Prostate cancer identification via photoacoustic spectroscopy and machine learning

Yingna Chen a,1, Chengdang Xu b,1, Zhaoyu Zhang c, Anqi Zhu c, Xixi Xu a, Jing Pan a, Ying Liu b, Denglong Wu b, Shengsong Huang b,**, Qian Cheng a,d,**

a Institute of Acoustics, School of Physics Science and Engineering, Tongji University, Shanghai, China
b Department of Urology, Tongji Hospital, Tongji University School of Medicine, Shanghai, China
c School of Software Engineering, Tongji University, Shanghai, China
*d Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai, China

ARTICLE INFO

Keywords:
- Photoacoustic physio-chemical analysis
- Prostate cancer
- Machine learning
- Unsupervised hierarchical clustering
- Supervised classification

ABSTRACT

Photoacoustic spectroscopy can generate abundant chemical and physical information about biological tissues. However, this abundance of information makes it difficult to compare these tissues directly. Data mining methods can circumvent this problem. We describe the application of machine-learning methods (including unsupervised hierarchical clustering and supervised classification) to the diagnosis of prostate cancer by photoacoustic spectrum analysis. We focus on the content and distribution of hemoglobin, collagen, and lipids, because these molecules change during the development of prostate cancer. A higher correlation among the ultrasonic power spectra of these chemical components is observed in cancerous than in normal tissues, indicating that the microstructural distributions in cancerous tissues are more consistent. Different classifiers applied in cancer-tissue diagnoses achieved an accuracy of 82% (better than that of standard clinical methods). The technique thus exhibits great potential for painless early diagnosis of aggressive prostate cancer.

1. Introduction

Prostate cancer is a disease with a high incidence and low cure rate after metastasis. According to the 2019 cancer statistics, prostate cancer is the second most commonly diagnosed cancer in men worldwide (following lung and bronchus cancer) [1]. The medical-imaging diagnostic techniques most frequently used for prostate cancer, such as ultrasound or magnetic resonance imaging (MRI), fail to provide information on chemical composition, exhibit poor resolution, and incur high cost [2–4]. Thus, satisfying the criteria of high accuracy, sensitivity, and minimal invasiveness for prostate cancer diagnosis is still difficult.

Photoacoustic physio-chemical analysis (PAPCA) has exhibited great potential in the diagnosis of prostate cancer. Because of differences in molecular bonds and vibration modes [5–7], different biomacromolecules have unique light-absorption spectra and can be detected under different light exposures. Meanwhile, biological tissues can also be characterized according to their acoustic characteristics using the ultrasound power spectra of photoacoustic signals. Thanks to its ultrasonic and optical characteristics, this method can simultaneously assess the chemical compositions and microscopic histological characteristics of prostate tissue with high resolution and sensitivity with minimal invasiveness. Progress in this field will entail addressing the limitations of the existing imaging techniques in the diagnosis of prostate cancer, exploring pathological detection and analysis, and developing diagnostic tools.

PAPCA has been used for a variety of medical diagnostic purposes, including the detection of inflammation and fibrosis associated with Crohn’s disease [8], evaluation of liver conditions [9,10], bone assessment [11,12], microvascular imaging [13], and prostate cancer diagnosis [14,15]. Previous studies have shown that prostate cancer can be identified by photoacoustic detection of hemoglobin or lipids [16–18]. However, most of the existing research on the application potential of time-domain or frequency-domain photoacoustic spectra in the diagnosis of prostate cancer is focused on extracting physical quantization parameters [14,19–21]. The information provided by the extraction of
individual parameters for the evaluation of prostate cancer is limited, and the accuracy of this method needs to be improved. Adequate amount of information would help further improve diagnostic accuracy. However, additional work is required in this area because of the challenges associated with existing data analysis and mining methods.

The rapid development of machine learning provides an opportunity to analyze photoacoustic physio-chemical spectra completely. Unsupervised machine-learning algorithms, such as hierarchical clustering analysis, are widely used in gene-expression profile analysis. These algorithms can uncover the intrinsic correlations of gene expressions that are difficult to identify directly otherwise, illustrating the great potential of deep data mining \cite{22,23}. Supervised machine-learning methods, such as classification algorithms, can establish a classification model for parameter optimization depending on training-data samples and labels, thereby effectively improving the accuracy of classification. Machine learning has exhibited considerable potential in clinical applications, including the diagnosis of lung cancer, breast cancer, and brain tumors \cite{24–27}. Therefore, it appears reasonable that machine learning could be used to analyze the data of photoacoustic spectra and improve the accuracy of prostate cancer diagnoses.

In this study, PAPCA was combined with machine learning to analyze the characteristic variations in a photoacoustic physio-chemical spectrum corresponding to the evolution of prostate cancer. We aimed to clarify the feasibility of prostate cancer diagnostics based on the photoacoustic physio-chemical spectrum. To this end, 101 samples were obtained from 22 cases of radical prostatectomy. Then, based on unsupervised clustering \cite{28}, a correlation analysis was conducted on the power spectra of normal and tumor prostate tissues at different wavelengths. First, we aimed to explore the variations in the microstructural distributions of different biological macromolecules during the evolution of prostate cancer. Second, we aimed to clarify the feasibility of realizing the diagnosis of prostate cancer based on the photoacoustic physio-chemical spectrum. Subsequently, based on supervised linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) algorithms, we conducted dimensionality reduction and characteristic parameter extraction of the photoacoustic physio-chemical spectrum to diagnose prostate cancer.

