ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

DEVELOPMENT OF A BIOMEDICAL IMAGE ANALYSIS FRAMEWORK, BASED ON WEB SERVICES

A dissertation submitted
to the Department of Digital Systems,
of University of Piraeus in complete
fulfillment of the requirements for the
degree of Doctor of Philosophy

by

Theodosios E. Goudas

Graduation Date: 20-02-2015

7-member Committee

Maglogiannis Ilias - Assistant Professor, University of Piraeus (supervisor)

Plagianakos Vassilis - Associate Professor, University of Thessaly

Hatziioannou Aristotle – Research Associate Professor of the NHRF

Themistocleous Marinos - Associate Professor, University of Piraeus

Prentza Andriana - Associate Professor, University of Piraeus

Kyriazis Demosthenis - Lecturer, University of Piraeus

Delibasis Konstantinos - Assistant Professor, University of Thessaly

Dissertation written by

Theodosios E. Goudas

M.S. Informatics and Management, Schools of Informatics and Economic Sciences, Aristotle University of Thessaloniki, Thessaloniki, Greece, 2009

B.S. Information and Communication Systems Engineering Department, University of the Aegean, Samos, Greece, 2007

Approved by

Maglogiannis Ilias - Assistant Professor, Department of Digital Systems, University of Piraeus (supervisor)

Plagianakos Vassilis - Associate Professor, Department of Computer Science and Biomedical Informatics, University of Thessaly

Hatziioannou Aristotle – Research Associate Professor, National Hellenic Research Foundation

Accepted by

Themistocleous Marinos - Associate Professor, Department of Digital Systems, University of Piraeus

Prentza Andriana - Associate Professor, Department of Digital Systems, University of Piraeus

Kyriazis Demosthenis - Lecturer, Department of Digital Systems, University of Piraeus

Delibasis Konstantinos - Assistant Professor, Department of Computer Science and Biomedical Informatics, University of Thessaly

DEDICATION

I dedicate this Ph.D. thesis to my beloved parents.

ACKNOWLEDGEMENTS

I am using this opportunity to express my gratitude to everyone who supported me throughout the accomplishment of this Ph.D. thesis. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the Ph.D. subject.

First of all, I would like to special thank my supervisor, Ilias G. Maglogiannis, for his continuous guidance and assistance in the context of the Ph.D. thesis.

I would also like to thank my 2nd Doctoral Dissertation Committee Member, Vassilis P. Plagianakos, for his assistance on different types of requests.

I would also like to thank my 3rd Doctoral Dissertation Committee Member, Aristotle Chatziioannou, who assisted me with his knowledge around biomedical aspects of the problems we encountered.

I would like to deeply thank my beloved parents, who supported me emotionally and economically from the day I decided to target the Ph.D. thesis achievement and generally from the beginning of my life.

I would like to thank Dimitris Konsoulas for his friendship and his support on tech and platform-based questions around cross-platform collaboration of the applications.

Last but not least, I would like to thank Effie Amanatidou for her endless support, love and the provided proofreading of the submitted papers and this thesis, as well.

Theodosios E. Goudas
20-02-2015, Athens, Piraeus

SYNOPSIS

Data mining of biomedical images is a complex task, because it requires several steps, like color adjustment, image filtering, segmentation, feature extraction, characterization, etc. and appropriate calibration on each of them, so as to achieve a satisfying result.

This Ph.D. thesis focuses on the development of a framework, containing the necessary data mining and image analysis techniques, organized into entities-operators, capable to deal with the complex biomedical data mining tasks. The proposed framework enables the building of workflow schemes, capable of dealing with these mining tasks. Furthermore, it enables the generation of multiple workflow versions of the core workflow scheme by combining all the operators with each possible way and auto-selecting the optimal one. In this thesis, a number of applications, exploiting web services and applying ontological modelling are presented, allowing the intelligent creation of image mining workflows and the optimal workflow scheme selection for each of them.

The choice of the optimal workflow is based either to the comparison with the ground truth, or via the log-likelihood distance metric of the clustered salient objects. This biomedical image analysis framework may require advanced knowledge of data mining and image analysis theory, but it requires only fundamental programming skills for the development of an intermediate level workflow scheme. It can be directly integrated to TAVERNA or similar workflow management platforms.

Additionally, the biomedical problems examined during this thesis and their corresponding solutions are also presented. Each of the proposed image mining methodologies utilizes the developed framework. All the proposed methodologies achieve satisfying performance through the proposed framework. Furthermore, some additional image mining scenarios are applied through the framework, demonstrating the optimal workflow proposal mechanism.

ΠΕΡΙΛΗΨΗ

Η εξόρυξη γνώσης από βιοϊατρικές εικόνες είναι μια πολύπλοκη και χρονοβόρα διαδικασία, γιατί απαιτούνται πολλά βήματα, όπως η ρύθμιση του χρώματος, το φιλτράρισμα της εικόνας, η τμηματοποίησή της, η εξαγωγή χαρακτηριστικών γνωρισμάτων, ο χαρακτηρισμός της, κ.λπ. Κάθε ένα από αυτά τα βήματα απαιτεί κατάλληλη βαθμονόμηση, ώστε συνολικά να πετύχουν το βέλτιστο αποτέλεσμα.

Αυτή η διδακτορική διατριβή, εστιάζει στην ανάπτυξη ενός πλαισίου, το οποίο περιέχει τις απαραίτητες τεχνικές εξόρυξης και ανάλυσης εικόνας, οργανωμένες σε οντότητες, για την επίλυση των σύνθετων βιοϊατρικών προβλημάτων ανάλυσης εικόνας. Το προτεινόμενο πλαίσιο επιτρέπει το σχεδιασμό διαγραμμάτων ροών εργασίας, για να επιλύσει αυτά τα προβλήματα. Επιπλέον, παρέχει τη λειτουργία της αυτόματης δημιουργίας παράλληλων πολλαπλών εκδόσεων (multiple parallel instances) του διαγράμματος ροής εργασίας που σχεδιάστηκε, πραγματοποιώντας όλους τους πιθανούς συνδυασμούς των τελεστών που προστέθηκαν στο διάγραμμα, για να επιλεγεί αυτόματα ο βέλτιστος συνδυασμός ροής εργασίας. Για την υλοποίηση αυτού του πλαισίου αξιοποιήθηκε η τεχνολογία υπηρεσιών δικτύου (web services technology), σε συνδυασμό με τη μοντελοποίηση τεχνικών ανάλυσης και εξόρυξης εικόνας (image mining and analysis techniques), σε ανεξάρτητες οντότητες.

Η επιλογή της βέλτιστης ροής εργασίας πραγματοποιείται, είτε συγκρίνοντας τα αποτελέσματα της τμηματοποίησης με την πραγματικότητα (Ground Truth), είτε με τη χρήση του μέτρου της απόστασης log-likelihood των ομαδοποιημένων εντοπισμένων αντικειμένων (clustered salient objects). Για τη χρήση αυτού του πλαισίου απαιτούνται βασικές – και ίσως σε κάποιες περιπτώσεις προχωρημένες – γνώσεις ανάλυσης εικόνας, αλλά δεν απαιτούνται προγραμματιστικές γνώσεις. Αυτό το πλαίσιο μπορεί να ενσωματωθεί στο πρόγραμμα διαχείρισης ροών εργασίας TAVERNA ή σε οποιαδήποτε άλλη παρόμοια πλατφόρμα.

Επιπλέον, σε αυτήν τη διδακτορική διατριβή, παρουσιάζονται όλα τα βιοϊατρικά προβλήματα ανάλυσης εικόνας που εξετάσθηκαν κατά τη διάρκεια της εκπόνησής της. Καθεμιά από τις προτεινόμενες προσεγγίσεις εξόρυξης εικόνας, χρησιμοποιεί το προτεινόμενο πλαίσιο. Όλες οι προσεγγίσεις εξήγαγαν ικανοποιητικά αποτελέσματα, χρησιμοποιώντας τις δυνατότητες του προτεινόμενου πλαισίου. Επιπλέον, ορισμένα πρόσθετα σενάρια εξόρυξης εικόνας μοντελοποιήθηκαν στο προτεινόμενο πλαίσιο, αποδεικνύοντας την αποδοτική λειτουργία της εύρεσης της βέλτιστης ροής εργασίας.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Research Area

This work is about the development of a web service-based image-mining framework. It can be easily understood by just reading the title of this Ph.D thesis that this work is related with the Image Mining, Workflow Management and Web Services fields that are briefly described below.

1.1.1 Image Mining

Image Processing and Computer vision are closely related to the study of biological vision. The field of biological vision studies and models the physiological processes behind visual perception in humans and other animals. Computer vision, on the other hand, studies and describes the processes implemented in software and hardware, behind artificial vision systems. Interdisciplinary exchange between biological and computer vision has proven fruitful for both fields.

Image analysis methods are utilized in order to offer solution to complex biomedical imaging problems, such as the quantification of pathogenic areas, the segmentation and the characterization of salient objects etc.

Biomedical image characterization is a complex procedure since it requires several stages as data acquisition, preprocessing, segmentation, feature extraction, training and classification (see Figure 1) of the corresponding data [1]. Proper modification of the above stages and suitable image analysis methods usage – for each specific biomedical problem – is required. There are also cases that large amount of images must be analysed and characterized.

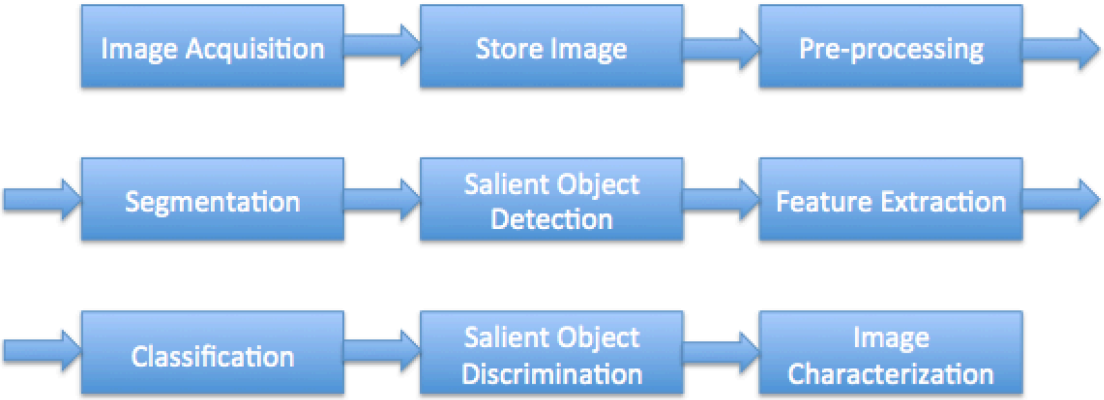


Figure 1. Generic Workflow Diagram of a Standard Image Characterization Procedure

Nowadays, the trend in image processing and software engineering, in general, is towards the development of algorithms and tools that provide each one of these functions distributed [2] – especially in form of Web Services – a technology that enables developers to programmatically access heterogeneous, distributed resources, providing easier integration and interoperability between data and applications.

1.1.2 Workflow Management and Web Services

A potential solution for the biomedical image characterization procedure would be the creation of workflow diagrams, whose operators – entities will be parts of the data mining tasks. Based on this solution, automated characterization of large and complex datasets can be achieved through custom workflow builds. Even by utilizing the above solution, the process of combining all the relevant operators for the achievement of the optimal characterization result, it still remains a time consuming procedure. Thus, an evaluation system for the performance of these workflows is required. The main advantage of a workflow-based approach for the task of image mining is that people with fundamental programming skills can easily deal with difficult characterization tasks. Computer scientists and expert physicians, who deal with large datasets, need the feature of repeating this complex procedure for many samples in a short time and sometimes they need to evaluate multiple solutions, in order to find the optimal solution for a specific issue. Both of the above needs can be accomplished through a workflow-based approach.

Web services technology became popular the last few years, because of its ability to integrate the features of a developed application into any system, regardless the programming language or the architecture that was utilized for the development of the specific system. Web services are embedded in web pages, cloud applications and many other web-based applications.

The major feature of web services is the interoperability. The element that characterizes the web service technology is the XML based messages that must follow a specific SOAP format. These XML messages can be easily sent from computer to computer. This ease occurs due to the following features:

- The client application can be written in any programming language.
- The locations of the sender and receiver machine are irrelevant.
- It is no necessary for the client to know the type of SOAP processor that is installed on the server.
- It does not matter what is the type of the computer or the operating system that is running on the sender or the receiver of that type of messages.

It is a great expectation that through the utilization of web services technology each developed application in the world will be able to exchange SOAP messages with each other application. Web services have also low operational costs. They can easily adapt to any system. They require little or no code changes to the old system, in order to integrate to it.

Taking into account the advantages of the workflow and web services technologies, their combination developed powerful tools for several scientific fields. MyExperiment and Biocatalogue repositories (http://www.myexperiment.org and http://www.biocatalogue.org respectively), hold a lot of workflows and web services that can be accessed through TAVERNA [2], a workflow manager, which provides the feature of building workflows, based on custom or other users' services.

There are many additional workflow managers in our days, but only a few of them are able to manage scientific data streams and can adapt their modeled processes to the needs of the corresponding scientific field. These are (except TAVERNA): Kepler [3], BASCIIS [4], VisTrails [5], ESRI ArcGIS ModelBuilder [6], Microsoft Visio [7], OpenMI [8], Pegasus [9], Platform LSF [10], Science Pipes [11], DaltOn [12] and Triana [13]. It must be mentioned, that TAVERNA workflow manager is the most eligible to customizations and integrating modules, enabling the creation of web-based operators that implement their own rules (input format, output format, etc.) to the core platform.

1.2 Research Contribution and Structure

In this work, a novel biomedical image-mining framework is proposed. Specifically, the framework is based on intelligent planning that enables the creation of several dynamic image processing and analysis workflows, instead of black-box tools. These workflows have the ability to adjust to specific image data mining tasks. In more details, the proposed framework allows the parallel image processing of large amount of images through the user's developed workflow scheme, combined with feature of multiple instance generation of the same workflow - each of them initiated by different settings - but maintaining the initial structure.

The dynamic formulation of image processing workflows provides flexibility to the users for creating their own analysis schemas and methodologies, testing them in practice and optimizing their performance through the optimal workflow selection feature. To the best of our knowledge, there is no other workflow based image-mining framework, supporting multiple workflow instance generation and optimal schema selection. The Ph.D thesis is structured as follows:

In chapter 2, the related work of the field is presented. In the first subsection of the chapter, the state of the art literature is briefly presented, while in the second chapter a brief analysis of the related software tools is provided, as well.

In chapter 3, the development steps and the features of the developed framework are described analytically. More specifically, the development needs for each image-mining operator are presented. The software and hardware requirements are also denoted.

In chapter 4, we present a number of case studies that utilize our proposed framework. These case studies present novel solutions for four biomedical problems. More specifically, the obstructive nephropathy detection, the breast cancer quantification, the skin lesion detection and the block detection and quantification of fibrotic areas methodologies are analytically described. The chapter also contains the background information for these pathologies and the corresponding datasets.

In chapter 5, the experimental results from the utilization of the proposed framework are presented and analyzed. More specifically, the image mining results, of the biomedical image analysis problems presented in chapter 4, are analyzed. These results exported from the utilization of the developed framework's functionalities, described in chapter 3.

Finally, chapter 6 concludes this work by discussing its performance and usability, while it proposes future work for the specific scientific field.

1.3 References

[1] M.R. Smith, X. Wang, R.M. Rangayyan, Evaluation of the sensitivity of a medical data-mining application to the number of elements in small databases, Biomedical Signal Processing and Control, Volume 4, Issue 3, New Trends in Voice Pathology Detection and Classification - M & A of Vocal Emissions, July 2009, Pages 262-268.

[2] Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A., Li, P. Taverna: A tool for the composition and enactment of bioinformatics workflows (2004) Bioinformatics, 20 (17), pp. 3045-3054.

[3] Ludäscher, B., et al. Scientific Workflow Management and the Kepler System. Concurrency and Computation: Practice & Experience 18(10), 1039–1065 (2006).

[4] Ben-Miled, Z., Li, N., Baumgartner, M., Liu, Y. A decentralized approach to the integration of life science web databases. Informatica (Slovenia) 27(1), 3–14 (2003).

[5] Callahan, S.P., Freire, J., Santos, E., Scheidegger, C.E., Silva, C.T., Vo, H.T. VisTrails: Visualization meets data management (2006) Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 745-747.

[6] Jiménez-Perálvarez, J.D., Irigaray, C., El Hamdouni, R., Chacón, J. Building models for automatic landslide-susceptibility analysis, mapping and validation in ArcGIS (2009) Natural Hazards, 50 (3), pp. 571-590.

[7] Microsoft Visio can be fount at: www.microsoftstore.com/Visio_2013.

[8] Gijsbers, P.J.A., Gregersen, J.B. OpenMI: A glue for model integration (2005) MODSIM05 - International Congress on Modelling and Simulation: Advances and Applications for Management and Decision Making, Proceedings, pp. 648-654.

[9] Kee, Y.-S., Byun, E., Deelman, E., Vahi, K., Kim, J.-S. Pegasus on the virtual grid: A case study of workflow planning over captive resources (2008) 2008 3rd Workshop on Workflows in Support of Large-Scale Science, WORKS 2008, art. no. 4723961.

[10] Platform LSF can be found at: http://www.platform.com/workload-management/high-performance-computing.

[11] Science Pipes Workflow Manager can be found at: http://sciencepipes.org/.

[12] Jablonski, S., Curé, O., Rehman, M.A., Volz, B. DaltOn: An infrastructure for scientific data management (2008) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5103 LNCS (PART 3), pp. 520-529.

[13] Taylor, I., Shields, M., Wang, I., Harrison, A. The triana workflow environment: Architecture and applications (2007) Workflows for e-Science: Scientific Workflows for Grids, pp. 320-339.

CHAPTER 2

Related Work

2.1 Review of Scientific Literature

In this chapter, the state-of-the-art works of the field are briefly presented. Additionally, the utilized software tools for the accomplishment of this Ph.D thesis are also presented, along with a brief description of the way that were exploited in the context of this research.

2.1.1 Image Analysis

The field of biomedical image analysis has occupied several research teams and significant research work may be found in literature in this field. Phukpattaranont and Boonyaphiphat [1] presented an algorithm for the segmentation of cancer cells on microscopic images of immune-histologically stained slides from breast cancer. Their method was based on color contents, neural networks classification and cell size consideration. Maglogiannis et al [2] successfully classified biological microscopic images of lung tissue sections with idiopathic pulmonary fibrosis. Tosun and Gunduz-Demir [3] proposed an effective algorithm for segmenting histo-pathological images. Niwas et al [4] investigated the contribution of wavelet transformation in detecting breast cancer. They also examined, if complex wavelets are capable of achieving higher accuracy results than the more common real valued wavelet transformation. Feature extraction performed by wavelets, served as input to a K-Nearest Neighbor classifier, which provided satisfying results. [5], [6], [7] and [8] propose fully automated methods, in order to segment and classify cancer cell nuclei and cytoplasms. Xu et al [9], performed segmentation, using Support Vector Machines on preprocessed angiogenic chorioallantoic membrane images and by utilizing automatic counting techniques, they achieved angiogenesis quantification. Wu et al [10], proposed a frequency analysis algorithm, based on Discrete Fourier Transformation, in order to achieve the segmentation of textured cells. Chaabane et al [11] achieved color edge detection on cancer cells. Their method is based on statistical features and a threshold technique. Sagonas et al [12] proposed a method for the analysis of Fluorescent In Situ Hybridization (FISH), based on cell nuclei and red/green spot modeling, utilizing Radial Basis Function.

Despite the complexity of the automated recognition problem in biological images, there are many revolutionary researches in the literature. El Abbadi and Miry [13], presented a novel method for the segmentation of dermoscopic images, including the filtering of issues encountered in their trial of achieving efficient segmentations, like a variety of lesion shapes, sizes, color, changes, due to different skin types and textures and presence of hairs. In [14] and [15], authors estimated the diagnostic efficiency of dermoscopic criteria on the biomedical problem of the differentiation of superficial basal cell carcinoma from other basal cell carcinoma types. Madooei et al [16], achieved the quantification of melanin and hemoglobin content in dermoscopy images by utilizing a novel methodology, which employs the stochastic Latent Topic Models framework. Rand et al [17], produced x-ray scatter images of hepatocellular carcinoma labelled with a nanoparticle contrast agent, by utilizing a numeric processing technique. Boyaci et al [18], performed retrospective analysis of multidetector computed tomography findings, in order to specify the dominance of paratracheal air cysts and their correlation with emphysema and bronchiectasis. Lou et al [19], utilized a novel software on patients with lower extremity peripheral arterial occlusive disease evaluated the results of the quantitative analysis of preoperative and postoperative hemodynamic change exported by the corresponding software. Degenhart et al [20], investigated the localization of the right adrenal veins (RAV) in primary aldosteronism patients through the utilization of multidetector computed tomography venous mapping. In [21], authors examined wrist radiographs samples, in order to identify any anatomical or pathological characteristics that predict incorrect red dot classification. Additionally, they proved the dominance of red dot markers on these samples.

Sotaquira et al [22], segmented and morphologically quantified the mitral annulus (MA) and mitral leaflets in closed valve configuration from RT3-D TEE volumes, through the utilization of a novel algorithm. Lorenzi et al [23], presented a non-linear registration-based method for the estimation and the analysis of the contributions from aging and pathology through the observation of brain morphologies. They accomplished their task by utilizing a longitudinal model of the brain's normal aging process from magnetic resonance imaging scans. In [24], authors evaluated a number of algorithms through the task of the segmentation of cortical grey matter, non-myelinated and myelinated white matter, brainstem, basal ganglia and thalami, cerebellum, and cerebrospinal fluid in the ventricles and in the extracerebral space separately. In [25] and [26], authors achieved automated localization of breast cancer on dynamic contrast-enhanced magnetic resonance images, including the filtering of artefact objects for the enhancement of the segmentation task. In [27], authors utilized a gamma transformation approach, in order to achieve segmentation and detection of brain haemorrhage in MRI scans of the brain. In [28], authors propose a tool for the diagnosis and the automatic segmentation of brain tumor. The achieved segmentation is accomplished through the use of the gray-level co-occurrence matrix features that are extracted from the MR images. Support Vector Machines Classifier was utilized for the classification of the images as tumorous or non-tumorous. Khalvati et al [29], achieved segmentation of the breast boundary in 3D MR images through a robust atlas-based segmentation algorithm. Mitra et al [30], adopted Watershed methodology to overcome the issues of touching objects on the segmentation of MRI skull bone lesions.

P. Wang et al [31] classified endoscopic images based on texture feature. They chose the neural network classifier, since they found that it was the ideal for the solutions of their problem. LaConte et al [32] applied Support Vector Machines to block design fMRI. As for the audio section, Olmez and Dokur [33] used Artifial Neural Network Classifier, in order to classify heart sounds. They used the first and the second Heart sound, in order to extract some features to train their model. Wolberg et al [34] developed a system for evaluating cytologic features derived directly from a scan of breast fine needle aspirate slide. Bar et al [35] tested an automated tissue pre-classification approach for telepathology. They developed image analysis and ranking techniques, in order to improve the accuracy of their proposed system. Masseroli et al [36] designed and developed an image analysis method, in order to achieve automatic quantification of liver fibrosis.

Continuing the literature review, it was found that the field of microscopy image analysis has occupied several research teams and significant research work may be found. Several robust tools have also been proposed for the assessment of liver fibrosis [37]-[40], the study of micro vascular circulating leukocytes [41], the assessment of testicular interstitial fibrosis, [42], [43], or that of lung fibrosis [44] and angiogenesis [45]. The use of pattern recognition or classification methods – like Support Vector Machines or Neural Networks – could enable the design of decision-making algorithms, appropriate to microscopic data. Within this context, a method for evaluation of electron microscopic images of serial sections based on the Gabor wavelets and the construction of a mapping between the “model” and the “target” image has been proposed in [45].

2.1.2 Workflow Management and Web-Services related Works

In the workflow management and web-services technology field, Kreftin et al [46], proposed the grid integration of medical image processing applications as grid workflows, where the workflow manager is responsible for the execution of all tasks related to grid communication and the developer is responsible for setting the access rights on his code and defining the workflow manager what to do with it. Coarse-grained parallelization of processing steps can be applied, in order to achieve runtime reduction. Kooper et al [47], presented a novel solution for reconstructing 3D medical volumes proposing the web service implementation as an additional layer to a dataflow framework. Koulouzis et al [48], used image-based analysis of vascular disorders as a case study and compared two transport models, a centralized (data are located in a centralized data repository) and a disturbed one (data are delivered directly from producing to consuming services), showing that in the disturbed data the flow model achieved a satisfying speed up of the workflow execution. In [49], a fast and reliable collaborative framework for ophthalmologists and other experts in the field, containing retinal image-based applications, is proposed. It presents a semi-automated methodology for the analysis of retinal microcirculation. Glatard et al [50], were based on Taverna workflow manager to deploy a data intensive application, which supports image registration algorithms, grid enabled workflow manager and a grid middleware for performing the computations. Olabarriaga et al [51] developed a set of medical image analysis tools and described the requirements, in order to intergrade the existing systems (frameworks) with them. In a later work Olabarriaga et al [53], created a virtual lab, which integrates many image analysis tools and platforms into a user-friendly system. Snel et al [52] developed a distributed workflow management system, which offers a set of image analysis tools (and data logistic management tools). MIAKT [54] uses grid and web services, in order to support several image processing functionalities. For example, it allows different clinicians to collaborate on a diagnosis for patients that suffer breast disease. DeVIDE [55] is a modular framework, which supports the development of medical image processing algorithms. Its advantage is that the user can interact with every level of the system. For example, the user executes an algorithm and during its runtime stage he can edit the code. The changes that will be applied, affect instantly the output results. LONI Pipeline [56], provides an image analysis tool for neuroimaging applications. This framework allows the parallelization of data independent components. However, all the surveyed systems focus on automated and fast or parallel execution of image processing workflows and none of them support their intelligent annotation, planning and building.

Alexopoulos et al [57], implemented a concept of a web-based workflow system, capable to support and coordinate the production engineering activities, utilizing a variety of digital tools. In [58], authors achieved the analysis of complex data from the proteomic field by utilizing a number of novel web based tools, intergraded in Taverna.

Xudong et al [59] presented the development of a workflow framework, utilizing the advantages of SHOP2 (Simple Hierarchical Ordered Planner 2) and CSP (Constraint Satisfaction Problems), in order to achieve intelligent service composition. Cheng et al [60], developed a prototype service oriented based system in order to manage construction supply chains.

In [61], authors present Ergatis, a workflow management system, which allows users to build and run their own pipeline application builds for the analysis of the complex genomics data. Another similar work in the genomics data field is presented by Lushbough et al [62], who developed Bioextract Server with the additional capability to access large genomics databases. In [63], a chemistry inspired workflow management system, capable to deal with several workflow patterns, based on the chemical model, is proposed. Kim [64] presented a web service wrapper for the processing of chemical information via the chemoinformatics workflows providing independent applications from each combination of the given web services.

In [65], authors present GATE Teamware, a collaborative text annotation framework based on web services, capable of dealing with several text processing tasks as corpus annotation via the use of web services workflows. Zhao et al [66] presented the Living Human Digital Library project, a set of web services that allows collaboration in the data-processing level on the biomedical field.

Table 1. State-of-the-art web-based platforms

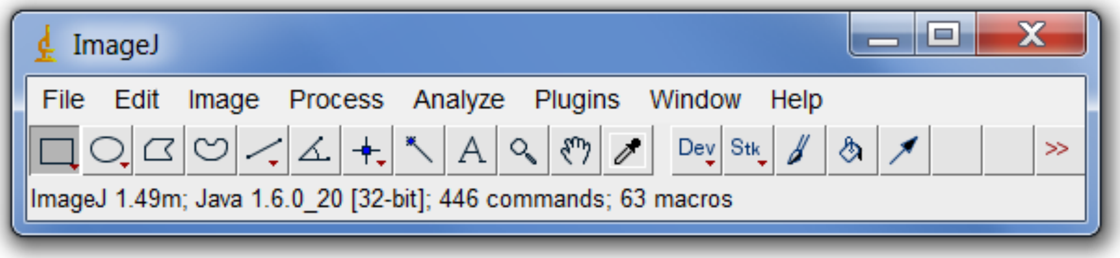
Reference	Data Type	Technology	Features
[46]	Generic	Grid Systems	Parallel execution
[47]	3D medical volumes	SOA-architecture	Reconstruction of 3D image
[48]	Microscopy Images	SOA-architecture	Data distribution - speed up
[49]	Retinal Images	SOA-architecture	Ophthalmology tools
[50]	Generic	Grid Systems	Image registration algorithms
[51]	Biomedical Images	Client-Server	Image Analysis Tools collection
[52]	Generic	SOA-architecture	Data distribution - speed up
[53]	Generic	Client-Server	Virtual Lab (tool collection)
[54]	Biomedical Images	Grid Systems	Parallel connections to the same instance of the framework
[55]	Generic	Client-Server	Algorithm development - Live debugging
[56]	Neuroimaging data	Grid Systems	Parallel execution
[57]	Production Engineering Activities	Web-Services	Production Engineering tool collection
[58]	Proetomic data	SOA-architecture	Data distribution - speed up
[59]	Generic	Web-Services	Intelligent Service Composition
[60]	Generic	SOA-architecture	Construction supply chains management
[61]	Genomics data	SOA-architecture	Intelligent flow management
[62]	Chemical data	SOA-architecture	Intelligent flow management
[63]	Chemical data	SOA-architecture	Several workflow patterns support
[64]	Chemical data	Web-Services	Data distribution - speed up
[65]	Text	SOA-architecture	Corpus integration
[66]	Biomedical Images	SOA-architecture	List of existing biomedical tools

2.2 Related Software

In this section we present the tools that were utilized in the development procedure of the proposed framework. A brief description of their features, along with their contribution to thesis goal is presented below.

2.2.1 ImageJ

ImageJ [67] is an application developed by National Institute of Health. It is able to run either as an applet, either as a downloadable application, or locally on any pc or mac, with Java 1.4 and above installed. ImageJ is supported by the Microsoft Windows, the Mac OS and Linux platforms. It can also run on a Sharp Zaurus PDA.



ImageJ 1.49m; Java 1.6.0_20 [32-bit]; 446 commands; 63 macros

Figure 2. ImageJ graphic user interface

ImageJ is a powerful image-processing tool written in Java and it is used widely in academic societies for scientific projects. More specifically, more than 1400 academic users have utilized it for scientific researches. This is a tool application, since it is open-source and it can be easily integrated to java-based application. Its open-source feature makes it eligible for rapid integration with many systems by utilizing web services technology.

It can analyze and process almost every type of known image file format. In more details, it can open and process TIFF, GIG, JPEG, BMP and ASCII images, while it can read FITS, PNG, PGM and DICOM images. JPEG, TIFF, GIF and DICOM images can be also accessed through a URL link. It can process the 8-bit gray scaled or indexed color, the 16-bit integer, the 32-bit floating-point and the RGB color images. Speaking of processing, it is the fastest image-processing tool in the world [68].

Apart from the standard functionality of viewing images, ImageJ provides a number of tools that satisfy common and advanced needs around image processing procedures. In cases where images contain salient objects that need to be distinct, ImageJ provides several image enhancement methods. Additionally, geometrical processes are available, in order to enhance the efficiency of the tool. There are also automated procedures, offering all the common image processing techniques. In cases where the image processing techniques must be focused on a specific object or area of the image, any of the provided image selection tools can be utilized. From a simple rectangular selection to a freehand selection, there is a great variety of selection tools (see Figure 3).

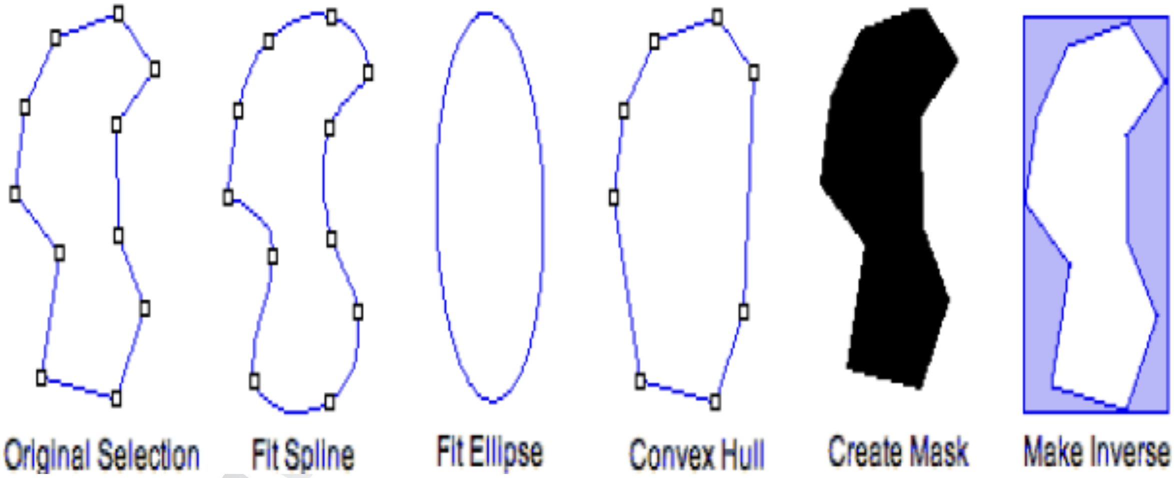


Figure 3. *Some of the selection tools provided by ImageJ*

The java code of ImageJ is written in such a way, in order to extend its features through the plugin applications and the macro-commands that also supports. A plugin consists of standardized java based functions, where users may develop their own code and integrate it, without any further programming knowledge, into the ImageJ main functionality enhancing its capabilities.

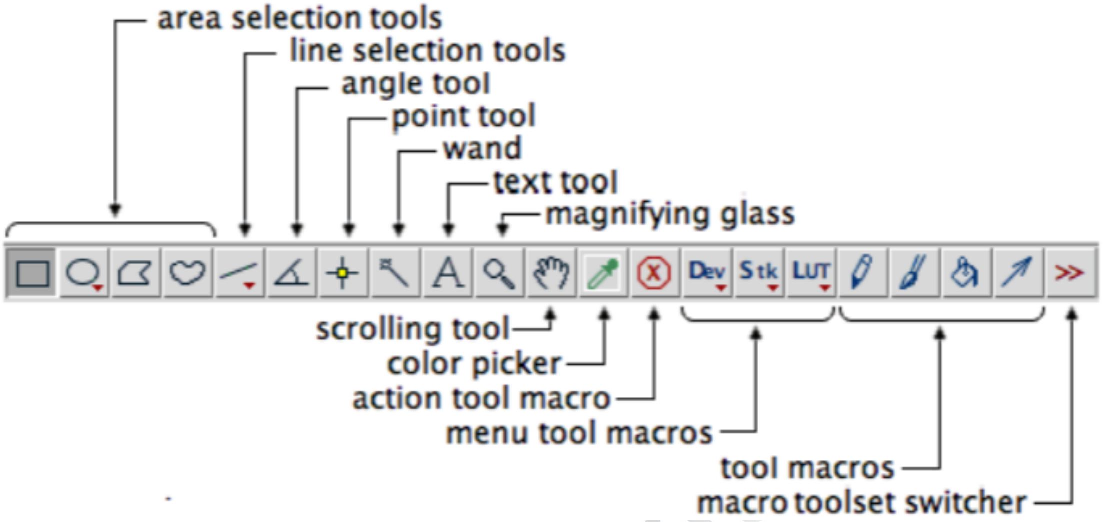


Figure 4. The available ImageJ tools

Contribution to thesis Goal

ImageJ is written in Java as it mentioned above. This makes the integration of every ImageJ functionality to the developed framework. Although ImageJ is a powerful tool, it contains the basic image analysis functionalities, which are accessible through its GUI or through Java code. The proposed framework contains additional advanced image analysis tools and requires no programming skills to access its features. Additionally, the proposed framework automates the image characterization procedures. Thus, large amount of Images can be characterized automatically. The proposed framework can also be enhanced with additional functionalities through the web-services technology.

2.2.2 Weka

Weka [69] is a well-known machine learning software that was developed in New Zealand’s Waikato University. It is equipped with a number of data preprocessing and machine learning algorithms, while it contains capable visualization tools, which enable the data mining and visualization functionalities. The Weka application is written in Java and it remains open-source under a GPL license.

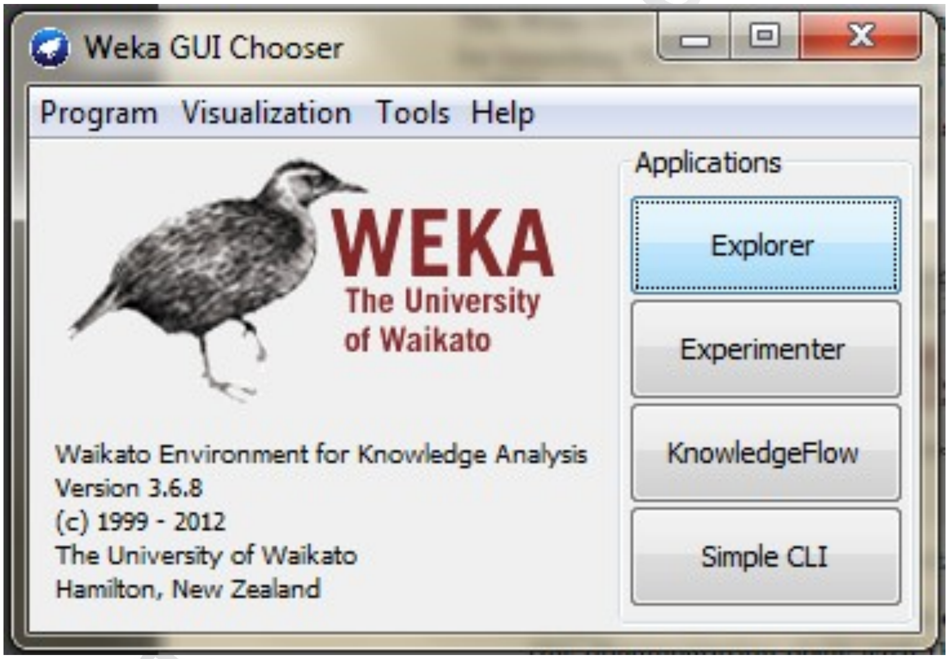
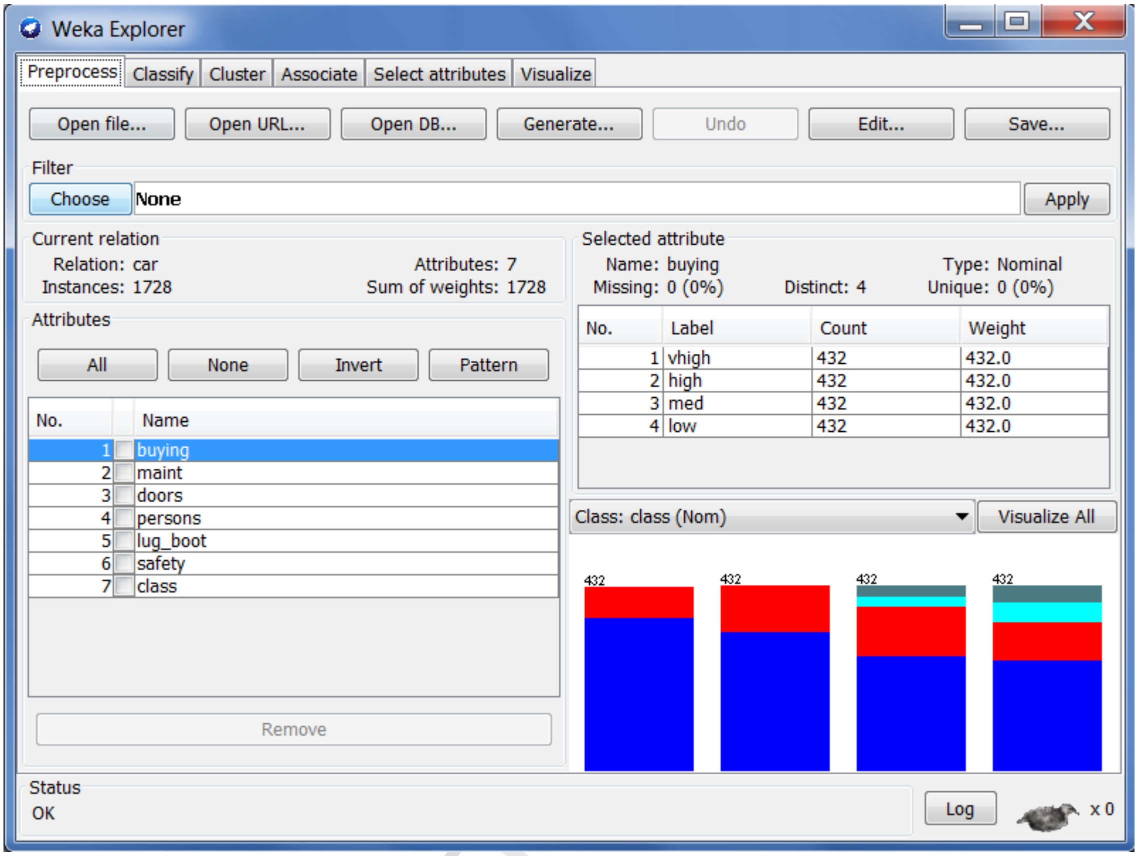

Weka GUI Chooser
Program Visualization Tools Help
Applications
Explorer
Experimenter
KnowledgeFlow
Simple CLI
WEKA
The University of Waikato
Waikato Environment for Knowledge Analysis
Version 3.6.8
(c) 1999 - 2012
The University of Waikato
Hamilton, New Zealand

Figure 5. The graphic user interface chooser of weka

The data can be imported through ARFF, CSV, DATA and BSI files. Weka contains multiple graphic user interfaces (GUIs). Each of them holds different capabilities. The available GUIs are the following:

Explorer: the most commonly used GUI of the application (see Figure 6). It contains all the available data preprocessing and machine learning algorithms. It also contains visualization features for further data analysis. It supports the preprocessing, the classification, the clustering, the feature selection and the visualization of the provided data.



Filter

None

Current relation

Relation: car
Attributes: 7
Instances: 1728
Sum of weights: 1728

Selected attribute

Name: buying
Type: Nominal
Missing: 0 (0%)
Distinct: 4
Unique: 0 (0%)

No.	Label	Count	Weight
1	vhigh	432	432.0
2	high	432	432.0
3	med	432	432.0
4	low	432	432.0

Attributes

No.		Name
1	□	buying
2	□	maint
3	□	doors
4	□	persons
5	□	lug_boot
6	□	safety
7	□	class

Class: class (Nom)

Status

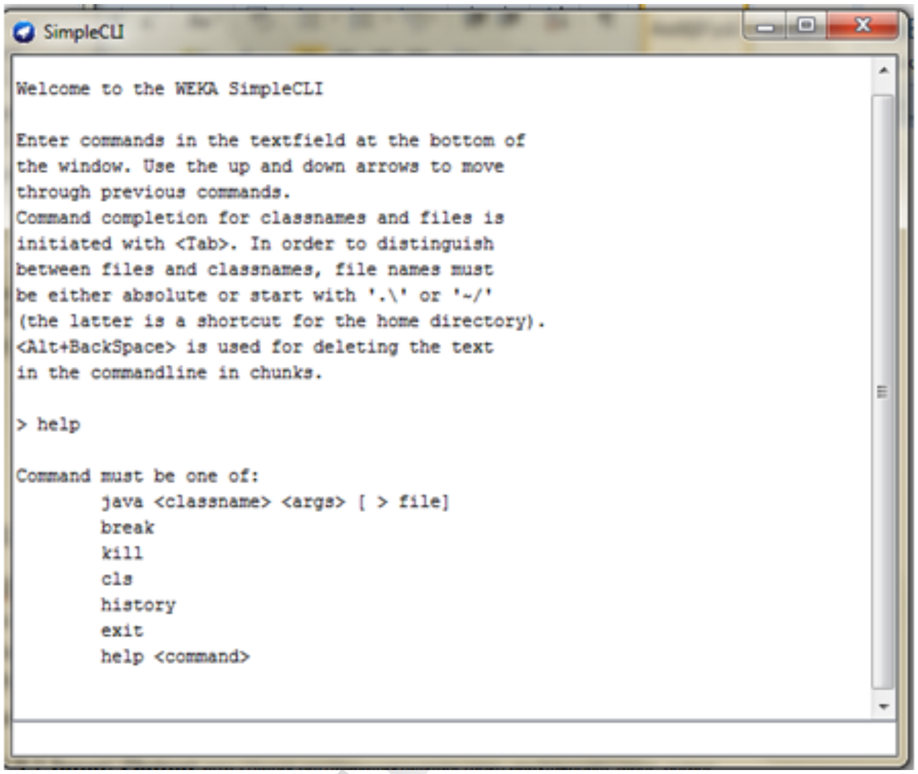
OK

Figure 6. The Explorer GUI of the Weka application

Experimenter: a GUI enabling statistical tests on learning algorithms and advanced analysis of the results. It is available in two versions.

The simple GUI version allows access to the standard functionalities.

The advanced GUI version enables the utilization of advanced functionalities. Both versions enable the basic experimentation functionalities, which can be established locally or remotely (to another computer). The result outputs are in the form of ARFF or CSV file format, while the JDBC database output option is available.

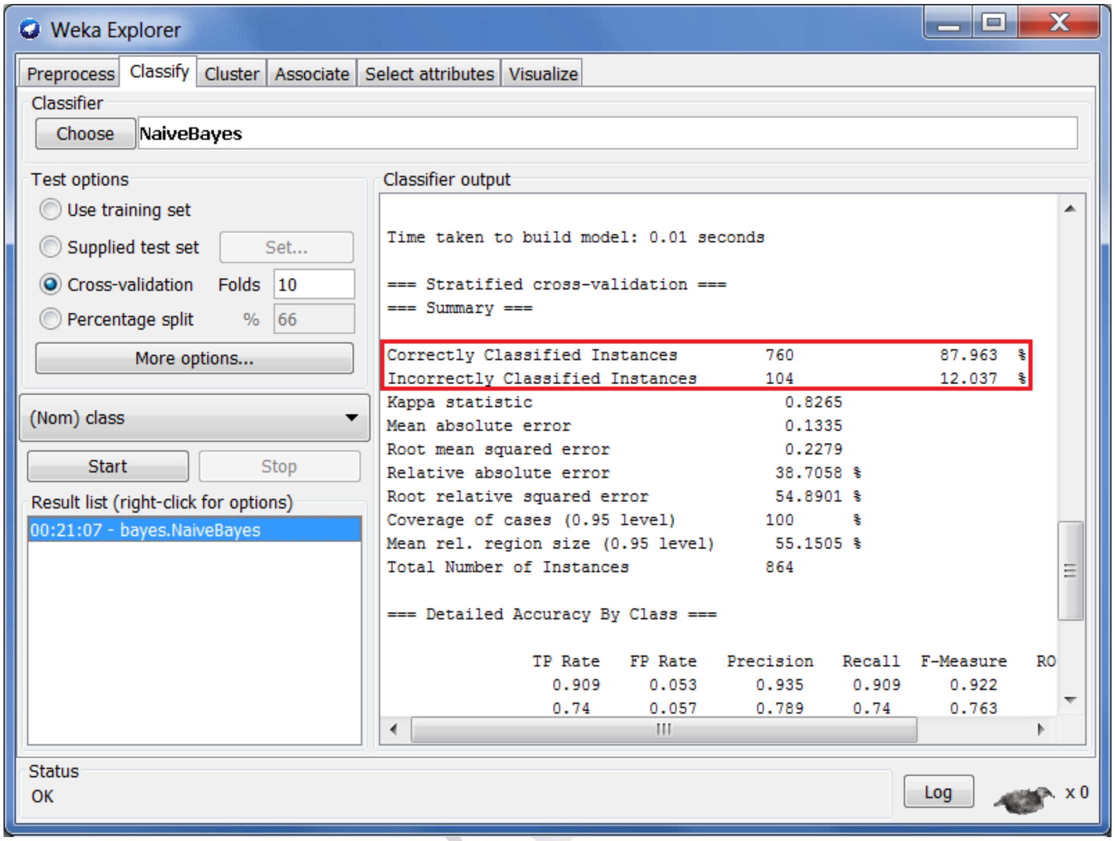


```
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with '.\' or '~/'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

> help

Command must be one of:
        java <classname> <args> [ > file]
        break
        kill
        cls
        history
        exit
        help <command>
```

Figure 7. The command prompt window of Weka (Simple CLI GUI)

Knowledge Flow: a GUI enabling drag n' drop capabilities combined with the Explorer's features. Each machine learning algorithm, data preprocessing procedure and add-on (plugin) is represented by an operator object.

Simple CLI: a command prompt GUI (see Figure 7) that enables the utilization of Weka's machine learning algorithms through commands and macros. It provides access to any classifier, clusterer or filter contained in Weka application. The corresponding results are exported in a terminal style form (inside the command prompt).



