Research Article | Vol. 12, No. 2 / 1 February 2021 / Biomedical Optics Express 1020

Biomedical Optics EXPRESS

Malignant and benign thyroid nodule differentiation through the analysis of blood plasma with terahertz spectroscopy

MARIA R. KONNIKOVA,[1,2] OLGA P. CHERKASOVA,[1,3,*] MAXIM M. NAZAROV,[4] DENIS A. VRAZHNOV,[5] YURI V. KISTENEV,[6,7] SERGEI E. TITOV,[8,9] ELENA V. KOPEIKINA,[10] SERGEI P. SHEVCHENKO,[10] AND ALEXANDER P. SHKURINOV[1,2]

[1] *Institute for Problems of Laser and Information Technologies of the Russian Academy of Sciences, Branch of Federal Scientific Research Center, "Crystallography and Photonics" of the RAS, Shatura 140700, Russia*
[2] *Faculty of Physics, Lomonosov Moscow State University, 119991, Moscow, Russia*
[3] *Institute of Laser Physics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, 630090, Russia*
[4] *National Research Centre Kurchatov Institute, Moscow, 123182, Russia*
[5] *Institute of Strength Physics and Materials Science of the Siberian Branch of the Russian Academy of Sciences, Tomsk, 634055, Russia*
[6] *Tomsk State University, Tomsk, 634050, Russia*
[7] *Siberian State Medical University, Tomsk, 634050, Russia*
[8] *Institute of Molecular and Cellular Biology of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, 630090, Russia*
[9] *Novosibirsk State University, Novosibirsk, 630090, Russia*
[10] *Municipal Clinical Hospital No. 1, Novosibirsk, 630047, Russia*
[*] *o.p.cherkasova@gmail.com*

Abstract: The liquid and lyophilized blood plasma of patients with benign or malignant thyroid nodules and healthy individuals were studied by terahertz (THz) time-domain spectroscopy and machine learning. The blood plasma samples from malignant nodule patients were shown to have higher absorption. The glucose concentration and miRNA-146b level were correlated with the sample's absorption at 1 THz. A two-stage ensemble algorithm was proposed for the THz spectra analysis. The first stage was based on the Support Vector Machine with a linear kernel to separate healthy and thyroid nodule participants. The second stage included additional data preprocessing by Ornstein-Uhlenbeck kernel Principal Component Analysis to separate benign and malignant thyroid nodule participants. Thus, the distinction of malignant and benign thyroid nodule patients through their lyophilized blood plasma analysis by terahertz time-domain spectroscopy and machine learning was demonstrated.

© 2021 Optical Society of America under the terms of the OSA Open Access Publishing Agreement

1. Introduction

Thyroid nodules are the most widespread among the pathologies of the endocrine system. About 5–15% of them are malignant and require surgical intervention, the decision for which is to be taken based on a preoperative diagnosis [1,2]. For this purpose, various imaging techniques are used, including computed [3], magnetic resonance [4,5], and positron emission tomographies [6]. However, ultrasonography [7] and thyroid nodules biopsy under ultrasonography control with fine-needle aspiration cytology (FNAC) [8,9] are the most widely used. The latter requires highly skilled executing personnel and has a risk of errors [10]. Even after adequate sampling, the conclusion turns out to be indefinite in 15-30% of cases because cytological features are

#412715 https://doi.org/10.1364/BOE.412715
Journal © 2021 Received 28 Oct 2020; revised 11 Jan 2021; accepted 15 Jan 2021; published 26 Jan 2021

Biomedical Optics EXPRESS

insufficient to distinguish between benign and malignant thyroid nodules. In these cases, it is necessary to carry out lobectomy to establish the concluding diagnosis [11].

Diagnostics accuracy increases when using immunocytochemistry and thyroid cancer's molecular biomarker evaluation [12–14]. The latter includes the detection of somatic mutations and translocations, determination of expression pattern of protein-encoding genes, and specific micro-RNA (miRNA) [15], as well as oncometabolites [16]. MiRNA is a class of small endogenous non-coding RNA functioning as the negative regulators of gene expression. These miRNAs may contribute to major cell processes in carcinogenesis, namely the growth, differentiation, apoptosis, and adhesion [17]. The miRNA tissue profiles help differentiate benign and malignant thyroid nodules; however, it uses an invasive biopsy procedure [18,19]. Quite the contrary, analysis of miRNA [20] and metabolites in body fluids (blood serum or plasma, urine) possess potential in early cancer detection [21].

Despite a broad list of existing methods, thyroid cancer early detection is not still solved. Optical spectroscopy allows one to measure molecular biomarkers in various samples of biological origin. For example, Raman or Fourier-transform infrared spectroscopy combined with multidimensional analysis distinguished pathologically changed and normal thyroid nodules [22–24].

Terahertz time-domain spectroscopy (THz-TDS) is attractive for designing new noninvasive or minimally invasive diagnostic tools [25–27]. The THz-TDS gives a possibility of direct refractive index measurement, in addition to the absorption coefficient. Herewith, the sample's dielectric function can be restored [28,29], providing a more informative analysis [30–32]. The THz-TDS proved to be a sensitive method of determining structural changes in tumor tissues [33–35], particularly in vivo [36]. The high sensitivity of THz spectroscopy to blood content was established [37–45]. In particular, the diabetes patients' blood absorption in the THz range depends linearly on the glucose level [41,43]. THz spectroscopy distinguished the blood plasma of healthy and diabetic rats [38,44]. Blood plasma THz absorption was decreased for mice with grafted Ehrlich's carcinoma compared to healthy ones [46]. The absorption in the 0.05−1.0 THz range of rat's blood serum was changed during the hepatic cancer development, correlated with blood total protein concentration [47]. It confirms that pathologies, including cancer, cause essential changing blood optical characteristics in the THz range.

This work investigates the abilities of THz-TDS and machine learning in differentiating malignant and benign thyroid nodules by analyzing blood plasma. The liquid and lyophilized blood plasma were studied. The spectral data analysis was conducted by the Kernel Principal Component Analysis (PCA), multidimensional scaling, and Uniform Manifold Approximation and Projection (UMAP) methods. The prognostic models were developed by the linear kernel Support Vector Machine (SVM). The spectral data were also compared with a biochemical composition of blood and the tissue's miRNA profile.

2. Material and methods

2.1. Sample preparation

The study was conducted following the Russian Federation's legislation; each patient signed informed consent; the clinical data were depersonalized. The Ethics Committee of Municipal Clinical Hospital No.1 (Novosibirsk, Russia) approved the study protocol.

Healthy volunteers (n = 6) and patients with thyroid nodules (n = 10), provided by the Municipal Clinical Hospital No. 1 of Novosibirsk, were recruited. Peripheral blood samples were collected in vacutainers with EDTA (Sarstedt AG & Co. KG, Germany). The plasma was separated by centrifugation at 2800 g for 15 min at +4 °C. Blood plasma samples were frozen and stored at -80 °C. Standard biochemical blood tests were conducted at the Municipal Clinical Hospital No.1 (Novosibirsk).

For all patients with thyroid nodules, FNAC was performed, and molecular markers in the biopsy material were determined. Histological examination was carried out for 5 samples. We

Biomedical Optics EXPRESS

ranked the importance of these methods (from more reliable to least) as follows: histological, molecular, and cytological analysis. A minimum set of markers was extracted (levels of *HMGA2* mRNA and miR-375, -221, and -146b combined with the mitochondrial-to-nuclear DNA ratio) and yielded highly accurate discrimination between benign and malignant thyroid nodules [19]. The final clinical diagnoses were made by combining FNAC with histological examination, analysis of molecular markers, and clinical data. Based on this, patients with thyroid nodules were divided into two groups: group 1, with benign thyroid nodules (1b – 5b), and group 2, with malignant thyroid nodules (1m –5m) (Table 1). The detailed patients' description is shown in Table S1.

