Image Classification of Human Carcinoma Cells Using Complex Wavelet-Based Covariance Descriptors

Furkan Keskin¹, Alexander Suhre¹, Kivanc Kose¹, Tulin Ersahin², A. Enis Cetin¹, Rengul Cetin-Atalay²

¹ Electrical and Electronics Engineering Department, Bilkent University, Ankara, Turkey, ² Department of Molecular Biology and Genetics, Bilkent University, Ankara, Turkey

Abstract

Cancer cell lines are widely used for research purposes in laboratories all over the world. Computer-assisted classification of cancer cells can alleviate the burden of manual labeling and help cancer research. In this paper, we present a novel computerized method for cancer cell line image classification. The aim is to automatically classify 14 different classes of cell lines including 7 classes of breast and 7 classes of liver cancer cells. Microscopic images containing irregular carcinoma cell patterns are represented by subwindows which correspond to foreground pixels. For each subwindow, a covariance descriptor utilizing the dual-tree complex wavelet transform (DT-CWT) coefficients and several morphological attributes are computed. Directionally selective DT-CWT feature parameters are preferred primarily because of their ability to characterize edges at multiple orientations which is the characteristic feature of carcinoma cell line images. A Support Vector Machine (SVM) classifier with radial basis function (RBF) kernel is employed for final classification. Over a dataset of 840 images, we achieve an accuracy above 98%, which outperforms the classical covariance-based methods. The proposed system can be used as a reliable decision maker for laboratory studies. Our tool provides an automated, time- and cost-efficient analysis of cancer cell morphology to classify different cancer cell lines using image-processing techniques, which can be used as an alternative to the costly short tandem repeat (STR) analysis. The data set used in this manuscript is available as supplementary material through http://signal.ee.bilkent.edu.tr/cancerCellLineClassificationSampleImages.html.

Introduction

Automatic classification of biomedical images is an emerging field, despite the fact that there is a long history of image recognition techniques [1]. Automated classification of carcinoma cells through morphological analysis will greatly improve and speed up cancer research conducted using established cancer cell lines as in vitro models. Distinct morphologies of different types and even sub-types of cancer cells reflect, at least in part, the underlying biochemical differences, i.e., gene expression profiles. Moreover, the morphology of cancer cells can infer invasiveness of tumor cell and hence the metastatic capability. The change in morphologies upon treatment with agents that induce cellular responses such as cell death or cell growth arrest [2]. Table 1 shows a summary of the different morphologies for the cancer cell lines in the dataset. In addition, an automated morphological classification of cancer cells will enable the correct detection and labelling of different cell lines. In molecular biology studies, experimenters deal with a large number of specimens whose identity have to be checked recurring during different stages of the experiment. Therefore, predicting labels of cancer cell lines in a fast and accurate manner via a pattern classification approach will greatly enhance biologists’ ability to identify different types of cell lines without the need to scrutinize each and every microscopic image one by one. Although cell lines are being used widely as in vitro models in cancer research and drug development, mislabeling cell lines or failure to recognize any contamination may lead to misleading results. Short tandem repeat (STR) analysis is being used as a standard for the authentication of human cell lines. However, this process takes a long time and has to be carried out by an expert. Automated analysis, on the other hand, will provide the scientists a fast and easy-to-use tool that they can use in their own laboratories to verify their cell lines.

Modelling of cell morphology has been studied by several groups, for example for fission yeast in [3] and for e. coli bacteria in [4]. In the fission yeast case, differential expression of protein affects the cell size and, therefore, cell fate, while in the e. coli case, the topological organization is analyzed with respect to the underlying signaling network. To the best of our knowledge there have been no studies that have used morphology of different human cancer cell lines for classification.

Feature parameters are computed using the dual-tree complex wavelet transform (DT-CWT). In addition, directional difference scores and covariance descriptors are deployed in support vector machines (SVM) for analysis and classification of carcinoma cell line images. Detailed descriptions of these methods can be found in the feature extraction and classification sections; below we perform a literature search on how these techniques are applied in the medical domain. DT-CWT is a recently developed image decomposition method that possesses orientation selectivity and shift invariance properties lacking in the classical discrete wavelet.
In the biomedical image analysis literature, DT-WT is used to predict the histological diagnosis of colorectal lesions in colonoscopy images by employing a probabilistic framework where a joint statistical model for complex wavelet coefficient magnitudes is proposed [5]. In [6], authors model the marginal distributions of DT-WT coefficient magnitudes by Rayleigh and Weibull probability density functions to classify the zoom-endoscopy images for colorectal cancer diagnosis. In [7], MR images of human brain and wrist are classified using textural features extracted via DT-WT decomposition. Directional difference scores are first introduced in this article and applied to our classification problem. Normalized versions of covariance descriptor, which is a matrix-form feature describing an image region are used. In the medical domain, covariance descriptors are utilized for classification of colonic polyps in CT colonography images [8]. Our study is one of the first studies to apply the covariance descriptors to medical image analysis domain. SVM is a well-known machine learning algorithm that learns the decision boundaries between classes using separating hyperplanes. SVM is used in [9] for automated prostate cancer grading on histology images. In [10], a segmentation framework for cell microscopic images is proposed that adopts segmentation-by-classification approach and uses SVM for pixel classification. In [11], computer-aided classification of renal cell carcinoma subtypes is performed by using SVM. A fully automated system is presented for human cell phenotype monitoring in [12] and subcellular phenotypes on human cell arrays are automatically classified via SVM.

