Journal of Biomedical Optics

BiomedicalOptics.SPIEDigitalLibrary.org

Tissue multifractality and hidden Markov model based integrated framework for optimum precancer detection

Sabyasachi Mukhopadhyay
Nandan K. Das
Indrajit Kurmi
Asima Pradhan
Nirmalya Ghosh
Prasanta K. Panigrahi

SPIE.

Sabyasachi Mukhopadhyay, Nandan K. Das, Indrajit Kurmi, Asima Pradhan, Nirmalya Ghosh, Prasanta K. Panigrahi, "Tissue multifractality and hidden Markov model based integrated framework for optimum precancer detection," *J. Biomed. Opt.* **22**(10), 105005 (2017), doi: 10.1117/1.JBO.22.10.105005.

Tissue multifractality and hidden Markov model based integrated framework for optimum precancer detection

Sabyasachi Mukhopadhyay,[a] Nandan K. Das,[a,b] Indrajit Kurmi,[c] Asima Pradhan,[c,d] Nirmalya Ghosh,[a] and Prasanta K. Panigrahi[a,*]
[a]Indian Institute of Science Education and Research Kolkata, Mohanpur, West Bengal, India
[b]Nanyang Technological University, School of Chemical and Biomedical Engineering, Singapore
[c]Indian Institute of Technology Kanpur, Department of Physics, Kanpur, Uttar Pradesh, India
[d]Indian Institute of Technology Kanpur, Center for Lasers and Photonics, Kanpur, Uttar Pradesh, India

Abstract. We report the application of a hidden Markov model (HMM) on multifractal tissue optical properties derived via the Born approximation-based inverse light scattering method for effective discrimination of precancerous human cervical tissue sites from the normal ones. Two global fractal parameters, generalized Hurst exponent and the corresponding singularity spectrum width, computed by multifractal detrended fluctuation analysis (MFDFA), are used here as potential biomarkers. We develop a methodology that makes use of these multifractal parameters by integrating with different statistical classifiers like the HMM and support vector machine (SVM). It is shown that the MFDFA-HMM integrated model achieves significantly better discrimination between normal and different grades of cancer as compared to the MFDFA-SVM integrated model. [DOI: 10.1117/1.JBO.22.10.105005]

Keywords: inverse analysis on light scattering; tissue characterization; support vector machine; hidden Markov model; multifractal detrended fluctuation analysis.

Paper 170310RRR received May 12, 2017; accepted for publication Sep. 29, 2017; published online Oct. 19, 2017.

1 Introduction

Disease diagnosis through optical methods is an area of considerable research interest.[1] Optical tools are sensitive and are hence potentially capable of discriminating different stages of disease progression.[1–7] However, tissue being a complex medium, with several fluorophores, scatterers, and absorption domains, makes it difficult for proper diagnosis through optical means.[1] Hence, identifying reliable markers for accurately depicting the tissue condition through noninvasive optical methods has received significant attention.[1] For this purpose, recent approaches have focused on extracting intrinsic fluorescence[2] and tissue multifractality, characterizing the morphological changes by multifractal detrended fluctuation analysis (MFDFA).[3] Other approaches make use of principal component analysis[4] for identification of underlying spectral correlation and other image processing tools like wavelets[5] for pin pointing parameters that faithfully capture the disease progression. Clinical application of this approach, not only depends on these biomarkers but also crucially depends on the validation of the diagnosis outcome through a suitable diagnostic algorithm, which can accurately classify the measured spectra from an unknown tissue, using the stored database of spectra of tissues of known histopathologic classification. This will supplement and augment the histopathological approach, the current industry gold standard. Over the last few decades, a variety of diagnostic algorithms have been developed for optical diagnosis of cancer.[7] Classification schemes like artificial neural network[6] and support vector machine (SVM)[7] have been found promising in binary classification, e.g., normal versus cancer.

However, in a clinical situation, it is often required to classify the tissue site as normal and different grades of cancer. Hence, there is strong interest on statistical classifiers to extract the information content of the entire spectral data in order to get the best diagnostic features and enhance accuracy in classification of tissues into corresponding histopathologic categories.[8] Total principal component regression developed by Tan et al.[8] has classified various cancers, based on gene expression profiles and provided the optimized results compared to other methods for multiclass classifications. SVM has been deployed for multiclass cancer classification with classwise optimized gene.[9] Its usefulness for optical diagnosis is yet to be explored. In recent times, hidden Markov models (HMMs) are being widely used in biological sequence analysis as a robust method.[10] An HMM has been deployed to analyze hyperspectral images and a new HMM-based spectral measure has been referred to as the HMM information divergence in order to characterize the spectral properties.[11]

Here, we demonstrate the efficacy of MFDFA-HMM integrated framework for optical diagnosis of cancer. More specifically, it is found that HMM on the multifractal light scattering properties of the tissues shows remarkable efficiency in differentiating normal and different stages of precancer. It is particularly effective when applied on global fractal parameters like generalized Hurst exponent and the corresponding singularity spectrum width/strength of multifractality, characterizing the global morphological conditions of the tissues for multiclass cancer classification, as compared to the MFDFA-SVM integrated framework under same application.

*Address all correspondence to: Prasanta K. Panigrahi, E-mail: pprasanta@iiserkol.ac.in

1083-3668/2017/$25.00 © 2017 SPIE

2 Methods and Materials

2.1 *Sample Preparation*

Biopsied cervical precancer tissue slices (the histopathologically characterized grade I, grade II, and grade III precancer tissue) and normal tissues were collected from Ganesh Shankar Vidyarthi Memorial (GSVM) Medical College, Kanpur, India (age of patients between 35 and 60 years; $n_{\text{total}} = 35$, with $n_{\text{gradeI}} = 14$, $n_{\text{gradeII}} = 6$, $n_{\text{gradeIII}} = 9$; and six biopsies from the normal counterparts, $n_{\text{normal}} = 6$). The standardized histological preparation of the excised tissues involving fixation, dehydration, embedding in wax, sectioning under a rotary microtome with thickness $\sim 5\ \mu$m, and lateral dimension $\sim 4\ \text{mm} \times 6\ \text{mm}$, is followed by performing subsequent dewaxing. The consent for the use of all the intact tissue (human cervix with cancer and normal) samples in our study was obtained from the Ethical Committee, GSVM Medical College and Hospital, Kanpur, India. The sample preparation methods follow approved guidelines in our study.

