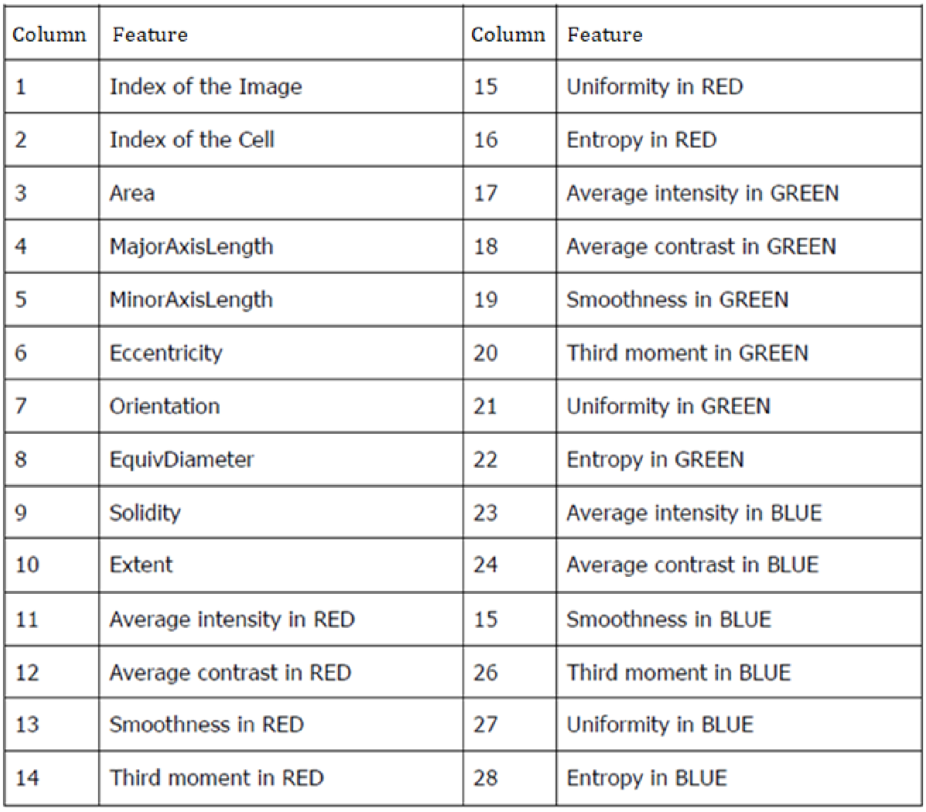
The SIPAKMED database is publicly available and it can be used for experimental purposes with the request to cite the following paper:

M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, A. Charchanti, SIPAKMED: A new dataset for feature and image based classification of normal and pathological cervical cells in Pap smear images, IEEE International Conference on Image Processing (ICIP) 2018, Athens, Greece, 7-10 October 2018.

Description of Cell Features

In each image of our database, the boundaries of the regions of interest, i.e. the area of the cytoplasm and the nucleus of the cells, were manually defined by expert observers. The coordinates of the contour of each area are provided for both the images of cell clusters (in im_(*class)* directories) and the isolated cell images (in the CROPPED subdirectory of each class).

In every region of interest, we calculate 26 features concerning the intensity (average intensity, average contrast), the texture (smoothness, uniformity, third moment, entropy in all three color channels), and the shape (area, major and minor axis length, eccentricity, orientation, equivalent diameter, solidity and extent). These features were calculated for both the region of the nucleus and the cytoplasm of each cell and they are stored in 10 tables (cytoplasm and nuclei features for each class) of 28 columns (the two added fields denote the number of the image and the cell correspondingly). The feature tables are stored in Feature_CELL directory and they have the following structure:



Column	Feature	Column	Feature
1	Index of the Image	15	Uniformity in RED
2	Index of the Cell	16	Entropy in RED
3	Area	17	Average intensity in GREEN
4	MajorAxisLength	18	Average contrast in GREEN
5	MinorAxisLength	19	Smoothness in GREEN
6	Eccentricity	20	Third moment in GREEN
7	Orientation	21	Uniformity in GREEN
8	EquivDiameter	22	Entropy in GREEN
9	Solidity	23	Average intensity in BLUE
10	Extent	24	Average contrast in BLUE
11	Average intensity in RED	15	Smoothness in BLUE
12	Average contrast in RED	26	Third moment in BLUE
13	Smoothness in RED	27	Uniformity in BLUE
14	Third moment in RED	28	Entropy in BLUE

Description of Image and Deep Features

The trained VGG-19 Convolution neural network was used as a feature extractor by feeding it an input cell image and using the intermediate layer pre-activations to construct a feature vector.

File ConvFeatures.mat :

In this file there are features extracted by using the pre-activations of the last convolutional layer (layer conv5).

These features are then aggregated by sum-pooling to form a feature vector of size 512 to describe each cell image. For every cell there are 3 types of features: raw features 512 size, compressed features 256 size, compressed features 32 size

File FcFeatures.mat:

In this file there are features extracted by using the pre-activations of the first fully-connected layer (layer fc6).

These features are of size equal to the number of neurons in the corresponding layer, i.e. 4096. For every cell there are 3 types of features: raw features 4096 size, compressed features 256 size, compressed features 32 size.

Both files shared the following structure:

-- Folds 1-5 (5-fold cross-validation)

-- Cell Classes (Superficial-Intermediate, Parabasal, Koilocytotic, Dyskeratotic, Metaplastic)

-- Size of cell features (Raw values, PCA 256, PCA 32)

-- Feature for each cell (Cell cluster image id, cell id, [features])

This work was co-financed by the European Union (European Regional Development Fund- ERDF) and Greek national funds through the Operational Program THESSALY- MAINLAND GREECE AND EPIRUS-2007-2013 of the National Strategic Reference Framework (NSRF 2007-2013).