Classifier: NaiveBayes

Test options:

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Result list (right-click for options)

00:21:07 - bayes.NaiveBayes

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         760               87.963  %
Incorrectly Classified Instances       104               12.037  %
Kappa statistic                          0.8265
Mean absolute error                      0.1335
Root mean squared error                  0.2279
Relative absolute error                 38.7058 %
Root relative squared error             54.8901 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      55.1505 %
Total Number of Instances              864

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  RO
                 0.909    0.053    0.935     0.909    0.922
                 0.74     0.057    0.789     0.74     0.763
```

Status

OK

Figure 8. Result window of the Explorer GUI

Contribution to thesis Goal

Weka is a powerful machine-learning platform. Since it is written in Java, it was easily integrated to our proposed system. Weka offers advanced machine learning methods that through their integration with the proposed framework will have their representative ontology-based operator. This will enable the smart combination of machine learning algorithms along with advances image analysis technques. This feature is robust, since the combination of machine learning along with image analysis for an image-mining task, is a difficult procedure that requires deep image analysis knowledge and advance programming skills.

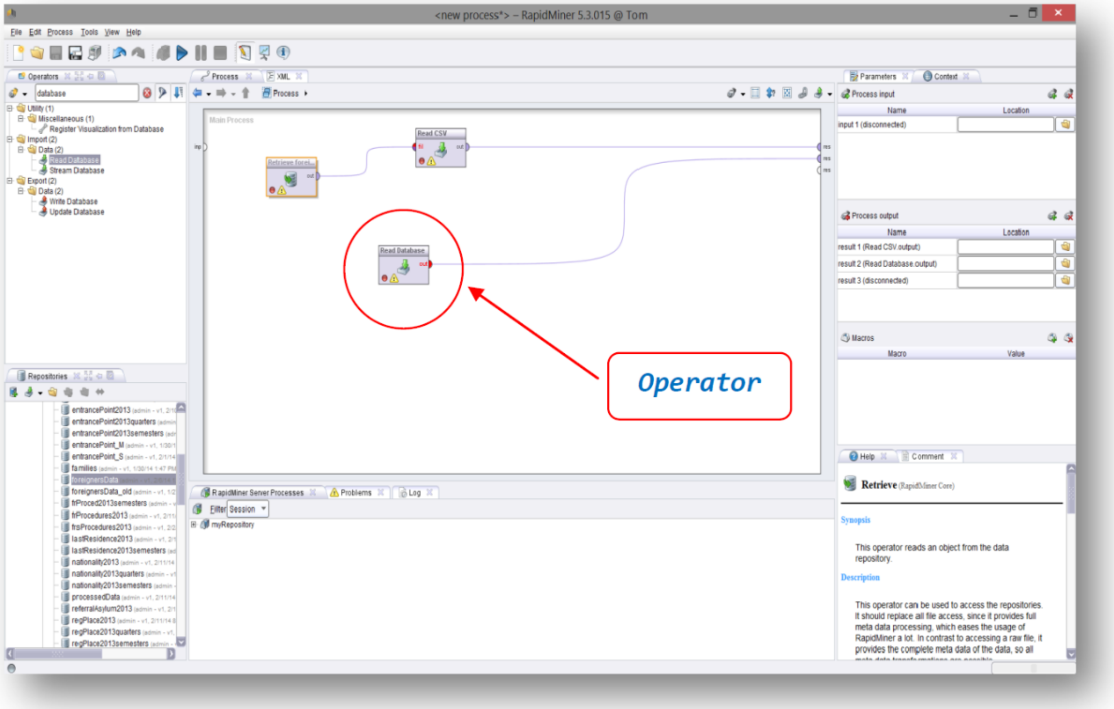
2.2.3 RapidMiner

RapindMiner [70], formerly known as YALE app (Yet Another Learning Environment), is a machine learning environment, enabling data mining and predictive and business analysis. It is widely used for research purposes, academic training, software development and other applications.

RapidMiner was among the most popular data mining application in 2009 [71], while in 2010 conquered the first place. It is a Java based open source software under an AGPL license. The application was developed by Ralf Klinkenberg, Ingo Mierswa and Simon Fisher in 2001. Ingo Mierswa and Ralf Klinkenberg founded Rapid-I in 2006.

RapidMiner has multiple applications in many research fields. These are biomedical engineering, energy consumption, telecommunications, banking services, informatics, information technology, etc.

The application can extend its features through the application of additional plugins. At the moment, there are 16 official plugins that enhance the performance of the tool by focusing on the following domains: Image analysis, text mining, time series analysis, web mining, optimal processing distribution, parallel processing, data visualization, datastream and multimedia mining, tracking drifting concepts etc.



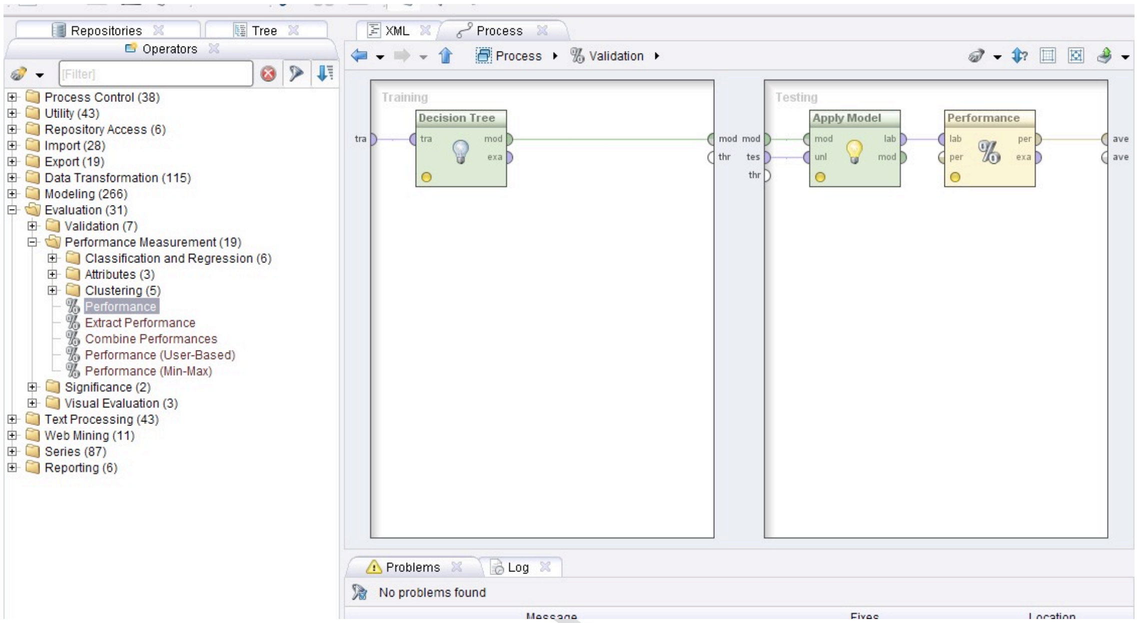
This operator reads an object from the data repository.

This operator can be used to access the repositories. It should replace all file access, since it provides full meta data processing, which eases the usage of RapidMiner a lot. In contrast to accessing a raw file, it provides the complete meta data of the data, so all

Figure 9. RapidMiner Operator example

Each of the data processing and mining capabilities of the platform are represented by an operator object, which contains some input and/or output ports. Additionally, the available operator objects represent file inputs, models, evaluators and each other functionality that can be integrated to RapidMiner. Therefore, the RapidMiner platform allows the creation of workflow chains for a large amount of machine learning tasks. Thus, every processing step of the data is accessible by the users. Some specific operators contain internal operators. Evaluation operators are usually in this form (containing internal training and testing operators – see Figure 10).

RapidMiner's operators (see Figure 9) support a variety of machine learning algorithms, all the current functionalities of Weka, data pre-processing, feature extraction and selection algorithms, meta-operator functionalities (that optimize the exported results), evaluation processed and visualization capabilities.



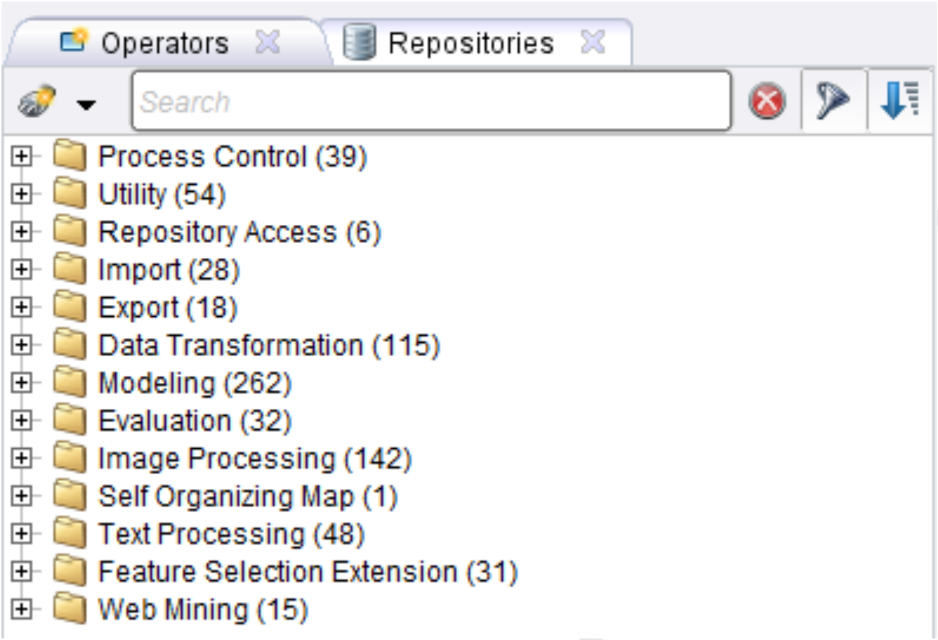
No problems found

Figure 10. RapidMiner Performance Operator

Data and workflow schemes can be represented via an XML file. This XML file contains the necessary information, in order to represent the exact constructed workflow scheme designed on the RapidMiner platform. This feature (see Figure 12) enables the easy transfer of the constructed workflows from one personal computer to another. Except the XML file format, RapidMiner can accept other type of files, as well (i.e. arff, csv, bibtex, C4.5, dBase etc.).

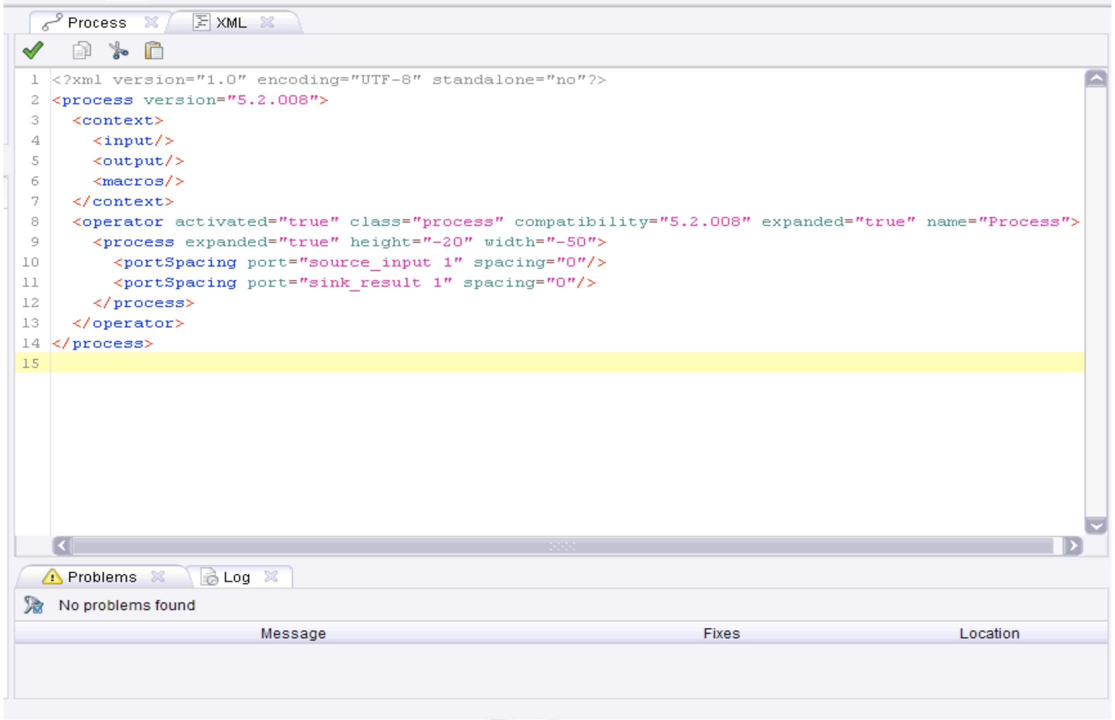
Process validation feature ensures the stability of the constructed workflow schemes, since it denotes the incorrectly connected operators and proposes the default solution, in order to execute the constructed scheme. Additionally, checkpoint operators can be placed between the operators, so as to achieve the debugging of the constructed flow. Thus, the construction of complex mining tasks becomes easier.

Following the execution of a workflow scheme, the results view panel of the RapidMiner denotes the exported results. This panel contains the final and the checkpoint results.



- Process Control (39)
- Utility (54)
- Repository Access (6)
- Import (28)
- Export (18)
- Data Transformation (115)
- Modeling (262)
- Evaluation (32)
- Image Processing (142)
- Self Organizing Map (1)
- Text Processing (48)
- Feature Selection Extension (31)
- Web Mining (15)

Figure 11. RapidMiner's operators menu

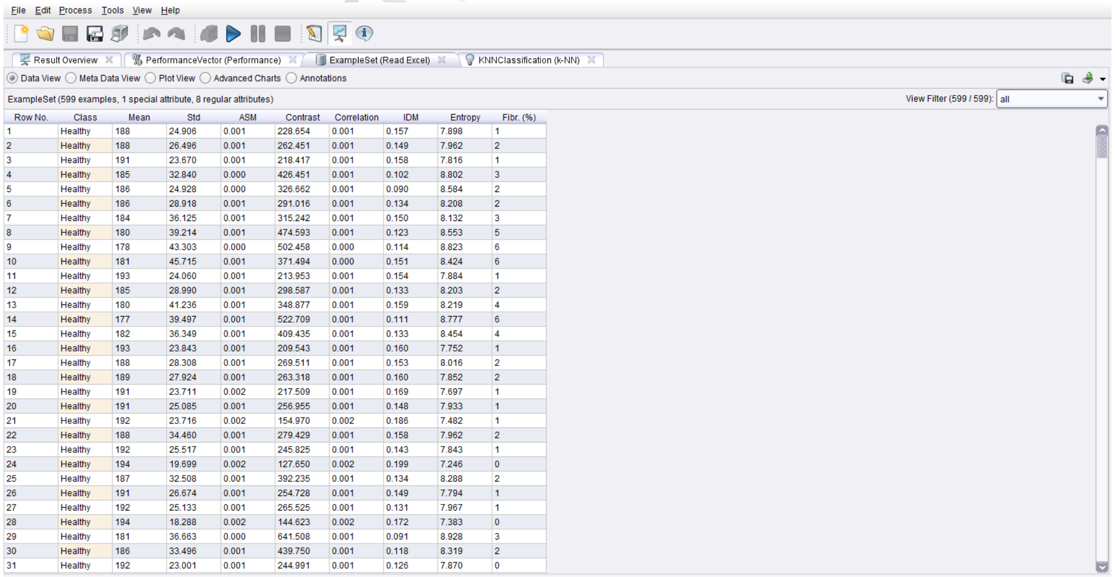


```xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.2.008">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.2.008" expanded="true" name="Process">
    <process expanded="true" height="-20" width="-50">
      <portSpacing port="source_input 1" spacing="0"/>
      <portSpacing port="sink_result 1" spacing="0"/>
    </process>
  </operator>
</process>
```

No problems found

Message	Fixes	Location

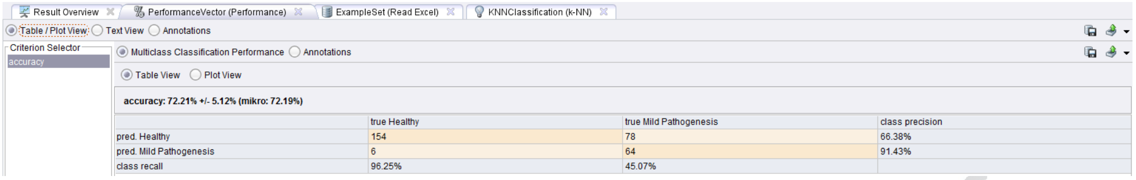
Figure 12. RapidMiner's XML representation mode


File Edit Process Tools View Help
Result Overview
PerformanceVector (Performance)
ExampleSet (Read Excel)
KNNClassification (k-NN)
● Data View ○ Meta Data View ○ Plot View ○ Advanced Charts ○ Annotations
View Filter (599 / 599): all

ExampleSet (599 examples, 1 special attribute, 8 regular attributes)

Row No.	Class	Mean	Std	ASM	Contrast	Correlation	IDM	Entropy	Fibr. (%)
1	Healthy	188	24.906	0.001	228.654	0.001	0.157	7.898	1
2	Healthy	188	26.496	0.001	262.451	0.001	0.149	7.962	2
3	Healthy	191	23.670	0.001	218.417	0.001	0.158	7.816	1
4	Healthy	185	32.840	0.000	426.451	0.001	0.102	8.802	3
5	Healthy	186	24.928	0.000	326.662	0.001	0.090	8.584	2
6	Healthy	186	28.918	0.001	291.016	0.001	0.134	8.208	2
7	Healthy	184	36.125	0.001	315.242	0.001	0.150	8.132	3
8	Healthy	180	39.214	0.001	474.593	0.001	0.123	8.553	5
9	Healthy	178	43.303	0.000	502.458	0.000	0.114	8.823	6
10	Healthy	181	45.715	0.001	371.494	0.000	0.151	8.424	6
11	Healthy	193	24.060	0.001	213.953	0.001	0.154	7.884	1
12	Healthy	185	28.990	0.001	298.587	0.001	0.133	8.203	2
13	Healthy	180	41.236	0.001	348.877	0.001	0.159	8.219	4
14	Healthy	177	39.497	0.001	522.709	0.001	0.111	8.777	6
15	Healthy	182	36.349	0.001	409.435	0.001	0.133	8.454	4
16	Healthy	193	23.843	0.001	209.543	0.001	0.160	7.752	1
17	Healthy	188	28.308	0.001	269.511	0.001	0.153	8.016	2
18	Healthy	189	27.924	0.001	263.318	0.001	0.160	7.852	2
19	Healthy	191	23.711	0.002	217.509	0.001	0.169	7.697	1
20	Healthy	191	25.085	0.001	256.955	0.001	0.148	7.933	1
21	Healthy	192	23.716	0.002	154.970	0.002	0.186	7.482	1
22	Healthy	188	34.460	0.001	279.429	0.001	0.158	7.962	2
23	Healthy	192	25.517	0.001	245.825	0.001	0.143	7.843	1
24	Healthy	194	19.699	0.002	127.650	0.002	0.199	7.246	0
25	Healthy	187	32.508	0.001	392.235	0.001	0.134	8.288	2
26	Healthy	191	26.674	0.001	254.728	0.001	0.149	7.794	1
27	Healthy	192	25.133	0.001	265.525	0.001	0.131	7.967	1
28	Healthy	194	18.288	0.002	144.623	0.002	0.172	7.383	0
29	Healthy	181	36.663	0.000	641.508	0.001	0.091	8.928	3
30	Healthy	186	33.496	0.001	439.750	0.001	0.118	8.319	2
31	Healthy	192	23.001	0.001	244.991	0.001	0.126	7.870	0

Figure 13. RapidMiner's Result View


Result Overview
PerformanceVector (Performance)
ExampleSet (Read Excel)
KNNClassification (k-NN)
● Table / Plot View ○ Text View ○ Annotations
Criterion Selector
accuracy
● Multiclass Classification Performance ○ Annotations
● Table View ○ Plot View

accuracy: 72.21% +/- 5.12% (mikro: 72.19%)

	true Healthy	true Mild Pathogenesis	class precision
pred. Healthy	154	78	66.38%
pred. Mild Pathogenesis	6	64	91.43%
class recall	96.25%	45.07%	

Figure 14. RapidMiner's Result Mode - Confusion Matrix example

Contribution to thesis Goal

RapidMiner is a robust machine learning operator-based tool. Practically, we didn't integrate its libraries to our proposed framework. We adopted its operator-based architecture and the way that presents the data to the user. Our proposed framework offers robust ontology functionalities (such as, multiple workflow generation, optimal workflow selection, etc.), which, to the best of our knowledge, there aren't in any other workflow-based platform.

2.2.4 Taverna

Taverna [72] is a workflow manager that supports the construction and execution of scientific workflow-based processes. Taverna was developed under myGrid [73] project and it is written in Java and it is open source under a LGPL license. Taverna is a domain independent application and it is supported by the Microsoft Windows, the Mac OS and Linux platforms.

Taverna workbench was utilized by scientific applications of different scientific fields, such as medicine, bio-chemistry, astronomy, music, bioinformatics, etc.

Taverna workbench is utilized by over 300 academic and commercial organizations. A number of services is available via BioCatalogue [74], a public repository of Life Science Web services. Additionally, Taverna users can share their developed web services through myExperiment [75], a scientific web site. Both, myExperiment and BioCatalogue are products of the myGrid project.

Taverna's architecture consists of three parts: The Workbench, the Enactor and the Simple Conceptual Unified Language (SCUFL).

Taverna workbench

Workbench accesses various available web services the same way a web browser accesses web sites. Through the workbench, web services can be integrated to a constructed workflow as a drag 'n drop eligible operator (see Figure 15). The operator connects with another operator with a drag 'n drop procedure, as well.

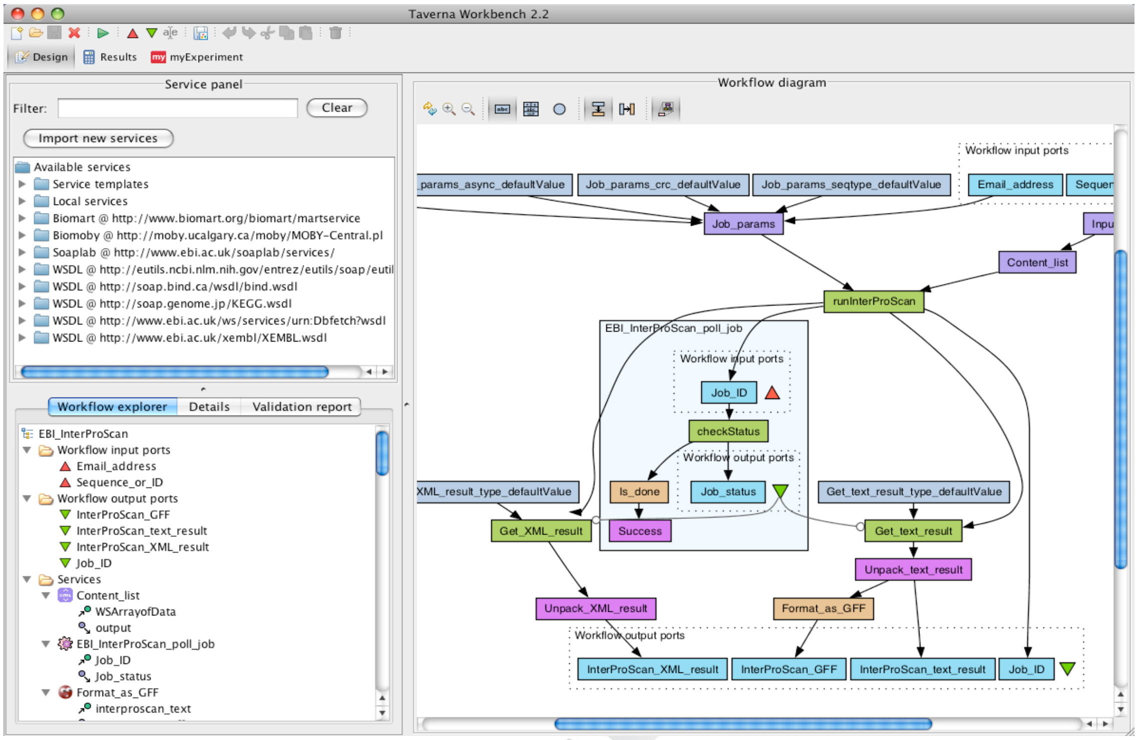


Figure 15. Taverna Wofkflow Scheme example

The workbench contains a number of automated methods that facilitate the experimental procedures of the users' constructed workflows. Taverna's included automated methods are accessed locally or remotely (via the utilization of web services technology).

Taverna workbench allows users to integrate their own applications, through the web services technology, in the form of Taverna operators. This feature makes eligible the construction of complex workflows, related to data processing, even from users with fundamental programming skills or with limited hardware resources.

The application can invoke both SOAP and RESTful web services and integrate them into its workbench in the form of editable workflow operators. R statistical services can be also integrated. Taverna enables the live monitoring of the running workflow scheme, denoting each operator as currently running, successfully executed or erroneous. Thus, becomes easy for the users to understand, which part of the workflow is erroneous.

Taverna Enactor

Taverna Enactor is basically the heart of Taverna application, since it is responsible for the execution of the workflow processes and for the corresponding data input and output stages. The enactor is the core of the entire application; the workflow construction is achieved through the workbench.

Simple Conceptual Unified Language (SCUFL)

Simple Conceptual Unified Language (SCUFL) is the utilized language for the programming representation of the workflow. This is an XML based language, which also has a graphical representation through the Workbench. The Taverna application doesn't need to have SCULFL programming knowledge, since they interact with the graphical user interface, which provides full access to the integrated services.

Contribution to thesis Goal

Taverna workbench was selected as the core platform for our proposed framework among the other workflow managers that support scientific data streams and can adapt their modeled processes to the needs of the biomedical image-mining field. Specifically, we exploited the application's infrastructure, in order to integrate the advanced image analysis and machine learning techniques. Additionally, we integrated our novel techniques (such as, multiple workflow generation, optimal workflow selection, etc.), which – to the best of our knowledge – they cannot be found to similar frameworks.

2.3 References

[1] P. Phukpattaranont, S. Limsiroratana, P. Boonyaphiphat. Computer-Aided System for Microscopic Images: Application to Breast Cancer Nuclei Counting. International Journal of Applied Biomedical Engineering vol.2, no.1 2009 pp 69-74.

[2] Maglogiannis, H. Sarimveis, C. Kiranoudis, A.A. Chatzioannou, N. Oikonomou, V. Aidinis, "Radial Basis Function neural networks classification for the recognition of idiopathic pulmonary fibrosis in microscopic images." *IEEE Transactions on Information Technology in Biomedicine*, 12(1), pp. 42-54, 2008.

[3] A.B. Tosun, C. Gunduz-Demir, "Graph Run-Length Matrices for Histopathological Image Segmentation." *IEEE Transactions on Medical Imaging,* vol. 30, No.3 March pp. 721-732, 2011.

[4] S. Issac Niwas, P. Palanisamy, K. Sujathan, E. Bengtsson, Analysis of nuclei textures of fine needle aspirated cytology images for breast cancer diagnosis using Complex Daubechies wavelets, Signal Processing, Volume 93, Issue 10, October 2013, Pages 2828-2837

[5] Chen X, Zhou X, Wong ST. Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy, IEEE Trans Biomed Eng. 2006 Apr;53(4):762-6.

[6] Lindblad, J., Wählby, C., Bengtsson, E. and Zaltsman, A. (2004), Image analysis for automatic segmentation of cytoplasms and classification of Rac1 activation. Cytometry, 57A: 22–33. doi: 10.1002/cyto.a.10107

[7] Hiremath, P.S. ; Gulbarga Univ., Gulbarga ; Iranna Y., H. Automated Cell Nuclei Segmentation and Classification of Squamous Cell Carcinoma from Microscopic Images of Esophagus Tissue. Advanced Computing and Communications, 2006. ADCOM 2006. International Conference on . pp 211-216.

[8] Al-Kofahi, Y. ; Dept. of Electr., Comput. & Syst. Eng. (ECSE), Rensselaer Polytech. Inst., Troy, NY, USA ; Lassoued, W. ; Lee, W. ; Roysam, B. Improved Automatic Detection and Segmentation of Cell Nuclei in Histopathology Images. Biomedical Engineering, IEEE Transactions on (Volume:57 , Issue: 4). April 2010. Pp.841-852.

[9] Zhongyu Xu ; Coll. of Comput. Sci. & Eng., Changchun Univ. of Technol., Changchun, China ; Fen Hu ; Hongcheng Guo ; Quansheng Dou. Support vector

machine image segmentation algorithm applied to angiogenesis quantification. Natural Computation (ICNC), 2010 Sixth International Conference on (Volume:2) pp. 928-931.

[10] Wu, H Fiel, M.I. ; Schiano, T.D. ; Ramer, M. ; Burstein, D. ; Gil, J. Segmentation of textured cell images based on frequency analysis. Image Processing, IET (Volume:5 , Issue: 2) March 2011. Pp.148-158.

[11] Chaabane SB, Fnaiech F. Color edges extraction using statistical features and automatic threshold technique: application to the breast cancer cells. BioMedical Engineering OnLine 2014, 13:4 doi:10.1186/1475-925X-13-4.

[12] Sagonas C, Marras I , Kasampalidis I, Pitas I, Lyroudia K, Karayannopoulou G. FISH image analysis using a modified radial basis function network. Volume 8, Issue 1, January 2013, Pages 30–40

[13] El Abbadi, N.K., Miry, A.H. Automatic segmentation of skin lesions using histogram thresholding (2014) Journal of Computer Science, 10 (4), pp. 632-639.

[14] Lallas, A., Tzellos, T., Kyrgidis, A., Apalla, Z., Zalaudek, I., Karatolias, A., Ferrara, G., Piana, S., Longo, C., Moscarella, E., Stratigos, A., Argenziano, G. Accuracy of dermoscopic criteria for discriminating superficial from other subtypes of basal cell carcinoma (2014) Journal of the American Academy of Dermatology, 70 (2), pp. 303-311.

[15] Longo, C., Lallas, A., Kyrgidis, A., Rabinovitz, H., Moscarella, E., Ciardo, S., Zalaudek, I., Oliviero, M., Losi, A., Gonzalez, S., Guitera, P., Piana, S., Argenziano, G., Pellacani, G. Classifying distinct basal cell carcinoma subtype by means of dermatoscopy and reflectance confocal microscopy (2014) Journal of the American Academy of Dermatology, . Article in Press.

[16] Madooei, A., Drew, M.S. A probabilistic approach to quantification of melanin and hemoglobin content in dermoscopy images. (2014) Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, 17 (Pt 1), pp. 49-56.

[17] Rand, D., Walsh, E.G., Derdak, Z., Wands, J.R., Rose-Petruck, C. A highly sensitive x-ray imaging modality for hepatocellular carcinoma detection in vitro (2015) Physics in Medicine and Biology, 60 (2), pp. 769-784.

[18] Boyaci, N., Dokumaci, D.S., Karakas, E., Yalcin, F., Kurnaz, A.G.O. Paratracheal air cysts: Prevalence and relevance to pulmonary emphysema and bronchiectasis using thoracic multidetector CT (2015) Diagnostic and Interventional Radiology, 21 (1), pp. 42-46.

[19] Lou, W., Gu, J., Su, H., He, X., Chen, L., Chen, G., Song, J., Shi, W., Wang, T., Zhao, B. Hemodynamic changes of lower extremity peripheral arterial disease following interventional therapy: primary application of iFlow quantitative evaluation (2015) Chinese Journal of Radiology (China), 49 (1), pp. 57-60.

[20] Degenhart, C., Strube, H., Betz, M.J., Pallauf, A., Bidlingmaier, M., Fischer, E., Reincke, M., Reiser, M.F., Wirth, S. CT mapping of the vertebral level of right adrenal vein (2015) Diagnostic and Interventional Radiology, 21 (1), pp. 60-66.

[21] Kranz, R., Cosson, P. Anatomical and/or pathological predictors for the "incorrect" classification of red dot markers on wrist radiographs taken following trauma(2015) British Journal of Radiology, 88 (1046), art. no. 20140503.

[22] Sotaquira, M., Pepi, M., Fusini, L., Maffessanti, F., Lang, R.M., Caiani, E.G. Semi-automated Segmentation and Quantification of Mitral Annulus and Leaflets from Transesophageal 3-D Echocardiographic Images(2015) Ultrasound in Medicine and Biology, 41 (1), pp. 251-267.

[23] Lorenzi, M., Pennec, X., Frisoni, G.B., Ayache, N. Disentangling normal aging from Alzheimer's disease in structural magnetic resonance images(2015) Neurobiology of Aging, 36 (S1), pp. S42-S52.

[24] Išgum, I., Benders, M.J.N.L., Avants, B., Cardoso, M.J., Counsell, S.J., Gomez, E.F., Gui, L., Huppi, P.S., Kersbergen, K.J., Makropoulos, A., Melbourne, A., Moeskops, P., Mol, C.P., Kuklisova-Murgasova, M., Rueckert, D., Schnabel, J.A., Srhoj-Egekher, V., Wu, J., Wang, S., de Vries, L.S., Viergever, M.A. Evaluation of automatic neonatal brain segmentation algorithms: The NeoBrainS12 challenge (2015) Medical Image Analysis, 20 (1), pp. 135-151.

[25] Gubern-Mérida, A., Martí, R., Melendez, J., Hauth, J.L., Mann, R.M., Karssemeijer, N., Platel, B. Automated localization of breast cancer in DCE-MRI(2015) Medical Image Analysis, 20 (1), pp. 265-274.

[26] Gubern-Mérida, A., Kallenberg, M., Mann, R.M., Martí, R., Karssemeijer, N. Breast segmentation and density estimation in breast MRI: A fully automatic

framework (2015) IEEE Journal of Biomedical and Health Informatics, 19 (1), art. no. 6762834, pp. 349-357.

[27] Roy, S., Ghosh, P., Bandyopadhyay, S.K. Contour extraction and segmentation of cerebral hemorrhage from MRI of brain by gamma transformation approach(2015) Advances in Intelligent Systems and Computing, 328, pp. 383-394.

[28] Srinivasan, S.V., Narasimhan, K., Balasubramaniyam, R., Rishi Bharadwaj, S. Diagnosis and segmentation of brain tumor from MR image (2015) Advances in Intelligent Systems and Computing, 325, pp. 687-693.

[29] Khalvati, F., Gallego-Ortiz, C., Balasingham, S., Martel, A.L. Automated segmentation of breast in 3-D MR images using a robust atlas(2015) IEEE Transactions on Medical Imaging, 34 (1), art. no. 6878440, pp. 116-125.

[30] Mitra, A., De, A., Bhattacharjee, A.K. MRI skull bone lesion segmentation using Distance based Watershed segmentation (2015) Advances in Intelligent Systems and Computing, 328, pp. 255-261.

[31] Wang P et al. (2001) Classification of Endoscopic images based on Texture and Neural Networks. Biomedical Engineering Research Center. Nanyang technological university, Singapore, 639798, Accession number ADA409511.

[32] LaConte S et al. (2005) Support Vector Machines for temporal classification of block fMRI data. Biomedical Engineering, Georgia Institute of Technology, Emory University, 531 Asbury Circle, Suite N305, Atlanta, GA 30322, USAdsds Neurolmage, volume 26, issue 2, pp. 317-329.

[33] Olmez T, Dokur Z.(2003) Classification of heart sounds using an artificial neural network. Department of Electronics and Communication Engineering, Istanbul Technical University, 80626 Maslak, Istanbul, Turke, Pattern Recognition Letters v.24, issues 1-3, p617-629.

[34] Wolberg W, Street W, and Mangasarian O. (1993) Breast cytology diagnosis with digital image analysis. Analytical Quantitative Cytology Histology, vol. 15 no.6 pp. 396-404.

[35] Barr M, McClellan S, Winokur T, and Vaughn. (2004) An automated Tissue Preclassification Approach for Telepathology: Implementation and

Performance Analysis. IEEE Transaction on Information Technology in Biomedicine, vol. 8 no 2, pp. 97-102.

[36] Masseroli M, Caballero T, O'Valle F, Del Moral R M G, Perrez-Milena, and Del Moral. (2000) Automatic quantification of liver fibrosis: design and validation of a new image analysis method. Comparison with semi-quantitative indexes of fibrosis. Journal of Hepatology, vol. 32, pp. 453-464.

[37] T. Caballero, A. Pérez-Milena, M. Masseroli, F. O'Valle, F. J. Salmerón, R. M. G. Del Moral, and G. Sánchez-Salgado, "Liver fibrosis assessment with semiquantitative indexes and image analysis quantification in sustained-responder and non-responder," *Journal of Hepatology*, vol. 34, pp. 740-747, 2001.

[38] M. Masseroli, T. Caballero, F. O'Valle, R. M. G. Del Moral, A. Pérrez-Milena, and R. G. Del Moral, "Automatic quantification of liver fibrosis: design and validation of a new image analysis method. Comparison with semi-quantitative indexes of fibrosis.," *Journal of Hepatology*, vol. 32, pp. 453-464, 2000.

[39] M. Yagura, S. Murai, H. Kojima, H. Tokita, H. Kamitsukasa, and H. Harada, "Changes of liver fibrosis in chronic hepatitis C patients with no response to interferon-α therapy: including quantitative assessment by a morphometric method.," *Journal of Gastroenterology*, vol. 35, pp. 105-111, 2000.

[40] P. Bedossa, D. Dargere, and V. Paradis, "Sampling variability of liver fibrosis in chronic hepatitis C.," *Hepatology*, vol. 38 no. 6, pp. 1449-1457, 2003.

[41] M. A. Hussain, S. N. Merchant, L. S. Mombasawala, and R. R. Puniyani, "A decrease in effective diameter of rat mesenteric venules due to leukocyte margination after a bolus injection of pentoxifylline—digital image analysis of an intravital microscopic observation.," *Microvascular Research*, vol. 67, pp. 237-244, 2004.

[42] K. Shiraishi, H. Takihara, and K. Naito, "Quantitative analysis of testicular interstitial fibrosis after vasectomy in humans," *Aktuelle Urol.*, vol. 34 no. 4, pp. 262-264, 2003.

[43] K. Shiraishi, H. Takihara, and K. Naito, "Influence of interstitial fibrosis on spermatogenesis after vasectomy and vasovasostomy," *Contraception*, vol. 65 no. 3, pp. 245-249, 2002.

[44] G. Izbiki, M. J. Segel, T. G. Christensen, M. W. Conner, and B. R., "Time course of bleomycin-induced lung fibrosis," *Int J. Exp. Path*, vol. 83, pp. 111-119, 2002.

[45] C. Doukas, I. Maglogiannis, A. Chatziioannou, “Computer Supported Angiogenesis Quantification Using Image Analysis and Statistical Averaging” IEEE Transactions on Information Technology in Biomedicine 12 (2008), pp. 650-658.

[46] D. Kreftin, M. Vossberg, A. Hoheisel, T. Tolxdorff. “Simplified implementation of medical image processing algorithms into a grid using a workflow management system” Journal Future Generation Computer Systems, Volume 26 Issue 4, April, 2010 Pages 681-684

[47] R. Kooper, A. Shirk, Sang-Chul Lee, A. Lin, R. Folberg, P. Bajcsy. “3Dmedicalvolumereconstruction using webservices”. Computers in Biology and Medicine, Volume 38, Issue 4, April 2008, Pages 490–500

[48] S. Koulouzis, E. Zudilova-Seinstra, A. Belloum. “Data Transport between Visualization Web Services for Medical Image Analysis” Procedia Computer Science Volume 1, Issue 1, May 2010, Pages 1727–1736 ICCS 2010

[49] [4] M. Ortega, N. Barreira, J. Novo, M.G. Penedo, A. Pose-Reino, F. Gσmez-Ulla. “Sirius: A web-based system for retinal image analysis”. International Journal of Medical Informatics Volume 79, Issue 10, October 2010, Pages 722–732

[50] T. Glatard, J. Montagnat, X. Pennec. “Grid-enabled workflows for data intensive medical applications”. Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium,June 2005 pp. 537 - 542

[51] S. D. Olabarriaga, J. G. Snel, C. P. Botha, and R. G. Belleman. “Integrated Support for Medical Image Analysis Methods: From Development to Clinical Application”. Information Technology in Biomedicine, IEEE Transactions on, Jan. 2007 Volume: 11 , Issue: 1 pp. 47 - 57

[52] [7] J.G. Snel, S.D. Olabarriaga, J. Alkmade, H. Gratama van Andel, A.J Nederveen, C.B. Majoie G.J. den Heeten, M. van Straten, R.G. Belleman. “A Distributed Workflow Management System for Automated Medical Image Analysis and Logistics” Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium pp. 733 - 738

[53] S. D. Olabarriaga, T. Glatard, and P. T. de Boer. "A Virtual Laboratory for Medical Image Analysis". Information Technology in Biomedicine, IEEE Transactions, July 2010 Volume 14 , Issue 4 pp. 979 – 985

[54] N. Shadbolt, P. Lewis, S. Dasmahapatra, D. Dupplaw,B.Hu, andH. Lewis, "MIAKT: Combining grid and web services for collaborative medical decision making" presented at the UK e-Science All Hands Meeting, Nottingham, U.K., 2004.

[55] C. Botha, "DeVIDE—The Delft visualization and image processing development environment" Technical University of Delft, Delft, The Netherlands, Tech. Rep., May 2005.

[56] D. Rex, J.Ma, andA. Toga, "The LONI pipeline processing environment" NeuroImage, vol. 19, no. 3, pp. 1033–1048, Jul. 2003.

[57] K. Alexopoulos, S. Makris, V. Xanthakis, G. Chryssolouris. A web-services oriented workflow management system for integrated digital production engineering. CIRP Journal of Manufacturing Science and Technology Volume 4, Issue 3, 2011, Pages 290–295

[58] de Bruin JS, Deelder AM, Palmblad M. Scientific Workflow Management in Proteomics. Mol Cell Proteomics. 2012 Jul;11(7):M111.010595. doi: 10.1074/mcp.M111.010595.

[59] Xudong Song, Wanchun Dou, Jinjun Chen. A workflow framework for intelligent service composition. Future Generation Computer Systems. Volume 27, Issue 5, May 2011, Pages 627–636.

[60] Jack C.P. Cheng, Kincho H. Law, Hans Bjornsson, Albert Jones, Ram Sriram. A service oriented framework for construction supply chain integration. Automation in Construction. Volume 19, Issue 2, March 2010, Pages 245–260.

[61] Orvis J[1], Crabtree J, Galens K, Gussman A, Inman JM, Lee E, Nampally S, Riley D, Sundaram JP, Felix V, Whitty B, Mahurkar A, Wortman J, White O, Angiuoli SV. Ergatis: a web interface and scalable software system for bioinformatics workflows. Bioinformatics. 2010 Jun 15;26(12):1488-92. doi: 10.1093/bioinformatics/btq167. Epub 2010 Apr 22.

[62] Carol M. Lushbough, Douglas M. Jennewein, and Volker P. Brendel. The BioExtract Server: a web-based bioinformatics workflow platform. Nucleic Acids Res. Jul 1, 2011; 39 W528–W532.

[63] Hector Fernandez, Cédric Tedeschi, and Thierry Priol. A Chemistry-Inspired Workflow Management System for Scientific Applications in Clouds. ESCIENCE '11 Proceedings of the 2011 IEEE Seventh International Conference on eScience. Pages 39-46

[64] Jungkee Kim. Web services for a chemical information clustering. Computer Sciences and Convergence Information Technology (ICCIT), 2011 6th International Conference. Nov. 29 2011-Dec. 1 2011 pp 140-143.

[65] Kalina Bontcheva, Hamish Cunningham, Ian Roberts, Angus Roberts, Valentin Tablan, Niraj Aswani, Genevieve Gorrell. GATE Teamware: a web-based, collaborative text annotation framework. Language Resources and Evaluation December 2013, Volume 47, Issue 4, pp 1007-1029.

[66] Zhao X, Liu E, Clapworthy GJ, Viceconti M, Testi D. SOA-based digital library services and composition in biomedical applications. Comput Methods Programs Biomed. 2012 Jun; 106(3):219-33. doi: 10.1016/j.cmpb.2010.08.009. Epub 2010 Sep 17.

[67] Rasband, W.S., ImageJ, U. S. National Institutes of Health, Bethesda, Maryland, USA, http://imagej.nih.gov/ij/, 1997-2014.

[68] Schneider CA, Rasband WS, Eliceiri KW (2012). "NIH Image to ImageJ: 25 years of image analysis". *Nat Methods* **9** (7): 671–675.

[69] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[70] Markus Hofmann, Ralf Klinkenberg, "RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)," *CRC Press*, October 25, 2013.

[71] Guido Deutsch, "RapidMiner from Rapid-I at CeBIT 2010," *Data Mining Blog*, March 18, 2010.

[72] Katherine Wolstencroft, Robert Haines, Donal Fellows, Alan Williams, David Withers, Stuart Owen, Stian Soiland-Reyes, Ian Dunlop, Aleksandra Nenadic, Paul Fisher, Jiten Bhagat, Khalid Belhajjame, Finn Bacall, Alex Hardisty, Abraham Nieva de la Hidalga, Maria P. Balcazar Vargas, Shoaib Sufi, and Carole Goble (2013): "The Taverna workflow suite: designing and executing workflows of

Web Services on the desktop, web or in the cloud”, *Nucleic Acids Research*, 41(W1): W557-W561.doi:10.1093/nar/gkt328

[73] "myGrid.org.uk official website". Retrieved 2012-06-20.

[74] Bhagat, J.; Tanoh, F.; Nzuobontane, E.; Laurent, T.; Orlowski, J.; Roos, M.; Wolstencroft, K.; Aleksejevs, S.; Stevens, R.; Pettifer, S.; Lopez, R.; Goble, C. A. (2010). "BioCatalogue: A universal catalogue of web services for the life sciences". *Nucleic Acids Research* **38** (Web Server issue): W689–W694. doi:10.1093/nar/gkq394. PMC 2896129. PMID 20484378

[75] Goble, C. A.; Bhagat, J.; Aleksejevs, S.; Cruickshank, D.; Michaelides, D.; Newman, D.; Borkum, M.; Bechhofer, S.; Roos, M.; Li, P.; De Roure, D. (2010). "MyExperiment: A repository and social network for the sharing of bioinformatics workflows". *Nucleic Acids Research* **38** (Web Server issue): W677–W682. doi:10.1093/nar/gkq429. PMC 2896080. PMID 20501605

CHAPTER 3

The Proposed Image Mining Framework

This chapter presents information, related to the tools and methods used, for developing the proposed framework and exposing its functionality. More specifically, ontological planning is introduced, followed by communication of the modules through web services. Proof of concept is demonstrated through the exploitation of freely distributed, popular workflow management tools, like the TAVERNA [1] and Rapid Miner [2] platforms, in order to deploy the framework. Both RapidMiner and Taverna workflow managers, constitute established solutions in the field of scientific computing, that fully comply with the Data Mining Workflow (DMWF) standard [44]. They have already been widely embraced by several research communities, as that of the data miners (the former) or the latter by the broader bioinformatics community, including biologists, engineers, medical doctors and generally scientists in the field of Biomedicine. The repositories myExperiment (http://www.myexperiment.org) and Biocatalogue (http://www.biocatalogue.org), host an impressive number of workflows and web services all accessible through Taverna, for a very wide range of disparate biomedical data processing tasks. In this way, Taverna constitutes a huge leap in the direction of bringing advanced hyper-computing capabilities to the hands of the biomedical community, enabling accessibility to this type of processing even to people with rudimentary programming skills. Adopting good practices and tools from the world of scientific computing, like the aforementioned, enables the exploitation of the tools and methodologies developed in this work and provided by the image mining framework, by a broader user community, efficiently, rapidly, for different purposes and allowing the implementation of hybrid processing patterns integrating image related tasks with other types of processing (textual, multimedia, decision-making, artificial intelligence).

These workflows provide great variety of disparate biomedical tasks and are capable of dealing with very specific biomedical issues. The proposed framework provides global support for several types of image analysis. In this way, the proposed workflow-based framework provides to the biomedical community advanced hyper-computing capabilities, which allow accessibility to the web-services (processes) of the framework, even to people with fundamental programming skills.

One of the most powerful provided features is the ability to generate additional scheme instances based on a core data flow. This feature allows users to perform multiple tests with a single run, by applying a number of different functionalities with different combinations, which are integrated on the core scheme automatically. All the above features of the proposed framework are combined, in order to provide multiple approaches to a specific image analysis problem. The most robust feature of the framework is the selection of the optimal workflow scheme. The workflow scheme proposal feature can be applied through a comparison with the ground truth dataset, or through a clustering procedure.

3.1 Ontological Planning

Data Mining (DM) workflows generally represent a set of DM operators, which are executed and applied on data or models. In most of the DM tools, users are only working with operators and setting their parameters (values). Data is implicit, hidden in the connectors, whereas the user provides the data and applies the operators, but after each step, new data is produced. In the presented approach, a distinction is made between all components of the DM workflow, operators, data and parameters.

In order to enhance the system's co-operability with the user on workflow design, an ontology-relying formalization of the DM workflows has been developed. The definition of a DM workflow, solicits the detailed formulation of the Data Mining Workflow (DMWF) ontology, since workflows are stored and represented in DMWF format.

The versatility, disparity and user-related subjectivity of the DM tasks, which imply non-monotonic optimality for the vast majority of the solution areas of the relevant problems addressed, yet the striking similarities of the underlying complexity patterns, preponderate for the introduction and utilization of ontological planning. Moreover, in this way, novel knowledge inference related to the particular data mining challenge may be attained, borrowed from similar in their technical description cases, which could be deployed and validated for its efficiency ad hoc. The DMWF ontology has several classes that contribute in the description in adequate detail, of the DM world, Input/Output Objects (IOO), MetaData, Operators, Goals, Tasks and Methods (see Table 3). The DMWF ontology enables the semantic annotation of image mining operators and thus the utilization of planning and reasoning tools, which in turn render feasible, workflow task automation and optimization. To our best knowledge there is no other ontological approach available for annotating image mining operators and tasks.

Table 2. Main classes of the DMWF ontology

Class	Description	Examples
IOObject	Input and output used by the operator	Data, Model, Report
MetaData	Characteristics of the IOObjects	Attribute, AttributeType, DataColumn
Operator	DM operators	ImageFeatureExtraction, FeatureClassification
Goal	A DM goal	PatternDiscovery,
Task	A task that is used to achieve a goal	ImageEnhancement, FeatureTraining
Method	A method is used to solve a task	AutoThreshold

Table 3. Main roles of the DMWF ontology

Properties	Domain	Range	Description
uses - usesData - usesModel	Operator	IOObject	Defines input for an operator
produces - producesData - roducesModel	Operator	IOObject	Defines output for an operator
parameter	Operator	MetaData	Defines other parameters for operators
simpleParameter	Operator	data type	
solvedby	Task	Method	A task is solved by a method
worksOn -inputData -outputData	TaskMethod	IOObject	The IOObject elements the Task or Method works on
worksWith	TaskMethod	MetaData	The MetaData elements the Task or Method worksWith
decomposedTo	Method	Operator/Task	A Method is decomposed into a set of steps

The entities stemming from the DWMF ontology are connected through roles or properties as shown in Table 6. The operator parameters as well as some basic data characteristics are values (integer, double, string, etc.) in terms of data properties, e.g. number of records for each data table, number of missing values for each column, mean value and standard deviation for each scalar column, number of different values for nominal columns, etc. Having them modeled in the ontology enables the planner to use them in planning. More information on the ontological planning and the planner tool can be found in [3] and [4]. The DMWF ontology has been developed in the context of the eLICO project and it can be found in http://www.e-lico.eu/ontologies/dmo/e-Lico-eProPlan-DMWF-HTN.owl.

In the current version of the framework, eight types of operators – based on their functionality – were developed (See Table 4). Each of the developed ontologies indicates the type of data that is required for its input. As operator is being considered each object, which corresponds to a developed process of the web service and contains input or/and output ports. The input and output ports support the flow of image data, or text data (settings, commands, etc.). Each operator which accepts images and more than one processing method as input is considered as core operator. Additionally, each operator that can be integrated into a core operator, in order to provide additional functionalities, is considered as secondary operator.

In this framework, two novel features were implemented, in order to enhance the user's collaboration with the proposed framework. The first one allows the creation of multiple workflow instances by making eligible to import additional operators in each of the core operators (Filter, Segmentation, Clustering and Feature extraction – see Table 4). More specifically, the above mentioned feature performs multiple independent parallel runs, with different combinations of image processing and segmentation algorithms, based on the core workflow scheme that was constructed by the user.

The second proposed feature utilizes the output of all the generated workflow instances and evaluates their performances, in order to propose the optimal workflow scheme for the specific type of dataset. The evaluation process can be achieved either by utilizing a ground truth dataset or by clustering the detected salient objects, based on a feature extraction process of the specific areas.

Table 4. The types of the operators of the proposed framework

Operator type	Functionality	Description
Feature Extraction	Single / Multi	Operators responsible for the feature extraction of the Image-s used as input
Filter	Single / Multi	Operators responsible for the application of Image analysis techniques and processing of the Image-s
Segmentation	Single / Multi	Operators responsible for the segmentation tasks for the coresponding Image-s
Clustering	Single / Multi	Operators responsible for the clustering of the salient objects' features. The output from this process can be used by the Evaluation type Operators
Convertion	Single	Assistive Operators responsible for the Convertion of the Input Image-s to the selected format (RGB, Grayscale etc)
Utility	Single	Assistive Operators that provide versatility services for the indermediate users (Add images, Separate Images etc)
Evaluation	Single / Multi	Operators that propose the optimal workflow scheme for a specific task (Ground Truth Cehecker, Scheme proposal, etc)
Input / Output	Single	Operators responsible for the Importation and Exportation of Images from the designed workflow

3.2 Ontology Features

The proposed framework provides functionalities that support image acquisition, sampling and storing of the processed images, complete program control and easy integration into image-enabled applications that utilize databases. Additionally, it includes functions that support image transformation, color processing, feature extraction and image enhancement. More specifically, the following functionalities are included into the services of the proposed framework.

Mass image acquisition and storage: these functions refer to the ability of the proposed system to import multiple images to the workflow and to the acquisition of the processed images and their automated storage in the local drive.

Filter functions: these functions carry out the image manipulation and enhancement tasks. The toolbox provides basic functionalities for image enhancement (such as background subtraction, background correction, de-noising, smoothing, spatial and median filtering, histogram equalization etc.), color processing and image texture analysis.

Class separation: this feature provides the automated separation of the different classes of a segmented image by acquiring the informative parts of each class from the initial image and generating new images, each of them containing only the informative parts of the corresponding class.

Multiple methods integration and independent instance generation: this feature enables the integration of more than one method in a single operator and generates a number of generated instances, based on the number of implemented methods, which are completely independent. This feature provides rapid evaluation of the different settings of a specific customized workflow.

Smart Image Conversion: this feature refers to the ability of the operators to correct and readjust the provided input to the correct format. For example, the watershed filtering task, will ensure that the input of the flow will be a binary image by checking it, and in case of incompatibility, it will auto-converting it to binary, using the automated threshold technique [5].

Default value auto fill: Each operator contains one or more input ports and one output port. Excluding the image dataset ports (the ports that require image-s as input), in case that an input port is not set, the NULL operator service will fill the default values for this port, so as to avoid possible system exceptions or crushes during the running of a constructed workflow. This feature ensures the stability of the proposed system and provides better collaboration with the user.

Feature extraction, Classification and Rapid characterization: It refers to tools and functions, related to the high level understanding and content retrieval from a digital image.

Clustering: It refers to operators that utilize the clustering techniques, in order to cluster the salient objects, which are detected by the combination of image analysis operators. Prerequisite procedure for this functionality is the feature extraction of the salient objects.

Evaluation – Workflow Scheme Selection: These operators calculate the necessary metrics and perform comparisons between the different workflows' images, in order to export the workflow instance with the best performance.

3.2.1 Multiple Parallel Workflow Instances Generation

The parallel instances generation sounds a complex task, but it is quite user friendly and can be easily constructed by a user with fundamental programming skills. This easiness lies to the fact that the additional instant generation feature is enabled automatically through a drag 'n drop procedure of a secondary operator onto a core operator. More specifically, each operator – which corresponds to segmentation, filtering, feature extraction or clustering process (see Figure 16) – contains additional input ports, allowing the integration of more than one method for the specific action. The integration of one methodology into a main operator does not affect the integration or the implementation of another integrated methodology into the same core operator. This feature achieves the storage of the earlier connected operators' results and connects them with the next operators. Additionally, it enables the production of a number of independent instances – based on the number of the methods applied on the specific operator. This feature allows the user to perform multiple parallel independent test runs and maintain the initial workflow structure. The implementation of this feature makes the processing of large datasets and the optimal evaluation of the specific workflow eligible.

The main functionalities of the proposed framework are illustrated in Figure 16 and a brief description for each of them is included in Table 4. The number of generated workflow instances equals the number of possible combinations, based on the type and number of integrated operators. For example, if there is a developed workflow scheme containing a tier of filtering, with three methods embedded, and a tier of segmentation, with two segmentation methods embedded, there will be six generated workflow instances running simultaneously. In this case, the Scheme Selection process will export the optimal flow and the corresponding processed data.

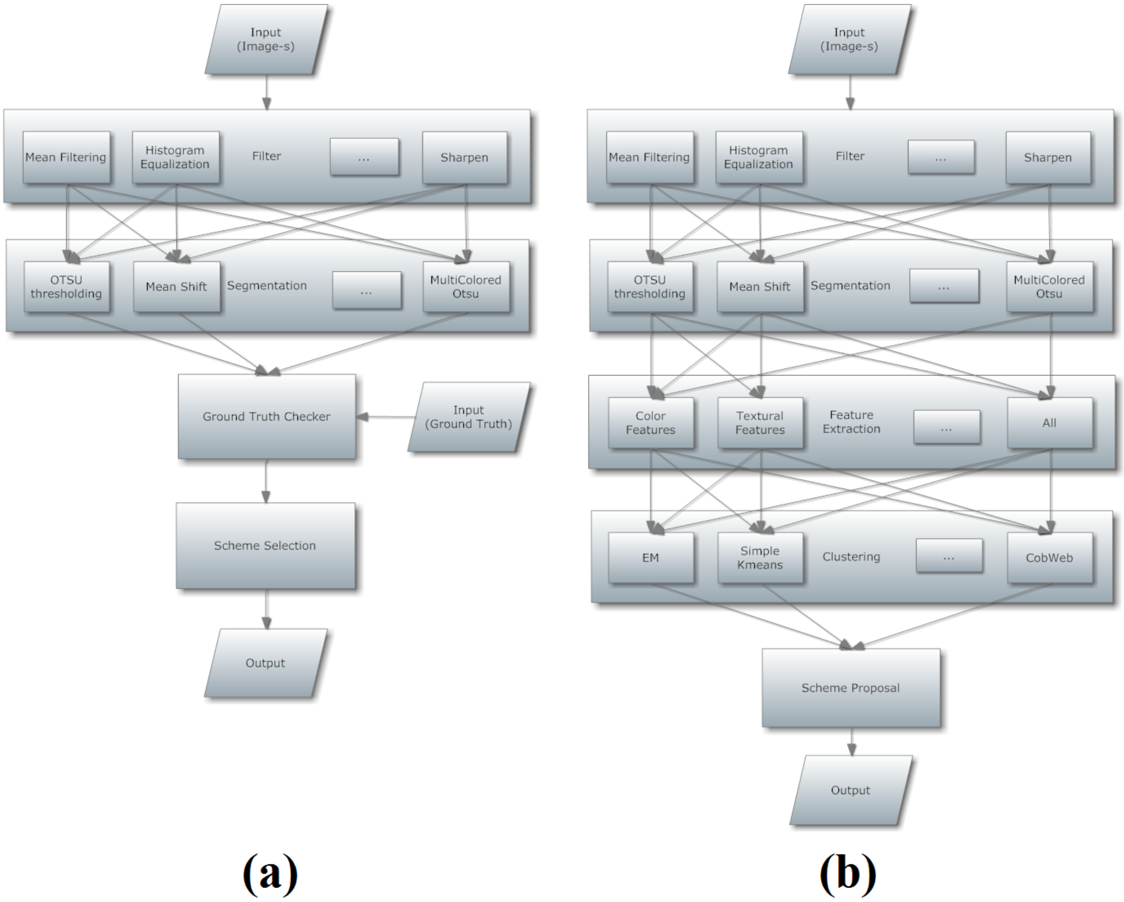


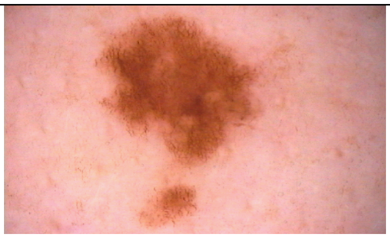
Figure 16. Workflow schemes demonstrating the main features of the proposed framework (multiple independent instance generation, mass local storage, feature extraction, segmentation, editing, filtering, clustering, scheme proposal etc.). In (a) a workflow demonstrating the optimal workflow proposal through the utilization of a Ground Truth dataset. In (b) a workflow demonstrating the optimal workflow proposal through the clustering of the detected objects.

The TAVERNA application provides a graphical user interface for both creating and running workflows. In TAVERNA, a workflow is considered to be a graph of processors, each of which transforms a set of data inputs into a set of data outputs.

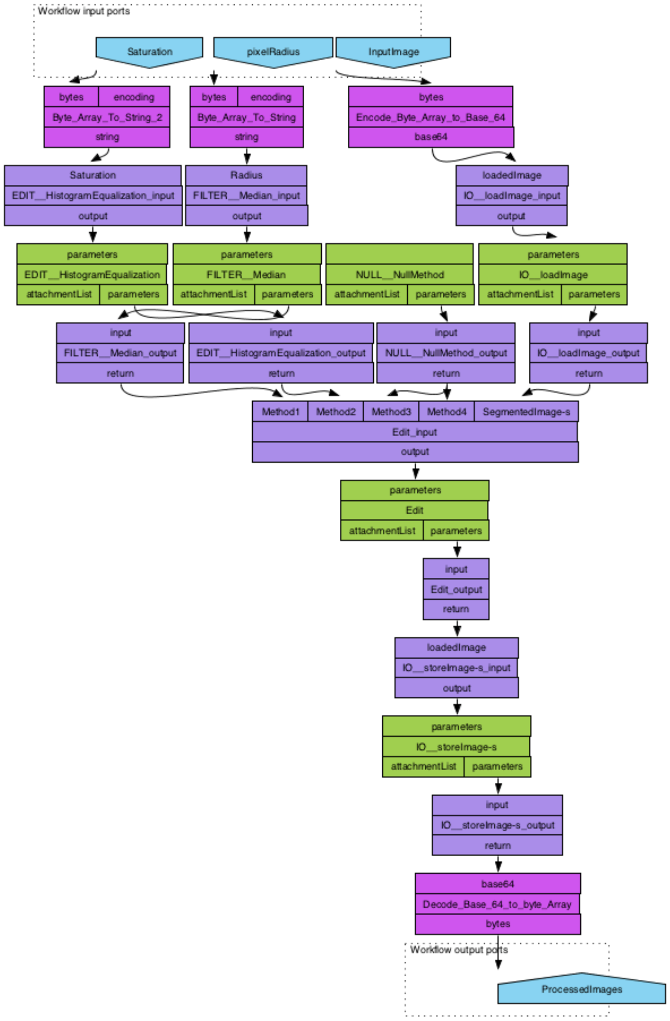
The proposed framework includes baseline image processing tools like subtraction, background correction, de-noising, smoothing, peak harvesting and peak alignment, segmentation, analysis, registration, feature extraction and recognition.