In this study, discrimination of 14 classes of biomedical images is achieved, which are all images of cancer cell lines. The dataset at hand consists of two major types of cancer cell lines, namely breast cancer and liver cancer (hepatocellular carcinoma) with 7 sub-classes, respectively. The dataset consists of 840 images, i.e., 60 per sub-class. Our approach aims to carry out the automated analysis by extracting a feature vector from the images. These feature parameters reflect the large morphological diversity of the images. Notice, however, that our software learns the specific covariances of these features from the training set, so the model for each image class is not rigid and therefore allows for larger variation in the image data, while maintaining its high effectivity.

| Cell Line     | Shape         | Shape          | Growth properties | Source          | Classification | Disease        |
|---------------|---------------|----------------|-------------------|-----------------|----------------|----------------|
| BT-20         | epithelioid   | stellate       | adherent          | mammary gland   | Basal A        | Adenocarcinoma |
| CAMA-1        | epithelioid   | grape-like     | adherent          | mammary gland   | Luminal        | Adenocarcinoma |
| MDA-MB-157    | epithelioid   | stellate       | adherent          | mammary gland   | Basal B        | Medullary carcinoma |
| MDA-MB-361    | epithelioid   | grape-like     | adherent          | mammary gland   | Luminal        | Metastatic adenocarcinoma |
| MDA-MB-453    | epithelioid   | grape-like     | adherent          | mammary gland   | Luminal        | Metastatic carcinoma |
| MDA-MB-468    | epithelioid   | grape-like     | adherent          | mammary gland   | Basal A        | Metastatic adenocarcinoma |
| T47D          | epithelioid   | mass           | adherent          | mammary gland   | Luminal        | Invasive ductal carcinoma |
| FOCUS         | fibroblastoid | polygonal to   | adherent          | liver           | poorly differentiated | Hepatocellular carcinoma |
| Hep40         | epithelioid   | polygonal      | adherent          | liver           | well differentiated | Hepatocellular carcinoma |
| HepG2         | epithelioid   | polygonal, grow as clusters | adherent | liver | well differentiated | Hepatocellular carcinoma |
| Huh7          | epithelioid   | polygonal      | adherent          | liver           | well differentiated | Hepatocellular carcinoma |
| Mahlavu       | fibroblastoid | polygonal to   | adherent          | liver           | poorly         | Hepatocellular carcinoma |
|               |               | spindle-shaped |                   |                 | differentiated | carcinoma |
| PLC           | epithelioid   | polygonal      | adherent          | liver           | well differentiated | Hepatocellular carcinoma |
| SkHep1        | fibroblastoid | polygonal to   | adherent          | liver           | poorly differentiated | Hepatocellular carcinoma |
|               |               | spindle-shaped |                   |                 |                |                |

doi:10.1371/journal.pone.0052807.t001

| Breast cancer cell line | Liver cancer cell line |
|-------------------------|------------------------|
| BT-20                   | FOCUS                  |
| CAMA-1                  | Hep40                  |
| MDA-MB-157              | HepG2                  |
| MDA-MB-361              | Huh7                   |
| MDA-MB-453              | Mahlavu                |
| MDA-MB-468              | PLC                    |
| T47D                    | SkHep1                 |

doi:10.1371/journal.pone.0052807.t002
This paper is organized as follows: We first present the experimental results and then offer a brief discussion. In the Materials section, the used cell cultures are described. In the feature extraction section steps are described comprising image decomposition method by the dual-tree complex wavelet transform (DT-CWT), directional difference score computation and covariance matrix construction. In the classification section, SVM based covariance matrix classification algorithm is explained along with the foreground-background segmentation by EM algorithm and random subwindow selection.

Results
The dataset used in this study consists of 280 microscopic human carcinoma cell line images with each of the 14 classes having 20 images. Images in the dataset were acquired at 10×, 20× and 40× magnification. The size of each image was 3096×4140 pixels. 7 classes belonged to breast cancer cell lines and the other classes belonged to liver cancer. Each cell type has a specific phenotype in terms of nuclei (spherical vs. ovoid), nucleoli (prominent vs. hardly noticeable), size (large vs. small) and shape (round vs. cell pods) [1]. The names of the cancer cell lines used in our study are shown in Table 2 and example images of all 14 classes are shown in Figure 1. Aggressive cancer cells with metastatic properties switch from an epithelial-like (epithelioid) morphology to a spindle-shaped fibroblast-like (fibroblastoid) morphology during epithelial-mesenchymal transition (EMT), which is an indication of the invasiveness and metastatic capability of cancer cells. While epithelioid cells have polygonal shape with regular dimensions and sharp boundaries, fibroblastoid cells have elongated shapes and are bipolar or multipolar.