Table 1. Characteristics of patients[a]

Name	G	Age	The final clinical diagnoses	Glucose, mM	Total protein, g/L	miRNA-146b
1b	F	37	toxic goiter	5.03	61	-2.32
2b	F	63	FTA	5.8	82.4	-32.51
3b	F	34	FTA	5.64	72.3	-1.57
4b	F	57	diffuse non-toxic goiter	6.07	80	-9.24
5b	F	71	Hürthle cell adenoma	5.74	72.2	-3.32
1m	F	66	PTC	4.8	74.9	17.44
2m	F	66	FTC	4.47	69.6	1.40
3m	F	32	PTC	4.52	83.8	3.01
4m	F	64	Follicular variant of PTC	5.2	75	18.89
5m	M	59	Hürthle cell carcinoma	5.2	70.8	-1.67

[a]G- Gender, b – benign, m – malignant, follicular thyroid adenoma (FTA), papillary thyroid carcinoma (PTC), follicular thyroid carcinoma (FTC).

The blood plasma was lyophilized by freeze-drying VaCo 2 (ZirBus, Germany) at -50°C and a pressure of 3 Pa. Pellets with a diameter of 5 mm had been made from lyophilized plasma samples using laboratory press «Specac, GS15011 (Great Britain)» at the pressure of 1 ton. Two variants of the pellets of different thickness («thick» and «thin») from each sample were made. The pellets' weight and thickness were measured by analytic scales «A&D, GR-200 (Japan)» and micrometer «MK-25 0.01(Russia)». The thick pellets' weight and thickness were in the range of 21.1-26.2 mg and 0.75-1.11 mm, respectively, and of the thin pellets were 9.3-13.2 mg and 0.385-0.565 mm, respectively. The density of pellets was in the range of 1.128-1.562 mg/mm^3.

2.2. Experiment setup

The measurements were carried out with a THz-TDS spectrometer based on a femtosecond Spectra Physics Tsunami laser system ($\lambda = 800$ nm, $\tau = 80$ fs, 5 nJ, rep. rate is 80 MHz) [48]. A commercial multi-dipole photoconductive antenna was used as the emitter of THz radiation. The radiation power of the latter was within 70 ÷ 75 mW, and the bias voltage was 15 V. The detector was a 4 mm ZnTe crystal. The repetition period of the THz pulses was 12 ns; the THz pulse energy was 10^{-13} J. The latter is below a tissue damage threshold [49]. A detailed description of the spectrometer was presented in our previous works [44,47,48].

Measurements were carried out at a temperature of (21 ± 1)°C. A temporal averaging over 1024 sample scans with a 25 ps total duration was used. The registered spectral range was from 0.2 to 2 THz for lyophilized samples and from 0.2 to 1.6 THz for liquid plasma samples with a spectral resolution of 40 GHz. One measurement takes about 3 minutes. For increasing reliability, each serum sample was analyzed in triplicates, and then all spectral scans were averaged. The algorithm of the measurements is shown in Fig. 1.

Biomedical Optics EXPRESS

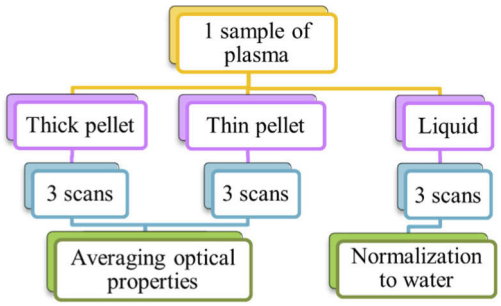

Fig. 1. The algorithm of the blood plasma sample measurements.

The pellets from lyophilized blood plasma were placed on a holder and measured in a transmission mode. Liquid plasma analysis was carried out using two identical cells of 120 µl volume (A145, Bruker Optics) with a spacer of 0.12 mm thickness [50]. One cell was filled with distilled water, the second one - with blood plasma. The plasma was thawed at 4 °C and centrifuged at 1000 g for 10 minutes to remove any precipitate. Transmission of the cell with water was used as a reference signal, making it possible to eliminate water's dominant contribution in blood plasma [47]. The cell transmission measurements had been repeated several times to reduce the influence of laser radiation power temporal drift.

2.3. Methods

2.3.1. Spectral data extraction

The blood plasma transmittance T_w was normalized to air for dry plasma and to water for liquid plasma:

$$T_w(f) = \frac{E_{sample}(f)}{E_{reference}(f)} \tag{1}$$

where $E_{sample}(f)$ is the spectrum of the THz wave transmitted through a dry or liquid sample, $E_{reference}(f)$ is the transmission spectrum of the reference THz wave transmitted through an empty frame for pellets or a cell with water for liquids. The use of water to measure a reference signal makes it possible to consider the reflection losses and re-reflections on the cell walls.

A plasma sample absorption coefficient and the refractive index were calculated using the simplified formulas [47]:

$$\alpha = \frac{-\ln(T_w(f)) + \ln(1 - R^2)}{d}, \tag{2}$$

$$n = n_{average} + arg\frac{E_{reference}(f)}{E_{sample}(f)} \cdot \frac{c}{2 \cdot \pi \cdot f \cdot d}, \tag{3}$$

where $n_{average} = \frac{c \cdot \Delta t}{d} + 1$ is the averaged refractive index, $R = \frac{n_{average}-1}{n_{average}+1}$ is the reflection loss, Δt is the shift of the pulse transmitted through the sample relative to the reference signal, d is the sample thickness, f is the frequency, c is the speed of light.

We do not consider multiple reflections in thin pellets since this effect is negligible due to the strong absorption of the pellet material. The absorption coefficient for the field amplitude and not for the power was used, which is twice as large and generally accepted. Instrumental errors associated with the accuracy of the cell thickness measurement and the baseline drift ultimately lead to a total error of up to 5%. The error deteriorates up to 15% for liquid plasma at the frequency spectrum edges (0.05–0.07 and 1.2–1.6 THz) and lyophilized plasma less than 10% for the total spectral range.

Biomedical Optics EXPRESS

The approximate formulas (2), (3) cause an error in absorption calculation of less than 1% for frequencies above 0.05 THz and for liquid plasma layer of at least 0.12 mm and lyophilized plasma pellets thickness from 0.385 to 1.11 mm. The variation in pellet thickness causes a total error of optical characteristics below 15%, with the method described above.

2.3.2. Spectral data analysis methods

The following dimensionality reduction methods were used: Kernel PCA, multidimensional scaling, and UMAP methods [51–53]. We also applied a Composite Multiscale Entropy (CMSE) method for the THz data cluster analysis [54]. These methods were chosen due to the difference in respective basic ideas and the application fields' versatility. Kernel PCA with linear, radial based function kernel and original kernel, derived from the Green's function for the Ornstein-Uhlenbeck (OU) process [55] was used. The latter has the following form:

$$K(x, y) = \exp(-\gamma \cdot \|x - y \cdot \exp(-\lambda)\|), \tag{4}$$

where parameters γ, λ define the shape of the kernel. We used the following values: $\gamma = 0.001$, $\lambda = 0.01$, and the pairwise Euclidean distance as a norm $\| \ldots \|$.

The UMAP method was applied with the number of neighbors 4, minimal distance 0.7. Multidimensional scaling was used with the default set of parameters. CMSE has the embedded dimension 2, threshold value 0.15, and scale numbers were assigned to 10 because we have lower spectral resolution than in the reference work. All these computations were conducted in Python 3.8.2 and Scikit-learn package 0.23.2.

The amount of the data is not large, and the classification could suffer overfitting. Here, we used Supervised Learning methods to estimate the linear separability of the data, for example, by using linear kernel SVM [56]. To validate estimation of the data separability, we used a standard approach for Supervised Learning: initial data were randomly divided into 3 splits with the same distribution of the data among classes as in the original dataset. Linear SVM was trained on each split with class weight balancing. Obtained quality metrics such as sensitivity, specificity, accuracy, and precision were averaged. Also, Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) analyses were performed for each data split and the averaged ones.

A two-stage ensemble scheme was proposed for the data analysis (see Fig. 2). The first stage aims to separate healthy and thyroid nodules, enabling to estimate of the potential of sensitivity. The second stage was aimed to separate benign and malignant thyroid nodules, allowing evaluating the potential of specificity. By applying them consequently, the possibility of constructing a good prediction model arises.

