Annals of the University of Craiova, Mathematics and Computer Science Series
Volume 45(1), 2018, Pages 156–166
ISSN: 1223-6934

On supporting cancer grading based on histological slides using a limited number of features

Dumitru Bogdan Mateis Boruz and Catalin Stoean

Abstract. A computer assisted cancer diagnosis approach based on histopathological images is outlined in the current study. It mainly detects the lighter components that appear in each slide and computes simple morphological features for them, as well as basic statistic about these. This relatively short numerical data set is subsequently fed to a classifier for learning the correspondence between the features and their classes. The classification accuracy reaches outputs similar to other more complex methodologies and therefore encourages future investigation.

image classification, feature extraction, support vector machines
[2010]Primary 60J05; Secondary 60J20.

1. Introduction

There are more and more frequent computational applications that put forward decision making algorithms meant to assist automated diagnosis or prognosis prediction in medicine [6], [10], [14], [15], [25], [28], [29]. The diagnosis of cancer makes no exception, especially as this is a disease that can be cured only through early identification.

Histology (also called microscopic anatomy) is the study of the form or structures obtained from tissues seen under a light or electron microscope. Biopsies presume the extraction of tissue from a patient's body for examination to determine if it is healthy or the degree of extent of a disease. After it had been removed, the tissue is treated using chemicals in order to prevent decay, then it is placed under the microscope, where an image of it is captured, resulting in a histological image [9]. Examination of histological images represents the current usual practice in establishing an accurate diagnosis for cancer.

Colorectal cancer is the third most commonly diagnosed malignancy and the fourth leading cause of death in the world [1], [8]. Age and family history are primary risk factors, hence periodical screening for persons past certain age, depending on the cancer type, is recommended by medical professionals. This occurs because the recovery, treatment procedure and survival chances heavily depend on the development stage of the disease [4]. Considering that roughly a third of the population in the US is over 50, the number of screenings produced per year ends up being substantially high, even if not all people actually do yearly medical examinations [5], [36]. This leads to acknowledging the economical significance of a potential digital pathology type of application which can rapidly diagnose a patient, saving thus valuable time for both the patient and the pathologist in the short term and providing noticeable cost savings in the long term [12]. Therefore, if this type of diagnostic assistance were used at least for identifying the healthy tissues, it could give the pathologist time to focus on the more complicated samples.

The goal of the current research is to investigate whether a minimalistic feature extraction procedure that uses only a limited amount of morphological features, taking into account only lighter spots (glands, adipose tissue) based on the intensity of the pixels, and further employs a support vector machine (SVM) classifier is able to achieve accurate results. A methodology applied for the same data set that extracts significantly more features [26] is used for comparison.

2. Data Set

The data set used [24] for learning in this research is extracted from a set of digitalized histopathological slides of healthy tissue and diseased tissue, separated in grades 1, 2 and 3. The images are at 10x magnification level with a 800x600 resolution and are obtained from the Emergency County Hospital of Craiova, Romania. The healthy tissue is further refered as G0, and G1, G2 and G3 are stages of cancer ordered by severity. There are 357 images in total, with 62 healthy tissue images and 96 or more records for each of the three grades of cancer.

Figure 1 shows pairs of initial images and the same samples with contours for each of the four classes in turn.

3. State of the Art

There are numerous research efforts to sustain cancer diagnosis by means of microscopical image analysis [2], [3], [19], [30].

The traditional methodologies comprise the following stages [7], [11], [16]:

(1) **Preprocessing**. This regards the sampling of the slides and also color and illumination normalization. This way, corrections can be added to the initial images prior to the identification of the important components.
(2) **Segmentation**. It refers to the identification of the main components that are usually studied within the histological images, i.e. nuclei and glands. The identification is based either on region or on boundary of the components.
(3) **Feature extraction and feature selection**. Various attributes are extracted from the segmented components like morphological ones, textural, fractal, topological or based on intensity. As usually this leads to a very high number of attributes, feature selection techniques are applied to reduce the dimensionality.
(4) **Diagnosis**. A classifier is employed to deal with the numerical data set. It is usually an algorithm from the supervised learning class, but it can be also unsupervised or semi-supervised.

