



OPEN

Identifying metastatic ability of prostate cancer cell lines using native fluorescence spectroscopy and machine learning methods

Jianpeng Xue¹, Yang Pu², Jason Smith^{3,4}, Xin Gao⁵, Chun Wang⁶ & Binlin Wu^{3✉}

Metastasis is the leading cause of mortalities in cancer patients due to the spreading of cancer cells to various organs. Detecting cancer and identifying its metastatic potential at the early stage is important. This may be achieved based on the quantification of the key biomolecular components within tissues and cells using recent optical spectroscopic techniques. The aim of this study was to develop a noninvasive label-free optical biopsy technique to retrieve the characteristic molecular information for detecting different metastatic potentials of prostate cancer cells. Herein we report using native fluorescence (NFL) spectroscopy along with machine learning (ML) to differentiate prostate cancer cells with different metastatic abilities. The ML algorithms including principal component analysis (PCA) and nonnegative matrix factorization (NMF) were used for dimension reduction and feature detection. The characteristic component spectra were used to identify the key biomolecules that are correlated with metastatic potentials. The relative concentrations of the molecular spectral components were retrieved and used to classify the cancer cells with different metastatic potentials. A multi-class classification was performed using support vector machines (SVMs). The NFL spectral data were collected from three prostate cancer cell lines with different levels of metastatic potentials. The key biomolecules in the prostate cancer cells were identified to be tryptophan, reduced nicotinamide adenine dinucleotide (NADH) and hypothetically lactate as well. The cancer cells with different metastatic potentials were classified with high accuracy using the relative concentrations of the key molecular components. The results suggest that the changes in the relative concentrations of these key fluorophores retrieved from NFL spectra may present potential criteria for detecting prostate cancer cells of different metastatic abilities.

The ability to metastasize is a fateful characteristic of certain malignant tumors, which currently account for the majority of cancer-related deaths. One third of all people worldwide will be diagnosed with a form of cancer during their lifespan. Currently, one third of these diagnosed cases will result in death due to metastasis. However, only a few notable methods have been developed to measure the metastatic ability present in a variety of cancers. In particular, prostate cancer afflicts more men than any other form in the western world—ranking third in terms of mortality after lung cancer and colorectal cancer¹. The detection of prostate cancer at an early stage is paramount for improving patient prognosis, as prostate cancer detected early has a significantly higher chance of successful treatment. One promising non-invasive method to diagnose cancers without removing tissue is based on optical spectroscopy, which was shown to be able to determine the state of tissue *ex vivo*^{2–5} as well as *in vivo*^{6–8} in previous studies. To characterize the properties of normal, benign, and malignant tissue and cells, one major focus in optical biopsy is measuring native fluorescence (NFL) spectra, which plays an important role in discrimination of cancerous tissue from normal tissue since the early studies of Alfano in 1980s².

¹The Engineering Research Center of Synthetic Peptide Drug Discovery and Evaluation of Jiangsu Province, China Pharmaceutical University, No. 639 Longmian Avenue, Jiangning District, Nanjing 211198, China. ²Davinci Applied Technologies Inc, 476 Expressway Dr. S., Medford, NY 11763, USA. ³Physics Department and CSCU Center for Nanotechnology, Southern Connecticut State University, New Haven, CT 06515, USA. ⁴Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA. ⁵Natural Sciences Department, LaGuardia Community College, City University of New York, Long Island City, NY 11101, USA. ⁶Center for Optics Research and Engineering of Shandong University, Jinan 211198, China. ✉email: wub1@southernct.edu

It is widely acknowledged that the emission spectrum of a tissue/cell is a superposition of spectra of various salient fluorophores^{9–11}. The main building-block fluorophores present in tissue include tryptophan, reduced nicotinamide adenine dinucleotide and its phosphorylated form [NAD(P)H], flavin adenine dinucleotide (FAD), collagen, and elastin. These molecules appear with different amounts and structure in tumor evolution and these changes can be revealed by NFL spectra^{2–5}.

Tryptophan is an essential amino acid. It cannot be produced by cells and must be supplied in the diet. Tryptophan is needed for protein synthesis. It is transported via large amino acid transporter system (LAT1/CD98) into the cells, where it is degraded to kynurenine by the enzyme indoleamine-2,3-dioxygenase^{12,13}. Many studies have shown that the degradation of tryptophan is a mechanism that tumors select to achieve immune escape^{12–15}. The failure of immune system control on the growth of tumor cells leads to the fast development of a tumor^{12,13,16}. Since there are more large amino acid transporters on the cell membrane of aggressive cancer cells, tryptophan can be taken up more efficiently in such cells than others from the surrounding environment.

Tryptophan is required for T lymphocyte effector functions^{17,18}. The T cells in immune system are particularly susceptible to low tryptophan concentrations, which results in energy and apoptosis so that cancer cells can escape from the immune detection and survive^{12,13}. Moreover, a pair of receptor and ligand proteins, PD-1 receptor on T cells and PD-L1/2 on cancer cells, have been observed to permit cancer cells to escape the immune system when PD-1 receptor is activated in low tryptophan environment. When the PD-1 binds to PD-L1/2, it suppresses the T cell activity, causing T cell apoptosis. These interactions help cancer cells escape from the immune detection and develop toward increasingly aggressive forms. Therefore, direct monitoring of the tryptophan level in cells/tissue can be used to investigate the immune escaping ability of the cancer cells and the metastasis ability of prostate and other cancer cells with low, mild and high aggressiveness.

We use NADH to denote both NADH and NADPH, since (a) the emission spectra from NADH and NADPH are identical¹⁹, and (b) the cellular content and fluorescence quantum yield of NADH are higher than that of NADPH, therefore we expect that our results mostly indicated the concentration of NADH²⁰. NADH is the principal electron donor in cellular metabolism. The nicotinamide adenine dinucleotide (NAD⁺), a natural coenzyme, regulates immune responses and creates homeostasis via a novel signaling pathway^{13,21}. Interestingly, quinolinic acid, which is a neurotoxic catabolite of the kynurenine pathway, is the major pathway for the de novo NAD⁺ synthesis. More importantly, the immunoregulatory properties of NAD⁺ are strongly related to the overexpression of tryptophan hydroxylase 1 (Tph1).

