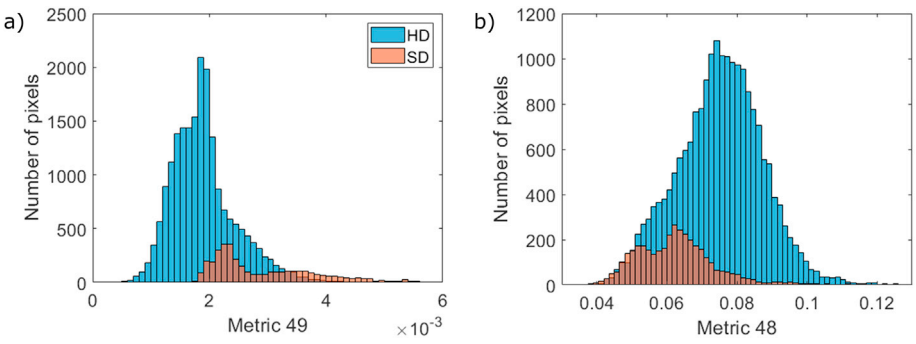

Fig. 4. Histograms of Blood class pixels distributions for metric: a) 49 and b) 48.

pancreatic H&E stained tissue image with histopathological annotations of three classes: Lobules, Fibres, and Blood, is presented. Moreover, in the same Fig. 1a part of tissue slice imaged with FT-IR in HD is overlaid, using Amide A frequency at 3282 cm^{-1}. Mean spectra for analyzed classes are compared in Fig. 1b. Different sizes and shapes of analyzed structures influence physical effects distorting spectra, i.e. light scattering. It can be observed based on the baseline shape that scattering is increased for blood cells, which is expected because of their circular shape. The scattering effect in spectra measured in SD is slightly higher than in HD, which can be explained by different NA of the objectives. From a classification point of view, differences in band shapes and positions between analyzed classes influence the classifier outcome and should be compared. In the raw spectra for HD and SD data sets, differences between classes in the fingerprint region, especially in bands shape/-position of DNA/RNA region (940-1155 cm^{-1} and 1180-1295 cm^{-1}) and proteins - Amide I/II (1490-1700 cm^{-1}), can be observed. The spatial resolution of FT-IR images influences the visibility of small tissue structures like blood cells, therefore differences between HD and SD images were investigated in Fig. 1c. Structure evident in Fig. 1c - proteinaceous fluid, is much better visible in HD image than SD. In HD FT-IR image fibers and their direction in connective tissue can be conveniently distinguished.

For further exploration of differences between the internal data structure of HD and SD for a given class, the Principal Component Analysis (PCA) was applied on baseline corrected and normalized spectra (results presented in Fig. 2). SD pixels (darker color palette) for all classes are more concentrated in PC1 and PC2 space than in the case of HD data (brighter color palette) where points are highly spread out. For each class SD pixels are contained in space covered by HD pixels, therefore, the variability of the HD data set is much higher than for the SD data set. It is interesting that each class differs in terms of the spreads – the largest differences are present for Blood class. Loadings for PC1 and PC2 for each class are available in Supplementary Materials.

3.2. Random forest classification for the whole spectral range

Classification results in the form of predicted FT-IR SD and HD images along with confusion matrices for each resolution are presented in Fig. 3. When predictions for all 28 patients are under consideration, the general conclusion is that tissue structures predicted for the SD data set are more consistent than in the case of HD data. Confusion matrices support this, as the highest True Positive Rate (TPR) for all classes is achieved by the SD model. The class with the worst prediction is Blood (only 65% TPR for HD data set). For majority of cores the predicted images are of very similar quality. However, some differences can be observed. M12 biopsy predicted image is a good example of such situation. In the case of SD data Lobules tissue is more consistent, also Blood class highly corresponds to the dense blood regions in H&E image, while in the predicted HD biopsy image, additional pixels predicted as Blood are present. Therefore, it may be concluded that HD model has the ability to detect much smaller blood cells clumped in blood vessels. Low Blood TPR values for HD data set can be connected with classification as Blood pixels from background (between blood cells) are hard to remove. In HD image improper prediction is visible in biopsies J5 and J6 where their edges are classified as Blood - it may result from higher light scattering of the biopsies edges. The predicted image of biopsy J9 is a good example of correct pixel prediction for Blood class in SD image (blood placed in vessels is very consistent), and much worse prediction of HD image where only individual erythrocytes are predicted properly. Nevertheless, there are also cases, e.g. K2 biopsy, where HD image presents a better prediction results for Blood class. In the H&E image, multiple blood cells and small blood vessels are visible. The HD model is more sensitive to these small structures and is able to predict them between lobules in the image. Moreover, in the H&E image of this biopsy, a nerve is visible and

