How to cite this paper:

Albashish, D., Sahran, S., Abdullah, A., Alweshah, M., & A., Adam. (2018). A hierarchical classifier for multiclass prostate histopathology image gleason grading. *Journal of Information and Communication Technology, 17 (2),* 323-346.

A HIERARCHICAL CLASSIFIER FOR MULTICLASS PROSTATE HISTOPATHOLOGY IMAGE GLEASON GRADING

[1]Dheeb Albashish, [2]Shahnorbanun Sahran, [2]Azizi Abdullah, [3] Mohammed Alweshah & [2]Afzan Adam

[1&3] Prince Abdullah Ben Ghazi Faculty of Information Technology
Al-Balqa Applied University, 19117 Al-Salt, Jordan
[1&2] Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, Selangor, Malaysia

bashish@bau.edu.jo;{shahnorbanun@bau.edu.jo; azizia@bau.edu.jo; afzan@ukm.edu.my; weshah@bau.edu.jo

ABSTRACT

Automated classification of prostate histopathology images includes the identification of multiple classes, such as benign and cancerous (grades 3 & 4). To address the multiclass classification problem in prostate histopathology images, breakdown approaches are utilized, such as one-versus-one (OVO) and one-versus-all (Ovall). In these approaches, the multiclass problem is decomposed into numerous binary subtasks, which are separately addressed. However, OVALL introduces an artificial class imbalance, which degrades the classification performance, while in the case of OVO, the correlation between different classes not regarded as a multiclass problem is broken into multiple independent binary problems. This paper proposes a new multiclass approach called multi-level (hierarchical) learning architecture (MLA). It addresses the binary classification tasks within the framework of a hierarchical strategy. It does so by accounting for the interaction between several classes and the domain knowledge. The proposed approach relies on the 'divide-

Received: 2 September 2017 **Accepted:** 4 January 2018

and-conquer' principle by allocating each binary task into two separate subtasks; strong and weak, based on the power of the samples in each binary task. Conversely, the strong samples include more information about the considered task, which motivates the production of the final prediction. Experimental results on prostate histopathological images illustrated that the MLA significantly outperforms the Ovall and OVO approaches when applied to the ensemble framework. The results also confirmed the high efficiency of the ensemble framework with the MLA scheme in dealing with the multiclass classification problem.

Keywords*:* Multiclasss classification, hierarchical classification, image classification, ensemble classification**.**

INTRODUCTION

Prostate cancer is an affliction that is becoming increasingly common in elder males. The most important stage of prostate cancer diagnosis is the examination of the microscopic images of biopsy specimens. This approach is favored due to the flexibility they afford the pathologist, who can manipulate and examine tissue images on a monitor and assign a grade to the cancer. The most widespread system for histological image grading of prostate is the Gleason grading system (Tabesh et al., 2007).

The Gleason grades determine the degree of malignancy of prostate cancer based on the pattern(s) present in the ROI (Tabesh et al., 2007). The Gleason grade system taxonomize the prostate tumor using Grades 1 - 5. Higher Gleason grades means a higher malignancy level, and vice versa. Practically, Grades 1 & 2 are considered as benign tissue (Farjam, Hamid, Kourosh, & Reza, 2007), where the glands are a well-differentiated, while Grades 3, 4, & 5 are considered as a malignant tissue. Specifically, Grade 3 has small and smooth tissue components of the gland that infiltrates the stroma (Epstein, 2010), while the glands of Grade 4 are fused; they are not well-disjointed by stroma. Finally, Grade 5 corresponds to a poorly differentiated tumor. However, it should be pointed out that it is uncommon in prostate tissue databases (Nguyen, Sarkar, & Jain, 2014). Indeed, both malignant (Grades 3, 4, 5) and benign (Grade 1, 2) tissue images share similar colors, but its distribution gradually diverges. In malignant tissue images, the texture characteristics of the surface are finer than the benign tissue; this is due to the spread of the nuclei tissue component in the malignant images. Subsequently, this paper concentrates on the three-class classification issue: Grades 3, 4 and benign prostate tissue types, and compare them to the multiclass classification problem in machine learning.

The multiclass classification problem was solved in machine learning based on two categories. The first category combines all the classification constraints simultaneously within a single-classifier method, such as k-nearest neighbors (kNN), the decision trees (DT), and artificial neural network (ANN). However, such methods significantly increases complexity, which in turn requires advanced optimization techniques (Honeine, Noumir, & Richard, 2013). The second category is called the decomposition strategy, which divides the multiclass classification problem into several binary sub-problems. All these sub-problems are solved separately, and their results are combined to elucidate the final result. The most popular strategies in decomposing a multiclass problem into multiple binary ones are the one-versus-all (Ovall) (Honeine et al., 2013), the one-versus-one (OvO) approaches (Zhang et al 2016) and hierarchal approaches (Silla & Freitas, 2011). They mainly utilize the binary classifier scheme, such as support vector machines (SVM), linear discriminant analysis (LDA), and logistic regression (LR), which were originally designed for binary classification tasks. The OvO category is more capable of discriminating the multiclass classification problems compared to the Ovall (Zhang et al., 2016). Although Ovall and OvO approaches are used in prostate histopathology multiclass classification, they fail to take into account the information supplied by the pathologists. For instance, cases of Grade 3 vs. Grade 4 and Benign vs. Grade 3 share a specific characteristic in the classification task. They also suffer from numerous limitations. For example, in the case of Ovall, an artificial class imbalance is introduced, which degrades the classification performance, while in OVO, the correlation between different classes not classified as a multiclass problem is broken into multiple independent binary problems. Thus, the hierarchical approaches are utilized instead of OvO and Ovall in multiclass classification.

In the state-of-art of the multiclass classification, the hierarchical classification approaches have also been introduced for addressing the multiclass classification tasks (Silla & Freitas, 2011). The hierarchical classification involves a tree structure of $(m - 1)$ binary classifiers rather than a single layer of m classifiers (in the Ovall approach) or $\frac{m(m-1)}{2}$ classifiers (in the OvO approach) (Casasent & Wang, 2005). Thus, the number of required calculations is significantly reduced in this approach.

This paper focuses on solving the three-class classification problem in prostate cancer grading, i.e. Grade 3, Grade 4, and Benign. It starts by proposing a new hierarchical multiclass approach called multi-level learning architecture (MLA), which addresses the binary classification tasks in the hierarchical strategy. It does so by taking into account the domain knowledge and the

correlation between different classes. MLA also relies upon the divide-and-conquer' concept, and work by dividing each binary task into strong and weak sub-tasks based on the power of the samples in each original binary task. This help the strong samples predict the final result, as they have more information about the task. The proposed MLA is applied within our previous ensemble framework (Albashish et al. 2016) for prostate grading.

RELATED WORK

These histopathology images allow the pathologist to examine and manipulate tissue images on a monitor and assign a grade to the cancer region that represents its level of malignancy. This procedure is referred to as histopathological image grading.