Study details including sample collection, PAPCA data acquisition, unsupervised clustering methods, and supervised classification models are listed in the Materials and Methods section. The Results and Discussion section presents the corresponding results and related analysis. Finally, the Conclusion section provides a summary of the study.

2. Materials and methods

2.1. Sample collection and experimental protocol

Twenty-two male patients aged 61–85 years, with an average age of 71.9 years, volunteered to participate in the study. The number of patients were chosen based on the power study performed. The detailed results are presented in the attachment. All provided their informed consent, and all procedures were approved by the Institutional Review Committee of Tongji Hospital. The experiment was conducted in Tongji Hospital, Shanghai, China, and involved 22 prostate tissue specimens from patients who had not received treatment before surgery. Ex vivo prostates were surgically removed and then subjected to a sterile gauze to wipe the blood on their surfaces. The tissues were transported in an ice bag at 0 °C to a photoacoustic laboratory within 1 h after resection. The entire photoacoustic detection process was completed within 2 h. Under the guidance of experienced physicians and based on preoperative pathology, interstitial measurements by needle PA probes were acquired at a total of 101 locations. The measurement locations were marked by a syringe needle after PA measurement. The prostates were then sent back to Tongji Hospital for pathological diagnosis. The pathology confirmed that 52 normal sites and 49 cancerous sites were sampled (Table A2).

2.2. Photoacoustic signal acquisition

The signal acquisition setup (Fig. A4) used in this experiment comprised laser triggering and ultrasonic signal acquisition, that could simultaneously collect the photoacoustic signals of the samples and of blackbody. Here, we choose black body to calibrate laser pulse energy instead of photodiode devices because the bandwidth of optical wavelength is exceeding the spectral response range of a common silicon photodiode \cite{29}. The laser used in this experiment was a tunable optical parametric oscillator (Phocus Mobile, OPOTEK, Carlsbad, CA), which scanned wavelengths at intervals of 10 nm in the ranges 690–950 nm and 1200–1690 nm. This band can well cover the targeted chromophores including deoxymyoglobin, oxyhemoglobin, lipids, and collagen. The pulse duration was 2–5 ns and the pulse repetition rate was 10 Hz. In the optical path, a 90:10 beam splitter was used to split the laser beam, and the 10 % split light was irrigated on a blackbody. The 90 % split light was focused by the lens and coupled to a fiber diffuser (radius: 300 μm, length: 2 cm) we developed earlier \cite{30} for laser irradiation. Considering the optical energy attenuation through the lens group and the fiber coupling was approximately 70 % and the optical output area of the fiber diffuser was approximately 0.377 cm², we can determine the laser fluence in mJ cm⁻² for each wavelength, as shown in Fig. 1. The input energy to the tissue surface with the maxmiun of 11.4 mJ cm⁻² at 720 nm is satisfied within the ANSI limit.

The ultrasonic signal generated by the pulsed laser was collected using a needle hydrophone (HNC1500, ONDA Corp., Sunnyvale, CA) having a frequency response ranging from 0 to 20 MHz. The ultrasound signals of the blackbody were received by a focused transducer with a central frequency of 4.86 MHz and a -6 dB bandwidth of 69.79 % (V307-SU, Olympus Corp., Tokyo, Japan) and then a pulse receiver with a high-pass filter of 1 MHz and an amplifier of 25 dB (5073PR, Olympus Corp., Tokyo, Japan). Here, the 1 MHz high-pass filter was used to suppress the relative low frequency noise produced by the scattered pulsed laser irradiated to the surface of the hydrophone. The data were then collected using an oscilloscope (TDS3034B, Tektronix, Ohio, USA) at a sampling rate of 250 MHz and stored in a computer (ThinkPad S3–5440, Lenovo, China). 128 times average of signal is used to obtain sufficient signal-to-noise ratio (SNR). With the laser’s frequency at 10 pulses per second, we developed a program to automatically switch the wavelength after every 15 s.
2.3. Photoacoustic physio-chemical spectra

The data processing method we employed has been described in detail in our previous work [15]. The tissue time-domain signal was divided by the peak-to-peak value of the blackbody signal, and the laser energy of the tissue signal was calibrated. Subsequently, the Welch method was used to measure the power spectra of the photoacoustic signals via MATLAB 2019B. By default, we used a Hamming window with a length of 2500 sampling points, and an overlap rate of 90% for the sliding calculation of the window. The frequency response range of the hydrophone was in the range 0–20 MHz. After a 1 MHz high-pass filtering, the frequency range of analysis was determined to be 1–20 MHz and the frequency spectral resolution was 0.1 MHz. The photoacoustic physio-chemical spectrum was obtained by arranging the measured power spectra corresponding to the wavelengths. By averaging the power spectra at different wavelengths of the photoacoustic physio-chemical spectrum, the intensity spectrum of the photoacoustic signal was obtained, which reflected the content of the photoacoustic source. Then machine learning methods are used to analyze the characteristic of the spectra. Fig. 2 is a block diagram of the machine learning algorithm applied.