More recently, deep learning algorithms, such as convolutional neural network, start the analysis from feature extraction [22]. Moreover, the features are learned solely from data and not handcrafted based on general human knowledge. However, although this class of methods is generally significantly more accurate, it needs a large amount of data to train on and the sizes of the images have to remain low. Additionally, they require large computational power even for low resolution images, which,

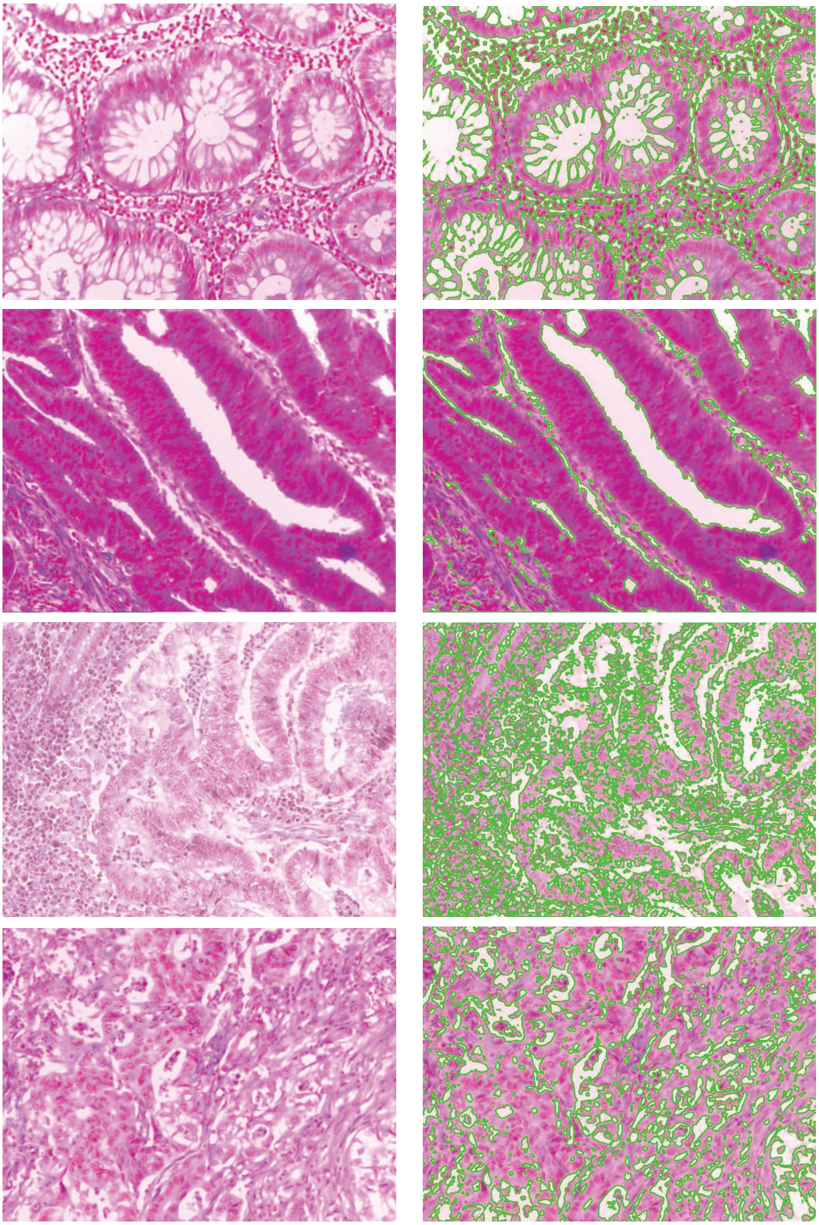

FIGURE 1. Samples for healthy tissue and each of the G1-G3 grades in the left column and their detected contours on the right column.

naturally, cannot contain all the details necessary to distinguish between classes, at least for the human experts.