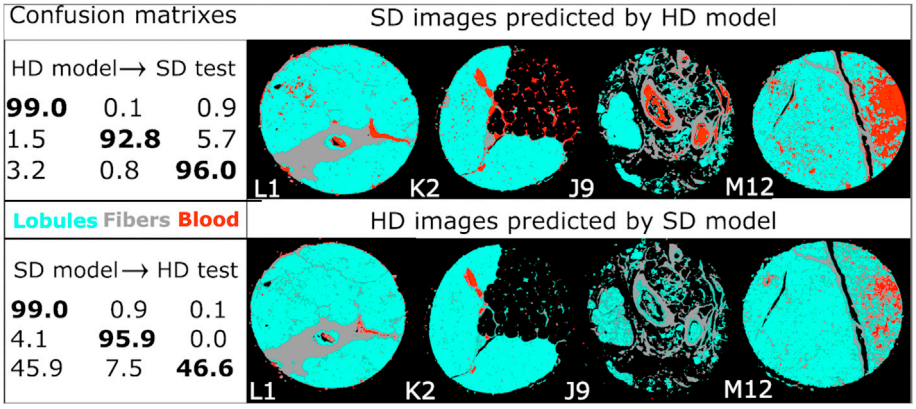

Fig. 5. Cross-definitions models classification results for predicted images: SD upper figure panel and HD lower figure panel. Confusion matrices are presented in left figure panel.

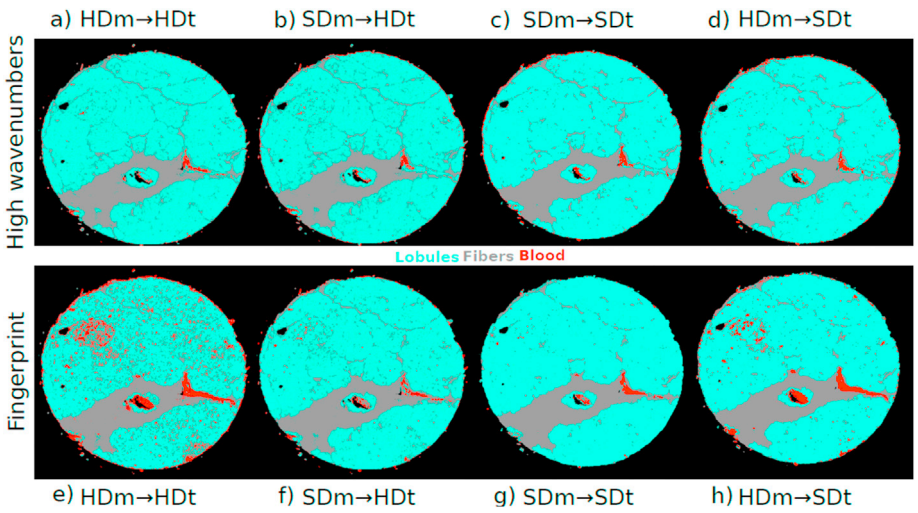

Fig. 6. Classification results for high wavenumbers spectral region: a) HD model (HDm) predicted on HD data (HDt), b) SD model (SDm) predicted on HD data, c) SD model predicted on SD data (SDt), d) HD model predicted on SD data; and low wavenumbers spectral region: e) HD model predicted on HD data, f) SD model predicted on HD data, g) SD model predicted on SD data, h) SD model predicted on HD data.

it is interesting that in the SD predicted image nerve was classified as Fibers, while in HD image nerve is predicted as Blood and Lobules, which shows the more cellular affinity of this definition. The last considered biopsy L1 presents differences in the prediction of small connective tissue originated fibers: HD data set allows the detection of small fibers between Lobules which cannot be observed in SD predicted images.