The main goal of this work is to create a collection of image processing tools, provided through web services, which may be used in various workflows administered by workflow management tools, such as the TAVERNA workflow manager. These image-processing pipelines can be applied on several types of images, like the microscopy images of melanoma and breast cancer in the following sections. The implementation of the image mining toolbox, as a Web Service, allows the easy integration of its whole functionality into TAVERNA.

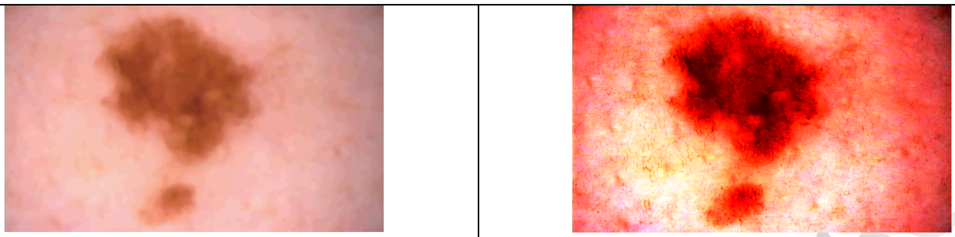
A typical example exploiting some of the core features of the proposed framework is depicted in Figure 17. In this example, the histogram equalization function and the median filtering function are applied on the initial image, providing two independent instances of the same workflow scheme. The framework is user friendly and provides easy creation of complex image analysis workflows. The specific workflow is simple (Figure 17 b), but demonstrates a robust feature that allows experts to perform multiple tests with different settings and methods without changing the initial workflow diagram. In this way, rapid evaluation is achieved and the user chooses the instance with the optimal settings and methodologies.



a)


Workflow input ports
Saturation
pixelRadius
InputImage
bytes
encoding
Byte_Array_To_String_2
string
Byte_Array_To_String
Encode_Byte_Array_to_Base_64
base64
Saturation
EDIT__HistogramEqualization_input
output
Radius
FILTER__Median_input
loadedImage
IO__loadImage_input
parameters
EDIT__HistogramEqualization
FILTER__Median
NULL__NullMethod
IO__loadImage
attachmentList
input
FILTER__Median_output
EDIT__HistogramEqualization_output
NULL__NullMethod_output
IO__loadImage_output
return
Method1
Method2
Method3
Method4
SegmentedImage-s
Edit_input
Edit
Edit_output
IO__storeImage-s_input
IO__storeImage-s
IO__storeImage-s_output
Decode_Base_64_to_byte_Array
Workflow output ports
ProcessedImages

b)



c) d)

Figure 17. a) The initial image that will be the input of the workflow
b) A workflow demonstrating the generation of the two independent instances of the same scheme
c) The median filtering instance output image
d) The histogram equalization output image

3.2.2 Feature Extraction

The feature extraction operators serve multiple purposes. They can provide additional characterization to the segmented objects, or they can provide the mandatory input to the clustering operators, in order to perform clustering of the detected objects and afterwards to propose the optimal workflow scheme. The functionality of the core feature extraction operator is quite simple. It receives the annotation areas from the segmented binary images and exports the selected features from the corresponding areas of the initial images. The user can select the extracted features. The user can integrate textural and/or color feature extraction operators. The supported color features are the mean red, mean green and mean blue values of each detected object. The supported textural features are the area, the mean (gray value), the standard deviation, the modal gray value, the minimum and maximum gray values, the perimeter, the ellipse, the major and minor axis, the angle, the circularity, the ellipse aspect ratio, the integrated density, the median, the skewness, the kurtosis, the area fraction, the roundness and the solidity values.

3.2.3 Workflow Evaluation

The optimal workflow proposal functionality can be achieved through the integration of Scheme proposal operator at the end of the composed scheme. This operator takes the input from the clustering or the ground truth checker operator. The output it provides, is the sequence of the integrated operators that are necessary for the optimal solution of the issue. The methodologies of both clustering and ground truth checking functionalities are described on the following subsections.

Evaluation through Ground Truth dataset

The performance of a workflow scheme is defined by the evaluation of the applied processing operators along with the segmentation algorithm on how accurately detected the salient object (see Figure 16 a). In order to provide an objective evaluation of the different generated workflow instances, a metric that is not affected by the amount of difference between the TP and TN pixels is required. Thus, Matthews Correlation Coefficient [6] was utilized, because it is not affected by the possible large amount of true negative pixels.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (1)$$

The ground truth checker algorithm requires the original image containing red colored areas that correspond to the salient objects. Afterwards, it makes a comparison of the segmented image with the provided ground truth image and calculates the above mentioned metric. The image that succeeds the greater MCC value is chosen as optimal and servers as input to the Scheme Proposal operator that recognizes which operators processed the specific image. Furthermore, it recognizes the application sequence of the utilized operators and provides the complete optimal workflow for the detection and recognition of the salient objects of the specific image dataset.

Evaluation through Clustering

There are plenty of cases that the salient objects of the image dataset are too many to count or annotate. This fact makes the creation of the ground truth dataset impossible. In these cases, clustering of the detected salient object is achieved through a selected feature extraction process. The user may select one or more clustering algorithms (see Figure 30 b), in order to perform the clustering of the salient objects of each image separately.

The produced clusters are evaluated via the log likelihood distance of the clusters metric [7], which is probability-based distance. This metric is able to deal with not only continuous, but also with categorical variables, as well. The longer the distance between two classes, the greater the log likelihood distance metric becomes. In order to calculate the log likelihood metric, it is assumed that the categorical or the continuous variables it deals with, are independent to each other. Scheme proposal operator will choose the optimal version of each image, based on the log-likelihood metric, and will detect the operators processed the specific image. More specifically, the evaluation operator will choose as optimal workflow instance the one that will achieve the greater log-likelihood distance between the clusters formed by the detected objects. By detecting the utilized operators, the framework exports the optimal operator sequence for the specific dataset. In cases that more than one classifier were utilized for the clustering task, the average value of the log-likelihood for each clusterer is calculated. Thus, the workflow that scores the greater average log-likelihood, is selected as the optimal one.

3.3 Web service communication protocol and Security Considerations

Web services became popular as a promising technology, providing the feature of building distributed system applications. It is an implementation of Service Oriented

Architecture (SOA) [8] that supports the concept of loosely-coupled, open-standard language and platform-independent systems. There are two ways to exploit the web services technology. One way is through Simple Object Access Protocol (SOAP). SOAP is based on Web Service Description Language, which sets the rules for accessing each web service and defines what, each of the services offers. The other way to access web services features is REpresentational State Transfer (REST).

Some people believe that one of them is better and overlaps the other. This is false statement, since every protocol offers different features with the corresponding pros and cons. Both protocols contain a set of rules for the exploitation of the web services. Particularly, both of the protocols determine different ways of communication with the server, requesting the major information for the web services establishment.

Microsoft developed SOAP in 1998. It is a protocol that supports all the features of web services standards and it is a long-term supported protocol. On the other hand, REST, was developed a couple of years later, under the purpose of filling the gaps and solve the issues of SOAP, providing a simpler utilization of web services technology. However, REST is sometimes more complex than the SOAP implementation and although it was created to solve the issues of SOAP, created issues of its own.

A brief analysis of both protocols follows, in order to evaluate both of them and select the most appropriate for our research.

3.3.1 Simple Object Access Protocol (SOAP)

The initial motivation for SOAP's development was to replace the older technologies that couldn't collaborate efficiently with the Internet (Distributed Component Object Model – DCOM, Common Object Request Broker Architecture – CORBA, etc.), which were based on binary type messages. SOAP supported the HTML format, which collaborates efficiently with the Internet.

SOAP's client-server communication is mostly based on HTML. The HTML basically determines the format of the exchanged messages. SOAP was submitted to the

Internet Engineering Task Force (IETF) and became a worldwide standard. Its architecture supports the extensibility of its current methods. Thus, a huge variety of extensions, containing the acronym WS-, are available for the corresponding use (i.e. WS-Security, WS-Policy, etc.). The current range of SOAP's extensibility is huge. The key point is that every application utilizes only the part it needs from it, depending on its needs. For example, each application utilizes a different combination of the WS packs. The strong part of SOAP, the XML, may also become its weakness. There are cases that the XML part of a web service may become very complex. The greater the XML complexity, the greater the code error possibility becomes.

Another strong part of this protocol is the Web Service Description Language (WSDL). It is commonly known as WSDL document. This document contains the entire functionality of the web service, such as the type and number of the argument inputs of the service, the type of the output, the names of the arguments, etc. WSDL offers language and system independency to the collaborating services. Practically, is like a translator between the web service applications.

Additionally, SOAP provides all the necessary information for an active web service through the Universal Discovery Description Integration (UDDI) feature. In more details, UDDI provides the analytical instructions that are required for the collaboration of the active web service and brief description of the service's capabilities. Thus, the potential collaborator knows if the active service serves its needs and it is able to achieve the initial communication with it. The first contact includes the download of the corresponding WSDL document.

One of the most robust features of the SOAP protocol is the build in error handling. This feature ensures the debugging easiness of the applications that would like to contact with the developed service. For example, in case of the development of a collaborative application, which has no ownership rights on the code of the active web service, the build-in error handler provides a full report with the exact error during the debugging process. Without this feature, it becomes easily understood that the debugging procedure of different ownership collaborative services would be impossible.

3.3.2 REpresentational State Transfer (REST)

After the release of SOAP, some developers criticized it as hard to use and not so versatile due to its XML based nature. Specifically, for the collaboration of SOAP with a JavaScript application, a lot of XML code is required, even for the simplest task. Rest is a lighter alternation of SOAP. Instead of relying onto XML based messages for the client-server communication, REST, in most cases, depends only on a URL address. In just a few cases the provision of additional information is required. The most REST-based web services require only the URL.

The protocol utilizes 4 HTTP v1.1 verb style commands (GET, DELETE, PUT, and POST) for the task accomplishment of a generated REST service. The middleman tool for the server's response provision is not an XML element. In REST services the middleman role can be played either by JavaScript Object Notation (JSON), or by Command Separated Value (CSV), or by Really Simple Syndication (RSS) file format. These formats are responsible for the response of the developed web service application. Thus, ensures the versatility of the protocol.

Additionally, REST contains a similar feature with SOAP. The feature of Web Application Description Language (WADL) is the corresponding WSDL of REST, with a small difference. In REST protocol architecture, WADL is not a mandatory element. The WADL document is similarly formatted as the WSDL document is. WADL was developed as an answer to the WSDL (of SOAP protocol), in order to provide platform and language independent compatibility, but it still is in early stage and it is utilized only through Java code.

3.3.3 Protocol Selection

The optimal protocol selection depends to which protocol is covering our needs. In this research, we want to develop an image-mining framework, which will be able to data mine different kind of biomedical images.

The biomedical data and the possible extracted prediction by the proposed framework should be secured. This makes necessary the WS-Security extension of SOAP architecture. REST does not support any advanced security suite. Additionally, SOAP contains a build-in error handler, which in our case, is also necessary, because we would like to provide to the future users an analytical feedback for the debugging process of the workflow construction process. Thus, in case of connecting incompatible operators, a proper message will be shown, indicating a proposed solution. This feature enhances the easiness of the utilization of our system, even to users with fundamental programming skills.

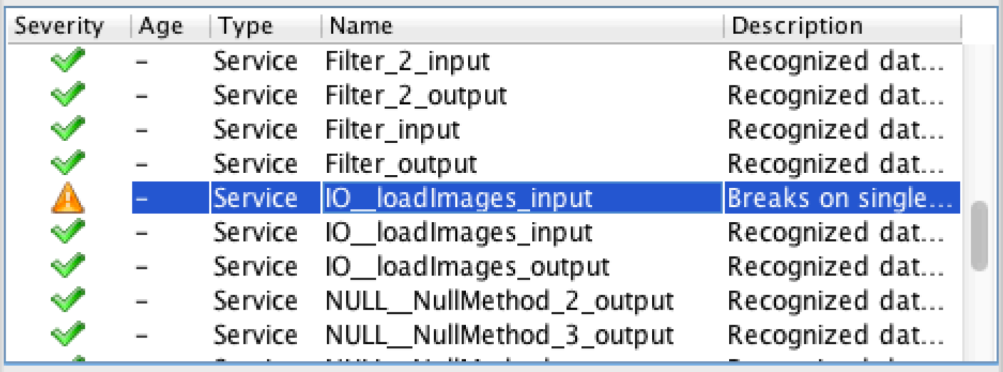
Considering that we would like to offer an open framework for image mining, accessible to anyone, we would like to ensure its versatility. More specifically, the framework must be platform and language independent. This feature is also supported by the SOAP protocol, instead of the REST one, which is depended by the usage of the HTTP.

By taking under consideration all the above needs, SOAP protocol fits the requirements of our proposed system. Additionally, the proposed system will be able to run more efficiently in distributed systems via the utilization of SOAP approach.

The development of the proposed framework was based on SOAP, because it enables the sharing of one operation (the functionality of an operator), to multiple process instances. Additionally, SOAP provides extended privacy [8] and advanced security options. The proposed framework provides functionality and access to resources for the mining of various types of biomedical data. Thus, proper access-control to the resources and secure data transmission is required. The WS-Security kit (Rampart) [9] has been utilized, in order to address both issues. WS-Security is a standard methodology for adding security to SOAP Web service message exchanges. It uses a SOAP message-header element to attach the security information to messages, in the form of tokens, conveying different types of claims (which can include names, identities, keys, groups, privileges, capabilities, and so on), along with encryption and digital-signature information. On top of the WS-Security kit, the SSL [10] protocol has been used for the proper encryption of the data during transmission, between the service consumer and the web service itself.

An Image Mining Web Service has been developed, in order to develop the entire framework, which provides all the above features. The service provides all the functionalities of the image processing and mining framework, described in previous sub-sections along with additional features that enable the integration of the services into third party tools. The stability of the constructed schemes is ensured via each operator's provided feedback with the TAVERNA workflow management. More specifically, when an operator receives an invalid input, it generates a warning message (see Figure 18). Thus, the debugging task of a constructed workflow scheme becomes an easy and user-friendly procedure.

The service has been implemented in Java, using the Apache MAVEN3 toolkit [40] and it is currently hosted on a Virtual Machine powered by GRNET-Okeanos.



Severity	Age	Type	Name	Description
✓	–	Service	Filter_2_input	Recognized dat...
✓	–	Service	Filter_2_output	Recognized dat...
✓	–	Service	Filter_input	Recognized dat...
✓	–	Service	Filter_output	Recognized dat...
⚠	–	Service	IO__loadImages_input	Breaks on single...
✓	–	Service	IO__loadImages_input	Recognized dat...
✓	–	Service	IO__loadImages_output	Recognized dat...
✓	–	Service	NULL__NullMethod_2_output	Recognized dat...
✓	–	Service	NULL__NullMethod_3_output	Recognized dat...

Figure 18. The debugging mode of Taverna in collaboration with the developed framework.

3.4 Ontology Development and Functionality

In this subsection of the chapter, we will discuss the usability and some development details of the developed operators. It can be noticed, that these operators can be utilized for different biomedical image mining issues. In Table 5 the developed operators are presented along with their respective type, role and a brief explanation of their functionality. Further details about each of them, are given to the following subsections.

Table 5. The Developed Operators of the Proposed Workflow

Name	Type	Role	Function
loadImage-s	Input/Output Operator	Independent Operator	It uploads images to the system
storeImage-s	Input/Output Operator	Independent Operator	it downloads images from the system
Filter	Filtering Operator	Core-primary	the core Operator for the image filtering procedure
Mean	Filtering Operator	Secondary	Applies mean filtering to the image
Median	Filtering Operator	Secondary	Applies median filtering to the image
Min	Filtering Operator	Secondary	Applies min filtering to the image
Max	Filtering Operator	Secondary	Applies max filtering to the image
Gaussian	Filtering Operator	Secondary	Applies Gaussian filtering to the image
Sharpen	Filtering Operator	Secondary	Sharpens the foreground elements of the image
ContrastEnhancement	Filtering Operator	Secondary	Enhances the contrast of the image
HistogramEqualization	Filtering Operator	Secondary	Performs histogram equalization
findEdges	Filtering Operator	Secondary	Detect the edges of the foreground objects
Smooth	Filtering Operator	Secondary	Smooths the image through the reduce of the contrast
Watershed	Filtering Operator	Secondary	Performs watershed filtering on a binary image
BackgroundSubtraction	Filtering Operator	Secondary	Removes the background elements of the image
MajorityVoting	Filtering Operator	Secondary	Applies majority voting technique to an image
MinimumObjectSize	Filtering Operator	Secondary	Removes the masks of the foreground objects that are smaller than the minimum size
Segmentation	Segmentation Operator	Core-primary	the core Operator for the image segmentation procedure
Otsu	Segmentation Operator	Secondary	Performs Otsu Segmentation
Mean Shift	Segmentation Operator	Secondary	Performs Mean Shift Segmentation
Li	Segmentation Operator	Secondary	Performs Li Segmentation
Huang	Segmentation Operator	Secondary	Performs Huang Segmentation
Triangle	Segmentation Operator	Secondary	Performs Triangle Segmentation
Yen	Segmentation Operator	Secondary	Performs Yen Segmentation
FeatureExtraction	Feature Extraction Operator	Core-primary	the core Operator for the feature extraction procedure
ColorFeatures	Feature Extraction Operator	Secondary	Exports the mean brightness values of each channel for the segmented objects
TexturalFeatures	Feature Extraction Operator	Secondary	Exports the textural features of the segmented objects
AllFeatures	Feature Extraction Operator	Secondary	Exports the color and textural features of the segmented objects
SeparateClasses	Utility Operator	Independent Operator	Generates additional Images, each of the containing only one class from the imported image
Annotation	Utility Operator	Independent Operator	Annotates the segmented objects on the original image
AddImages	Utility Operator	Independent Operator	Combines the streams from two or more workflows into one stream
Proposed_for_KUP	Segmentation Operator	Secondary	Perfoms the characterization of Kidney biopsy Images
Proposed_for_CC	Segmentation Operator	Secondary	Perfoms the characterization of Breast Cancer biopsy Images
Convert_toGray	Converter Operator	Independent Operator	Converts an image to 8-bit GrayScale
Convert_toIndexed	Converter Operator	Independent Operator	Converts an image to 8-bit indexed (based on the number of available colors) image
Convert_toRGB	Converter Operator	Independent Operator	Converts an image to 24-bit RGB
Ground_Truth_Checker	Evaluation Operator	Independent Operator	It evaluates a segmented image based to the Matthews Correlation Coefficient
Scheme_Proposal	Evaluation Operator	Independent Operator	It exports a document, with the optimal workflow
Clustering_Evaluation	Evaluation Operator	Independent Operator	It evaluates a segmented image based to the LogLikelihood metric of the segmented objects' Clustering

3.4.1 Input / Output File Operators

Since the operators and the entire functionality of the proposed framework is based on web services, proper operator for the importation and exportation of the image files is required. In low-level programming language, the upload of image files onto a web service is a complex procedure that is accomplished through a Message Transmission Optimization Mechanism (MTOM) [11]. MTOM enables the web transfer of binary type data to and from an active web service. The proposed framework, in high-level programming language, enables the upload and download of the imported and processed image to and from the service, with a simple drag 'n drop operator.

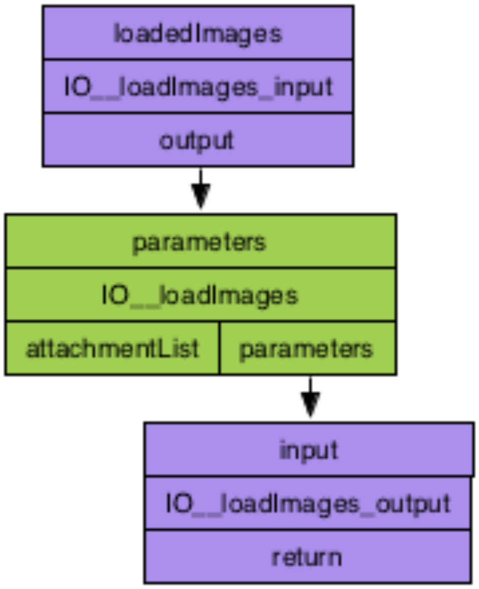


Figure 19. Load Image-s Operator in Taverna Framework

The task of uploading the image is accomplished through the load images operator (see Figure 19), which receives the encoded to base64 [12] type data and converts them into usable data available for the image processing techniques. On the other hand, the task of downloading the workflow-processed data to a local storage drive is accomplished via the store images operator (see Figure 20), which receives the data in their workflow-based form and exports them into binary data. These data are stored to the local drive of the personal computer that executes the proposed framework.

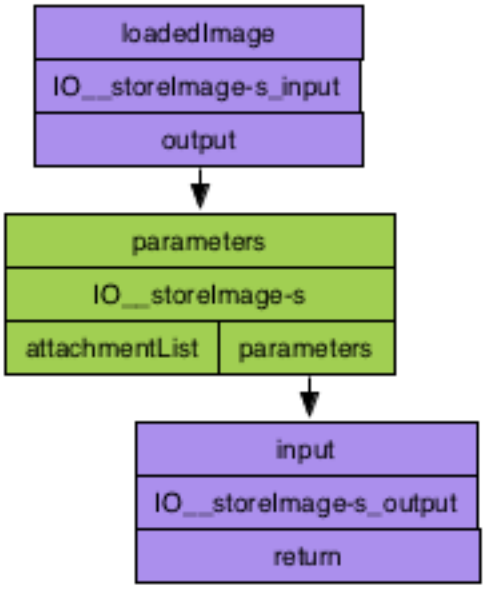


Figure 20. Store Image-s Operator

3.4.2 Filtering Operators

In this section, we will discuss the operators that correspond to the filtering process. We need to determine that the term "Filtering Operators" involves all the advanced processes of the proposed framework, which are related with the pre-segmentation and meta-segmentation methods that enhance the efficiency of the image-mining task.

Core Filtering Operator

In order to apply a filtering technique on an image, the integration of the corresponding operator onto the core operator (see Figure 21) is required. Specifically, the core operator receives as input one (at least) or up to 4 filtering methods. The number of the integrated methods indicates the number of the different-independent generated instances of one image. The operator exports the above-mentioned number of processed images, separated into different flows, in order to distinct later the most effective one.

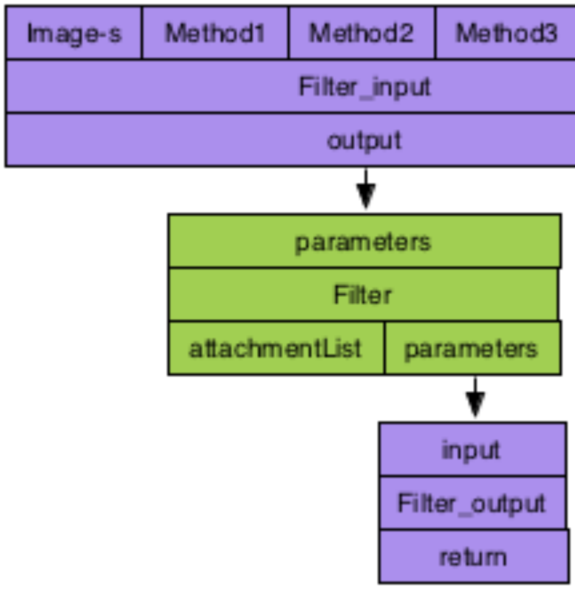


Figure 21. Filtering Operator, along with its available settings

Mean Filtering

Mean Filter Operator (see Figure 22) is one of the most common linear filtering techniques. It is widely used for the smoothing and moderate noise reduction. Mean filtering technique [13] changes each pixel's brightness with the mean brightness of its neighbor pixels. Thus, noise removal is achieved and in most cases, a blurring effect is denoted.

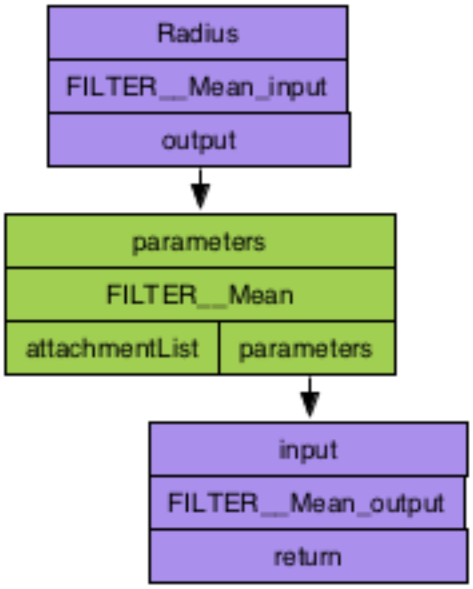
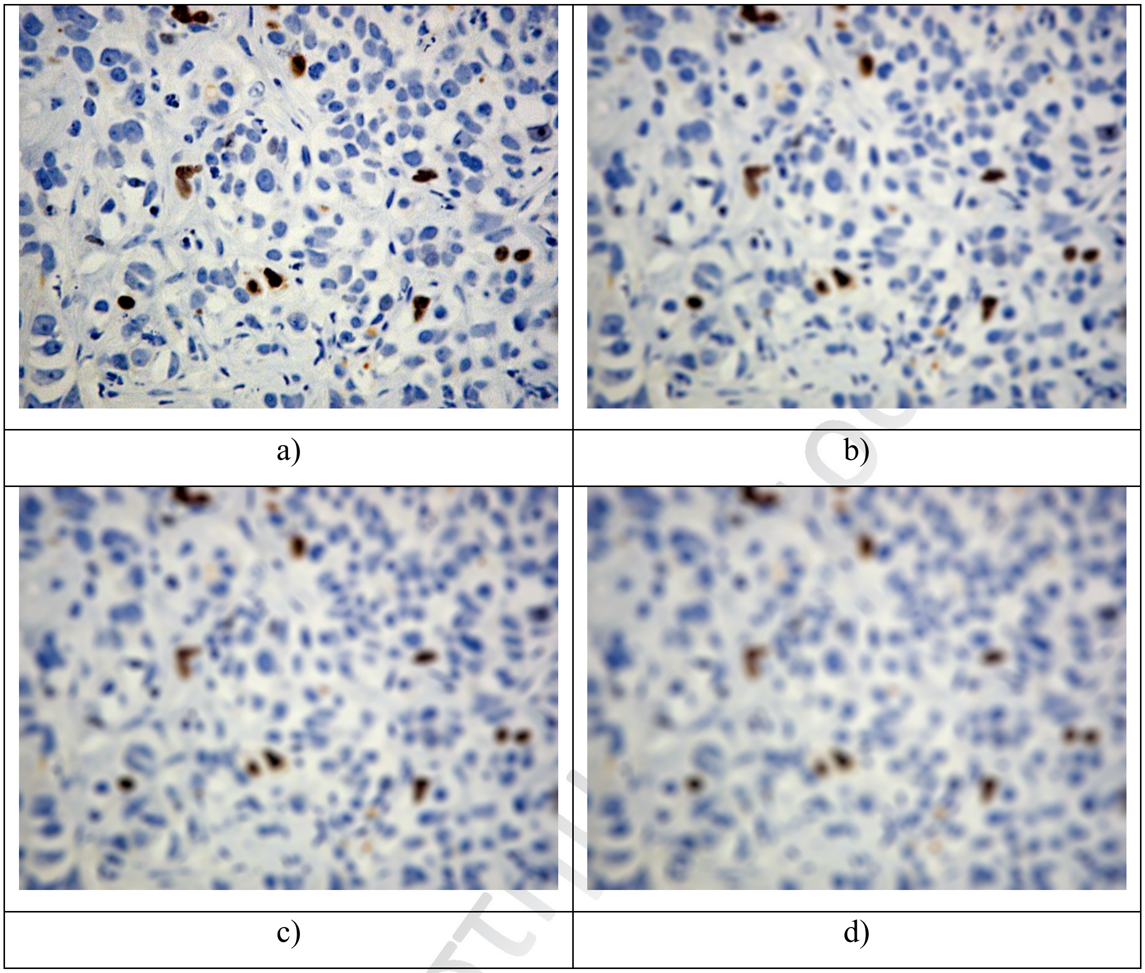


Figure 22. Mean Filter Operator

The mean filtering operator receives as input the Radius of the pixel area around the pixel that will be considered as its neighborhood. This is the area that will determine the value of the specific pixel. The larger the neighborhood area, the greater the achieved normalization and information loss is (see Figure 23). This filtering technique is considered as a low-pass filtering, which practically deletes the outlier values of the image. Therefore, the details of the image are reduced.



a)

b)

c)

d)

Figure 23. Output result after the application of Mean Filtering operator with input argument b) 3 (3x3), c) 5 (5x5) and d) 7 (7x7) on image a).

Median Filtering

Median Filtering [14] operator applies a linear technique for the noise reduction in image files, based on a simple idea. Specifically, the brightness value of a pixel equals to the median value of an ascending dataset, generated by its neighbor pixel area. In case the dataset holds an odd number of ascending ordered elements, the algorithm selects the median value. In case the number of the elements is even, the algorithm calculates the average value of the two median elements and changes the brightness of the pixel to this value.

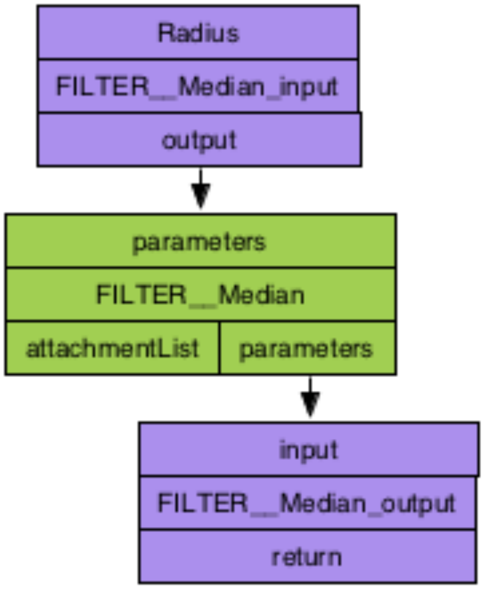


Figure 24. Median Filter Operator

Median Filtering operator (see Figure 24) is utilized, in order to reduce the noise and provide a moderate smoothing of the edges of the salient objects. In Figure 25 the output results of the application of the filter on a breast cancer biopsy are presented. Similarly to the Mean Filter operator, the result is depends on the size of the pixel Radius argument – the area of the neighbor pixels (see Figure 25).

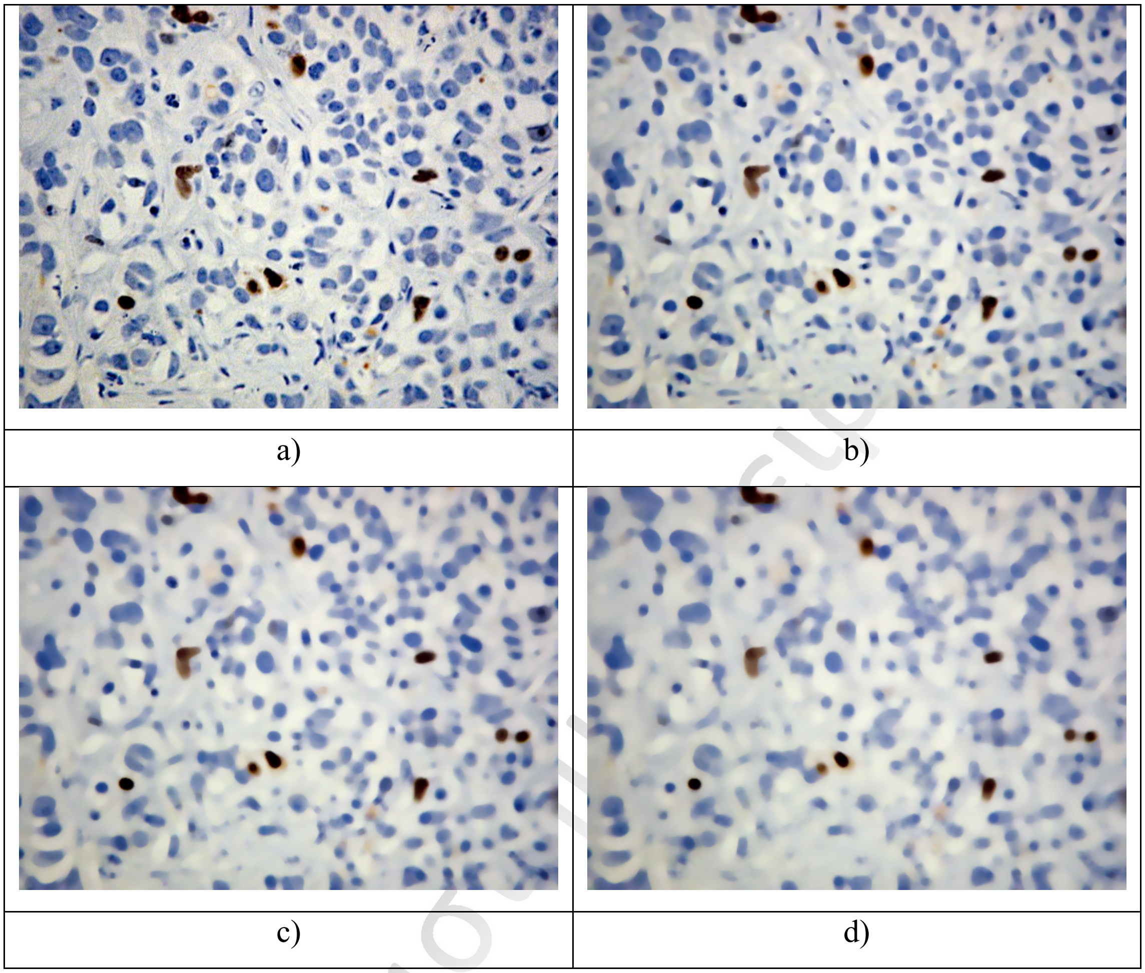

a)
b)
c)
d)

Figure 25. Output result after the application of Median Filtering operator with input argument b) 3 (3x3), c) 5 (5x5) and d) 7 (7x7) on image a).

Min and Max Filtering

Min and Max filtering [15] operators correspond to non-linear filtering techniques that can be implemented to the image filtering procedure via the utilization of masks. These filtering techniques are considered as sharpening filters, since they can emphasize the unclear salient objects of an image. The functionality of these filters is based on the brightness comparison of each pixel with its maximum or minimum brightness of its neighborhood.

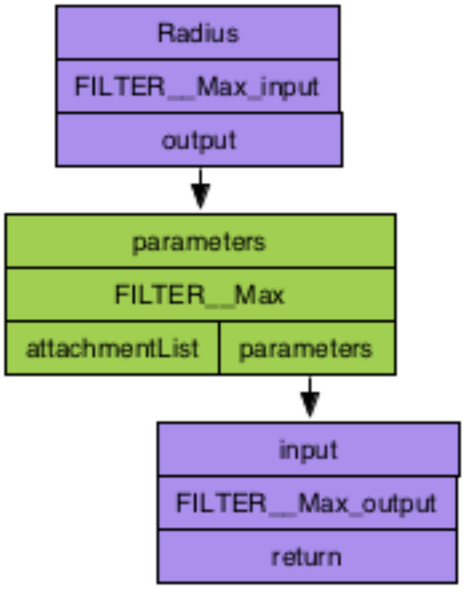


Figure 26. Max Filter Operator

More specifically, the Min Filtering operator stretches the dark areas and shrinks the lighted ones. Thus, it can be utilized for the white noise removal.

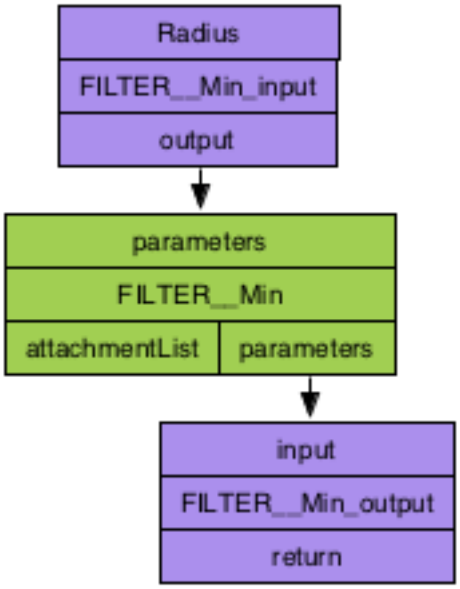
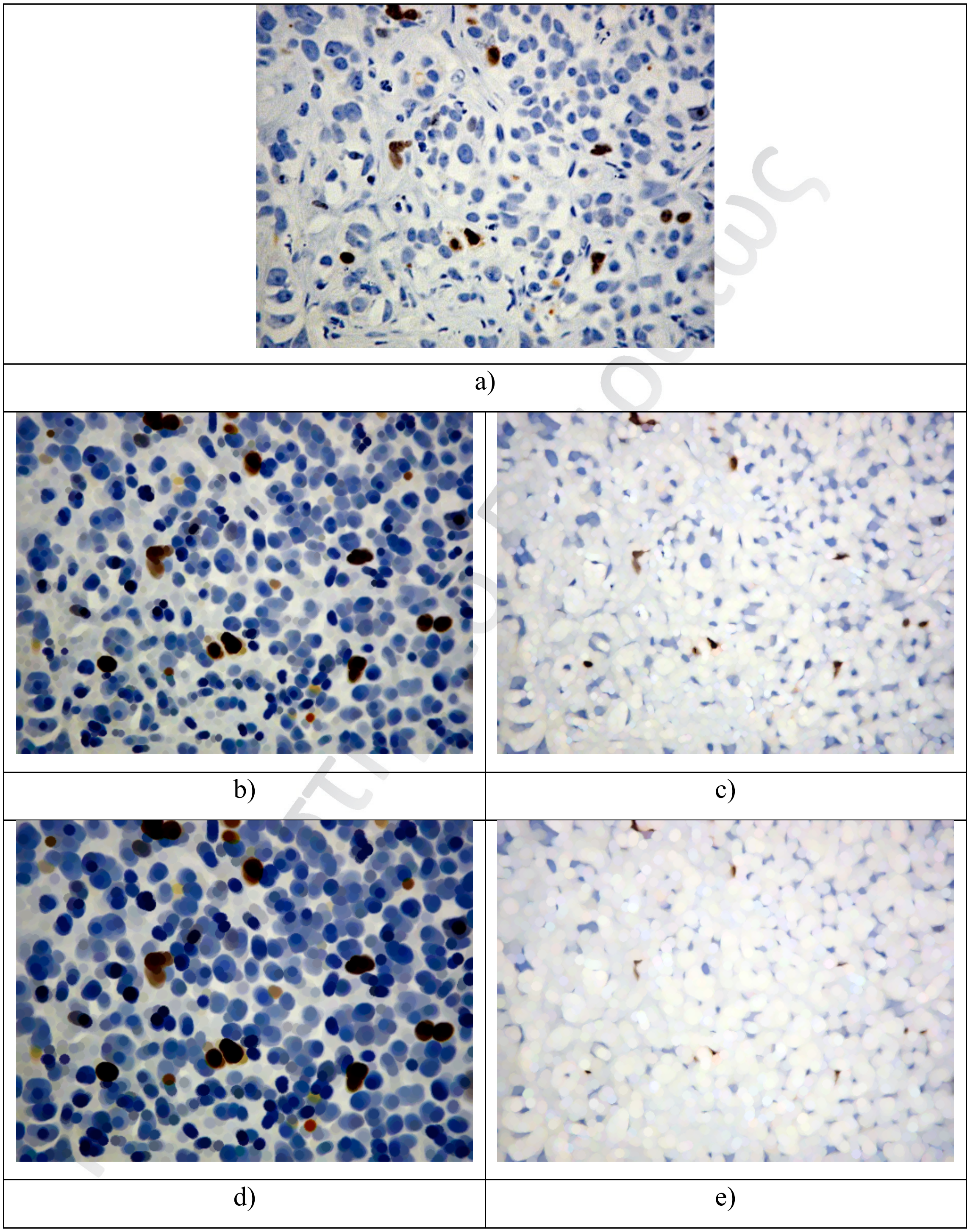


Figure 27. Min Filter Operator

The Max Filtering operator enhances the lighted-white areas, while it shrinks the darker ones. Max filtering technique is utilized for the dark and background noise removal.

Min and Max Filtering operators can be utilized, in order to enhance or make more distinctly the salient objects of a biomedical image. Similarly to the mean and median filters, the results are affected by the size of the pixel radius argument – window (see Figure 28).



a)

b)

c)

d)

e)

Figure 28. a) original image b) Min Filter with a 3 pixel radius setting c) Max Filter with a 3 pixel radius setting d) Min Filter with a 5 pixel radius setting e) Max Filter with a 5 pixel radius setting

Gaussian Filtering

The Gaussian Blur operator (see Figure 29) represents the Gaussian filtering methodology [16]. It belongs to the low-pass filtering techniques, therefore, except for the noise removal; it adds a blurring effect in the image.

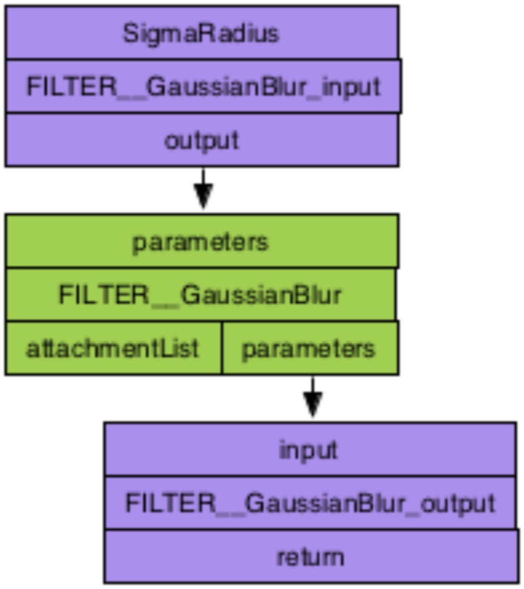


Figure 29. Gaussian Blur Operator

This technique also receives as input the Radius of the neighbor pixels (see Figure 44). The larger the radius, the greater the blurring effect (see Figure 30 b) and c)). Practically, the application of Gaussian Blur operator is the convolution of an image with a Gaussian function [16]. The Gaussian blurring reduces the high frequency components of an image. Thus, it can be utilized to enhance the foreground objects of an image.

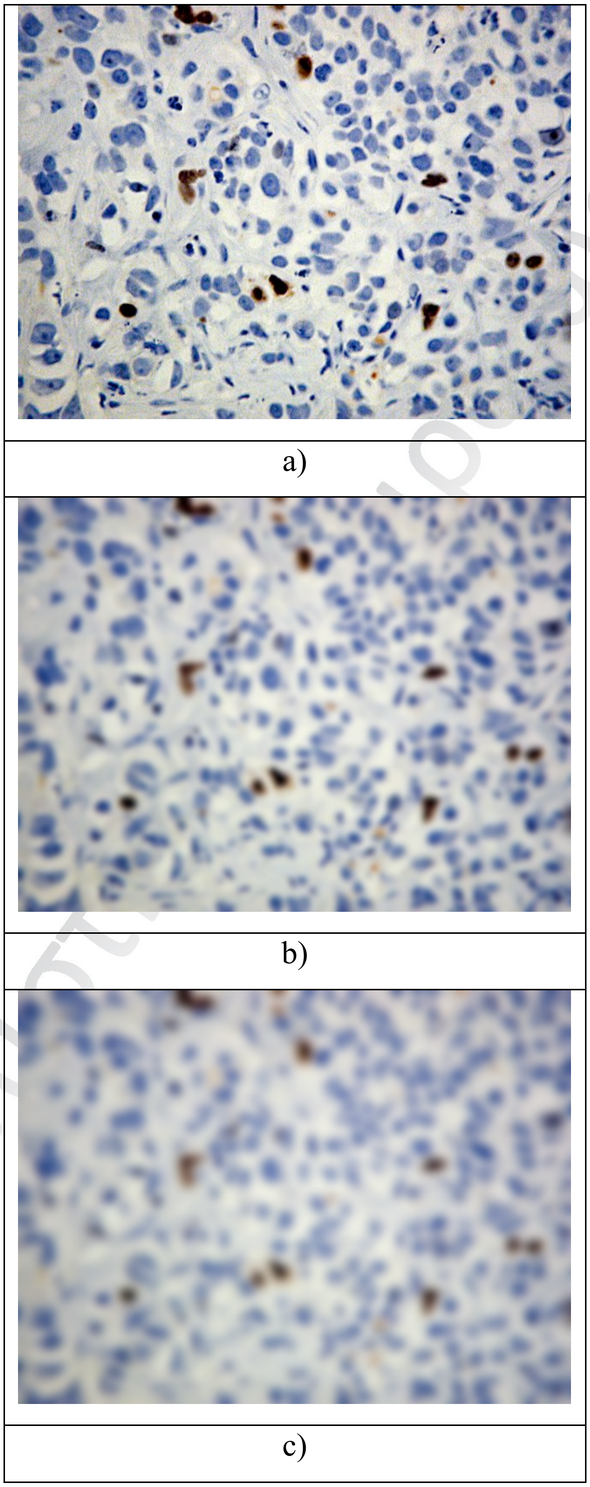


Figure 30. *a) original image b) Gaussian filtering with a 3 pixel sigma radius setting c) Gaussian filtering with a 5 pixel sigma radius setting*

Sharpen

Sharpen operator (see Figure 31) represents the un-sharp masking technique [17], which it is based on the removal of the smoothed areas of the image. Its functionality is practically an emulation of the traditional technique used for the improvement of the photographic or biomedical images taken.

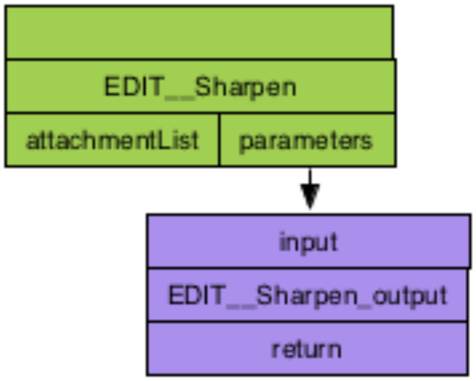


Figure 31. Sharpen Operator

The specific technique enables the combination of high and low frequent ncy elements of the image. Sharpen operator can be applied iteratively, enhancing the sharpening effect for each application loop (se Figure 32 b) and c)).

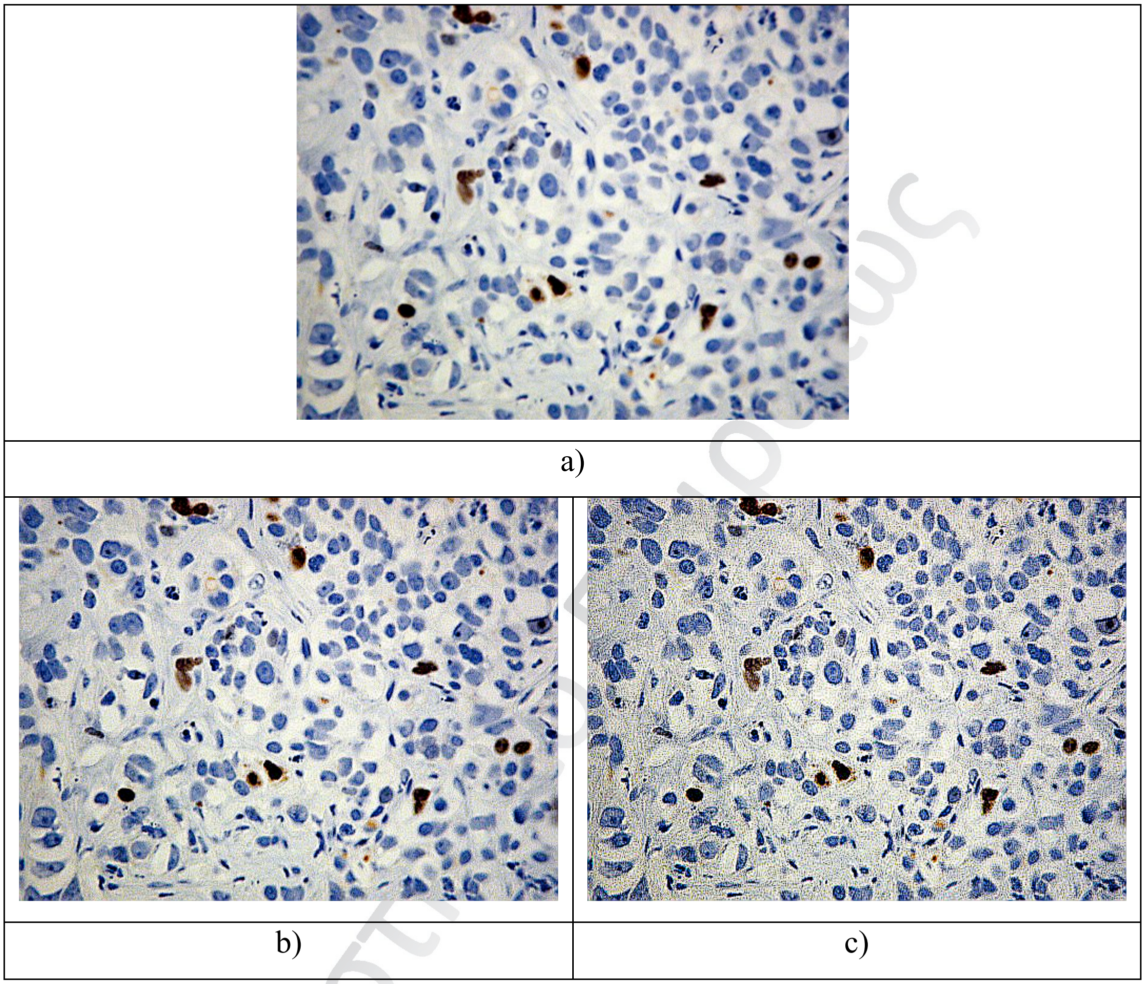


*Figure 32. a) original image b) the original image after the application of 1*Sharpen filtering operator c) the original image after the application of 2*Sharpen filtering operator.*

Enhance Contrast and Histogram Equalization Operator

The histogram of an image is defined as the distribution of the gray (or each channel's intensity in case of RGB images) tones of the images, and in most cases, it uniquely determines the image. A histogram is a graph, in which the horizontal axis contains all the brightness values from 0 to 255 and the vertical axis holds the frequency of each of brightness value of the image. Depending on the application, the vertical axis can be normalized by the maximum value of the histogram. If the maximum value of the vertical axis is equal to 1, then the corresponding histogram depicts the density probability distribution of the gray tones in the image.

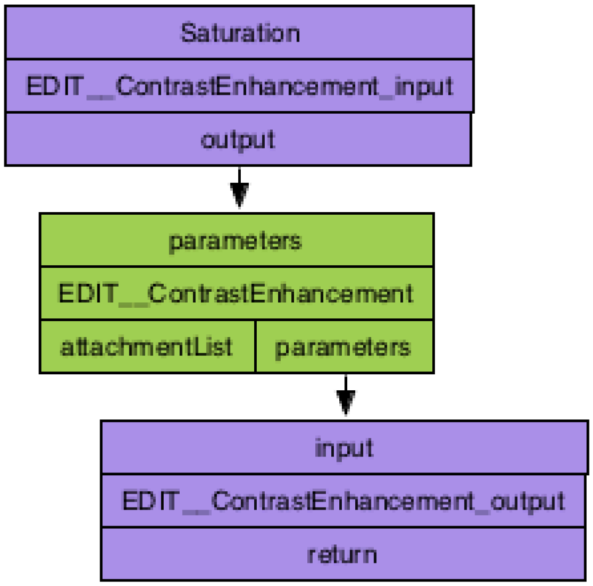

Saturation
EDIT__ContrastEnhancement_input
output
parameters
EDIT__ContrastEnhancement
attachmentList
parameters
input
EDIT__ContrastEnhancement_output
return

Figure 33. Contrast Enhancement Operator

The Saturation argument (see Figure 33) defines the percentage of the image's pixels that are permitted to become saturated. The greater the argument value, the highest the increment of the image's contrast (see Figure 34). If the user set the value to 0, no change will be applied on the image.

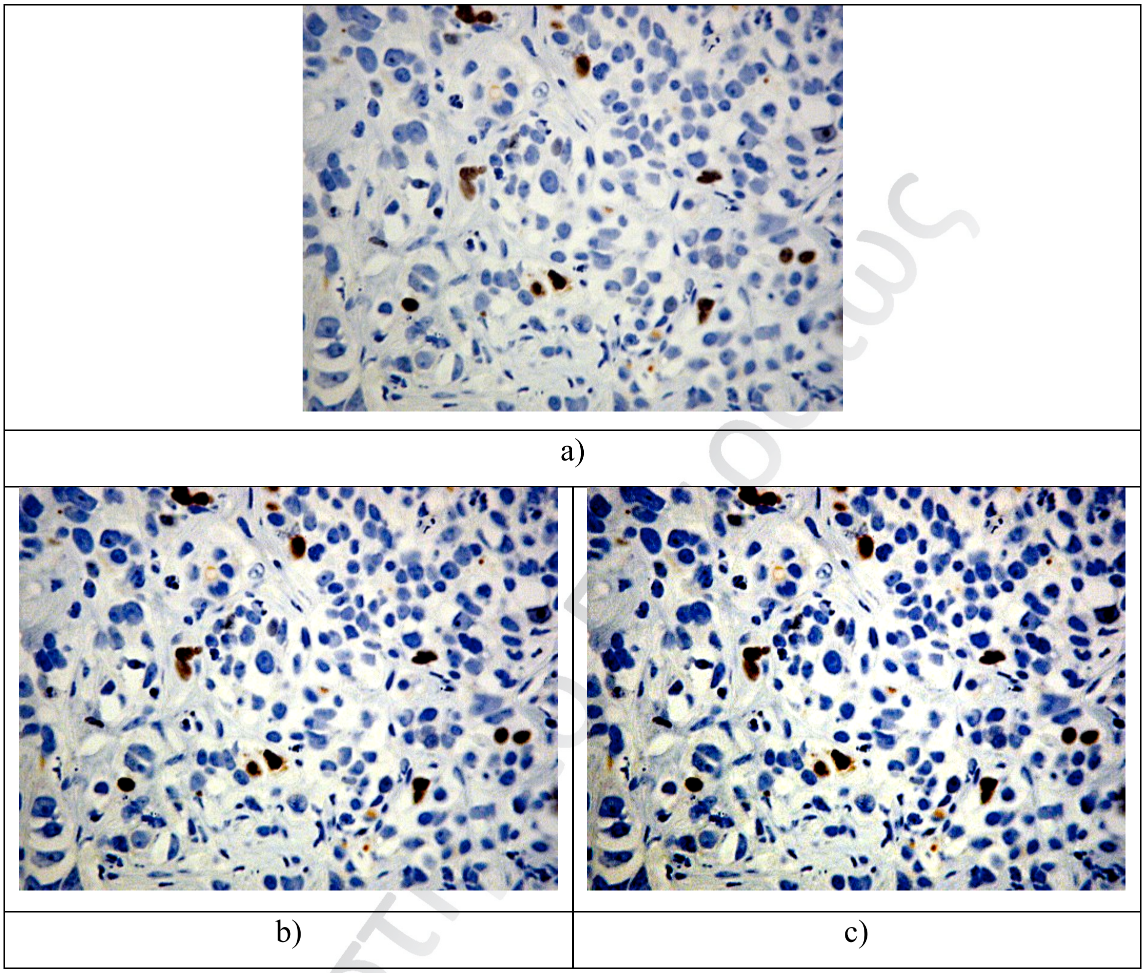


Figure 34. a) Original image b) the original image after the application of Contrast Enhancement with a 5% saturation input argument c) the original image after the application of Contrast Enhancement with a 15% saturation input argument.

Histogram equalization [18] operator transforms the brightness of each channel, in cases of RGB images, or the gray pixel brightness, in cases of gray scale images, so that they are evenly distributed through the rang of brightness [18]. Through this technique, increment of the contrast is usual achieved. This operator receives as input the percentage of the saturated pixels (see Figure 35). The higher the value, the greater the contrast increment is (see Figure 36).

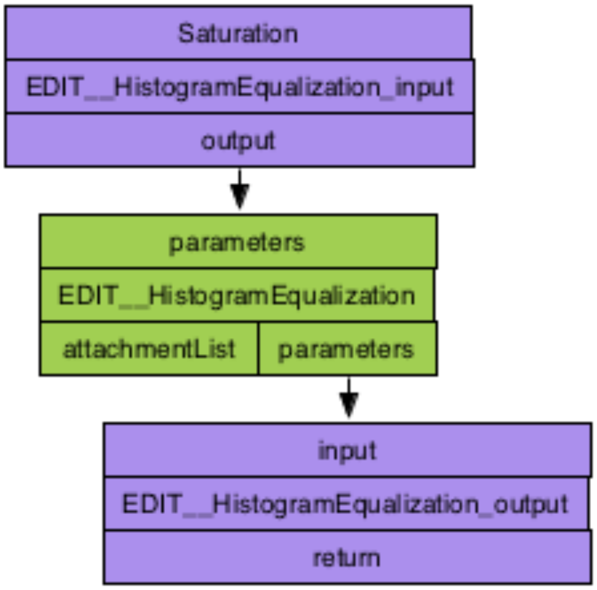

Saturation
EDIT__HistogramEqualization_input
output
parameters
EDIT__HistogramEqualization
attachmentList
parameters
input
EDIT__HistogramEqualization_output
return

Figure 35. *Histogram Equalization Operator*

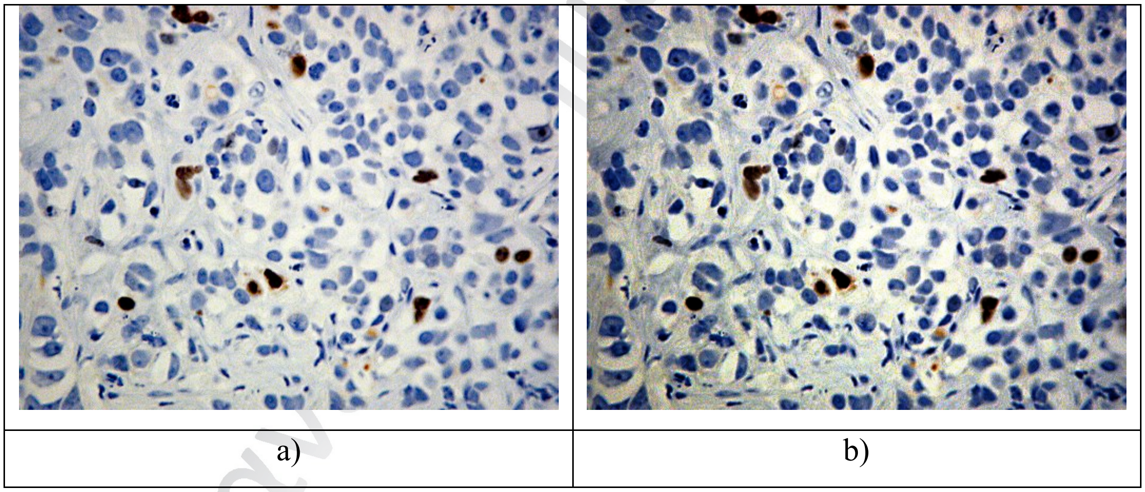


Figure 36. *a) original image b) the original image after the application Histogram Equalization operator*

Find Edges Operator

The functionality of Find Edges operator (see Figure 37) is based on the Canny's edge detection algorithm [19], which is the most efficient between the edge detection algorithms. Basically, Canny's algorithm is not a regular edge detection algorithm. Specifically, it is a methodology for the optimal edge detection and drawing based on a number of criteria. Particularly, these criteria are the correct detection, the positioning and the valid drawing of each edge.

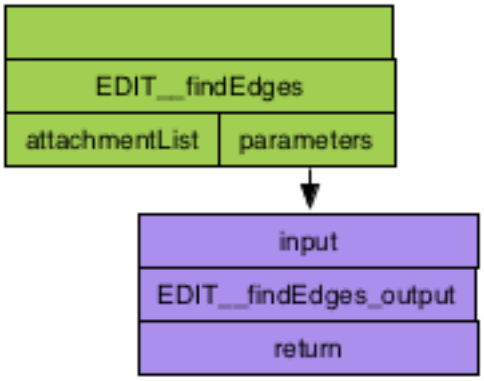


Figure 37. Find Edges Operator

The algorithmic steps of Canny's algorithm are the following:

1. Applying a Gaussian Filter of 0 mean and with a specified standard deviation.
2. Determination of the image's gradient by utilizing Sobel masks technique.
3. Suppression of the mid-range (non-maximum) values.
4. Hysteresis Thresholding achieved through further pixels filtering of the generated image, utilizing the technique described in [19] (after the application of the above steps).

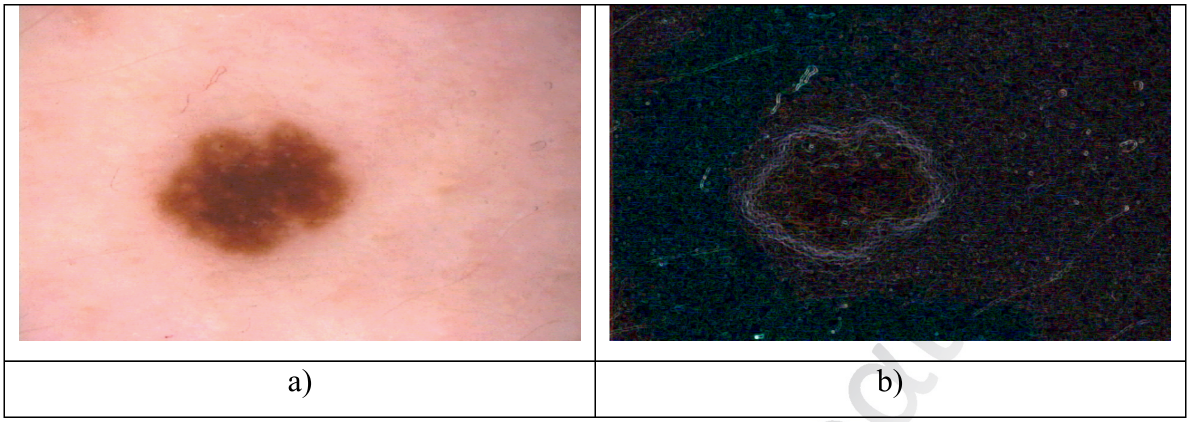

a)
b)

Figure 38. a) The original figure b) The figure after the application of Find edges operator

Smooth Operator

Similar to the Mean Filtering operator, it applies a blurring effect to the image, by replacing the value of each pixel with the mean value of its 3x3 neighborhood. This operator was developed for users with moderate image analysis skill.

Watershed Operator

Watershed operator is mostly a meta-segmentation filtering procedure, utilizing the watershed-filtering algorithm [20]. More specifically, it is utilized, in order to separate the adjacent salient objects of the image that either are collided, or are near enough to confuse the segmentation algorithm and miss-characterize them as one object. It is applied mostly on binary images and requires no input-setting arguments.

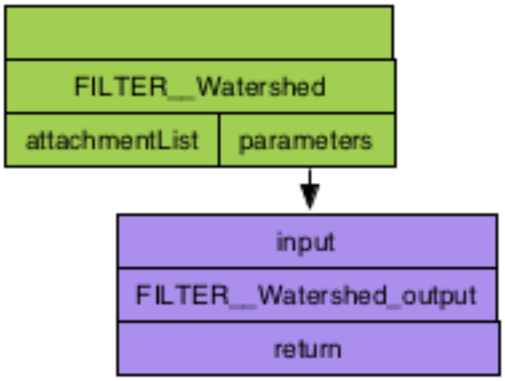
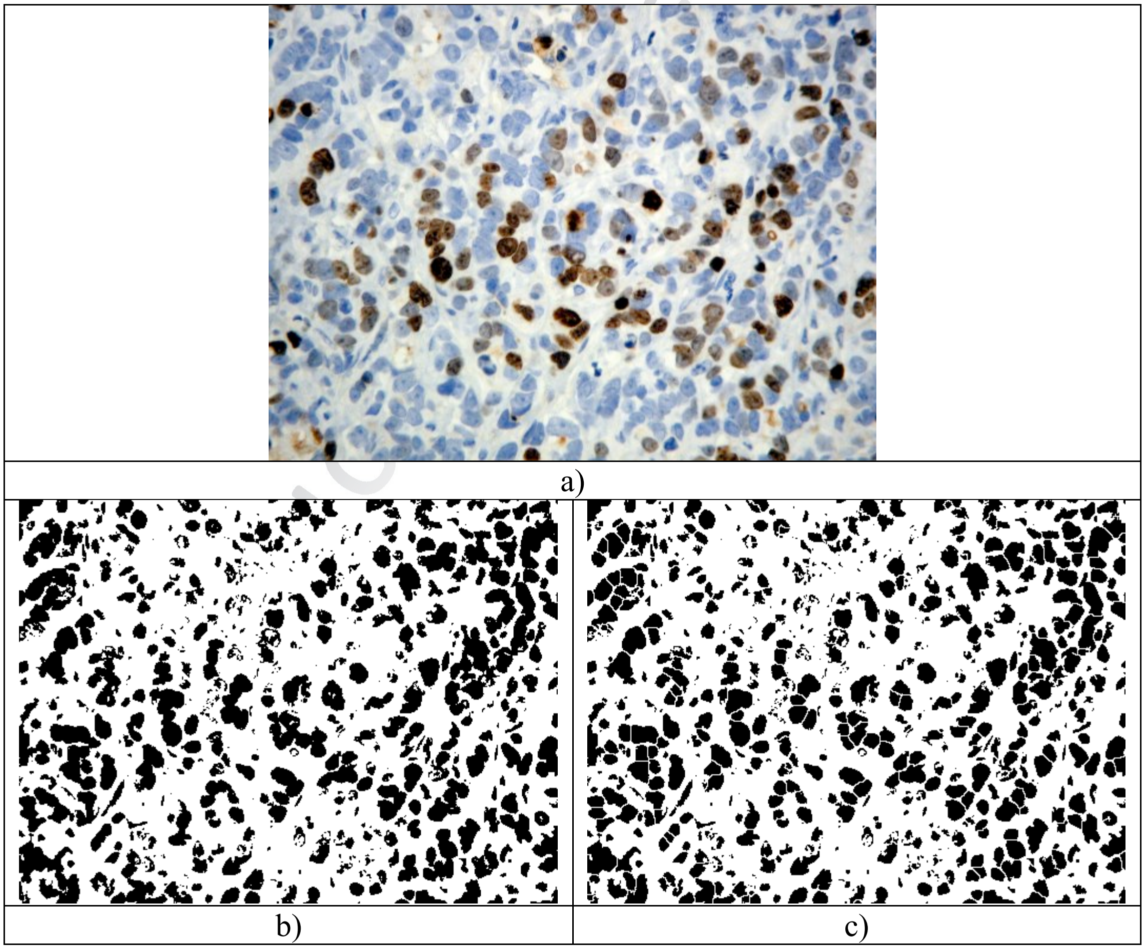


Figure 39. Watershed Operator


a)

b)

c)

Figure 40. a) Original Image b) Otsu segmented image c) watershed filtering of the segmented image

Afterwards, the watershed filter is applied on this image, in order to accurately cut the merged particles and provide a clear image for the characterization or quantification procedures.

Background Subtraction Operator

The background algorithm subtraction operator (see Figure 41) is based on the idea of the 'rolling ball' algorithm described in [21]. Based on the rolling ball algorithm, the grayscale image is considered as a 3D surface, where the third dimension is the value of each pixel of the image. Thus, a rough surface is generated. A ball with a specific radius is rolled over the raw surface. The ball's size must be equal or bigger than the larger foreground item. Each point - reachable by the ball - is the background. Thus, for every pixel, a specific background value is set, based on the local average of the ball's radius. These values are subtracted from the original image, providing an image with a clear (bright) background.

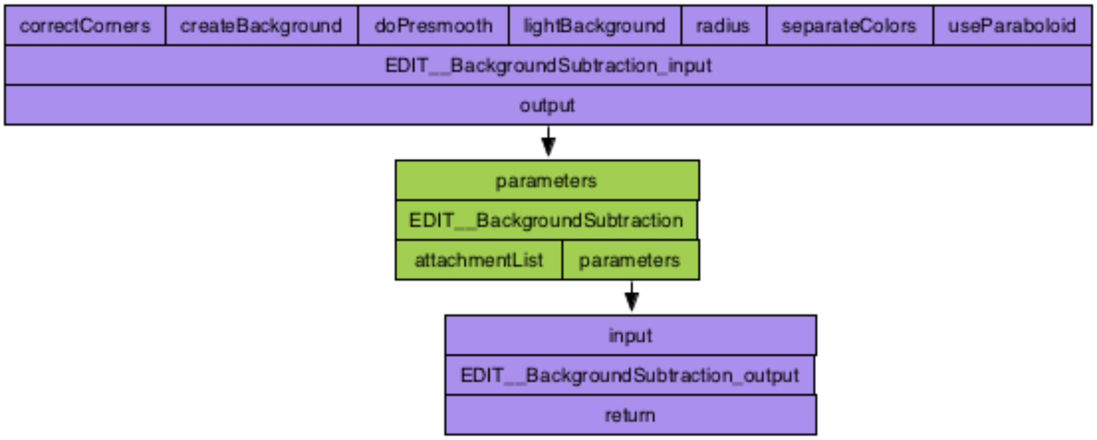


Figure 41. Background Subtraction Operator

The operator receives the following arguments (see Figure 41):

Rolling Ball Radius: This argument sets the radius of the paraboloid's curvature, which is basically the radius of the imaginary rolling ball described above. The smaller the radius is, the greater the effect that will be applied on the image. Of course, this does not mean that the optimal radius is the smallest. Very small radius values may increase the background noise, instead of removing it.

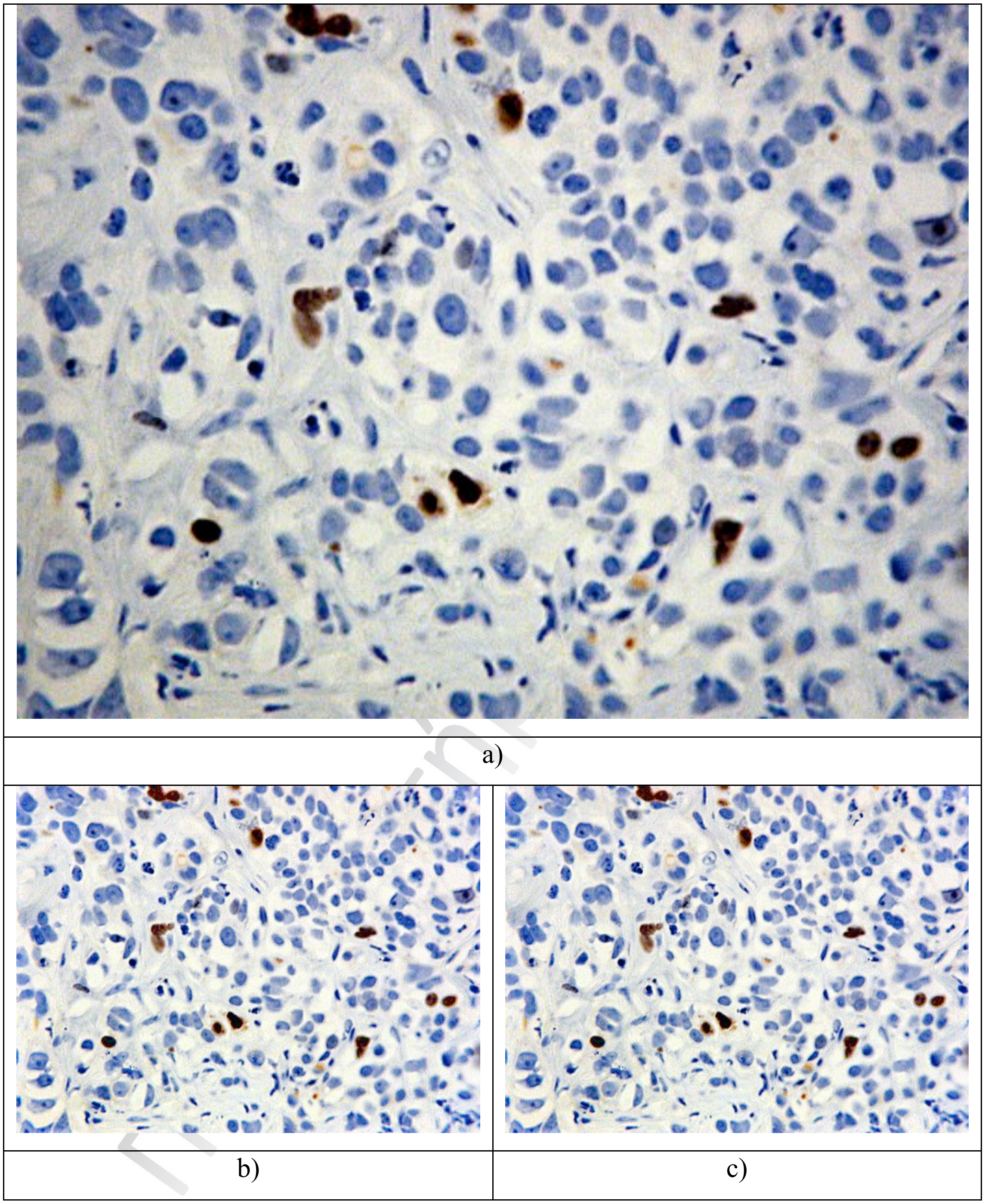
Light Background: This is a Boolean argument and determines if corresponding background is darker or lighter than the salient objects.

Separate Colors: This is a Boolean argument that determines if the functionality of the algorithm will affect only the brightness and leave the saturation and the hue of the image unchanged.

Create Background (Don't Subtract): This is a Boolean argument, which, if it is true, it exports only the background. Practically it provides an opposite result of the algorithm. It is useful in cases that the examination of the background is crucial.

Sliding Paraboloid: The application of background subtraction algorithm removes the areas that are considered as background. Due to the rolling ball functionality, it may leave some artifact objects. These objects belong to background, but due to this error, may be miss-considered as foreground. The sliding paraboloid does not use downscaling, a procedure step of the rolling ball algorithm. By setting this parameter as true, the downscaling procedure is removed, thus, there are no generated artifact objects.

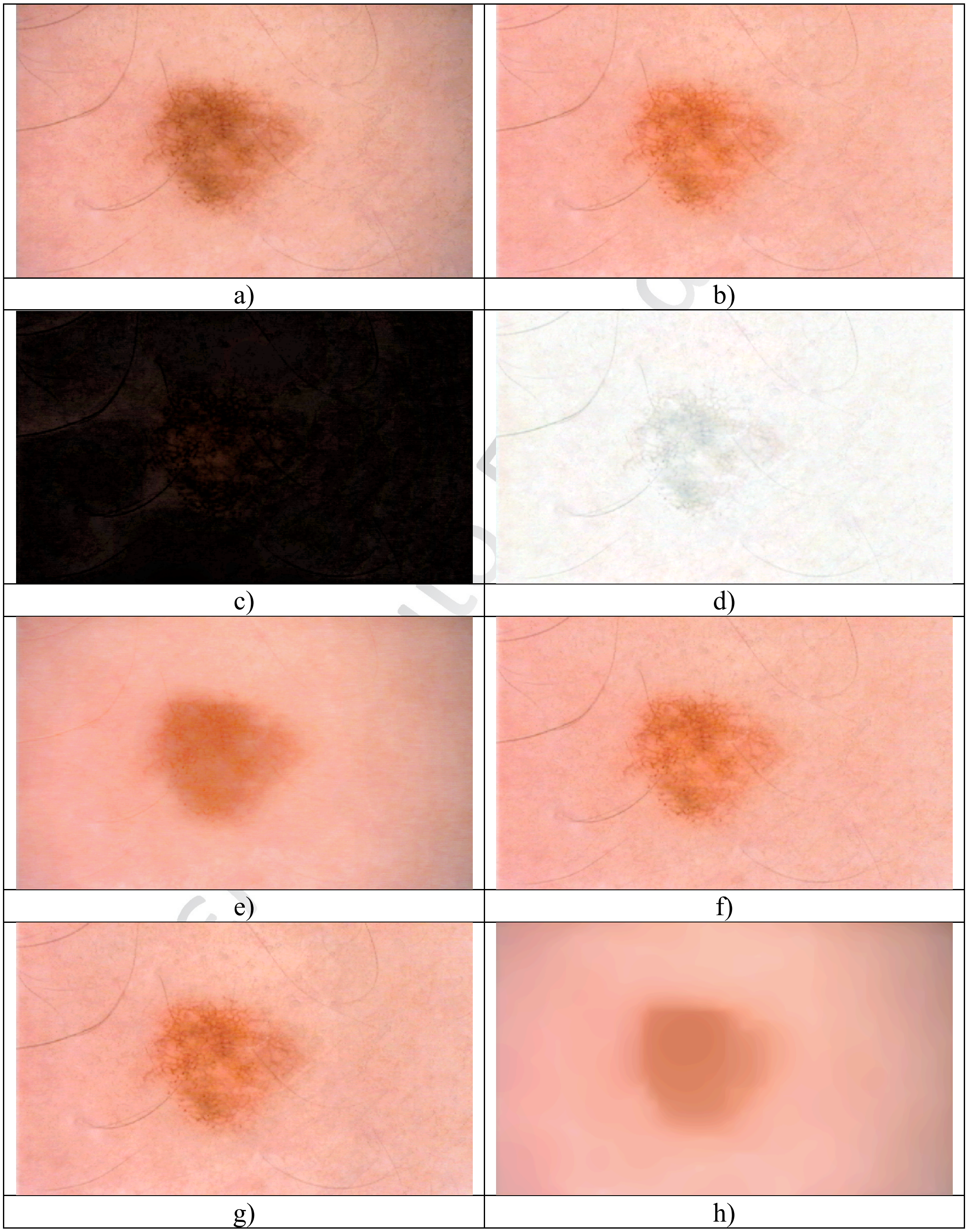
Disable Smoothing: Following the background subtraction, images are filtered with a max-filter (3x3), in order to remove potential generated artifact objects. By setting this argument as true, the max-filtering process is removed.


a)

b)

c)

Figure 42. a) Original image b) background subtraction application with 50 pixels ball radius and light background enabled


a)

b)

c)

d)

e)

f)

g)

h)

Figure 43. a) Original Image b) Background Subtraction application with 50 pixels ball radius and light background enabled c) Background Subtraction application *with 50 pixels ball radius and light background disabled d) Background Subtraction application with 50 pixels ball radius and light background and separate colors enabled e) Background Subtraction application with 50 pixels ball radius and create background light background enabled f) Background Subtraction application with 50 pixels ball radius and sliding paraboloid and light background enabled g) Background Subtraction application with 50 pixels ball radius, light background enabled and smoothing disabled h) Background Subtraction application with 50 pixels ball radius and create background, separate colors and light background enabled.*

Majority Voting Operator

In some segmentation or classification tasks, some pixels or larger areas may be misclassified. In order to remove misclassification issues, a morphological filtering is required. A majority vote technique described in [22], was utilized, in order to accomplish this task. The value of the central segmentation square is set by the majority vote of its eight neighbor segmentation blocks. The majority limit is set to five, which means that if five or more neighbor segmentation squares have the same value, which differs from the value of the central one; it is automatically set to the same value with its neighbors. The operator receives the size of the window of the neighborhood area.

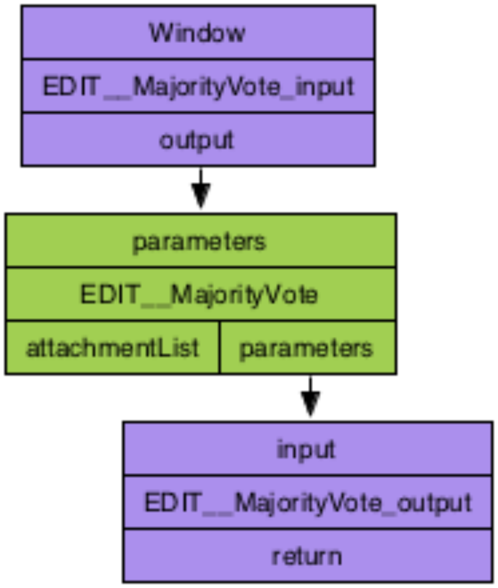
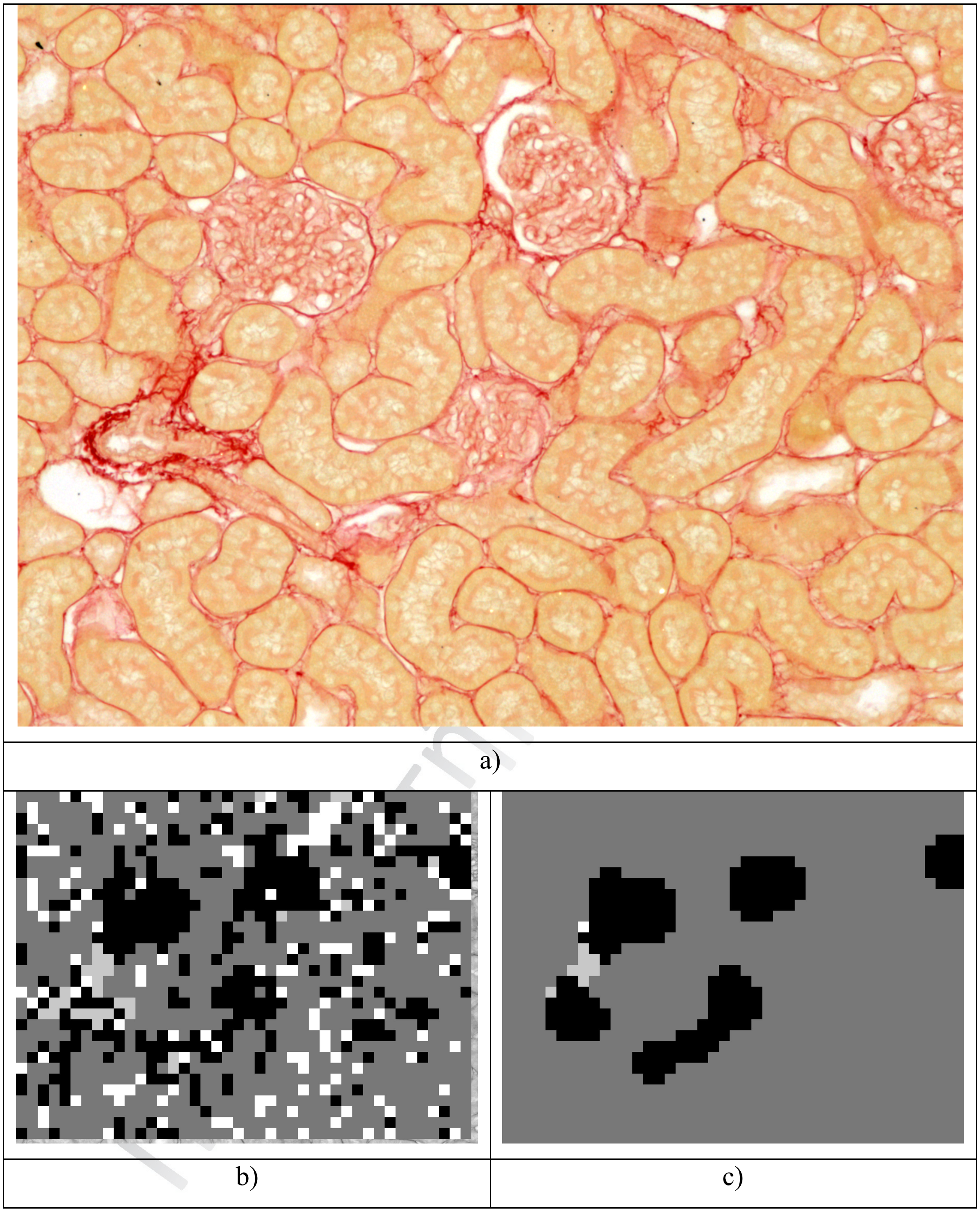


Figure 44. Majority Voting Operator

Consider the case of a segmented image, containing misclassification issues (see Figure 45 b)). The application of majority voting operator will provide clearer segmentation results (see Figure 45 c)), and in most cases, more accurate results.



a)

b)

c)

Figure 45. a) Original Kidney Biopsy Image b) Segmented Image containing misclassification issues c) Segmented image after the application of majority voting operator

Minimum Object Removal Operator

Size is another major factor for salient objects in biomedical images. Size may characterize entirely an object. For that reason, an analyze particle method, similar to a component labelling algorithm [23], was developed. In our first try to develop a customized component labelling algorithm [24], we developed a similar one, which worked perfectly on 4-connected patterns, but it had some issues dealing with 8-component patterns. The new enhanced algorithm provides faster scanning and noise removal. This algorithm may apply a two-pass check, but it is quite faster than the previous algorithm. More specifically, the developed algorithm scans the image from right to left, beginning from the top and moving to the bottom, until it finds a pixel of a foreground object, in our case a black one, since we deal with a binary image. The algorithm checks the specific's pixel neighbor pixels. It first checks the left pixel if it has already been labelled. If it is labelled, it assigns the same label to the current pixel. If it is not, it continues a clockwise searching of neighbor pixels up to the upper right pixel. If none of these four neighbor pixels have an assigned label, then a new label is created and assigned to the current pixel. This procedure continues for every foreground pixel (black) of the binary image. When it is complete, a second scan begins, but this time, it scans the image from right to left, but this time from the bottom and moving to the top. Considering that this time every foreground object has already an assigned label, we won't have the case of assigning a new label and the checks applied will be readjusted. When the algorithm finds a pixel of a foreground object (a labelled pixel), it checks, if the left pixel is labelled. If it is, it changes the current label to the left pixel's label. If not, it continues a counter-clockwise searching of neighbour pixels up to the lower right pixel. If none of these four neighbor pixels has an assigned label, then it leaves the current label unchanged.

The operator that exploits the above algorithm receives as input the minimum acceptable size of a foreground object (see Figure 46).

3.4.2.1.1 Pseudocode of the object removal algorithm