We adopt a 20-fold cross-validation strategy for the experiments. The dataset is divided into 20 disjoint subsets and each subset consisting of 14 images is used exactly once as the test set. For $k = 1 \ldots 20$, the $k^{th}$ subset is formed by taking the $k^{th}$ indexed image of each class. We run 20 experiments, choosing each image as the test image only once for each class, and obtain the average image classification accuracy over 20 runs. The number of selected random subwindows is taken to be $s = 100$. We perform the above experiment for both covariance and normalised covariance matrices, and for four different mapping functions in (10)-(13). SVM RBF kernel parameters are chosen as $y = 0.5$ and $C = 1000$. Experimental results are shown in Tables 3 for 10×, Table 4 for 20× and Table 5 for 40×. These tables show that normalised covariance matrix-based method outperforms the covariance method for all mapping functions, achieving an accuracy above 95%. Complex wavelet and directional difference features based classification methods (10)-(12) have higher accuracies than the classical covariance method in (13). Example images that were incorrectly classified are shown in Figure 2.

For comparison, similar experiments were carried out with scale-invariant feature transform (SIFT) [13] features. Table 6 shows the performance of those features. While the accuracy for discriminating between two cancer cell lines is 100%, the SVM classifier ($y = 1.3 \times 10^{-3}$ and $C = 1.3$) performs more poorly with each added cancer cell line. Furthermore, we investigated the effect of only using the diagonal of the normalised covariance matrix from Equation 7, i.e., the variance values of the features, as input for the SVM. Results can be seen in Table 7. The accuracy rates drop by approximately 10%. Therefore, using the covariance matrix from Equation 7, i.e., the variance values of the features, as inputs is vital for a good performance of the system.

Discussion
The proposed automated system for human breast and liver cancer cell line images can aid the biologist as a second reader and avoid the need for costly and time-consuming biochemical tests. The dual-tree complex wavelet transform and region covariance based computational framework is successfully applied to classify the cancer cell line images. We adopt a covariance-based approach by exploiting pixel-level attributes to construct local region descriptors encoding covariances of several attributes inside a region of interest. Pixel attributes are extracted using directional difference scores and the DT-CWT features. Since background regions occur frequently in a cancer cell line image, we randomly sample subwindows from the foreground image regions after foreground-background segmentation and each microscopic image is represented by correlation matrices of certain number of subwindows sampled randomly from the whole image. Finally, an SVM classifier with RBF kernel is trained to learn the class boundaries.

Figure 2 juxtaposes example images of cell line A that gets misclassified as cell line B, with examples of both cell lines A and B. All images were recorded at 20×. The three cell lines shown in the figure that get misclassified are MDAMB-468, Mahlavu and SKHep1. Some MDAMB-168 images get misclassified as MDAMB-361. Both are breast-cancer cell lines. From Figure 2, one understands that both images have layers, i.e., they have a 3-D structure, indicated by the white areas around the cell. This may be the reason why they get confused with one another. The liver cancer cell lines Mahlavu and SkHep1 are both misclassified as FOCUS, which is also a liver cancer cell-line. In the Mahlavu case, the image that gets misclassified shows several structures of significant length but short width, informally called “pods”. The FOCUS cell line has similar properties but, Mahlavu generally doesn’t. Also, the misclassified image in the figure shows less informative morphological properties, other than most Mahlavu images. In the case of SkHep1, the example image shows a sparser structure than most SkHep1 images. In the second column of the figure there are two different example images from the FOCUS cell line in order to demonstrate its varying pod morphology bearing poor differentiation. In addition, this preliminary observation indicates that when the cell lines are poorly differentiated (as in FOCUS, Mahlavu and SkHep1), their morphology may vary, hence they are more prone to be misclassified [14]. This observation can be further investigated in the future with a larger dataset specific to these kind of undifferentiated cell lines.

We demonstrate that automatic classification of microscopic carcinoma cell line images can be reliably performed using DT-CWT and correlation descriptors. Covariance descriptors are computed for features extracted from 2-D DT-CWT subbands and directional difference scores. Promising classification results were obtained by our experiments, which reveal the ability of the
proposed features to characterize breast and liver carcinoma cell line textures.