2.4. Pearson correlation coefficients map and hierarchical clustering

Overall, 77 wavelengths (690–950 nm, 1200–1690 nm, Δλ =10 nm) were numbered from zero to 76 by size. All machine-learning computations were implemented in Python 3.7, and visualizations were created on Baidu’s open-source visualization site, Echarts. The Python Pandas module calculated the Pearson correlation coefficients (PCCs) [31] for the power spectra of different wavelengths. Thus, the PCC matrix between the power spectra of all wavelengths was obtained. The dimension of the matrix was 77 × 77 (a symmetric diagonal matrix). By employing an unweighted pair group method with arithmetic mean (UPGMA), a cluster analysis was conducted on the PCC matrix, and the correlation-coefficient matrix after clustering was obtained. The clustering group was divided into six groups. Each sample point had its corresponding cluster’s PCCs matrix. By extracting the wavelength clustering of all samples, clusters corresponding to normal and tumorous samples were identified.

To better observe and analyze the differences between normal and tumor prostate tissues, we used the correlation-coefficient network diagram for further statistical analysis and visualization. Here, 77 wavelengths (690–950 nm, 1200–1690 nm, Δλ =10 nm) were regarded as 77 nodes, and all were distributed on the circumference of a circle. We calculated the proportion of the samples with a correlation coefficient R greater than 0.9 for every two wavelengths in the total sample size of this group. The proportion parameter was defined as the correlation weight WR. When the WR of two wavelengths exceeded 70%, we added a line between the nodes representing them. After computing the WRs of all wavelengths, the corresponding correlation-coefficient network diagram was constructed. In the correlation-coefficient network diagram, three indicators were used to represent the differences in correlation:

1) Connectivity: connectivity between the nodes only exists when WR > 0.7.
2) Node size: if a node has more connections, it has a larger size.
3) Node label (optional): the node label is only added when the node size is greater than the threshold.

We divided all samples into the normal and tumor groups, calculated the correlation-coefficient network of the two groups, and visualized the results.
2.5. Supervised classification: LDA

The raw photoacoustic physio-chemical spectrum was a two-dimensional 77 × 191 matrix, containing 77 wavelengths (690–950 nm, 1200–1690 nm, Δ\(\lambda\) =10 nm) and 191 frequency points (1–20 MHz, \(\Delta f\) =0.1 MHz). LDA was used to solve the dichotomy problem by calculating a one-dimensional matrix between classes and reducing this one-dimensional matrix to a single numerical value, called the mapping value. In this study, we considered the power-spectrum values of a single frequency point at different wavelengths as the elements of the initial 77-element one-dimensional matrix, which we reduced to a single value through LDA. After this mapping value was normalized, normal and tumor tissues could be identified by the corresponding mapping value being greater than zero. We calculated the mapping values of all frequency points at different wavelengths as the elements of the initial 77-element one-dimensional matrix, which we reduced to a single value through LDA. After this mapping value was normalized, normal and tumor tissues could be identified by the corresponding mapping value being greater than zero. We calculated the mapping values of all frequency points in advance and extracted those points that could completely distinguish normal and tumor samples based on the calculated default values. The mean values of these frequency-point mappings were used as eigenvalues for the final diagnosis of prostate cancer.

In the classifier test, 90 samples out of 101 were randomly selected for training to obtain an appropriate linear mapping. The remaining 11 samples were used for testing the model. For a single hold-out set, where 90 % of data are used for training and 10 % used for testing, the test set is very small, resulting in large variations in the performance estimate for different samples of data, or for different partitions of the data to form training and test sets. In our experiment, we conducted the 10-fold cross-validation estimation where we divided the data into ten groups with only one group used for testing and tested each divided group as a test set. With the k-fold cross-validation, we have trained k LDA models and have obtained the average accuracy of all models. By averaging over k different partitions, the variance is reduced, making the performance estimate less sensitive to the partitioning of the data. Therefore, the robustness of the model is guaranteed. The choice of k has no formal rule. We chose 10-fold cross-validation for its unbiasedness and low variance as per Ron Kohavi’s study on cross-validation [32]. The model was evaluated by 10-fold cross-validation, which was repeated three times, and the average value was calculated to obtain the accuracy of the final model.

2.6. Supervised classification: QDA

Considering that the photoacoustic spectrum parameters and the malignancy of prostate cancer may be nonlinear, we also used QDA as a second classification method. The optimal quadric surface was used to separate the subspaces. The use of the photoacoustic physio-chemical spectrum was consistent with LDA. Principal component analysis (PCA) was used to reduce the dimension of the 1 × 77 matrix at different frequency points, and the QDA quadric surface was optimized. QDA anticipated the classification performance of a single frequency point and predicted the combination of multiple frequency points with a classification accuracy higher than 65 %. The test of the classifier was also consistent with LDA. The training set included 90 randomly selected samples out of 101. The remaining 11 samples comprised the test set. The model was evaluated by 10-fold cross-validation, which was repeated three times, and the average value was calculated to obtain the accuracy of the final model.

2.7. Statistical analysis

According to the conventional photoacoustic spectrum analysis methods we proposed before [33], we extracted the slope and intercept as the conventional quantification parameters for prostate cancer identification. Summarized quantification results as means ± standard error of the means were shown in the attachment (Fig. A5). A one-sided \(P\)-value of less than 0.05 was performed for a significance analysis. Then, we calculated the area under cure (AUC) to compare the performance of conventional analysis methods and machine learning. The conventional quantification parameters extraction was performed using Matlab and all AUC calculation was performed using GraphPad Prism software and Python (3.7).