2.2 *Experimental System*

The spatial distribution of tissue refractive index (RI) was recorded by a differential interference contrast (DIC) microscope (Olympus IX-81, United States). At a magnification of 60×, these DIC images were recorded by a CCD camera (ORCA-ERG, Hamamatsu, 1344×1024 pixel dimension 6.45 μm). The elastic scattering spectra from the multiple sites of the biopsied tissue sections were recorded by the angle resolved spectral light scattering measurements (Fig. 1). In brief, light emitted from a Xe-lamp (HPX-2000, Ocean Optics, United States) was collimated by a combination of lenses and illuminated the tissue sample at the center of a goniometric arrangement (spot size ~1-mm-diameter). The collimated scattered light from the sample was focused into a collecting fiber probe coupled to a spectrometer (USB4000FL, Ocean Optics) for wavelength resolved signal detection. The recordings of spectra were performed (360 to 800 nm) with a spectral resolution of 2.05 nm, where the angular range was kept at 10 deg to 150 deg with an interval of 10 deg. For the inverse multifractal study, the spectra were recorded at backscattering angle $\theta = 150$ deg (Fig. 3).

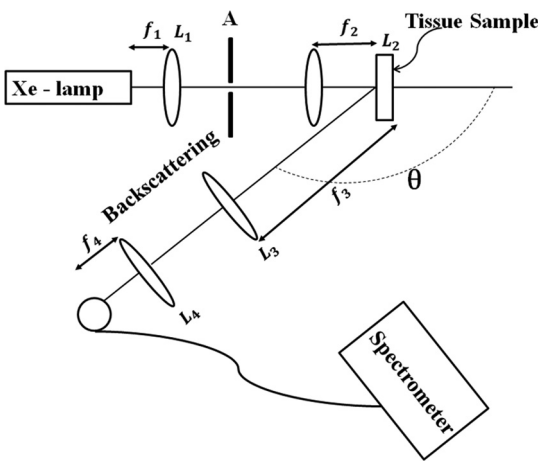

Fig. 1 Schematic of the spectral light scattering measurement. Xe lamp: excitation source; A: aperture; L_1: collimating lens; L_2: illuminating lens; L_3 and L_4: collecting lenses; f_1: focal length of collimating lens L_1; f_2: focal length of illuminating lens L_2; and f_3 and f_4: focal lengths of collecting lenses L_3 and L_4, respectively.

2.3 *Background Analysis*

In the first step, the elastic scattering spectra data were processed through Fourier domain preprocessing via the Born approximation, followed by the multifractal analysis. For analyzing the dataset, 6 normal, 14 grade-I, 6 grade-II, and 9 grade-III samples were taken.

2.3.1 *Light scattering-based inverse analysis in Born approximation*

At first, we extract the fractal parameters from the spatial variation of RI, as manifested through corresponding light scattering data.[12] We have followed the inverse analysis strategy[12] for quantification of multifractal signature from the scattered light signal through MFDFA, before applying an HMM on them. The normalized RI fluctuations of a weakly fluctuating scattering medium are given by $\Delta n(r) \sim \frac{n(r)-n_0}{n_0}$, where n_0 and r are the average RI of the medium and location within the volume, respectively. The fluctuation part $\Delta n(r)$ is responsible for phase distortion and scattering. It is known that the elastic scattering field for scalar excitation can be related to $\Delta n(r)$ in first-order Born approximation via Fourier transform.[12] For continuous random media, such as tissues, the expression for scattered intensity is given by

$$I(\beta) \approx k^4 \sigma^2 \left| \int \eta(r) \mathrm{e}^{i(\beta r)} d^3 r \right|^2, \tag{1}$$

where λ is the wavelength; $k = 2\pi/\lambda$; θ is the scattering angle; $\eta(r)$ is the spatial inhomogeneity distribution of index; and $\sigma = n_0 \delta n$ is the strength of fluctuation index, where the amplitude of fluctuating index is δn. The scattering vector $\boldsymbol{\beta}$ has the modulus $\beta = 2\text{k}\ \sin(\theta/2)$, where β is related to the spatial frequency ν via $\beta = 2\pi\nu$. For a medium exhibiting self-similarity in index fluctuations, the information on multifractality can be extracted from the scattering signal as follows:

$$\eta'(\rho) \approx \int k^{-2} \sqrt{I(\beta = 2\pi)} e^{-i(\boldsymbol{\beta} r)} d^3\beta. \tag{2}$$

The parameter $\eta'(\rho)$ exhibits the index inhomogeneity distribution with spatial scale $\rho = |r - r'|$, where ρ is the distance between any two points in the medium. It represents the randomness of the medium in statistical sense and embodies the essential multifractal features of index fluctuations in complex systems, such as tissues. This is subsequently analyzed using MFDFA[13] to yield the multifractal tissue optical properties.