**Materials and Methods**

### 1 Cell Culture

The six hepatocellular carcinoma, one hepatoblastoma and seven breast cancer cell line images obtained from the following sources: FOCUS ([15]), Hep40 ([16]), Huh7 ([17]), SkHeP1 ([18]), SkHeP1 ([19]), GLC-82 ([20]), MDA-MB-231 ([21], HU-HC-10 ([20]), MDA-MB-453 ([21], MDA-MB-468 ([22], T47D ([23], ATCC HTB-131), MDA-MB-468 ([24], ATCC HTB-132), T47D ([25], ATCC HTB-133). The cell lines were seeded into dishes with 20% confluence and grown at 37°C under 5% CO₂ in standard Dulbecco’s modified Eagle’s medium (DMEM) supplemented with 10% FBS, 1% Non-Essential Aminoacid and 1% penicillin/streptomycin (GIBCO Invitrogen) up to 70% confluency. The authentication of the cell lines was regularly checked by STR profiling. Pictures were taken with Olympus CKX41 inverted microscope using Olympus DP72 camera with 20X objective.

### 2 Feature Extraction

#### 2.1 Dual-Tree complex wavelet transform.

The dual-tree complex wavelet transform (DT-CWT) has been recently used in various signal and image processing applications ([19], [20]) and [21]. It has desirable properties such as shift invariance, directional selectivity and lack of aliasing. In the dual-tree CWT, two maximally decimated discrete wavelet transforms are executed in parallel, where the wavelet families of two different trees form an approximate Hilbert transform pair [22]. Filterbanks for DT-CWT are shown in Figure 3. Low-pass analysis filters in real and imaginary part of the 2nd level detail wavelet filters are used to obtain a single intensity image which is lowpass filtered with $\frac{1}{4} + \frac{1}{4}$ filters and downsampling by 4 in both directions to obtain a single intensity image $I_W(x, y)$ which represents the original image and will be used as the image to be classified. Let $W_0(x, y)$ and $W_0(x, y)$ denote, respectively, the real and imaginary part of the 2nd level complex wavelet

| Feature mapping function | Covariance-based classification | Normalised Covariance-based classification |
|--------------------------|--------------------------------|------------------------------------------|
| $\phi_1(x, y)$           | 96.8                          | 97.5                                     |
| $\phi_2(x, y)$           | 96.8                          | 98.6                                     |
| $\phi_3(x, y)$           | 96.4                          | 97.1                                     |
| $\phi_4(x, y)$           | 77.5                          | 86.1                                     |

| Feature mapping function | Covariance-based classification | Normalised Covariance-based classification |
|--------------------------|--------------------------------|------------------------------------------|
| $\phi_1(x, y)$           | 97.5                          | 99.3                                     |
| $\phi_2(x, y)$           | 96.8                          | 98.6                                     |
| $\phi_3(x, y)$           | 97.9                          | 99.3                                     |
| $\phi_4(x, y)$           | 77.9                          | 85.7                                     |

Microscopic cancer cell line images contain significant amount of oriented singularities. Recently, a Bayesian classification method that uses the sparsity in a transform domain is developed to classify cancer cell lines [26]. Attributes like orientation selectivity and shift invariance render DT-CWT a good choice for the processing of microscopic images with lots of edge- or ridge-like singularities. We incorporate the complex wavelet transform into recently proposed region covariance descriptors [27] for feature extraction from microscopic images. In the region covariance framework each pixel is mapped to a set of pixel properties which’s covariances are measured and used as a region descriptor. We use DT-CWT complex coefficient magnitudes in detail subbands as pixel features and compute covariance descriptors. Augmenting covariance matrices with directional information through the use of 2-D DT-CWT helps to improve the discriminative power of descriptors.

2-D DT-CWT of an image is obtained by four real separable transforms [27]. Real-part and imaginary-part analysis filters are applied successively to rows and columns of the image. By addition and subtraction of corresponding detail subbands, we obtain a total of 16 subbands consisting of 6 real detail subbands, 6 imaginary detail subbands and 4 approximation subbands. Two-dimensional dual-tree decomposition is an oversampled transform with a redundancy factor of $\frac{1}{2}$ for d-dimensional signals. In our work, we perform two-level 2-D DT-CWT decomposition of each biomedical image of size $m \times n$ and use only the 2nd level detail subband coefficients to better exploit the analyticity of DT-CWT.