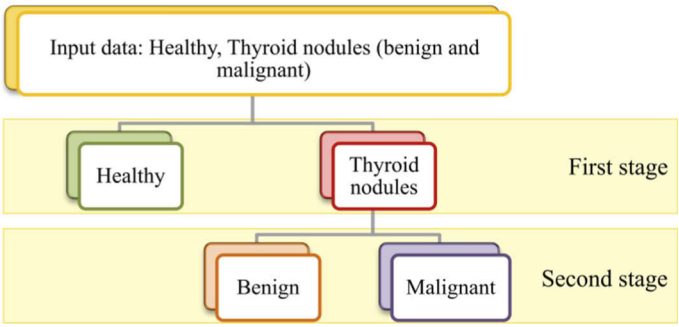

Fig. 2. The two-stage scheme of the data separation.

Biomedical Optics EXPRESS

2.3.3. Statistical data analysis

The OriginLab Corp. statistical software package was used to process the data. Mann–Whitney test was used for paired comparisons. The Pearson correlation coefficient was used to estimate linear relations between biochemical parameters and THz optical characteristics. The null hypothesis feasibility was taken at 5% of significance.

3. Results and discussions

3.1. Liquid blood plasma

Liquid blood plasma samples were analyzed without additional processing. Their absorption coefficient and refractive index spectra are shown in Fig. 3.

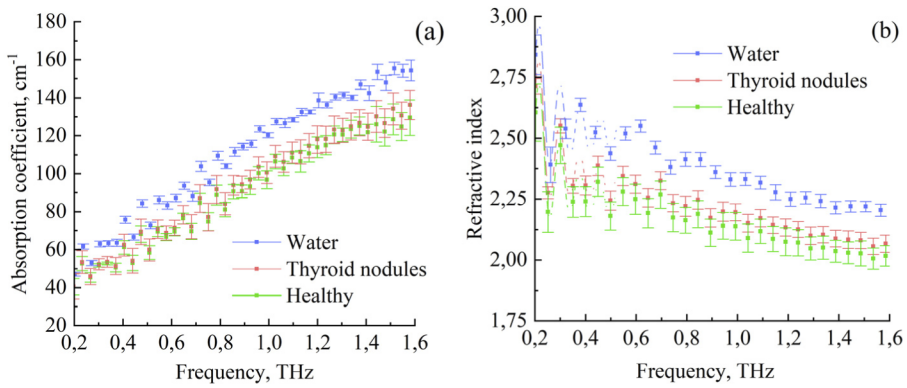

Fig. 3. Average absorption coefficient a) and refractive index b) spectra of water (1, blue), healthy individuals (2, green), and patients with thyroid nodules (3, red) with error bars. Averaging was carried out over 30 measurements for blood plasma of patients with thyroid nodules, over 15 measurements for both blood plasma healthy individuals and water.

The absorption coefficient and refractive index values of blood plasma from patients with thyroid nodules are slightly higher than those from healthy individuals, but there are no statistically significant differences.

The shapes of blood plasma spectra and the water spectrum are the same; however, the amplitude of blood plasma absorption and refractive index spectra is much less than that of water. We observed a similar phenomenon in human and animal blood plasma [38,44,47,57]. The coincidence of blood plasma and water spectra shape is explained by the fact that water makes up more than 90% of blood plasma. At the same time, such blood plasma components as glucose, proteins, and various salts have a significant effect on the structure of water (it mainly bind water molecules to a state with a lower THz absorption) and change its THz response [29,48,50,58–60]. Despite the high absorption of THz radiation by water, the THz TDS is sensitive to small variations of the liquid blood plasma component concentrations in the low-frequency range [40,41,43,47].

It is convenient using spectra subtraction to visualize the differences between the optical characteristics of the blood plasma:

$$\alpha_{dif} = \alpha_{plasma} - \alpha_{water} \tag{5}$$

$$n_{dif} = n_{plasma} - n_{water} \tag{6}$$

where α_{dif}, n_{dif} - difference of absorption and refraction values spectra of liquid plasma, α_{plasma}, n_{plasma}– plasma spectrum values for the case when the reference signal $E_{reference}(f)$ is air, α_{water}, n_{water}– spectral characteristics of water.

Biomedical Optics EXPRESS

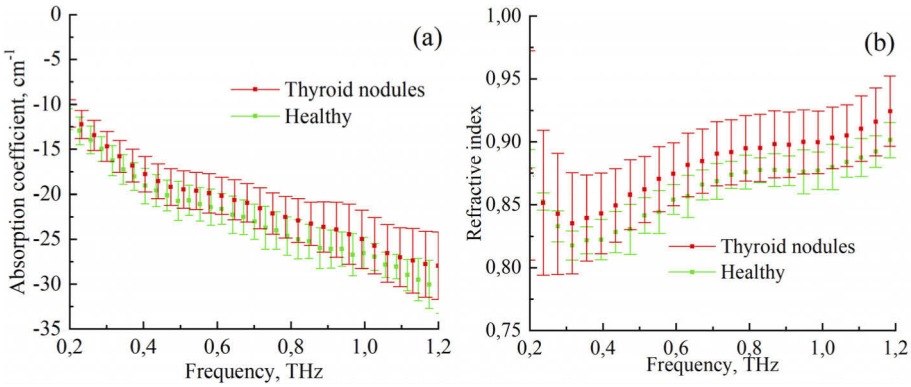

Fig. 4. Average difference spectra of absorption coefficient a) and refractive index b) of blood plasma healthy individuals (green squares) and patients with thyroid nodules (red squares) with error bars. Averaging was carried out over 30 measurements for blood plasma of patients with thyroid nodules and over 15 measurements for healthy individuals' blood plasma (each measurement with an independent reference signal obtained by passing through a cell with water).

The difference of absorption coefficient and refractive index values spectra of blood plasma of healthy individuals (green squares) and patients with thyroid nodules (red squares) are shown in Fig. 4. Spectral differences of liquids are small compared to the total error.

Figure 5 shows the difference in absorption coefficient and refractive index values spectra for blood plasma from group 1 (purple squares) and 2 (orange squares).

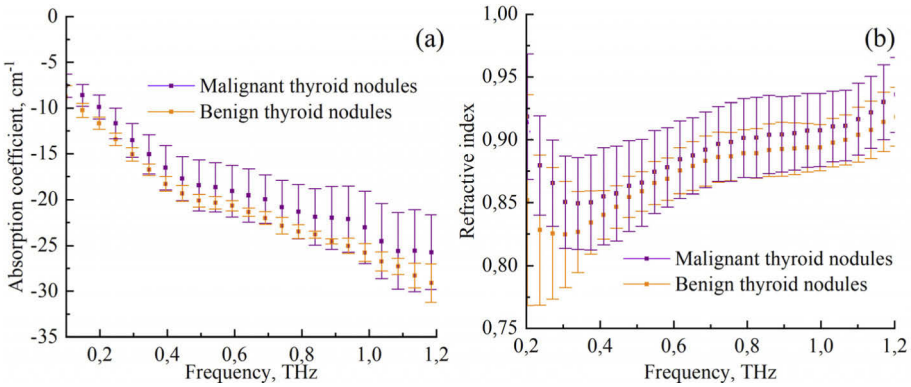

Fig. 5. Averaged difference spectra of absorption coefficient and refractive index for the liquid plasma of two groups: malignant thyroid nodules (purple squares) and benign thyroid nodules (orange squares) normalized to water. Averaging was carried out over 15 measurements (each measurement with an independent reference signal obtained by passing through a cell with water).

Group 1 is characterized by significantly higher blood glucose levels than group 2 (5.66 ± 0.17 mM versus 4.84 ± 0.16 mM, $p < 0.05$) and 1.2 times higher protein levels. The absorption coefficient and refractive index, normalized to water, are lower for group 1 than group 2. The differences are most significant at 0.2 THz. Earlier, the most reliable transmission spectral range of 0.07 to 0.5 THz for blood serum was found for rat experimental liver cancer [47].