2.3.2 *Multifractal detrended fluctuation analysis*

MFDA is an effective tool for quantitative estimation of the generalized Hurst exponents. In MFDFA, one first generates a profile function $Y(i)$ (spatial series of length N, $i = 1, \ldots, N$) from the one-dimensional spatial index fluctuations. Here, length of the series $N = 256 \times 256$ pixels. Subsequently, the profile is divided into $N_s = \text{int}(N/s)$ nonoverlapping segments b of equal segment length s. The segment length s varies from 16 to 128. The local trend of the series $[y_b(i)]$ is determined for each segment b by least square polynomial fitting and then subtracted

from the segmented profiles to yield detrended fluctuations. The resulting variance of the detrended fluctuation is determined for each segment as follows:

$$F^2(b,s) = \frac{1}{s}\sum_{i=1}^{s}\{Y[(b-1)s+i] - y_b(i)\}^2. \quad (3)$$

The moment (q) dependent fluctuation function can be obtained by taking the average over all the segments:

$$F_q(s) = \left\{\frac{1}{2N_s}\sum_{b=1}^{2N_s}\left[F^2(b,s)\right]^{\frac{q}{2}}\right\}^{\frac{1}{q}}, \quad (4)$$

where q may vary from negative to positive values with fraction or integer values. Since length N of the series is often not an integer multiple of variable segment length s, in order to take into account all the data points, the procedure was performed twice on the series, starting from either end of the series.[13] The scaling behavior can be obtained by analyzing the variations of $F_q(s)$ versus s for each value q, assuming the general scaling function as follows:

$$\mathrm{F_q(s)} \sim \mathrm{s}^{h(q)}. \quad (5)$$

The relation between the generalized Hurst exponent $h(q)$ and the multifractal scaling exponent $\tau(q)$ can be demonstrated as follows:

$$\tau(q) = qh(q) - 1. \quad (6)$$

The Hurst exponent (H) is defined as $h(q=2)$ and the values $H > 0.5$, $H = 0.5$, and $H < 0.5$ correspond to long range correlation, uncorrelated random fluctuations, and anticorrelated behavior, respectively.[11] The Hurst exponent $h(q)$ and the scaling exponent $\tau(q)$, along with the singularity spectrum $f(\alpha)$ completely, characterize any nonstationary multifractal fluctuation series. Here, $f(\alpha)$ is related to $\tau(q)$ via a Legendre transformation:

$$\alpha = \frac{d\tau}{dq}, \quad f(\alpha) = q\alpha - \tau(q), \quad (7)$$

where α is the singularity strength and the width ($\Delta\alpha = |\alpha_1 - \alpha_2|$) [considered at $f(\alpha) = 0$] of $f(\alpha)$ is a quantitative measure of multifractality.[13]

2.4 *Data Analysis*

The obtained two global fractal parameters, generalized Hurst exponent and the corresponding singularity spectrum width, were subjected to (a) SVM-based multiclass classification and (b) HMM-based multiclass classification. For analyzing the dataset, 6 normal, 14 grade-I, 6 grade-II, and 9 grade-III samples were taken.

2.4.1 *Support vector machine*

SVMs are powerful statistical classifiers under the supervised learning scheme. The central idea behind SVM operation is to separate classes with a surface that maximizes margin between them by avoiding overfitting to form an optimal separating hyperplane (OSH). Hence, by following structural risk minimization (SRM) of statistical learning makes prediction on a function $f(x)$ as $f(x) = \sum_{i=1}^{N} w_i k(x, x_i) + w_0$, where $k(x, x_i)$ is the kernel function defined on a basis function, $\{w_i\}$ is the corresponding model weights, and w_0 is the bias weight.

The training data points lie far away from the OSH, does not participate in the specification and hence receives zero weight. Data point that lies close to decision boundary receives nonzero weights. These training data points are "support vectors."[14,15] If we remove these points, it will change the boundary location. Unlike relevance vector machine, there are restrictions while choosing of kernels in SVM. An appropriate selection of kernel function is an important aspect as it defines the accuracy level of SVM-based operation while determining training data classification. The kernel function will produce optimum results in classification as long as it obeys the Mercer's theorem.[14,15] Figure 2 displays the simplified workflow of SVM-based multiclass classification on extracted multifractal parameters from tissue samples.

2.4.2 *Hidden Markov model*

HMM[16] is a statistical Markov model with hidden states and is also the simplest dynamic Bayesian network. An HMM is closely related with mixture models, which are statistically independent. An HMM can be efficiently employed for a time series data, where actual parameters are unknown and only observational information are known. From this series of observations, the probabilities of parameters giving such observations and the transmission probabilities can be found by the Baum–Welch algorithm.[17] The basic principle of an HMM can be described as follows. For $s_0, \ldots, s_t$ states as input to the model, the s_{t+1}'th state can be predicted by the traditional Markov model, where given the present input, the future is independent of the past:

$$p(s_{t+1}|s_t, \ldots, s_0) = p(s_{t+1}|s_t). \quad (8)$$

If q_n'th observation can be made on the basis of q_{n-1}'th observation, it is a first-order assumption, which generally is used for Bayesian modeling. Using the Markov assumption, we also can write it as follows:

$$p(q_{1,\ldots,}q_n) = \prod_{i=1}^{n} p(q_i|q_{i-1}). \quad (9)$$

According to the Bayes' formula:

$$p(q_i/x_i) = \frac{p(x_i/q_i)p(q_i)}{p(x_i)}. \quad (10)$$

In a more general way, the above Eq. (8) can be rewritten as follows:

$$p\left(\frac{q_{1,\ldots,}q_n}{x_{1,\ldots,}x_n}\right) = \frac{p\left(\frac{x_{1,\ldots,}x_n}{q_{1,\ldots,}q_n}\right)p(q_{1,\ldots,}q_n)}{p(x_{1,\ldots,}x_n)}. \quad (11)$$

The measure of probability can be achieved by likelihood parameter L, which is proportional to the probability:

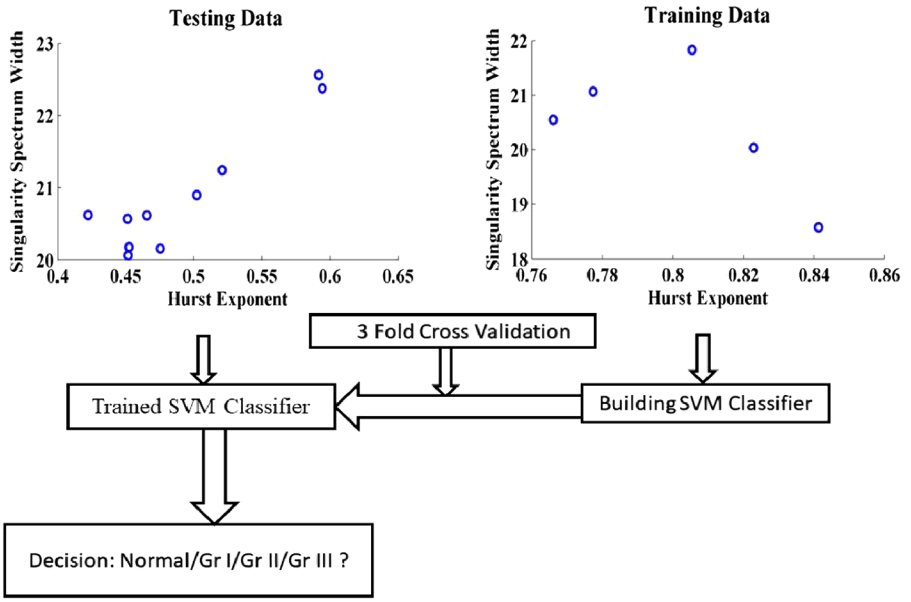

Fig. 2 SVM-based multiclass classification on extracted multifractal parameters from tissue samples.

$$p\left(\frac{q_{1,\ldots,}q_n}{x_{1,\ldots,}x_n}\right)\infty L\left(\frac{q_{1,\ldots,}q_n}{x_{1,\ldots,}x_n}\right)=p\left(\frac{x_{1,\ldots,}x_n}{q_{1,\ldots,}q_n}\right)p(q_{1,\ldots,}q_n). \tag{12}$$

In the case of an HMM, the model is prepared with the training data $\Theta=\{\pi,A,B\}$ and a sequence of N states is $S=\{s_{1,\ldots,}s_N\}$. Here, π is defined as the prior probability, A is the transition probability, and B is the emission probability. The probability of a state sequence $Q=\{q_1,\ldots,q_n\}$ obtained from an HMM with parameters Θ can be expressed as a product of the transition probabilities:

$$p\left(\frac{Q}{\Theta}\right)=\pi_{q_1}\prod_{n=1}^{N-1}a_{q_n,q_{n+1}}=\pi_{q_1}a_{q_1,q_2}a_{q_2,q_3}\ldots a_{q_{N-1},q_N}. \tag{13}$$

For an observational sequence $X=\{x_1,\ldots,x_N\}$, a (hidden) state sequence $Q=\{q_1,\ldots,q_N\}$ can be determined from an HMM with parameter Θ. Hence, the likelihood of X along the path Q takes the form:

$$p\left(\frac{X}{Q},\Theta\right)=\prod_{n=1}^{N}p\left(\frac{x_n}{q_n},\Theta\right)=b_{q_1,x_1}.b_{q_2,x_2}\ldots b_{q_N,x_N}. \tag{14}$$

It can be expressed as the product of the emission probabilities computed along the considered path.

With the likelihood of an observed sequence $X=\{x_1,\ldots,x_N\}$ and the parameter Θ defined by an HMM, the probability p can be expanded as follows:

$$p\left(\frac{X}{\Theta}\right)=\sum p\left(X,\frac{Q}{\Theta}\right). \tag{15}$$

Using the Baum–Welch algorithm, the hidden parameters in an HMM are found. This algorithm utilizes the expectation maximization (EM) algorithm for finding the maximum likelihood estimation of the parameters of an HMM, given a set of observed feature vectors.

As we mentioned earlier, hidden Markov chain can be represented as $\theta=(A,B,\pi)$. Here, the stochastic transition matrix $A=\{a_{ij}\}=p(X_t=j|X_{t-1}=i)$, where X_t is the discrete variable. The emission probability $B=[b_j(y_t)]=p(Y_t=y_t|X_t=j)$, where Y_t is the observational sequence. The EM algorithm defines a local maximum for $\theta^*=\text{argmax}[p(Y/X)]$. After defining the initial condition, the Baum–Welch algorithm follows the forward and backward procedure to find the proper estimation of the predicted results.

In the case of forward procedure, let the probability of viewing $y_1,\ldots,y_t$ at state i in time t is $\alpha_i(t)=p(Y_1=y_1,\ldots,Y_t=y_t|X_t=i,\theta)$.

Under recursion procedure, $\alpha_i(1)=\pi_i b_i(y_1)$ and $\alpha_j(t+1)=b_j(y_{t+1})\sum_{i=1}^{N}\alpha_i(t)a_{ij}$.

Similarly, let the probability of viewing $y_{t+1},\ldots,y_T$ at state i in time t is $\beta_i(t)=p(Y_{t+1}=y_{t+1,\ldots,}Y_T=y_T|X_t=i,\theta)$. Under recursion procedure, $\beta_i(T)=l$ and $\beta_i(t)=\sum_{j=1}^{N}\beta_j(t+1)a_{ij}b_j(y_{t+1})$.

In the final step, according to the Bayes' theorem, the probability of the observed sequence Y and the parameters θ in state i at time t given as $\gamma_i(t)=p(X_t=i|Y,\theta)=\frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N}\alpha_j(t)\beta_j(t)}$.

The probability of being in state i and j at times t and $t+1$, respectively, given the observed sequence Y and parameters θ:

$$\varepsilon_{ij}(t)=p(X_t=i,X_{t+1}=j|Y,\theta)=\frac{\alpha_i(t)a_{ij}\beta_j(t+1)b_j(y_{t+1})}{\sum_{k=1}^{N}\alpha_k(T)}.$$

Hence, θ can be updated at expected frequency spent in state i at time l as $\pi_i^*=\gamma_i(l)$. The expected number of transitions from state i to state j compared to the expected total number of transitions away from state i is $a_{ij}^*=\frac{\sum_{t=1}^{T-1}\varepsilon_{ij}(t)}{\sum_{t=1}^{T-1}\gamma_i(t)}$.