Quantification of the key biomolecular components within tissues/cells may present potential criteria for cancer detection and classification. Since the fluorescence signal is a mixed signal due to multiple fluorophores and the spectrum possesses high dimensional data, it is important to unmix the NFL spectra and reduce the dimension so that key components in the signal can be retrieved and analyzed. The measurement of the oxidation state of pyridine nucleotide [NAD(P)H] and the relative concentrations of other fluorophores is possible by decomposing the cellular spectral signal into individual fluorophores²². To unmix the NFL data, and retrieve the spectra and relative concentrations of the components of our interest, algorithms such as principal component analysis (PCA)^{23–25} and nonnegative matrix factorization (NMF)^{5,24,26} have been used.

In this study, the native fluorescence spectra of low metastatic (LNCaP), moderately metastatic (DU145), and advanced metastatic (PC-3) human prostate cell lines^{27,28} were studied using the selected excitation wavelength of 300 nm to investigate the key molecules such as tryptophan and NADH.

The basis spectra of these key fluorophores were obtained along with their relative concentrations, which were subsequently used to classify the samples using support vector machines (SVMs)²⁹. Our study provides a possible diagnosis method based on the changes of relative concentrations of tryptophan and NADH in cells which indicate the metastasis competence and the risk levels of cancer in patients.

This study was focused on the classification of prostate cell lines with different metastatic ability based on NFL and machine learning. The purpose was to detect invasiveness (metastatic potential) of prostate cancer cell lines, and find a quantitative relationship between the metastatic potential and the molecular information revealed by fluorescence spectroscopy. We first used PCA for linear unmixing. The component spectra and the relative concentrations retrieved by PCA may include negative values and cannot be attributed to particular fluorophores. Besides PCA, we also used NMF, which can potentially retrieve components for individual fluorophores due to the nonnegativity constraint.

Samples and methods

Sample preparation and cell lines. DU145 (ATCC, Virginia) cells were cultured in DMEM medium (Sigma, Missouri), supplemented with 10% fetal bovine serum (FBS) (Thermo Scientific, Massachusetts), 25 units/mL penicillin, and 25 mg/mL streptomycin (Sigma, Missouri). The LNCaP (ATCC, Virginia) cells were cultured in RPMI 1640 medium with 2 mM L-glutamine and 10% FBS. The PC-3 (ATCC, Virginia) cells were cultured in F-12K Medium (Sigma, Missouri) with 10% FBS. All cell lines were incubated under 5% CO₂ atmosphere at 37 °C. The cells were harvested at more than 95% cell confluence. The 0.25% trypsin-ethylenediaminetetraacetic acid (EDTA) (Sigma, Missouri) were used to treat the cell for less than 1 min. Then 5 mL cell culture media were used to wash the cells away from the bottom of the flask. The solutions were centrifuged at 2000 revolutions per minute (rpm) for 5 min. The supernatant with remaining trypsin were aspirated. The cells were resuspended with 5 mL phosphate buffered saline (PBS, Sigma, Missouri), and centrifuged again. The supernatant (almost all the solution) was removed. The remaining centrifuged cells were resuspended to $\sim 3.6 \times 10^6$ cells/mL with PBS. The cells were then transferred to a quartz cuvette (NSG Precision Cells, New York) with inner dimension of 1 cm × 1 cm × 4 cm for the subsequent fluorescence experiments. The cell suspensions were vortexed to mix evenly before each measurement. After the experiments, the rate of living cells was estimated using trypan blue solution (Sigma, Missouri).

Experimental parameters and data acquisition. In this study, a Fluorolog-3 spectrofluorometer system (HORIBA Scientific, Piscataway, New Jersey) was used to measure the NFL spectra of the cells. The excitation light was set to be 300 nm with 0.5 μ W power deposition and 5 nm spectral slit, and used to shine on samples with a spot size of $\sim 3 \text{ mm} \times 1 \text{ mm}$. The scan speed was set to be 300 nm per minute with each scan taking less than 1 min. The fluorescence signal with a spectral resolution of $\sim 1 \text{ nm}$ was collected in the range of 320 nm and 580 nm. To examine the amount of the background fluorescence, the emission spectrum of the supernatant from cell samples was also measured in the quartz cuvette. The intensity of the fluorescence signal at 340 nm measured from the supernatant was only 0.1 to 1.5% of that collected from cell samples. This background spectrum was deducted from the spectra for the cell samples before subsequent analysis. The excitation wavelength of 300 nm was selected because it was shown to be an optimal wavelength to collect both tryptophan and NADH fluorescence signals³⁰. Especially, tryptophan showed a quantum yield (QY) of 0.2 to 0.35, higher than the other fluorophores in cells and tissue³⁰.

Analytic methodology. The spectral data were first analyzed using PCA and NMF to unmix the fluorescence signals to reduce data dimensionality and acquire valuable features. PCA is a matrix decomposition technique which finds the uncorrelated orthogonal components (principal components or PCs) which account for the largest variances in the data. Usually, the PCs do not directly correspond to the basis spectra of the key fluorophores, but are instead linear combinations of them. The PCs are found as eigenvectors for an eigenvalue equation of the covariance matrix of the spectral data matrix. In the matrix form, $X^{m \times n} = W^{m \times r} \times H^{r \times n}$, where X is the input data matrix of dimension m wavelength values by n spectra, W is the matrix containing r basis spectra (PC loadings), and H is the corresponding weight matrix (PC scores). In practice, the eigenvectors can be found by singular value decomposition (SVD) of mean-centered data matrix X' . In matrix form, $X' = U \Sigma V^T$, where U and V are orthonormal left and right singular vectors respectively, and Σ is a diagonal matrix containing singular values. If $W = U$, H can be found by projecting data matrix X onto the PCs in W , i.e., $H = W^T X$. The PC scores in H are then used for potential classification.