Biomedical Optics EXPRESS

3.2. Lyophilized blood plasma

Since the water contribution to the THz blood plasma spectrum dominates over the other components, lyophilization was used to increase THz spectroscopy's informativity. Figure 6 (a, b) shows the THz absorption coefficient of lyophilized blood plasma pellets.

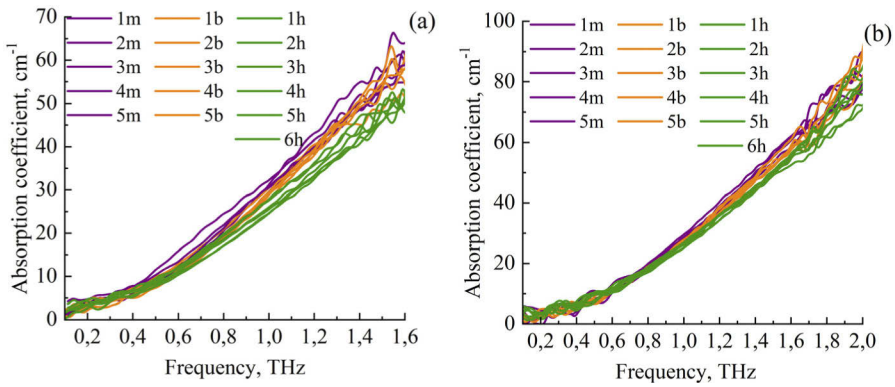

Fig. 6. The absorption coefficient of thick (a) and thin (b) pellets of lyophilized human blood plasma: green line - a group of healthy individuals, 1h-6h; orange line - group 1, 1b-5b; purple line - group 2, 1m-5m.

Using pellets of various thicknesses improves the optical parameters evaluation accuracy and expands the data's spectral range. Thus, when thickness increases, the transmitted pulse is more attenuated, and the error in thickness determining affects less. But at high frequencies, the spectral intensity is attenuated below the noise level. The advantage of thin pellets is more transparency at the high-frequencies of the THz spectral range and the additional ability to determine the pellet's thickness and the average refraction through internal re-reflections.

Figure 6 shows that the absorption coefficient is the same for both types of pellets independently of the sample type. Moreover, "thick" pellets expand the reliable spectral range to the region from 0.04 to 0.1 THz, and "thin" ones to the region from 1.5 to 2 THz. Note that no difference was found in the refractive index between thick and thin pellets. It allows us to average each sample over two pellet types. The resulting spectra are shown in Fig. 7.

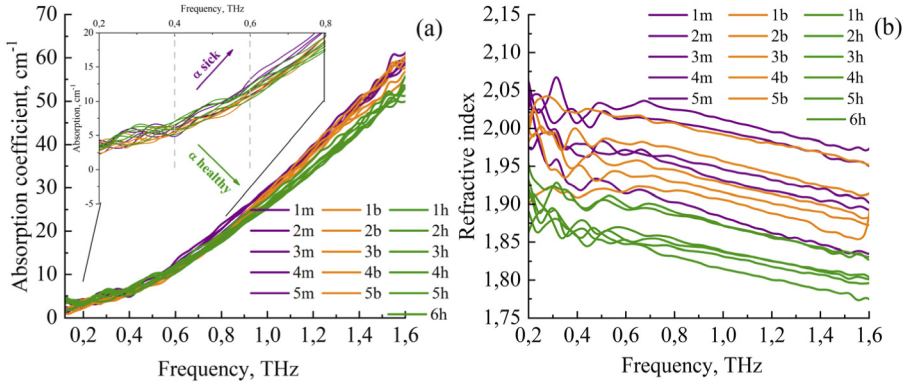

Fig. 7. The absorption coefficient (a) and refractive index (b) with each sample averaging over thick and thin pellets: green line - a group of healthy individuals, 1h-6h; orange line - group 1, 1b-5b; purple line - group 2, 1m-5m.

Biomedical Optics EXPRESS

The absorption is higher in blood plasma from healthy individuals up to a frequency of 0.4 THz. In the range of 0.4-0.6 THz, there is a change in the dynamics of absorption. Samples from individuals with thyroid nodules exhibit stronger absorption, which is higher from 0.6 THz and more. The same trend was observed in the absorption spectra of liquid samples (see Fig. 3 and Fig. 4).

The pellet spectra do not have any prominent features. Glucose has phonon lines at 1.45, 2.12, 2.5 THz [50]. However, there is evidence that, upon lyophilization of liquid samples containing proteins and sugars, bonds are formed between them, and resonances disappear [61]. The refractive index spectrum does not change, but the absolute values correlate with the disease's presence (see Fig. 7(b)).

The spectra were averaged over groups 1 and 2 to identify spectral differences depending on benign or malignant thyroid nodules (see Fig. 8). The samples from malignant nodule patients had higher absorption compared with that of benign nodule patients. The same trend was observed when studying liquid samples (see Fig. 5). However, spectral differences in liquids are small compared to the total error.

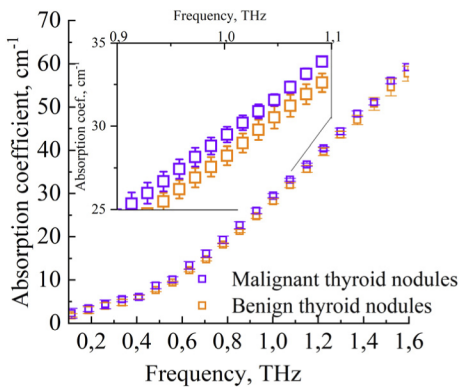

Fig. 8. The absorption coefficient of lyophilized blood plasma of two groups with malignant thyroid nodules (purple squares) and benign thyroid nodules (orange squares).

Thus, both liquid and lyophilized blood plasma samples with malignant thyroid nodules have higher absorption than samples with benign thyroid nodules. For dry pellets, the differences are significant. The shapes of the spectra showing the increase in absorption with frequency are the same for all samples.

3.3. Correlation of spectral data and blood plasma composition

The absolute value of the Pearson correlation coefficient |R| is a measure of the strength of the linear relationship between the values of a data pair. A value of |R| equal to 1 means that there is a perfect linear relation; |R|>0.5 represents a moderate to a strong relationship, 0.3< |R| <0.5 represents a weak to reasonable relationship, and |R|<0.3 represents a weak relationship [41].

The Pearson correlation coefficient was applied to study the relationship between the blood plasma THz absorption spectra and glucose or protein concentrations. We used the two-tail test, which allots half of the α in one direction (r>0 side) and half of the α in the other direction (r<0 side). We used p=0.05 as a threshold of significant correlation. If p <0.05 was true, we took an |R| to indicate their correlation strength. If p> 0.05, no correlation was found. The strength of the linear correlation between glucose concentration and the absorption at 1 THz is shown in Fig. 9.

For glucose, the calculated Pearson's correlation coefficient is R=-0.84, p=0.0047 in liquid blood plasma (Fig. 9(a)) and R=-0.79, p=0.01 in lyophilized blood plasma (Fig. 9(b)), which

Biomedical Optics EXPRESS

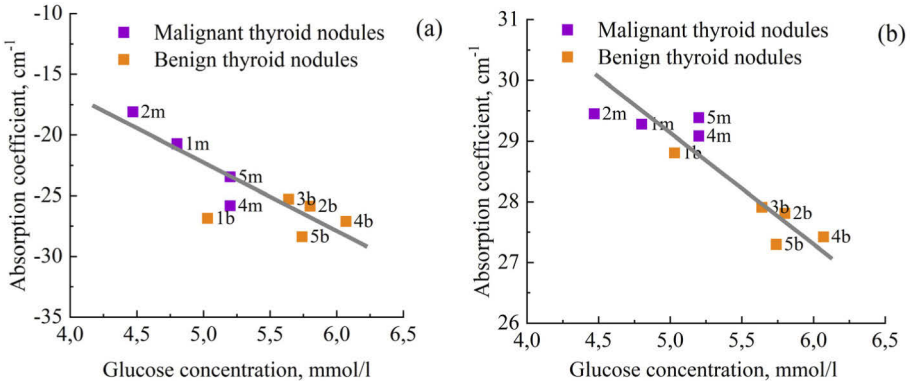

Fig. 9. Dependence of the absorption coefficient of liquid a) and lyophilized b) blood plasma of two groups with malignant thyroid nodules (purple squares) and benign thyroid nodules (orange squares) at a frequency of 1 THz on the concentration of glucose in the samples.

means the correlation is high and negative. Correlation dependencies for total protein and miRNA are presented in Table 2.