Each subband at the 2nd level is of size $\frac{m}{4} \times \frac{n}{4}$. The original image is lowpass filtered with $\frac{1}{4} + \frac{1}{4}$ filters and downsampling by 4 in both directions to obtain a single intensity image $I_W(x, y)$ which represents the original image and will be used as the image to be classified. Let $W_0^R(x, y)$ and $W_0^I(x, y)$ denote, respectively, the real and imaginary part of the 2nd level complex wavelet.
Table 5. Average classification accuracies (in %) of 40× carcinoma cell line images over 20 runs using SVM with RBF kernel.

| Feature mapping function | Covariance -based classification | Normalised Covariance -based classification |
|--------------------------|----------------------------------|--------------------------------------------|
| ϕ₁(I, x, y)              | 89.3                             | 95.7                                       |
| ϕ₂(I, x, y)              | 90.0                             | 96.4                                       |
| ϕ₃(I, x, y)              | 92.5                             | 96.8                                       |
| ϕ₄(I, x, y)              | 61.2                             | 85.0                                       |

doi:10.1371/journal.pone.0052807.t005

Coefficient at the position (x, y) corresponding to directional detail subbands at orientation θ, where $θ \in \{±15, ±45, ±75\}$. The magnitude of the complex wavelet coefficient is then given by

$$M_d(x, y) = \sqrt{W_R(x, y)^2 + W^0_R(x, y)^2}$$  (1)

Hence, for each pixel in the average image $I_d(x, y)$, six complex wavelet coefficient magnitudes $M_d(x, y)$ representing six different orientations of DT-CWT are extracted. These magnitudes will be utilized as features in the covariance matrix computation for randomly sampled regions of the image $I_d(x, y)$. The computational complexity of DT-CWT is $O(M \cdot N)$, where $M \cdot N$ refers to the number of pixels in the image.

2.2 Directional differences. In order to account for the large morphological variation of the images in our dataset, we evaluated differences between pixels in various directions. Consider a point $p_1$ on a two-dimensional function $I(x, y)$. Now consider a second point $p_2$. The Euclidean distance between $p_1$ and $p_2$ is $d$ and $p_2$ lies on a circle that has an orientation of angle $\alpha$ with respect to the $x$-coordinate, i.e., $p_2$ lies on a circle, which’s center point is $p_1$ and has a radius $d$. The difference between $p_1$ and $p_2$ can be written as

$$T(d, \alpha) = |I(x, y) - I(x + d \cdot \cos \alpha, y + d \cdot \sin \alpha)|.$$  (2)

Now consider we want to compute a couple of difference values for equidistant concentric circles where the largest circle has radius $R$ and the smallest has radius $R/A$, where $A$ is an integer with values ranging from $[1, R]$. When the parameters $R$ and $A$ are fixed, we can rewrite the above equation as

$$T(i, \alpha) = |I(x, y) - I(x + i \cdot \frac{R}{A} \cos \alpha, y + i \cdot \frac{R}{A} \sin \alpha)|.$$  (3)

where $i=1,2,...,A$. We can compute a score for each $\alpha$ value by computing a function with respect to $i$, as

$$s_{\alpha} = F_j(T(i, \alpha)).$$  (4)

For example, $F_j$ can be the median function. In that case $s_{\alpha}$ is simply the median of all the differences between the center pixel and the points at distances $\frac{R}{A}$ at the fixed orientation $\alpha$. We use these scores as features in covariance matrix computation. Three different functions, namely median, max and mean functions, are employed for $F_j$ in this study. For each image $I_d(x, y)$ obtained according to the dual-tree complex wavelet section, 8 output images of the same size are generated as the result of the function $F_j$, corresponding to 8 different orientations when the radius $d$ is chosen as 5 in the experiments. Hence, in addition to DT-CWT features, each pixel $(x, y)$ of the image $I_d$ has 8 attributes, which denote the scores $s_{\alpha}$ for 8 different $\alpha$ values.

The computational complexity of the directional difference operation is $O(n^2 \alpha^2)$, where $n$ and $\alpha$ refer to the number of digits of the pixels and the number of considered angles, respectively.

2.3 Covariance matrices for cell line description. Successfully employed in texture classification

Figure 2. Examples of misclassified images (20×). Misclassified images are shown in the first column. Examples from their true cell line are given in the second column. Images in the third column show examples of the cell line that the images got misclassified into.

doi:10.1371/journal.pone.0052807.g002
In this study, normalized covariance matrices are used as in [32], pedestrian detection [29] and flame detection [30], covariance descriptors enable the combination of different features over an image region of interest. Given an intensity image \( I \) of size \( m \times n \), we define a mapping \( \phi \) from image domain to feature domain as

\[
F(x,y) = \phi(I,x,y)
\]

where each pixel \( (x,y) \) is mapped to a set of features and \( F \) is the \( m \times n \times d \) dimensional feature function. For a given subwindow \( R \) consisting of \( n \) pixels, let \( \{f_k\}_{k=1}^n \) be the \( d \)-dimensional feature vectors extracted from \( R \). Then, the covariance matrix of region \( R \) can be computed as

\[
C = \frac{1}{n-1} \sum_{k=1}^{n} (f_k - \mu)(f_k - \mu)^T
\]

where \( \mu \) is the mean of the feature vectors inside the region \( R \). The covariance matrix is symmetric positive-definite and of size \( d \times d \). There exists a very efficient multiplier-less implementation of covariance descriptors, called co-difference matrices, which have been shown to yield comparable performances to the original ones [31].