NMF has been widely used within various fields of facial recognition²⁶, imaging^{31,32}, and time-series³³. The approach has also been employed for biomolecular spectral decomposition^{5,24}. Unlike PCA, NMF only uses nonnegativity constraints and is thus particularly well-positioned to retrieve the individual basis spectra for the key fluorophores³⁴. Using this non-negative technique on an inherently positive spectral dataset has been shown to allow for the extraction of intrinsic fluorescence spectra^{5,35}. NMF allows for the choice of internal dimensionality (r -value) for the solution matrix pair. The best-performing choice in r -value has been shown to be problem-specific. The problem that NMF attempts to solve is a multiplicative reconstruction of a given input matrix by minimizing the Frobenius norm, reaching a convergence in the process. This is shown mathematically as $X^{m \times n} = W^{m \times r} \times H^{r \times n}$, with $\|X - WH\|_F < t$, where t notates some threshold value. To find W and H , the values of W and H are initialized randomly, and updated using a certain algorithm such as a multiplicative update rule or alternating least squares^{26,34}. Thus, non-unique convergence is inherent. Therefore, Tikhonov regularization is commonly used to achieve convergence. Once convergence has been reached, matrix W is the resultant “feature matrix” containing r characteristic spectra that allow it to re-represent the spectra within the input matrix X using various weights contained in matrix H . Since the update condition turns every negative value to zero, the obtained matrix W has increased sparsity and resultant columns which give interpretable extracted spectral information.

Once the weights (scores) are obtained using either PCA or NMF, support vector machines (SVMs) are used to classify the spectra based on the weights in H . SVMs are widely used in classification problems to determine the boundary (commonly referred to as a hyperplane) that results in the largest separation between classes based on the closest data points which are called support vectors (SVs). Both linear and Gaussian radial basis function (RBF) kernels were employed.

To avoid bias in the classification, leave-one-out cross-validation (LOOCV)^{36,37} was used with SVM. LOOCV is a widely used technique for validation of discriminative performance. LOOCV begins by discarding a data point from the set. It then fits an SVM separation boundary using the remaining data and subsequently determines whether the previously left-out data point would be correctly grouped using the newly defined SVM separation. This process is iterated through every data point and provides an overall performance accuracy. This method provides a more robust evaluation of the classifier.

To further investigate the above models including PCA-SVM and NMF-SVM, the optimal number r_{opt} of the most relevant components for classification, which is considered a hyperparameter, was also evaluated. The process for hyperparameter optimization is similar to feature selection. To find r_{opt} , a nested leave-one-out cross validation (LOOCV) method³⁷⁻⁴⁰ was used to evaluate this hyperparameter. The internal LOOCV was used to evaluate this parameter. To avoid bias in the decomposition, the decomposition was performed only on the training set in the internal LOOCV to find the component loadings during hyperparameter optimization. The component loadings were then used to find a set of component scores for the test set in the internal LOOCV for validation. The optimal value of the hyperparameter was determined based on the highest accuracy³⁷. Once the optimal hyperparameter was found, the final evaluation was then performed using the external LOOCV. Similarly, new component loadings were found using the training set in the external LOOCV and used to find new component scores for the test set in the external LOOCV for final evaluation.

The performance of a two-class classification was evaluated using statistical measures such as sensitivity, specificity and accuracy⁴¹. A receiver operating characteristic (ROC) curve was used as another method to evaluate the performance of the model. The ROC curve was plotted as Sensitivity vs. 1-Specificity by varying the discrimination threshold for the binary classifier⁴². The area under the ROC curve, referred to as AUROC was used as a statistical measure to describe the predictive performance of the model^{43,44}. For multi-class classification

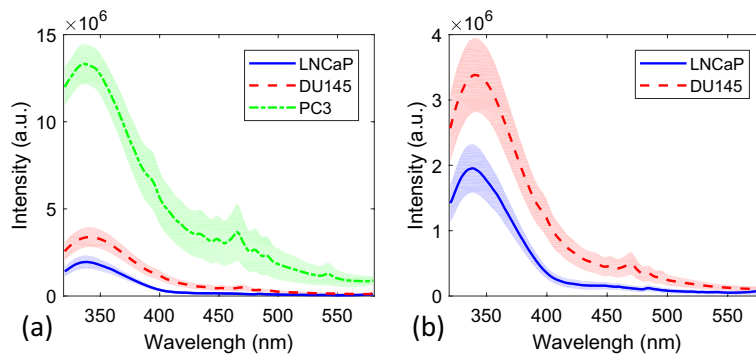


Figure 1. Average and error bounds of raw NFL spectra from (a) all three cell lines and (b) DU145 and PC-3.

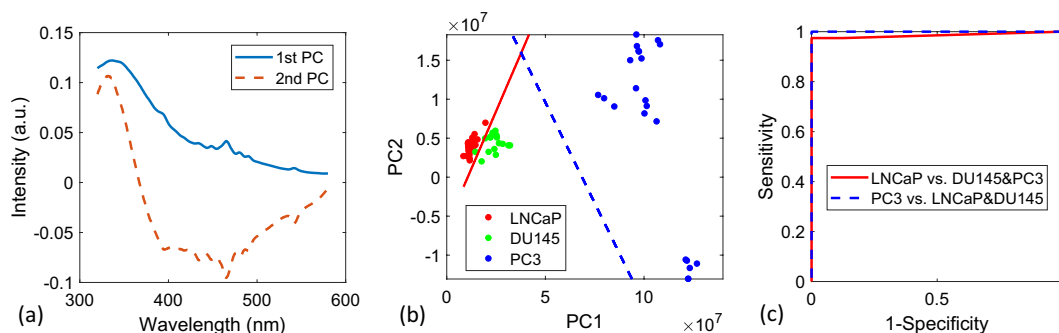


Figure 2. (a) Overlay of the first two extracted PCs. (b) Scatterplot of the first two PC scores for each NFL spectrum. The solid line separates the least advanced cell line (LNCaP) from the more advanced (DU145 and PC-3) and the dashed line shows the SVM boundary between the cell line possessing highest aggression (PC-3) and the two lowest (LNCaP and DU145). (c) The corresponding ROC curves for classifications with LOOCV.

with more than two classes involved, accuracy for classifying each class and an overall accuracy were used to evaluate the performance of the classification.