Table 2. The correlations between the absorption coefficients at 1 THz and the examined variables. R is the Pearson correlation coefficient; p is the two-tail p-value

Examined Variables	Liquid blood plasma	Lyophilized blood plasma
Glucose	**R=-0.84, p=0.0047**	**R=-0.79, p=0.01**
Proteins	R=-0.1, p=0.79	R=-0.34, p=0.368
miRNA-146b	R=0.41, p=0.23	**R=0.63, p=0.049**

Notably, the THz response change depending on glucose concentration is more significant than expected from an effective medium parameters model at such a low volume fraction of glucose. It can be presumably caused by rebuilding protein structure due to glucose presence during lyophilization [61]. Probably, for this reason, no correlation between protein concentration and the absorption coefficient was found. A strong correlation was observed between the absorption coefficient and glucose concentration both for liquid and lyophilized samples. Glucose and protein make the most significant contribution to the blood plasma absorption in the THz range [41,43,44,47].

The dependence of the malignancy degree of thyroid nodules on the several miRNAs content in tissue and blood plasma has been established [18–20]. Overexpression of miRNA-146b in papillary and follicular thyroid carcinoma compared to levels in normal thyroid tissues was reported [19]. This upregulation was positively correlated with tumor aggressiveness. Essential miR-146b content differences between groups ($p<0.05$) were observed. Group 2 had higher miRNA-146b values compared to Group 1 (see Table 1). The level of miRNA 146 had a positive correlation with the lyophilized plasma absorption at 1 THz ($R = 0.63$, two-tail p-value = 0.049) (see Table 2).

3.4. Machine learning application

For the first stage of the ensemble scheme (Fig. 2), kernel PCA, multidimensional scaling, and UMAP showed the best results (see Fig. 10). The exception was CMSE, which produced features with a dimension of scale number (10 in our case), and additional dimensionality reduction was required.

Biomedical Optics EXPRESS

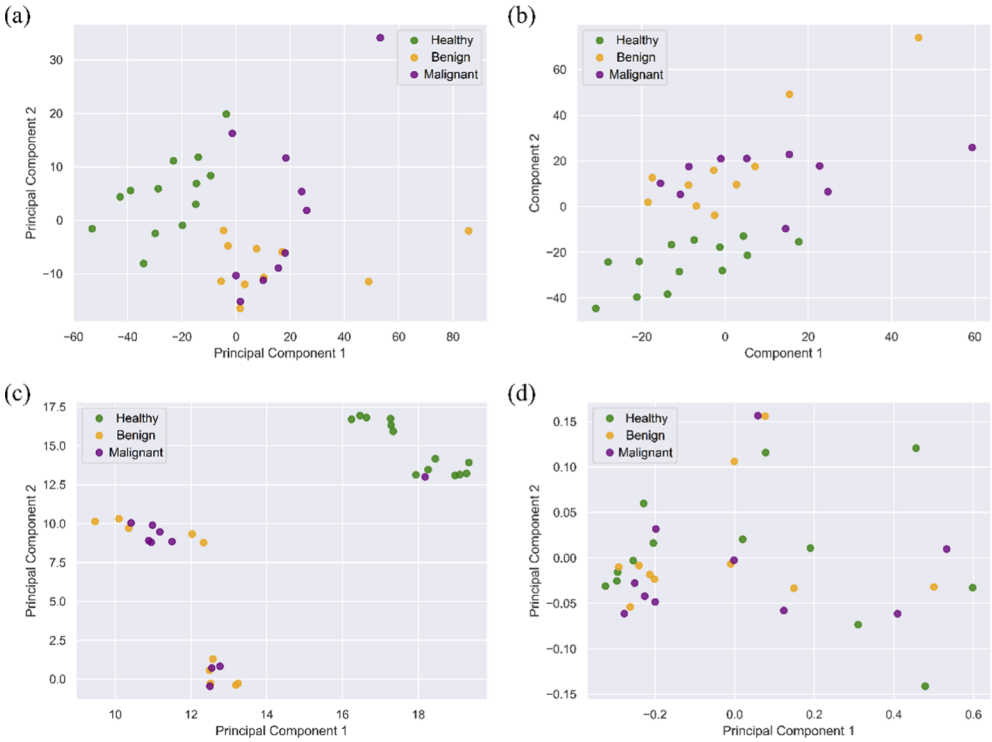

Fig. 10. Visualization of linear kernel PCA (a), multidimensional scaling (b), UMAP (c), and CMSE with linear PCA (d).

Linear separability of the healthy versus benign and malignant thyroid nodules classes was verified by linear SVM, which was applied to the data, preliminary transformed by the linear kernel PCA (see Fig. 11 (a)). The averaged quality metrics: sensitivity was equal to 0.92 ± 0.04; specificity was equal to 0.85 ± 0.05 accuracy was equal to 0.88 ± 0.04; precision was equal to 0.80 ± 0.03. The mean ROC curve and AUC analysis for the linear SVMs is shown in Fig. 11 (b). The results demonstrate the potential of developed models for the separation of healthy and thyroid nodules groups.

Also, linear SVM allows obtaining the most informative features, based on the proximity of the feature vector to the separating hyperplane. An averaged relative features' importance was estimated for the three train/test splits and linear SVMs. The most significant absorption frequency for group separating was 1.11095 THz; relative importance was 0.008. The latter was low due to the normalization of a large number of the feature vector components. Figure 12 illustrates the downfall of relative features importance and most informative THz absorption frequencies.

Next, we developed a method to test the separability of the benign and malignant groups. The thick and thin blood plasma pellets were found to give different information for malignant and benign thyroid nodules distinguishing. For example, data after multidimensional scaling and linear kernel PCA are not linearly separable; UMAP performs well on both data. We suppose that it can be related to the high complexity of the UMAP model and the low data volume.

The OU Kernel PCA provided explicit separability of groups of benign and malignant thyroid nodules when lyophilized blood plasma pellets with a thickness of 0.385-0.565 mm had been used (see Fig. 13(a)). We believe that it relates to their "transparency" in the THz high-frequency range, as was mentioned earlier. That is why the thickness of the pellets should be taken into

Biomedical Optics EXPRESS

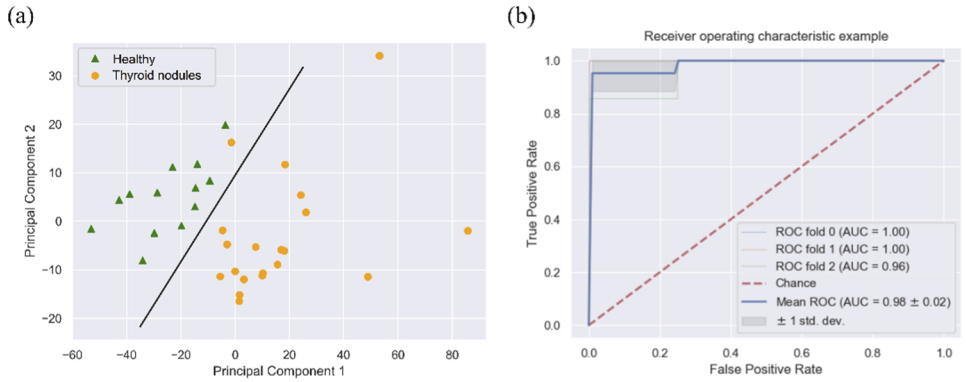

Fig. 11. Linear separability of the healthy, benign, and malignant thyroid nodus data by the SVM with a linear kernel. Regularization parameter C=1 (a); Mean ROC curve for linear SVMs for the 3 splits (b).