The conventional pathologist's visual examination of histopathology images and assigning of Gleason grading face a number of difficulties, i.e. the heterogeneous nature of the tissue structure (nuclei, lumen, stroma, cytoplasm), thus, not all images belonging to a certain grade look similar (Al-Kadi, 2010). Moreover, the diagnosing performance depends on the personal experience and the skills of the pathologist. Also, examining the tissue under a microscope is time-consuming and tedious.

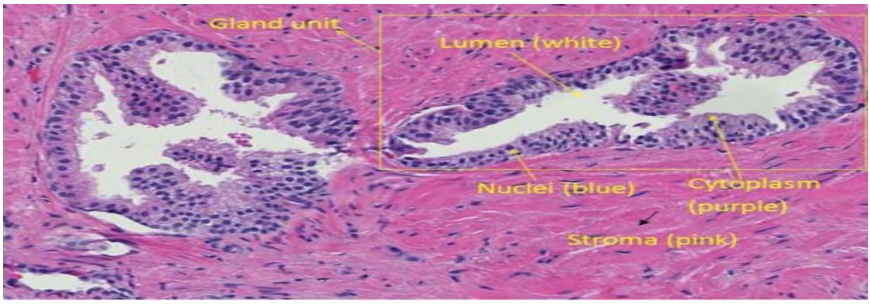

Figure 1. Cyber risk assessment (CRA) conceptual framework.

In most works on digital pathology CADs for prostate cancer grading, there are two approaches based on feature description: tissue-structure-based and texture-based CAD systems (Mosquera-Lopez, Agaian, Velez, & Thompson, 2015). The texture-feature CADs employ quantities of spatial variation in pixel intensities to characterize grade patterns, while in the tissue-structure CADs, the structure features are computed from the measurements of size, shape, and tissue structures, such as nuclei lumen and cytoplasm to help distinguish between different Gleason grades.

In the existing CADs of PCa grading based on texture features, the most widely seen features are co-occurrence(Haralick & Shanmugam, 1973), first-order statistics, and fractal analysis and wavelet [13,14]. For instance, Tabesh et al (Tabesh et al., 2007) combined the local and image texture features of the tissue images captured at 20X magnifications. They extracted the local features from the tissue components, such as the nuclei and lumen, and extracted the texture, fractal, and color features from the images. The developed CAD obtained a rate of 96.7% accuracy for tumor vs. non-tumor classification, and 81% accuracy for high vs. low-grade classification. However, these accuracies are only realized when they are applied on small spots of an image.

Usman et al. (Ali, Shaukat, Hussain, Ali, & Khan, 2016) used Discrete Wavelet Packet Decomposition (DWPD) and gray level co-occurrence matrix (GLCM) features from the gray level histogram image to identify tissue regions as either Grades 3, 4, or 5 (a three-class problem). The experimental result on 13 test prostate tissue images using SVM reported an accuracy of 92.4%. However, the developed CAD suffered from the curse of dimensionality problem due to its usage of many features.

The second approach in CADs of PCa grading is based on the tissue-structure. In this approach, the main tissue components (e.g., lumen, nuclei, cytoplasm, and stroma) or glands (as illustrated in Figure 1) are first segmented. After that, the features are extracted from individual tissue segments. For instance, Naik et al. (Naik et al., 2008) used the level set algorithm to extract the glands from the image. Then, the lumen area and various shape feature were extracted from each gland. The experiment result on 44 images reported a 95.1% accuracy for Grade 3 vs. Grade 4. However, the number of samples were small compared to the number of the extracted features. They ignored the nuclei structure, which is an important component for determining the Gleason patterns. A segmentation-based method was proposed in (Nguyen, Sabata, & Jain, 2012b). They first used K-means to identify the main tissue components (lumen, nuclei, cytoplasm, and stroma). Then, they extracted the structure and contextual features from the glands. These features include the lumen area, nuclei mean and standard deviation while the contextual features include the shape and size similarity and the neighborhood crowdedness. A tissue image is classified into cancer vs. noncancerous glands via the SVM classifier. By using a dataset that includes 48 images, they reported a result of 79% accuracy. However, the manual labelling of the glands was time-consuming and tedious for the pathologists. Moreover, all of these methods are used to detect the cancer, but cannot be used to discriminate between the Gleason grades due to the glands of higher grades, such as Grade 4, are meager, which makes their structural features inscrutable.

Although the state-of-the-art tissue structure-based CAD systems rely on the presence of tissue components, some basic tissue components, such as lumen, are occluded by cytoplasm (Nguyen, Sabata, & Jain, 2012a); consequently, accurate high- level features measurement cannot be acquired. In our previous researches, this limitation was circumvented by introducing an ensemble framework (Albashish et al., 2016), which was based on the texture features of the tissue components (mainly, lumen, cytoplasm, nuclei, and stroma). By using a dataset of 97 images at 40X magnification, this framework achieved 93.59% AUC performance for Grade 3 vs. Grade 4, which is the most challenging classification task in this domain. However, this framework was applied for the discrimination between two classes only. This paper proposes a new extension in order to address the multiclass classification task (three-class classification).

The multiclass classification problem in prostate CADs involve more than two classes (Nguyen et al., 2014)[17,18], which is more difficult than the two-class classification problem. This is because the decision function of a multiclass classification task tends to be more complicated than that of two-class classification task (Kang, Cho, & Kang, 2015), there are two techniques for addressing the multiclass classification problem in machine learning. On one hand, some classifier algorithms are direct. There are two techniques for addressing the multiclass classification problem in machine learning. On one hand, some classifier algorithms directly address the multiclass classification problems, such as artificial neural networks (ANN) and k-nearest neighbors (k NN), while on the other, it addresses the binarization techniques. These techniques divide the multiclass classification task into sets of two-class classification sub-tasks, then it learns a different binary classifier model for each new sub-task, and finally combine the result using a specific strategy to obtain a final result. The binarization technique solves several easier problems instead of a single complex problem.

Regarding the binarization strategy, the two popular used approaches are ‘‘one-vs-all” (Ovall) and ‘‘one-vs-one’’ (OVO) (Galar, Fernández, Barrenechea, Bustince, & Herrera, 2011). These strategies use the binary classification model to effectively deal with a multiclass problem. They seek to address several binary sub-problems, which are simpler than the original problem (Kang et al., 2015).

In the OVO approach, an m -class problem is decomposed into $\frac{m(m-1)}{2}$ binary sub-problems, as illustrated in Figure 2. Each independent binary problem is solved by a binary classifier, which is responsible for discriminating between each pair of classes. The training data for each pair of classes are those samples

from the original multiclass dataset, whose class label belongs to one of both classes and the samples with different class labels are neglected in this binary problem. With such an appropriate consideration, the multiclass classification problem is divided into simpler binary sub-problems, which are expected to address the multiclass task using binary classifier models (Galar, Fernández, Barrenechea, et al., 2011).

$$\begin{pmatrix} - & r_{12} & \cdots & r_{1m} \\ r_{21} & - & \ldots & r_{2m} \\ \vdots & & & \vdots \\ r_{m1} & r_{m2} & \ldots & - \end{pmatrix}$$

In order to predict a new sample (the testing phase), each base classifier model produces a prediction for a test sample. The prediction output for each pairwise classes (*i, j*) is given by $r_{ij} \in [0,1]$, where the r_{ij} is the confidence degree of the favorite class *i* given by a base classifier. The confidence in favor of class j is computed by $r_{ji} = 1 - r_{ij}$, while the confidence degrees for all base classifier can be represented by a score-matrix R, as shown in Equation (1).