If $b_i^*(v_k)$ is the expected number of times, the output observations have been equal to v_k while in state i over the expected

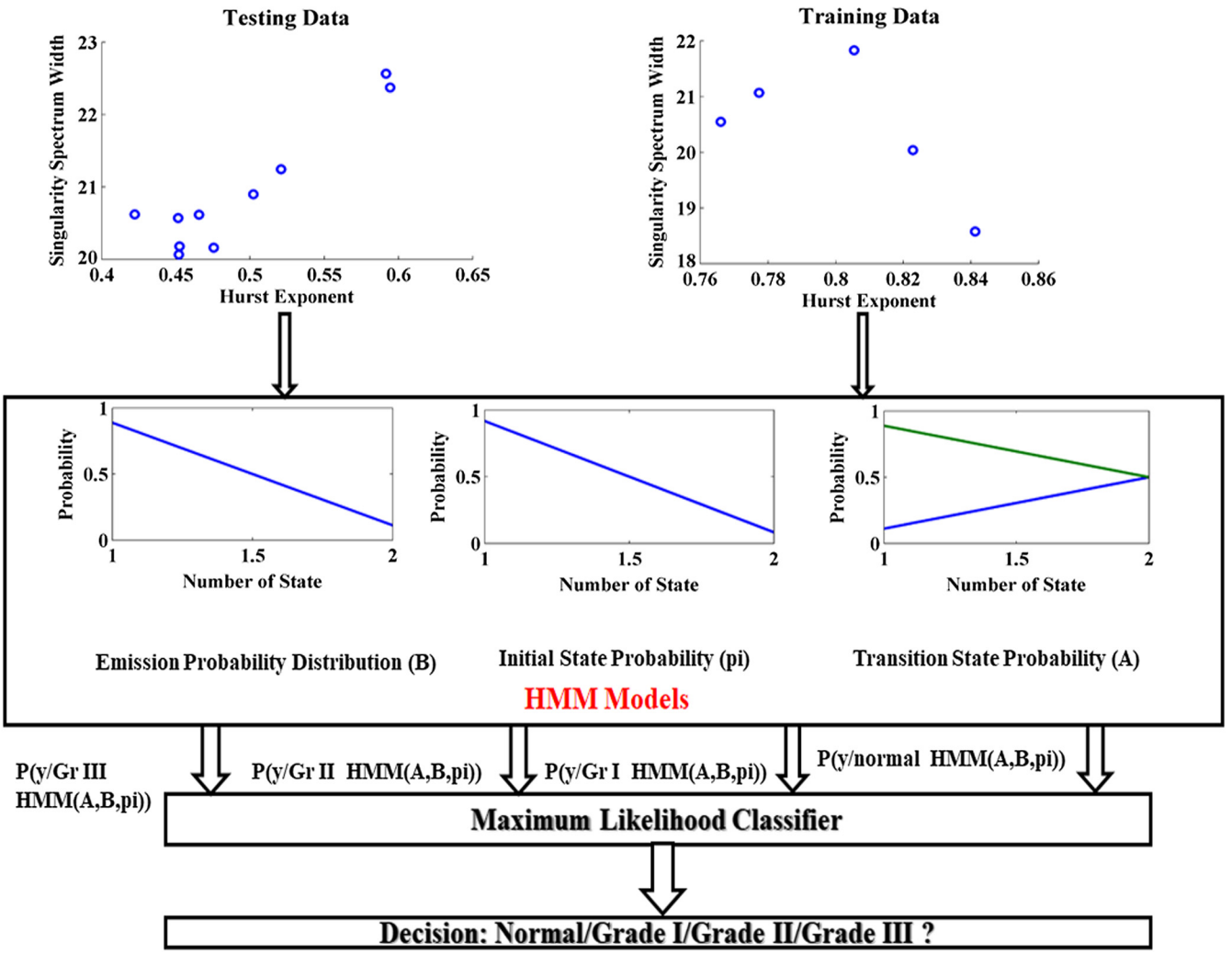

Fig. 3 The flowchart of an HMM-based model on multifractal tissue optical properties derived from light scattering spectra. First row depicts the singularity spectrum versus generalized Hurst exponent for testing and training datasets. Second row shows the processing of test data and training data through emission probability density, initial state probability, and transition probability (model trained using the training data) has been performed on the normal and different grades of cancer, respectively. The output has been processed by the maximum likelihood classifier method to perform the HMM-based classification.

total number of times in state i, then $b_i^*(v_k) = \frac{\sum_{t=1}^{T} l_{y_t=v_k}\gamma_i(t)}{\sum_{t=1}^{T}\gamma_i(t)}$. Here, the indicator function $l_{y_t=v_k} = 1$ exists only for $y_t = v_k$.

These steps are iterated until a desired level of convergence is achieved.

In this paper, an HMM is applied on the known multifractal fractal parameters, which leads to a significant improvement on the prediction accuracy. For experimentation, we first train the multifractal parameters in an HMM for each of the categories. Prior probabilities are first selected as a random function. A and B are modeled as Gaussian densities, with mean 0 and variance 1. Subsequently, a representative data is trained on the model iteratively to fit and modify the model using EM algorithm. The model is optimized using Lagrange multipliers. We use forward and backward algorithms to compute a set of sufficient statistics for our EM step tractably. Once the model is sufficiently trained for a given sequence of data we calculate the likelihood of sequence for each category, i.e., as $P(X/\theta_i)$, when the sum of the joint likelihoods of the sequence over all possible state sequences Q, allowed by the model for each category. The maximum likelihood gives the prediction for the sequence data.

The proposed HMM-based model on light scattering derived from multifractal tissue optical properties has been demonstrated in Fig. 3. The HMM-based data analysis steps for normal and different cancerous grades have been shown here in detail for the ease of understanding.