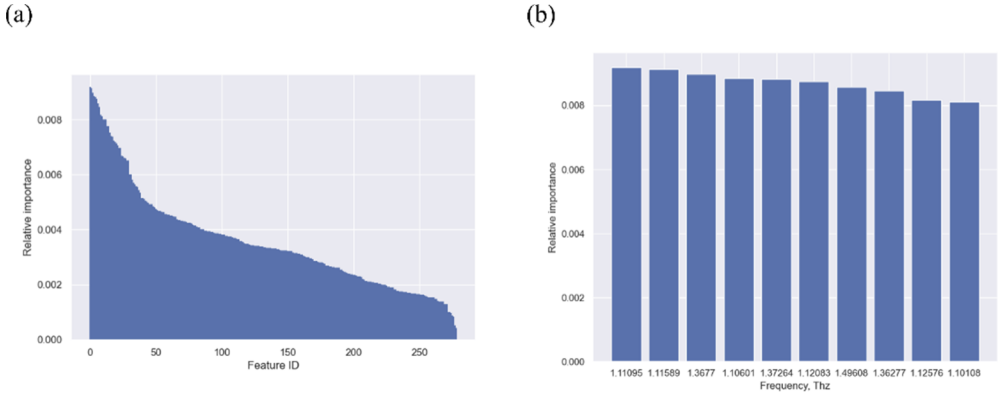

Fig. 12. Feature IDs sorted by relative features importance (a). Feature ID corresponds to a specific THz absorption frequency. The first most informative THz absorption frequencies (b).

account: thin ones are more suitable for the groups under study distinction. This result is consistent with the established spectral data and blood plasma composition correlations.

The benign and malignant groups' size is small. Therefore, the creation of a predictive model cannot be reliable. Still, the confirmed separability of these groups is very promising.

Biomedical Optics EXPRESS

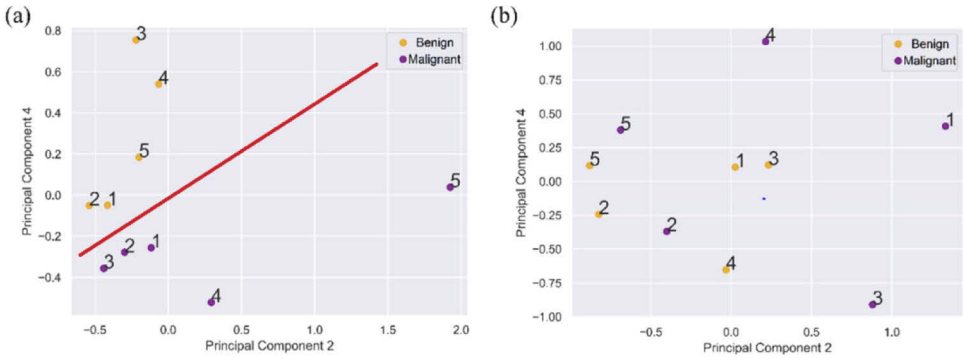

Fig. 13. Thin (a) and thick (b) benign and malignant thyroid nodules plasma pellets transformed by the OU kernel PCA. The numbers show the sample ID.

4. Conclusion

We investigated blood plasma samples of healthy individuals and patients with benign and malignant thyroid nodules by THz-TDS. A strong correlation between the glucose concentration and the absorption for liquid and lyophilized blood plasma samples was established. We also demonstrated the correlation between miRNA-146b level and the absorption at 1 THz for the lyophilized blood plasma samples.

In total, THz spectra differences of liquid blood plasma samples were comparable with the experimental error value. The study of lyophilized blood plasma, pressed into pellets, showed a reliable separation of healthy individuals and patients with thyroid nodules and separation of patients with benign and malignant thyroid nodules through THz absorption coefficient and refractive index values. This separation was confirmed by machine learning as follows. The raw data were projected to the lower-dimensional space by kernel PCA, multidimensional scaling, and UMAP methods. All these methods give similar results except CMSE. Linear separability between healthy and thyroid nodules classes was demonstrated by the linear SVM with 92% sensitivity, 85% specificity, and 88% accuracy. The informative feature histogram was obtained to show the most informative frequencies for the classification.

In our case, an individual algorithm was not sufficient to distinguish patients with benign and malignant thyroid nodules using THz spectra of lyophilized blood plasma samples. A two-stage ensemble scheme was proposed. On a first step corresponded to a sensitivity estimation, the thyroid nodule group is separated from the healthy one. On a second step corresponded to a specificity estimation, benign and malignant groups are separated.

The current number of blood plasma samples from patients with benign and malignant thyroid nodules is not enough to create a robust predictive model. Still, the confirmed separability of these groups is very promising. Thus, we proved that THz TDS could be sensitive to blood composition changes, depending on the thyroid nodule malignancy degree. Collecting enough samples can give an opportunity for differential diagnosis of thyroid nodules through blood plasma analysis by THz spectroscopy and machine learning.

Funding. Ministry of Science and Higher Education of the Russian Federation (0307-2019-0007, 075-15-2019-1950, 2020-220-08-2389); Russian Foundation for Basic Research (17-00-00275 (17-00-00186), 17-00-00275 (17-00-00270), 19-52-55004); Government Council on Grants, Russian Federation (III.23.2.10).

Acknowledgments. This work was supported by the Ministry of Science and Higher Education within the State assignment FSRC "Crystallography and Photonics" RAS, Interdisciplinary Scientific and Educational School of Moscow University «Photonic and Quantum Technologies. Digital Medicine». This work was supported by the Government of the Russian Federation (proposal No. 2020-220-08-2389 to support scientific research projects implemented under the supervision of leading scientists at Russian institutions, Russian institutions of higher education) in the part of

Biomedical Optics EXPRESS

machine learning implementation. This work was supported by the Government of the Russian Federation proposal No. 075-15-2019-1950 to support scientific research projects implemented under the supervision of leading scientists at Russian institutions in part of sample preparation. The work was performed under the government statement of work for ISPMS Project No. III.23.2.10.

Disclosures. The authors declare that there are no conflicts of interest.

Supplemental document. See Supplement 1 for supporting content.