3. Results and discussion

3.1. Photoacoustic physio-chemical spectra of prostate tissues

The measured photoacoustic power spectra were arranged by wavelength to obtain the original photoacoustic physio-chemical spectrum. The photoacoustic physio-chemical spectra of typical normal and cancer tissues are shown in Fig. 3a. The spectra reveal the differences between normal and tumor tissues. In both regions, the photoacoustic physio-chemical signal of cancer tissues is considerably stronger than that of normal tissues (orange dashed box). In the 690–950 nm band, the main source of light and sound is hemoglobin. The color of tumor tissues in this band is significantly stronger than that in normal tissues because the proliferation of blood vessels during cancer evolution increases the overall hemoglobin content of tumor tissues. In the 1200–1370 nm band, the main photoacoustic sources are lipids and collagen. Their
Fig. 4 a is highly consistent with the light-absorption group of different tree are not arranged in the order of 0 all tissue samples were statistically analyzed, as shown in Fig. 4 c. No (71–1260 nm) biomolecules in Fig. 4 b. Furthermore, different bio photograph physio-chemical spectra

3.2. Correlation analysis of biomacromolecules of prostate tissues through
photoacoustic physio-chemical spectra

Using points of different shading to represent proximity and raw data matrices is not new to researchers in the fields of taxonomy and cluster analysis, and has proven useful for recovering missing structure or obtaining better structure [37]. The 77 wavelengths were clustered into six wavelength groups using the PCCs and UPGMA (Fig. 4a). The clustering results were then compared with the light-absorption spectra of the biomolecules (Fig. 4b). Although the 77 terminal nodes of the cluster tree are not arranged in the order of 0–76, the wavelength grouping in Fig. 4a is highly consistent with the light-absorption group of different biological macromolecules in Fig. 4b. Furthermore, different biomacromolecules have unique light-absorption spectra because of differences in molecular bonds and vibration modes. UPGMA helps to analyze the differences in photoacoustic spectra and define the detection wavelength range in which photoacoustic signals are generated by different biomacromolecules dominate. The groups are as follows: V1 (#0–11; 690–800 nm): deoxyhemoglobin; V2 (#12–26; 810–950 nm): oxyhemoglobin; V3 (#27–32; 1200–1250 nm): lipid 1; V4 (#33–47; 1260–1400 nm): collagen; V5 (#48–70; 1410–1630 nm): water; V6 (#71–76; 1640–1690 nm): lipid 2. The wavelength clustering results of all tissue samples were statistically analyzed, as shown in Fig. 4c. No significant difference in clustering results between normal samples and cancerous samples was observed because the main biomacromolecules in normal or cancerous tissues did not vary.

Fig. 5a–d show typical cluster diagrams of normal and cancerous tissue. The different intensities of pseudo-color reflect the intergroup and intragroup correlation coefficients of various molecular groups. In the clustering process of UPGMA, the relative positions of the groups were different, but the results were not affected, therefore, we combined V3–V6 when displaying the analysis. From these samples, the vascular correlation groups (V1–V2) showed a low correlation with other groups (V3–V6) because the vascular framework was different from the distribution of other components. In addition, by comparing the normal and cancerous samples, we observed that the intergroup correlation between V1–V2 and V3–V6 was increased (the color was deepened) in the cancerous tissues. Because of their high correlation, V1 and V2 were placed in the same category (the red group in Fig. 5c). This correlation is even more evident in the network maps (Fig. 5e–f) representing the statistical results of all samples. In terms of all three indicators mentioned earlier, we can see an increase in the correlation between different groups: 1) A network map of cancerous tissues contains more edges (938 in the figure) than normal tissues (673). (In the figure, the green arrow points to some extra links.) 2) The node sizes in the cancerous-tissue network map increases considerably (see the nodes by the orange arrows, for example). 3) More labels are present in the network map of cancerous tissues (see box in figure indicated by purple arrow).

The correlation increase between lipids and collagens is especially evident due to the microstructural changes of these two biological macromolecules during the development of prostate cancer. Normally, collagen fibers are the structural network supporting prostate tissues [38,39], while lipids are the main components of cell membranes that exist in exosomes of the extracellular environment [40]. They are distributed in different areas of the organization and therefore have different distribution structures. However, when prostate tissues are cancerous, the environment changes and metabolism is abnormal [41]. The complex tumor environment increases the complexity of photoacoustic spectrum and blurs the distribution characteristics of single substance, thus increasing the correlation of photoacoustic spectrum.

Based on correlation and cluster analyses, wavelengths could be grouped depending on power-spectral similarity, providing an exact reference for identification of different biological macromolecules. In addition, the above analysis shows that the photoacoustic physiochemical spectrum can effectively reflect the variations in the microstructure of biological macromolecules during the evolution of prostate cancer; therefore, the diagnosis of prostate cancer based on the photoacoustic physio-chemical spectrum is effective and feasible.

3.3. Identification of prostate cancer through LDA and QDA

Although the correlation and cluster analyses demonstrate that the photoacoustic physio-chemical spectrum can effectively reflect the microstructural changes of different biological macromolecules, the overall correlation of the correlation-coefficient graph is relatively high (> 0.75), indicating that redundant information is present. This high
correlation is not conducive to solving the classification problem. (The diagnosis of prostate cancer is essentially the binary classification problem of distinguishing normal tissues from tumor tissues.) Therefore, we have introduced LDA and QDA models to remove redundant information in the wavelength dimension, extract parameters that could better reflect tumor characteristics, and improve the diagnostic accuracy.