4. Proposed Methodology

A limited number of attributes are extracted from the histological slides and machine learning techniques are applied in order to classify them. The components that are analysed are represented only by glands. Contours for them are found using thresholding. The used morphological features are the perimeter and area of the found contours and for each slide statistics like the minimum, maximum, average and standard deviation for both measures are computed. This leads to a number of 8 numerical attributes for each image.

The task can be expressed in the form of a supervised learning problem, with each image assigned to a class. The data set is separated into training and test sets. The classifier learns the correspondence from the training samples and their classes and is subsequently applied to identify the outputs for the test images.

The data collection process starts by finding the appropriate thresholding parameters which offered the clearest contours for the images. Afterwards measures about every contour in each image such as perimeter and area are gathered. Relevant information such as the highest and lowest values for area and perimeter as well as their mean and standard deviation are preserved.

The procedure for collecting data from an image works in the following manner:

(1) The image is transformed in grayscale format.
(2) A thresholding procedure is applied to the resulted image. This highlights key features in the reached image depending on the values for the parameters used in the thresholding.
(3) The contours in the thresholded image are found and the area and perimeter are calculated and stored for each one to be used in creating the statistical information described above.

Although SVM are known to deal well with a high number of attributes [13], in the current work we keep the number of features low, referring only those that are considered to be essential, as others might add noise to the numerical data set. Additionally, the number of contours found was added, but the pre-experimental results showed that the results were not significantly improved.

When extracting features from the dataset, special attention is set on fine tuning the thresholding weight so the glaring differences between grades are visible, especially as concerns the extracted numerical data. This allows the SVM algorithm to more easily find the differences for classification.

5. Experimental Results

The current section begins with the setup of the experiments, then shows the main results and is followed by a subsection where the outputs are discussed.

5.1. Experimental setup. The training data represents two thirds of the entire data set and the last third is used for prediction. Each setting is validated in 30 repeated runs. The separation into the training/test sets is made randomly in every repetition. The random sampling will minimize the bias that could appear during testing. The average accuracy is obtained by calculating the accuracy over the 30 repetitions of the application of the SVM.

Various parameters will be tweaked or changed during testing in order to find the best ones, namely all the SVM kernels are tested along with various values for cost and gamma. The parameter tuning is made both manually and automatically with similar results. Another parameter outside of the learning function that will be tested is the thresholding value, specifically values between 100 and 230 will be tested in order to transform the input image into a binary one.

The approach for finding the best parameters in this case was both trial and error and making use of the tools in R designed for fine tuning. The optimal parameters were manually found to be a radial kernel, a cost of 10 and a gamma value of 0.1. The automated *tune* function which has a brute force attempt to find the best parameter in the given range - in this case between 1-15 for cost, incrementing with .5 for each attempt, and 0.1 and 2.0 for gamma, incrementing with .1 per cycle - found different parameters to be optimal. However, after testing those parameters, the accuracy values were not visibly different from the ones found manually.

5.2. Results and Visualization. Figure 2 shows the importance of every attribute with respect to each class, as discovered by a random forest model.

Figure 3 illustrates the classification test accuracy for the SVM with default parameters and with tuned ones when the thresholding value is varied from 100 to 230. As observed, more intermediary steps are tried where the best results are achieved.

Tables 1 and 2, as well as Figure 4, show various statistics computed per class.

Class	Sensitivity	Specificity	Precision	F1	Balanced Acc.
G0	0.889	0.961	0.800	0.842	0.925
G1	0.727	0.955	0.857	0.787	0.841
G2	0.909	0.886	0.750	0.822	0.898
G3	0.865	0.988	0.970	0.914	0.926

Table 1. Statistical values computed per class.

Class	Neg Pred Value	Prevalence	Detection Rate	Detection Prev.
G0	0.980	0.149	0.132	0.165
G1	0.903	0.273	0.198	0.231
G2	0.963	0.273	0.248	0.331
G3	0.943	0.306	0.264	0.273

Table 2. Secondary table of statistical measures.