3 Results and Discussion

The DIC images of different pathological grades have been presented in Fig. 4 for comparisons. The histopathologically characterized tissue samples were provided by the pathologists of GSVM Medical College and Hospital, Kanpur, where cancer stages were defined by the pathologists.

The results of the inverse analysis on the light scattering spectra recorded (using the spectral light scattering measurement system in Fig. 1) from a grade-1 dysplastic cervical tissue (corresponding to Fig. 1) are displayed in Fig. 5. The large variation of the slopes of log $F_q(s)$ versus log s [Fig. 5(c)] is the

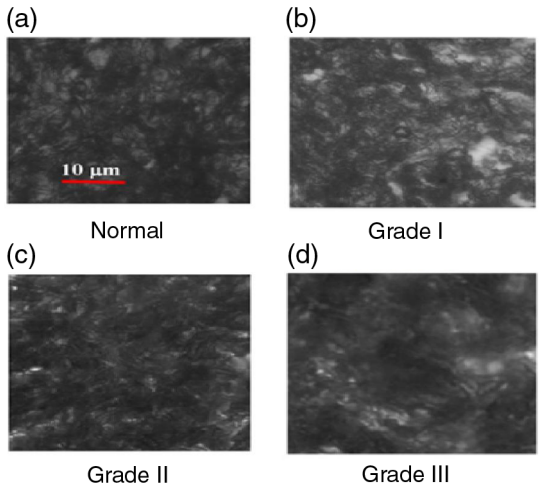

Fig. 4 The representative DIC images of (a) healthy (normal) and (b–d) different stages (histopathologically characterized grade I, II, III) of precancer tissues, respectively.

evidence of strong multifractality. The derived $h(q)$ spectrum and the singularity spectrum $f(\alpha)$ [Figs. 5(d) and 5(e)] demonstrate the strong multifractality in the spatial variations of tissue refractive indices. The parameter $\eta'(\rho)$, obtained by Fourier preprocessing, contains submicron level spatial index fluctuations information [as evident from Fig. 5(b)]. The observed small-scale fluctuations of $h(q)$ [Fig. 5(d)] and the resulting width of $f(\alpha)$ [Fig. 5(e)] possibly originate from the microarchitecture of the fibrous network of connective tissue. To summarize Fourier preprocessing of light scattering in Born approximation and its subsequent analysis through MFDFA documents, there are small spectral variations as signature of subtle or hidden changes in the refractive indices spatial distribution via multifractal parameters.

After MFDFA analysis, the observed trends are $h(q=2) = 0.63 \pm 0.02$, 0.56 ± 0.05, 0.48 ± 0.03, 0.41 ± 0.04, and $\Delta\alpha = 0.86 \pm 0.01$, 0.90 ± 0.03, 0.96 ± 0.04, 0.99 ± 0.01 for normal and different grades of cancer, respectively. Here, it can be observed that there exists an overlapping of multifractal parameters like Hurst exponent and singularity spectrum width among normal and different grade of precancerous tissues. Hence, the supervised classifiers like SVM and HMM have been applied here for multiclass classification purpose.

Our initial dataset consists of 35 samples. We have used Monte Carlo cross-validation, where we randomly split the dataset into training and testing dataset. This process has been repeated 100 times. The size of training and testing dataset varies for each split; we just ensured that a minimum two number of samples have been included in both training and testing dataset for each split. For each such split, the model has been fit to the training data, and predictive accuracy has been assessed using just the validation data (testing data). Then, the results of the testing data have been averaged over the splits. The advantage of this method is that the proportion of the training/validation split is not dependent on the number of iterations, and the