Random Forest offers easy insight into variable importance. One would expect differences between variables utilized by models based on different data structures. The importance of metrics calculated for earlier established spectral regions were inspected (Table 1). Taking into account 20 of the best metrics which have the highest importance for classification, 10 of them are important for both SD and HD resolutions. Almost all of these metrics come from the fingerprint region. In the HD data set only one metric is based on 3000-3025 cm^{-1} high-frequency spectral range, while for SD data two metrics origin from 3025 to 3100 cm^{-1} spectral range. It can be concluded that there are some dissimilarities in class differentiation in metrics space for both data sets. Therefore, we decided to study it closer, to check if it is possible to create a classifier that will be able to properly predict data, based on lower number of metrics coming from high wavenumber or fingerprint spectral region.

Taking into account, that in RF in each tree node a single metric is used for split (finally leading to an object assignment to a class in leaf node), distributions of pixels coming from class Blood for two individual metrics were inspected in Fig. 4. We compared also pixels distribution for class Blood for 10 common metrics for HD and SD data set (presented in Supplementary Materials). For the majority of metrics the pixels distributions are similar to metric no. 48 in Fig. 4b (ratio of integration of 1190–1290 cm^{-1} band to amide I integration), where distribution of pixels coming from HD is covering SD pixel distribution. However, there are also pixels distributions similar to metric no. 49 (ratio of integration of 1295–1320 cm^{-1} band to Amide I integration) in Fig. 4a where most of the SD pixels are covered by HD pixels distribution, but it is slightly moved in direction of higher or lower metrics values.

Finally, to check if data space created by the SD model has the ability to cover variation in HD data set and vice versa, we explored classification results for two models. The first one was trained on SD data set, and prediction on HD data set was done; the second one was trained on HD data set, and prediction on SD data set was done. Comparison of prediction results for those cross-definition models for HD and SD data sets (for the same biopsies chosen in Fig. 3) was done and presented in Fig. 5. HD model predicting SD data set (upper section in Fig. 5) has the ability to correctly classify a higher number of blood cells than SD model (upper section in Fig. 3), and at the same time it retains high consistency for Lobules and Fibers classes. Prediction results for SD model on HD data set (lower section in Fig. 5) are worse than for HD model (lower section in Fig. 3). Small fibers in biopsy L1 are predicted properly but it can be observed in the example of biopsy J9 that Lobules are frequently predicted as Fibers. Furthermore, prediction results for class Blood are much worse than for models created using HD dataset, which is supported by low Blood class TPR value.

Table 2
True Positive Rate (TPR) values for all model-test combinations for high wavenumber (HW) and fingerprint (FP) regions for analyzed classes: Lobules, Fibers, Blood. With green and red color the highest and the lowest TPR values for each class are marked by an asterix.

Model_Test_region	Lobules	Fibers	Blood
HD_HD_HW	95.5	94.5	24.3
SD_SD_HW	98.0	97.1	**20.9***
HD_SD_HW	97.8	94.4	35.0
SD_HD_HW	**94.3***	93.3	22.1
HD_HD_FP	97.3	95.5	74.3
SD_SD_FP	**99.7***	**99.1***	92.6
HD_SD_FP	98.3	**90.7***	**98.2***
SD_HD_FP	98.9	95.6	49.2

3.3. Random forest classification for fingerprint and high wavenumber ranges

Considering pixel size and sampling, which in the case of HD imaging is sufficient for the full mid-IR spectral range, whereas in the case of SD is only optimal for fingerprint region, classification results for two data sets with metrics based on: high wavenumber (HW - 4000-2700 cm^{-1}) and fingerprint (FP - 1900-800 cm^{-1}) regions were investigated. In Fig. 6 prediction results on the biopsy L1 image for standard and cross-definitions models created for HW and FP regions are shown. In the case of images predicted for the HW data sets, even small fibers are properly classified but blood cells are poorly classified, while for FP data sets those classification results are reverse. For HW data, HD cross-definition model with a prediction on SD data (Fig. 6d) provides slightly better classification results for small fibers than in the case of a standard model where prediction on SD data is done with SD model (Fig. 6c). In the case of prediction on HD image for HW data, both standard (Fig. 6a) and cross-definition (Fig. 6b) models classify small fibers very well. HD models based on FP metrics (Fig. 6e,h) can properly

predict clustered and individual blood cells as in the case of previously investigated whole spectral range data. Furthermore, when the HD model is applied to predict SD image (Fig. 6h) improvement in Blood class classification in comparison to the SD model (Fig. 6g) is evident. SD model (Fig. 6f) is able to predict class Fibers and Lobules on HD image very well but prediction of Blood cell class is the worst in comparison to remaining FP predicted images.