Finally, an aggregation strategy is used to infer the output class for a tested pattern from the aforementioned R (Equation (1)). There are a number of aggregations strategies proposed in literature. The powerful and simplest strategy is the Max-Wins rule, which is also called the voting strategy (Galar, Fernández, Bustince, & Herrera, 2011). It considers the output (vote) of each classifier. Each classifier produces a vote for the predicted class; then, the votes received by each class are counted. The final prediction for the tested pattern is the class that has the largest number of votes, as illustrated in Equation (2).

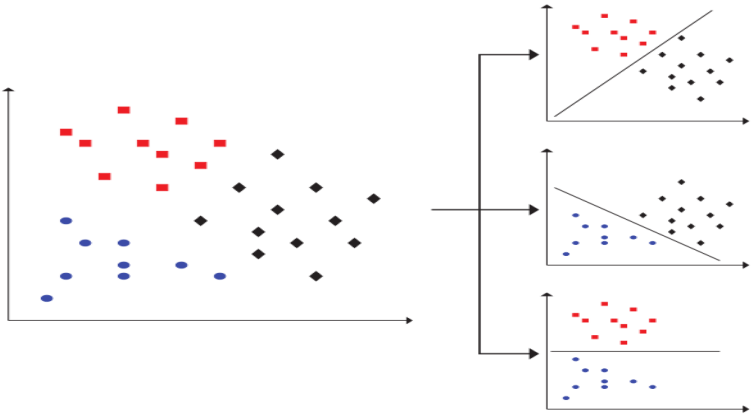

Figure 2. An example of OVO approach for decomposition of a three-class problem into binary class sub-problems (Zhang et al., 2016).

$$class = \arg\max_{i=1,\ldots,m} \sum_{1 \le i \ne j \le m} S_{ij},$$

where,

$$S_{ij} = \begin{cases} 1 & r_{ij} > r_{ji} \\ 0 & otherwise \end{cases}.$$

The OVO approach has been applied in PCa grading CADs to classify multiple Gleason grades or differentiate between different groups of prostate glands based on the prostate histopathology images. Most of these studies utilized LibSVM (Chang & Lin, 2011) as a multiclass classifier tool for SVM and the OVO approach. The authors in (Nguyen et al., 2014) and (Nguyen et al., 2012a) solved the three-class problem, which included the most common grades among the tissues, namely normal (Benign), Grade 3, and Grade 4, based on the LibSVM tool, while the authors in (Huang & Lee, 2009) utilized the LibSVM tool to differentiate between Gleason grades 1, 3, 4, and 5. In (Ali et al., 2016), they used OVO to distinguish between Grades 3, 4, and 5; they reported an overall accuracy of 92.2% for these three classes. Moreover, in (Khelifi, Adel, & Bourennane, 2012), the OVO approach was used to discriminate between stroma, benign, prostatic intraepithelial neoplasia, and prostatic carcinoma. Although OVO is utilized to solve the multiclass classification problem, it failed to capture the correlation between the different classes.

The second well-known approach for solving the multiclass classification is the one-against-all (Ovall) approach, which is also known as the one-versus-rest approach (Anand, Mehrotra, Mohan, & Ranka, 1995). In Ovall, the multiclass problem is decomposed into m two-class sub-problems (Figure 3). Let $C = \{c_1,...c_i ..., c_k]$ denote the set of k classes. For the k multiclass problem, k two-class classifier models (e.g. SVM) are constructed $km = \{m_1, ... m_i, ...m_k]$, where model is trained to separate the class m_i (positive class) from the rest classes $\{C - m_i]$ (negative class). Figure 3 displays how Ovall decompose a three-class problem into two-class sub-problems. In the test phase, a test pattern x is presented to each binary classifier model $m_i \in km$; the positive value of the decision function of the model m_i ($f_i(x)$) indicates that the test pattern x belongs to the class c_i, while the negative value of ($f_i(x)$) indicates that the test sample x belongs to the rest classes. In Ovall, the voting scheme is used where each classifier m_i votes for the corresponding class c_i based on its output (i.e. the posterior class probability) value $f_i(x)$, i=1,…, k. Unlike the OVO approach that has a scoring matrix, Ovall has a scoring vector Equation (4) that represents the values of the decision functions for all of the models (Moustakidis & Theocharis, 2012).

The final prediction for the test sample x is assigned to the class with the maximum decision value as illustrated in Equation (5)

In CADs of PCa grading, some studies employed the OVA approach to classify the multiple Gleason grades or different groups in the prostate histopathology images. For example, Doyle et al. (Doyle et al., 2012) compared the OVA approach and their proposed cascade approach to discriminate between seven classes, including Benign and Grades 3, 4, 5; the experiment results showed that OVA outperformed their cascade. In (Almuntashri et al., 2011), the Ovall approach was used to discriminate between the Gleason Grades 3, 4, and 5. However, in this study, Grade 4 reported the lowest performance.

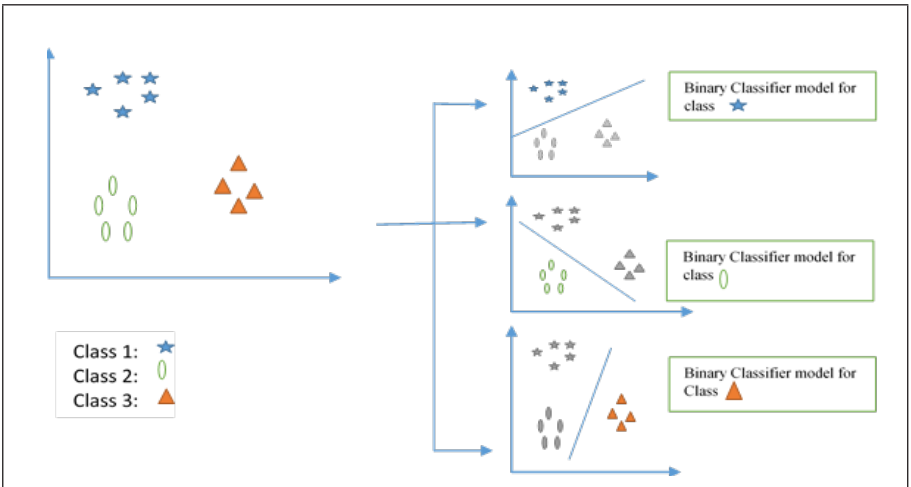

Figure 3. An example of Ovall approach for decomposition of a three-class problem into binary class sub-problems.