5.3. Observations and Discussions. Initially the thresholding parameter was incremented by 10 for each collection. After finding the sweet spot between 200-230, smaller increments were made in that interval until the best parameter for collecting data was found at 217. This is more visually apparent in Figure 3.

Beside considering all found contours, a couple of other contour detection types were attempted, like considering only the exterior ones, or going only two levels deep (i.e. ignoring all contours that were below the second level). However, the results were not affected to a large extent by these changes, only showing a moderate decrease in

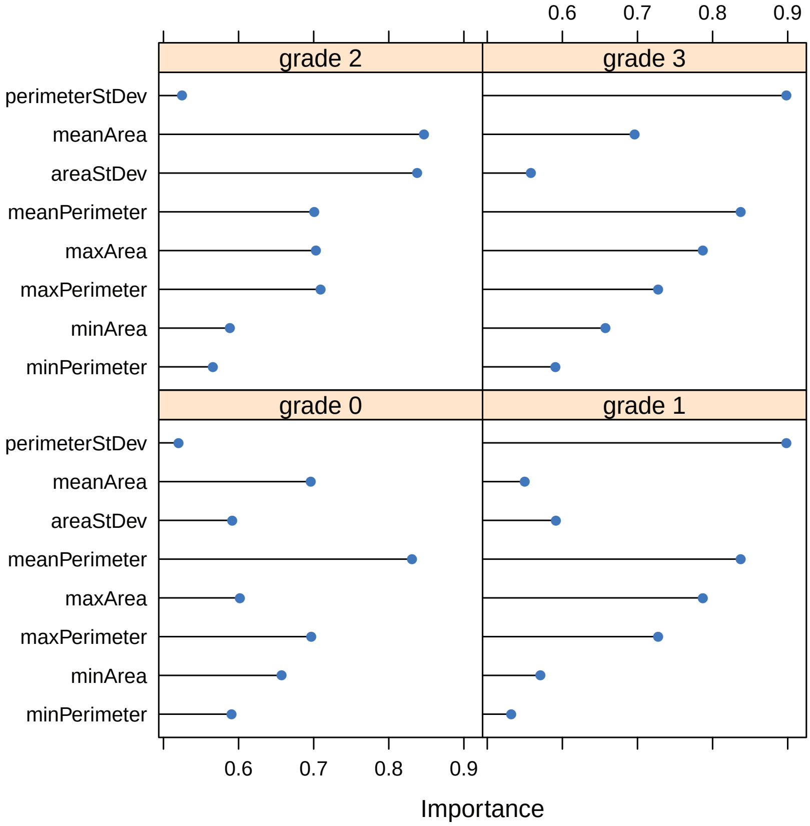

FIGURE 2. The level of importance for every attribute for each class as found by a random forest model.

accuracy for some threshold values, while remaining unchanged for the remaining majority.

Another attempt to add more information to the numerical data collection was to combine two types of results into one and have that submitted for classification with SVM. For instance adding a dataset obtained through the custom method described earlier to another dataset collected in the standard manner was tried. This effectively doubled the amount of attributes for each observation, but this had no significant effect on the results in either direction.

Besides the vertical enlargement (doubling the number of attributes), horizontal addition was also tried, specifically doubling the amount of observations without success as concerns the increase in the test classification accuracy. The total number of contours per image was also considered as attribute, but without improving the accuracy, so it was set aside.

Figure 2 is the result of an analysis that was made using a random forest model to investigate the effect of the set of parameters that achieved the highest accuracy. The

