EPJ Web of Conferences **195**, 10006 (2018) https://doi.org/10.1051/epjconf/201819510006
TERA-2018

Applications of THz laser spectroscopy and machine learning for medical diagnostics

Yu. V. Kistenev[1,2], A. V. Borisov[1,2], A.I.Knyazkova[1], E. A. Sandykova[1,2], V. V. Nikolaev[1], D. A. Vrazhnov[1]

[1]Tomsk State University, Tomsk, Russia, yuk@iao.ru
[2]Siberian State Medical University, Tomsk, Russia

THz spectroscopy allows to analyze molecular rotations associated with hydrogen bond breaking. But, the identification of pure compounds using molecular signatures with THz spectroscopy is still not straightforward because of the inherently broad spectral signatures in biotissue. A smooth shape THz of spectra causes a necessity to use the machine learning for tissue diagnosis using THz spectroscopy.

Typical machine learning pipeline includes the following steps (Fig.1) [1]:

- preprocessing of data;
- selection of informative features;
- development of predictive models for new data classification.

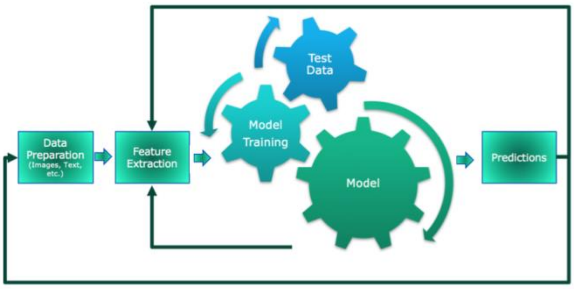


Fig.1. Typical machine learning pipeline.

In this report we present examples of peculiarities of machine learning applications for 2D THz image analysis. The results have been obtained using THz time-domain spectrometer (TDS) T-Spec by EXPLA in the range 0.2-3 THz.

At the preprocessing stage, in addition to filtering, various approaches to allocating areas of interest also should be considered. Automatic search of characteristic structures in the image is based on their formalized mathematical description of the image textures.

The example of preprocessing stage image transform connected with 2D THz TDS absorption spectra of formalin-fixed paraffin-embedded prostate cancer biopsy tissues is presented below. The goal is to remove artifacts of plastic substrate and paraffin from the image.

The Fig. 2 shows the spatial distribution of 2D THz image for a paraffin block without a sample and for a plastic substrate at frequencies 0.90 THz (Fig. 1a) and 1.05 THz (Fig. 1b).

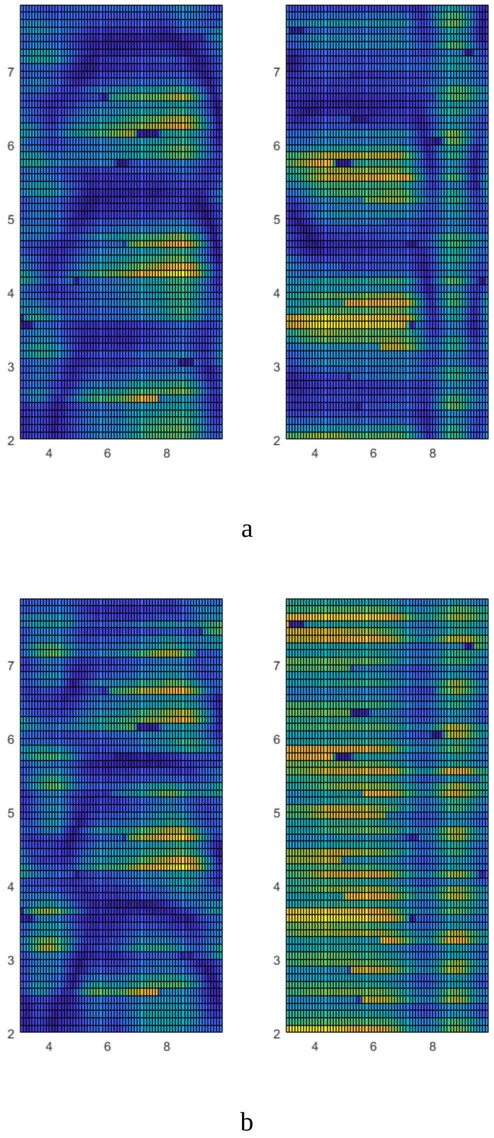
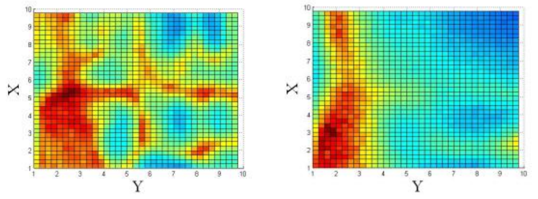