Experimental results

The viability of all the three cell samples used in this study were $\sim 98\%$ before spectral measurements using the cells in a quartz cuvette. This study was focused on the efficacy of the technique. Therefore, we require the optical signal be reliable and not change significantly during the signal integration. This was confirmed in the repeated measurements (data not shown here).

A total of 24, 17 and 23 NFL spectra were collected from the LNCaP, DU145 and PC-3 cell samples respectively by using 300 nm excitation. The average spectra of these acquisitions with error bounds are shown in Fig. 1a. For all three groups, the strongest fluorescence peak is located at ~ 340 nm, which is the characteristic emission peak of tryptophan. Because of high fluorescence intensity from PC-3 much stronger than the other two samples, a second plot is given in Fig. 1b with PC-3 removed to provide a higher clarity for the other two cell types. As observed in these two plots, a peak at ~ 465 nm is prominent in the DU145 spectra as well as in PC-3, which is considered to mainly contributed by NADH. It should be noted that across all wavelengths, the average fluorescence intensity collected from LNCaP is observed to be of lowest value, followed by DU145 and PC-3.

PCA and NMF were then utilized to determine the effectiveness of NFL for distinguishing metastatic potentials among different prostate cells. A 2D SVM boundary was first tested based on the first two components to classify between the LNCaP group and the other two (lowest risk vs. two highest) groups as well as between the PC-3 group and the remaining groups (highest risk vs. two lowest).

The first two basis spectra retrieved by PCA and a scatter plot of the weights with trained SVM classifiers are shown in Fig. 2a,b, respectively. The variances for the first two PCs contribute 99.97% of the total variance.

Figure 2a illustrates that the first two PCs are mainly attributed to tryptophan (~ 340 nm) and NADH (~ 465 nm) fluorescent signals. As illustrated by Fig. 2b, classification using these PCs, especially the first PC, provides strong evidence for their effectivity in prostate cell line discrimination. Since PCA allows for the existence of negative values, the 2nd PC possesses both a positive tryptophan peak and a broad negative peak above ~ 370 nm. The broad negative peak contains a peak about 460 nm which is attributed to NADH, along with a shoulder at ~ 400 nm which may be attributed to lactate⁴⁵. More insightful understanding needs be obtained to confirm the source of this peak. Although the spectra of the fluorophores in cells are not expected to be exactly the same as those for the corresponding pure chemicals, we collected and show the fluorescence spectra of the aqueous solution of the chemicals in Fig. 3 below for comparison with the retrieved component spectra. The spectra in

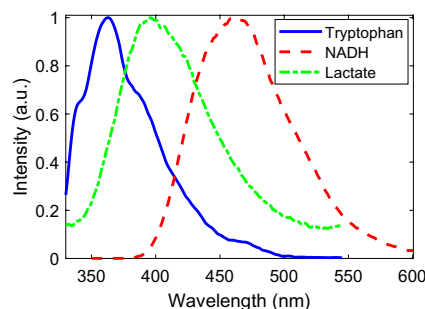


Figure 3. Fluorescence spectra of aqueous solutions of the three fluorophores.

(PCA) Classifiers	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUROC
Least aggressive vs. rest (LNCaP vs. DU145&PC-3)	97.4	100.0	98.4	0.986
Most aggressive vs. rest (PC-3 vs. LNCaP&DU145)	100.0	100.0	100.0	1.0

Table 1. Cross-validated sensitivity, specificity, accuracy and AUROC for LOOCV SVM using PCA.

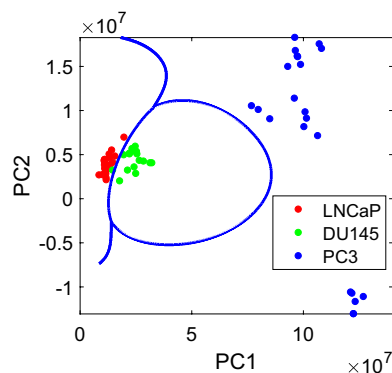


Figure 4. A multi-class SVM classification.

Fig. 3 are normalized to the spectra maxima, and consistent with literature data with emission maxima shown around ~360 nm, ~395 nm, ~460 nm for the three chemicals of tryptophan, lactate and NADH, respectively.

The SVM classifier was further evaluated with LOOCV. The corresponding ROC curves are shown in Fig. 2c. The resulting values of sensitivity, specificity, accuracy and AUROC are shown in Table 1.

A multi-class SVM classification was also performed for PC1 and PC2 using a Gaussian kernel²⁵. The classifier trained using all data is shown in Fig. 4. The LOOCV sensitivity, specificity, and accuracy were found to be 100.0%, 94.1%, and 100.0%, along with a total accuracy of 98.4%.

To further evaluate the model, the optimal number of PCs, r_{opt} , as a hyperparameter was evaluated using multi-class classification by SVM and a nested LOOCV. The r_{opt} value for PCA-SVM was found to be 3 based on the maximum accuracy. The external LOOCV was used for a final evaluation. The classification accuracies for all the cell types were found to be 95.8%, 100%, and 100% for LNCaP, DU145 and PC-3 respectively. A total accuracy was found to be 98.4%.