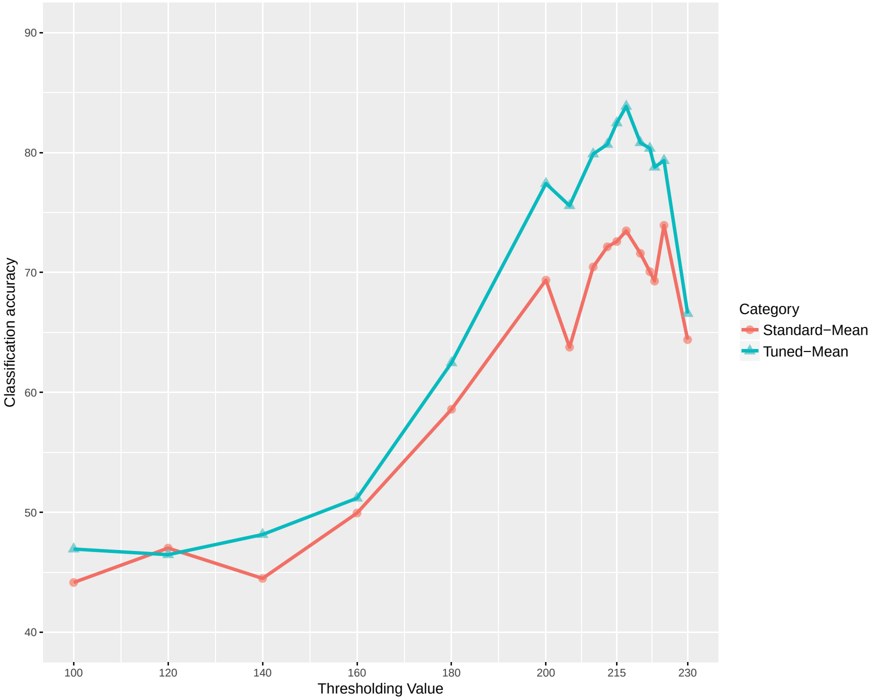

Figure 3. Classification accuracy for various values of the threshold parameter.

most important attributes for each class are found through the use of the R package caret [17]. This led to interesting findings concerning how each class is differently affected by them. While the average perimeter seems to have a consistent influence on each class, other attributes importance varies significantly depending on grade.

After covering the data processing and analysis, the SVM function becomes the next object to test upon. After the pre-experimental phase phase, the best kernel for this dataset proved itself to be the radial one, while the worst one by a large margin was the sigmoid kernel, outputting significantly worse numbers, namely between 25% and 46% . The polynomial and linear kernels were closer to the radial one but offered still unsatisfactory results, up to 65% for the linear one and up to 68% for the polynomial kernel.

Since establishing that the radial kernel offers the best results, the most natural course of action is to focus on finding the best parameters for it. The key parameters here proved themselves to be the cost and gamma. After observing an increase and decrease in accuracy depending on the value of these parameters, a trial and error approach was taken in order to find the best one for this dataset. After multiple runs with various combinations, the limit of improvement from this method was reached at a value of 83.97% accuracy with a cost value of 9 and gamma at 0.1. Naturally, more complex methods for this task like a multimodal optimizer could be employed [21], [23], but that would complicate the methodology and is outside the scope of this