In this study, normalized covariance matrices are used as in [32].

\[
\tilde{C}(i,j) = \begin{cases} 
\sqrt{C(i,j)}, & \text{if } i = j \\
\sqrt{C(i,j)C(j,i)}, & \text{otherwise}
\end{cases}
\]

With

\[
M_0(x,y) = [M_{01}(x,y) \ldots M_{0k}(x,y)]
\]

and

\[
s_0^k(x,y) = [s_{01}^k(x,y) \ldots s_{0y}^k(x,y)]
\]

where \( f_1 \ldots f_k \) correspond to the six orientations of DTF-CWT detail subbands \( \{ \pm 15, \pm 45, \pm 75 \} \), \( M_k(x,y) \) is as defined in Equation (1), \( z_1 \ldots z_8 \) correspond to the eight orientations of directional difference score estimation, and \( k = 1,2,3 \) denote, respectively, the median, max and mean functions \( \mathcal{F}_j \) in the directional differences section, feature mapping functions employed in this study are

\[
\phi_1(I,x,y) = [I_{d}(x,y)I_{x}||I_{y}||I_{xx}||I_{yy}||M_0(x,y)s_0^2(x,y)]^T, \quad (10)
\]

\[
\phi_2(I,x,y) = [I_a(x,y)I_{x}||I_{y}||I_{xx}||I_{yy}||M_0(x,y)s_0^2(x,y)]^T, \quad (11)
\]

\[
\phi_3(I,x,y) = [I_a(x,y)I_{x}||I_{y}||I_{xx}||I_{yy}||M_0(x,y)s_0^2(x,y)]^T, \quad (12)
\]

\[
\phi_4(I,x,y) = [I_a(x,y) |I_x| |I_y| |I_{xx}| |I_{yy}|]^T \quad (13)
\]

where \( |I_x| \) and \( |I_{xx}| \) denote the first- and second-order derivatives at \( (x,y) \) of the image \( I_a \). The computational complexity of covariance matrix computation is \( O\{d^2\} \), where \( d \) refers to the number of features in the subimage.

### 3 Classification Using a Multiclass SVM

The images in our dataset show a large amount of background pixels. Clearly, the background is not discriminative. Therefore, we address the issue of segmenting the images into foreground and background before classification. For our dataset, a simple thresholding scheme is not sufficient for segmentation, since foreground pixels have a large variance and may therefore have values higher and lower than the background pixels. We modeled the image as a mixture of two Gaussians, representing the foreground and background pixels, respectively. Using this model, an Expectation-Maximization (EM) algorithm was applied for segmentation. The result is noisy, so a morphological closing operation was applied, followed by median filtering. We obtained the sizes of the closing and median filter kernels by comparing the scores of the segmentation results of various kernel sizes. The used score was first described in [33] and evaluated in [34]. Examples can be seen in Figure 4.

Since it is necessary to focus on foreground-like regions in carcinoma cell line images, \( s \) analysis square windows are randomly selected, as in [35], from each image with the two constraints: the percentage of the foreground pixels in the selected region of an image must be above 50 and the variance of the selected region must exceed an image-dependent threshold, which is the variance of the whole image.

For each subwindow, a covariance matrix is computed using Equation (6) for each of the feature mapping functions in (10)-(13). The image signature is composed of \( s \) covariance matrices of the

### Table 6. Classification accuracies for SIFT features.

| Number of cell lines | Classification accuracy in % |
|----------------------|-------------------------------|
| 2                    | 100.00                        |
| 3                    | 80.00                         |
| 4                    | 66.25                         |
| 5                    | 60.00                         |
| 6                    | 51.67                         |
| 7                    | 56.43                         |
| 8                    | 47.50                         |
| 9                    | 42.22                         |
| 10                   | 38.50                         |
| 11                   | 35.91                         |
| 12                   | 34.23                         |
| 13                   | 32.00                         |
| 14                   | 36.07                         |

| Magnification | Classification accuracy in % |
|---------------|-------------------------------|
| 10×           | 84.60                         |
| 20×           | 84.60                         |
| 40×           | 80.00                         |