Fig. 2. Spatial distributions of the THz signal intensity for a paraffin block without a sample (on the left in each figure) and for a plastic substrate (on the right in each figure) at the frequencies: a - 0.90 THz and b - 1.05 THz.

The difference of absorption spectra allows to remove similar artifacts from the image. To realize it, an optimization algorithm was developed and implemented [3]. This algorithm allows to select pixels on the THz image with minimal influence of the paraffin and plastic substrates. The results of selection of a biopsy tissue on the 2D THz image are shown in the Fig.3.



a b

Fig.3. Selection of a biopsy tissue on the 2D THz image: initial image (a), image after artifacts removing (b).

A key step in image analysis is the informative features selection, because the quality of the created predictive model is defined by the ability of spatial separation of the various groups under study in the feature space. One of the most effective method of this task solution is the principal component analysis (PCA). The basic idea of PCA is to find the reduced number of new variables, termed the principal components, which are enough for recovery of the initial variables, possibly with insignificant errors [2].

The PCA applications for 2D THz image analysis was done on animal model (rats) lymphedema tissue.

Lymphedema is a chronic progressive disease of the lymphatic system caused by abnormal accumulation of tissue fluid with a high protein content. Early diagnosis of this disease helps to choose the right treatment and prevent its further development. The existing methods of lymphedema diagnosing at early stages are not strict and consistent. The invention of THz microscopy opens up new possibilities for lymphedema tissue analysis in vivo.

Using the optimization algorithm, mentioned above, we carried out the classification of THz spectra of the most informative areas obtained in-vivo from the lymphedema affected leg tissue (result of surgery) and obtained from healthy leg tissue. The results show good enough the separation of lymphedema tissues from and healthy tissues in the space of the principal components (see Fig. 4).

The principal components are built using the THz spectra in the 0.8-1.0 THz spectral range. Note that the separation of the groups using THz imaging became possible after three weeks from the lymphedema surgery initiation.

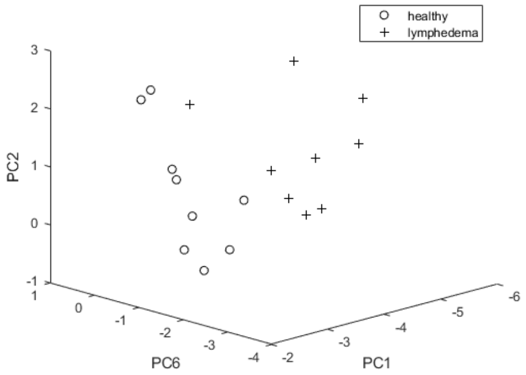


Fig. 4. Spatial distribution of the THZ spectra from lymphedema affected leg tissues and healthy leg tissues in the principal component space.

The work was carried out under partial financial support of the Russian Fund of Basic Research (grant No.17-00-00186, grant No. 18-42-703012 р_мол_а).

References

1. https://blog.westerndigital.com/machine-learning-pipeline-object-storage/machine-learning-pipeline/

2. *L. Pomerantsev and O. Ye. Rodionova,* Concept and role of extreme objects in PCA/SIMCA//*Journal of Chemometrics.* 2014. **28**(5), P.429-438.

3. *Kistenev, Yu.V., Borisov, A.V., Knyazkova, A.I., Ilyasova, E.E., Sandykova, E.A., Gorbunov, A.K., Spirina, L.V.* Possibilities of cytospectrophotometry of oncological prostate cancer tissue analysis in the TGz spectral range // Proc. of SPIE – XIII International Conference on Atomic and Molecular Pulsed Lasers (AMP17). - AMP100-94, 2018. No.10614-94.