```
INPUT: Binary Image after Adaptive Thresholding
INITIALIAZATION: set the values of the 2D integer arrays Label (which contains the label for each pixel of the image) and Object_Array (which contains each foreground pixel’s label and its coordinates) to zero, set the values of the 1D integer array Size to zero, set the integer Pos to 0, integer Counter_of_forgroundpixels is set to 0, integer Lab is set to 0 and integer new_label is set to 1. Integer Acceptable_size is set by the user.
FOR y=0 to the Image Height-1 DO
    FOR x=0 to the Image Width-1 DO
        IF Pixel[x,y] belongs to a foreground object (pixel is black) THEN
            Counter_of_forgroundpixels++
            IF the left pixel’s label is different than 0 THEN
                Label[x,y]= left pixel’s label
            ELSE IF upper left pixel’s label is different than 0 THEN
                Label[x,y]= upper left pixel’s label
            ELSE IF upper pixel’s label is different than 0 THEN
                Label[x,y]= upper pixel’s label
            ELSE IF upper right pixel’s label is different than 0 THEN
                Label[x,y]= upper right pixel’s label
            ELSE
                Label[x,y]= new_label
                new_label ++
            END IF
        END IF
    END FOR
END FOR
//second pass
FOR y= Image Height-1 to 0 DO
    FOR x=0 to the Image Width-1 DO
        IF Label[x,y]!=0 THEN
            IF the left pixel’s label is different than 0 THEN
                Label[x,y]= left pixel’s label
            ELSE IF lower left pixel’s label is different than 0 THEN
                Label[x,y]= lower left pixel’s label
            ELSE IF lower pixel’s label is different than 0 THEN
                Label[x,y]= lower pixel’s label
            ELSE IF lower right pixel’s label is different than 0 THEN
                Label[x,y]= lower right pixel’s label
            END IF
            //set values to Object_Array which will help us to remove objects smaller than the acceptable
        //size
            Object_Array[0,Pos] = Label[x,y]
            Object_Array[0,Pos] = x
            Object_Array[0,Pos] = y
        END IF
    END FOR
END FOR
//Invalid object removal
FOR i = 0 to Counter_of_forgroundpixels-1 DO
    Lab = Object_Array[0,i]  // it reads the label of the object
    Size[Lab] = Size[Lab] + 1 // it increases the size of the label assigned to a specific object
END FOR
FOR i = 0 to new_label DO
    IF Size[i] < Acceptable_size THEN
        Remove each pixel assigned with the specific label i
    END IF
END FOR
```

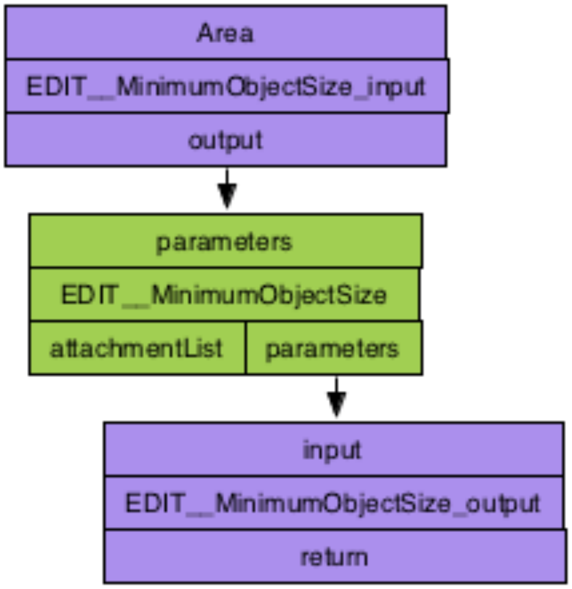


Figure 46. Minimum Object Removal Operator

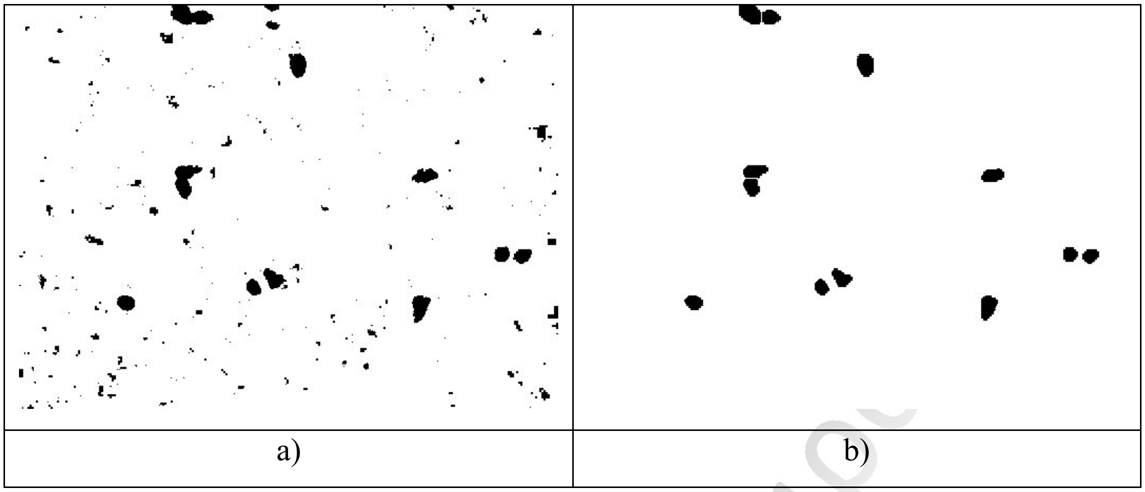


Figure 47. a) Segmented image containing noisy areas b) The segmented image after the application of Minimum Object Removal Operator

3.4.3 Segmentation Operators

Segmentation operators are another operator category that depends on core-type operators, in order to apply their functionalities on the workflow-imported images.

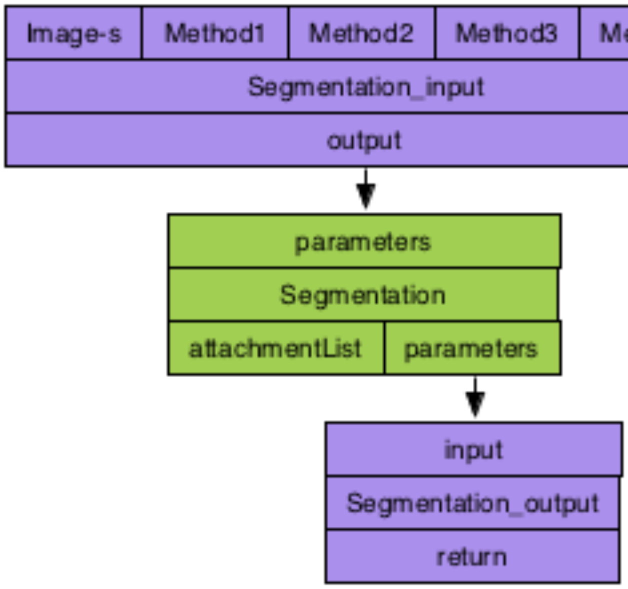


Figure 48. Core Segmentation Operator, along with its available settings

More specifically, the developed Segmentation operator (see Figure 48) receives as input one (at least) or up to 4 filtering methods. The number of the integrated methods indicates the number of the different-independent generated instances of one image. The operator requires the images that will be processed by the one or more integrated segmentation techniques. The operator exports the above-mentioned number of processed images, separated into different flows, in order to distinct later the most effective one.

Otsu Segmentation

Otsu's segmentation algorithm [25] was proposed in 1979 and it is considered one of the most efficient thresholing segmentation algorithms. The selection of the optimal threshold value is based on the maximization of the difference between the dark and the light areas. Otsu's approach assumes that there are only two classes in the image (usually foreground and background pixels). There is also an extension of the Otsu's methodology that allows the detection of more than two classes in the image. It is called Multi-Otsu [26].

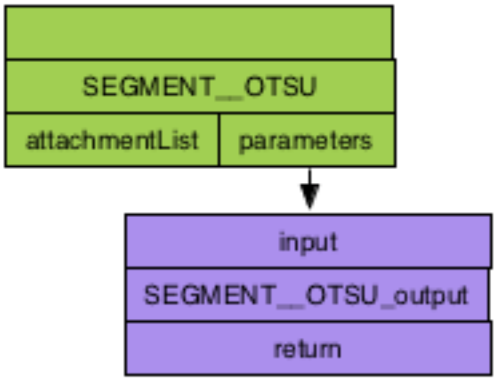


Figure 49. Otsu Segmentation Operator

The algorithmic steps are the following:

The intra-class variance calculation is presented below, which it is typically the weighted variance of the foreground pixels class plus the weighted variance of the background pixels class:

$$\sigma_w^2 = w_b(t)\sigma_b^2(t) + w_f(t)\sigma_f^2(t)$$

Where,

b: the background class

f: the foreground class

w: the probability of each class

t: the separating threshold

σ_x^2: the corresponding variance of each class

b: denotes the background pixels class

f: denotes the foreground pixels class

The probability of each class is calculated via the image's histogram.

$$w_b(t) = \sum_0^t p(i)$$

The class mean equals:

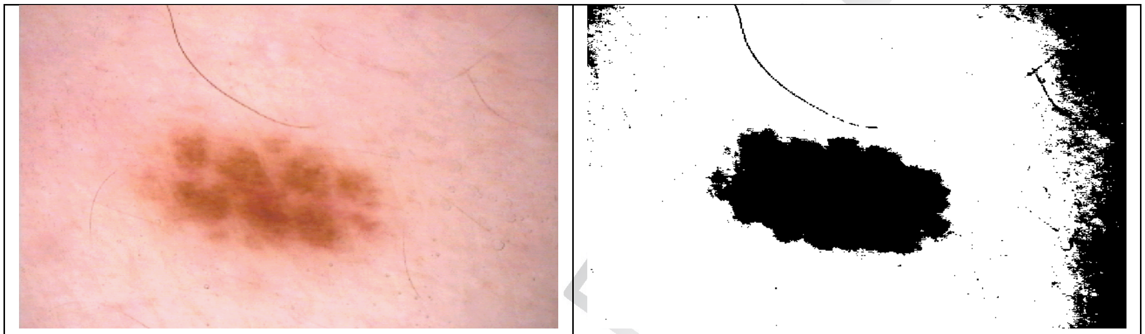
$$\mu_b(t) = \frac{\sum_0^t p(i)x(i)}{w_b}$$

Where,

x(i): the i^{th} histogram bin's center value.

Particularly, the algorithm calculates the histogram and exports the probabilities of the classes for each brightness level. Following the initialization of the class probability (from $w_i(0)$) and the class mean (from $\mu_i(0)$), an iterative procedure checking each potential threshold value, starting from 1 and reaching the maximum brightness value, is taking place. Specifically, it iteratively updates the w and μ values and calculates the corresponding σ_b^2 until it finds the maximum difference between the foreground and background pixels classes. This implies to the greatest values of a σ_b^2 (threshold1_value) and σ_f^2 (threshold2_value). Thus, the optimal threshold value equals:

$$Optimal\,Threshold = \frac{\text{threshold1\_value} + \text{threshold2\_value}}{2}$$



a) b)

Figure 50. a) original Image b) after the application of Otsu segmentation operator

Mean Shift Operator

The Mean Shift operator (see Figure 51) is based on the corresponding mean shift algorithm [30]. This segmentation algorithm is capable of handling any kind of feature space (color space, scale space, etc.) and does not require any prior knowledge about the number of the clusters. It is based on Gradient Ascent theory and it was introduced by Fukunaga and Hostetler in 1975.

3.4.3.1.1 Gradient Ascent

Gradient Ascent [27] is a global approach for the solution of optimization issues, where there is a necessity of maximizing functions of continuous parameters. For instance, in order to find the maximum value of a function *y=f(x)* (consider *f(x)* as a difficult equation). The algorithm for finding the maximum is the following:

- *y=f(x)* has a local maxima at x_{max}
- x_1 a random value from the data range

- calculate $f(x_1)$
- if there is a positive slope of y ($f'(x_1)$>0) then the x_{max} is greater than x_1 and it is found at the right of x_1.
- on the other hand, if *f ’(x1)*<0 then the x_{max} will be less than x_1, which implies that it is found at the left of x_1.

In this way, the direction in which $x_{1,}$ should move to, in order to approach x_{max} is known.

$x_1 \leftarrow x_1 + nf'(x_1)$ (1)

Where *n* is a positive constant. In case of extreme small *n*, there is also a local maxima for *f*, the above rule will converge to it, after a specific number of loops.

3.4.3.1.2 Mean Shift Segmentation

The sampling of the data points is performed from an underlying probability density function. When non-parametric density estimation is applied on the input data points, a discrete probability density function representation is produced. On the other hand, if a non-parametric gradient estimation is applied to the same source, a full probability density function analysis is produced.

Consider $\{x_i\}_{i=1\dots n}$ as a set of n points in a N-dimensional Euclidian space. The kernel density estimation [27] is calculated for the point x with a kernel K(x) [28].

$\hat{f}(\mathrm{x}) = \frac{1}{nh^N}\sum_{k=0}^{n} \mathrm{K}\left(\frac{x-x_i}{h}\right)$(2)

where h, is the window radius.

The Epanechnikov kernel provides the minimum mean integrated square error.

$$K_E(x) = \begin{cases} \frac{1}{2}c_N^{-1}(N+2)(1-x^T x), \ if\ x^T x < 1 \\ 0, \qquad\qquad\qquad\qquad\qquad otherwise \end{cases} (3)$$

where c_N is the volume of the N-dimensional sphere unit [28].

Mean shift’s calculation is based on the gradient of the kernel density estimation (2).

$\hat{\nabla} f(\mathrm{x}) \equiv \nabla \hat{f}(\mathrm{x}) = \frac{1}{nh^N}\sum_{i=1}^{n} \nabla K\left(\frac{x-x_i}{h}\right)$(4)

The mean shift of data point x is calculated through the formula below.

$$m(x) = \frac{\sum_{i=1}^{n} g\left(\frac{x-x_i}{h}\right) x_i}{\sum_{i=1}^{n} g\left(\frac{x-x_i}{h}\right)} - x(5)$$

where g(x)= - K'(x). Thus, a typical mean shift iteration procedure includes the computation of mean shift vector $m(x_i^t)$ and the move of the density estimation window by $m(x_i^t)$ for each data point x_i. The necessary formulas for the proof of the mean shift's formula are included in [27].

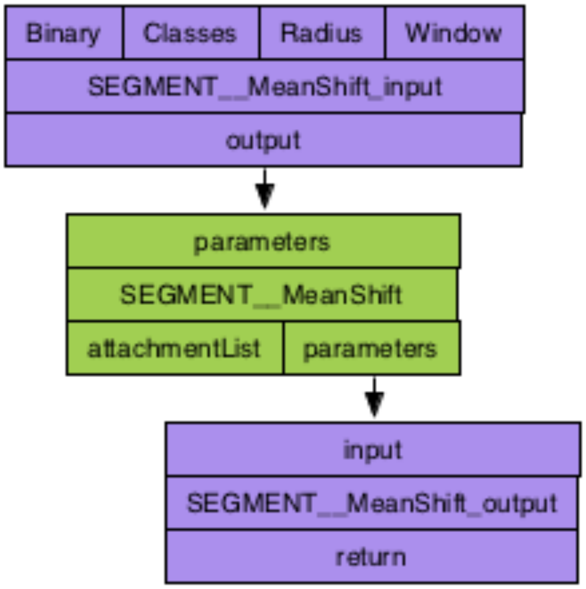
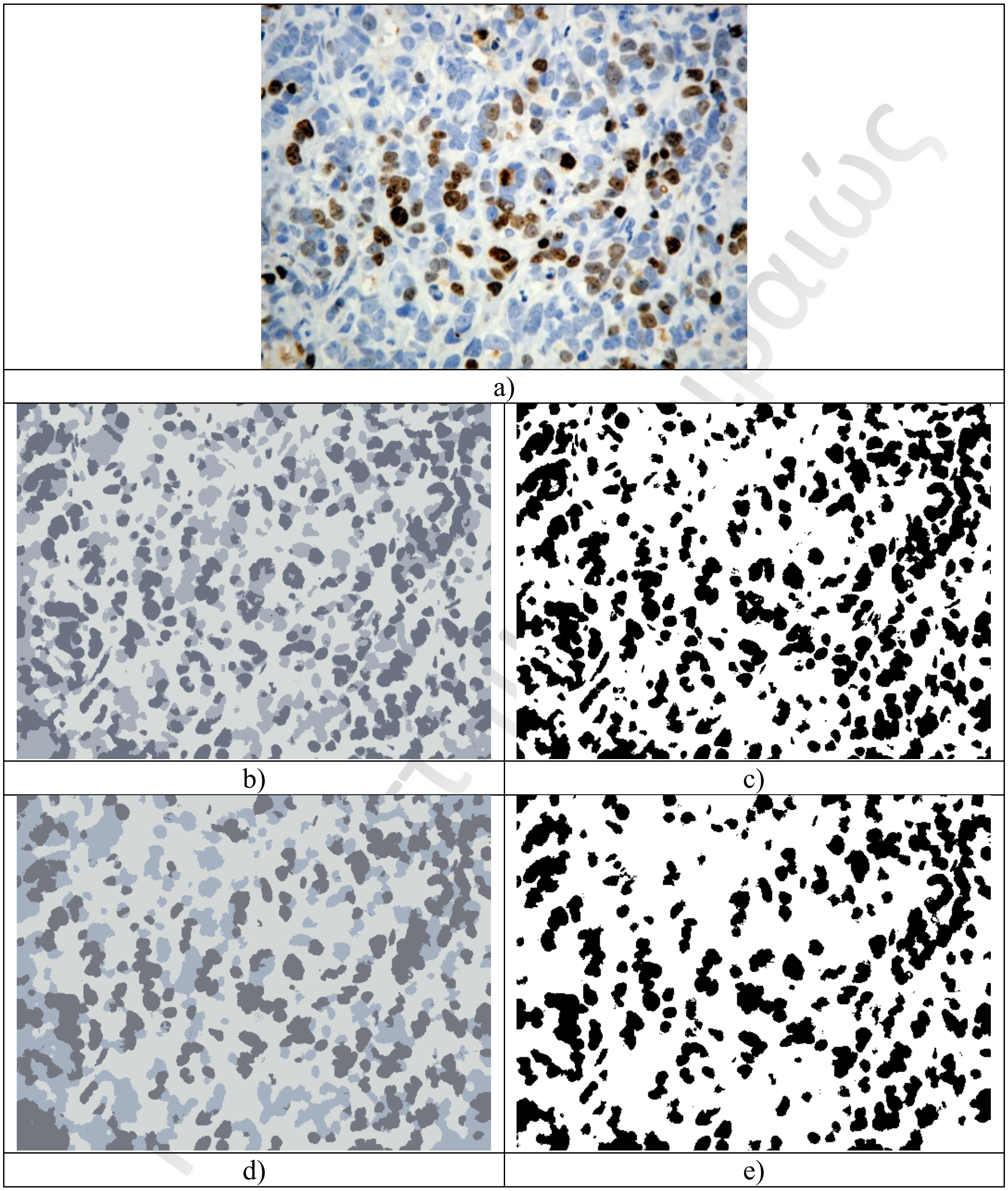


Figure 51. Mean Shift segmentation operator along with its available settings

The mean shift operator (see Figure 51) receives four inputs. Three of them are mandatory and the other one is optional (task specified-focused). The first mandatory input is the spatial radius, the second one is the color distance value and the third one indicates the number of classes for the segmentation process. Although the multicolored output of the mean shift algorithm, the operator contains the option to provide binary results in cases of two-class problems. The default value for this feature provides the binary version of the mean shift algorithm. Thus, the binary optional input receives a Boolean argument for the binary version.


a)

b)

c)

d)

e)

Figure 52. a) Original Image b) Mean Shift application with 5 spatial radius and 25 color range (Binary input false) – customized for 3 classes c) Mean Shift application with 5 spatial radius and 25 color range (Binary input true) d) Mean Shift application with 7 spatial radius and 50 color range (Binary input false) –

customized for 3 classes e) Mean Shift application with 7 spatial radius and 50 color range (Binary input true).

Li Segmentation

Li method [29] is another thresholding segmentation method. It is similar to Otsu's algorithm, but it selects the optimal threshold value with a different approach. This algorithm is based on the minimum cross entropy metric.

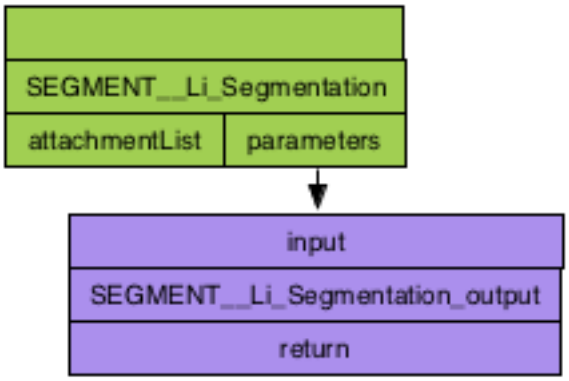
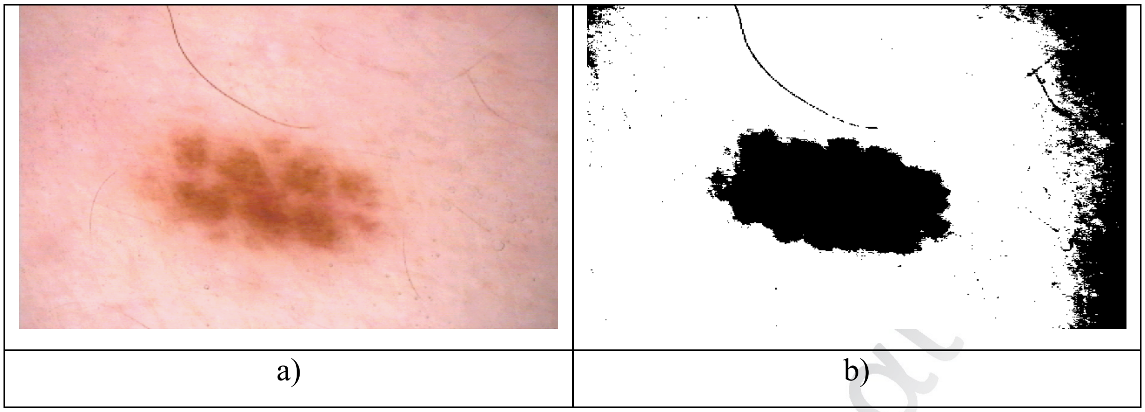


Figure 53. Li Segmentation Operator

More specifically, the optimal threshold value is the one that achieves the minimum cross entropy between the original and the thresholded image. Otsu's methodology is based on the variance metric. The cross entropy metric allows the optimal threshold selection, regardless the populations of the two distributions. In cases where there is no prior knowledge for the image mining of a dataset, the Li segmentation method will provide the most unbiased estimation of the threshold-based segmentation of the dataset's images (see Figure 54).


a)

b)

Figure 54. a) Original Image b) the initial image after the application of Li segmentation operator

Huang Segmentation

Huang thresholding segmentation algorithm [30] is another Otsu-like algorithm. This algorithm is based on fuzzy sets theory [31] and membership function. The difference to the Otsu's algorithm lies to the threshold selection phase, which applies a measure of fuzziness on the image that is going to be segmented.

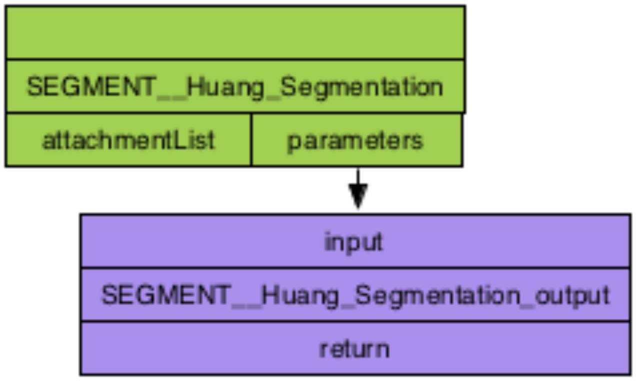
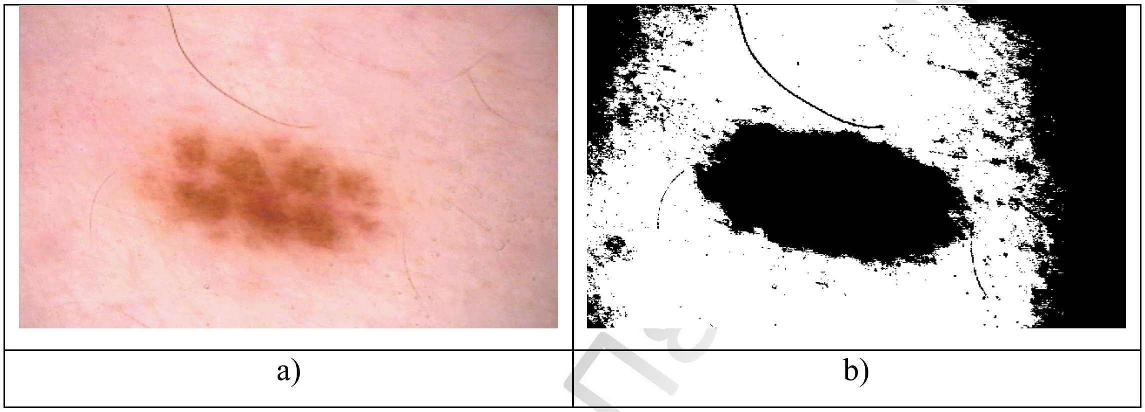

SEGMENT__Huang_Segmentation
attachmentList
parameters
input
SEGMENT__Huang_Segmentation_output
return

Figure 55. Huang Segmentation Operator

The fuzziness metrics are utilized by this method, are the Shannon's function and Yager's measure analyzed in [30]. These metrics assist to the detection of the deepest valley of the histogram.


a)

b)

Figure 56. a) Original Image b) the image after the application of Huang Segmentation Operator

Triangle Segmentation

Triangle segmentation algorithm [32] was proposed by Zack et al and it is another threshold-based algorithm. The algorithm follows a simple, but effective procedure, in order to find the optimal threshold value that will separate the two classes (foreground and background).

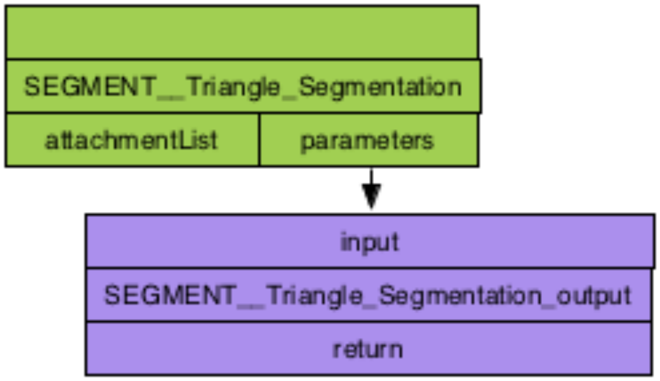

SEGMENT__Triangle_Segmentation
attachmentList
parameters
input
SEGMENT__Triangle_Segmentation_output
return

Figure 57. Triangle Segmentation Operator

The first step of the algorithm constructs an imaginary triangle by drawing a line in the image's histogram between the minimum brightness value and the maximum frequency. The distance between the line and the frequency of each value of the histogram is calculated. The maximum achieved distance denotes the optimal threshold value. This algorithm is very effective, in cases where the image's pixels construct weak peaks in the corresponding histogram.

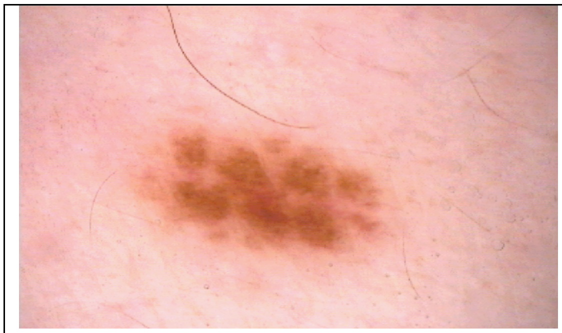
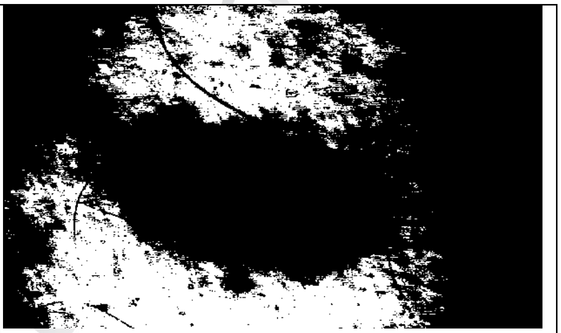
	
a)	b)

Figure 58. a) Original Image b) the image after the application of Triangle segmentation operator

Yen Segmentation

Yen's segmentation algorithm is a multi-level threshold-based algorithm [33]. The selection of the optimal threshold value is based on two criteria. The first one is a cost function [33], while the second one is the number of the required bits for the corresponding thresholded image representation.

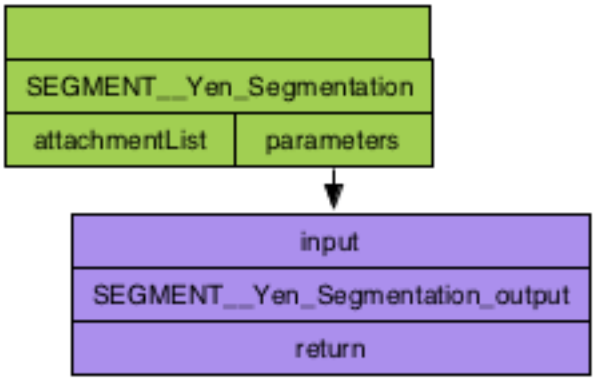
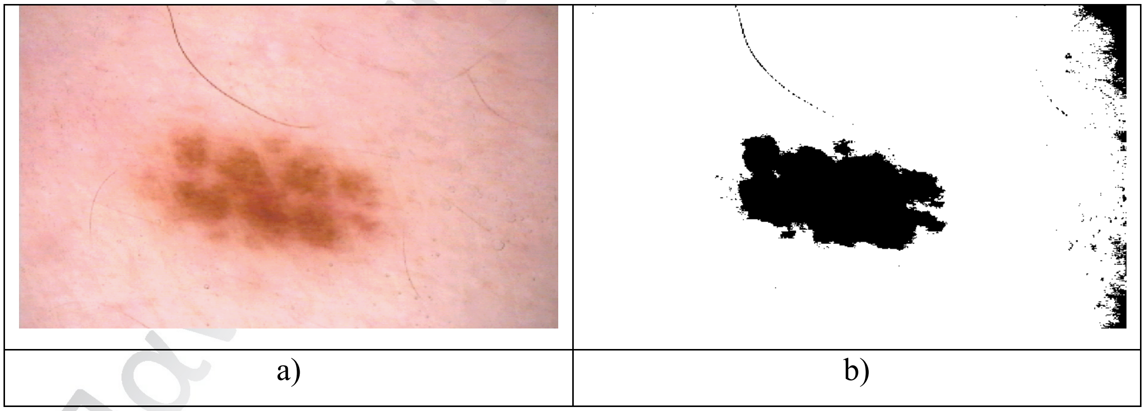


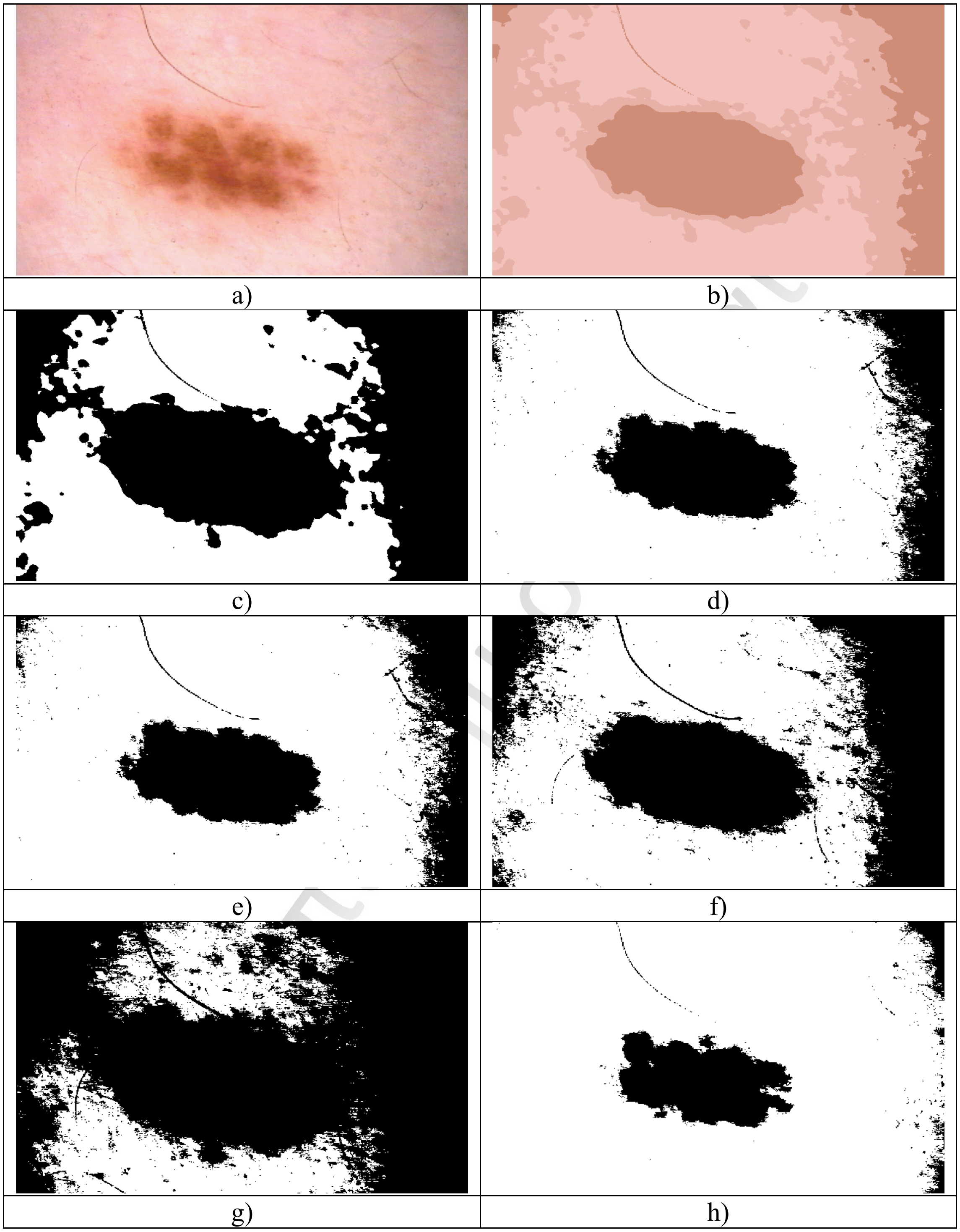
Figure 59. Yen Segmentation Operator

The number of the gray level classes of image's pixels is defined automatically. In cases where binary level segmentation is required, the maximum correlation criterion [33] between the initial and the thresholded and the initial image is utilized. The multilevel thresholding feature deals with the drawbacks of the similar threshold-based algorithms.



a) b)

Figure 60. a) Original Image b) the initial image after the application of Yen Segmentation Algorithm



a) b)

c) d)

e) f)

g) h)

Figure 61. a) Original Image b) After the application of Mean Shift operator with spatial radius of 5, color distance set at 25 c) After the application of Mean Shift operator with spatial radius of 5, color distance set at 25 (binary input set True) d)

After the application of Otsu Operator e) After the application of Li segmentation operator f) After the application of Huang Segmentation Operator g) After the application of Triangle segmentation operator h) After the application of Yen segmentation operator.

3.4.4 Feature Extraction Operators

Feature extraction operators are another operator category that depends on core-type operators, in order to apply their functionalities on the workflow-imported images.

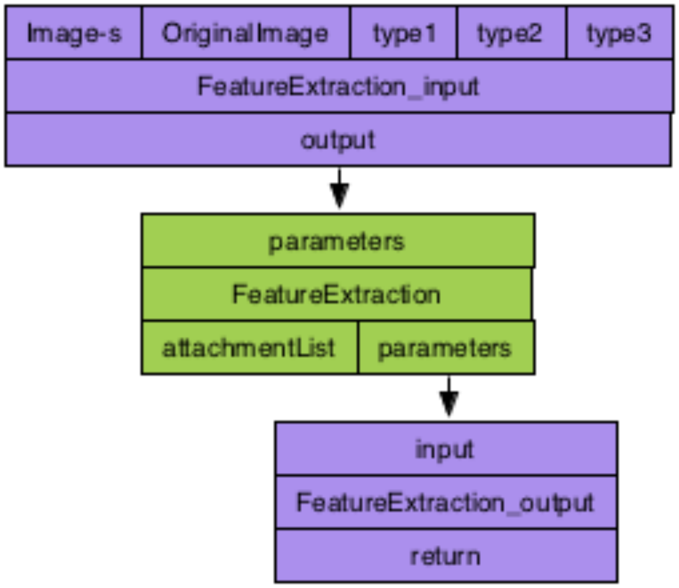


Figure 62. Feature Extraction core operator

The core Feature Extraction operator (see Figure 62) receives as input the segmented (binary) image-s, the corresponding original image-s and the selected features that are going to be extracted. Each type denotes a different instance of the workflow.

The feature extraction operators (see Figure 63) serve multiple purposes. They can provide additional characterization to the segmented objects, or they can provide the mandatory input to the clustering operators, in order to perform clustering of the detected objects and afterwards, to propose the optimal workflow scheme. The functionality of the core feature extraction operator is quite simple. It receives the annotation areas from the segmented binary images and exports the selected features from the corresponding areas of the initial images. The user can integrate textural and/or color feature extraction operators. There are three different operators for the parallel workflow instances generation: The Color Feature Extraction Operator, the Textural Feature Extraction Operator and the ALL Feature Extraction (which combines the previous two operators).

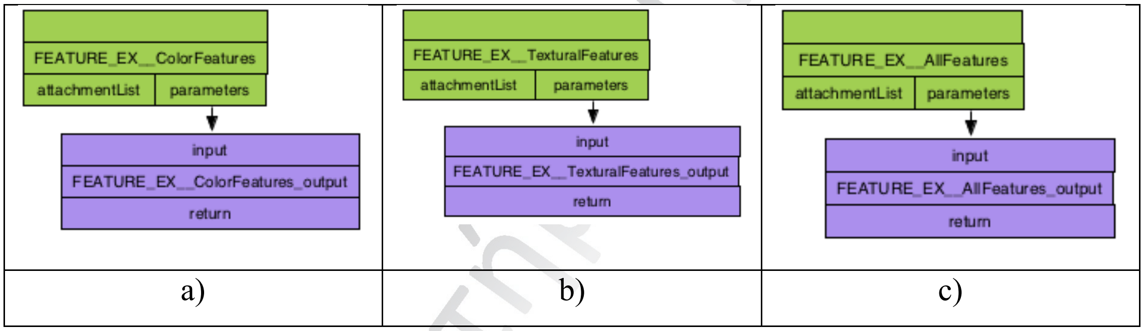

FEATURE_EX__ColorFeatures
attachmentList
parameters
input
FEATURE_EX__ColorFeatures_output
return
FEATURE_EX__TexturalFeatures
attachmentList
parameters
input
FEATURE_EX__TexturalFeatures_output
return
FEATURE_EX__AllFeatures
attachmentList
parameters
input
FEATURE_EX__AllFeatures_output
return
a)
b)
c)

Figure 63. The available feature extraction options of the core Feature Extraction Operator

Color Features and Textural Features Operators

In order to achieve rapid characterization of the salient objects of the corresponding examined image, a set of features is extracted. These features are calculated for each detected salient object (from the corresponding binary image, containing their masks). There are two feature categories, the color and the textural features.

3.4.4.1.1 Color Features

Mean Red value: denotes the mean value of the red channel for the pixels of the corresponding mask.

Mean Green value: denotes the mean value of the green channel for the pixels of the corresponding mask.

Mean Blue value: denotes the mean value of the blue channel for the pixels of the corresponding mask.

3.4.4.1.2 Textural Features

Area: Area is the number of pixels of each object. The average value provides an estimation of the mean size of the detected salient object.

Mean (gray) value: This feature is the mean value of the average gray of each annotated object. In RGB images, pixels are represented by three values of the corresponding colors, each of them ranges from 0 to 255. Thus, the following formula is used, in order to convert each pixel to gray scale:

$$\mathrm{gray}_{\mathrm{value}} = 0.299 * \mathrm{red}_{\mathrm{value}} + 0.587 * \mathrm{green}_{\mathrm{value}} + 0.114 * \mathrm{blue}_{\mathrm{value}} \quad (1)$$

Standard Deviation: This is the standard deviation of the gray values of the pixels of a specific object-mask. This feature indicates the diversity of each type of salient objects.

Modal gray value: This feature indicates the most frequently occurring pixel value of each object-mask.

Minimum and maximum gray value: This feature exports the minimum and maximum gray values of the pixels of each detected object. The motivation of this feature is to examine the gray value ranges of each salient object's type.

Perimeter: This feature, along with the Area feature, provides an estimation of what is the size and the length of the boundaries of each salient object.

Ellipse: This feature is based to the theory of trying to fit an ellipse to the cell area.

Major and minor: These two values indicate the length (in pixels) of the major and minor axis of the ellipse surrounding each cell.

Angle: This is the angle between the major axis of the ellipse and X-axis of the image.

Circularity: Circularity is a metric that indicates how similar to a circle is an object. The formula that calculates the circularity of an object is the following:

$$Circularity = 4\pi \times \frac{Area}{Perimeter^2} \quad (2)$$

Ellipse Aspect Ratio: The aspect ratio of an ellipse is the ratio of the length of its major axis over the length of its minor axis.

Integrated Density: This feature indicates the sum of the pixels of a mask. This was used as a metric of how bright each salient object's type is.

Median: Median indicates the median gray value of the pixels of each mask.

Skewness: Skewness indicates how the gray values of the cell's pixels differ in shape from a fully symmetrical distribution. It also denotes to which direction (positive or negative), deviates from the symmetrical distribution. If skewness equals 0, then the distribution of the values of the specific object, draw symmetrical distribution and mean, mode and median features are equal. If the frequency of the brighter gray value is greater than the darker ones, that indicates a negatively skewed distribution. Thus, the opposite indicate a positively skewed distribution.

Kurtosis: Kurtosis measures how rough is the peak of the above described distribution, which indicates a measure on how the pixels of each mask are distributed in the acceptable value range, versus the center of the distribution.

Area Fraction: This feature indicates the percentage of each salient object's type. This feature provides an estimation of which type of treatment the image belongs.

Roundness: Similar to Circularity metric with the difference that this metric is based on the fitting ellipse features. The formula that calculates the roundness of a detected object is the following:

$$Roundness = 4 \times \frac{Area}{\pi \times MajorAxis^2} \quad (3)$$

Solidity: This feature indicates the ratio of the area of the object-mask over the convex area. This characterizes the geometry of each type of cells. This feature - combined with the Circularity and Roundness feature - it can characterize how sharp the edges of a cell type are, despite the fact they are in circular forms.

3.4.5 Utility Operators

Utility operators have been developed, in order to increase the proposed framework's easiness of use and offer advanced workflow features, so as to deal with complex biomedical mining tasks. The utility operators are independent and they do not require integration to core type operators, in order to apply their algorithms onto the workflow's images. Description of the developed utility type operators is given to the following sub-chapters.

Separate Classes Operator

There are cases that the segmentation algorithm (i.e. the breast cancer biopsy segmentation of the cancer and apoptotic cells) produces multiclass results. This fact requires the existence of an operator, capable to distinct the different classes and separate them into different images, in order to make them available for processing by operators that require binary type of images (i.e. Ground Truth Checking, Annotation, Feature Extraction, etc.).

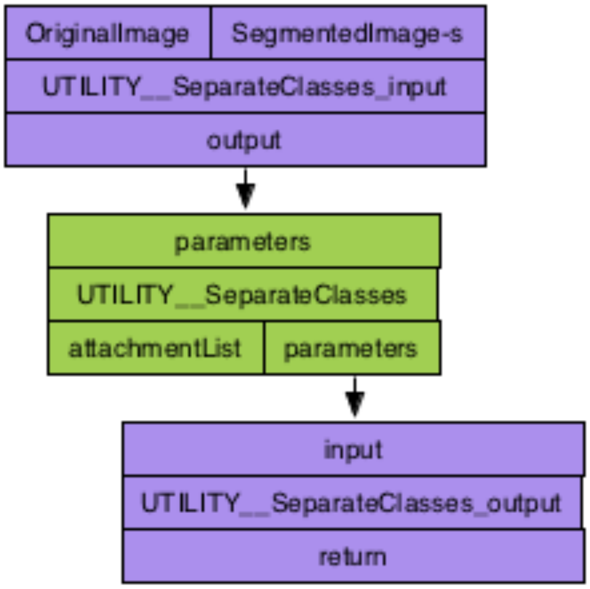
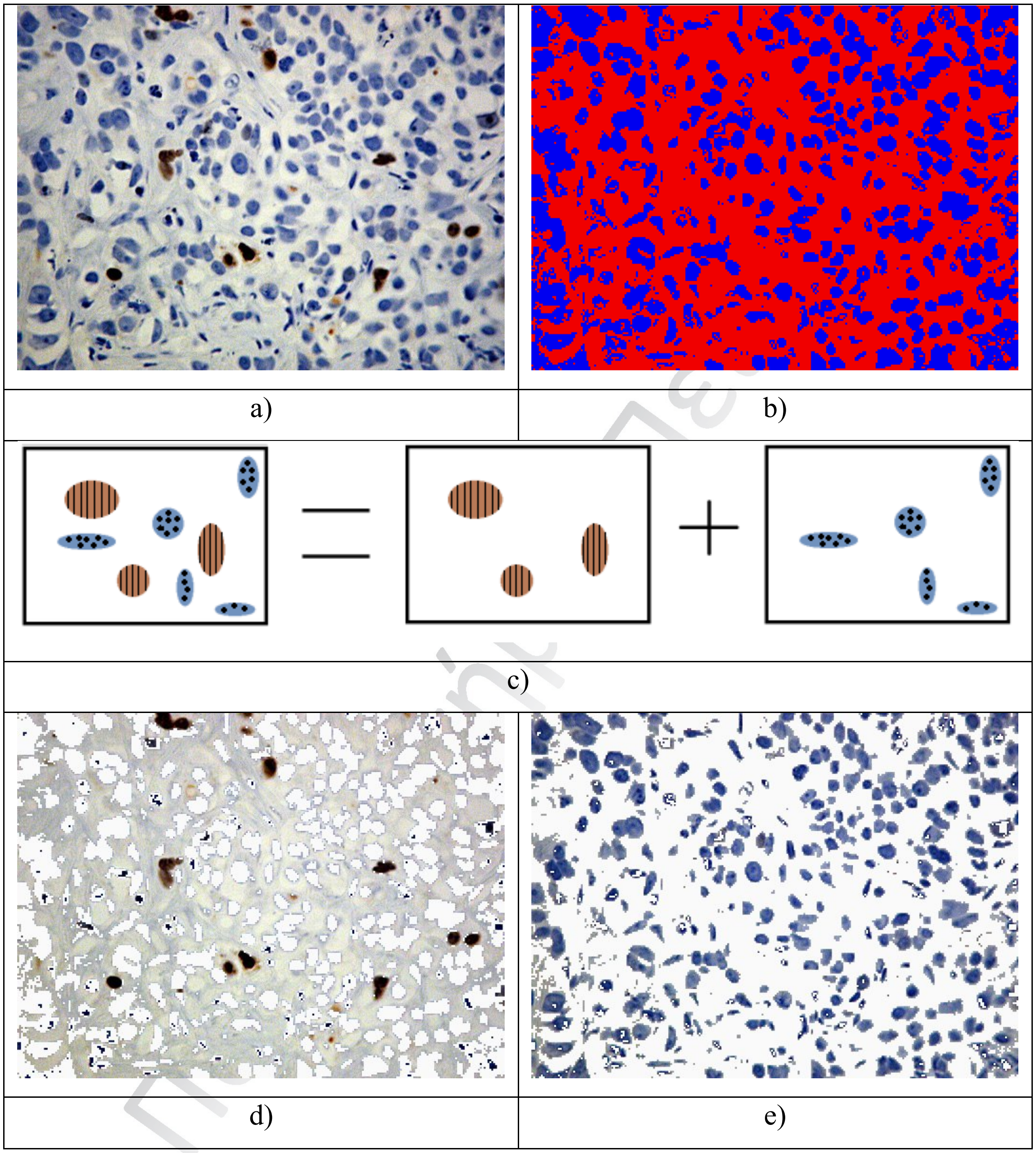


Figure 64. Separate Classes Operator, along with its available settings

The Separate Classes operator receives as input the segmented and the corresponding original image-s, in order to take the informative part from each image and generate additional images, each of them containing objects only from each respective class. Specifically, it detects the number of different colors (see Figure 65) on an image and generates the corresponding number of new images. Following this procedure, each image contains the original pixels from the initial image that correspond to the colored pixel areas of the specific class. Thus, the generation of the new images is accomplished.



a)

b)

c)

d)

e)

Figure 65. a) Original Image b) Segmented Image c) Separate Class Operator Functionality d) Generated image containing only cancer cells (red class) e) Generated image containing only apoptotic cells (blue class).

Annotation Operator

In salient object detection task, the annotation of the objects on the original image is required. The annotations on the original image provides a user friendly feature and makes the segmentation results understandable even to people with average image analysis knowledge. Thus, the image annotation operator (see Figure 66) was developed as an independent utility operator.

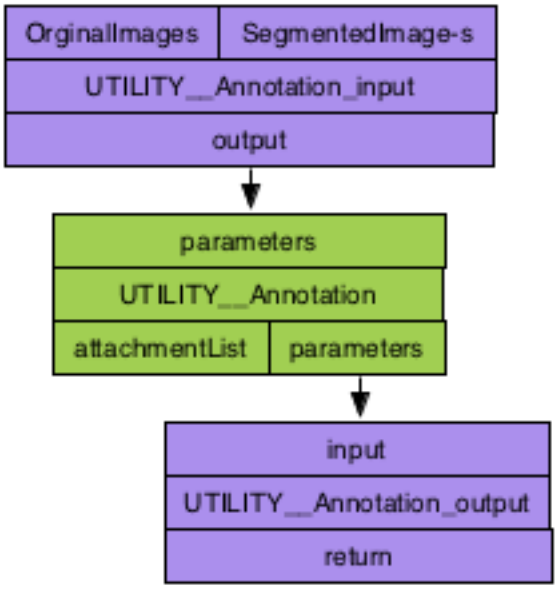


Figure 66. Annotation Operator

The operator receives the segmented and the corresponding original image-s as inputs. The output of the specific operator provides the original images with the annotation marks on the salient objects, based on the class they belong to (see Figure 67).

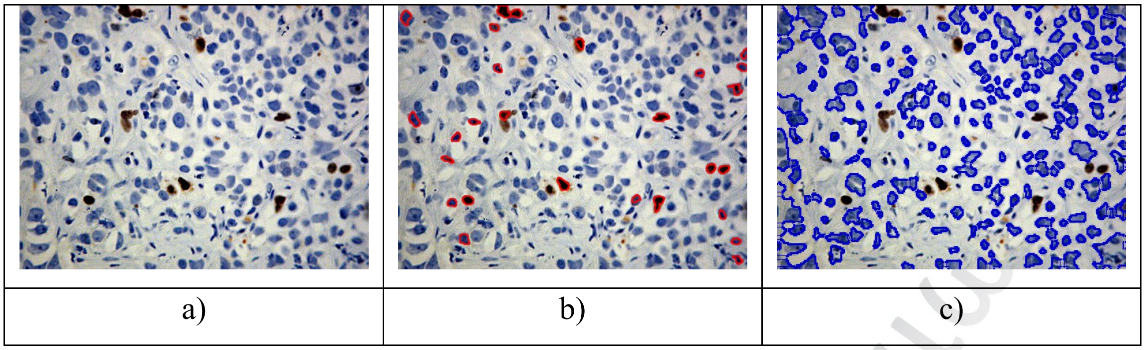


Figure 67. a) Original Image b) Cancer cells annotations b) Apoptotic cells annotations

Add Images Operator

Add Images operator was developed, in order to combine two or more different workflow image lists into one. Specifically, in cases where there are more than one workflow schemes into one construction object, it combines their exported images into one.

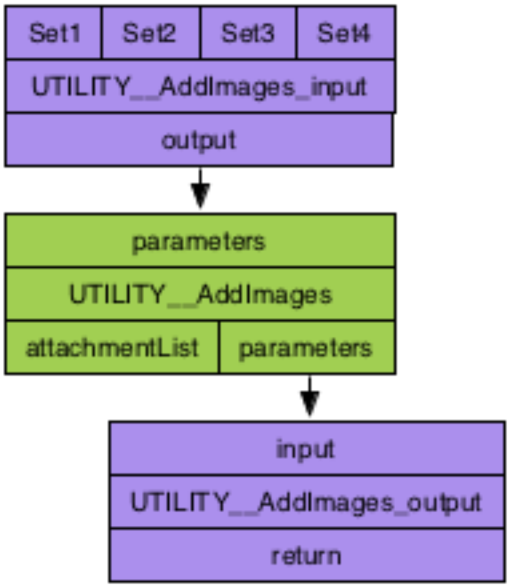


Figure 68. Add Images Operator

3.4.6 Converter Operators

The most commonly used image-processing operator in image processing cases is the converter from an image bit format to another. Thus, the necessary converters for the image-s of the workflow are required. Since we would like to provide a versatile open framework for the image mining, a variety of converters is provided. Specifically, the following converters have been developed:

ConvertToGray: From any type of image format the images in the flow are converted to a gray scale version of the image.

ConvertToIndexed: There are some cases where specific number of colors must exist on a specific image. More specifically, it converts the image-s of the workflow into 8-bit indexed color by utilizing Heckbert's median-cut color quantization algorithm [40]. For example, see Figure 89 (b), where the initial image is limited to a 3-color range.

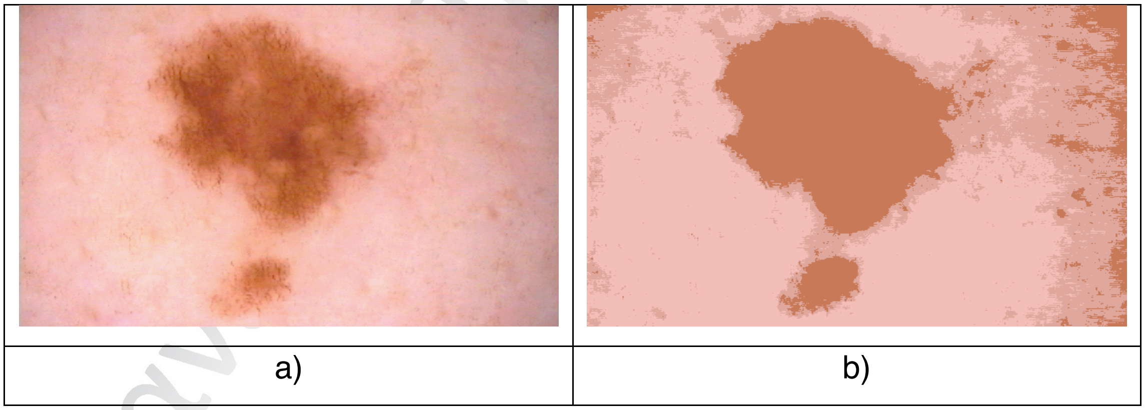


Figure 69. a) The original Image b) The image after the application of median cut algorithm with a 3-color range setting.

ConvertToRGB: From any type of image format the images in the flow are converted to a RGB version of the image. In cases where a gray scale is converted to RGB, the perspective of the image will be the same. Nevertheless, the image matrix will convert from 2D of [0-255] to 2D of ([0-255],[0-255], [0-255]).

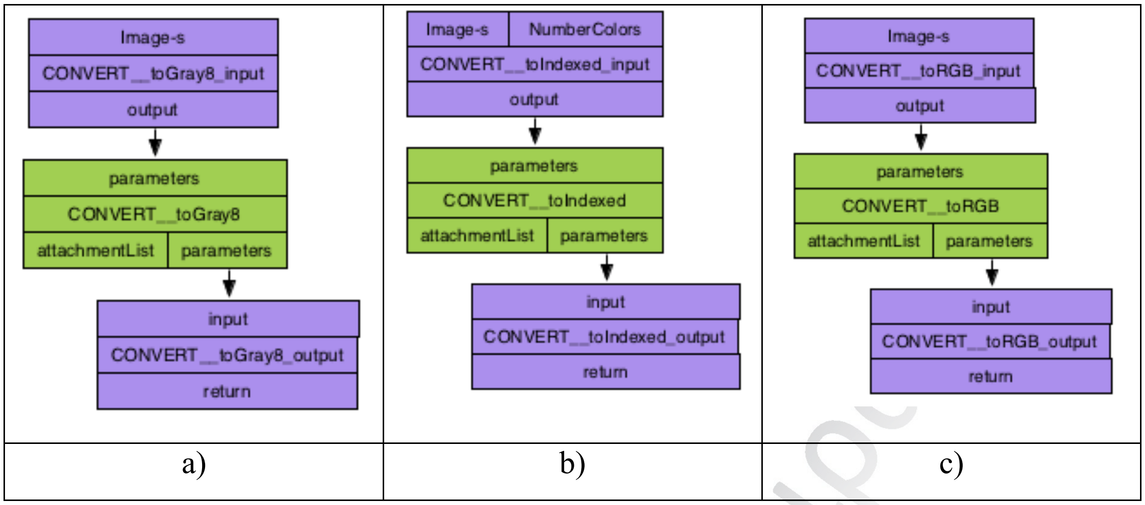


Figure 70. a) the grayscale converter b) the indexed image converter c) the RGB converter

3.4.7 Evaluator Operators

Ground Truth evaluator

In order to provide an objective evaluation of the different generated workflow instances, a metric that is not affected by the amount of difference between the TP and TN pixels is required. Thus, Matthews Correlation Coefficient was utilized, because it is not affected by the possible large amount of true negative pixels.

The specific operator (see Figure 71) requires the original image-s containing red colored areas (masks) that cover the salient objects and the corresponding segmented images. Afterwards, it makes a comparison of the segmented image with the provided ground truth image and calculates the above-mentioned metric.

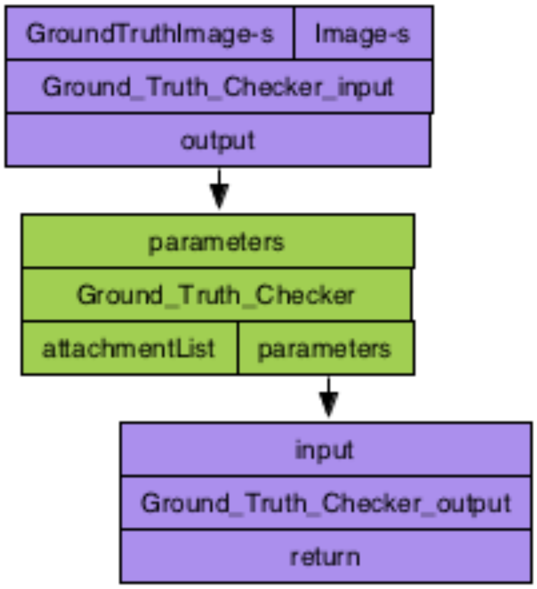


Figure 71. Ground Truth Checker Operator.

The image that succeeds the greater MCC value, is chosen as optimal and serves as input to the Scheme Proposal operator, which recognizes which operators, processed the specific image. Furthermore, it recognizes the application sequence of the utilized operators and provides the complete optimal workflow for the detection and recognition of the salient objects of the specific image dataset.

Clustering Operators

There are plenty of cases that the salient objects of the image dataset are too many to count or annotate. This fact makes the creation of the ground truth dataset impossible. In these cases, clustering of the detected salient object is achieved through a selected feature extraction process.

The user may select one or more clustering algorithms (see Figure 72), in order to perform the clustering of the salient objects of each image separately. The produced clusters are evaluated via the log likelihood distance of the clusters metric that was discussed in previous sections of this chapter. The longer the distance between two classes, the greater the log likelihood metric becomes. Scheme proposal operator (see

Figure 93) will choose the optimal version of each image, based on the log-likelihood distance metric, and will detect the operators processed the specific image. More specifically, the evaluation operator will choose as optimal workflow instance the one that will achieve the greatest log-likelihood distance between the clusters formed by the detected objects.

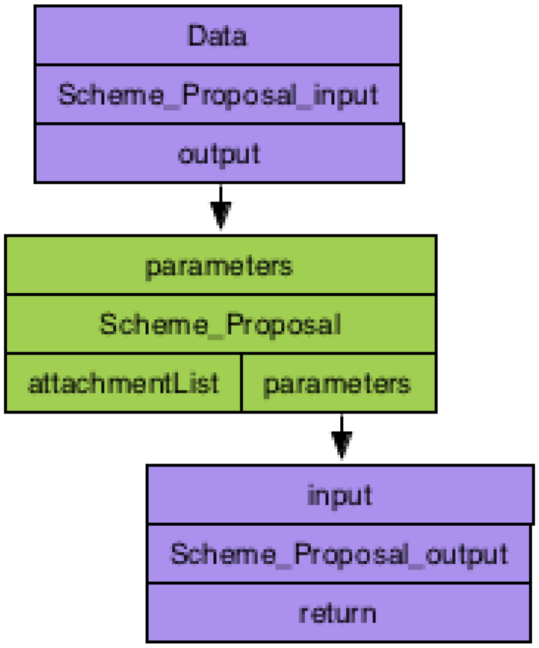


Figure 72. Scheme Proposal Operator.

In cases that more than one classifier were utilized for the clustering task, the average value of the log-likelihood distance for each clusterer is calculated. Thus, the workflow that scores the greater average log-likelihood, is selected as the optimal one.

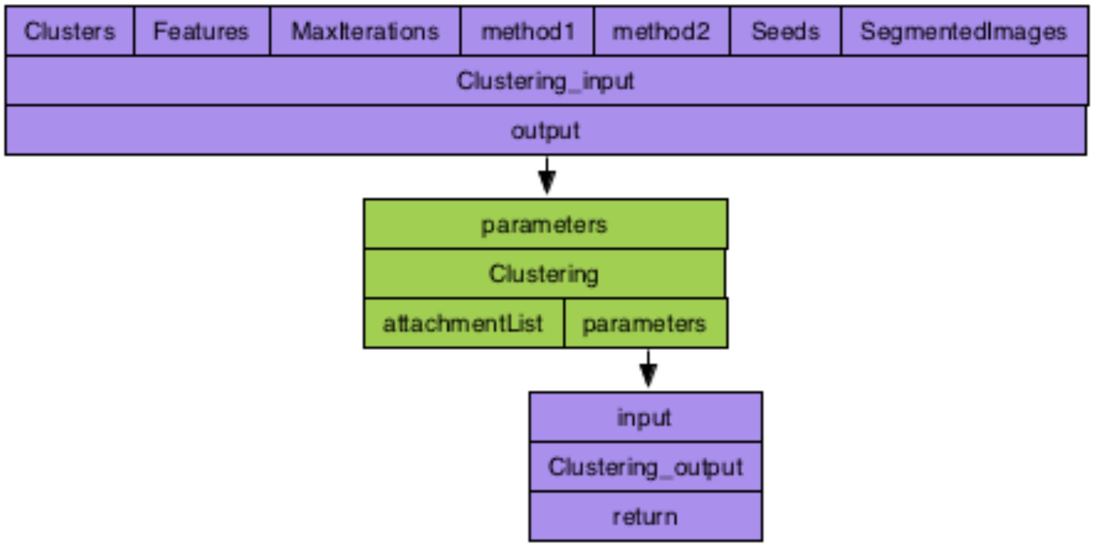
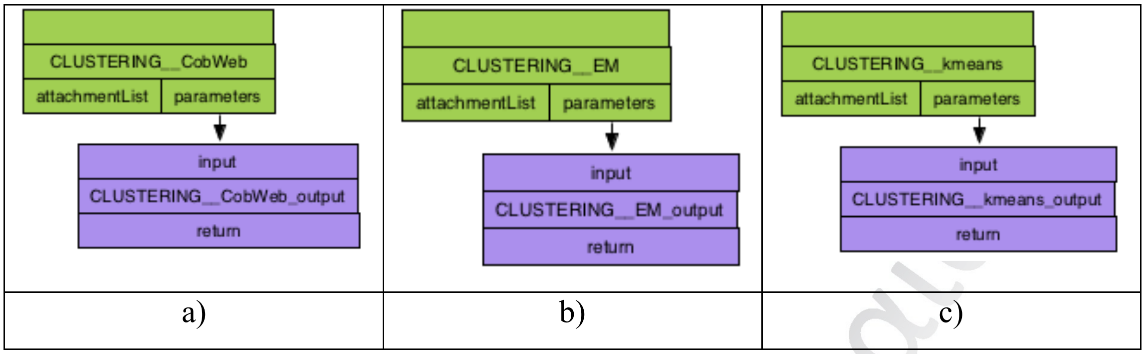


Figure 73. *The Clustering Operator, which is responsible for the log-likelihood distance estimation of each clustering task.*

The Clustering Operator receives as input the number of clusters, the feature extraction operator's output, the number of maximum eligible iterations, the number of seeds, the segmented image-s of the flow and the number of integrated clustering methods for this operator task.

In the current version of the framework three developed clustering algorithms were developed (see Figure 94). The K-Means [41], the CobWeb [42] and the Expectation Maximization [43] clustering algorithms can be integrated to the core Feature Extraction operator and generate parallel workflow instances, each of them providing the corresponding results.



a) b) c)

Figure 74. a) CobWeb b) Expectation Maximization and c) KMeans clustering operators

3.5 References

[1] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R. Pocock, Anil Wipat and Peter Li, “Taverna: a tool for the composition and enactment of bioinformatics workflows”, Bioinformatics, vol. 20, no. 17, pp. 3045-3054, June 2004.

[2] Miersw Ingo, Wurst Michae, Klinkenberg Ralf, Scholz Martin and Euler Timm. “YALE: Rapid Prototyping for Complex Data Mining Tasks”, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006.

[3] Jörg-Uwe Kietz, Floarea Serban and Abraham Bernstein. eProPlan: A Tool to Model Automatic Generation of Data Mining Workflows In: Pavel Brazdil, Abraham Bernstein, Jörg-Uwe Kietz (eds.): Proc. of the 3rd Planning to Learn Workshop (WS9) at ECAI 2010.

[4] Jörg-Uwe Kietz, Floarea Serban, Abraham Bernstein and Simon Fischer. Data Mining Workflow Templates for Intelligent Discovery Assistance and Auto-Experimentation. In: Proc. of the ECML/PKDD-10 Workshop on Third Generation Data Mining: Towards Service-Oriented Knowledge Discovery (SoKD-10).

[5] Shitu Luo, Qi Zhang, Feilu Luo, Yanling Wang, Zhiyong Chen. An improved moment-preserving auto threshold image segmentation algorithm. International Conference Proceedings. Information Acquisition, 2004. Pp 316 – 318.

[6] Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000, 16, 412–424.

[7] Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya. Comparison the various clustering algorithms of WEKA Tools. International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 5, May 2012 pp. 73-80.

[8] Newcomer Eric, Lomow Greg, Understanding SOA with Web Services. Addison Wesley. ISBN 0-321-18086-0. Publication Date: 2004

[9] Todica V., Vaida M.F., “SOA-based medical image processing platform”, in Proc. of IEEE International Conference on Automation, Quality and Testing, Robotics, 2008, May 2008, vol. 1, pp. 398-403.

[10] The OpenSSL Project, information available online at: http://www.openssl.org/.

[11] Gruschka, N., Lo Iacono, L. Server-Side Streaming Processing of Secured MTOM Attachments. Web Services (ECOWS), 2010 IEEE 8th European Conference on 1-3 Dec. 2010, pp 11-18

[12] Josefsson S. The Base16, Base32, and Base64 Data Encodings. IETF. October 2006. RFC 4648. Retrieved March 18, 2010.

[13] Jain, R., Kasturi, R. and Schunck, B.G., Machine Vision, McGraw-Hill, 1995.

[14] Tukey, J.W., Exploratory Data Analysis, Addison-Wesley, Reading MA, 1971.

[15] Kramer, H. and Bruchner, J., “Iteration of a non-linear transformation for the enhancement of digital images”, Pattern Recognition, Vol. 7, pp. 53-58, 1975.

[16] Shapiro, L. G. & Stockman, G. C: "Computer Vision", page 137, 150. Prentice Hall, 2001

[17] Rafael C. Gonzalez and Richard E. Woods, Digital Image Processing, 3rd Edition, Prentice Hall, 2008.

[18] Ritter, G.X. and Wilson, J.N., Handbook of Computer Vision Algorithms in Image Algebra, CRC Press, 1996.

[19] Canny, J., *A Computational Approach To Edge Detection*, IEEE Trans. Pattern Analysis and Machine Intelligence, 8(6):679–698, 1986.

[20] Fernand Meyer. Un algorithme optimal pour la ligne de partage des eaux. Dans *8^{me} congrès de reconnaissance des formes et intelligence artificielle*, Vol. 2 (1991), pages 847–857, Lyon, France.

[21] Stanley Sternberg, "Biomedical Image Processing", IEEE Computer, Volume 16, Issue 1 January 1983, pp 22-34.

[22] Balazs Harangi, Rashid Jalal Qureshi, Adrienne Csutak, Tünde Petö, András Hajdu. Automatic detection of the optic disc using majority voting in a collection of optic disc detectors. In Proceedings of ISBI'2010. pp.1329-1332

[23] Suzuki K, Horiba I, Sugie N (2003) Linear-time connected-component labeling based on sequential local operations. Computer Vision and Image Understanding, vol. 89 Issue 1, pp.1-23, January 2003.

[24] Goudas T, Maglogiannis I (2012) Cancer cells detection and pathology quantification utilizing image analysis techniques. Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE pp. 4418-4421.

[25] Otsu, N., "A Threshold selection method from gray-level histograms", IEEE Tran. on System Man and Cybernetics, SMC-9 (1), pp. 62-69, 1979.

[26] Ping-Sung Liao and Tse-Sheng Chen and Pau-Choo Chung (2001). "A Fast Algorithm for Multilevel Thresholding". *J. Inf. Sci. Eng.* **17** (5): 713–727.

[27] Fukunaga, K. ; Hostetler, L. "The estimation of the gradient of a density function, with applications in pattern recognition", Information Theory, IEEE Transactions on (Volume:21 , Issue: 1) January 1975, pp. 32-40.

[28] Comaniciu, D. and Meer, P. Mean shift analysis and applications. Computer Vision. The Proceedings of the Seventh IEEE International Conference on (Volume:2), 1999. Pp 1197-1203.

[29] Li CH, Tam PKS. (1998) An Iterative Algorithm for Minimum Cross Entropy Thresholding, Pattern Recognition Letters, 18(8): 771-776.

[30] Huang L-K, Wang M-J J. (1995) Image Thresholding by Minimizing the Measures of Fuzziness, Pattern Recognition, 28(1): 41-51.

[31] Alkhazaleh, S., Salleh, A.R. and Hassan, N. Soft Multisets Theory. Applied Mathematical Sciences, v. 5, No. 72, 2011, pp. 3561–3573

[32] Zack, G. W., Rogers, W. E. and Latt, S. A., 1977, Automatic Measurement of Sister Chromatid Exchange Frequency, Journal of Histochemistry and Cytochemistry 25 (7), pp. 741-753.

[33] Yen JC, Chang FJ, Chang S. (1995) A New Criterion for Automatic Multilevel Thresholding, IEEE Trans. on Image Processing, 4(3): 370-378

[34] Cortes C, Vapnik V. (1995) Support-Vector Networks, Machine Learning, 20, pp.273-297.

[35] Friedman N, Geiger D, Moises et al. (1997) Bayesian Network Classifiers, Machine Learning, pp. 131-163.

[36] Roussopoulos N, Kelley S, Vincent F. (1995) Nearest Neighbor Queries, SIGMOD '95 Proceedings of the 1995 ACM SIGMOD international conference on Management of data ISBN:0-89791-731-6, p71-79.

[37] T. Mitchell, "Decision Tree Learning", in T. Mitchell, Machine Learning, The McGraw-Hill Companies, Inc., 1997, pp. 52-78

[38] Kohavi R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence P.1137—1143

[39] Balazs Harangi, Rashid Jalal Qureshi, Adrienne Csutak, Tünde Petö, András Hajdu. Automatic detection of the optic disc using majority voting in a collection of optic disc detectors. In Proceedings of ISBI'2010. pp.1329-1332

[40] Kruger, Anton. "Median-cut color quantization." Dr Dobb's Journal-Software Tools for the Professional Programmer 19.10 (1994): 46-55.

[41] Hartigan, J. A.; Wong, M. A. (1979). "Algorithm AS 136: A K-Means Clustering Algorithm". *Journal of the Royal Statistical Society, Series C* **28** (1): 100–108.

[42] Fisher, Douglas (1987). "Knowledge acquisition via incremental conceptual clustering". *Machine Learning* **2** (2): 139–172.

[43] Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society, Series B* **39** (1): 1–38.

[44] Wegener, D., May, M. Specification of Distributed Data Mining Workflows With DataMiningGrid (2009) Data Mining Techniques in Grid Computing Environments, pp. 165-178.

CHAPTER 4
Building Image Mining Workflows With the Proposed Framework

This Chapter presents the proposed methodologies developed for several biomedical image-mining issues, in the context of the Ph.D. thesis. The approaches to these case studies are considered novel, compared to the current state-of-the-art literature of the field. Each of the proposed solutions utilizes the proposed framework's features. The proposed framework will assist the creation of efficient smart workflows. These workflows will automate the complex image mining procedures and provide fast, accurate and reproducible results to the expert physicians/pathologist or to computer science engineers.

4.1 Installation and Utilization of the proposed Framework

The following instructions require the installation of Taverna Workflow Manager version 2.4 or above. The installation of Taverna itself is based on an auto-installer, so no further instructions are required for its installation. Additionally, for the utilization of our framework, a broadband connection is required.

Following the prerequisites stage, the user can start Taverna application. The application itself needs a couple of minutes to load all the necessary initial web services that are pre-installed to the Manager. When the platform is ready for use the graphic, user interface of the application appears (see Figure 75).

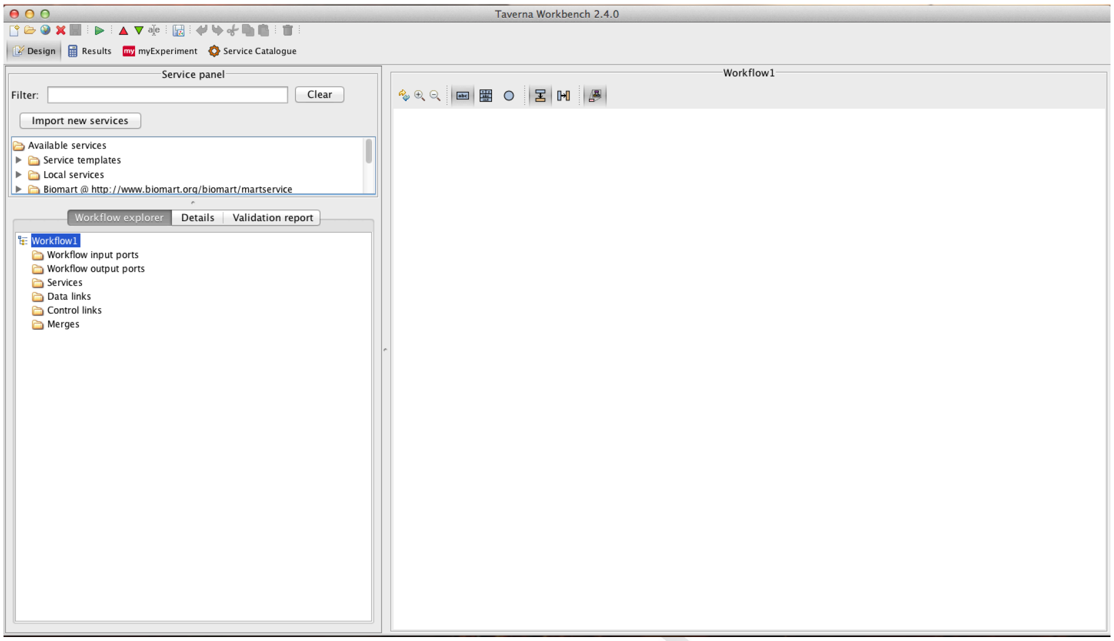


Figure 75. The Graphic User Interface of the Taverna Application.