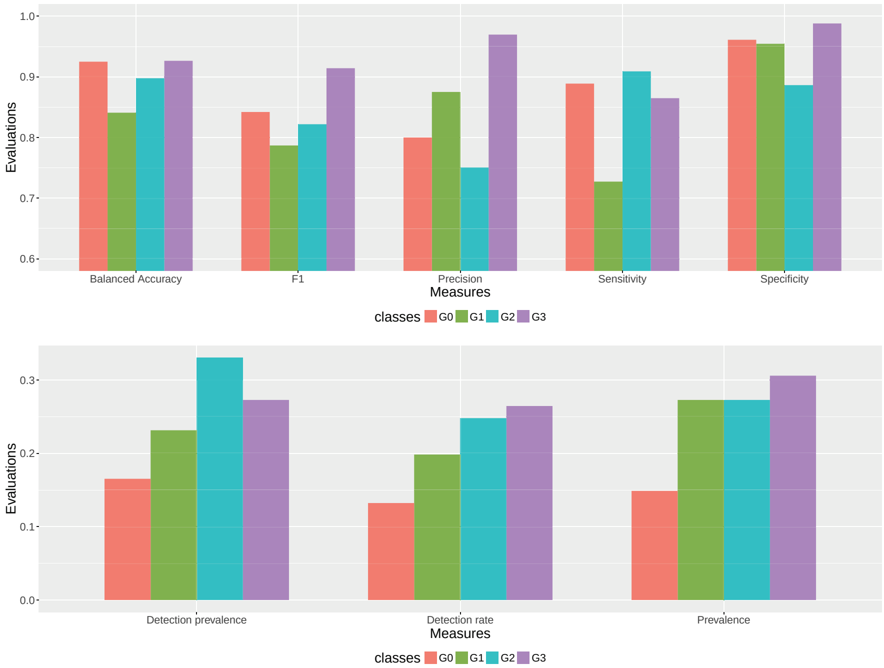

FIGURE 4. Statistical measures computed for each class in turn.

article. These parameters were also tested with the datasets with different thresholding values and showed a major increase in accuracy at around 10% across the board, as pointed out in Figure 3. A notable observation is individual accuracies over 90%, which could point to more potential ways of increasing precision.

Relatively similar approaches that involve the segmentation of the structures in the histological images, feature extraction followed by classification for the same data set are introduced in [25] and [26], which are derived from the study in [31]. However, these procedures assume the specific identification of glands and of nuclei and extract around 80 features from them, that is 10 times more than in the current work. Additionally, they use feature selection and reach accuracies of nearly 80% and 84%, respectively. To give them justice, it has to be mentioned that in these works no special attention was paid to the parameter setting of the classifier, but rather default settings were kept. Nevertheless, the results are very similar to these achieved in the current work, but with less computational effort herein. On a different level, a convolutional neural network is applied in [20] which reached up to 91.44%, further enhanced to reach 92% in [33], and a combination of classifiers in Mathematica [27] applied directly on the images as well achieved the top performance so far of 95.65%. Note however that neither of the latter two approaches do not extract features and are very expensive regarding the computational costs.

Tables 1 and 2, as well as Figure 4, illustrate that the classes that are most accurately classified are G0 and G3, e.g. see F1 and balanced accuracy. The accurate results are encouraging and push the methodology as a relevant one able to provide a second opinion for the medical experts that need to distinguish between the histological slides.

6. Conclusions

The current work puts forward a methodology that makes use of a very limited amount of morphological features that are not directly connected to components like nuclei or glands in the histological images, but rather only to the existing lighter components using intensity-based thresholding. Surprisingly, some simple statistics about the perimeters and areas of those contours, aligned to a proper fine tuning of the classifier, leads to results similar to those of more complex approaches that extract larger amounts of features and are even endowed with feature selection to keep only the most relevant attributes for the classification step.

Besides traditional SVM, other classifiers could be employed in future work for the discrimination between classes based on the extracted features. Among them, an evolutionary algorithm for the optimization of the SVM was illustrated to work well in [32], [34], but an ensemble of different classifiers [18], [35] might also lead to significant improvements in results.

References

[1] M. Arnold, M.S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global patterns and trends in colorectal cancer incidence and mortality, *Gut* **66** (2015) no. 4, 683-691.

[2] R. Bhargava, A.V. Florea, M. Pelmus, M.W. Jones, M. Bonaventura, A. Wald, M. Nikiforova, Breast Tumor Resembling Tall Cell Variant of Papillary Thyroid Carcinoma: A Solid Papillary Neoplasm With Characteristic Immunohistochemical Profile and Few Recurrent Mutations, *American Journal of Clinical Pathology* **147**, no. 4 (2017), 399-410.

[3] R. Bhargava, A. Madabhushi, Emerging themes in image informatics and molecular analysis for digital pathology, *Annual review of biomedical engineering* **18** (2016), 387–412.

[4] K. Bibbins-Domingo, D.C. Grossman, S.J. Curry, K.W. Davidson, J.W. Epling, F.A.. R. García, M.W. Gillman, et al., Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement, *JAMA* **315** (2015), no. 23, 256475. `https://doi.org/10.1001/jama.2016.5989`.