Table 7. Classification accuracies for variance values only.
same size. Each class is represented by $s \times \#(\text{images in each class})$ covariance matrices. Covariance matrices are symmetric positive-definite and do not lie in the Euclidean space; so, they are vectorized resulting in $d(d+1)/2$-dimensional vectors for $d \times d$ matrices. A multiclass SVM classifier is trained with RBF kernel in the $d(d+1)/2$-dimensional vector space using the training points. SVM algorithm is implemented using LIBSVM library [36]. For each test subwindow, the corresponding covariance descriptor is vectorized and fed into the trained SVM model for prediction. Therefore, there exist $s$ labels for each microscopic image corresponding to $s$ subwindows, and the image in question is assigned the label that gets the majority of votes among $s$ labels. The above process is re-executed using normalised covariance matrices instead of unnormalised covariance matrices. In order to compare the discriminative power of our features with more traditional one, we carried out similar experiments with SIFT [13] features for the $20 \times$ images. In SIFT, feature points are extremas in scale-space, i.e., a difference-of-gaussians (DoG) pyramid. The method is invariant to scale, orientation and location of the features, which makes it a commonly-used method in the field of computer vision. In our experiments, SIFT features are computed on the foreground that is found according to the description above. The resultant feature vectors for the images were then fed into an SVM. Table 6 shows the performance of those features. While the accuracy for discriminating between two cancer cell lines is 100%, the SVM classifier performs more poorly with each added cancer cell line.

The computational complexity of SVM classification in the test phase is $O\left(\left(d(d+1)/2\right) \cdot S\right)$ [37], where $d$ and $S$ refer to the number of features and the number of support vectors, respectively.

**Availability and Future Directions**

The software can be tested at http://signal.ee.bilkent.edu.tr/cancerCellLineClassificationEngine.html. The datasets used in this study can also be downloaded from there and can be used by fellow researchers in future studies. Images to be uploaded should be recorded using either $10 \times$, $20 \times$, $40 \times$ magnification and should be in JPG format. The authors are currently working on making the described procedure more computationally efficient by using a single-tree approximation to the dual-tree complex wavelet transform used in this study.

**Supporting Information**

Data S1 The supporting information consists of a RAR file named ‘Data S1.rar’. This file includes several MATLAB files that can be used to evaluate the identity of test images provided by the user. Note that an online version of this program is available at http://signal.ee.bilkent.edu.tr/cancerCellLineClassificationEngine.html and a dataset of images is available at http://signal.ee.bilkent.edu.tr/cancerCellLineClassificationSampleImages.html. (RAR)

**Author Contributions**

Implementation and design of software: FK AS KK AEC. Conceived and designed the experiments: FK AS TE RCA AEC. Performed the experiments: FK AS TE. Analyzed the data: FK AS KK TE RCA AEC. Contributed reagents/materials/analysis tools: FK AS KK TE RCA AEC. Wrote the paper: FK AS TE RCA AEC.
References

1. Dundar M, Badve S, Raykar V, Jain R, Sertel O, et al. (2010) A multiple instance learning approach toward optimal classification of pathology slides. In: Pattern Recognition (ICPR), 2010 20th International Conference on. 2732–2735. doi:10.1109/ICPR.2010.669.

2. Buontempo F, Eraslan T, Misirli S, Senturk S, Etro D, et al. (2011) Inhibition of akt signaling in hepatoma cells induces apoptotic cell death independent of akt activation status. Investigational New Drugs 29:1303–1313.

3. Vilela M, Morgan JJ, Lindahl PA (2010) Mathematical model of a cell size checkpoint. PLoS Comput Biol 6: e1001036.

4. Steuer R, Waldherr S, Sooraj V, Kollmann M (2011) Robust signal processing in living cells. PLoS Comput Biol 7: e1002218.

5. Kuyt R, Ud A, Hafner M, Gangl A, Wbra F, et al. (2010) Predicting the histology of colorectal lesions in a probabilistic framework. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. 103–110. doi:10.1109/CVPRW.2010.5543146.

6. Kuyt R, Ud A (2007) Modeling the marginal distributions of complex wavelet coefficients magnitudes for the classification of zoom-endoscopy images. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. pp. 1–8. doi:10.1109/ICCV.2007.4699170.

7. Aydogan D, Hannula M, Arola T, Hyttinen J, Dastidar P (2008) Texture based classification and segmentation of tissues using dt-cwt feature extraction methods. In: Computer-Based Medical Systems, 2008. CBMS ’08. 21st IEEE International Symposium on. pp.614–619. doi: 10.1109/CBMS.2008.46.

8. Kilic N, Kursun O, Ucan O (2010) Classification of the colonic polyps in ct-colonography using region covariance as descriptor features of suspicious regions. Journal of Medical Systems 34:101–105.

9. Doyle S, Hwang M, Shah K, Madabhushi A, Feldman M, et al. (2007) Automated grading of prostate cancer using architectural and textural image features. In: Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on. 1204–1207. doi: 10.1109/ISBI.2007.37094.

10. Lebrun G, Charrier C, Lezoray O, Mourie C, Cardot H (2007) A fast and efficient segmentation scheme for cell microscopic image. Cellular And Molecular Biology 53:51–61.

11. Raza S, Parry R, Sharma Y, Caudry Q, Moffitt R, et al. (2010) Automated classification of renal cell carcinoma subtypes using bag-of-features. In: Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE. 6749–6752. doi: 10.1109/EMBC.2010.5626009.