EXTENSION OF THE ENSEMBLE FRAMEWORK TO MULTI-CLASS CLASSIFICATION PROBLEM

Based on the OVO and Ovall approaches(Galar, Fernández, Barrenechea, et al., 2011) discussed in the previous section, our latest published binary ensemble framework (Albashish et al., 2016) can be extended to address the three-class problem in PCa grading. Based on these reported approaches, two kinds of extensions for the ensemble framework can be introduced. One utilizes the OVO multiclass within the ensemble framework called ensemble_OVO, while the other utilizes the Ovall multiclass within ensemble framework called ensemble_OVall.

The general structure of the ensemble_Ovall is shown in Figure 4. In this ensemble, four independent three-class classification problems based on the number of the tissue components (lumen, nuclei, cytoplasm, and stroma) are created, where each tissue component has one corresponding three-class problem. Each three-class problem is solved using the OvA approach. Particularly, three binary classifier models (e.g., SVM) for training are built; Benign (positive) versus all other classes (negative), Grade 3 versus all other classes, and Grade 4 versus all other classes. From this three-decision functions, the posterior class probabilities can be estimated and used to classify a new sample using the Maximum-a-posterior rule (Hsu & Lin, 2002). Generally, the combination of individual outputs in Ovall is created by determining the closest class. Then, the decisions of the four Ovall approaches are combined to produce the global decision for the ensemble_Ovall. The ensemble_Ovall combines the four OVall results in order to obtain the final decisions by using the majority voting strategy.

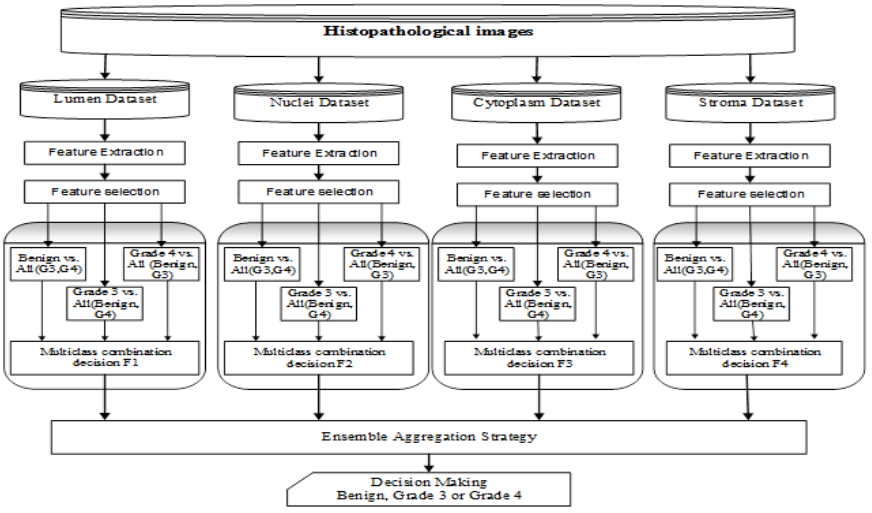

Figure 4. The architecture of the Ovall with the ensemble framework (ensemble_Ovall) for addressing the three-class classification problem in PCa grading.

The second extension for the ensemble framework is the ensemble_OVO (as shown in Figure 5). Similar to the ensemble_Ovall, it consists of four independent three-class classification problems based on the number of the tissue components, where each tissue component has one corresponding three-class problem. Each three-class problem is addressed using the OVO approach. Particularly, the $3(3-1)/2$ classifier models (e.g., SVM) are built in the training phase; Benign vs.

Grade 3, Benign vs. Grade 4, and Grade 3 vs. Grade 4. The output result for each OVO approach is obtained via the simplest strategy called Max-Win voting strategy (Galar, Fernández, Barrenechea, et al., 2011). The Max-Win strategy consider a vote for the output (predicted) class produced by the binary classifier. Then, votes received by each class are counted, the final output for OVO is the final class that obtains the largest number of votes. The decisions of the individual four OVO approaches are then combined to create the global decision for the ensemble_OVO. The ensemble_OVO applies the simple majority voting strategy to obtain the final decision for the four OVO results.

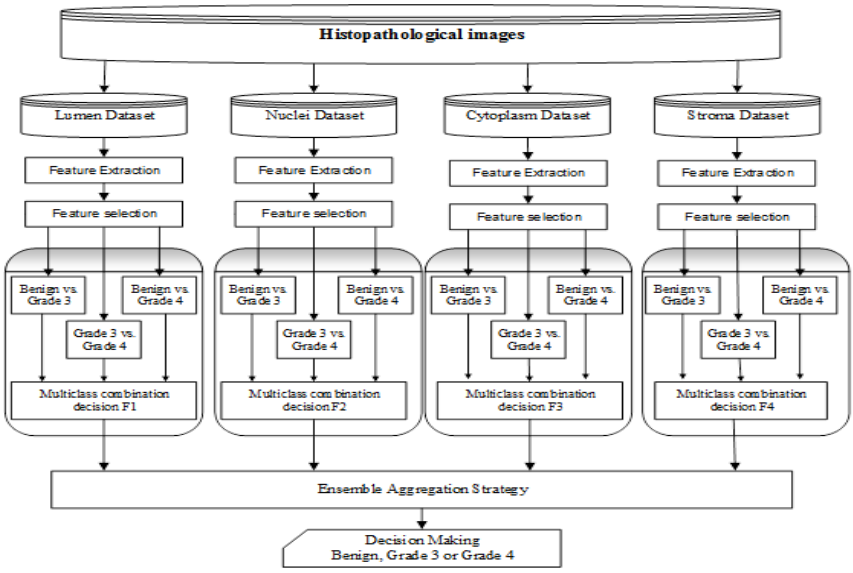

Figure 5. The architecture of the OVO with the ensemble framework (ensemble_OVO) for addressing the three-class classification problem in PCa grading.

THE PROPOSED MULTI-LEVEL LEARNING APPROACH

In this section, a new supervised multi-level (hierarchical) learning architecture (MLA) for solving the three-class classification problem in prostate based on the histopathology images is introduced. This architecture is based on

divide and conquer concept. In MLA, the three-class classification problem is decomposed into series of binary classification sub-problems based on the domain knowledge and organized in a multi-levels scheme.

The MLA classify binary sub-problems into a set of nodes, where each node represents a classifier model for a pair of meta-classes. The root node of the MLA includes the simplest discriminative classification task, Benign vs. Grade 3 based on the domain knowledge. The next levels represent the less discriminative binary tasks, Benign vs. Grade 4 and Grade 3 vs. Grade 4. Each binary task is divided into two subtasks (e.g., strong and weak) based on the power of the samples. In level 2, the classifier models are constructed based on the strong samples, while in level 3, the models are constructed based on the weak samples.