In order to enhance Taverna with the functionalities of the proposed framework, we need to import our developed services through the button Import new services (see Figure 76). Afterwards, the user must select the WSDL service option. The pop-up window of Figure 77 will open, asking for a valid WSDL service location. The current WSDL location of the proposed framework is at:

http://83.212.100.219:8080/teoservice/imageserver?wsdl

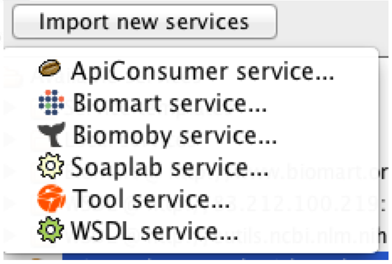


Figure 76. Import new services menu.

Following the integration of the WSDL service of our proposed framework, a folder that contains all the available, enriches the "Available services" menu (see Figure 78).

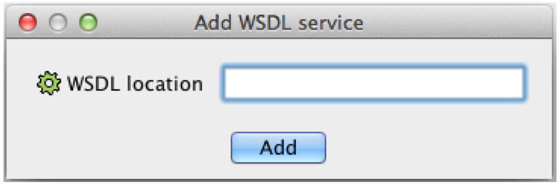
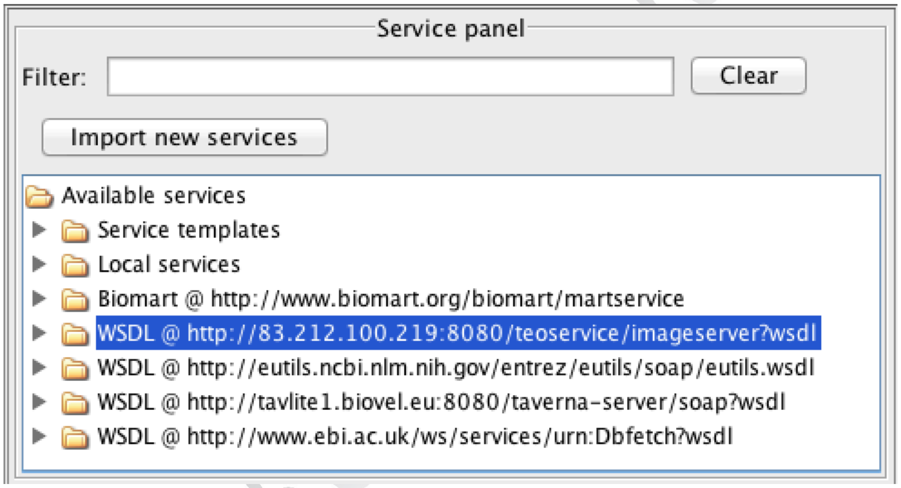


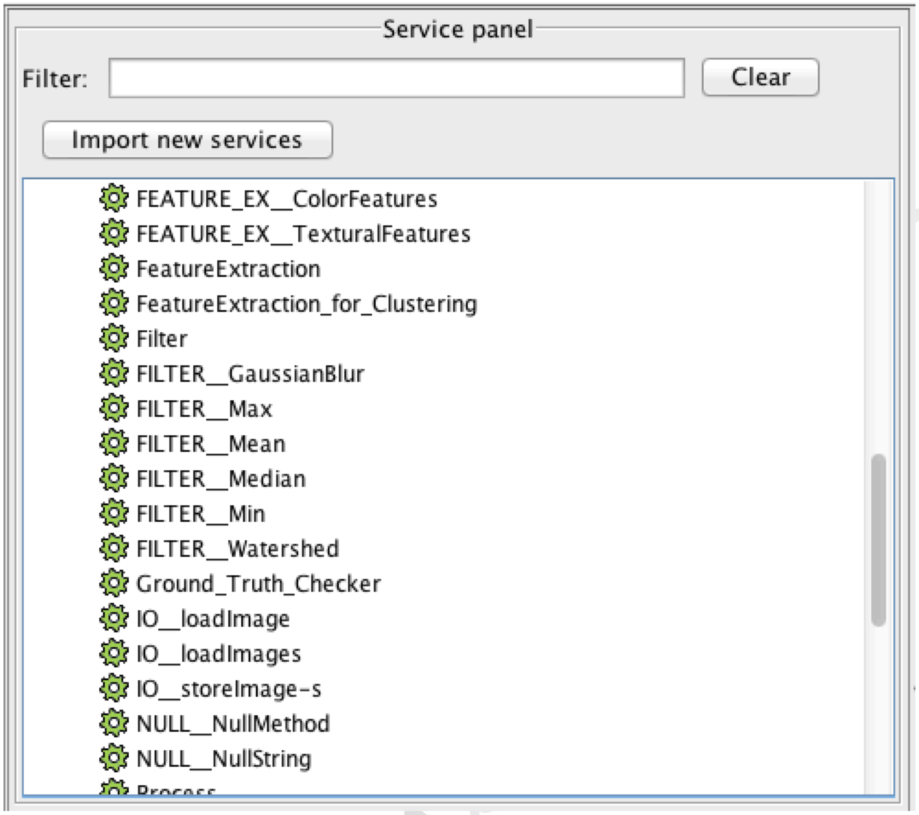
Figure 77. Setting the location of the WSDL service.



- Available services
  - Service templates
  - Local services
  - Biomart @ http://www.biomart.org/biomart/martservice
  - WSDL @ http://83.212.100.219:8080/teoservice/imageserver?wsdl
  - WSDL @ http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/eutils.wsdl
  - WSDL @ http://tavlite1.biovel.eu:8080/taverna-server/soap?wsdl
  - WSDL @ http://www.ebi.ac.uk/ws/services/urn:Dbfetch?wsdl

Figure 78. Service Panel Menu.

We can notice that the new folder that was integrated into service panel menu contains all the developed functionalities of our proposed framework (see Figure 79). The added folder along with the corresponding operators is permanently installed to the Taverna application.



Service panel

Filter: Clear

Import new services

- FEATURE_EX__ColorFeatures
- FEATURE_EX__TexturalFeatures
- FeatureExtraction
- FeatureExtraction_for_Clustering
- Filter
- FILTER__GaussianBlur
- FILTER__Max
- FILTER__Mean
- FILTER__Median
- FILTER__Min
- FILTER__Watershed
- Ground_Truth_Checker
- IO__loadImage
- IO__loadImages
- IO__storeImage-s
- NULL__NullMethod
- NULL__NullString
- Process

Figure 79. Integrated Operators of the Proposed Framework.

The construction of a simple flow is a quite easy procedure. The minimum flow requires at least one input and one output (see Figure 80). Let's suppose the simplest scenario, in which we would like to convert an RGB image to a gray scale one. The first thing we must do, is to import the images to the system through the loadImages operator (see Figure 81). Thus, we drag and drop the corresponding operator into the Workflow Area.

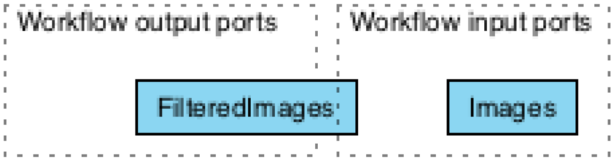


Figure 80. Taverna's Input and Output ports.

The initial form of the operator can be depicted in Figure xx. Each integrated operator requires the exportation of the XML input and output splitter, which is responsible for the interoperability of the data. In order to perform this exportation, we simply right click on the operator and select the corresponding options (see Figure 82).

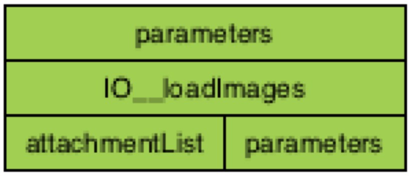


Figure 81. loadImages operator.

A last, but not least prerequisite is to transform the data from their binary format to an XML coded format. For this reason, we will utilize the base64 encoding [38], which is available in an operator form through Taverna's available services (see Figure 83).

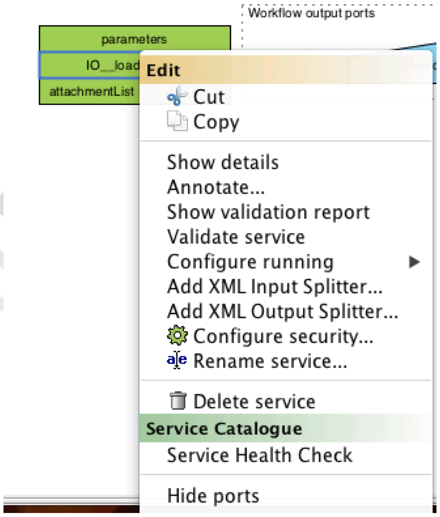


Figure 82. Add XML input Splitter and XML output Splitter options.

Each operator was developed, in order to interact with the Taverna application (messages, debugging, etc.). Thus, it can be noticed that on XML input splitters the name of the expected data is mentioned. In case of wrong data input, the corresponding operator sends a message to the Validation Report Section of the application, denoting the exact problem and a proposed solution (see Figure 84).

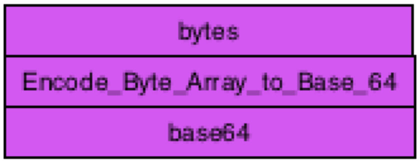
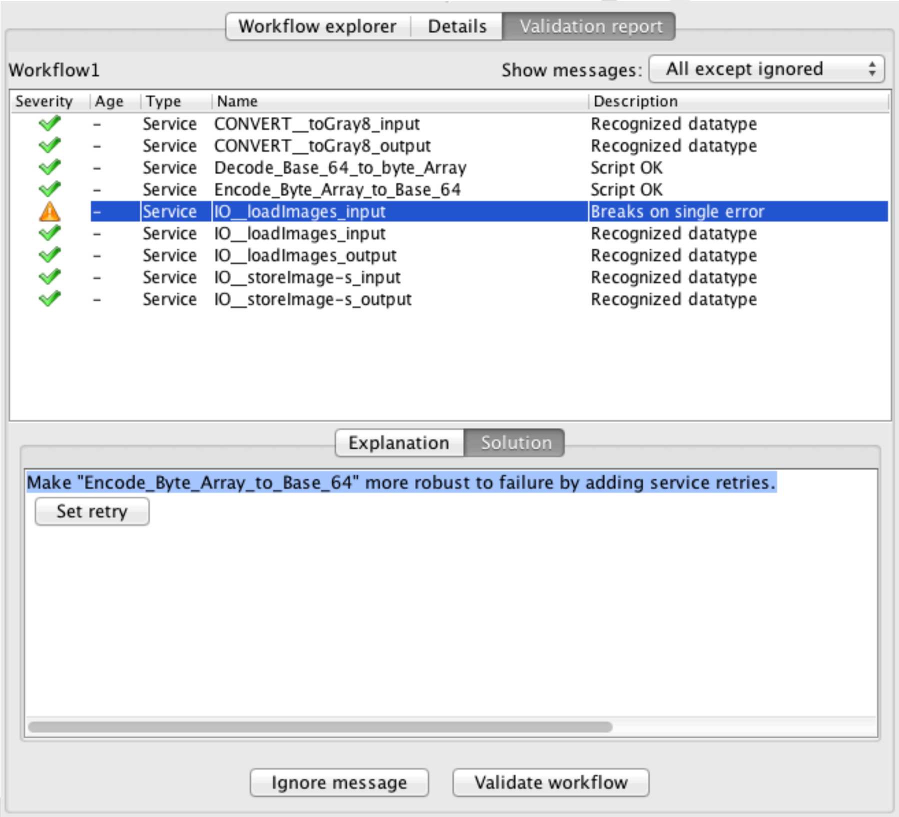


Figure 83. base64 Encoder.

Each operator's XML output splitter, usually is connected with the input splitter of the next operator. So as to get the processed by the constructed workflow data, the storeImages operator is required.



Workflow1

Show messages: All except ignored

Severity	Age	Type	Name	Description
✓	–	Service	CONVERT__toGray8_input	Recognized datatype
✓	–	Service	CONVERT__toGray8_output	Recognized datatype
✓	–	Service	Decode_Base_64_to_byte_Array	Script OK
✓	–	Service	Encode_Byte_Array_to_Base_64	Script OK
⚠	–	Service	IO__loadImages_input	Breaks on single error
✓	–	Service	IO__loadImages_input	Recognized datatype
✓	–	Service	IO__loadImages_output	Recognized datatype
✓	–	Service	IO__storeImage-s_input	Recognized datatype
✓	–	Service	IO__storeImage-s_output	Recognized datatype

Make "Encode_Byte_Array_to_Base_64" more robust to failure by adding service retries.

Figure 84. Validation report window of the Taverna application.

Following the construction of the customized workflow scheme, the decoded of the data is required, in order to get their binary form back. Thus, the base64 decoding operator is required. The summary workflow scheme is depicted in Figure 85.

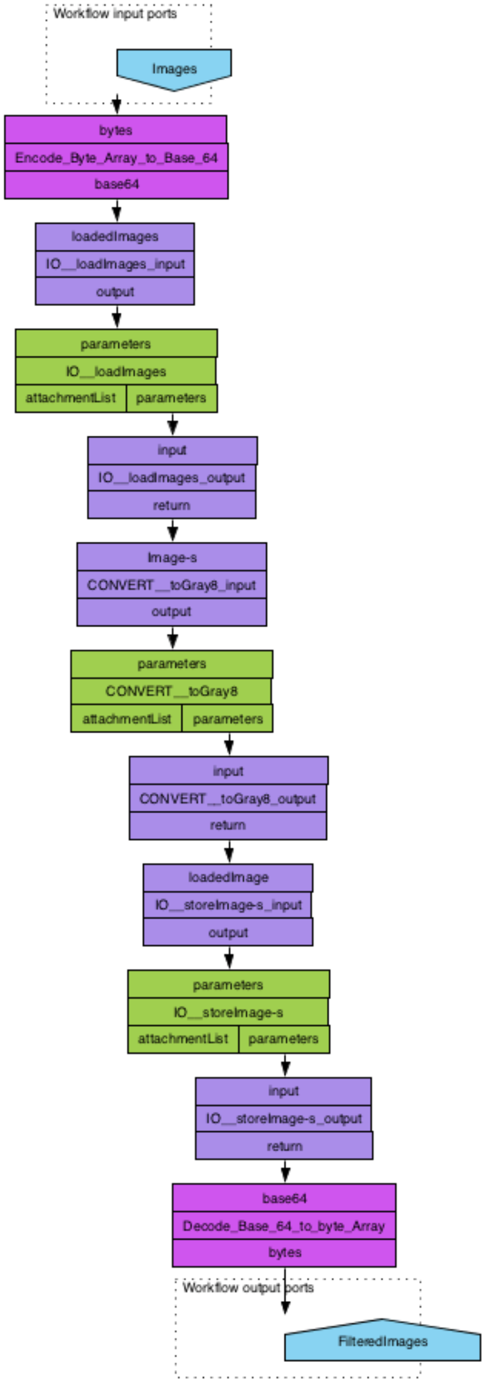
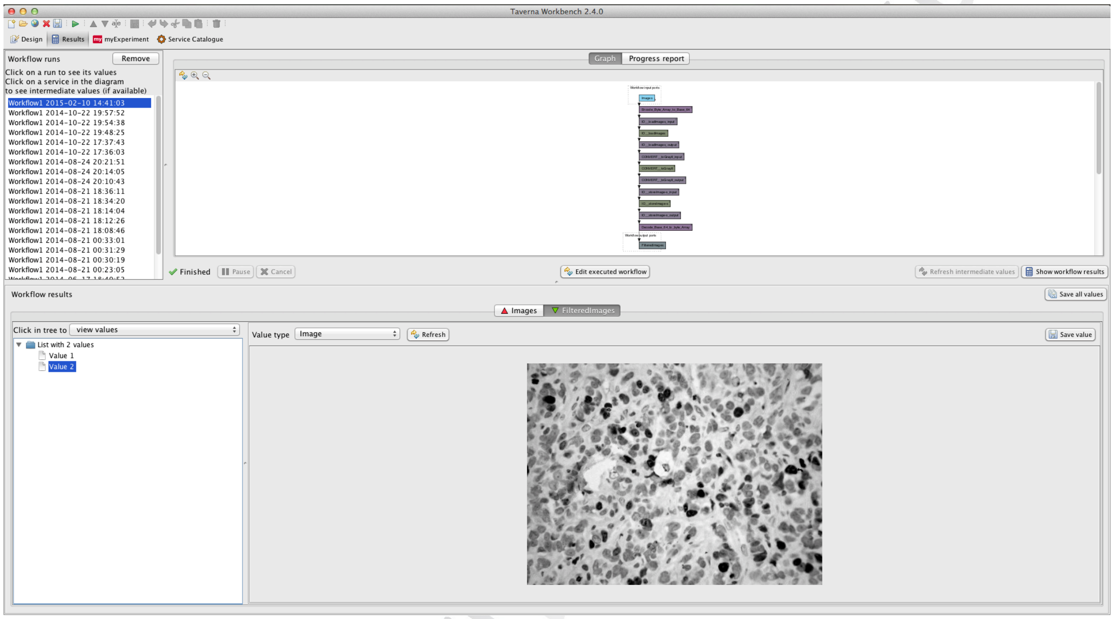


Figure 85. Constructed workflow scheme.

Concluding the construction of the flow, we are able to run the corresponding scheme. Following the execution of the scheme, the results are depicted in the Workflow Results panel (see Figure 86). Additionally, there is an option to save a specific, or every data inputted or exported by the specific flow, through the Save value or Save all values button (see Figure 86).



Click on a run to see its values
Click on a service in the diagram
to see intermediate values (if available)

Workflow1 2015-02-10 14:41:03
Workflow1 2014-10-22 19:57:52
Workflow1 2014-10-22 19:54:38
Workflow1 2014-10-22 19:48:25
Workflow1 2014-10-22 17:37:43
Workflow1 2014-10-22 17:36:03
Workflow1 2014-08-24 20:21:51
Workflow1 2014-08-24 20:14:05
Workflow1 2014-08-24 20:10:43
Workflow1 2014-08-21 18:36:11
Workflow1 2014-08-21 18:34:20
Workflow1 2014-08-21 18:14:04
Workflow1 2014-08-21 18:12:26
Workflow1 2014-08-21 18:08:46
Workflow1 2014-08-21 00:33:01
Workflow1 2014-08-21 00:31:29
Workflow1 2014-08-21 00:30:19
Workflow1 2014-08-21 00:23:05

List with 2 values
Value 1
Value 2

Figure 86. Results window of the Taverna application.

4.2 Detection of Obstructive Nephropathy on different set of Kidney Biopsy Microscopy Images

The kidney is a multicellular, heterogeneous, multi-structural organ that is responsible for a part of the complex process of blood's purification. As blood flows through the kidneys, waste materials, chemicals and unneeded water are removed out of the body as urine. Kidneys are being affected by many chronic diseases often leading to a slow deterioration of this organ, which implies to an inappropriate cleaning of the blood. Obstructive Nephropathy [1] is the main cause of renal failure, which occurs in all ages, but is often met in children and infants. It is caused by obstruction of the urinary tract, with hydronephrosis (which is dilation of the renal pelvis and calyses resulting from obstruction to flow of urine), slowing of the glomerular filtration rate and tubular abnormalities.

Considering that obstructive nephropathy is not a rare disease [2], automated detection of the pathogenic areas on a kidney biopsy image is very useful for the proper assessment of the disease. In this context, we have developed a computer-based application, which is able to recognize salient objects (i.e. see in Figure 87 non-Pathogenic Glomerulus, Pathogenic Glomerulus, non-Pathogenic Tubulus and Pathogenic Tubulus) among other regions, aiming at the quantification of the depicted in the image disease. In this section, we present the details of the implemented application along with an initial evaluation.

4.2.1 The Dataset

The utilized image dataset has been obtained from healthy and pathogenic kidney biopsies. In the context of our study, 60 images were utilized, 30 control and 30 pathological. Samples are stained with the Sirius Red technique, which is one of the best-understood techniques of collagen histo-chemistry. In bright-field microscopy collagen are red on a pale yellow, while nuclei are ideally black but may often be grey or brown.

In the examined kidney images, the pathological findings are connected with alterations in the imaging of the 2 major salient objects: Tubulus and Glomerulus.

A Nikon Eclipse E400 microscope was used with a Nikon lens Plan Fluor 20x / 0.50; DIC M; ∝ / 0.17 WD 2.1 combined with a Microfire by Optronics camera with the following settings, in order to capture the kidney biopsy images. Settings: exposure were as following: 10ms; Red: 105; Green: 100; Blue: 100; Gain: 1; Luminosity: 50; Contrast: 60.

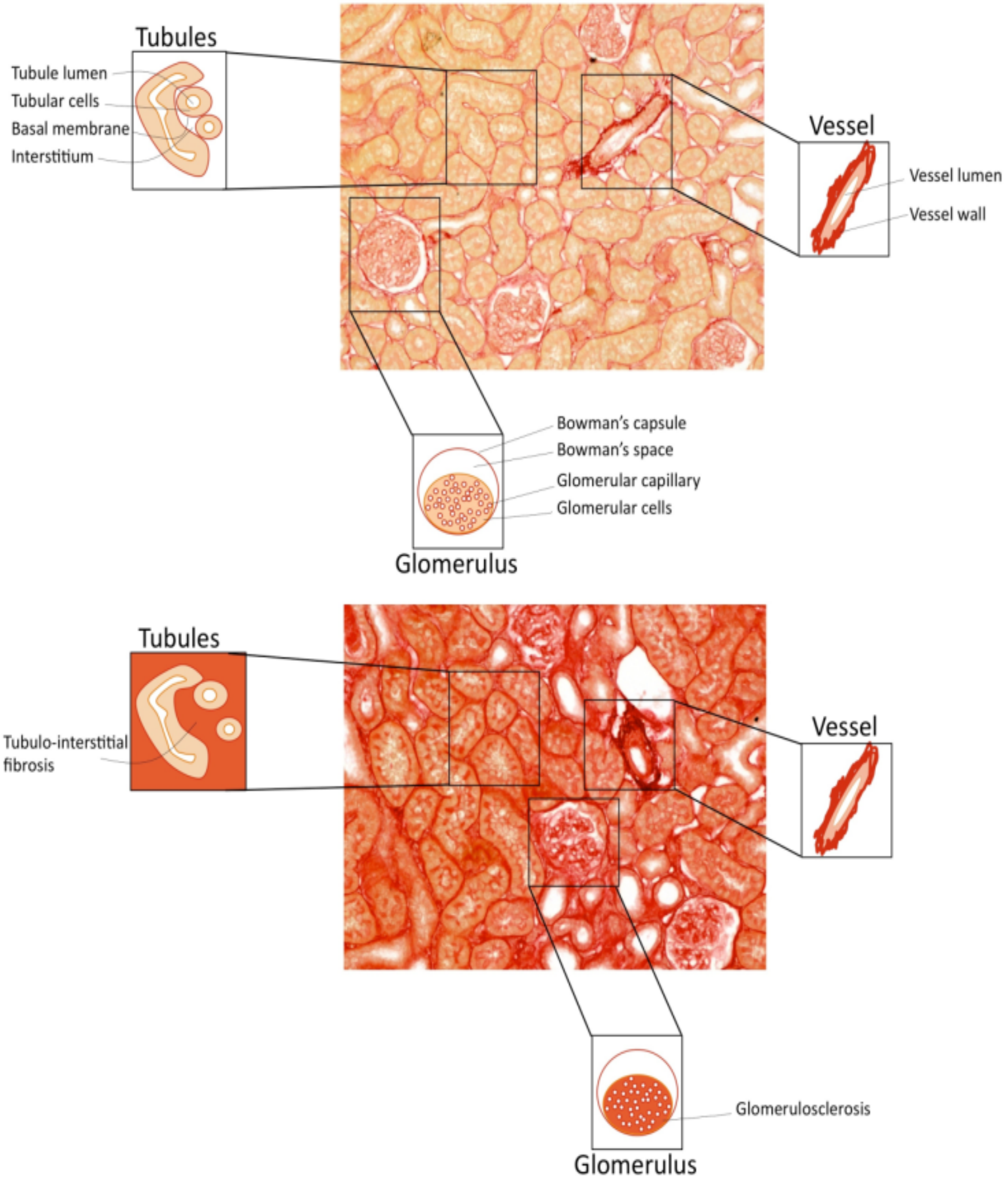


Figure 87. Regions of Interest and their characterization on a kidney biopsy image.

4.2.2 The proposed approach

An overview of the designed and implemented image processing pipeline is depicted in Figure 88. The image is divided in blocks (squares) and each block is allocated through a classification procedure to a specific salient object or background. Then, a morphological operator based on majority voting is utilized, in order to remove erroneously classified segments of the image.

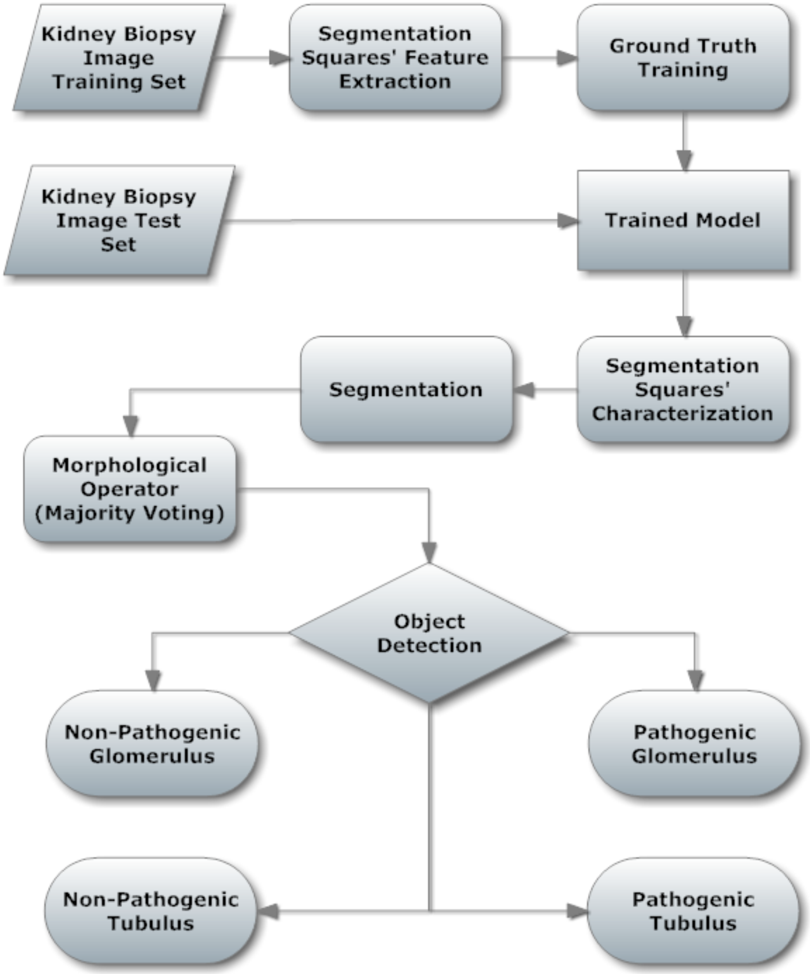

Kidney Biopsy Image Training Set
Segmentation Squares' Feature Extraction
Ground Truth Training
Kidney Biopsy Image Test Set
Trained Model
Segmentation
Segmentation Squares' Characterization
Morphological Operator (Majority Voting)
Object Detection
Non-Pathogenic Glomerulus
Pathogenic Glomerulus
Non-Pathogenic Tubulus
Pathogenic Tubulus

Figure 88. Flow Diagram of the proposed application for the recognition of salient objects in Kidney Biopsy Images

Salient Object Detection

In the specific images four (4) types/classes of salient objects are recognized. They are namely: non-Pathogenic Glomerulus, Pathogenic Glomerulus, non-Pathogenic Tubulus and Pathogenic Tubulus. Non-pathogenic and pathogenic Glomerulus has a diameter ranging from 50 to 120 μm [3]. The tubules of the nephrons are 30 – 55 mm long [4] with an average diameter of 50μm.

The proposed operator (see Figure 89) was developed, in order to deal with specific kidney biopsy images, which include four types/classes of salient objects. They correspond to the most important kidney structures namely Glomerulus, and Tubulus (see Figure 87).

The goal is to classify regions as pathogenic or not. Accurate and mainly reproducible diagnosis of obstructive nephropathy by an expert is a complex, non-trivial task, and since it represents a disease, which prevalence subjects is an important epidemiological threat, with severe implications (kidney transplantations) for children and infants, an analytical workflow, able to provide fast and reproducible results, like the one proposed here, is considered an invaluable diagnostic contribution.

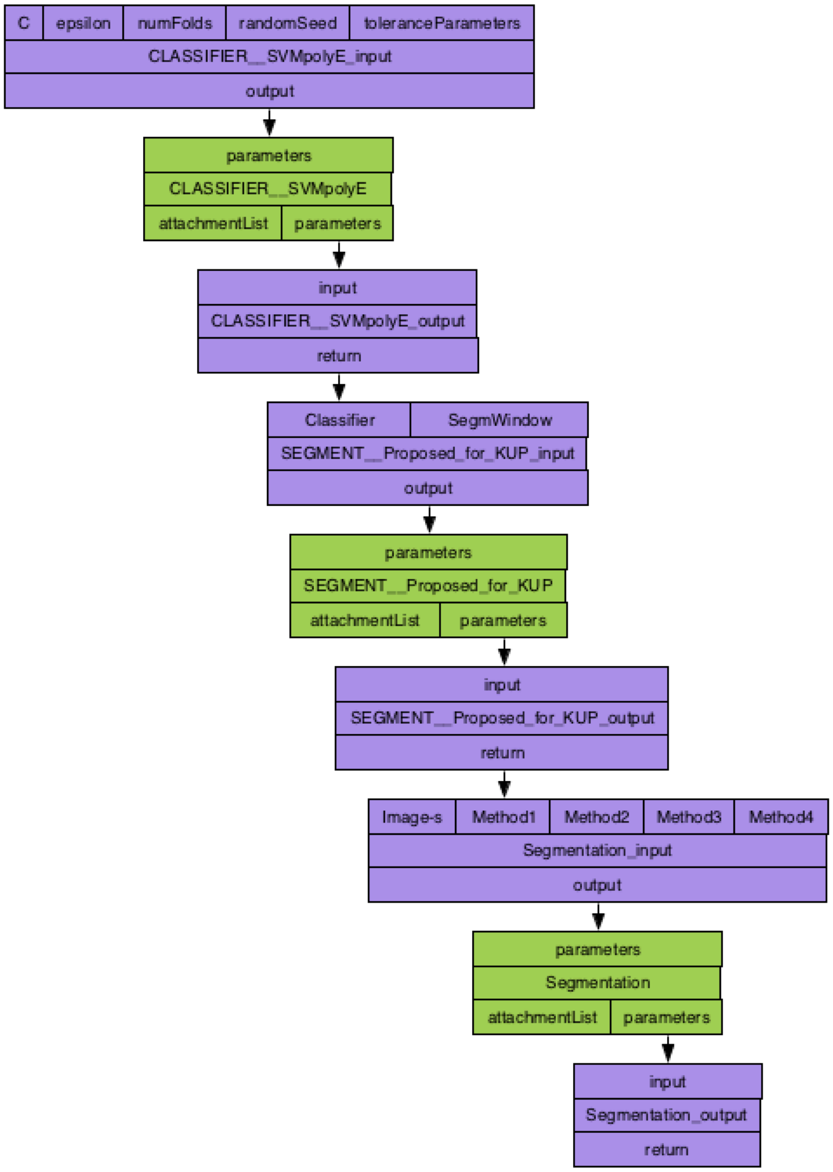


Figure 89. Developed operators for the Kidney Biopsy Images dataset.

Since the edges of the targeted regions are not clear in the biopsy images, a block based segmentation approach is adopted. The image is divided in blocks (squares). Segmentation squares are much smaller than the aforementioned objects. The size of the block is set 37x37. This value has been selected heuristically, as the most appropriate for providing satisfying accuracy and acceptable processing times, based on conducted experiments.

For the feature extraction procedure, customized code was developed, in order to separate the ROI into smaller square segments of a specific width (defined by the user and 37x37 in our case). An appropriate feature extraction workflow was planned in Taverna, using the proposed framework. Mean, Standard deviation, Contrast, Inverse Difference Moment, Correlation, Entropy and Angular Second Moment are the features, which were used as inputs for the classification of each segmentation block. The above features have been selected as the most appropriate textural features for image classification [5]. The only preprocessing step of the image data concerns gray scale conversion. Some additional preprocessing techniques were tested (i.e. histogram equalization and contrast enhancement etc.) but they were omitted at the end, since they provided lower accuracy.

Each block is allocated through the classification procedure to a specific class. Four widely known classifiers were examined, namely: Support Vector Machines (SVM) [6], Naïve Bayes [7], K-Nearest Neighbor [8] and Decision trees [9].

During the training procedure of the classification model, representative regions were selected through the above feature extraction procedure. The expert biologists, participating in this research, did the selection. This resulted in a training set of about 1200 segmentation squares (400 of each class). The ten-fold cross validation [10] has been adopted as a method for testing the accuracy. As depicted in Table 1, the results are satisfactory. The best performing classifier for the current problem is the SVM classifier achieving 94.7% accuracy, according to ground truth blocks, defined by the experts.

Table 6. Results Of The Classification Models Of Segmentation Window 37x37

Classifiers		true non-Pathogenic Glomerulus	true non-Pathogenic Tubulus	true Pathogenic Glomerulus	true Pathogenic Tubulus	class precision	Total Accuracy
SVM	pred. non-Pathogenic Glomerulus	381	1	12	12	93.842	
	pred. non-Pathogenic Tubulus	4	244	0	0	98.387	
	pred. Pathogenic Glomerulus	8	0	312	12	93.976	
	pred. Pathogenic Tubulus	4	3	14	314	93.731	
	class recall	**95.970**	**98.387**	**92.308**	**92.899**		**94.701**
KNN	pred. non-Pathogenic Glomerulus	371	7	12	16	91.379	
	pred. non-Pathogenic Tubulus	6	232	0	5	95.473	
	pred. Pathogenic Glomerulus	11	0	311	17	91.740	
	pred. Pathogenic Tubulus	9	9	15	300	90.090	
	class recall	**93.451**	**93.548**	**92.012**	**88.757**		**91.900**
Decision Trees	pred. non-Pathogenic Glomerulus	374	2	13	20	91.443	
	pred. non-Pathogenic Tubulus	4	244	0	4	96.825	
	pred. Pathogenic Glomerulus	11	0	315	11	93.472	
	pred. Pathogenic Tubulus	8	2	10	303	93.808	
	class recall	**94.207**	**98.387**	**93.195**	**89.645**		**93.565**
Naïve Bayes	pred. non-Pathogenic Glomerulus	341	0	4	12	95.518	
	pred. non-Pathogenic Tubulus	3	243	0	0	98.780	
	pred. Pathogenic Glomerulus	26	0	318	11	89.577	
	pred. Pathogenic Tubulus	27	5	16	315	86.777	
	class recall	**85.894**	**97.984**	**94.083**	**93.195**		**92.127**

The produced segmented images are quite noisy, thus a morphological filtering is required. A simple pixel majority vote [11] recursive technique with a dynamic vote limit proved sufficient for this task. Regarding the Glomerulus object, a threshold of a minimum area of 7 blocks is set, because as it noticed through experiments and observation, there is no glomerulus smaller than 9 blocks. Each step of the proposed approach is presented in Figure 90.

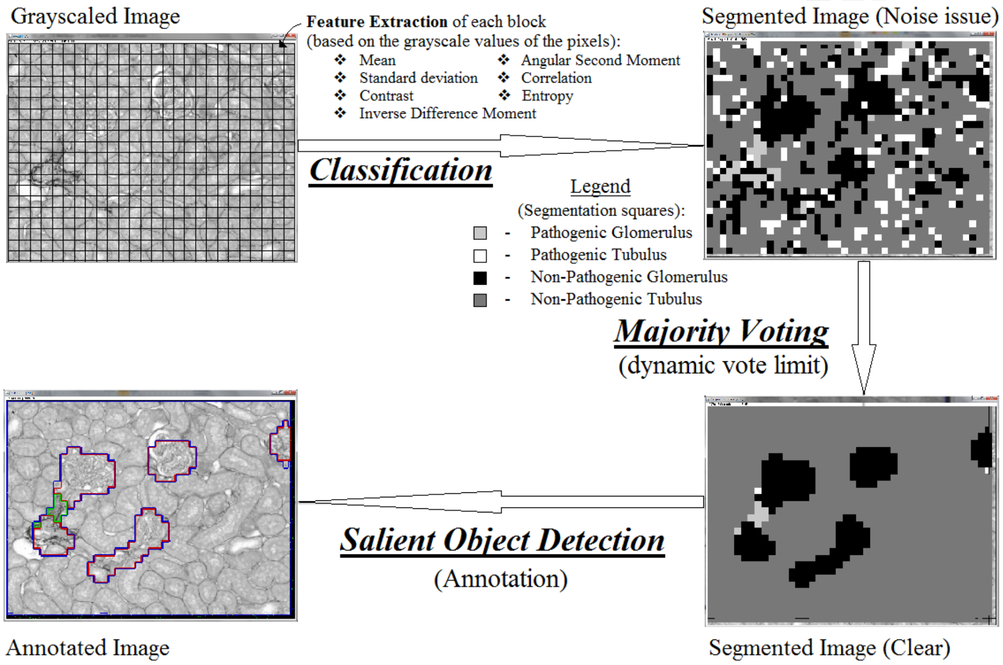


Figure 90. The steps of the proposed Obstructive nephropathy detection approach

Image Characterization

Through the classification of the blocks and the assistance of majority vote technique, salient object detection is achieved. Following the image segmentation, features of the detected total regions may be calculated. The calculation of 23 features is supported by the developed application (Mean, Standard Deviation, Correlation, Angular Second Moment, Inverse Difference Moment, Contrast, Entropy, Minimum Grayscale value, Maximum Grayscale Value, Mode, ROI's height, ROI's width, ROI's percentage, Area (in pixels), Median, Kurtosis, Skewness, Histogram's minimum value, Histogram's maximum value, Area Fraction, Centroid, Angle and Center of Mass). The above set of features provides additional information about the characterization of a salient object.

4.3 Advanced Block Detection and Quantification of Fibrotic Areas in Microscopy Images of Obstructive Nephropathy

Image characterization, is a complex procedure, since it requires several stages as data acquisition, preprocessing, segmentation, feature extraction, training and classification of the corresponding data. Proper modification of the above stages and suitable image analysis methods usage, for each specific biomedical problem, is required.

Specialist pathologists can accurately diagnose the pathogenesis of obstructive nephropathy, but they have to deal with many samples from many subjects. Therefore, is necessary the development of a fast and accurate tool for the detection and quantification of the pathogenesis.

An overview of the designed and implemented image-processing pipeline is depicted in Figure 91. The image is divided into large blocks (squares) followed by a non-informative block filtering preprocessing. Each block is allocated through a classification procedure to a specific type of pathogenesis status. Then, a two-way recognition of the pathogenesis' status is applied; one regarding to the whole informative image and the other is based on the majority of the segmentation's block recognition. The subsections below, discuss the details of the implemented workflow along with an initial evaluation.

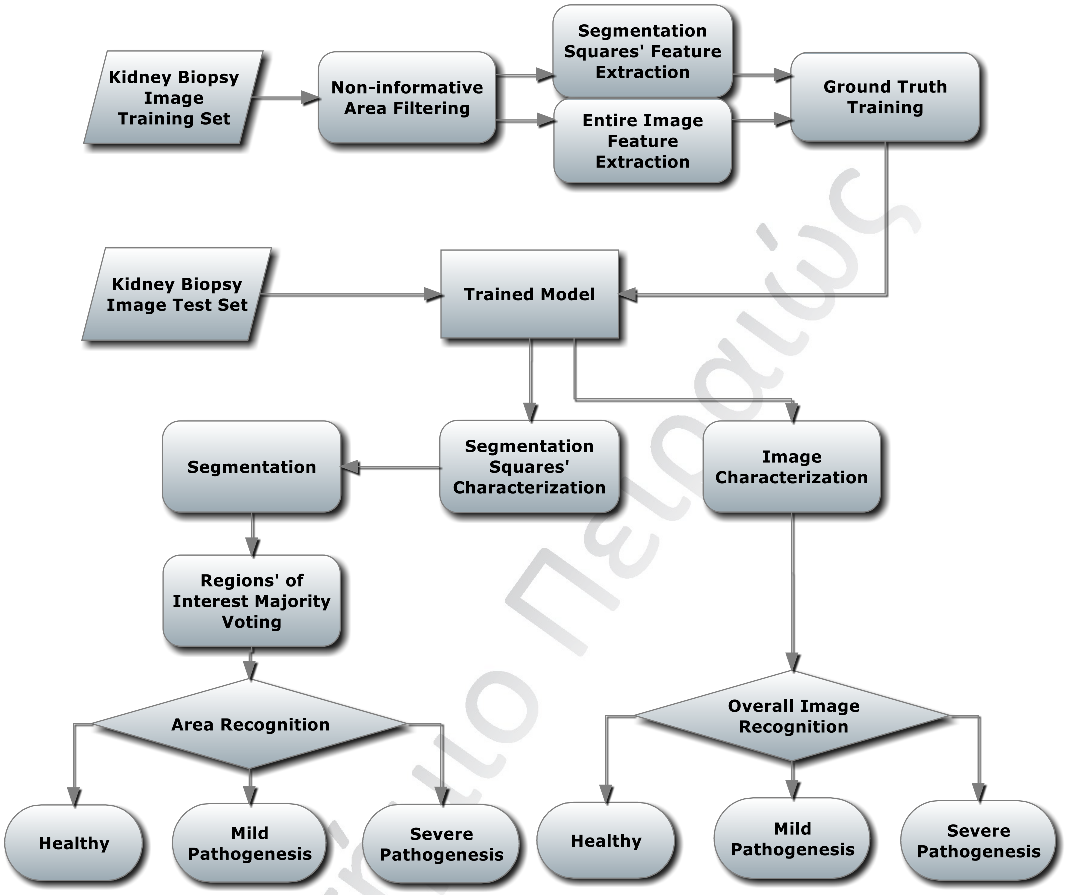

Kidney Biopsy Image Training Set
Non-informative Area Filtering
Segmentation Squares' Feature Extraction
Entire Image Feature Extraction
Ground Truth Training
Kidney Biopsy Image Test Set
Trained Model
Segmentation
Segmentation Squares' Characterization
Image Characterization
Regions' of Interest Majority Voting
Area Recognition
Overall Image Recognition
Healthy
Mild Pathogenesis
Severe Pathogenesis
Healthy
Mild Pathogenesis
Severe Pathogenesis

Figure 91. Dataflow Diagram of the proposed tool.

4.3.1 Dataset

Image dataset, which is used in this research, was acquired from mice kidney biopsies. Images were taken from healthy status of kidney into four different time periods of the obstructive nephropathy pathogenesis, and have been treated following Masson's trichrome. A magnification of 200, aperture of 0.5, 10 ms exposition and gain of 1.0 have been used as shooting settings. Thus, the dataset used for this research contains 5 images of the control status, 5 images of one day (1D), 10 images of three days (3D), 15 images of five days (5D) and 5 images of eight days (8D) after the application of the fibrotic pathogenesis.

Fibrosis is visualized in brown area (see Figure 92), between the salient objects, and is contained inside the kidney's biopsy image. In case there is a high amount of fibrosis, it is considered as a major sign of the obstructive nephropathy disease. The amount of the fibrotic areas characterizes both healthy and pathogenic biopsy images and causes separation difficulties.

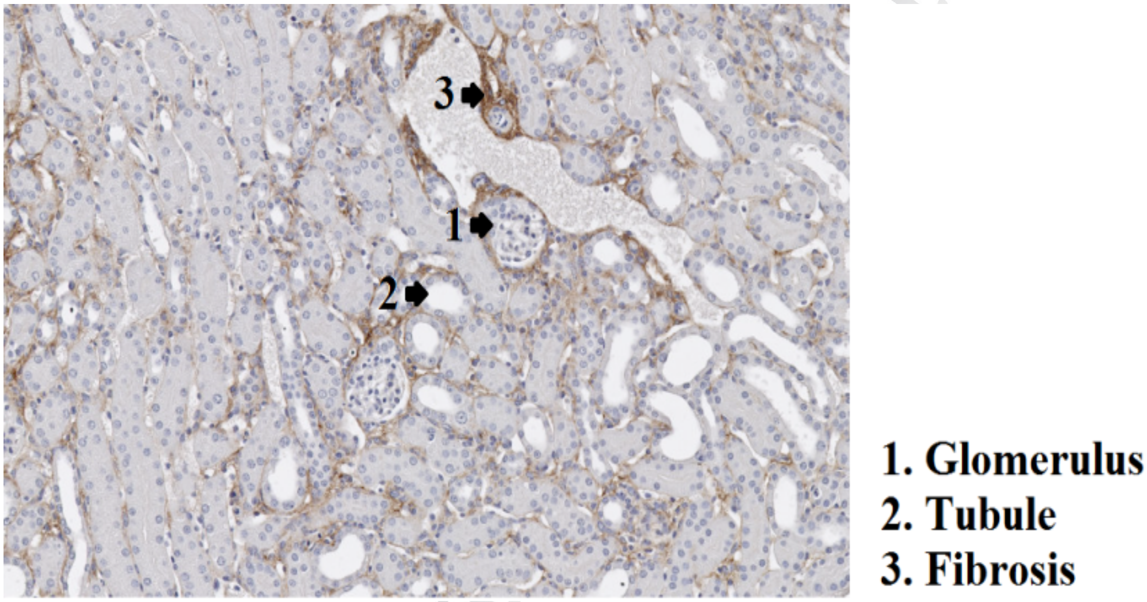


Figure 92. Salient Objects in a Kidney Biopsy

4.3.2 Pathogenesis levels clustering

Although images were taken from 5 different time periods of the disease, the problem does not seem to be a 5-class problem. That fact is justified, based on the indication of our pathologist experts, by the control and 1D images, which are not pathogenic at all. On the other hand, 5D and 8D images are both near kidney necrosis status. Hence, the problem’s classes are clustered into three categories. Control and 1D images, assigned to “Healthy” status, 3D images assigned to “Mild pathogenesis” status and 5D and 8D images assigned to “Severe pathogenesis” status. K-means algorithm [12] is used for clustering and a metric of clusters over classes (see Tables 7, 8) was adopted, in order to evaluate the applied clustering. Figure 93 indicates the deviation of each feature based on the above clustering. As it was depicted in Table 7 and Table 8, for the clustering of the segments and entire images respectively, the most clustered instances agree with the labeled instances.

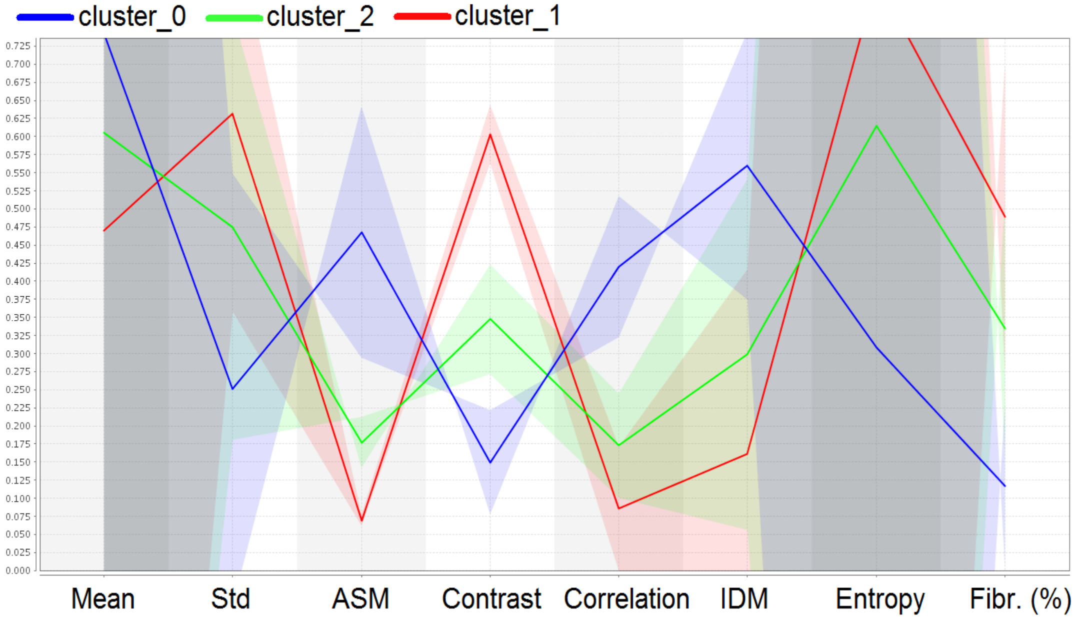


Figure 93. Mean values and Deviations of each feature on each cluster

Table 7. Clustering Accuracy of the block based segmentation approach

Clustering (based on segments)		Clusters		
		Healthy	Mild	Severe
Classes	Healthy	107	46	7
	Mild	30	85	27
	Severe	5	131	161
Accuracy (Clustered / Labeled instances) (%)		75.35	32.44	82.56
Overall Accuracy (%)		58.93		

Table 8. Clustering Accuracy of the entire image scan approach

Clustering (based on entire image)		Clusters		
		Healthy	Mild	Severe
Classes	Healthy	5	5	0
	Mild	1	3	6
	Severe	0	0	20
Accuracy (Clustered / Labeled instances) (%)		83.33	37.50	76.92
Overall Accuracy (%)		70.00		

4.3.3 Area Detection and Non-informative blocks and pixels removal

Firstly, an informative block is considered as a segmentation block, whose mean value pixels differs of white (255,255,255 RGB values correspond to white color), so the block must have values R≤235 && G≤235 && B≤235. The value 235 was selected, because the background is not exactly white). As it was mentioned above, the specific images offer 2 stages of the fibrotic pathogenesis plus the control status. Therefore, this work deals with 3-class problem. Classes included are: “Healthy”, there is no significant amount of fibrosis, “Mild pathogenic”, the fibrotic levels are a bit higher than normal and “Severe pathogenic” status of the kidney, is a high amount of fibrosis and leads kidney to its necrotic stage.

Considering the complexity of the images of the above dataset, the solution is based on a 3-stage solution. The first, utilizes feature extraction method, in order to classify the entire informative part of the image. The second stage uses feature extraction method for segments of the same informative part of the image. The third stage outputs a pathogenesis score based on the second stage and the classification of the segments. The feature extraction method is the same on both approaches and follows a detailed description below.

Since there is a difference, in percentage of fibrosis, between sections of the same image, block based segmentation was adopted for classification of each region and for non-informative areas removal as well. The image is divided in large blocks (squares). The size of each block was tested at 1000x1000, 600x600 and 200x200 pixels.

The sizes of the blocks were selected because of the high image resolution of dataset (7680x4320). The 1000x1000 size was selected as the maximum segmentation block because it covers approximately 8 to 10% of the kidney’s informative area, which is an acceptable representative of an area for a specific pathogenesis status based on the ground truth provided by our experts. The window size 200x200 was selected in order to cover the slightest details of the informative part of the kidney biopsy and provide higher resolution results. Last but not least, 600x600 size was selected as the value between 200 and 1000 in order to check the behavior of the proposed methodology.

Due to the type of the images used in this research, only the informative areas and pixels were taken under consideration. Non-informative areas are the white sections around the kidney biopsy, which have no useful information about the fibrosis issue. The non-informative pixels are the pixels which value is close to white (Red≥235 and Green≥235 and Blue≥235). Non-informative pixels are not part of the tissue and sometimes, when large areas of them exist, are the holes-tubules on it. In any case they need to be removed in order to calculate the ratio of fibrosis over the informative and valid area. The mean values of Red, Green and Blue channel of a segmentation block must be about 235 in order to be characterized as a non-informative block.

Customized code was developed, for the feature extraction procedure, in order to separate the ROI into smaller square segments of a specific width. “Mean”, “Standard deviation”, “Contrast”, “Inverse Difference Moment”, “Correlation”, “Entropy”, “Angular Second Moment” and “Fibrosis percentage” are the features, which were used as inputs for classification of each segmentation block. First seven features have been selected as the most appropriate textural features for image classification [5]. The “Fibrosis percentage” feature of each segmentation block and the whole image is a measurement of the corresponding brown pixels over the informative pixels of the block and the entire image respectively. The only preprocessing step of the image data concerns the non-informative area removal. Some additional preprocessing techniques were tested (i.e. histogram equalization and contrast enhancement) but they were omitted at the end, since they provided lower accuracy.

Each block is allocated through classification procedure to a specific class. Although several classifiers were tested, KStar [13], K-Nearest Neighbors [14], Random Forest [15] and Decision Trees [16] achieved the highest accuracy. During the training procedure representative images from each fibrotic pathogenesis status were selected and scanned by application of non-informative areas and pixels filtering. The labeling of each image and each segmentation block was accomplished by an expert pathologist. This resulted in a training set of about 250 segmentation squares (50 of each class). The fibrotic sections were selected manually and accurately from our experts, in order to retrieve the mean values of the RGB channels of these areas and allow the calculation of their percentage over the entire image and segmentation block respectively. Ten-fold cross validation [17] has been adopted as a method for testing the accuracy.

The best performing classifier for classification of segmentation blocks is Random Forest classifier, achieving 84.64% accuracy after comparison with ground truth blocks, defined by the expert. Random Forest was proved also the best classifier for the goal of entire image characterization is Random Forest classifier as well, achieving 97.5% accuracy.

4.3.4 Segments characterization and Score System for pathogenesis quantification

Following the classification of the segmentation blocks, features of each classified informative object are calculated. Except the features used in feature extraction procedure additional features are calculated as well. These are the below: “Mean”, “Standard deviation”, “Contrast”, “Inverse Difference Moment”, “Correlation”, “Entropy”, “Angular Second Moment”, “Fibrosis percentage”, “Integrated density”, “Median”, “Minimum”, “Max” and “Modal grey value”. The additional features are provided in order to texturally characterize and extract more information about each informative area of the examined kidney biopsy.

Except the segments classification, a score system is applied in order to characterize an image. More specifically, if a segment belongs to the "Healthy" class adds 0 points to the total score, if it belongs to the "Mild Pathogenesis" class it adds 1 point to the total image score and if it belongs to the "Severe Pathogenesis" class it adds 2 points to the total image score. Since the informative object's (the kidney biopsy) size differs on each image the total image score is divided by the number of the informative blocks. The result characterizes the kidney biopsy image and provides quantification of the obstructive nephropathy pathogenesis, as it is not an on-off stage disease.

4.4 Breast Cancer Cells Detection and Quantification on Breast Cancer Biopsies

Breast cancer is one of the most common cancers that can be diagnosed and be cured, however, it still remains a threat for women. For instance, this type of cancer is the prime cause of death in the United States [18]. Radiotherapy is one of the most common treatments on breast cancer cases [19]. Combined modality treatments are developed through empirical approaches, using specific drugs, in order to stop the growth of tumors and cause the cancer cell to enter the apoptotic phase with the assistance of the external beam radiation therapy. Researches on Vitamin D [20] proved effectiveness against a broad range of tumor cell types. The histological imaging domain, where images obtained by optical or electronic microscopy, is a scientific field, where image analysis and image processing techniques may apply.

4.4.1 Dataset Description

The microscopy images dataset was obtained from the National Cancer Institute tumor repository (Frederick, MD) [30]. These images were taken from six-week-old mice, which were injected the MCF-7 human breast cancer type. Subconfluent cultures, which were grown in RPMI 1640 on 37^{o}C, were fed with fresh medium, washed with PBS, trypsinized, resuspended in medium, and pooled. After centrifugation, cells were resuspended in Matrigel and cold RPMI 1640 for s.c. inoculation in mice.

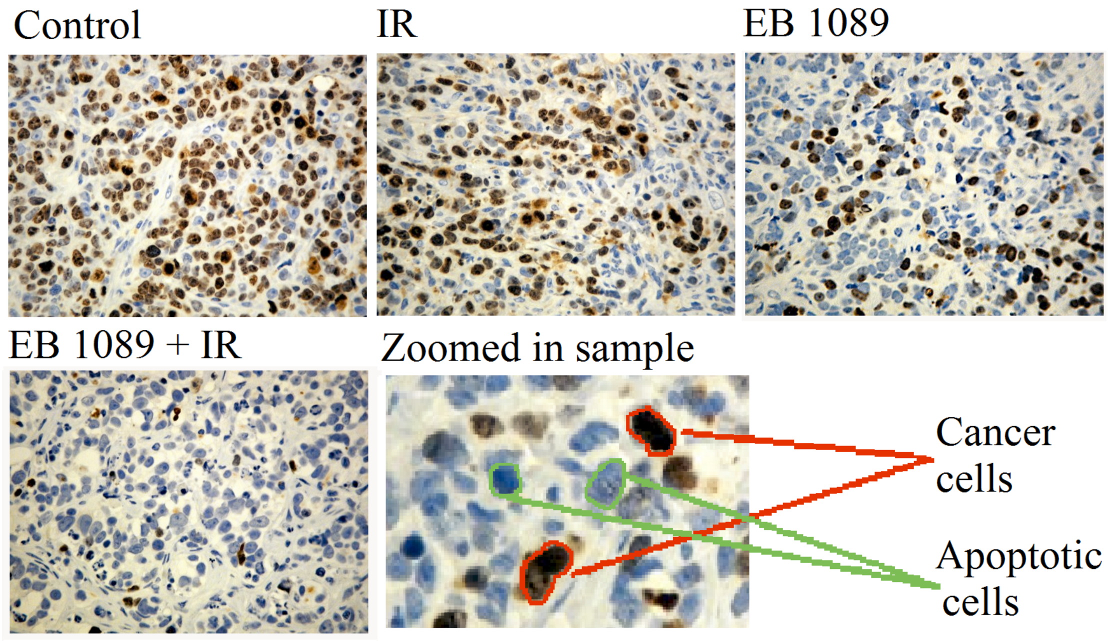


Figure 94. Four treatment groups of cancer cell images and indication of Cancer and Apoptotic Cells

When the tumor volumes reached 150–200 mm^3, tumor-infected mice were randomly selected to receive 3 different treatments. The first group's mice received EB 1089 alone (45 pmol/24 h for 8 days), the second group received radiation alone and the third EB 1089 followed by radiation. So, the image dataset provided four groups of datasets, Control, IR (Radiation), EB 1089 (Drug) and EB 1089 + IR combined (see Figure 94). As also depicted in Figure 94, cancer and apoptotic cells have a circular form [21] and sometimes are merged, making harder their quantification.

4.4.2 Block Based Segmentation Approach

In this work, an advanced image analysis tool for the fast detection and quantification of the cancer and the apoptotic cells in the microscopy image is presented. To the best of our knowledge, there is no other tool than can achieve the recognition and quantification of more than one cell type in one step. The proposed tool provides physicians the feature of repeating this complex procedure for many samples in a short time. The proposed tool was used for the evaluation of the influence of the vitamin D3 analogue EB1089 [22] with fractionated radiation on growth and apoptosis human breast cancer tumour MCF -7 [23] cells injected in six-week-old mice.

Proposed Approach

The proposed approach for the characterization of the cells and the quantification of breast cancer is presented in this section. This method is based on block based segmentation and image analysis techniques. More specifically, the separation of the classes of the cells is achieved through the use of Support Vector Machines classifier and some morphological techniques for noise removal and size correction as illustrated in the workflow diagram of Figure 95.

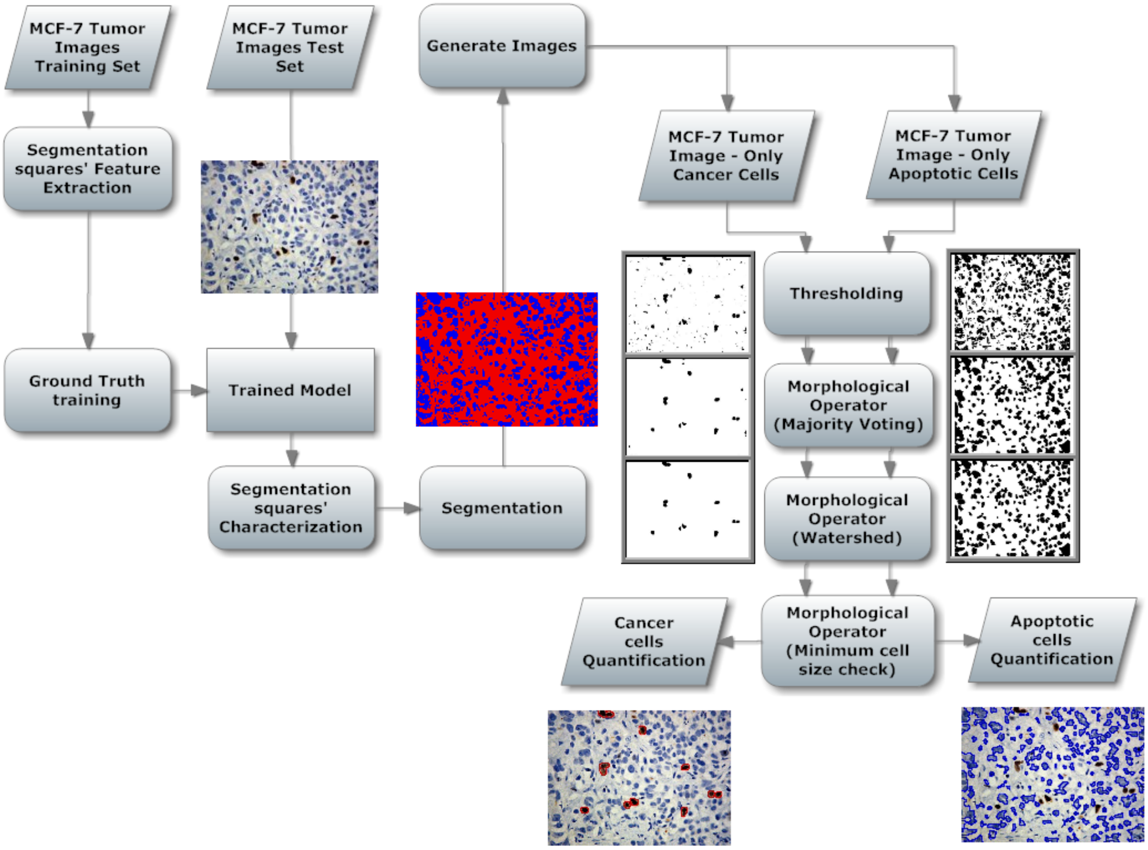

MCF-7 Tumor Images Training Set
MCF-7 Tumor Images Test Set
Generate Images
MCF-7 Tumor Image - Only Cancer Cells
MCF-7 Tumor Image - Only Apoptotic Cells
Segmentation squares' Feature Extraction
Thresholding
Ground Truth training
Trained Model
Morphological Operator (Majority Voting)
Segmentation squares' Characterization
Segmentation
Morphological Operator (Watershed)
Cancer cells Quantification
Morphological Operator (Minimum cell size check)
Apoptotic cells Quantification

Figure 95. Workflow Block Diagram of the block based approach

Following the import of an image in the system, it is first edited by the classification model, which separates the different types of cells into individual generated images. An adaptive thresholding segmentation technique is applied to enhance the visibility of the foreground objects of each generated image separately. Minor noise - generated through misclassification of small segments - is removed by adopting morphological operators, such as Majority Voting and Watershed filtering. A size correction procedure was developed for the valid and accurate quantification of each type of cells. The training of the above model is based on the ground truth, provided by expert pathologists.

Classification

Support Vector Machines (SVM) classifier was selected as the most efficient one, after experimenting with several classifiers such as Naïve Bayes, k-Nearest Neighbor and Decision Trees. The mean Red, Green and Blue channel values were utilized as image features. Each instance was labelled with the corresponding class, in order to make the SVM model capable of recognizing similar instances from unknown images and classify them. The ten-fold cross validation has been adopted for testing the classification accuracy. The width of the segmentation square was selected heuristically, from a 3 to 9 pixels range, and set at 3.

An independent operator (see Figure 96) was developed in order to characterize and quantify the cells from the corresponding. This operator is based on block-based segmentation and image analysis techniques. More specifically, the entire image is scanned into segmentation squares classified by the SVM classifier. Each segmentation square - belonging to a specific class - is assigned with the corresponding color (see Figure 97). Basically, one class contains everything but apoptotic cells (red segmentation squares in Figure 97) and the other everything but cancer cells (blue segmentation squares in Figure 97). The red color segmentation squares generate an image, which discards all the apoptotic class cells (see Fig. 98 b). The blue color segmentation squares do the same on the cancer class cells (see Figure 98 c). The result is the generation of two images, each of them containing one kind of cells.

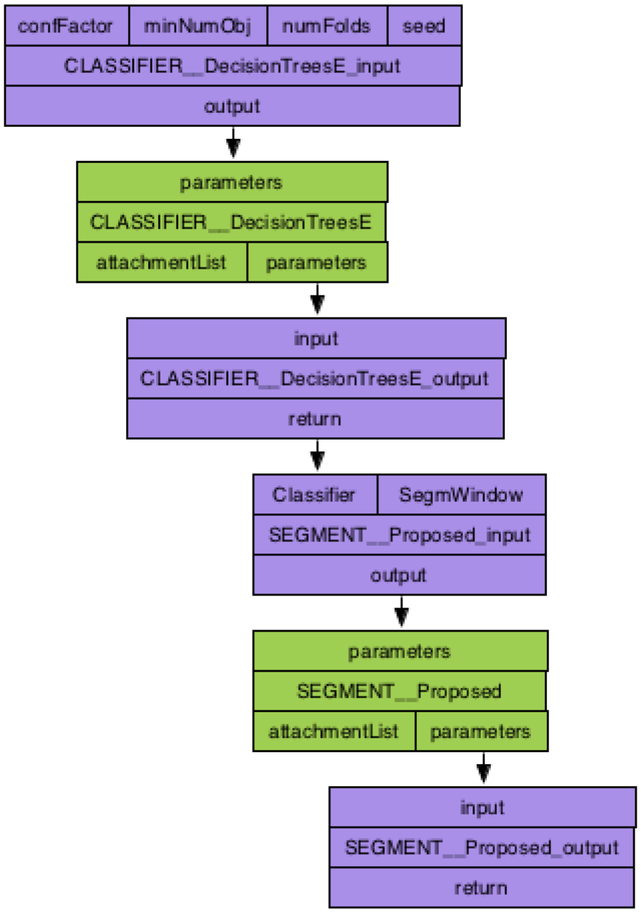


Figure 96. Developed operators for the Breast Cancer Biopsy Images dataset.

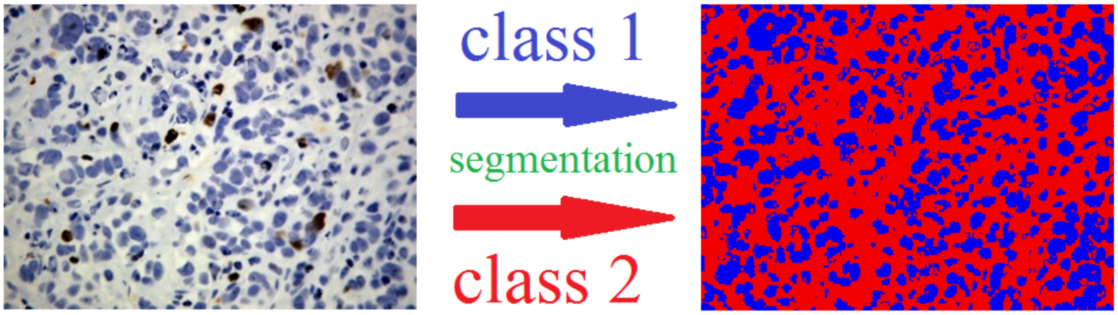


Figure 97. Original and Segmented Breast Cancer Biopsy Image

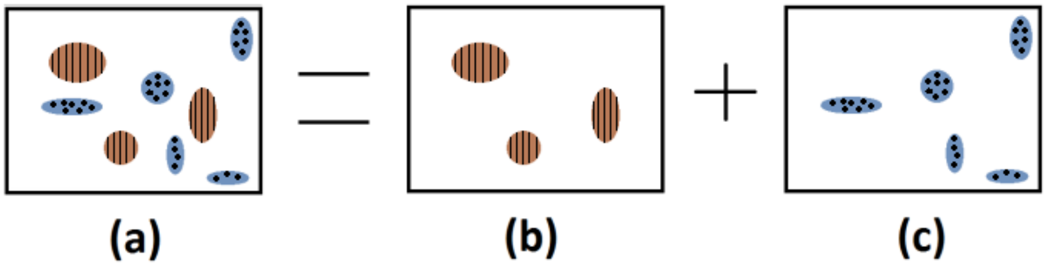


Figure 98. a) Original Image b) Red pixels in Figure 97 c) Blue pixels in Figure 97

Thresholding

Adaptive thresholding segmentation [24] is applied on the grayscale versions for each of the generated images. In order to find the optimal value for the threshold [25], we adopted the following procedure: An initial threshold value is set at the minimum possible value (1), which separates the foreground from the background pixels. The average values of the pixels up to the threshold value (the foreground objects) and the pixels above (the background objects) are calculated. Afterwards, threshold value is increased and the process is repeated, until the threshold value is greater than the composite average. The optimal threshold for the cancer cells image, utilizing the above technique, is T1=170 on the 0-255 scale. The same procedure is applied to the apoptotic image, as well. Due to the fact that the apoptotic cells are lighter than the dark cancer cells, require higher values to the threshold, in order to distinct them from the background. Therefore, the above procedure ended up with a T2=196 threshold value for the apoptotic cells image.

Majority Voting Technique

A morphological filtering is required, in order to remove misclassification issues. The majority vote technique described in [26], was utilized, in order to accomplish this task. The value of the central segmentation square is set by the majority vote of its eight neighbor segmentation blocks. The majority limit is set to five, which means, if five or more neighbor segmentation squares have the same value, which differs from the value of the central one; it is automatically set to the same value with its neighbors.

Watershed Filtering

Then, the watershed filter is applied on this image, in order to accurately cut the merged particles and provide a clear image for the quantification procedure. Majority vote technique was applied before watershed, since we firstly need to enhance the foreground objects and then try to separate the merged ones.

Size-based Correction and Labeling

Size is another major factor, except for the color, that characterizes a cancer or apoptotic cell. Based on the assistance of our expert pathologists, the size of a true cancer cell is interpreted as an approximately 100±15 pixels area on the image. For that reason, an analyze particle method, similar to a component labelling algorithm [27], was developed. In our previous work, [28] we developed a similar algorithm, which worked perfectly on 4-connected patterns, but it had some issues dealing with 8-component patterns. The new enhanced algorithm provides faster scanning and noise removal. This algorithm may apply a two-pass check, but it is quite faster than the previous algorithm. More specifically, the developed algorithm scans the image from right to left, beginning from the top and moving to the bottom, until it finds a pixel of a foreground object, in our case a black one, since we deal with a binary image. The algorithm checks the specific's pixel neighbor pixels. It first checks the left pixel if it has already been labelled. If it is labelled, it assigns the same label to the current pixel. If it is not, it continues a clockwise searching of neighbor pixels up to the upper right pixel. If none of these four neighbor pixels have an assigned label, then a new label is created and assigned to the current pixel. This procedure continues for every foreground pixel (black) of the binary image. When it is complete, a second scan begins, but this time, it scans the image from right to left, but this time from the bottom and moving to the top. Considering that this time every foreground object has already an assigned label, we won't have the case of assigning a new label and the checks applied will be readjusted. When the algorithm finds a pixel of a foreground object (a labelled pixel), it checks, if the left pixel is labelled. If it is, it changes the current label to the left pixel's label. If not, it continues a counter-clockwise searching of neighbor pixels up to the lower right pixel. If none of these four neighbor pixels has an assigned label, then it leaves the current label unchanged.

Pseudocode of the object removal algorithm