The other multivariate analysis method tested in this study was NMF. An overlay plot of the first two NMF-extracted nonnegative components (NCs) i.e., feature spectra is shown in Fig. 5a. In the initial analysis, r -value of 2 was used. As shown in Fig. 2a, the 1st NMF spectral feature shows a strong similarity with the fluorescent emission profile of tryptophan¹⁰. The NADH peak about 460 nm is not prominent. The 2nd NMF feature spectrum shows a NADH peak along with a peak at ~400 nm, which again might be due to lactate. Due to the inherent non-negativity of NFL spectra, the resulting feature spectra obtained by utilizing the positively-constrained technique, NMF, are more easily interpretable than with PCA. Figure 5b indicates both the relative concentrations of tryptophan and NADH and/or lactate (NADH/ lactate) increasing with the aggressiveness of the cancer cells.

Table 2 illustrates the LOOCV sensitivity, specificity and overall accuracy based on the two NCs. The LOOCV ROC curves are shown in Fig. 5c.

Similarly, a multi-class SVM classification was also performed with a Gaussian kernel using NC1 and NC2, as shown in Fig. 6. The LOOCV sensitivity, specificity, and accuracy were found to be 100.0%, 94.1%, and 100.0%, along with a total accuracy of 98.4%.

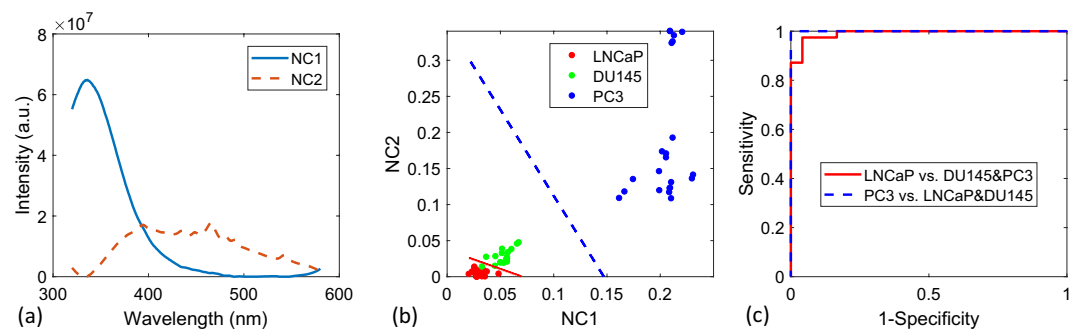


Figure 5. (a) Overlay of the first two extracted NMF feature spectra (solid and dashed corresponding to 1st and 2nd respectively). (b) Scatterplot of the corresponding NMF-related weight values for each NFL spectrum. The solid line separates the least advanced cell line (LNCaP) from the more advanced (DU145 and PC-3) and the dashed line shows the SVM boundary between the cell line possessing highest aggression (PC-3) and the two lowest (LNCaP and DU145). (c) The corresponding ROC curves for classification with LOOCV.

(NMF) Classifiers	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUROC
Least aggressive vs. rest (LNCaP vs. DU145&PC-3)	97.4	100.0	98.4	0.992
Most aggressive vs. rest (PC-3 vs. LNCaP&DU145)	100.0	100.0	100.0	1.0

Table 2. Cross-validated sensitivity, specificity, accuracy and AUROC for LOOCV SVM using NMF.

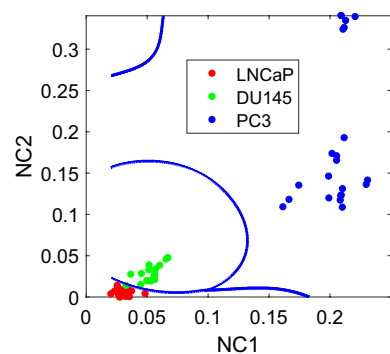


Figure 6. A multi-class SVM classification.

Classifiers	Accuracy (%) (LNCaP)	Accuracy (%) (DU145)	Accuracy (%) (PC-3)	Accuracy (%) Overall
PCA, SVM-LOOCV	100.0	94.1	100.0	98.4
PCA&SVM-LOOCV	95.8	100.0	100.0	98.4
NMF, SVM-LOOCV	100.0	94.1	100.0	98.4
NMF&SVM-LOOCV	100.0	94.1	100.0	98.4

Table 3. Cross-validated sensitivity and specificity values obtained from NFL classification using NMF.

The optimal number of NCs, r_{opt} was also evaluated using multi-class classification by SVM with a nested LOOCV. For simplicity, for any r -value used in NMF, all the r NCs will be then used for classification. The r_{opt} value for NMF-SVM was found to be 2. The classification accuracies based on the final evaluation in the external LOOCV were found to be identical to those above based on NMF without cross validation, i.e., 100.0%, 94.1%, 100.0%, and 98.4% for LNCaP, DU145, PC-3 and overall accuracy respectively.

The multi-class SVM classification with LOOCV for the two models based on PCA and NMF are summarized in Table 3 for comparison, where “PCA, SVM-LOOCV” means PCA was performed on the whole dataset and LOOCV was only used with SVM, while “PCA&SVM-LOOCV” means both PCA and SVM were performed on the training set in the cross-validation loops, as explained in detail above. Same is true for the classifier labels with NMF in the table.

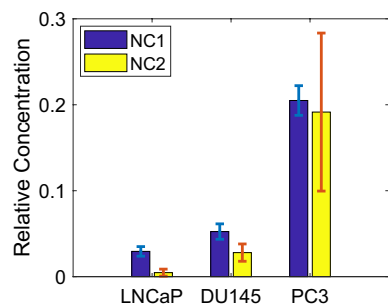


Figure 7. Relative concentrations of tryptophan and NADH/Lactate in three types of cells based on NC scores.