When TPR (Table 2) for all biopsies is considered it is clear that the best classification results are achieved for FP data sets. Cross-definition HD models predicted on SD data set give the highest TPR for Blood class within FP and HW data models. The worst TPR for the Blood class is given by HW cross-definition SD model predicted on SD data set. SD data set predicted by SD model gives the highest TPR for Lobules and Fibers classes but it can be related to the very high consistency of these structures – small blood cells between lobules and fibers were not detected by the model.

4. Conclusions

The data exploration step has shown differences in variability of spectra for individual classes in HD and SD datasets. HD dataset was found to be more differentiated - covering broader PC1 and PC2 space. Classification results based on the whole spectral region have shown that SD dataset gave higher TPR and more consistent tissue prediction, while HD dataset allows for the prediction of small blood cells and fibers. When cross-definition models are considered, it is clear that SD model is not able to cover the variability of Blood class in the predicted HD dataset. However, HD model significantly improves Blood class prediction on SD dataset. Classification results for high wavenumber region indicate that models based on under-sampled SD dataset give the worse prediction results for Blood class on HD and SD datasets. In the case of the fingerprint region, SD model is able to predict Blood class on SD image with a high TPR above 90%. This indicates that HD imaging in the high wavenumber region has a higher ability to properly predict small structures like blood cells, while in SD imaging sampling rate is not sufficient and gives worse classification results. Therefore, we suggest, if the research goal is recognition of individual blood cells or similar size tissue structures, HD definition should be applied. Especially if biochemical information critical for classification appears in high wavenumber region. When detection of tissue types and boundaries between them is considered, SD definition will be sufficient and allows for much faster measurement performance, which is critical for clinical application.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported the National Science Centre, Poland ("Three-dimensional macromolecule orientation by means of infrared (IR) imaging and its significance in cancer microenvironment", Grant No. 2018/31/D/ST4/01833). This research was performed using equipment purchased in the frame of the project co-funded by the Malopolska Regional Operational Program Measure 5.1 Krakow Metropolitan Area as an important hub of the European Research Area for 2007–2013, project no. MRPO.05.01.00-12-013/15.The authors declare no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2021.104407.

References

[1] M.A. Kiskowski, R.S. Jackson, J. Banerjee, X. Li, M. Kang, J.M. Iturregui, O.E. Franco, S.W. Hayward, N.A. Bhowmick, Role for stromal heterogeneity in prostate tumorigenesis, Canc. Res. 71 (2011) 3459–3470, https://doi.org/10.1158/0008-5472.CAN-10-2999.

[2] M.S. Condon, The role of the stromal microenvironment in prostate cancer, Semin. Canc. Biol. 15 (2005) 132–137, https://doi.org/10.1016/j.semcancer.2004.08.002.

[3] M. Fanous, A. Keikhosravi, A. Kajdacsy-Balla, K.W. Eliceiri, G. Popescu, Quantitative phase imaging of stromal prognostic markers in pancreatic ductal adenocarcinoma, Biomed. Opt Express 11 (2020) 1354, https://doi.org/10.1364/boe.383242.

[4] S. Mittal, K. Yeh, L. Suzanne Leslie, S. Kenkel, A. Kajdacsy-Balla, R. Bhargava, Simultaneous cancer and tumor microenvironment subtyping using confocal infrared microscopy for all-digital molecular histopathology, Proc. Natl. Acad. Sci. U.S.A. 115 (2018) E5651–E5660, https://doi.org/10.1073/pnas.1719551115.