We trained the LDA model at different frequency points and calculated the mapping values of all 101 samples under these LDA models to analyze their distribution. Based on the LDA obtained by training at Fig. 5. Comparison of correlation between normal and cancerous samples. (a) Clustering tree of a normal sample. (b) Heatmap of the normal sample. (c) Clustering tree of a cancerous sample. (d) Heatmap of the cancerous sample. (e) Correlation weight (WR) network map of normal group. (f) WR network map of tumor group. Note increased number of links (green arrow) and labels (purple arrow), as well as larger nodes (orange arrows). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Fig. 6. Distribution of mapping values of 101 samples at 2.8 and 5 MHz. The first 52 samples are normal samples, and the last 49 are cancerous samples. (a) Normal and tumor samples can be completely distinguished at 2.8 MHz; (b) Normal and tumor samples cannot be completely distinguished at 5 MHz.
different frequency points, the mapping-value distribution of samples had two main distributions, as shown in Fig. 6. The first 52 samples were normal samples, and the last 49 were tumor samples. As shown in Fig. 6a, in the first distribution, the calculated mapping value of the normal sample could be completely distinguished from the preset value of the tumor sample by a threshold (red dashed line). The corresponding graph for the second distribution is shown in Fig. 6b. No threshold was found that completely distinguished the normal sample from the tumor sample (red dashed line). It’s clear that the frequency points without overlapping areas of the mapping-value distribution were more effective in distinguishing between normal and tumor samples. The superficial reason for the difference in performance of different frequencies is the difference in data features. According to the theory of linear discriminant analysis (LDA) algorithm, finding a linear mapping can maximize the variance between classes and minimize the variance within classes to achieve target discrimination. Therefore, the data features at these two frequency points are different, which makes LDA realize different degrees of discrimination between them. The fundamental reason is that, under different frequency points, our target chromophore size is different, and its correlation with pathological features or characterization ability is different. Therefore, to improve the diagnostic accuracy, we adopted the method of combination evaluation, which entailed obtaining the mean value of the calculated mapping values at these frequency points as the final classification eigenvalue.

The selected frequency points and prostate cancer diagnoses based on these single frequencies are listed in Table 1. The combined assessment also demonstrated that the average accuracy could reach 76.8% after 10-fold cross-validation, repeated three times for 101 samples. This accuracy rate is on par with that of techniques commonly used in clinical practice; the accuracy rate of multi-mode ultrasound is approximately 71.7% [42], and the diagnostic accuracy rate of MRI is 80% [43]. Thus, with the LDA method, the accuracy of photoacoustic spectroscopy for prostate cancer diagnosis is comparable to that of MRI and ultrasound. Theoretically, LDA is more efficient in classifying data with the same covariance matrix. Therefore, it is reasonable to believe that as the accumulation of data increases, the characteristics of the data, including the covariance matrix, will become clearer as to the diagnostic upper limit that can be achieved with LDA. A more accurate linear mapping with maximum inter-class variance and minimum intra-class variance can be found with more data to further improve the diagnostic accuracy. Therefore, if significant differences exist in the covariance matrix of characteristic distribution between normal samples and tumor samples in the accumulated data, the LDA will reach the upper limit of judgment that can be realized by this algorithm.

LDA is suitable for small sample classification. Based on the assumption of linearity in the parameters and disease evolution, multi-dimensional parameters can be reduced to a single value for classification. In fact, the parameters of the photoacoustic physio-chemical spectrum and tumor evolution might be nonlinear. Therefore, we have also introduced a second method of discriminant analysis (QDA) as a classification method. Here, we use PCA to reduce the dimension of the wavelength at different frequency points. The portfolio is also assessed. We estimate the classification performance by considering different frequency points and dimensionality-reduction combinations. In the process of estimation, combinations with classification accuracy higher than 65% are used for assembly evaluation. The selected frequency points in the dimensionality-reduction combination are shown in Fig. 7, where the accuracy of any combination is represented by the size of the circle. In the same manner, after 10-fold cross-validation was repeated thrice for all 101 samples, the final average accuracy obtained was 81.7%, and the sensitivity and specificity were 78.2% and 87.1%, respectively. This accuracy rate exceeded 80%, which is an improvement over the accuracies obtained via MRI and multimodal ultrasound.

Finally, we compared the performance between machine learning-based methods and the results of our previous work on conventional signal-processing-based method. As we know, from the principle to methods, there is some difference between machine learning and conventional methods. Thus, here we listed a comparison table for two methods, as shown in Fig. 8.

In addition, based on the parameter AUC (Area Under the ROC Curve), we compared the classification performance of linear fitting quantification parameters of PA spectra at 700, 1210, and 1370 nm according to our previous conventional signal-processing-based methods and that of LDA and QDA. The results are shown in Table 2. As the closer AUC is to 1, the better the discrimination effect is. Therefore, it can be seen that the method based on machine learning does have a better distinguishing effect.