Table 3 shows the performances of the classification of cells with different metastatic potentials based on these two methods are almost identical. But the NCs retrieved by NMF are better interpretable. The relative concentrations of tryptophan and NADH/lactate indicated by NC1 and NC2 are plotted in Fig. 7, which clearly shows an increase in both components with the aggressiveness of the cancer cells. It is evident that errors also increase with the relative concentrations of the NCs across the cell lines. We hypothesize that this may be attributed to the increased inhomogeneity with increasing malignancy or aggressiveness of the cancer cells.

Summary and discussion

In this study, PC-3, DU145 and LNCaP cell lines were investigated using native fluorescence spectroscopy. Though all three cell lines used are prostate cancer cells, they are different cell lines with differences in biological characteristics^{27,28}. LNCaP cells are androgen-sensitive human prostate adenocarcinoma cells. DU145 cells are not hormone-sensitive and do not express prostate-specific antigen (PSA). PC-3 cells do not respond to androgens either, as well as to glucocorticoids or fibroblast growth factors.

All data were collected under the same experimental conditions and the raw data was analyzed to distinguish the cell types without normalization. It was assumed that the absolute signal levels directly reflected chemical concentrations. However, what we presented in the paper was still relative concentrations of the fluorophores among samples. We did not directly compare fluorescence signals between cells and homogeneous solutions to retrieve absolute molar concentrations of the fluorophores. The local environment of the fluorophores is not the same between live cells and homogeneous solutions, and the fluorescence properties of the molecules may not be the same. Therefore, usually they cannot be compared directly to retrieve the absolute concentrations of the fluorophores⁴⁶. This will be further verified during future investigation.

The spectral data were analyzed using machine learning based analysis methods PCA-SVM and NMF-SVM. In particular, multivariate analysis methods PCA and NMF were used to unmix the signal, reduce the data dimension, and detect spectral features. SVMs were used to classify different types of cell lines. This study provided further evidence for the potential of the NFL technique to discriminate among cancerous prostate cells of different metastatic ability with high accuracy. Classification between the most aggressive group (PC-3) with the other groups (LNCaP and DU145) was perfect for all methods illustrated herein. Results also showed that the accuracy for distinguishing the three different types of the cell samples using the two analytic methods was almost identical. A large sample size may provide a better comparison between the methods. But the two basis spectra and the coefficients obtained by NMF are more interpretable, and may be interpreted as the estimated fluorescence spectra of cell intrinsic biomolecules such as tryptophan, NADH and probably lactate, and their corresponding relative concentrations. Indeed, the relative concentrations of these biochemicals in PC-3 seemed to be much higher than those of DU145 and LNCaP, with the latter two being observed as significantly closer to each other (Fig. 7).

In this study, using a multi-class (three-class) classification, we found classifiers to separate every class from other two classes. The data and results of this study showed a direct relationship between the metastatic potential of the prostate cancer cells and the relative concentrations of the key biomolecules such as tryptophan and NADH retrieved from the NFL spectra.

The androgen receptor (AR) plays a critical role in the metastasis of prostate cancer cells, but its mechanism remains unknown⁴⁷. LNCaP cells are androgen-sensitive, while DU145 and PC-3 cells are not²⁸. But based on NFL spectral analysis, the difference between LNCaP and DU145 is small, while the difference between PC-3 and the combined group of LNCaP and DU145 is much larger, as evident in Figs. 3, 4, 5, and 6. Therefore, the classification of metastatic ability of the cells in our results cannot be entirely explained by hormone sensitivity. Even though understanding the mechanism of metastatic potential of prostate cancer cells are related to this study, the main purpose of this paper is to report a technique that may be used to detect cells with different metastatic potential.

Tryptophan is the main source of fluorescence signal in tissues and cells when excited at ~ 300 nm^{35,48,49}. The emission peak of tryptophan in aqueous solution was shown to be at about ~ 350 nm^{50,51}. This is consistent with our own observation in Fig. 3. Protein bound tryptophan has also been extensively studied in the literature^{52,53}. The emission peak wavelength of tryptophan in protein is sensitive to its local environment and ranges from ~ 308 to ~ 355 nm^{53,54}. In the results of this study, the peak assigned to NADH is usually mixed with another peak (shoulder) around ~ 400 nm, which we proposed to assign to lactate. Lactate is another key

molecule that is involved in carcinogenesis^{55,56}. Due to the Warburg effect, which is considered to be a hallmark of cancer, cancer cells show high rates of glycolysis in the hypoxia condition or even with oxygen^{55,56}. Higher rates of glucose metabolism through anaerobic glycolysis leads to a higher rate of L-Lactate release in cancer cells compared to normal cells^{57,58}.

In this study, the relative concentrations of tryptophan and NADH/lactate were found to correlate with the metastatic ability of the cancer cells. They may provide the criteria for prediction of tumor metastasis. Further, the combination of NFL spectroscopy and machine learning based analysis illustrates a high degree of capability for evaluation of tumor metastasis. NFL has the potential to be an alternative optical tool for medical armamentarium⁵⁹. Compared to other techniques for detection of metastatic potential of cancer cells such as those by Paidi et al.⁶⁰ and Bendau et al.⁶¹, the NFL technique can not only be used for measurements with cells in vitro and fresh tissue specimens ex vivo, but also for in vivo measurements. Since it does not involve spatial scanning (mapping), it works faster than those that require mapping^{60,61}, and can potentially be implemented for real-time diagnosis during surgery.

Since we only used three cell lines, whether or not the above-mentioned correlation between the relative concentrations of the key fluorophores and the metastatic potential of the cell lines universally exists for other prostate cell lines needs to be further verified. Future studies will continue to collect more data using different cell lines, obtain more insightful understanding of the spectra, and further test the fluorescence spectroscopic measurements using other excitation wavelengths.