References

1. D.-L. Kim, K.-H. Song, and S. K. Kim, "High prevalence of carcinoma in ultrasonography-guided fine needle aspiration cytology of thyroid nodules," Endocr. J. **55**(1), 135–142 (2008).
2. J. P. Brito, A. J. Yarur, L. J. Prokop, B. McIver, M. H. Murad, and V. M. Montori, "Prevalence of thyroid cancer in multinodular goiter versus single nodule: a systematic review and meta-analysis," Thyroid **23**(4), 449–455 (2013).
3. M. Bin Saeedan, I. M. Aljohani, A. O. Khushaim, S. O. Bukhari, and S. T. Elnaas, "Thyroid computed tomography imaging: pictorial review of variable pathologies," Insights Imaging **7**(4), 601–617 (2016).
4. T. Kang, D. W. Kim, Y. J. Lee, Y. J. Cho, S. J. Jung, H. K. Park, and H. J. Baek, "Magnetic resonance imaging features of normal thyroid parenchyma and incidental diffuse thyroid disease: a single-center study," Front. Endocrinol. **9**, 746 (2018)..
5. L. Aghaghazvini, P. Pirouzi, H. Sharifian, N. Yazdani, S. Kooraki, A. Ghadiri, and M. Assadi, "3 T magnetic resonance spectroscopy as a powerful diagnostic modality for assessment of thyroid nodules," Arch. Endocrinol. Metab. **62**(5), 501–505 (2018).
6. G. Treglia, A. S. Kroiss, A. Piccardo, F. Lococo, P. Santhanam, and A. Imperiale, "Role of positron emission tomography in thyroid and neuroendocrine tumors," Minerva Endocrinol. **43**(3), 341–355 (2018)..
7. V. Chaudhary and S. Bano, "Thyroid ultrasound," Indian J. Endocr. Metab. **17**(2), 219 (2013).
8. L. Xu, J. Gao, Q. Wang, J. Yin, P. Yu, B. Bai, R. Pei, D. Chen, G. Yang, S. Wang, and M. Wan, "Computer-aided diagnosis systems in diagnosing malignant thyroid nodules on ultrasonography: a systematic review and meta-analysis," Eur. Thyroid. J. **9**(4), 186–193 (2020).
9. C. K. H. Wong, X. Liu, and B. H. H. Lang, "" Cost effectiveness of the needle aspiration cytology (FNAC) and watchful observation for incidental thyroid nodules"," J. Endocrinol. Invest. **43**(11), 1645–1654 (2020).
10. C. Stevens, J. K. Lee, M. Sadatsafavi, and G. K. Blair, "Pediatric thyroid fine-needle aspiration cytology: a meta-analysis," J. Pediatr Surg. **44**(11), 2184–2191 (2009).
11. A. Matrone, M.C. Campopiano, A. Nervo, G. Sapuppo, M. Tavarelli, and S. De Leo, "Differentiated thyroid cancer, from active surveillance to advanced therapy: toward a personalized medicine," Front. Endocrinol. **10**, 884 (2020).
12. D. Shibru, K. W. Chung, and E. Kebebew, "Recent developments in the clinical application of thyroid cancer biomarkers," Curr. Opin. Oncol. **20**(1), 13–18 (2008).
13. N. D. Banks, J. Kowalski, H.-L. Tsai, H. Somervell, R. Tufano, A. P. B. Dackiw, and M. A. Zeiger, "A diagnostic predictor model for indeterminate or suspicious thyroid FNA samples," Thyroid **18**(9), 933–941 (2008).
14. C. V. Villabona, V. Mohan, K. M. Arce, J. Diacovo, A. Aggarwal, J. Betancourt, H. Amer, T. Jose, P. DeSantis, and J. Cabral, "Utility of ultrasound versus gene expression classifier in thyroid nodules with atypia of undetermined significance," Endocrine practice **22**(10), 1199–1203 (2016).
15. S. Sciacchitano, L. Lavra, A. Ulivieri, F. Magi, G. Paolo De Francesco, C. Bellotti, L. B. Salehi, M. Trovato, C. Drago, and A. Bartolazzi, "Comparative analysis of diagnostic performance, feasibility and cost of different test-methods for thyroid nodules with indeterminate cytology," Oncotarget **8**(30), 49421–49442 (2017).
16. F. Khatami, M. Payab, M. Sarvari, K. Gilany, B. Larijani, B. Arjmand, and S. M. Tavangar, "Oncometabolites as biomarkers in thyroid cancer: a systematic review," Cancer Manage. Res. **11**, 1829–1841 (2019).
17. J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz, and T. R. Golub, "MicroRNA expression profiles classify human cancers," Nature **435**(7043), 834–838 (2005).
18. P. Pallante, R. Visone, M. Ferracin, A. Ferraro, M. T. Berlingieri, G. Troncone, G. Chiappetta, C. G. Liu, M. Santoro, M. Negrini, C. M. Croce, and A. Fusco, "MicroRNA deregulation in human thyroid papillary carcinomas," Endocr Relat Cancer **13**(2), 497–508 (2006).
19. S. E. Titov, M. K. Ivanov, P. S. Demenkov, G. A. Katanyan, E. S. Kozorezova, A. V. Malek, Y. A. Veryaskina, and I. F. Zhimulev, "Combined quantitation of HMGA2 mRNA, microRNAs, and mitochondrial-DNA content enables the identification and typing of thyroid tumors in fine-needle aspiration smears," BMC Cancer **19**(1), 1010 (2019).
20. S. Yu, Y. Liu, J. Wang, Z. Guo, Q. Zhang, F. Yu, Y. Zhang, K. Huang, Y. Li, E. Song, X. Zheng, and H. Xiao, "Combined quantitation of HMGA2 mRNA, microRNAs, and mitochondrial-DNA content enables the identification and typing of thyroid tumors in fine-needle aspiration smears," BMC Cancer **19**(1), 1010 (2019).
21. W. Wojtowicz, A. Zabek, S. Deja, T. Dawiskiba, D. Pawelka, M. Glod, and P. Mlynarz, "Serum and urine 1H NMR-based metabolomics in the diagnosis of selected thyroid diseases," Sci. Rep. **7**(1), 9108 (2017).
22. J. Depciuch, A. Stanek-Widera, D. Skrzypiec, D. Lange, M. Biskup-Fruzyńska, K. Kiper, J. Stanek-Tarkowska, M. Kula, and J. Cebulski, "Spectroscopic identification of benign (follicular adenoma) and cancerous lesions (follicular thyroid carcinoma) in thyroid tissues," J. Pharm. Biomed. **170**, 321–326 (2019).