4. Conclusions

In this study, we used the photoacoustic physio-chemical spectra of normal and cancerous prostate tissue to establish diagnostic models for prostate cancer. Based on the UPGMA cluster analysis method, we analyzed the correlation differences between different macromolecules in normal and cancerous tissues, and distinguished the characteristic light-absorption bands of different biological macromolecules, V1-V6. We found that the power-spectrum correlation of hemoglobin, collagen, and lipids is considerably higher for cancerous than for normal prostate tissues, reflecting an increase in the microstructure similarity of the distribution of these biomolecules according to the visualization results of V1–V6 of normal and tumor samples. This can be that the
complex tumor environment increases the complexity of photoacoustic spectrum and blurs the distribution characteristics of single substance, thus increasing the correlation of photoacoustic spectrum. These findings indicate that the photoacoustic physio-chemical spectrum can effectively reflect the changes in biological macromolecules during tumor development, and thus be used to diagnose prostate cancer. The accuracy rate of LDA is 76.3 %, which is comparable to that of multimodal ultrasound and MRI commonly used in clinical practice. The accuracy rate of QDA reaches 81.7 %, further improving the accuracy of the prostate cancer diagnosis.

The results of this study prove the feasibility and effectiveness of the combination of machine learning and photoacoustic physio-chemical spectroscopy in exploring the variations in the microscopic structure and chemical composition of prostate cancer and thus facilitating its diagnosis. In the future, more samples should be accumulated, and the classification model should be further optimized to further improve the accuracy of the diagnosis. In addition, benign and malignant tumors should be distinguished in the classification to enhance the clinical utility of the method.

The prostate is located deep inside human body and is difficult to examine in human clinical trials. However, the prostate is close to the rectum and filled with tissue fluid for coupling, providing the possibility for two methods for clinical application. One method is to receive photoacoustic signals via transmission type. The light source enters from the urethra while the photoacoustic signal is received through rectum. Our previous research shows that this method can effectively detect the prostate with a radius of 2 cm [44]. The other method is reflection type where both the light source and ultrasound probe are inserted through the rectum. The existing transrectal ultrasound probe has a certain empty space to realize the embedding of fiber laser source; however, this is a more difficult process.

CRediT authorship contribution statement

Yingna Chen: Methodology, Formal analysis, Verification and Writing-Original Draft. Chendang Xu: Methodology, Investigation and Resources. Zhaozyu Zhang: Formal analysis and visualization. Anqi Zhu: Formal analysis. Xixi Xu: Formal analysis. Jing Pan: Formal analysis. Ying Liu: Resources. Denglong Wu: Conceptualization and Project administration. Shengsong Huang: Conceptualization, Writing-Reviewing and editing and Supervision. Qian Cheng: Conceptualization, Project administration and Funding acquisition.
Author contributions

Q.C. and S.H. conceived the project and designed the research; Y.C. and C.X. contributed equally to the design, execution, and analysis of the experiments; S.H. and D.W. collected the prostate samples; J.P. assisted with the execution of the experiment; C.X. and Y.L. followed up patients and registered clinical data; Z.Z., A.Z., and X.X performed the data processing by machine learning; and Q.C. supervised the project. All authors contributed to the preparation and revision of the manuscript.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Funding

This study was supported by the National Natural Science Foundation of China [grant numbers 12034015, 11674249]; the National Key Research and Development Program of China [grant number 2017YFC0111400]; Shanghai Health and Family Planning Commission Foundation [grant number 201640051]; the Science Lead Supporting Project of Shanghai [grant number 18411961100] and Clinical Research Plan of SHDC [grant number SHDC2020CR3074B].

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Appendix A

Power study to determine the number of patients needed

Based on the conventional signal processing results of photoacoustic spectrum at 1210 nm and 1370 nm (Fig. A1), we conducted the power studies using the software PASS to calculate the minimum number of samples required. In particular, two-sample T-test assuming equal variance were prepared. The results are shown as follows.

Based on results at 1210 nm, the mean values of slopes are -0.21 and -0.12 respectively and numeric results for two-sample T-test assuming equal variance are shown in Table 1 and Fig. 1.

Fig. A2. Power study results based on slopes of photoacoustic power spectrum at 1210 nm.

Fig. A3. Power study results based on slopes of photoacoustic power spectrum at 1370 nm.
and 0.073 respectively and numeric results for two-sample T-test assuming equal variance are shown in Table 2 and Fig. 2.

Alternative Hypothesis: $H_1: \delta = \mu_1 - \mu_2 < 0$

According to the power study results, the sample size from 7 to 34 can meet the minimum sample requirements. Thus, over the last two years, we collected as many samples as we could within the process,
Table A3
101 samples from 22 cases of whole prostate from rad prostatectomy.

| Prostate # | Pathology-P. | Pathology-N. | Whole Prostate Pathology |
|------------|--------------|--------------|--------------------------|
| 1          | 5            | 3            | 2                        | A                      |
| 2          | 5            | 2            | 3                        | A                      |
| 3          | 4            | 2            | 2                        | A                      |
| 4          | 2            | 2            | 2                        | A                      |
| 5          | 2            | 1            | 1                        | A                      |
| 6          | 3            | 2            | 1                        | A                      |
| 7          | 9            | 0            | 0                        | A                      |
| 8          | 3            | 0            | 3                        | NA                     |
| 9          | 5            | 3            | 2                        | A                      |
| 10         | 5            | 3            | 2                        | A                      |
| 11         | 6            | 3            | 3                        | A                      |
| 12         | 5            | 0            | 5                        | NA                     |
| 13         | 5            | 0            | 5                        | NA                     |
| 14         | 5            | 4            | 1                        | A                      |
| 15         | 6            | 0            | 6                        | NA                     |
| 16         | 5            | 2            | 3                        | A                      |
| 17         | 5            | 1            | 4                        | A                      |
| 18         | 5            | 2            | 3                        | A                      |
| 19         | 4            | 3            | 1                        | A                      |
| 20         | 4            | 4            | 0                        | A                      |
| 21         | 3            | 1            | 2                        | A                      |
| 22         | 3            | 2            | 1                        | A                      |