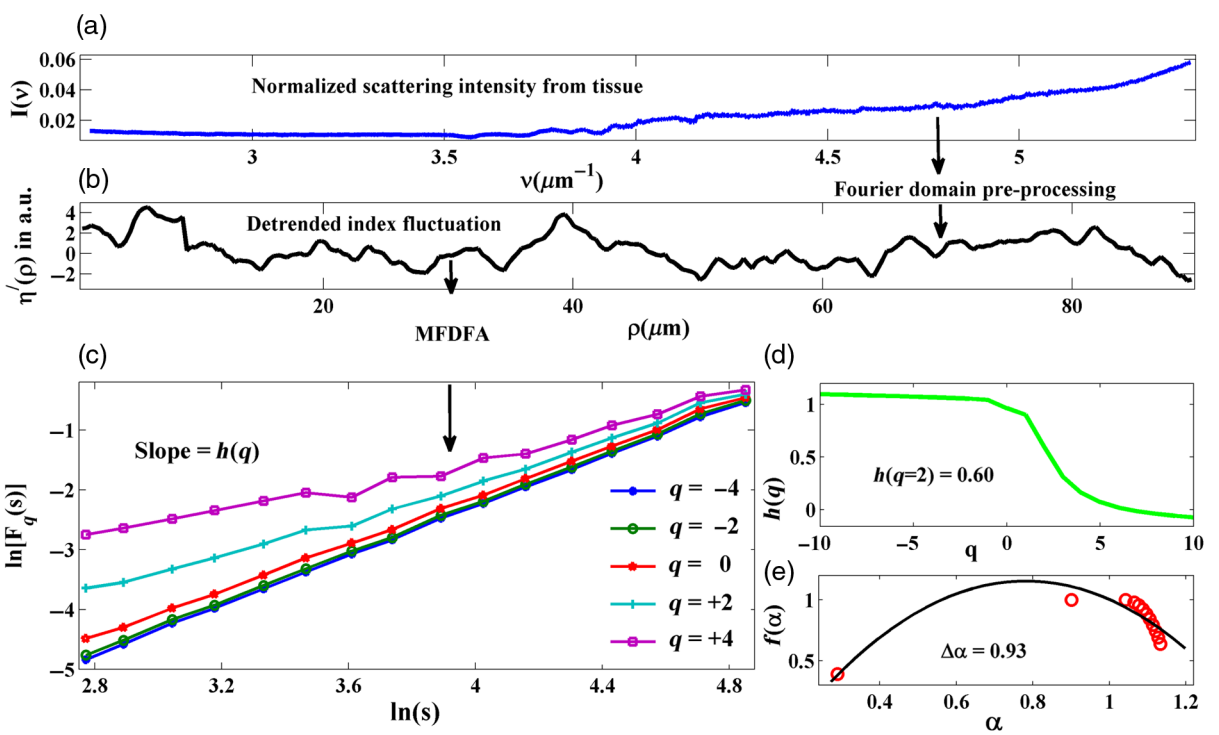

Fig. 5 Results of the inverse analysis performed on the light scattering spectra recorded from a grade-1 precancer human cervical tissue slice (corresponding to Fig. 1). (a) The recorded light scattering intensity [I(ν) versus ν, $\nu = \frac{2}{\lambda}\sin(\frac{\theta}{2})$, $\theta = 150$ deg, $\lambda = 360$ to 740 nm, shown here]. (b) The representative index fluctuations with spatial scale ρ [$\eta'(\rho)$] extracted via Fourier domain preprocessing [using Eq. (2)] on $I(\nu)$ (shown following polynomial detrending). (c)–(e) Results of the MFDFA inverse analysis on the detrended fluctuations $\eta'(\rho)$. (c) The variation of log $F_q(s)$ versus log s for different moments q ($= -4$ to $+4$ shown here). Multifractality in $\eta'(\rho)$ is evident from significant variations in the slopes with varying q. (d) The MFDFA-derived moment dependence of generalized Hurst exponent $h(q)$. (e) The resulting singularity spectrum $f(\alpha)$.

predictive accuracy is more or less independent of the samples used for training dataset. We additionally used nine unknown samples taken at different time than the dataset and the prediction done by our model has been compared by manual verification. Results of the unknown samples also have been averaged and presented in the results. The above process has been repeated for SVM and HMM.

The data in Table 1 depicts the mean and variance (mean ± variance) for each MFDA parameter of each grade for the entire dataset. The same dataset has been used for model generation in SVM and HMM.

There are nine unknown samples, which are considered as test samples for SVM and HMM classification purpose.

SVM creates an optimum manifold barrier with radial basis function kernel between healthy and different grades of cancer depending on MFDFA parameters. Figure 6 displays the prediction analysis carried out by training data in SVM.

From Fig. 6, it is clearly visible that SVM works very well in binary classification, i.e., between normal and grade III with 98.5% and 100% accuracy, respectively. While trying out SVM classification for multiclass classification, the MFDFA-SVM integrated framework performs poorly by degrading the overall performance of the system with 57.14% and 55% accuracy, respectively. This error in prediction has occurred due to wrongly predicting normal tissues as grade I tissues (1.5%), grade I as normal (14.29%), and grade II tissues (28.57%) as well as grade II tissues as normal (10%) and grade I tissues (35%).

Table 1 Summary of the multifractal tissue optical properties derived from light scattering spectra.

	Normal	Grade-I	Grade-II	Grade-III
Hurst exponent [$h(q=2)$]	0.63 ± 0.02	0.56 ± 0.05	0.48 ± 0.03	0.41 ± 0.04
Singularity spectrum width ($\Delta\alpha$)	0.86 ± 0.01	0.90 ± 0.03	0.96 ± 0.04	0.99 ± 0.01

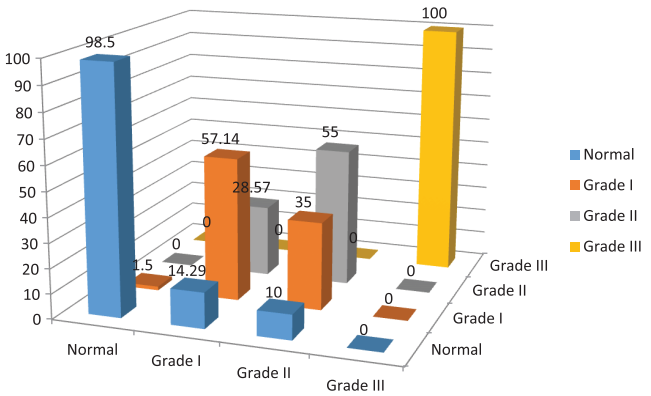

Fig. 6 SVM-based tissue classification based on light scattering-derived multifractal tissue optical properties. The horizontal surfaces of these figures have been sectioned into 4 × 4 rectangles. The accurate and inaccurate prediction of each stage (normal, grade I, grade II, and grade III) have been represented by diagonal and off diagonal rectangles, respectively.

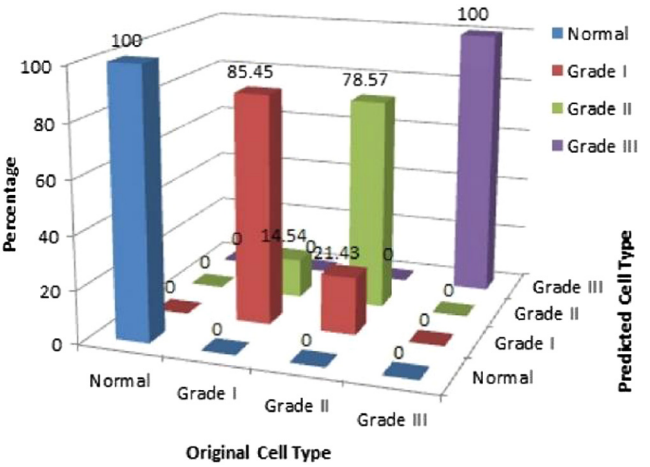

Fig. 7 HMM-based tissue classification based on light scattering – derived multifractal tissue optical properties. The horizontal surfaces of these figures have been sectioned into 4 × 4 rectangles. The accurate and inaccurate prediction of each stage (normal, grade I, grade II, and grade III) have been represented by diagonal and off diagonal rectangles, respectively.