[5] Bureau of Labor Statistics, Resident population of the United States by sex and age as of July 1, 2016 (in millions), *Statista. Web.* 12 April 2018.

[6] P. Cudek, W. Paja, M. Wrzesień, Automatic System for Classification of Melanocytic Skin Lesions Based on Images Recognition. In: *Czachórski T., Kozielski S., Stańczyk U. (eds) Man-Machine Interactions 2. Advances in Intelligent and Soft Computing* **103** Springer, Berlin, Heidelberg (2011), 189–196.

[7] C. Demir, B. Yener, Automated cancer diagnosis based on histopathological images: a systematic survey, Technical Report TR-05-09, Rensselaer Polytechnic Institute (2005).

[8] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D.M. Parkin, D. Forman, F. Bray, GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase no. 11, Lyon, France: International Agency for Research on Cancer (2013).

[9] L. Gartner, *Textbook of Histology, 4th Edition*, Elsevier, 2016.

[10] F. Gorunescu, S. Belciug, Boosting backpropagation algorithm by stimulus-sampling: Application in computer-aided medical diagnosis, *Journal of Biomedical Informatics* **63** (2016), 74-81.

[11] M.N. Gurcan, L.E. Boucheron, A. Can, A. Madabhushi, N.M. Rajpoot, B. Yener, Histopathological Image Analysis: A Review, *IEEE Reviews in Biomedical Engineering* **2** (2009), 147–171.

[12] J. Ho, S.M. Ahlers, C. Stratman, O. Aridor, L. Pantanowitz, J.L. Fine, A.V. Parwani, Can Digital Pathology Result in Cost Savings? A Financial Projection for Digital Pathology Implementation at a Large Integrated Health Care Organization, *Journal of Pathology Informatics* **5** (2014), no. 1, 5, `http://doi.org/10.4103/2153-3539.139714`

[13] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, *In European conference on machine learning* April 1998, Springer, Berlin, Heidelberg (1998), 137–142.

[14] B.L. Iantovics, Cooperative medical diagnosis elaboration by physicians and artificial agents. In *M.A. Aziz-Alaoui, C. Bertelle (eds) From System Complexity to Emergent Properties. Understanding Complex Systems* Springer, Berlin, Heidelberg (2009), 315-339.

[15] D. Iliescu, R. Dragusin, D. Cernea, C. Patru, M. Florea, S. Tudorache, Intrapartum ultrasound - an integrated approach for best prognosis, *Medical Ultrasonography* , **19** (2017), no. 1, 121.

[16] D. Komura, S. Ishikawa, Machine Learning Methods for Histopathological Image Analysis, *Computational and Structural Biotechnology Journal* **16** (2018), 34–42.

[17] M. Kuhn, Building Predictive Models in R Using the caret Package, *Journal of Statistical Software, Articles*, **28** (2008), no. 5, 1–26.

[18] L.I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, Second Edition, John Wiley & Sons, 2014.

[19] J.T. Kwak, S.M. Hewitt, A.A. Kajdacsy-Balla, S. Sinha, R. Bhargava, Automated prostate tissue referencing for cancer detection and diagnosis, *BMC Bioinformatics* **17** (2016), no. 1, 227.

[20] S. Postavaru, R. Stoean, C. Stoean, G. Joya Caparros, Adaptation of Deep Convolutional Neural Networks for Cancer Grading from Histopathological Images. In *I. Rojas, G. Joya, A. Catala (eds) Advances in Computational Intelligence. IWANN 2017*, Springer, Cham, (2017), 38-49.

[21] M. Preuss, C. Stoean, R. Stoean, Niching Foundations: Basin Identification on Fixed-Property Generated Landscapes, *The ACM Proceedings of the 13th annual conference on Genetic and evolutionary computation* (GECCO-2011), Dublin, Ireland (2011), 837-844.

[22] D. Shen, G. Wu, H.-I. Suk, Deep Learning in Medical Image Analysis, *Annual Review of Biomedical Engineering* **19** (2017), 221-248.