Total 101 49 52

Note: P: Pathologically positive; N: Pathologically negative; A: agressive cancer.
NA: Non-agressive tissue.

totaling 22 prostate tissue samples. Figs. A2 and A3

Tables A1 and A3

References
[1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2019, CA Cancer J. Clin. 69 (2019) 7–34, https://doi.org/10.3322/caac.21551.
[2] J.-M. Correas, E.J. Halpern, R.G. Barr, S. Ghai, J. Walz, C. Dariane, J. de Rosette, Advanced ultrasound in the diagnosis of prostate cancer, World J. Urol. (2020), https://doi.org/10.1007/s00246-020-01923-0.
[3] R.M. Hicks, J.P. Simko, A.C. Westphalen, H.G. Nguyen, K.L. Greene, L. Zhang, P. R. Carroll, T.A. Hope, Diagnostic accuracy of 99mGa-PaSMa-11 PET/MRI compared with multiparametric MRI in the detection of prostate Cancer, Radiology 289 (2013) 793–797, https://doi.org/10.1148/radiol.2013121877.
[4] A. Stabile, F. Giganti, A.B. Rosenkrantz, S.S. Taneja, G. Villeirs, I.S. Gill, C. Allen, M. Emberton, C.M. Moore, V. Kasivisvanathan, Multiparametric MRI for prostate cancer diagnosis: current status and future directions, Nat. Rev. Urol. 17 (2020) 41-61, https://doi.org/10.1038/s41571-019-0212-4.
[5] G. Herzberg, Mol. Spectra Mol. Struct, Read Books Limited, 2013, p. 680.
[6] S.F. Mason, Molecular electronic absorption spectroscopy[J], Quarterly Rev. Chem. Soc. 15 (1961) 287-371.
[7] R.C. Herman, R. Hofstadter, Vibration Spectra and Molecular Structure V. Infrared Studies on Light and Heavy Acetic Acids, J. Chem. Phys. 6 (1938) 534-540, https://doi.org/10.1063/1.1750306.
[8] G. Xu, L.A. Johnson, J. Hu, J.R. Dillman, P.R.D. Higgins, X. Wang, Detecting inflammation and fibrosis in bowel wall with photoacoustic imaging in a crun’s disease animal model, in: A.A. Orovesky, L.V. Wang (Eds.), San Francisco, California, United States, 2016, p. 97080I, https://doi.org/10.1117/12.2213064.
[9] G. Xu, A. Norton, T.A. Lister, J. Mesirov, D.S. Unger, D. W. West, A. Norton, T.A., A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Luss, A. Rosenwald, J. C. Boldrich, H. Sacket, T. Yan, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. H. Jr, L. Liu, D.B. Lewis, R. Tihbhirani, G. Sherlock, W.C. Chua, T.C. Greiner, B. K. Weisenburger, J.O. Armitage, R. Warke, T.L. Weyr, M. Graver, J. C. Byrd, D. Botstein, P.O. Brown, L.M. Staudt, Distinct types of diffuse large B-cell lymphomas identified by gene expression profiling, Nature 403 (2000) 9.
[10] M.A. Shipk, K.N. Ross, P. Tamayo, A.P. Weng, J.I. Katoh, R.C.T. Aguir, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, K.W. Last, A. Norton, T.A. Lister, J. Mesirov, D.S. Neurb, E.S. Lander, J. Aster, T.R. Golub, Diffuse large b-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning, Nat. Med. 8 (2002) 68-74, https://doi.org/10.1038/nm1062.
[11] A. A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Luss, A. Rosenwald, J. C. Boldrich, H. Sacket, T. Yan, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. H. Jr, L. Liu, D.B. Lewis, R. Tihbhirani, G. Sherlock, W.C. Chua, T.C. Greiner, B. K. Weisenburger, J.O. Armitage, R. Warke, T.L. Weyr, M. Graver, J. C. Byrd, D. Botstein, P.O. Brown, L.M. Staudt, Distinct types of diffuse large B-cell lymphomas identified by gene expression profiling, Nature 403 (2000) 9.
[12] S. Huang, Y. Qin, Y. Chen, J. Pan, C. Xu, D. Wu, W.-Y. Chao, J.T. Wei, S.A. Tomlins, X. Wang, J.B. Fowler, P.L. Carson, Q. Cheng, G. Xu, Interstitial assessment of agressive prostate cancer by photio-chemical photoacoustic: an ex vivo study with intact human prostates, Med. Phys. 45 (2018) 412-4132, https://doi.org/10.1002/mp.13061.
[13] G. Xu, M.C. Davis, J. Siddiqui, S.A. Tomlins, S. Huang, L.P. Kunji, J.T. Wei, X. Wang, Quantifying Gleason scores with photoacoustic spectroscopic and Validation study with human tissues, Biomed. Opt. Express 6 (2015) 4781, https://doi.org/10.1117/12.2213064.
[14] T. Feng, K. Kozloff, M. Cao, Q. Cheng, J. Yuan, X. Wang, Study of photoacoustic measurement of bone health based on clinically relevant models, in: B. Choi, N. Kolliss, L. Zeng, H.W. Kang, B.J.F. Wong, J.F. Ilgner, G. Tearney, K. W. Gregory, L. Marcu, M.C. Skala, P.J. Campagnola, A. Mandelis, M.D. Morris (Eds.), San Francisco, California, United States, 2016, p. 96894F, https://doi.org/10.1117/12.2215136.
[15] X. Gao, C. Tao, X. Wang, X. Lii, Quantitative imaging of microvascularization in deep tissue with a spectrum-based photo-acoustic microscopy, Opt. Lett. 40 (2015) 970, https://doi.org/10.1364/OL.40.000970.
using native fluorescence spectroscopy with selective excitation wavelength,  
J. Biomed. Opt. 15 (2010), 047008, https://doi.org/10.1117/1.3463479.