An HMM creates abstract Markov models for the classification purpose on multifractal parameters. Figure 7 displays the prediction analysis carried out by training data in an HMM.

As can be seen from the graph, the parameters clearly show distinction between normal and different grades of precancer with singularity spectrum width ($\Delta\alpha$) and Hurst exponent $h(q)$. The normal and grade III tissues were correctly predicted. However, there is error while predicting grade I and II tissues (correct prediction rates are 85.45% and 78.57%, respectively). This error in prediction is also limited to wrongly predicting grade I tissues as grade II tissues (14.54%) as well as grade II tissues as grade I tissues (21.43%).

The results demonstrate that binary classification between normal and cancerous tissues (grade III) is very good both in SVM and HMM. Meanwhile, in multiclass classification cases, when precancerous grades (grade I, grade II) are to be classified along with normal and precancerous tissues (grade III), abstract parameters achieved using an HMM that performs better than the SVM. The presence of noise in the obtained signal damages the SVM performance as SVM clearly classifies based on the kernel formed after considering all the multifractal parameters. While in the case of HMM, the Markov model finds abstract parameters by controlling the actual multifractal parameters and produces a prediction based on the derived abstract parameters. As a consequence, an HMM avoids the noise added to the signal and able to produce better multiclass classification results than SVM.

In our current manuscript, we proceeded with Fourier domain preprocessing as our main focus was on global behavior analysis of time series through MFDFA-HMM and MFDFA-SVM integrated model classifications in early-stage cancer detection. It is known to us that wavelet domain analysis is more appreciable than Fourier domain (short-time Fourier transform) analysis in the case of local behavior analysis of time series.[18] In wavelet domain preprocessing, the performance of the SVM model can perform better than an HMM model as SVM has better generalization due to the principle of SRM than an HMM for system abnormality detection.[19] A comparative study between MFDFA-SVM and MFDFA-HMM integrated

model for local behavior analysis of time series through wavelet domain preprocessing in early-stage cancer diagnosis will be a part of our future study.

4 Conclusions

We have explored an integrated framework of light scattering-derived multifractal tissue optical properties (generalized Hurst exponent and width of singularity spectrum) along with a robust HMM for multiclass classification of different precancerous grades of human uterine cervix. The results clearly demonstrate that the use of HMM on the multifractal properties leads to significantly improved classification as compared to MFDFA-SVM based integrated model for multiclass classification. These MFDFA-HMM based classification results show considerable promise by exploring multifractal tissue optical properties as a biomarker for precancer detection. We are currently expanding our investigations toward *in-vivo* deployment of this integrated approach for precancer detection using tissue light scattering spectra. In general, the use of this MFDFA-HMM integrated model on elastic scattering spectroscopic data may lead to a diagnostic modality for the detection of other types of cancer.

Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors.

Acknowledgments

The authors thank Dr. Asha Agarwal, GSVM Medical College and Hospital, Kanpur, for providing the histopathologically characterized tissue samples.

References

1. P. N. Prasad, *Introduction to Biophotonics*, Wiley-Interscience, Hoboken, New Jersey (2003).
2. K. Pandey et al., "Fluorescence spectroscopy: a new approach in cervical cancer," *J. Obstet. Gynaecol. India* **62**(4), 432–436 (2012).
3. N. Das et al., "Probing multifractality in tissue refractive index: prospects for precancer detection," *Opt. Lett.* **38**, 211–213 (2013).
4. S. Devi, P. K. Panigrahi, and A. Pradhan, "Detecting cervical cancer progression through extracted intrinsic fluorescence and principal component analysis," *J. Biomed. Opt.* **19**, 127003 (2014).
5. A. H. Gharekhan et al., "Distinguishing cancer and normal breast tissue autofluorescence using continuous wavelet transform," *IEEE J. Sel. Top. Quantum Electron.* **16**, 893–899 (2010).
6. G. A. Rovithakis et al., "Artificial neural networks for discriminating pathologic from normal peripheral vascular tissue," *IEEE Trans. Biomed. Eng.* **48**(10), 1088–1097 (2001).
7. S. K. Majumder, N. Ghosh, and P. K. Gupta, "Support vector machine for optical diagnosis of cancer," *J. Biomed. Opt.* **10**(2), 024034 (2005).
8. Y. Tan et al., "Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data," *Nucleic Acids Res.* **33**(1), 56–65 (2005).
9. A. Anand and P. N. Suganthan, "Multiclass cancer classification by support vector machines with class-wise optimized genes and probability estimates," *J. Theor. Biol.* **259**(3), 533–540 (2009).
10. B. Yoon, "Hidden Markov models and their applications in biological sequence analysis," *Curr. Genomics* **10**, 402–415 (2009).
11. Q. Du and C. I. Chang, "Hidden Markov model approach to spectral analysis for hyperspectral imagery," *Opt. Eng.* **40**(10), 2277–2284 (2001).
12. N. Das et al., "Tissue multifractality and Born approximation in analysis of light scattering: a novel approach for precancers detection," *Sci. Rep.* **4**, 6129 (2014).
13. J. W. Kantelhardt et al., "Multifractal detrended fluctuation analysis of nonstationary time series," *Phys. A* **316**, 87–114 (2002).
14. V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York (1998).
15. C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discovery* **2**(2), 121–167 (1998).
16. L. E. Baum and T. Perie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Stat.* **37**(6), 1554–1563 (1966).
17. R. Hogg, J. McKean, and A. Craig, *Introduction to Mathematical Statistics*, 7th ed., Pearson Prentice Hall, Boston (2005).
18. F. Auger and P. Flandrin, "Improving the readability of time–frequency and time-scale representations by the reassignment method," *IEEE Trans. Signal Process.* **43**(5), 1068–1089 (1995).
19. Q. Miao, H. Huang, and X. Fang "A comparison study of support vector machines and hidden Markov models in machinery condition monitoring," *J. Mech. Sci. Technol.* **21**, 607–615 (2007).

Biographies for the authors are not available.