[5] S. Mittal, C. Stoean, A. Kajdacsy-Balla, R. Bhargava, Digital assessment of stained breast tissue images for comprehensive tumor and microenvironment analysis, Front. Bioeng. Biotechnol. 7 (2019) 1–9, https://doi.org/10.3389/fbioe.2019.00246.

[6] J.T. Kwak, A. Kajdacsy-Balla, V. Macias, M. Walsh, S. Sinha, R. Bhargava, Improving prediction of prostate cancer recurrence using chemical imaging, Sci. Rep. 5 (2015) 1–10, https://doi.org/10.1038/srep08758.

[7] S. Kumar, T.S. Shabi, E. Goormaghtigh, A FTIR imaging characterization of fibroblasts stimulated by various breast cancer cell lines, PloS One 9 (2014), https://doi.org/10.1371/journal.pone.0111137.

[8] J.N. Ouellette, C.R. Drifka, K.B. Pointer, Y. Liu, T.J. Lieberthal, W.J. Kao, J.S. Kuo, A.G. Loeffler, K.W. Eliceiri, Navigating the collagen jungle: the biomedical potential of fiber organization in cancer, Bioengineering 8 (2021) 1–19, https://doi.org/10.3390/bioengineering8020017.

[9] A. Gautieri, S. Vesentini, A. Redaelli, M.J. Buehler, Hierarchical structure and nanomechanics of collagen microfibrils from the atomistic scale up, Nano Lett. 11 (2011) 757–766, https://doi.org/10.1021/nl103943u.

[10] M. Diez-Silva, M. Dao, J. Han, C. Lim, S. Suresh, Shape and biomechanical human red blood cells in health, MRS Bull. 35 (2010) 382–388, https://doi.org/10.1557/mrs2010.571.

[11] E.J. Wood, Cellular and molecular immunology, in: fifth ed., in: A.K. Abbas, A.H. Lichtman (Eds.), Biochem. Mol. Biol. Educ., 32, 2004, pp. 65–66, https://doi.org/10.1002/bmb.2004.494032019997.

[12] T.P. Wrobel, P. Mukherjee, R. Bhargava, Rapid visualization of macromolecular orientation by discrete frequency mid-infrared spectroscopic imaging, Analyst 142 (2017) 75–79, https://doi.org/10.1039/c6an01086e.

[13] D. Liberda, M. Hermes, P. Koziol, N. Stone, T.P. Wrobel, Translation of an esophagus histopathological FT-IR imaging model to a fast quantum cascade laser modality, J. Biophot. (2020) 1–9, https://doi.org/10.1002/jbio.202000122.

[14] R. Richman, M.V. Wuthrich, Nagging Predictors, Risks, 8, https://doi.org/10.3390/risks8030083, 2020, 83.

[15] R.K. Reddy, M.J. Walsh, M.V. Schulmerich, P.S. Carney, R. Bhargava, High-definition infrared spectroscopic imaging, Appl. Spectrosc. 67 (2013) 93–105, https://doi.org/10.1366/11-06568.

[16] S. Mittal, R. Bhargava, A comparison of mid-infrared spectral regions on accuracy of tissue classification, Analyst 144 (2019) 2635–2642, https://doi.org/10.1039/c8an01782d.

[17] P. Koziol, M.K. Raczkowska, J. Skibinska, S. Urbaniak-Wasik, C. Paluszkiewicz, W. Kwiatek, T.P. Wrobel, Comparison of spectral and spatial denoising techniques in the context of High Definition FT-IR imaging hyperspectral data, Sci. Rep. 8 (2018) 1–11, https://doi.org/10.1038/s41598-018-32713-7.

[18] M.K. Raczkowska, P. Koziol, S. Urbaniak, C. Paluszkiewicz, W.M. Kwiatek, T.P. Wrobel, Influence of denoising on classification results in the context of hyperspectral data : high Definition FT-IR imaging, Anal. Chim. Acta 1085 (2019) 39–47, https://doi.org/10.1016/j.aca.2019.07.045.

[19] D.C.W. Hans, G. Beger, Andrew L. Warshaw, Ralph H. Hruban, Markus W. Buchler, Markus M. Lerch, John P. Neoptolemos, The Pancreas: an Integrated Textbook of Basic Science, Medicine, and Surgery, John Wiley & Sons, 2018.