One key issue in MLA is how to partition the samples of each class into strong and weak subsets. One of the measurements used to gauge information in a sample is its entropy (Shannon & Warren, 2001). It measures the amount of information that an event (i.e., instance) provides. Assume that we have an instance Y, which contains D features $Y = \{f_1, f_2, ..., f_D\}$. Then, the entropy $H(Y)$ or the amount of information (in bit) in this instance is measured using Equation (6), where $p(y)$ is the probability of the value y. If the entropy of an instance Y is less than a threshold, then this sample is dragged into the strong subset, or else, it belongs to the weak subset. The threshold for a class k is computed based on the average of entropies over all samples in this class (see Equation (7)),

$$H(Y) = -\sum_{y \in Y} p(y) \log_2(p(y), \tag{6}$$

where $\frac{1}{n}\sum_{i=1}^{n} H(Y_i))$ denotes the average of entropies over all of the samples (i.e., n samples) in class k. By using Equation (7), the samples of each class are divided into two separate subsets (i.e., strong and weak). Therefore, the dataset of each pairwise classes of Grade 3 vs. Grade 4 and Benign vs. Grade 4 are partitioned into two non-overlapping subsets (i.e., strong and weak). These two subsets can be used to train strong and weak models in levels 2 and 3, respectively.

To demonstrate the concept of the MLA, the researchers gave an example of how it can address the three-class classification problem in PCa grading. Both training and testing are illustrated in Figure 6. During training, the whole dataset is divided into three binary classification tasks into pairs of classes. However, the main distinction from the OVO approach is that the

MLA indicates the interclass relationship among different classes. Starting from the root node (level 1), a model representing the Benign vs. Grade 3 is constructed. Then, at level 2, strong models are constructed for Benign vs. Grade 4 and Grade 3 vs Grade 4 in the left and right nodes, respectively. Finally, in level 3, the weak models for these pairs of classes are constructed. As shown in Figure 6, the Benign vs. Grade 4 and Grade 3 vs. Grade 4 are classified based on two models instead of one in OVO and Ovall, indicating the accurate nature of MLA. Moreover, reducing the number of the samples in the training of each model decreases the training time (Kumar & Gopal, 2011). However, decreasing the number of samples while the number of features remain large could lead to overfitting problem. Thus, the researcher uses the FS method prior to training the MLA.

$$\begin{Bmatrix} weak\ sample, & H(Y) > \frac{1}{n}\sum_{i=1}^{n} H(Y_i) \\ strong\ sample, & H(Y) \leq \frac{1}{n}\sum_{i=1}^{n} H(Y_i) \end{Bmatrix}$$

With regards to the evolution of the MLA on an input sample $x \in Xx \in X$, the process starts at level 0 (root node), where the SVM model for Benign vs. Grade 3 is evaluated. The root node then exits via the left edge if the prediction is not Benign, or the right edge if the prediction is not Grade 3.

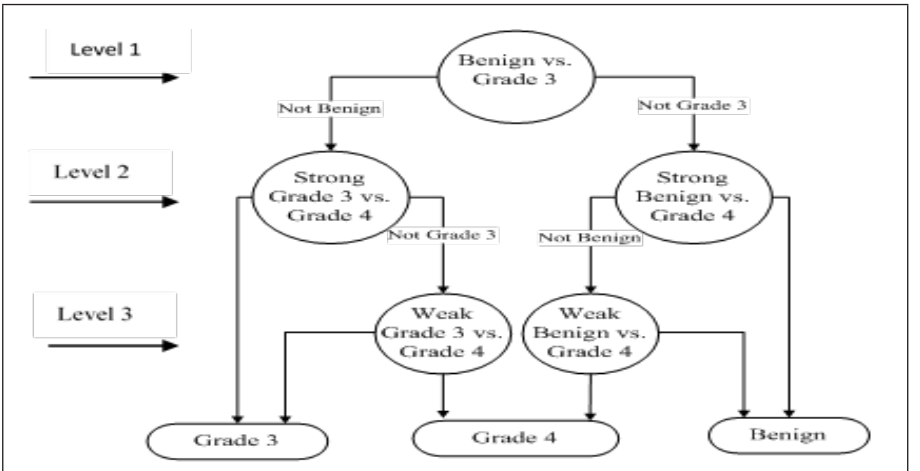

Figure 6. The structure of the MLA for solving the three-class classification of PC a grading.

Assume, for example, if the root predicts not Benign. Then, in the next level (level 1), the left node includes a strong predictor, which distinguishes between Grade 3 versus Grade 4. If the result, for example, is Grade 3, then we reach the leaf node, which is the final decision for the MLA. However, if the result is Grade 4, the evolution path goes to the next level (level 3), which includes the weak predictor for Grade 3 versus Grade 4 to redo the binary task. The value of the weak predictor is accepted as the final decision, which is either Grade 3 or 4. As evident from the evolution phase, each class in the MLA is checked at least twice. Therefore, this provides increased accuracy for the proposed multiclass classification approach.

The proposed MLA is extended for use with the ensemble framework (coded as ensemble_MLA). The ensemble_MLA is also made up of four multiclass classification tasks based on the number of tissue components in the ensemble framework. Each multiclass classification task is solved based on the MLA (see Figure 7). In testing a new pattern $x \epsilon X$ using the MLA, the final classification decision is given by the classification of the final leaf node classifier, which is either a strong or a weak classifier. Finally, the global decision for the ensemble_MLA for a test image is obtained via majority voting strategy, which was employed in the previous ensemble_OVO and ensemble_OVall.

EXPERIMENTAL RESULTS AND DISCUSSION

Dataset

To evaluate the efficiency of the proposed ensemble_MLA and state-of-the-art ensemble_OVall and ensemble_OVO in solving the three-class (Benign, Grades 3 and 4) classification problem in the PCa grading, two histopathology image datasets are used. The first dataset (Dataset I) is the self-collected datasets that had been collected by three pathologist from HUKM medical center (Malaysia). This dataset was created by selecting sub-images of Grade 3, benign, and Grade 4 carcinoma patterns at 40X magnification. The size of each sub-image is ~4140 x 3096 pixels. There are 52 sub-images of benign patterns, 41 sub-images of Grade 3 carcinoma, and 56 sub-images of Grade 4 carcinoma. These grades were confirmed by the three pathologists, rendering the results reports by this experiments clinically acceptable. The second prostate histopathology dataset (Dataset II) reported in (Farjam et al., 2007) and (Farjam, Soltanian-zadeh, Zoroofi, Ford, & System, 2004). This dataset contains 65 tissue region images ranging from Benign (Gleason Grades 1 and 2) to Gleason Grades 3 and 4. The number of samples categorized into these three classes are 19, 20 and 25, respectively. These tissue region images

have similar illuminations and magnifications. They were captured at *100X* magnification. In this research, the Gleason Grade 5 was excluded since it is not commonly present in prostate tissue databases (Nguyen et al., 2014). Table 1 shows the details of the two datasets.