Data availability

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

Received: 18 May 2020; Accepted: 8 January 2021

Published online: 26 January 2021

References

- Jemal, A. *et al.* Global cancer statistics. *CA Cancer J Clin.* **61**, 69–90 (2011).
- Alfano, R. R. *et al.* Laser induced fluorescence spectroscopy from native cancerous and normal tissue. *IEEE J. Quant. Electron.* **20**, 1507–1511 (1984).
- Pu, Y., Wang, W. B., Tang, G. C. & Alfano, R. R. Changes of collagen and nicotinamide adenine dinucleotide in human cancerous and normal prostate tissues studied using fluorescence spectroscopy with selective excitation wavelength. *J. Biomed. Opt.* **15**, 047008 (2010).
- Zhou, Y. *et al.* Human brain cancer studied by resonance Raman spectroscopy. *J. Biomed. Opt.* **17**, 116021 (2012).
- Wu, B., Gayen, S. K. & Xu, M. Fluorescence spectroscopy using excitation and emission matrix for quantification of tissue native fluorophores and cancer diagnosis. *Proc. SPIE* **8926**, 89261M (2014).
- Bergholt, M. S. *et al.* In vivo diagnosis of esophageal cancer using image-guided Raman endoscopy and biomolecular modeling. *Technol. Cancer Res. Treat.* **10**, 103–112 (2011).
- Panjehpour, M., Julius, C. E., Phan, M. N., Vo-Dinh, T. & Overholt, S. Laser-induced fluorescence spectroscopy for in vivo diagnosis of non-melanoma skin cancers. *Lasers Surg. Med.* **31**, 367–373 (2002).
- Alchab, L. *et al.* Towards an optical biopsy for the diagnosis of breast cancer in vivo by endogenous fluorescence spectroscopy. *J. Biophotonics* **3**, 373–384 (2010).
- Lakowicz, J. R. *Principles of Fluorescence Spectroscopy* (Plenum Press, New York, 1983).
- Wagnières, G. A., Star, W. M. & Wilson, B. C. In vivo fluorescence spectroscopy and imaging for oncological applications. *Photochem. Photobiol.* **68**, 603–632 (1998).
- Alfano, S., Wang, W. B. & Gayen, S. K. Lasers in cancer detection and diagnosis research: enabling characteristics with illustrative examples. *Technol. Cancer Res. Treat.* **4**, 663–673 (2005).
- Platten, M., Weller, M. & Wick, W. Shaping the glioma immune microenvironment through tryptophan metabolism. *CNS Oncol.* **1**, 99–106 (2012).
- Prendergast, G. C. Why tumours eat tryptophan. *Nature* **478**, 192–194 (2011).
- Labadie, B. W., Bao, R. & Luke, J. J. Reimagining IDO pathway inhibition in cancer immunotherapy via downstream focus on the tryptophan–kynurenine–aryl hydrocarbon axis. *Clin. Cancer Res.* **25**, 1462–1471 (2019).
- Platten, M., Nollen, E. A. A., Röhrig, U. F., Fallarino, F. & Opitz, C. A. Tryptophan metabolism as a common therapeutic target in cancer, neurodegeneration and beyond. *Nat. Rev. Drug. Discov.* **18**, 379–401 (2019).
- Böttcher, J. P. *et al.* NK cells stimulate recruitment of cDC1 into the tumor microenvironment promoting cancer immune control. *Cell* **172**, 1022–1037 (2019).
- Opitz, C. A. *et al.* An endogenous tumour-promoting ligand of the human aryl hydrocarbon receptor. *Nature* **478**, 197–203 (2011).
- Xu, L. *et al.* Tristetraprolin induces cell cycle arrest in breast tumor cells through targeting AP-1/c-Jun and NF- κ B pathway. *Oncotarget* **6**, 41679–41691 (2015).
- Huang, S., Heikal, A. A. & Webb, W. W. Two-photon fluorescence spectroscopy and microscopy of NAD(P)H and flavoprotein. *Biophys. J.* **82**, 2811–2825 (2002).
- Mujat, C. *et al.* Endogenous optical biomarkers of normal and human papillomavirus immortalized epithelial cells. *Int. J. Cancer* **122**, 363–371 (2008).
- Rodriguez, C. B. H., Vasudevan, A. & Elkhali, A. Aspects of tryptophan and nicotinamide adenine dinucleotide in immunity: A new twist in an old tale. *Int. J. Tryptophan Res.* **10**, 1–8 (2017).
- Chance, B., Cohen, P., Jobsis, F. & Schoener, B. Intracellular oxidation-reduction states in vivo. *Science* **137**, 499–508 (1962).
- Jolliffe, I. T. *Principal Component Analysis* (Springer, New York, 1986).
- Wu, B. *et al.* Statistical analysis and machine learning algorithms for optical biopsy. *Proc. SPIE* **10489**, 104890T (2018).
- Zhou, Y. *et al.* In *Neurophotonics and Biomedical Spectroscopy* (eds Alfano, R. R. & Shi, L.) 65–106 (Elsevier, Amsterdam, 2018).
- Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
- Pulukuri, S. M. *et al.* RNA interference-directed knockdown of urokinase plasminogen activator and urokinase plasminogen activator receptor inhibits prostate cancer cell invasion, survival and tumorigenicity in vivo. *J. Biol. Chem.* **280**, 36529–36540 (2005).
- Ravenna, L. *et al.* Distinct phenotypes of human prostate cancer cells associate with different adaptation to hypoxia and pro-inflammatory gene expression. *PLoS ONE* **9**, e96250 (2014).
- Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).