12. Conrad C, Ezile H, Wormalt P, Deagle N, Letch T, et al. (2004) Automatic identification of subcellular phenotypes on human cell arrays. Genome Research 14:1130–1136.
13. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vision 60: 91–110.

14. Sayan B, Emre N, Irmak M, Ozturk M, Getin-Atalay R (2009) Nuclear exclusion of p33ing1b tumor suppressor protein: explored in hcc cells using a new highly specific antibody. Hybridoma 28.

15. Le L, Iselebacker KJ, Wands JR, Goodman H, Shah C, et al. (1985) Establishment and characterization of a new human hepatocellular carcinoma cell line. J Cell Physiol 165: 459–467.

16. Boushabab B, Nishikawa Y, Simson D, Carr B (1984) Growth control and gene expression in a new hepatocellular carcinoma cell line, HepG2: inhibitory actions of vitamin K. In Vitro 20: 493–504.

17. Oelfinger P, Bronson D, Dreesman G (1981) Induction of hepatitis b surface antigen in human hepatoma-derived cell lines. J Gen Virol 53: 105113.

18. Selesnick IW, Li KY (2003) Video denoising using 2d and 3d dual-tree complex wavelet transforms. In: Wavelet Appl Signal Image Proc. X (Proc. SPIE 5207. pp.607–618.

19. Loo P, Kingsbury N (2000) Digital watermarking using complex wavelets. In: Image Processing, 2000. Proceedings. 2000 International Conference on. volume 3, pp. 29–32 vol.3. doi: 10.1109/ICIP.2000.899275.

20. Chen G, Bui T, Krzyzak A (2006) Palmprint classification using dual-tree complex wavelets. In: Image Processing, 2006 IEEE International Conference on. pp.2645–2648. doi: 10.1109/ICIP.2006.313053.

21. Thamarai M, Shanmugalakshmi R (2010) Video coding technique using swarm intelligence in 3-d dual tree complex wavelet transform. In: Machine Learning and Computing (ICMLC), 2010 Second International Conference on. pp.174–178. doi:10.1109/ICMLC.2010.39.

22. Selesnick I, Baraniuk R, Kingsbury N (2005) The dual-tree complex wavelet transform. Signal Processing Magazine, IEEE 22: 123–151.

23. Selesnick I (2001) Hilbert transform pairs of wavelet bases. Signal Processing Letters, IEEE 8: 170–173.

24. Selesnick I, Baraniuk R, Kingsbury N (2005) The dual-tree complex wavelet transform. Signal Processing Magazine, IEEE 22: 123–151.

25. Suhre A, Ersahin T, Getin-Atalay R, Getin AE (2011) Microscopic image classification using sparsity in a transform domain and Bayesian learning. In: 19th European Signal Processing Conference. pp.1005–1009.

26. Tuzel O, Porikli F, Meer P (2006) Region covariance: A fast descriptor for detection and classification. In: Leonardis A, Bischof H, Pinz A, editors, Computer Vision ECCV 2006, Springer Berlin/Heidelberg, volume 3952 of Lecture Notes in Computer Science. pp.589–600.

27. Kingsbury N (1997) Image processing with complex wavelets. Phil Trans Royal Society London A 357: 2543–2560.

28. Tuzel O, Porikli F, Meer P (2006) Region covariance: A fast descriptor for detection and classification. In: Proc. 9th European Conf. on Computer Vision. pp.689–696.

29. Tuzel O, Porikli F, Meer P (2008) Pedestrian detection via classification on riemannian manifolds. Pattern Analysis and Machine Intelligence, IEEE Transactions on 30: 1713–1727.

30. Habiboglu Y, Gunay O, Getin A (2011) Flame detection method in video using covariance descriptors. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. pp.1817–1820. doi:10.1109/ICASSP.2011.5946057.

31. Tuna H, Oneran I, Getin A (2009) Image description using a multiplier-less operator. Signal Processing Letters, IEEE 16: 731–733.

32. Habiboglu YH, Gunay O, Getin AE (2011) Real-time wildfire detection using correlation descriptors. In: 19th European Signal Processing Conference (EUSIPCO 2011). Special Session on Signal Processing for Disaster Management and Prevention. pp.894–898.

33. Nazif AM, Levine MD (1984) Low level image segmentation: An expert system. Pattern Analysis and Machine Intelligence, IEEE Transactions on PAMI-6: 553–577.

34. Chabrier S, Emile B, Rosenberger C, Laurent H (2006) Unsupervised performance evaluation of image segmentation. EURASIP J Appl Signal Process 2006: 217–217.

35. Maree R, Geurts P, Pieter J, Wehenkel L (2005) Random subwindows for robust image classification. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society 1 Conference on. volume 1, 34–40 vol. 1. pp.10.1109/CVPR.2005.

36. Chang CC, Lin CJ (2011) LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2: 27:1–27:27.

37. Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2: 121–167.