Table 1

Basic Information of the Datasets

Dataset	# of Benign	# of Grade 3	# of Grade 4
Dataset I	52	41	56
Dataset II	19	20	25

Experiment Setup

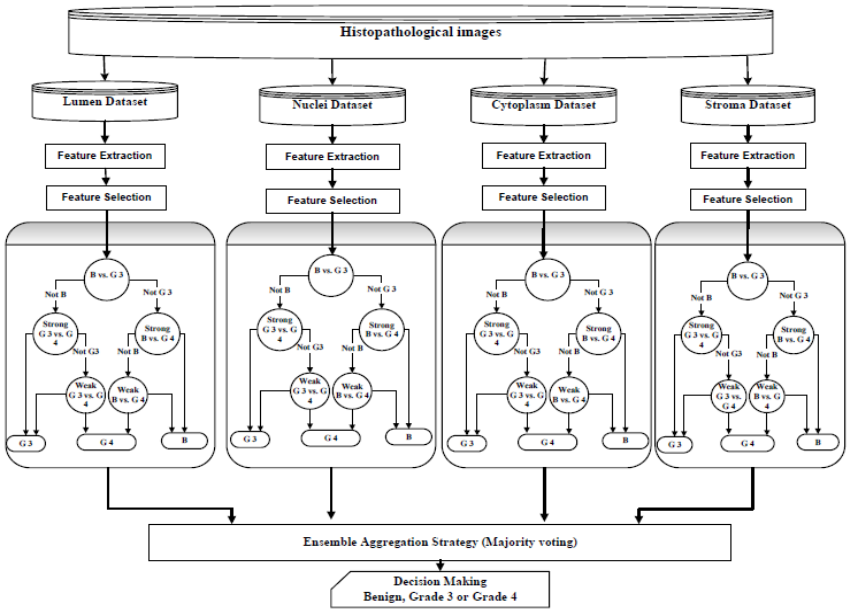

Figure 7. The architecture of the MLA with the ensemble framework (ensemble_MLA) for addressing the three-class classification problem in PCa grading, B: Benign, G3: Grade 3, and G4: Grade4.

In order to evaluate the efficiency of the proposed ensemble_MLA and state-of-the-art ensemble_OVall and ensemble_OVO in solving the three-class

(Benign, Grades 3 and 4) classification problem in the PCa grading, the histopathology image dataset is used. Each dataset is divided into four sub-datasets based on number of tissue components (lumen, nuclei, cytoplasm, stroma), where each tissue component (sub-dataset) is represented by 105 co-occurrence and 81 HOG features. When the ensemble_Ovall is tested, four multiclass Ovall approaches are used based on the number of tissue components, where each multiclass Ovall reports three SVMs. The final decision for the ensemble is computed via the aggregation of the four Ovall approaches based on majority voting.

The feature selection method is used as a pre-processing step before the multiclass approaches are applied. In this study, the extensions of SVM-RFE (Zhou & Tuck, 2007) and our previous feature selection method (Albashish et al., 2015) are used , and the top 50 features are selected for each multiclass approach. The SVM implementation is utilized by the LibSVM toolbox (Chang & Lin, 2011) , and the C and γ in the SVM are estimated using a grid search with different internal threefold cross-validations on the training dataset from the set $\{2^{-20},... 2^{1}\ldots, 2^{20}\}$.

Experiment Results and Discussion

The performance of the ensemble_OvO, ensemble_ovall, and ensemble_MLA/MV on two prostate histopathology image datasets for solving the three-class problem in PCa grading is measured in terms of accuracy, sensitivity, and specificity as tabulated in Figure 8, and Tables 2 and 3.

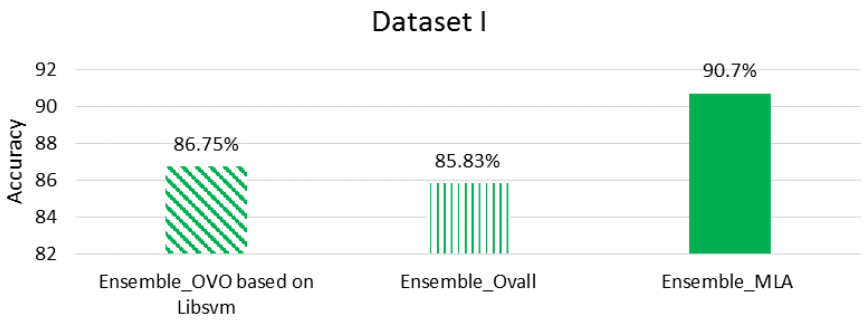

Figure 8 Average accuracy of ensemble_OvO, ensemble_ovall, and ensemble_MLA on the dataset I over 30 runs.

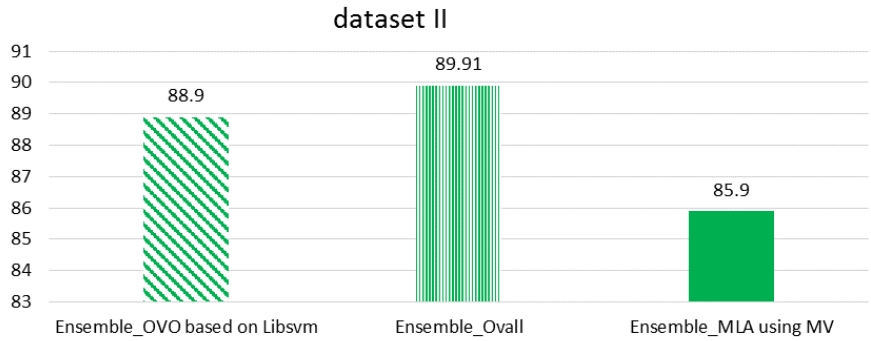

Figure 9. Average accuracy of ensemble_OvO, ensemble_ovall, and ensemble_MLA on the dataset II over 30 runs.

Figure 8 shows the overall accuracy of the comparison methods when they applied on the first dataset. At first glance, the proposed ensemble_MLA significantly outperformed the other approaches.

In support of this observation, comparing it to ensemble_OvO, the accuracy reported an increase of ~3.95%, and comparing it to the ensemble_Ovall, the proposed ensemble_MLA indicated an increased accuracy of ~4.87%. These values are attributed to the interaction between the classes utilized by the proposed MLA.

The classification accuracy of the ensemble_OvO remained very close to the ensemble_Ovall in the dataset I and dataset II as illustrated in Figures. 8 and 9, respectively. However, ensemble_MLA outperformed them in dataset I while in dataset II, the ensemble_MLA achieved the lowest accuracy. One reason for this low accuracy is the number of samples that were used in each binary classification task in the ensemble_MLA. Practically, number of training samples effect the performance of the classification (Yu, Yang, Wang, & Han, 2003).

The second measurement is the sensitivity, which determines the probability of the results that are true positive. Table 2 shows the sensitivity performance comparison of the ensemble_Ovall, ensemble_OVO, and the proposed ensemble_MLA. It can be seen that the overall sensitivity of the ensemble_MLA significantly outperformed the baseline ensemble_OVO and ensemble_Ovall in dataset I. In support of this observation, a classification improvement on the values of sensitivity is ~2.87% and 3.8% for the ensemble_OVO and

ensemble_Ovall, respectively. The ensemble_MLA apple to distinguish the Grade 4 from other grades easily comparing to the state of the art ensemble_ OVO and ensemble_Ovall.