Biomedical Optics EXPRESS

23. M. Sbroscia, M. Di Gioacchino, P. Ascenzi, P. Crucitti, A. di Masi, I. Giovannoni, F. Longo, D. Mariotti, A. M. Naciu, A. Palermo, C. Tafon, M. Verri, A. Sodo, A. Crescenzi, and M. A. Ricci, "Thyroid cancer diagnosis by Raman spectroscopy," Sci. Rep. **10**(1), 13342 (2020)..
24. J. N. Taylor, K. Mochizuki, K. Hashimoto, Y. Kumamoto, Y. Harada, K. Fujita, and T. Komatsuzaki, "High-resolution Raman microscopic detection of follicular thyroid cancer cells with unsupervised machine learning," J. Phys. Chem. B **123**, 4358 (2019).
25. J. H. Son, S. J. Oh, and H. Cheon, "Potential clinical applications of terahertz radiation," J. Appl. Phys. **125**(19), 190901 (2019).
26. A. Gong, Y. Qiu, X. Chen, Z. Zhao, L. Xia, and Y. Shao, "Biomedical applications of terahertz technology," Appl. Spectrosc. Rev. **55**(5), 418–438 (2020).
27. Y. Peng, C. Shi, X. Wu, Y. Zhu, and S. Zhuang, "Terahertz imaging and spectroscopy in cancer diagnostics: a technical review," BME Frontiers **2020**, 1–11 (2020).
28. K. I. Zaytsev, I. N. Dolganova, N. V. Chernomyrdin, G. M. Katyba, A. A. Gavdush, O. P. Cherkasova, G. A. Komandin, M. A. Shchedrina, A. N. Khodan, D. S. Ponomarev, I. V. Reshetov, V. E. Karasik, M. Skorobogatiy, V. N. Kurlov, and V. V. Tuchin, "The progress and perspectives of terahertz technology for diagnosis of neoplasms: a review," J. Opt. **22**, 013001 (2020).
29. O. Smolyanskaya, N. Chernomyrdin, A. Konovko, K. Zaytsev, I. Ozheredov, O. Cherkasova, M. Nazarov, J.-P. Guillet, S. Kozlov, Y. Kistenev, J.-L. Coutaz, P. Mounaix, V. Vaks, J.-H. Son, H. Cheon, V. Wallace, Y. Feldman, I. Popov, A. Yaroslavsky, A. Shkurinov, and V. Tuchin, "Terahertz biophotonics as a tool for studies of dielectric and spectral properties of biological tissues and liquids," Prog. Quantum Electron. **62**, 1–77 (2018).
30. Y. Peng, C. Shi, Y. Zhu, M. Gu, and S. Zhuang, "Terahertz spectroscopy in biomedical field: a review on signal-to-noise ratio improvement," PhotoniX **1**(1), 12 (2020).
31. O. P. Cherkasova, M. M. Nazarov, and A. P. Shkurinov, "Noninvasive blood glucose monitoring in the terahertz frequency range," Opt. Quantum Electron. **48**(3), 217–8919 (2016).
32. A. A. Gavdush, N. V. Chernomyrdin, G. A. Komandin, I. N. Dolganova, P. V. Nikitin, G. R. Musina, G. M. Katyba, A. S. Kucheryavenko, I. V. Reshetov, A. A. Potapov, V. V. Tuchin, and K. I. Zaytsev, "Terahertz dielectric spectroscopy of human brain gliomas and intact tissues ex vivo: double-Debye and double-overdamped-oscillator models of dielectric response," Biomed. Opt. Express **12**(1), 69–83 (2021).
33. E. Pickwell and V. P. Wallace, "Biomedical applications of terahertz technology," J. Phys. D: Appl. Phys. **39**(17), R301–R310 (2006).
34. M. Danciu, T. Alexa-Stratulat, C. Stefanescu, G. Dodi, B. I. Tamba, C. T. Mihai, G. D. Stanciu, A. Luca, I. A. Spiridon, L. B. Ungureanu, V. Ianole, I. Ciortescu, C. Mihai, G. Stefanescu, I. Chirilă, R. Ciobanu, and V. L. Drug, "Terahertz spectroscopy and imaging: a cutting-edge method for diagnosing digestive cancers," Materials **12**(9), 1519 (2019).
35. A.A. Gavdush, N.V. Chernomyrdin, K.M. Malakhov, S.-I.T. Beshplav, I.N. Dolganova, A.V. Kosyrkova, P.V. Nikitin, G.R. Musina, G.M. Katyba, I.V. Reshetov, O.P. Cherkasova, G.A. Komandin, V.E. Karasik, A.A. Potapov, V.V. Tuchin, and K.I. Zaytsev, "Terahertz spectroscopy of gelatin-embedded human brain gliomas of different grades: a road toward intraoperative THz diagnosis," J. Biomed Opt. **24**(2), 027001 (2019).
36. L. Yu, L. Hao, T. Meiqiong, H. Jiaoqi, L. Wei, D. Jinying, and Z. Yang, "The medical application of terahertz technology in noninvasive detection of cells and tissues: opportunities and challenges," RSC Adv. **9**(17), 9354–9363 (2019).
37. S. I. Gusev, P. S. Demchenko, E. A. Litvinov, O. P. Cherkasova, I. V. Meglinski, and M. K. Khodzitsky, "Study of glucose concentration influence on blood optical properties in THz frequency range," Nanosystems: Phys. Chem. Math. **9**(3), 389–400 (2018).
38. O. P. Cherkasova, M. M. Nazarov, I. N. Smirnova, A. A. Angeluts, and A. P. Shkurinov, "Application of time-domain THz spectroscopy for studying blood plasma of rats with experimental diabetes," Phys. Wave Phen. **22**(3), 185–188 (2014).
39. C. B. Reid, G. Reese, A. P. Gibson, and V. P. Wallace, "Terahertz time-domain spectroscopy of human blood," IEEE J. Biomed. Health Inform. **17**(4), 774–778 (2013).
40. C.K. Sun, H.Y. Chen, T.F. Tseng, B. You, M.-L. Wei, J.-Y. Lu, Y.-L. Chang, W.-L. Tseng, and T.-D. Wang, "High Sensitivity of T-Ray for Thrombus Sensing," Sci. Rep. **8**(1), 3948 (2018).
41. T.-F. Tseng, B. You, H.-C. Gao, T.-D. Wang, and C.-K. Sun, "Pilot clinical study to investigate the human whole blood spectrum characteristics in the sub-THz region," Opt. Express **23**(7), 9440–9451 (2015).
42. S. I. Gusev, P. S. Demchenko, O. P. Cherkasova, V. I. Fedorov, and M. K. Khodzitsky, "Influence of glucose concentration on blood optical properties in THz frequency range," Chinese Opt. **11**(2), 182–189 (2018).
43. H. Chen, X. Chen, S. Ma, X. Wu, W. Yang, W. Zhang, and X. Li, "Quantify glucose level in freshly diabetic's blood by terahertz time-domain spectroscopy," J. Infrared, Millimeter, Terahertz Waves **39**(4), 399–408 (2018).
44. O. P. Cherkasova, M. M. Nazarov, A. A. Angeluts, and A. P. Shkurinov, "Analysis of blood plasma at terahertz frequencies," Opt. Spectrosc. **120**(1), 50–57 (2016).
45. K. Jeong, Y.-M. Huh, S.-H. Kim, Y. Park, J.-H. Son, S. J. Oh, and J.-S. Suh, "Characterization of blood using terahertz waves," J. Biomed. Opt. **18**(10), 107008 (2013).

Biomedical Optics EXPRESS

46. O. A. Smolyanskaya, O. V. Kravtsenyuk, A. V. Panchenko, E. L. Odlyanitskiy, J. P. Guillet, O. P. Cherkasova, and M. K. Khodzitsky, "Study of blood plasma optical properties in mice grafted with Ehrlich carcinoma in the frequency range 0.1–1.0 THz," Quantum Electron. **47**(11), 1031–1040 (2017).
47. M. M. Nazarov, O. P. Cherkasova, E. N. Lazareva, A. B. Bucharskaya, N. A. Navolokin, V. V. Tuchin, and A. P. Shkurinov, "A complex study of the peculiarities of blood serum absorption of rats with experimental liver cancer," Opt. Spectrosc. **126**(6), 721–729 (2019).
48. M. M. Nazarov, O. P. Cherkasova, and A. P. Shkurinov, "A comprehensive study of albumin solutions in the extended terahertz frequency range," J. Infrared, Millimeter, Terahertz Waves **39**, 840 (2018).
49. O. P. Cherkasova, D. S. Serdyukova, A. S. Ratushnyak, E. F. Nemova, E. N. Kozlov, Y. V. Shidlovskii, K. I. Zaytsev, and V. V. Tuchin, "Effects of terahertz radiation on living cells: a review," Opt. Spectrosc. **128**(6), 855–866 (2020).
50. O. P. Cherkasova, M. M. Nazarov, M. Konnikova, and A. P. Shkurinov, "THz spectroscopy of bound water in glucose: direct measurements from crystalline to dissolved state," J. Infrared, Millimeter, Terahertz Waves **41**(9), 1057–1068 (2020).
51. Y. V. Kistenev, A. V. Shapovalov, A. V. Borisov, D. A. Vrazhnov, V. V. Nikolaev, and O. Y. Nikiforova, "Applications of principal component analysis to breath air absorption spectra profiles classification," Proc. SPIE **9810**, 98101Y (2015).
52. M. A. Cox and T. F. Cox, "Multidimensional scaling," in *Handbook of Data Visualization* (Springer, 2008), pp. 315–347.
53. L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction". arXiv preprint arXiv, 1802.03426 (2018).
54. H. Liu, K. Zhao, X. Liu, Z. Zhang, J. Qian, C. Zhang, and M. Liang, "Diagnosis of hepatocellular carcinoma based on a terahertz signal and VMD-CWSE," Biomed. Opt. Express **11**(9), 5045–5059 (2020).
55. D.A. Vrazhnov, V.V. Nikolaev, A. V. Shapovalov, and E. A. Sandykova, "The kernel construction for the biomedical data classification using support vector machine," Proc. SPIE **10614**, 106141Y (2018).
56. D. Elizondo, "The linear separability problem: Some testing methods," IEEE Trans. Neural Netw. **17**(2), 330–344 (2006).
57. A.A. Angeluts, A.V. Balakin, M.G. Evdokimov, M.N. Esaulkov, M.M. Nazarov, I.A. Ozheredov, D.A. Sapozhnikov, P.M. Solyankin, O.P. Cherkasova, and A.P. Shkurinov, "Characteristic responses of biological and nanoscale systems in the terahertz frequency range," Electron. Quantum. **44**, 614 (2014).
58. M. M. Nazarov, O. P. Cherkasova, and A. P. Shkurinov, "Study of the dielectric function of aqueous solutions of glucose and albumin by THz time-domain spectroscopy," Electron. Quantum. **46**(6), 488 (2016).
59. K. Shiraga, A. Adachi, M. Nakamura, T. Tajima, K. Ajito, and Y. Ogawa, "Characterization of the hydrogen-bond network of water around sucrose and trehalose: Microwave and terahertz spectroscopic study," The J. Chem. Phys. **146**(10), 105102 (2017).
60. R. J. Falconer and A. G. Markelz, "Terahertz spectroscopic analysis of peptides and proteins," J. Infrared Millim. Terahertz Waves **33**, 973−988 (2012).
61. S. Ohtake, Y. Kita, and T. Arakawa, "Interactions of formulation excipients with proteins in solution and in the dried state," Adv. Drug Delivery Rev. **63**(13), 1053–1073 (2011).