```
INPUT: Binary Image after Adaptive Thresholding
INITIALIAZATION: set the values of the 2D integer arrays Label (which contains the label for each pixel of the image) and Object_Array (which contains each foreground pixel's label and its coordinates) to zero, set the values of the 1D integer array Size to zero, set the integer Pos to 0, integer Counter_of_forgroundpixels is set to 0, integer Lab is set to 0 and integer new_label is set to 1. Integer Acceptable_size is set by the user.
        FOR y=0 to the Image Height-1 DO
            FOR x=0 to the Image Width-1 DO
                IF Pixel[x,y] belongs to a foreground object (pixel is black) THEN
                    Counter_of_forgroundpixels++
                    IF the left pixel's label is different than 0 THEN
                        Label[x,y]= left pixel's label
                    ELSE IF upper left pixel's label is different than 0 THEN
                        Label[x,y]= upper left pixel's label
                    ELSE IF upper pixel's label is different than 0 THEN
                        Label[x,y]= upper pixel's label
                    ELSE IF upper right pixel's label is different than 0 THEN
                        Label[x,y]= upper right pixel's label
                    ELSE
                        Label[x,y]= new_label
                        new_label ++
                    END IF
                END IF
            END FOR
        END FOR
        //second pass
        FOR y= Image Height-1 to 0 DO
            FOR x=0 to the Image Width-1 DO
                IF Label[x,y]!=0 THEN
                    IF the left pixel's label is different than 0 THEN
                        Label[x,y]= left pixel's label
```