Table 2

Average Sensitivity Over 30 Runs Comparison of the Proposed Ensemble_ MLA with Existing Multiclass Approaches for Dataset I and II

Dataset	Multiclass approach	Benign	Grade 3	Grade 4	Overall sensitivity
Dataset I	Ensemble_OVO based on LibSVM	97.17	**90.6**	74.28	87.35
	Ensemble_Ovall	97.94	89.0	72.14	86.36
	Ensemble_MLA	97.17	85.18	**88.33**	**90.22**
Dataset II	Ensemble_OVO based on LibSVM	88.0	80.0	96.9	88.3
	Ensemble_Ovall	82.0	**88.66**	**97.46**	**89.37**
	Ensemble_MLA	**100**	60.66	94.6	85.08

The ensemble_MLA distinguishes Grade 4 from the other classes in dataset I. These sensitivity results confirmed the accuracy of the ensemble_MLA in distinguishing Grade 4 tissue from Grade 3 and Benign tissues. Although the ensemble_MLA reported the highest overall sensitivity compared to the state-of-the-art approaches, its sensitivity value in discriminating Grade 3 reported a lower value (85.18%). This could be due to its high similarity between Grade 4 and the extracted features, which increased the ambiguity of discrimination of this class vis-à-vis the other classes.

In addition the ensemble_MLA achieved the highest sensitivity for distinguish the benign from other grades in dataset II. But, it achieved the lowest sensitivity in grades 3, and 4. One reason for this low sensitivity is number of training samples in dataset II as illustrated in Table 1. When number of training samples is low comparing to the number of feature, this is mostly attributed to underfitting issue, which manifested due to the construction of a classifier model without the suitable number of samples comparing to number of features issue (Yu et al., 2003).

To further understand how the various multiclass approaches performed across the three-class PCa grading, we look at the specificity measurement as shown in Figure 10. Specificity is a measure which determines the probability of the results that are true negative. Higher specificity value indicates superior performance among the ensemble methods.

Table 3

Average Specificity Over 30 Runs Comparison of the Proposed Ensemble_MLA with Existing Multiclass Approaches for Dataset I and II

Dataset	Multiclass approach	Benign	Grade 3	Grade 4	Overall specificity
Dataset I	Ensemble_OVO based on LibSVM	86.3	96.97	**95.23**	92.83
	Ensemble_Ovall	85.03	97.19	94.82	92.34
	Ensemble_MLA	**92.62**	**98.55**	93.61	**94.92**
Dataset II	Ensemble_OVO based on LibSVM	90.34	93.39	100	94.57
	Ensemble_Ovall	**94.54**	91.01	99.62	**95.05**
	Ensemble_MLA	79.67	**100**	**100**	93.22

Looking at the specificity results in Table 3, one can observe that the overall ensemble_MLA specificity results exceeded those obtained using the ensemble_OvO and ensemble_Ovall in dataset I. These results confirm that using the MLA approach via the ensemble framework is beneficial towards further improving the three-class grading specificity performance. The overall specificity results of the three-class for dataset I using the ensemble_MLA showed an increment of 2.09% and 2.58%, compared to the state-of-the-art ensemble_OvO and ensemble_Ovall, respectively.

Also, in dataset II, the proposed ensemble_MLA achieved superior results for distinguish Grades 3 and 4. However, for the benign cases, it achieved the lowest specificity due to heterogeneity nature of the samples and underfitting issue due to limited number of samples in benign case for dataset II.

The results confirmed that using the proposed MLA multiclass approach in the ensemble framework can improve the performance in the context of Gleason

grading of prostate histopathology. The proposed MLA is more robust than the OVO and Ovall schemes, because it performs each binary task in two iterations. The MLA starts by performing the binary tasks on the strong samples, i.e. high-quality samples, and repeats the same binary task using the weak samples to double check for each binary task in the MLA approach. This enhances the final prediction of the MLA.

This study also defined specificity and sensitivity for each class in order to compare the performance of the ensemble_MLA with the state-of-the-art multiclass binarization approaches (i.e., OvO and Ovall). For a class x, the sensitivity criterion is computed based on the number of test samples correctly classified as class x (i.e., TP) with respect to the total number of positive instances (TP/(TP+FN)). Contrastingly, specificity criterion is computed based on the number of correctly classified not class x (TN), divided by total number of negative cases TN/(TN+FP) (Fernández-Navarro, Hervás-Martínez, & Gutiérrez, 2011). [29]. Figures 9, 10 and Tables 1, 2 illustrate the sensitivity and specificity for each class from the three-class PCa grading problem when using the ensemble framework based on different multiclass binarization approaches.

CONCLUSION

In this paper, we introduced an approach called MLA to solve the multiclass classification problem in prostate histopathology image grading by splitting the multiclass problem into multi-level (hierarchical) binary subtasks. The proposed approach aims to obtain improved grading results by starting with the well-known binary classification (Benign vs. Grade 3) at level 1. In the next level, the rest of the binary classification tasks (Benign vs. Grade 4, and Grade 3 vs. Grade 4) are divided into two separate sub-tasks; strong and weak, based on the power of the samples in each binary task. In turn, this motivates the strong samples to produce the final prediction since they have more information about the considered task.

The performance of the MLA approach was evaluated based on its impact on the performance of the ensemble framework utilizing the main tissue components. The experimental results on the prostate histopathology image dataset were compared with the most popular approaches, such as the OVO and Ovall. It was revealed that this new approach outperforms the state-of-the-art OVO and Ovall. For future work, we intend to develop the analysis of other aggregation strategies in the proposed MLA scheme, such as product rule.

ACKNOWLEDGMENT

The authors would like to thank the Faculty of Information Science and Technology - Universiti Kebangsaan Malaysia for providing facilities and Financial Support under Fundamental Research Grant Scheme No. DIP-2015-023 and project No. ETP-2013-053.

REFERENCES

Al-Kadi, O. S. (2010). Texture measures combination for improved meningioma classification of histopathological images. *Pattern Recognition*, *43*(6), 2043–2053.

Albashish, D., Sahran, S., Abdullah, A., AbdShukor, N., & Hayati Md Pauz, S. (2016). Ensemble Learning of tissue components for prostate histopathology image grading. *Engineering and Information Technology (IJASEIT).*, *6*(6), 1132–1138.

Albashish, D., Sahran, S., Abdullah, A., Adam, A., Abd Shukor, N., & Hayati Md Pauz, S. (2015). Multi-scoring Feature selection method based on SVM-RFE for prostate cancer diagnosis. In *The 5th International Conference on Electrical Engineering and Informatics 2015* (pp. 682–686). Bali, Indonesia. 10-11 August.

Ali, U., Shaukat, A., Hussain, M., Ali, J., & Khan, K. (2016). Automatic cancerous tissue classification using discrete wavelet transformation and support vector machine. *Journal of Basic and Applied Scientific Research*, *6*(7), 15–23.