[23] C. Stoean, M. Preuss, R. Stoean, D. Dumitrescu, EA-Powered Basin Number Estimation by Means of Preservation and Exploration, Parallel Problem Solving from Nature (PPSN X), *Lecture Notes in Computer Science* **5199**, Springer Berlin / Heidelberg (2008), 569–578.

[24] C. Stoean, R. Stoean, A. Victor Sandita, D. Ciobanu, C. Mesina, Colorectal cancer histopathological image data set, DOI: `https://doi.org/10.6084/m9.figshare.4508672.v1` (2016).

[25] C. Stoean, R. Stoean, A. Sandita, D. Ciobanu, C. Mesina, C. L. Gruia, SVM-Based Cancer Grading from Histopathological Images using Morphological and Topological Features of Glands and Nuclei. In *G. Pietro, L. Gallo, R. Howlett, L. Jain (eds) Intelligent Interactive Multimedia Systems and Services 2016. Smart Innovation, Systems and Technologies* **55**, Cham, Springer (2016), 145-155.

[26] C. Stoean, In Search of the Optimal Set of Indicators when Classifying Histopathological Images, *IEEE Post-Proceedings of the 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, Timisoara (2016), 449-455.

[27] C. Stoean, D. Lichtblau, Classifier Result Aggregation for Automatically Grading Histopathological Images, *19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, IEEE Computer Society, Timisoara, in press.

[28] C. Stoean, R. Stoean, Evolution of Cooperating Classification Rules with an Archiving Strategy to Underpin Collaboration. In *H. N. Teodorescu, J. Watada, L. C. Jain (eds) Intelligent Systems and Technologies. Studies in Computational Intelligence* **217** Springer, Berlin, Heidelberg (2009), 47–65.

[29] C. Stoean, R. Stoean, Post-evolution of variable-length class prototypes to unlock decision making within support vector machines, *Applied Soft Computing* **25** (2014), 159–173.

[30] C. Stoean, R. Stoean, A. Sandita, C. Mesina, C.L. Gruia, D. Ciobanu, How Much and Where to Use Manual Guidance in the Computational Detection of Contours for Histopathological Images?, *Soft Computing*, `https://doi.org/10.1007/s00500-018-3029-9`.

[31] C. Stoean, R. Stoean, A. Sandita, C. Mesina, D. Ciobanu, C.L. Gruia, Investigation on Parameter Effect for Semi-Automatic Contour Detection in Histopathological Image Processing, IEEE

Post-Proceedings of the 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2015), IEEE Computer Society, Timisoara, Romania (2015), 445–452.

[32] C. Stoean, R. Stoean, *Support Vector Machines and Evolutionary Algorithms for Classification - Single or Together?*, Intelligent Systems Reference Library, Springer, **69**, 2014.

[33] R. Stoean, Analysis on the potential of an EAsurrogate modelling tandem for deep learning parametrization: an example for cancer classification from medical images, *Neural Computing and Applications* (2018), https://doi.org/10.1007/s00521-018-3709-5

[34] R. Stoean, C. Stoean, M. Preuss, D. Dumitrescu, Evolutionary Multi-class Support Vector Machines for Classification, *International Journal of Computers, Communications & Control, Supplementary Issue*, International Conference on Computers and Communications - ICCC 2006, Baile Felix Spa - Oradea, Romania (2006), 423–428.

[35] R. Stoean, C. Stoean, A. Sandita, D. Ciobanu, C. Mesina, Ensemble of Classifiers for Length of Stay Prediction in Colorectal Cancer, *Advances in Computational Intelligence, LNCS*, Springer, **9094** (2015), 444–457.

[36] The World Factbook 2018. Washington, DC: Central Intelligence Agency, 2018.

(Dumitru Bogdan Mateis Boruz, Catalin Stoean) Department of Computer Science,
University of Craiova, 13 A.I. Cuza Street, Craiova, 200585, Romania
E-mail address: bogdan.mateis@yahoo.com, catalin.stoean@inf.ucv.ro