[37] C.-H. Chen, Generalized Association Plots: Information Visualization Via Iteratively  
Generated Correlation Matrices, 2020, p. 24.

[38] C. Morrison, J. Thornhill, E. Gaffney, The connective tissue framework in the normal prostate, B.P.H and prostate cancer; analysis by scanning electron microscopy after cellular digestion, Urol. Res. 28 (2000) 304–307, https://doi.org/  
10.1007/s002400000123.

[39] Y. Pu, W. Wang, G. Tang, R.R. Alfano, Changes of collagen and nicotinamide  
adename dinucleotide in human cancerous and normal prostate tissues studied  
using native fluorescence spectroscopy with selective excitation wavelength,  
J. Biomed. Opt. 15 (2010), 047008, https://doi.org/10.1117/1.3463479.

[40] A. Llorente, T. Skotland, T. Sylvainne, D. Kauhanen, T. Rög, A. Orłowski,  
I. Vatralainen, K. Ekroon, K. Sandvig, Molecular lipidosomics of exosomes released by  
PC-3 prostate cancer cells, Biochim. Biophys. Acta - Mol. Cell Biol. Lipids 1831  
(2013) 1302–1309, https://doi.org/10.1016/j.bbalip.2013.04.011.

[41] X. Wu, G. Daniels, P. Lee, M.E. Monaco, Lipid metabolism in prostate cancer, (n.d.)  
10.

[42] L. Shuting, L. Jia, C. Bin, X. Shihao, Clinical value of transrectal multimodal  
ultrasound in diagnosis of prostate cancer, Chin. J. Med. Ultrasound Electron. Ed.  
17 (2020) 478–485, https://doi.org/10.3877/j.issn.1672-6448.2020.05.016.

[43] G. Murphy, M. Haider, S. Ghai, B. Sreeharsha, The expanding role of MRI in  
prostate Cancer, Am. J. Roentgenol. 201 (2013) 1229–1238, https://doi.org/  
10.2214/AJR.12.10178.

[44] H. Zhang, S. Huang, Y. Chen, W. Xie, M. Zhang, J. Pan, N. Sato, X. Wang, D. Wu,  
Q. Cheng, Examining the technical feasibility of prostate cancer molecular imaging  
by transrectal photoacoustic tomography with transurethral illumination, Exp.  
Biol. Med. 245 (2020) 313–320, https://doi.org/10.1177/1535370219884356.

Yingna Chen is currently a Ph.D. student in the Institute of Acoustics of School of Physics Science and Engineering at Tongji University at Shanghai, China. Her research interests are in the clinical translation of photoacoustic imaging and spectrum analysis.

Chengdang Xu is currently a Ph.D. student in the Department of Urology at Tongji Hospital of Tongji University at Shanghai, China. His research interests are in the basic prostate research and photoacoustic prostate diagnosis and evaluation.

Zhaoyu Zhang was once a senior student in the School of Software Engineering at Tongji University at Shanghai, China during the research. His major was image visualization and design.

Anqi Zhu was once a senior student in the School of Software Engineering at Tongji University at Shanghai, China during the research. Her major was machine learning and big data analysis.

Xixi Xu was once a senior student in the School of Physics Science and Engineering at Tongji University at Shanghai, China during the research. She was interested in the intelligent diagnosis of clinical diseases.

Jing Pan was once a master student in the Institute of Acoustics of School of Physics Science and Engineering at Tongji University at Shanghai, China during the research. She studies mainly in the photoacoustic signal analysis.

Ying Liu is currently a Ph.D. student in the Department of Urology at Tongji Hospital of Tongji University at Shanghai, China. She studies mainly in the basic prostate research.

Denglong Wu received the Ph.D. degree in Medicine from Shanghai Medical University (now School of Medicine at Fudan University) in 1998. He is currently the senior attending of urology at Tongji Hospital and a professor at Tongji University. He has been dedicated to the basic research prostate cancer and clinical diagnosis and treatment of it.
Shengsong Huang received the B.S. degree from Soochow University in 2005 and received his Ph.D. from Tongji University in 2019. He served as junior attending at Tongji Hospital. His work focuses on prostate cancer diagnosis and evaluation through photoacoustic imaging and spectrum analysis.

Qian Cheng received the B.S. degree in physics, M.S. and Ph.D. degree in acoustics from Tongji University, China, in 2000, 2003 and 2006, respectively. She is currently a professor at Tongji University. Since 2006, her research interests were optoacoustic phenomena, near-field acoustic imaging technique, Schlieren imaging technique and the development of the acoustic detecting instruments. Her most recent research has focused on the clinical translation of photoacoustic imaging and quantitative analysis, and in particular for tumor diagnosis and evaluation.