Almuntashri, A., Agaian, S., Thompson, I., Rabah, D., Zin Al-Abdin, O., & Nicolas, M. (2011). Gleason grade-based automatic classification of prostate cancer pathological images. In *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics* (pp. 2696–2701). Anchorage, Alaska, USA.9-12 October. doi.org/10.1109/ICSMC.2011.6084080

Anand, R., Mehrotra, K., Mohan, C. K., & Ranka, S. (1995). Efficient classification for multiclass problems using modular neural networks. *IEEE Transactions on Neural Networks*, *6*(1), 117–124. doi.org/10.1109/72.363444

Casasent, D., & Wang, Y. (2005). A hierarchical classifier using new support vector machines for automatic target recognition. *Neural Networks*, *18*(5), 541–548. doi.org/10.1016/j.neunet.2005.06.033

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACMTransactions on Intelligent Systems and Technology*, *2*(3), 1–27.

Doyle, S., Feldman, M. D., Shih, N., Tomaszewski, J., & Madabhushi, A. (2012). Cascaded discrimination of normal, abnormal, and confounder classes in histopathology: Gleason grading of prostate cancer. *BMC Bioinformatics*, *13*(1), 282–297. doi.org/10.1186/1471-2105-13-282

Epstein, J. I. (2010). An update of the gleason grading system. *The Journal of Urology*, *183*(2), 433–440. doi.org/10.1016/j.juro.2009.10.046

Farjam, R., Hamid, S.-Z., Kourosh, J.-K., & Reza, Z. (2007). An Image analysis approach for automatic malignancy determination of prostate pathological images. *Cytometry Part B :Clinical Cytometry*, *72B*(4), 227–240. doi.org/10.1002/cyto.b

Farjam, R., Soltanian-zadeh, H., Zoroofi, R. A., Ford, H., & System, H. (2004). Wavelet Based Determination of Malignancy of the Pathological Images of the Prostate. *WSEAS Trans Electronics*, *3*, 476–82.

Fernández-Navarro, F., Hervás-Martínez, C., & Gutiérrez, P. (2011). A dynamic over-sampling procedure based on sensitivity for multi-class problems. *Pattern Recognition*, *44*(8), 1821–1833. doi.org/10.1016/j.patcog.2011.02.019

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, *44*(8), 1761–1776. doi.org/10.1016/j.patcog.2011.01.017

Galar, M., Fernández, A., Bustince, H., & Herrera, F. (2011). *Aggregation schemes for binarization techniques Methods ' description*. *Technical Report, Dpto. de Autom´ atica y Computaci´ on, Public University of Navarre, Research Group on Soft Computing and Intelligent Information Systems*. Pamplona,Spain.

Haralick, R. M., & Shanmugam, K. (1973). Textural Features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, *3*(6), 610–621. doi.org/10.1109/TSMC.1973.4309314

Honeine, P., Noumir, Z., & Richard, C. (2013). Multiclass classification machines with the complexity of a single binary classifier. *Signal Processing*, *93*(5), 1013–1026. doi.org/10.1016/j.sigpro.2012.11.009

Hsu, C.-W., & Lin, C.-J. (2002). A Comparison of methods for multiclass support vector machines. *Neural Networks*, *13*(2), 415–425.

Huang, P.-W., & Lee, C.-H. (2009). Automatic classification for pathological prostate images based on fractal analysis. *IEEE Transactions on Medical Imaging*, *28*(7), 1037–50. doi.org/10.1109/TMI.2009.2012704

Kang, S., Cho, S., & Kang, P. (2015). Constructing a multi-class classifier using one-against-one approach with different binary classifiers. *Neurocomputing*, *149*(PB), 677–682. doi.org/10.1016/j.neucom.2014.08.006

Khelifi, R., Adel, M., & Bourennane, S. (2012). Multispectral texture characterization: application to computer aided diagnosis on prostatic tissue images. *EURASIP Journal on Advances in Signal Processing*, *2012*(1), 118–131. doi.org/10.1186/1687-6180-2012-118

Lopez, C. M., & Agaian, S. (2013). A new set of wavelet- and fractals-based features for Gleason grading of prostate cancer histopathology images. In *Proceedings SPIE-IS&T Electronic Imaging* (Vol. 8655, p. 865516). Burlingame california, USA. 3-7 February. doi.org/10.1117/12.1000193

Mosquera-Lopez, C., Agaian, S., Velez, A., & Thompson, I. (2015). Computer-aided prostate cancer diagnosis from digitized histopathology: A review on texture-based systems. *IEEE Reviews in Biomedical Engineering*, *8*, 98–113. doi.org/10.1109/RBME.2014.2340401

Moustakidis, S. P., & Theocharis, J. B. (2012). A fast SVM-based wrapper feature selection method driven by a fuzzy complementary criterion. *Pattern Analysis and Applications*, *15*(4), 379–397. doi.org/10.1007/s10044-012-0293-7

Naik, S., Doyle, S., Agner, S., Madabhushi, A., Feldman, M., & Tomaszewski, J. (2008). Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *In 5th IEEE International*

Symposium on Biomedical Imaging: From Nano to Macro (pp. 284–287). Paris, France. 14-17 May. doi.org/10.1109/ISBI.2008.4540988

Nguyen, K., Sabata, B., & Jain, A. K. (2012a). Prostate cancer grading: Gland segmentation and structural features. *Pattern Recognition Letters*, *33*(7), 951–961. doi.org/10.1016/j.patrec.2011.10.001

Nguyen, K., Sabata, B., & Jain, A. K. (2012b). Structure and context in prostatic gland segmentation and classification. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*. Nice, France. 1-5 October: Springer Berlin Heidelberg.

Nguyen, K., Sarkar, A., & Jain, A. (2014). Prostate cancer grading: Use of graph cut and spatial arrangement of nuclei. *IEEE Transactions on Medical Imaging*, *33*(12), 2254–2270. doi.org/10.1109/TMI.2014.2336883

Shannon, C. E., & Warren, W. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, *5*(1), 3–55.

Silla, C. N., & Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, *22*(1–2), 31–72. doi.org/10.1007/s10618-010-0175-9

Tabesh, A., Teverovskiy, M., Pang, H. Y., Kumar, V. P., Verbel, D., Kotsianti, A., & Saidi, O. (2007). Multifeature prostate cancer diagnosis and gleason grading of histological images. *IEEE Transactions on Medical Imaging*, *26*(10), 1366–1378. doi.org/10.1109/TMI.2007.898536

Yu, H., Yang, J., Wang, W., & Han, J. (2003). Discovering compact and highly discriminative features or combinations of drug activities using support vector machines. In *Bioinformatics Conference, 2003. CSB 2003. Proceedings of the 2003 IEEE* (pp. 220–228). IEEE.

Zhang, Z., Krawczyk, B., Garcìa, S., Rosales-pérez, A., & Herrera, F. (2016). Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data. *Knowledge-Based Systems*, *106*, 251–263. doi.org/10.1016/j.knosys.2016.05.048

Zhou, X., & Tuck, D. P. (2007). MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics*, *23*(9), 1106–14.