30. Wagnieres, G. A., Star, W. M. & Wilson, B. C. In vivo fluorescence spectroscopy and imaging for oncological applications. *Photochem. Photobiol.* **68**, 603–632 (1998).
31. Wu, B., Alrubaiee, M., Cai, W., Xu, M. & Gayen, S. K. Diffuse optical imaging using decomposition methods. *Int. J. Opt.* **2012**, 185435. <https://doi.org/10.1155/2012/185435> (2012).
32. Wu, B. & Gayen, S. K. Fluorescence tomography of targets in a turbid medium using non-negative matrix factorization. *Phys. Rev. E* **89**, 042708 (2014).
33. Rutkowski, T. M., Zdunek, R. & Cichocki, A. Multichannel EEG brain activity pattern analysis in time–frequency domain with nonnegative matrix factorization support. *Int. Congr. Ser.* **1301**, 266–269 (2007).
34. Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P. & Plemmons, R. J. Algorithms and applications for approximate nonnegative matrix factorization. *Comp. Stat. Data Anal.* **52**, 155–173 (2007).
35. Pu, Y. *et al.* Native fluorescence spectroscopic evaluation of chemotherapeutic effects on malignant cells using nonnegative matrix factorization analysis. *Technol. Cancer Res. Treat.* **10**, 113–120 (2011).
36. Wu, B., Nebylitsa, S. V., Mukherjee, S. & Jain, M. Quantitative diagnosis of bladder cancer by morphometric analysis of HE images. *Proc. SPIE* **9303**, 930317 (2015).
37. Jain, M., Robinson, B. D., Wu, B., Khani, F. & Mukherjee, S. Exploring multiphoton microscopy as a novel tool to differentiate chromophobe renal cell carcinoma from oncocytoma in fixed tissue sections. *Arch. Pathol. Lab. Med.* **142**, 383–390. <https://doi.org/10.5858/arpa.2017-0056-OA> (2018).
38. Bertini, I. *et al.* Metabolomic NMR fingerprinting to identify and predict survival of patients with metastatic colorectal cancer. *Cancer Res.* **72**, 356–364 (2012).
39. Kohavi, R. & John, G. H. Wrappers for feature subset selection. *Artif. Intell.* **97**, 273–324 (1997).
40. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **7**, 91 (2006).
41. Altman, D. G. & Bland, J. M. Diagnostic tests. 1: Sensitivity and specificity. *BMJ* **308**, 1552 (1994).
42. Metz, C. E. Basic principles of ROC analysis. *Semin. Nucl. Med.* **8**, 283–298 (1978).
43. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) Curve. *Radiology* **143**, 29–36 (1982).
44. Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* **27**, 861–874 (2006).
45. Liu, C.-H. *et al.* Optical Spectroscopic characteristics of lactate and mitochondrion as new biomarkers in cancer diagnosis: Understanding Warburg effect. *Proc. SPIE* **8220**, 82200Y (2012).
46. Ma, N., Digman, M. A., Malacrida, L. & Gratton, E. Measurements of absolute concentrations of NADH in cells using the phasor FLIM method. *Biomed. Opt. Express* **7**, 2441–2452 (2016).
47. Murthy, S. *et al.* Role of androgen receptor in progression of LNCaP prostate cancer cells from G1 to S phase. *PLoS ONE* **8**, e56692 (2013).
48. Baraga, J. J. *et al.* Laser induced fluorescence spectroscopy of normal and atherosclerotic human aorta using 306–310 nm excitation. *Lasers Surg. Med.* **10**, 245–261 (1990).
49. Liu, C.-H. *et al.* Resonance Raman of BCC and normal skin. *Proc. SPIE* **10060**, 100601B (2017).
50. Taniguchi, M. & Lindsey, J. S. Database of absorption and fluorescence spectra of >300 common compounds for use in PhotochemCAD. *Photochem. Photobiol.* **94**, 290–327 (2018).
51. Du, H., Fuh, R.-C.A., Li, J., Corkan, L. A. & Lindsey, J. S. PhotochemCAD: A computer-aided design and research tool in photochemistry. *Photochem. Photobiol.* **68**, 141–142 (1998).
52. Ghisaidoobe, A. B. T. & Chung, S. J. Intrinsic tryptophan fluorescence in the detection and analysis of proteins: A focus on Förster resonance energy transfer techniques. *Int. J. Mol. Sci.* **15**, 22518–22538 (2014).
53. Vivian, J. T. & Callis, P. R. Mechanisms of tryptophan fluorescence shifts in proteins. *Biophys. J.* **80**, 2093–2109 (2001).
54. Rehman, A. U. *et al.* Fluorescence quenching of free and bound NADH in HeLa cells determined by hyperspectral imaging and unmixing of cell autofluorescence. *Biomed. Opt. Express* **8**, 1488–1498 (2017).
55. Warburg, O. On respiratory impairment in cancer cells. *Science* **124**, 269–270 (1956).
56. Warburg, O. On the origin of cancer cells. *Science* **123**, 309–314 (1956).
57. Lopez-Lazaro, M. The Warburg effect: Why and how do cancer cells activate glycolysis in the presence of oxygen?. *Anticancer Agents Med. Chem.* **8**, 305–312 (2008).
58. De Bari, L., Chieppa, G., Marra, E. & Passarella, S. L-lactate metabolism can occur in normal and cancer prostate cells via novel mitochondrial L-lactate dehydrogenase. *Int. J. Oncol.* **37**, 1607–1620 (2010).
59. Alfano, R. R. & Pu, Y. In *Laser for Medical Applications* (ed. Jelinkova, H.) 325–367 (Woodhead Publishing Limited, New York, 2013).
60. Paldi, S. K. *et al.* Coarse Raman and optical diffraction tomographic imaging enable label-free phenotyping of isogenic breast cancer cells of varying metastatic potential. *Biosens. Bioelectron.* **2020**, 112863 (2020).
61. Bendau, E. *et al.* Distinguishing metastatic triple negative breast cancer from non-metastatic breast cancer using SHG imaging and resonance raman spectroscopy. *J. Biophotonics* **13**, e202000005. <https://doi.org/10.1002/jbio.202000005> (2020).

Acknowledgements

Graduate Research Fellowship at SCSU; Faculty Creative Activity Research Grant at SCSU; CSU-AAUP Research Grant.

Author contributions

J.X. and Y.P. designed and performed the experiments. J.S., X.G. and B.W. developed the algorithms and performed data analysis. J.X., Y.P., J.S. and B.W. wrote the manuscript. C.W. reviewed the manuscript. All authors contributed to and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021