```
                  ELSE IF lower left pixel's label is different than 0 THEN
                      Label[x,y]= lower left pixel's label
                  ELSE IF lower pixel's label is different than 0 THEN
                      Label[x,y]= lower pixel's label
                  ELSE IF lower right pixel's label is different than 0 THEN
                      Label[x,y]= lower right pixel's label
                  END IF
                  //set values to Object_Array which will help us to remove objects smaller than the
acceptable
              //size
                  Object_Array[0,Pos] = Label[x,y]
                  Object_Array[0,Pos] = x
                  Object_Array[0,Pos] = y
              END IF
          END FOR
      END FOR
      //Invalid object removal
      FOR i = 0 to Counter_of_forgroundpixels-1 DO
          Lab = Object_Array[0,i]  // it reads the label of the object
          Size[Lab] = Size[Lab] + 1 // it increases the size of the label assigned to a specific object
      END FOR
      FOR i = 0 to new_label DO
          IF Size[i] < Acceptable_size THEN
              Remove each pixel assigned with the specific label i
          END IF
      END FOR
```

Additional Features

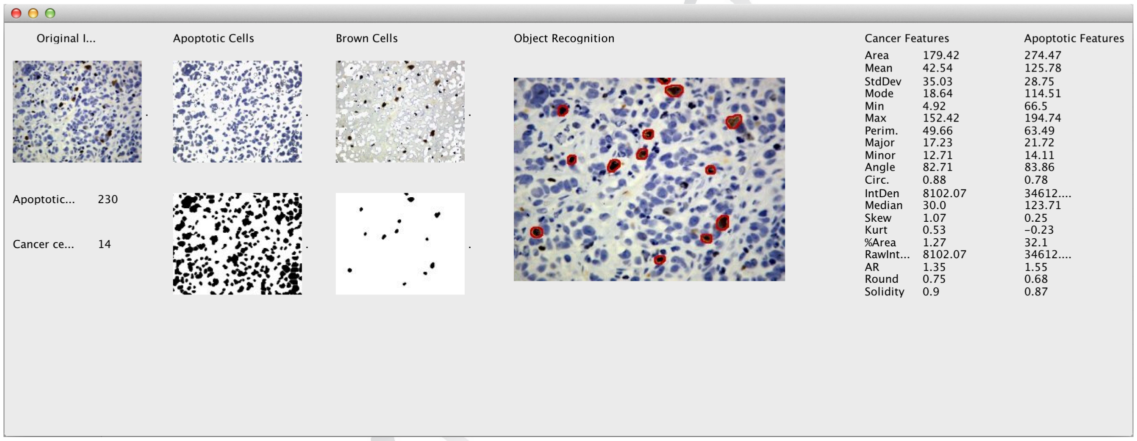
In order to improve the characterization accuracy of the cancer and apoptotic cells, a set of additional features is extracted after the classification. These features are calculated for each apoptotic and cancer cell. Following the extraction of the above features, a mean value of each feature is calculated for each type of cells. Thus, each group of images from each cancer treatment is further characterized. The features that lead to the summary results are presented in Table 9.

Table 9. Extracted features of the proposed application along with their definition.

Feature name	**Brief explanation**
Area	Area is the number of pixels of each object. The average value provides an estimation of the mean size of an apoptotic and cancer cell respectively in each case of the available treatments.
Mean (gray) value	This feature is the mean value of the average gray of each annotated cell. In RGB images, pixels are represented by three values of the corresponding colors, each of them ranges from 0 to 255. Thus, the following formula is used [29], in order to convert each pixel to gray scale: $\mathrm{gray}_{\mathrm{value}} = 0.299 * \mathrm{red}_{\mathrm{value}} + 0.587 * \mathrm{green}_{\mathrm{value}} + 0.114 * \mathrm{blue}_{\mathrm{value}}$ (1)
Standard Deviation	This is the standard deviation of the gray values of the pixels of a specific cell. This feature indicates the diversity of each type of cells.
Modal gray value	This feature indicates the most frequently occurring pixel value of each cell.
Minimum and maximum gray value	This feature exports the minimum and maximum gray values of the pixels of each cell. The motivation of this feature is to examine the gray value ranges of each cell type.
Perimeter	This feature, along with the Area feature, provides an estimation of what is the size and the length of the boundaries of each cell type.
Ellipse	This feature is based to the theory of trying to fit an ellipse to the cell area.
Major and minor	These two values indicate the length (in pixels) of the major and minor axis of the ellipse surrounding each cell.
Angle	This is the angle between the major axis of the ellipse and X-axis of the image

Circularity	Circularity is a metric that indicates how similar to a circle is an object. The formula that calculates the circularity of an object is the following: $Circularity = 4\pi \times \frac{Area}{Perimeter^2}$ (2)
Ellipse Aspect Ratio	The aspect ratio of an ellipse is the ratio of the length of its major axis over the length of its minor axis
Integrated Density	This feature indicates the sum of the pixels of a cell. This was used as a metric of how bright each type of the cells is
Median	Median indicates the median gray value of the pixels of each cell
Skewness	Skewness indicates how the gray values of the cell's pixels differ in shape from a fully symmetrical distribution. It also denotes to which direction (positive or negative), deviates from the symmetrical distribution. If skewness equals 0, then the distribution of the values of the specific cell, draw symmetrical distribution and mean, mode and median features are equal. If the frequency of the brighter gray value is greater than the darker ones, that indicates a negatively skewed distribution. Thus, the opposite indicate a positively skewed distribution.
Kurtosis	Kurtosis measures how rough is the peak of the above described distribution, which indicates a measure on how the pixels of each cell are distributed in the acceptable value range, versus the center of the distribution.
Area Fraction	This feature indicates the percentage of each cell type in a specific microscopy image. This feature provides an estimation of which type of treatment the image belongs.
Roundness	Similar to Circularity metric with the difference that this metric is based on the fitting ellipse features. The formula that calculates the roundness of an object is the following: $Roundness = 4 \times \frac{Area}{\pi \times MajorAxis^2}$ (3)
Solidity	This feature indicates is the ratio of the area of the cell over the convex area. This characterizes the geometry of each type of cells. This feature

	- combined with the Circularity and Roundness feature - it can characterize how sharp the edges of a cell type are, despite the fact they are in circular forms.
Cancer Factor	Cancer Factor is a feature that denotes the quantification of the breast cancer pathogenesis and is calculated through the following formula: $Cancer\ Factor = \frac{Number\ of\ valid\ cancer\ cells}{Number\ of\ valid\ apoptotic\ cells}$ (4)



Apoptotic... 230

Cancer ce... 14

Cancer Features		Apoptotic Features
Area	179.42	274.47
Mean	42.54	125.78
StdDev	35.03	28.75
Mode	18.64	114.51
Min	4.92	66.5
Max	152.42	194.74
Perim.	49.66	63.49
Major	17.23	21.72
Minor	12.71	14.11
Angle	82.71	83.86
Circ.	0.88	0.78
IntDen	8102.07	34612....
Median	30.0	123.71
Skew	1.07	0.25
Kurt	0.53	−0.23
%Area	1.27	32.1
RawInt...	8102.07	34612....
AR	1.35	1.55
Round	0.75	0.68
Solidity	0.9	0.87

Figure 99. Proposed tool's output containing the necessary characterization (the specific example is based on the block-based approach)

4.4.3 Mean Shift Segmentation Approach

The second segmentation approach is based on mean shift algorithm, in order to provide fast and valid segmentation of the several classes of each image - utilizing proper morphological techniques - in order to remove noise and other misclassification issues. Generally, the first approach is more accurate than the second one, but the latter achieves better time performance.

The Proposed Methodology

In this approach, images are first edited by mean shift filtering, which accomplishes the separation of the different type of cells into their respective generated images.

Mean shift [31], is based on a simple idea, but is a quite powerful and versatile tool, able to deal with complex tasks. It is used for finding nodes and for clustering tasks and quite frequently for image analysis tasks, like segmentation and tracking. Mean shift is a non-parametric iterative algorithm, introduced by Fukunaga and Hostetler in 1975, which can handle any kind of feature space (color space, scale space, etc.) and does not require any prior knowledge about the number of the clusters.

Mean shift was chosen due to its features, which make it eligible to deal with the specific task. Its simple idea makes it adaptable to many types of application. It is fast and suitable for real and heavy data analysis and does not require any prior knowledge or pattern on the specific field of application or prior shape on data clusters. It can easily handle color feature space. It has only one parameter to edit (h, the window size), which make it easily adaptable to the proposed application. The window size is quite crucial, since inappropriate window size can lead to merges of modes or potential generation of additional “fake” modes. Adaptive window size can deal with this issue.

Mean shift algorithm is based on Gradient Ascent method and tries to find the modes in a set of data points, denoting an underlying probability density function (PDF) in R^N. Thus, Gradient Ascent approach is applied on the local estimated density, until convergence - for each data instance.

Gradient Ascent

Gradient Ascent [32] is a global approach for the solution of optimization issues, where there is a necessity of maximizing functions of continuous parameters. For instance, in order to find the maximum value of a function *y=f(x)* (consider *f(x)* as a difficult equation). The algorithm for finding the maximum is the following:

- *y=f(x)* has a local maxima at x_{max}
- x_1 a random value from the data range
- calculate $f(x_1)$
- if there is a positive slope of y ($f'(x_1)$>0) then the x_{max} is greater than x_1 and it is found at the right of x_1.
- on the other hand, if *f '(x1)*<0 then the x_{max} will be less than x_1, which implies that it is found at the left of x_1.

In this way, the direction in which x_1, should move to, in order to approach x_{max} is known.

$$x_1 \leftarrow x_1 + nf'(x_1)$$

Where *n* is a positive constant. In case of extreme small *n*, and there is also a local maxima for *f*, the above rule will converge to it after a specific number of loops.

Mean Shift Segmentation Methodology

The sampling of the data points is performed from an underlying probability density function. When a non-parametric density estimation is applied on the input data points, a discrete probability density function representation is produced. On the other hand, if a non-parametric gradient estimation is applied to the same source, a full probability density function analysis is produced.

Consider $\{x_i\}_{i=1\ldots n}$ as a set of n points in a N-dimensional Euclidian space. The kernel density estimation [31] is calculated for the point x with a kernel K(x) [32].

$$\hat{f}(\mathrm{x}) = \frac{1}{nh^N}\sum_{k=0}^{n} \mathrm{K}\left(\frac{x-x_i}{h}\right)(1)$$

where h is the window radius.

The Epanechnikov kernel provides the minimum mean integrated square error.

$$K_E(x) = \begin{cases} \frac{1}{2}c_N^{-1}(N+2)(1-x^T x), & if\ x^T x < 1 \\ 0, & otherwise \end{cases} (2)$$

where c_N is the volume of the N-dimensional sphere unit [32].

Mean shift's calculation is based on the gradient of the kernel density estimation (1).

$$\widehat{\nabla} f(\mathrm{x}) \equiv \nabla \hat{f}(\mathrm{x}) = \frac{1}{nh^N}\sum_{i=1}^{n} \nabla K\left(\frac{x-x_i}{h}\right)(3)$$

The mean shift of data point x is calculated through the formula below.

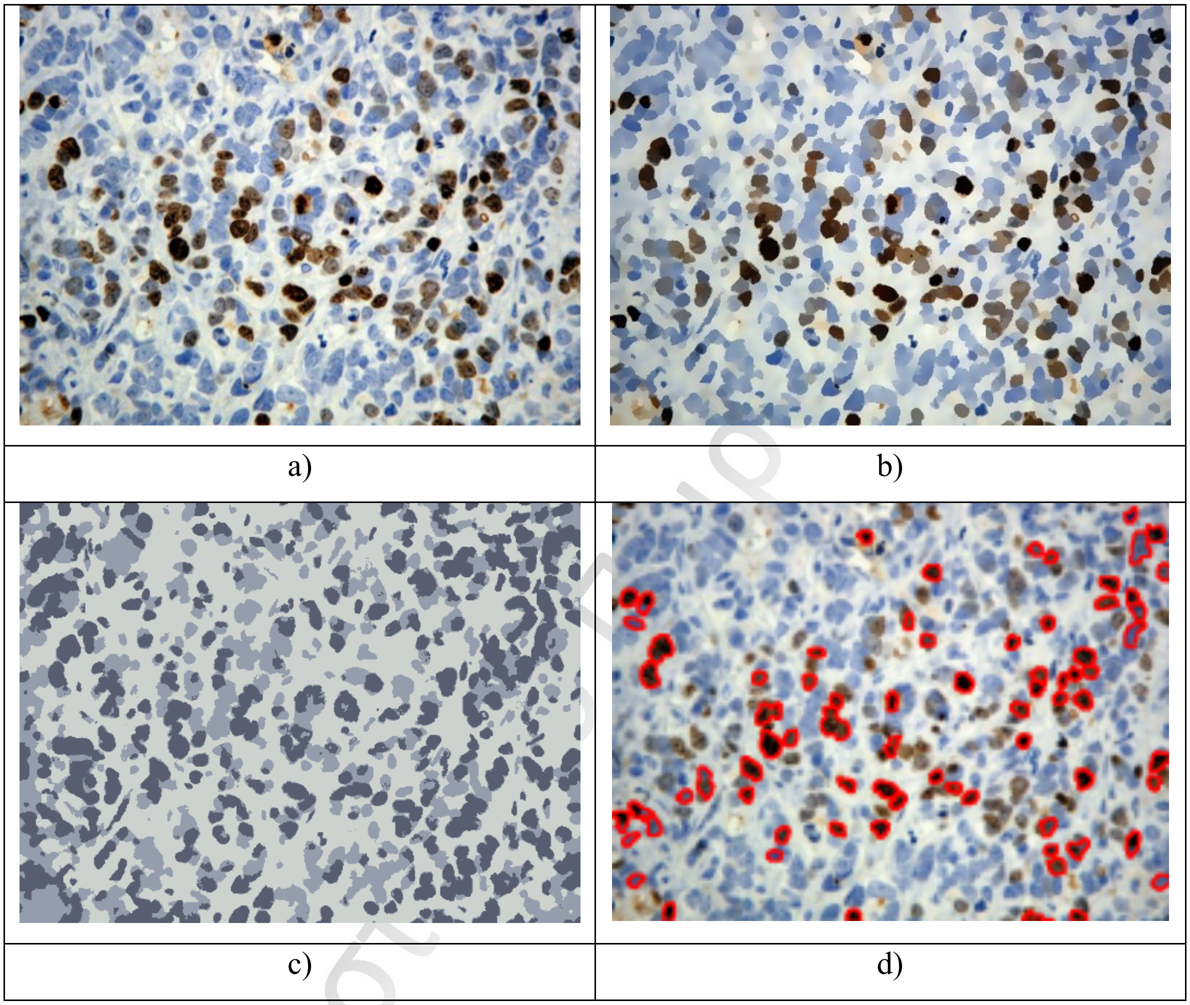
$$m(x) = \frac{\sum_{i=1}^{n} g\left(\frac{x-x_i}{h}\right)x_i}{\sum_{i=1}^{n} g\left(\frac{x-x_i}{h}\right)} - x(4)$$

where g(x)= - K'(x). Thus, a typical mean shift iteration procedure includes the computation of mean shift vector $m(x_i^t)$ and the move of the density estimation window by $m(x_i^t)$ for each data point x_i. The necessary formulas for the proof of the mean shift's formula are included in [31].

Since this work deals with microscopy images, the corresponding feature space is the color space. In color space the spatial location and the particular color of each pixel must be taken under consideration. For each pixel, a set of neighbor pixels, in a range of a specific spatial radius and a specific color distance is defined. For the initial set of neighboring pixels the new spatial color mean are calculated. These calculations will define the new center for the next iteration of the algorithm. This procedure repeats, until the new spatial and color means stay the same after the above calculations. The value of the color mean, from the last iteration, will be assigned to the initial position of the specific set of iterations.

Dataflow of the proposed approach

After the application of the customized mean shift algorithm that was described above, a novel methodology (see Figure 101) is performed, in order to acquire the informative part of the image. The informative part, only consists of cancer and apoptotic cells. The artifacts and the background are smartly removed. By utilizing the Median Cut quantization algorithm [35] (see Figure 100 c).



a)

b)

c)

d)

Figure 100. a) Original image b) after meanshift c) median cut quantization applied for three colors d) final annotation of the cancer cells

Automated thresholding, utilizing the Otsu methodology [36], is performed, in order to remove the minor remaining artefacts and the noisy fractions. A clear and sharp image is acquired through the previous steps. This image, in some cases, contains some wrongfully active pixels that survived from the above noise removal methods. For an even clearer result, an open [34] filtering is applied, providing excellent results. The specific image contains the mask from each salient object. The modified component-labelling algorithm is applied, in order to remove the invalid cells from the last image.

After the above-mentioned procedures, the masks of the valid salient object are used, in order to acquire the informative pixels from the mean shifted image and move them into a new clean white image. Median-cut color quantization [35] was utilized for the final color segmentation, in order to proceed to the generation of separate images for each class. More specifically, based on [35], there are only 2 colours in the entire image, which denote the different classes, plus the white background (see Fig. 100 b). After the generation of the images, the same feature extraction for the additional characterization of the cell is applied on each of them.

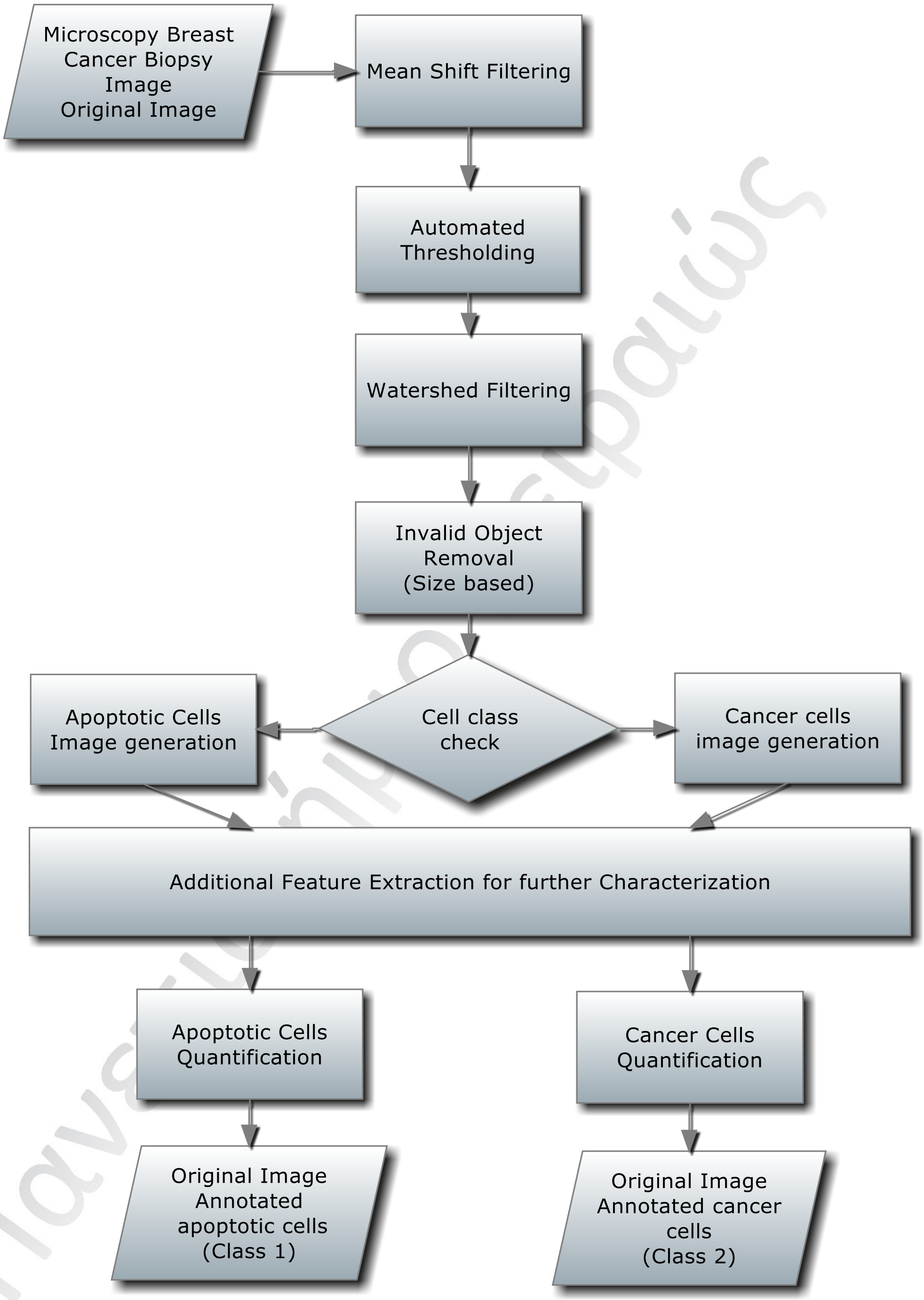

Microscopy Breast Cancer Biopsy Image Original Image
Mean Shift Filtering
Automated Thresholding
Watershed Filtering
Invalid Object Removal (Size based)
Cell class check
Apoptotic Cells Image generation
Cancer cells image generation
Additional Feature Extraction for further Characterization
Apoptotic Cells Quantification
Cancer Cells Quantification
Original Image Annotated apoptotic cells (Class 1)
Original Image Annotated cancer cells (Class 2)

Figure 101. Dataflow diagram of the mean shift based approach

4.5 Melanoma Detection and Annotation on Skin Lesion Images

Skin cancer is among the most frequent types of cancer and one of the most malignant tumors. Its incidence has increased faster than that of almost all other cancers and the annual rates have increased on the order of 3–7% in fair-skinned populations in recent decades [37]. Currently, between 2 and 3 million non-melanoma skin cancers and 132000 melanoma skin cancers occur globally each year. Skin cancer is the most common form of cancer in the United States.

Considering the above facts, a tool that will be able to handle the vast amounts of images that expert physicians deal with daily is required. The proposed tool will be able to annotate accurately the skin lesion image and perform a feature extraction that will characterize the salient area.In this approach (see Figure 102), images are first edited by a subtract background algorithm, in order to de-noise the images from the artifact objects and the lighting issues. Following this, mean shift segmentation algorithm is applied, in order to distinct the salient objects from the background. The diagram of the proposed methodology is depicted in Figure 102.

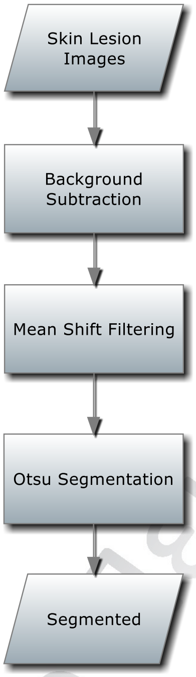


Figure 102. The workflow diagram of the proposed approach for the skin lesion detection.

4.5.1 Background Subtraction

Each of the examined images contains one salient object, in most cases, but their background may also contains some noise. The meaning of the noise term is the uneven illumination of the background, the few uncut hair and some artifact objects (pimples, blains, etc.) of high chromatic similarity with the skin lesion. A background subtraction algorithm is performed, in order to acquire the salient objects of the microscopy image.

The background algorithm subtraction method is based on the idea of the 'rolling ball' algorithm described in [33]. Based on the rolling ball algorithm, the grayscale image is considered as a 3D surface, where the third dimension is the value of each pixel of the image. Thus, a rough surface is generated. A ball with a specific radius is rolled over the raw surface. The ball's size must be equal or bigger than the larger foreground item. Each point - reachable by the ball - is the background. Thus, for each pixel, a specific background value is set, based on the local average of the ball's radius. These values are subtracted from the original image, providing an image with a clear (bright) and smoother background (see Figure 103 b).

4.5.2 Mean shift Segmentation

Following the background procedure, the mean shift algorithm accomplishes the accurate detection of the salient object. Mean shift, is based on a simple idea, but is a quite powerful and versatile tool, able to deal with complex tasks. It is used for finding nodes and for clustering tasks and quite frequently for image analysis tasks, like segmentation and tracking. Mean shift is a non-parametric iterative algorithm, introduced by Fukunaga and Hostetler in 1975, which can handle any kind of feature space (color space, scale space, etc.) and does not require any prior knowledge about the number of the clusters.

Its simple idea makes it adaptable to many image analysis applications. It is fast and suitable for real and heavy data analysis and does not require any prior knowledge or pattern on the specific field of application or prior shape on data clusters. It can easily handle color feature space. It has only one parameter to edit (h, the window size), which make it easily adaptable to the proposed application. The window size is quite crucial, since inappropriate window size can lead to merges of modes or potential generation of additional "fake" modes. Adaptive window size can deal with this issue.

Mean Shift theory was analytically described in Chapter 4.4.3., so no further analysis will be provided to the current chapter.

Since this work deals with microscopy images, the corresponding feature space is the color space. In color space, the spatial location and the particular color of each pixel must be taken under consideration. For each pixel, a set of neighbor pixels, in a range of a specific spatial radius and a specific color distance is defined. For the initial set of neighboring pixels, the new spatial color mean are calculated. These calculations will define the new center for the next iteration of the algorithm. This procedure repeats, until the new spatial and color means stay the same after the above calculations. The value of the color mean, from the last iteration, will be assigned to the initial position of the specific set of iterations. Mean shift was chosen due to its features, which make it eligible to deal with the objects segmentation task.

4.5.3 Auto-thresholding (Otsu segmentation)

An auto thresholding algorithm [36] is performed, in order to remove the minor remaining artefacts and the noisy fractions. Otsu's segmentation algorithm [36] was proposed in 1979 and it is considered one of the most efficient thresholing segmentation algorithms. The selection of the optimal threshold value is based on the maximization of the difference between the dark and the light areas. Otsu's approach assumes that there are only two classes in the image (usually foreground and background pixels).

The algorithmic steps are the following:

The intra-class variance calculation is presented below, which is typically the weighted variance of the foreground pixels class plus the weighted variance of the background pixels class:

$$\sigma_w^2 = w_b(t)\sigma_b^2(t) + w_f(t)\sigma_f^2(t)$$

Where,
b: the background class
f: the foreground class
w: the probability of each class
t: the separating threshold
σ_x^2: the corresponding variance of each class
b: denotes the background pixels class
f: denotes the foreground pixels class

The probability of each class is calculated via the image's histogram.

$$w_b(t) = \sum_{0}^{t} p(i)$$

The class mean equals:

$$\mu_b(t) = \frac{\sum_0^t p(i)x(i)}{w_b}$$

Where,
x(i): the i^{th} histogram bin's center value.

Particularly, the algorithm calculates the histogram and exports the probabilities of the classes for each brightness level. Following the initialization of the class probability (from $w_i(0)$) and the class mean(from $\mu_i(0)$), an iterative procedure checking each potential threshold value, starting from 1 and reaching the maximum brightness value, is taking place. Specifically, it iteratively updates the w and μ values and calculates the corresponding σ_b^2 until it finds the maximum difference between the foreground and background pixels classes. This implies to the greatest values of a σ_b^2 (threshold1_value) and σ_f^2(threshold2_value). Thus, the optimal threshold value equals:

$$Optimal\ Threshold = \frac{\text{threshold1\_value} + \text{threshold2\_value}}{2}$$

A clear and sharp image is acquired through the previous steps. This image, in some cases, contains some wrongfully active pixels that survived from the above noise removal methods. The specific image contains the mask from each salient object. The modified component-labelling algorithm is applied, in order to remove the invalid cells from the last image.

Following the above-mentioned procedures, the masks of the valid salient object are used, in order to acquire the informative pixels from the mean shifted image and move them into a new clean white image.

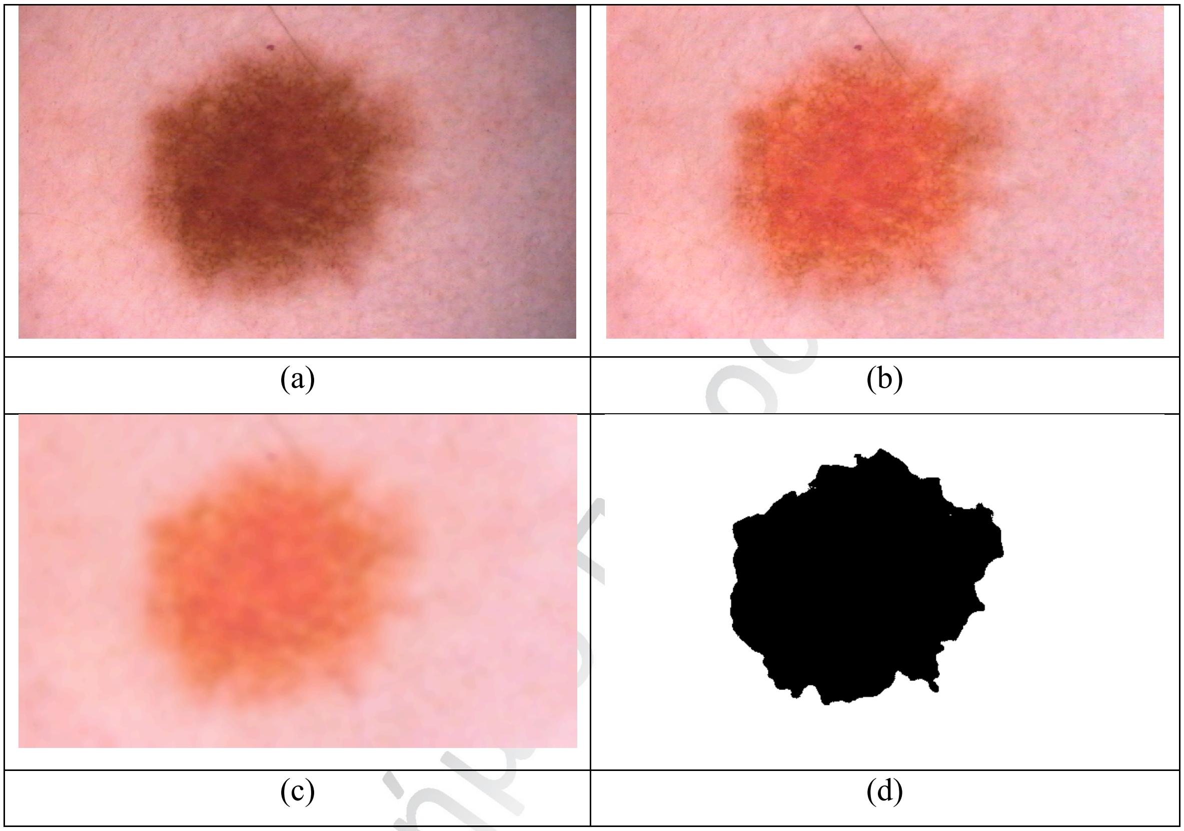

(a) (b) (c) (d)

Figure 103. a) The original mole image b) The mole image after the background subtraction c) The image from (b) after the application of mean shift filtering and d) the salient object mask extracted from the proposed workflow scheme.

4.6 References

[1] 1. S. Klahr, “Obstructive nephropathy”. Department of Internal Medicine, Barnes-Jewish Hospital (North Campus) at Washington University School of Medicine, *Internal medicine* pp. 355-361, 2000.

[2] T. Caballero, A. Pérez-Milena, M. Masseroli, F. O'Valle, F. J. Salmerón, R. M. G. Del Moral, and G. Sánchez-Salgado, "Liver fibrosis assessment with semi-quantitative indexes and image analysis quantification in sustained-responder and non-responder," *Journal of Hepatology*, vol. 34, pp. 740-747, 2001.

[3] J.P Royet, C. Souchier, F. Jourdan, H. Ploye, “Morphometric Study of the Glomerular Population in the Mouse Olfactory Bulb: Numerical Density and Size Distribution Along the Rostrocaudal Axis”, The journal of comparative neurology 270559-568 (1988)

[4] "Proximal convoluted tubule." *Encyclopædia Britannica. Encyclopædia Britannica Online*. Encyclopædia Britannica, 2011. Web.27Jun.2011http://www.britannica.com/EBchecked/topic/480781/proximal-convoluted-tubule.

[5] M. Haralick et al, (1973) Textural Features for Image Classification. IEEE Transactions on systems man and cybernetics Vol. SMC-3 pp. 610-621

[6] C. Cortes, V. Vapnik, “Support-Vector Networks”, *Machine Learning*, 20, pp.273-297, 1995.

[7] Friedman N, Geiger D, Moises et al., “Bayesian Network Classifiers”, *Machine Learning*, pp. 131-163, 1997.

[8] N. Roussopoulos, S. Kelley, F. Vincent, “Nearest Neighbor Queries”, SIGMOD '95 Proceedings of the 1995 ACM SIGMOD international conference on Management of data ISBN:0-89791-731-6, p71-79, 1995.

[9] T. Mitchell, "Decision Tree Learning", in T. Mitchell, *Machine Learning,* The McGraw-Hill Companies, Inc., 1997, pp. 52-78

[10] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.” *Proceedings of the 14th International Joint Conference on Artificial Intelligence* P.1137—1143, 1995.

[11] Balazs Harangi, Rashid Jalal Qureshi, Adrienne Csutak, Tünde Petö, András Hajdu. Automatic detection of the optic disc using majority voting in a collection of optic disc detectors. In Proceedings of ISBI'2010. pp.1329-1332

[12] J. A. Hartigan, M. A. Wong. “Algorithm AS 136: A K-Means Clustering Algorithm”**.** *Journal of the Royal Statistical Society. Series C (Applied Statistics)* Vol. 28, No. 1, pp. 100-108, 1979

[13] J. G. Cleary , L. E. Trigg. “K*: An Instance-based Learner Using an Entropic Distance Measure”. Proceedings of the 12th International Conference on Machine Learning**,** Volume: 5, Publisher: Morgan Kaufmann, pp. 108-114. 1995

[14] N. Roussopoulos, S. Kelley, F. Vincent, “Nearest Neighbor Queries”, SIGMOD '95 Proceedings of the 1995 ACM SIGMOD international conference on Management of data ISBN:0-89791-731-6, p71-79, 1995.

[15] L. Breiman. “Random Forests”. Machine Learning, Volume 45, Number 1, pp. 5-32, 2001

[16] T. Mitchell, "Decision Tree Learning", in T. Mitchell, Machine Learning, The McGraw-Hill Companies, Inc., 1997, pp. 52-78

[17] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.” Proceedings of the 14th International Joint Conference on Artificial Intelligence pp.1137—1143, 1995.

[18] German RR, Fink AK, Heron M, Johnson CJ, Finch JL, Yin D (2011) The Accuracy of Cancer Mortality Group: The accuracy of cancer mortality statistics based on death certificates in the United States. Cancer Epidemiology, vol. 35, Issue 2, pp. 126-131, April 2011.

[19] Loncaster J, Dodwell D. Adjuvant radiotherapy in breast cancer (2002) Are there factors that allow selection of patients who do not require adjuvant radiotherapy following breast-conserving surgery for breast cancer? Minerva Med. 93, pp. 101–107.

[20] Hansen CM, Hamberg KJ, Binderup E, Binderup L (2000) Seocalcitol (EB 1089): A vitamin D analogue of anticancer potential. Background, design, synthesis, preclinical and clinical evaluation. Curr. Pharm. Design, 6: pp. 803–828.

[21] Soule HD, Vazquez J, Long A, Albert S, Brennan M (1973) A human cell line from a pleural effusion derived from a breast carcinoma. Journal of the National Cancer Institute 51 (5), pp.1409–1416.

[22] Chen A, David BH, Bissonnette M, Scaglione-Sewell B, Brasitus TA (1999) 1, 25-Dihysdroxyvitamin D3 stimulates activator Protein- 1 dependent Caco-2 cell differentiation. J. Biol. Chem. 274: 35505–35513.

[23] Sundaram S, Sea A, Feldman S, Strawbridge R, Hoopes P, Demidenko E, Binderup L, Gewirtz A (2003) The Combination of a Potent Vitamin D3 Analog, EB 1089, with Ionizing Radiation Reduces Tumor Growth and Induces Apoptosis of MCF-7 Breast Tumor Xenografts in Nude Mice1. Clinical Cancer Research, vol. 9, pp. 2350-2356 June 2003.

[24] Batenburg KJ, Sijbers J (2009) Adaptive thresholding of tomograms by projection distance minimization. Pattern Recognition. Volume 42, Issue 10, pp.2297-2305, October 2009.

[25] Ridler TW, Calvard S (1978) Picture thresholding using an iterative selection method. IEEE Trans. System, Man and Cybernetics, SMC-8 pp.630-632, 1978.

[26] Harangi B, Qureshi RJ, Csutak A, Petö T, Hajdu A (2010) Automatic detection of the optic disc using majority voting in a collection of optic disc detectors. Proceedings of ISBI'2010, pp.1329-1332.

[27] Suzuki K, Horiba I, Sugie N (2003) Linear-time connected-component labeling based on sequential local operations. Computer Vision and Image Understanding, vol. 89 Issue 1, pp.1-23, January 2003.

[28] Goudas T, Maglogiannis I (2012) Cancer cells detection and pathology quantification utilizing image analysis techniques. Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE pp. 4418-4421.

[29] “ImageJ User Guide” http://imagej.nih.gov/ij/docs/guide/user-guide.pdf

[30] National Cancer Institute, http://web.ncifcrf.gov/

[31] Comaniciu, D. and Meer, P. Mean shift analysis and applications. Computer Vision. The Proceedings of the Seventh IEEE International Conference on (Volume:2), 1999. Pp 1197-1203.

[32] Fukunaga, K. ; Hostetler, L. “The estimation of the gradient of a density function, with applications in pattern recognition”, Information Theory, IEEE Transactions on (Volume:21 , Issue: 1) January 1975, pp. 32-40.

[33] Stanley Sternberg, "Biomedical Image Processing", IEEE Computer, Volume 16, Issue 1 January 1983, pp 22-34.

[34] Qing Liu and Cheng-yu Lai. «Edge detection based on mathematical morphology theory», Image Analysis and Signal Processing (IASP), 2011 International Conference on, 21-23 Oct. 2011. pp 151 – 154.

[35] Kruger, Anton. "Median-cut color quantization." Dr Dobb's Journal-Software Tools for the Professional Programmer 19.10 (1994): 46-55.

[36] Nobuyuki Otsu (1979). "A threshold selection method from gray-level histograms". *IEEE Trans. Sys., Man., Cyber.* **9** (1): 62–66. doi:10.1109.

[37] Marks R. “Epidemiology of melanoma”. Clin. Exp. Dermatol. vol. 25, pp.459–63, 2000.

[38] Josefsson S. The Base16, Base32, and Base64 Data Encodings. IETF. October 2006. RFC 4648. Retrieved March 18, 2010.

CHAPTER 5

Evaluation Results

In this chapter, a full description of the previously presented methodologies and framework features' results is given, in order to prove the easiness, versatility and efficiency of the developed framework. This chapter also presents the results of an additional image analysis score-based methodology on a complex kidney biopsy dataset.

5.1 Detection of Obstructive Nephropathy

In the specific images, four (4) types/classes of salient objects are recognized. They are namely: non-Pathogenic Glomerulus, Pathogenic Glomerulus, non-Pathogenic Tubulus and Pathogenic Tubulus.

The only preprocessing step of the image data concerns gray scale conversion. Some additional preprocessing techniques were tested (i.e. histogram equalization and contrast enhancement etc.), but they were omitted at the end, since they provided lower accuracy. The workflow scheme of the proposed methodology is depicted in Figure 104.

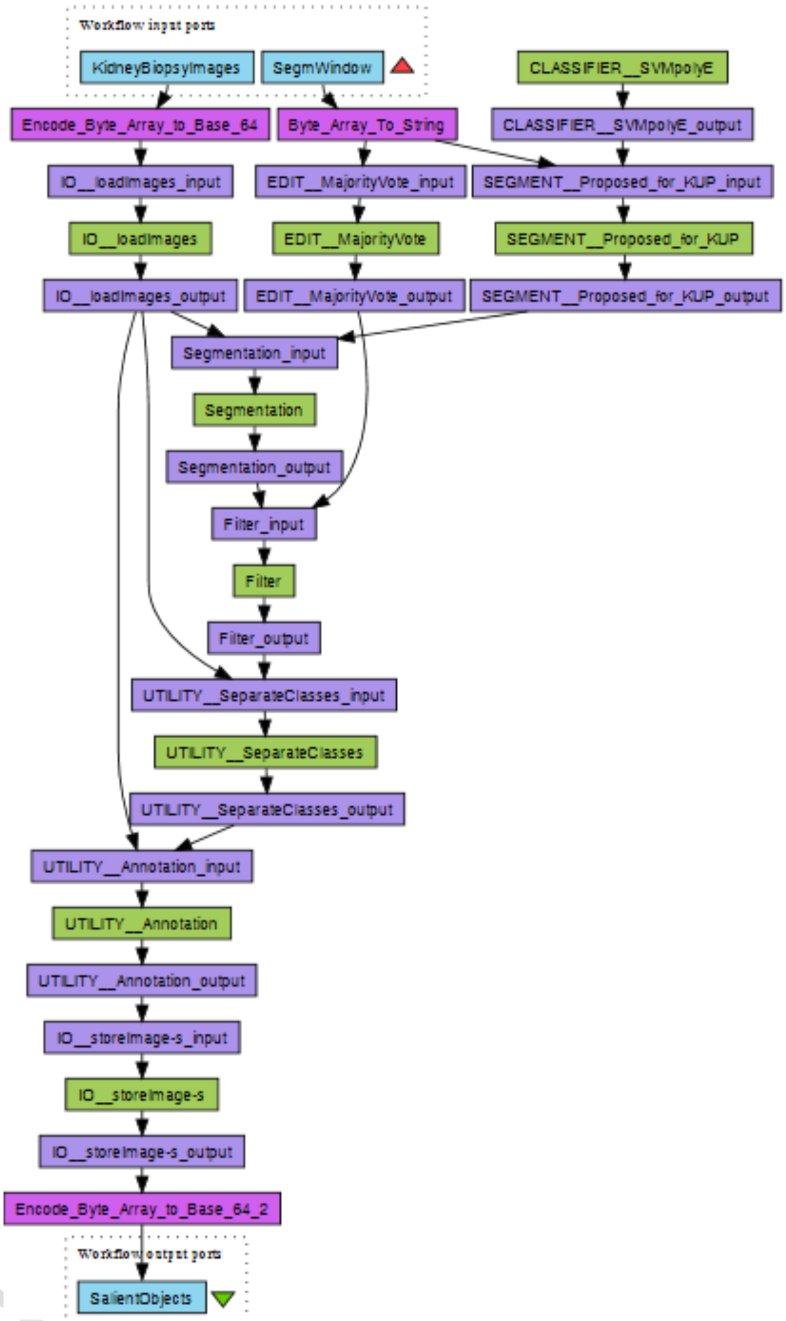


Figure 104. Workflow Scheme Designed in Taverna, utilizing the developed framework's operators.

Each block is allocated through the classification procedure to a specific class. Four widely known classifiers were examined, namely: Support Vector Machines (SVM) [4], Naïve Bayes [5], K-Nearest Neighbor [6] and Decision trees [7].

During the training procedure of the classification model, representative regions were selected through the above feature extraction procedure. The expert biologists participating in this research did the selection. The ten-fold cross validation [8] has been adopted as a method for testing the accuracy. As depicted in Table I, the results are satisfactory. The best performing classifier for the current problem is the SVM classifier achieving 94.7% accuracy according to ground truth blocks, defined by the experts.

The produced segmented images are quite noisy, thus a morphological filtering is required (see Figure 105). A simple pixel majority vote [9] recursive technique with a dynamic vote limit proved sufficient for this task. Regarding the Glomerulus object, a threshold of a minimum area of 7 blocks is set because as it noticed through experiments and observation, there is no glomerulus smaller than 9 blocks. The whole segmentation procedure is depicted in Fig. 4 for several images.

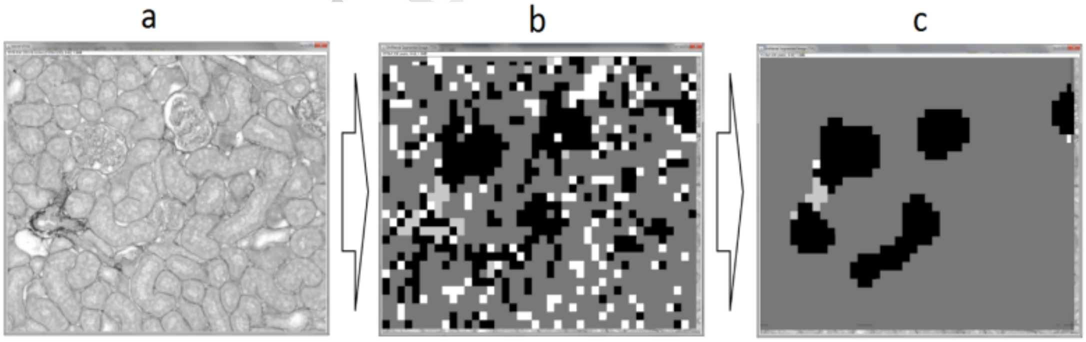


Figure 105. a) Original image, b) Segmented image, c) Majority vote enhanced

As it is depicted in Figure 105, the automated segmentation and characterization method manages to detect the salient objects in the kidney biopsy images. The tool was tested in eight (8) unknown images: four (4) healthy/control and four (4) pathogenic. The corresponding results are presented in Table IV, where Area Involved column stands for the percentage of the image that represents a specific salient foreground object.

The sensitivity, specificity and accuracy scores of the solution are presented in Table 11. As it can be noticed, since the background classes (Pathogenic Tubulus and Non-pathogenic Tubulus) cannot be united to one, the above values are presented only for the foreground objects. Rarely, mischaracterization of recognized objects to their opposing physiological state takes place in control images (false positives). This is a result of the extremely heterogeneous textural profile, perplexing even the 23-feature space, regarding the unambiguous characterization of salient object patterns. The situation is even more hurdled by the existence of artifact, not informative objects, which introduce additional “noise” to the images. Nevertheless, as can be seen from Tables 10 and 11 the overall object recognition performance is exceptional, managing to correctly recognize and discriminate salient geometries in both physiological states, characterized by extreme heterogeneity, with very high accuracy rates. This methodology aids significantly the efficient characterization and physiological scoring of the biopsies, providing a valuable tool for biomedical experts in the direction of efficiently performing batch image processing, ideally managing large data volumes.

5.1.1 User Perspective Evaluation of the Framework

The presented system provides a generic image-processing framework that can be used for different types of images. These tools fill image data pre-processing gaps that are critical for a broad range of image mining experiments. In order to provide analysts and users with an image processing toolbox for different variants of image data, including biological data, a suitable set of basic image processing tools and operators has been integrated. Furthermore, the proposed framework can incorporate workflow templates for common image processing and mining tasks. Processing workflows include image acquisition and display functionality, color conversion, transformation operations, image enhancement and segmentation. The mining workflows can be utilized either for generic image mining through using textural or transformation features or for domain specific mining workflows, such as the one presented in this work as case study.

The framework has been undergone a qualitative evaluation by three (3) biology collaborators against (i) system usability, (ii) speed and (iii) performance. Their opinion is that image mining until now is only achievable by existing image processing SW, which it has usually limited capabilities, it is non-customizable and it works as black box for them. In case they need to analyze specific image types and perform specific mining tasks, they have to employ specialized highly-trained professionals, i.e. computer scientists with image processing and computer vision skills.

In order to evaluate our proposed system's usability, we adopted the System Usability Scale (SUS) [16]. The SUS is a simple and quick, ten-item scale giving a global view of subjective assessments of usability. We collected 15 SUS questionnaires from people who tried our framework. The average score touched 72,6 in a range from 0 to 100, proving its usability and its user-friendly profile. A brief description of SUS questionnaire and its evaluation metric is provided in APPENDIX A.

The proposed framework reduces significantly the technical burden and allows them to perform customizable simple or advanced image mining task, without requiring detailed technical knowledge. Besides making image mining easier for inexperienced users, the utilization of the framework speeds-up their work. Our users received a short one-day training on basic image mining staff and on the system usage and then, they were asked to build some image mining workflows. The conceptual design and execution of image analysis tasks was just a matter of minutes (in case of using existing common workflows) or a few hours, as the experimentation with the proposed system showed. Finally, in terms of performance, the framework is capable to produce efficient workflows for quite difficult image mining tasks, such as the detection of salient objects in kidney biopsies and the quantification of obstructive nephropathy. The intervention, however, of image mining experts is occasionally needed for building complex workflows.

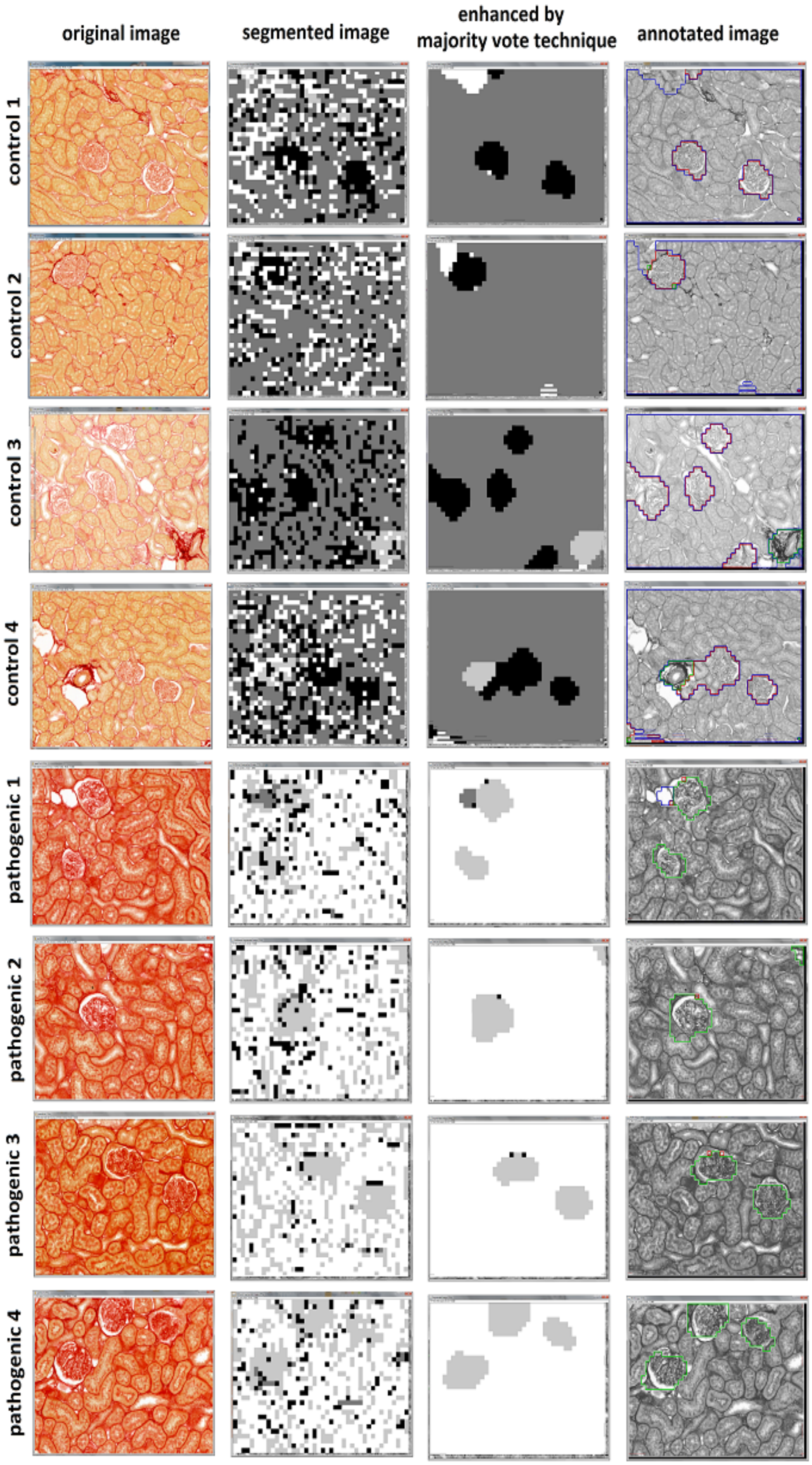


Figure 106. *Automated Segmentation and Annotation Results*

Table 10. Test of the Implemented Model

Image ID	Number of valid Salient Objects	Salient Objects Detected	Area Involved	Current run's accuracy
control1	2	2 * non-Pathogenic Glomerulus	6%	100%
control2	1	1 * non-Pathogenic Glomerulus	4%	100%
control3	5	4 * non-Pathogenic Glomerulus	10%	80%
		1 * Pathogenic Glomerulus	4%	
control4	2	2 * non-Pathogenic Glomerulus	10%	100%
pathogenic1	2	2 * Pathogenic Glomerulus	6%	100%
pathogenic2	2	2 * Pathogenic Glomerulus	6%	50%
pathogenic3	2	2 * Pathogenic Glomerulus	7%	100%
pathogenic4	3	3 * Pathogenic Glomerulus	6%	100%

Table 11. Workflow Accuracy, Specificity and Sensitivity

		Condition (as determined by "Ground Truth")		Predictive Values (%)
		Non-Pathogenic Glomerulus	Pathogenic Glomerulus	
Test Outcome	Non-Pathogenic Glomerulus	9	0	100
	Pathogenic Glomerulus	1	10	90.9
		90	100	95
		Sensitivity (%)	Specificity (%)	Total Accuracy (%)

5.2 Breast Cancer Biopsy Quantification Results

In order to represent the proposed approach describe in Chapter 4.4, we constructed the following workflow scheme (see Figure 107) in the Taverna Workbench, utilizing the operators of the developed framework. Firstly, we will evaluate the block-based segmentation approach, which succeeded better results and published as a novel accurate technique. Afterwards, the mean-shift based approach results will be presented. The mean-shift based approach was developed experimentally during the accomplishment of the Ph.D. thesis.

5.2.1 Evaluation of the block based segmentation approach

In order to evaluate the performance of the proposed tool, we will evaluate the proposed approach based on ground truth, provided by our pathologist experts. Before moving to the testing of the tool, the characterization ability of the proposed tool must be evaluated. Thus, feature extraction - based on the features analyzed in additional characterization section - was performed on each image of the dataset, through the proposed tool. Feature extraction based on the additional features was performed on each image of the dataset, using the proposed tool.

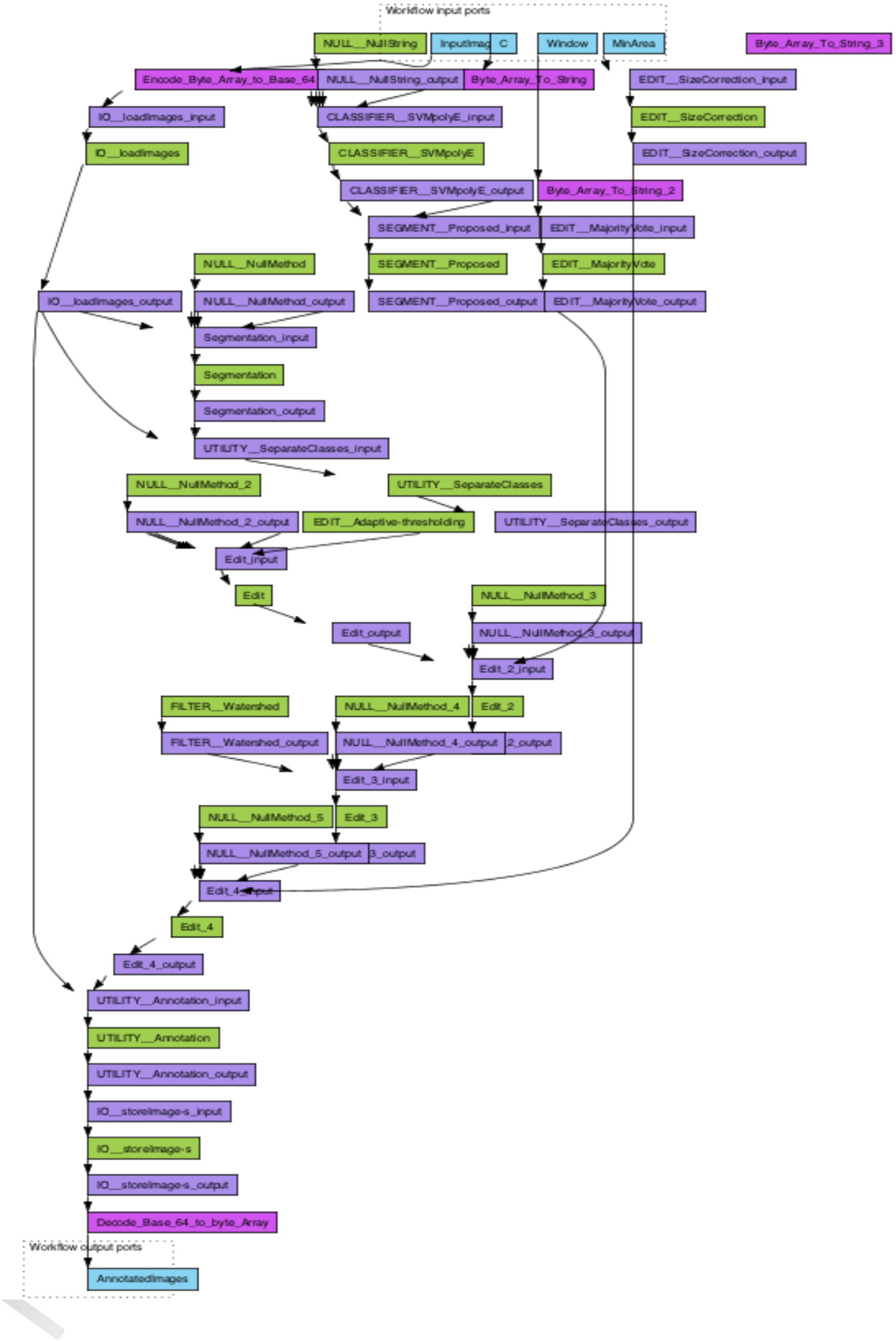


Figure 107. Workflow scheme for the breast cancer biopsies quantification and characterization, built in Taverna – utilizing the developed framework's operators.

Table 12. Extracted characterization features of cancer cells from each treatment group (mean values and standard deviation)

Treatment	Control		IR		EB 1089		IR + EB 1089	
	Mean	S. Deviation	Mean	S. Deviation	Mean	S. Deviation	Mean	S. Deviation
C/A factor(Cancer)	1,76	0,42	1,65	0,78	0,36	0,15	0,06	0,01
Mean R(Cancer)	91,75	4,72	87,02	2,93	72,57	7,30	61,97	16,77
Mean G(Cancer)	74,24	2,50	70,81	2,02	60,00	4,78	45,88	14,19
Mean B(Cancer)	64,85	0,71	57,27	2,49	50,08	3,35	40,11	9,74
Area(Cancer)	252,21	31,20	236,11	7,05	216,13	21,81	180,09	28,31
Mean(Cancer)	76,95	2,28	71,70	2,37	60,88	5,08	49,31	13,37
Std. Dev.(Cancer)	34,32	1,32	36,77	1,59	33,94	1,75	36,79	1,85
Mode(Cancer)	61,97	1,37	50,70	5,85	38,19	7,34	22,79	13,48
Min(Cancer)	14,25	1,47	11,06	3,24	11,52	4,23	8,55	7,35
Max(Cancer)	171,22	7,71	176,08	2,26	163,59	8,22	158,41	6,99
Perim.(Cancer)	61,87	4,10	59,83	1,24	56,89	3,09	51,34	3,82
Major(Cancer)	20,89	1,23	20,41	0,50	19,67	1,03	18,00	1,57
Minor(Cancer)	14,18	0,71	13,83	0,18	13,29	0,51	12,23	0,99
Angle(Cancer)	86,70	2,34	90,10	5,02	93,83	3,97	94,61	14,89
Circ.(Cancer)	0,79	0,02	0,79	0,01	0,81	0,01	0,83	0,03
Int. Den.(Cancer)	19500,49	2543,98	16789,31	796,94	13178,02	2252,82	8771,43	1368,46
Median(Cancer)	73,78	2,07	66,44	3,14	53,53	5,36	37,71	17,90
Skewness(Cancer)	0,39	0,06	0,55	0,06	0,78	0,06	0,74	1,07
Kurtosis(Cancer)	-0,07	0,10	0,01	0,11	0,39	0,28	7,76	14,87
%Area(Cancer)	24,23	3,58	20,42	4,32	8,05	2,81	1,10	0,18
RawIntDen(Cancer)	19500,49	2543,98	16789,31	796,94	13178,02	2252,82	8771,43	1368,46
AR(Cancer)	1,49	0,03	1,49	0,04	1,48	0,05	1,51	0,09
Round(Cancer)	0,70	0,01	0,70	0,01	0,70	0,02	0,70	0,03
Solidity(Cancer)	0,88	0,00	0,88	0,00	0,88	0,01	0,89	0,01
Cells(Cancer)	188,60	11,70	170,20	36,13	72,20	19,63	12,40	3,29

Table 13. Extracted characterization features of cancer cells from each treatment group (mean values and standard deviation)

Treatment	Control		IR		EB 1089		IR + EB 1089	
	Mean	S. Deviation	Mean	S. Deviation	Mean	S. Deviation	Mean	S. Deviation
Mean R(Apoptotic)	122,52	4,05	121,94	1,87	118,04	3,38	110,10	5,13
Mean G(Apoptotic)	127,88	4,43	134,45	1,22	134,26	2,78	118,53	3,50
Mean B(Apoptotic)	148,03	6,78	157,62	2,48	162,72	2,28	146,98	3,87
Area(Apoptotic)	192,01	28,30	168,69	10,51	264,18	38,78	288,04	23,32
Mean(Apoptotic)	132,82	4,98	138,01	1,31	138,34	2,69	125,20	2,08
Std. Dev.(Apoptotic)	24,98	2,07	24,53	1,09	24,96	2,76	28,10	2,97
Mode(Apoptotic)	128,61	6,54	133,61	1,82	132,15	3,84	115,77	3,30
Min(Apoptotic)	73,28	8,58	78,27	4,86	78,52	10,14	65,10	6,96
Max(Apoptotic)	190,70	3,28	195,31	1,51	200,43	5,43	193,67	4,20
Perim.(Apoptotic)	53,66	4,68	50,24	1,94	63,36	5,15	65,22	2,73
Major(Apoptotic)	18,45	1,53	17,41	0,71	21,36	1,64	21,90	0,83
Minor(Apoptotic)	12,29	0,57	11,61	0,35	13,96	0,94	14,58	0,61
Angle(Apoptotic)	89,16	6,03	90,25	10,42	83,25	4,46	87,21	3,17
Circ.(Apoptotic)	0,80	0,03	0,81	0,01	0,77	0,02	0,77	0,01
Int. Den.(Apoptotic)	25415,74	3180,71	23242,43	1240,78	36525,25	5404,13	36091,56	2323,89
Median(Apoptotic)	132,62	5,62	138,14	1,85	137,40	2,65	123,37	2,39
Skewness(Apoptotic)	-0,01	0,09	-0,02	0,11	0,21	0,14	0,26	0,12
Kurtosis(Apoptotic)	-0,38	0,41	-0,15	0,71	-1,98	2,94	-1,40	2,46
%Area(Apoptotic)	11,20	4,18	9,85	2,91	27,53	6,57	31,02	2,22
RawIntDen(Apoptotic)	25415,74	3180,71	23242,43	1240,78	36525,25	5404,13	36091,56	2323,89
AR(Apoptotic)	1,50	0,06	1,51	0,03	1,54	0,02	1,52	0,02
Round(Apoptotic)	0,69	0,03	0,69	0,01	0,67	0,00	0,68	0,01
Solidity(Apoptotic)	0,86	0,01	0,87	0,00	0,86	0,00	0,87	0,01
Cells(Apoptotic)	113,00	31,83	114,40	31,42	202,60	27,22	212,40	16,77

The mean values of the exported features of cancer and apoptotic cells are depicted in Table 12 and 13 respectively. Thus, in order to check the possible classes of the exported features, we imported them into a novel-clustering framework [10]. The proposed clustering framework assigns a label to each image of each treatment group. It predicted that there are 4 possible classes and achieved 85% accuracy in the labeling of each instance.

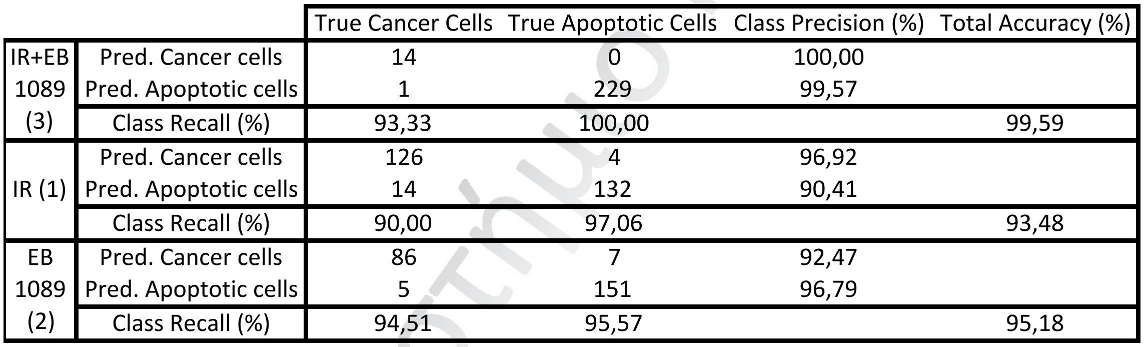
SVM classifier, with Polynomial kernel and a Complexity Constant of 1.5 (selected heuristically), was utilized for the initial segmentation and separation of the cancer and apoptotic cells.

Support Vector Machines' most common kernel, is the polynomial kernel, which exports a similarity metric for the training samples in a feature space over polynomials of the original variables, enabling the training process of non-linear models. Thus, the polynomial kernel looks at the provided features of the dataset along with their combinations between them.

$$K(x, y) = (< x, y > +1)^p \quad (5)$$

Complexity Constant represents a cost parameter constant that controls the number of support vectors and allows the control of the trade-off, between learning error and model complexity by the user. Due to the fact that the complexity is assumed during the training process, there is not much risk of over-fitting the training data in SVM.

Table 14. Confusion matrices of the three images (IR+EB 1089(3), IR(1) and EB 1089(2)).



		True Cancer Cells	True Apoptotic Cells	Class Precision (%)	Total Accuracy (%)
IR+EB 1089 (3)	Pred. Cancer cells	14	0	100,00	
	Pred. Apoptotic cells	1	229	99,57	
	Class Recall (%)	93,33	100,00		99,59
IR (1)	Pred. Cancer cells	126	4	96,92	
	Pred. Apoptotic cells	14	132	90,41	
	Class Recall (%)	90,00	97,06		93,48
EB 1089 (2)	Pred. Cancer cells	86	7	92,47	
	Pred. Apoptotic cells	5	151	96,79	
	Class Recall (%)	94,51	95,57		95,18

The application of the developed methodology in 15 images is depicted in Figure 108. The detection and quantification results are presented in Table 16. As depicted in Table 14, for three images (one of each type), the proposed tool achieved an average of 96.08% accuracy. Likewise, the tool proved sufficient in the testing of all the images of the dataset, achieving a 95.31% overall accuracy (see Table 15). Some miscounts may occur, due to the existence of extremely merged cells (which look like one round entity), and the watershed algorithm cannot separate them. The chromatic similarity of artifacts may cause, in rare cases, minor misclassification issues. The performed evaluation is in coincidence with the results reported in [11], indicating also the same efficiency of each treatment. In Fig. 7, only the cancer cells are annotated, in order to be easier for the physician to obtain an objective perception about the pathogenesis.

Table 15. Confusion matrix for all images (Mass Recognition / Quantification).

ALL TREATMENT IMAGES	True Cancer Cells	True Apoptotic Cells	Class Precision (%)	Total Accuracy (%)
Pred. Cancer cells	1172	102	91,99	
Pred. Apoptotic cells	82	2565	96,90	
Class Recall (%)	93,46	96,18		95,31

Table 16. Results of the Tool runs in treatment groups: IR (Radiation), EB 1089 (Drug) and EB 1089 + IR combined.

Image	IR + EB 1089 (1)	IR + EB 1089 (2)	IR + EB 1089 (3)	IR + EB 1089 (4)	IR + EB 1089 (5)	IR (1)	IR (2)	IR (3)
Cancer Cells	8	14	14	16	10	130	133	188
Apoptotic Cells	196	220	230	223	193	146	146	112
Image	**IR (4)**	**IR (5)**	**EB 1089 (1)**	**EB 1089 (2)**	**EB 1089 (3)**	**EB 1089 (4)**	**EB 1089 (5)**	
Cancer Cells	208	192	45	93	81	59	83	
Apoptotic Cells	77	91	207	156	206	224	220	

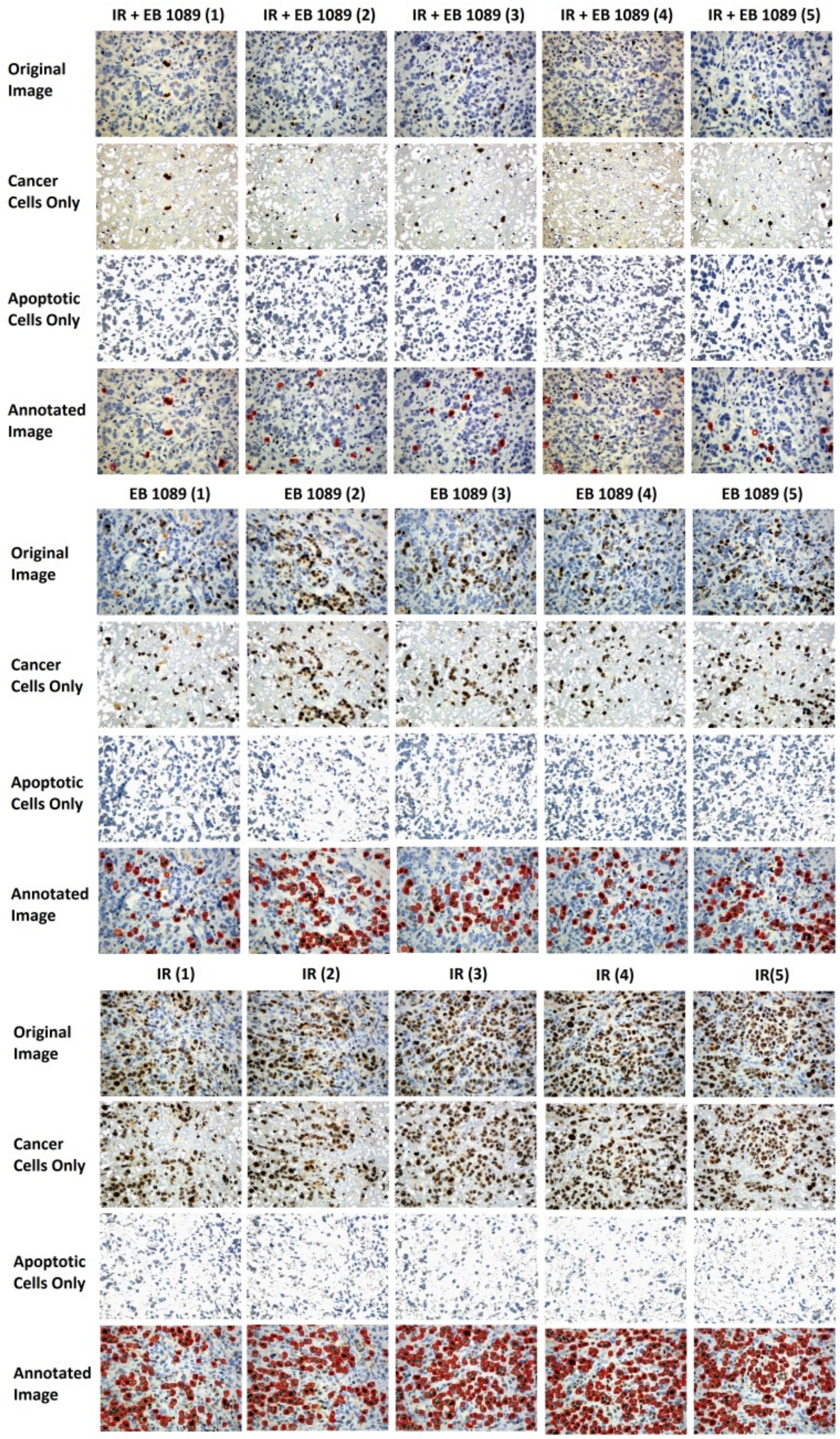
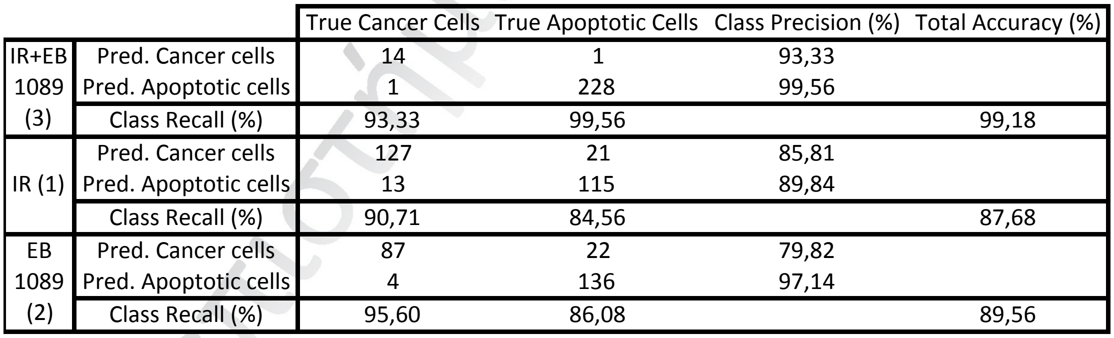


Figure 108. *Visualized Automatic Cell Recognition and Quantification Results.*

5.2.2 Evaluation of the mean-shift based approach

Although the above satisfying approach, the mean shift approach provides less accurate, but still descend results and also quite faster application of the necessary procedures for the later approach. More specifically, the required average time - for the proposed tool to complete all the necessary procedures on the same machine for an image size of 512x384 pixels - is 2.8 seconds, without proper multi-threading development of the application, due to the already low processing time. Additionally, the optimal value of the spatial radius variable is 5, white the optimal value for the color distance variable is 26. The optimal values for the specific issue were found heuristically. The workflow scheme for the specific approach is presented in Figure 109.

Table 17. Confusion matrices of the three images (IR+EB 1089(3), IR(1) and EB 1089(2)) – Mean Shift Approach.



		True Cancer Cells	True Apoptotic Cells	Class Precision (%)	Total Accuracy (%)
IR+EB 1089 (3)	Pred. Cancer cells	14	1	93,33	
	Pred. Apoptotic cells	1	228	99,56	
	Class Recall (%)	93,33	99,56		99,18
IR (1)	Pred. Cancer cells	127	21	85,81	
	Pred. Apoptotic cells	13	115	89,84	
	Class Recall (%)	90,71	84,56		87,68
EB 1089 (2)	Pred. Cancer cells	87	22	79,82	
	Pred. Apoptotic cells	4	136	97,14	
	Class Recall (%)	95,60	86,08		89,56

As depicted in Table 17, for three images (one of each type), the proposed tool achieved an average of 92.1% accuracy. Likewise, the tool proved sufficient in the testing of all the images of the dataset, achieving 89.93% overall accuracy (see Table 18). The few misclassifications issues lie to the fact that the very dark blue-apoptotic cells may be classified as cancer cells, due to the nature of mean shift classifier combined to median-cut color quantization algorithm, which sets them as the darker color foreground object, which implies to cancer cells. The performed evaluation is also in coincidence with the results reported in [14], indicating also the same efficiency with an insignificant deviation of each treatment. In Figure 110, only the cancer cells are annotated, in order to be easier for the physician to obtain an objective perception about the pathogenesis.

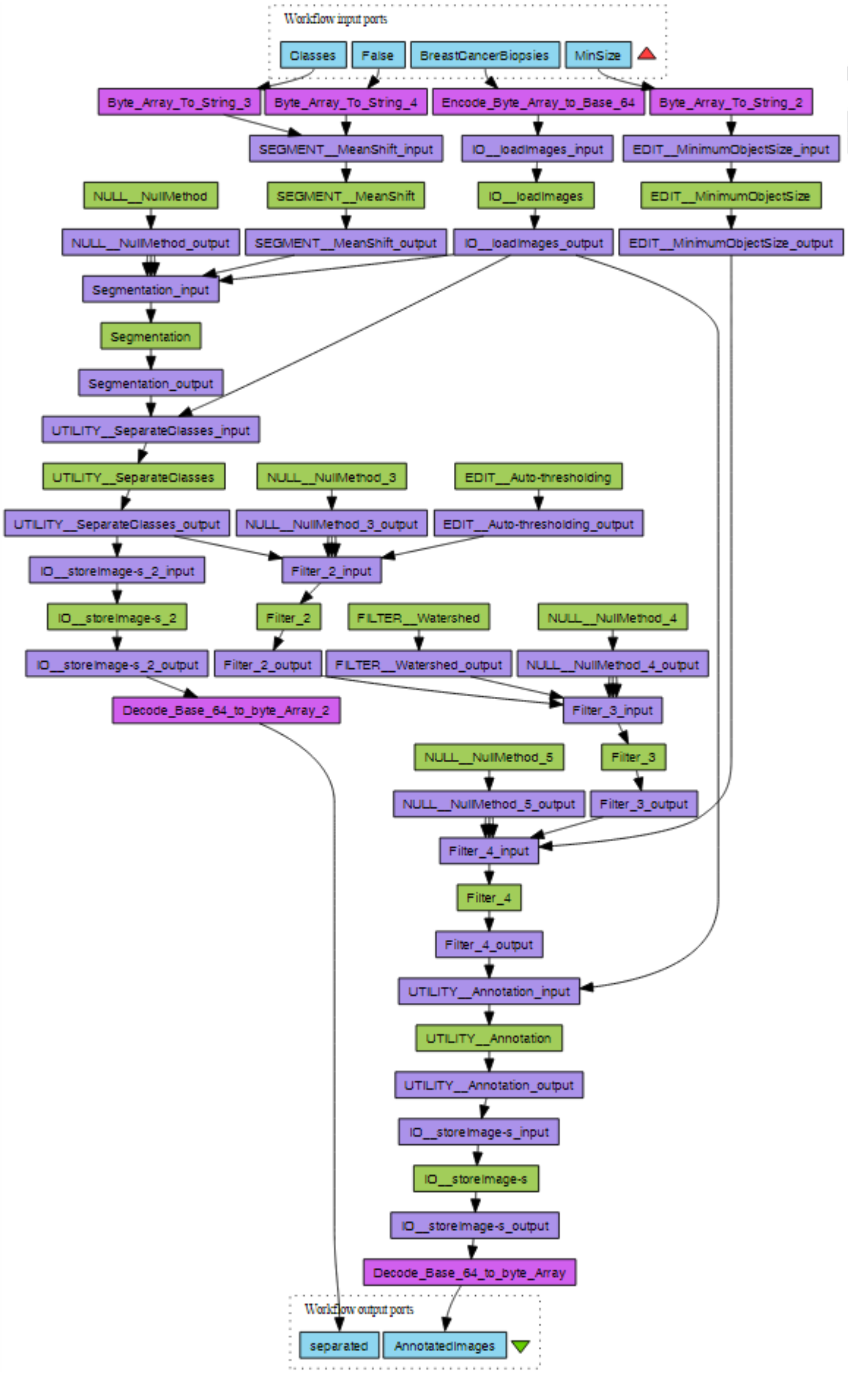


Figure 109. Workflow scheme for the mean-shift based approach

Table 18. **Confusion matrix for all the images (Mass Recognition / Quantification) – Mean Shift Approach.**

ALL TREATMENT IMAGES	True Cancer Cells	True Apoptotic Cells	Class Precision (%)	Total Accuracy (%)
Pred. Cancer cells	1181	322	78,58	
Pred. Apoptotic cells	73	2345	96,98	
Class Recall (%)	94,18	87,93		89,93

Table 19. **Results of the Tool runs in treatment groups: IR (Radiation), EB 1089 (Drug) and EB 1089 + IR combined – Mean Shift Based.**

Image	IR + EB 1089 (1)	IR + EB 1089 (2)	IR + EB 1089 (3)	IR + EB 1089 (4)	IR + EB 1089 (5)	IR (1)	IR (2)	IR (3)
Cancer Cells	18	16	16	26	11	148	149	208
Apoptotic Cells	186	218	228	213	192	128	130	92
Image	**IR (4)**	**IR (5)**	**EB 1089 (1)**	**EB 1089 (2)**	**EB 1089 (3)**	**EB 1089 (4)**	**EB 1089 (5)**	
Cancer Cells	239	209	59	109	113	81	101	
Apoptotic Cells	46	74	193	140	174	202	202	

It can also be noticed that the mean shift's misclassification issues contain mainly false positive instances and very few false negative (see Figure 111). This lies to the fact that in some cases, the dark pixels of blue cells are wrongfully equalized to the dark pixels of brown cells. Despite the fact that there are some misclassification issues, the mean shift approach, provides quite effective results that do not change dramatically the cancer ratio on each image.

The output results of the two approaches are almost identical, as it can be depicted in Figure 110. The misclassification issues are larger at IR treatment images, which are the images, which contain an increased number of cancer cells and may cause a minor confusion to the mean shift algorithm.

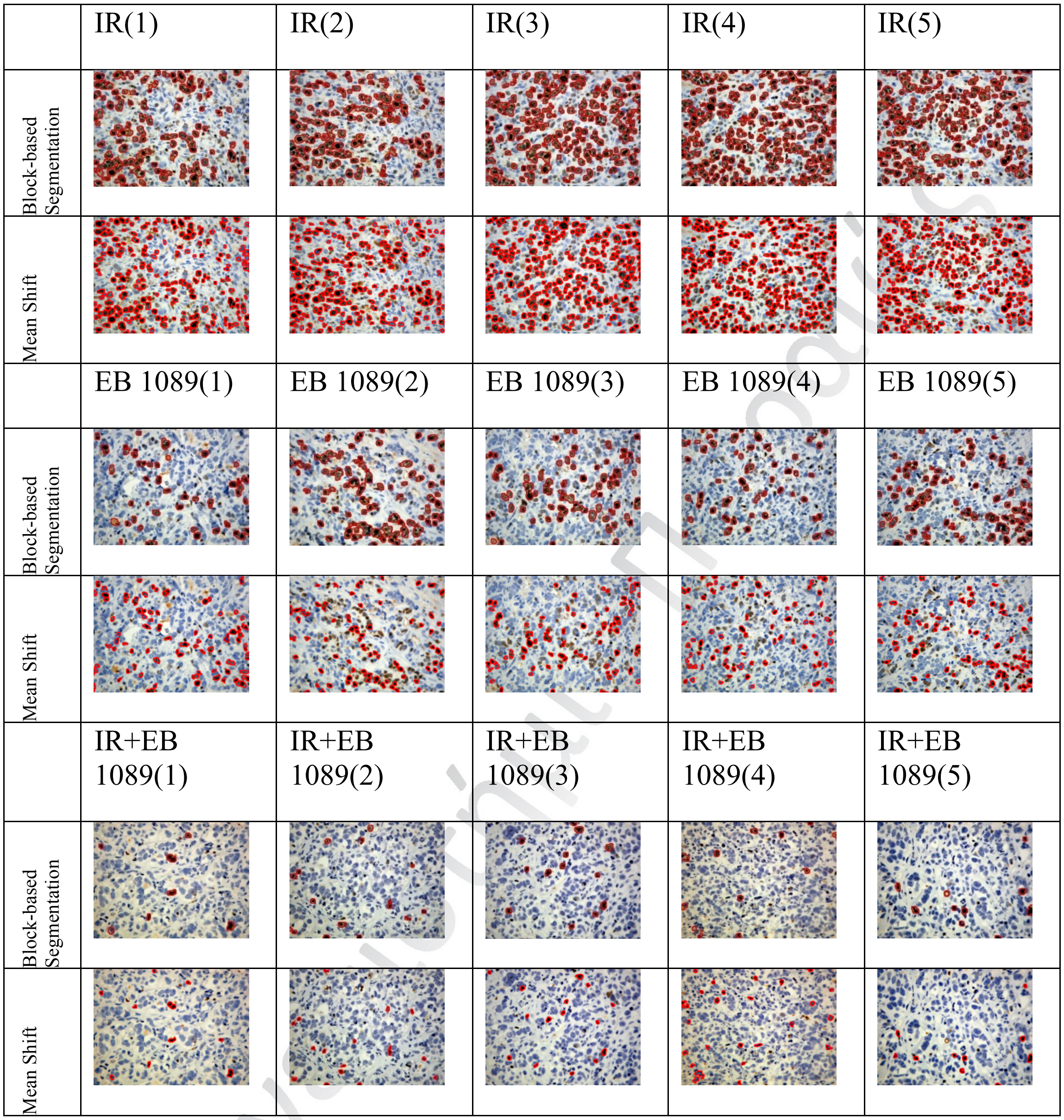


Figure 110. Comparison between the 2 proposed approaches. In the first row of each treatment the annotation achieved by the block based approach is presented. In the second row of each treatment the achieved annotation of the mean shift approach is presented. The mean shift approach appears to annotate more false positive cancer cells.

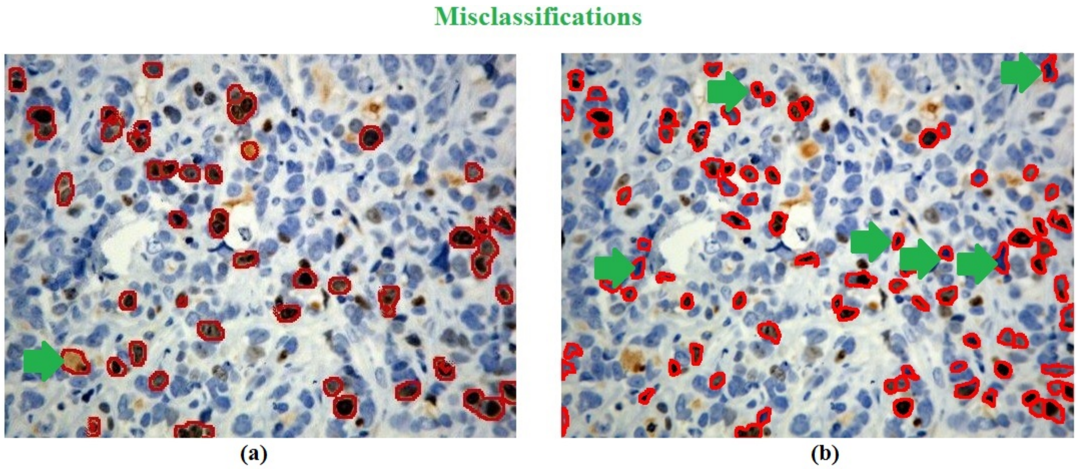


Figure 111. Classification issues of the 2 proposed approaches. In (a) an example of the block based approach is presented. The issue occurs due to the chromatic similarity to the dark brown of the cancer cells. In (b) an example of the mean shift approach is presented. In this case the issue occurs, due to the dark color of the apoptotic cells that misclassifies them as cancer cells.

5.3 Skin Lesion Detection and Annotation

The case study deals with a two-class problem, since a pixel may belong to the melanoma area or the background area. The proposed approach will be evaluated through its application on 20 skin lesion images. These images will be used as an input to the workflow scheme that will deal with the annotation of the salient objects (the melanoma objects in our case). The skin cancer pathology was chosen as a use case for the evaluation part, because it is one of the most frequent types of cancer that needs to be early predicted, in order to avoid severe health issues.

In Figure 112, the entire workflow scheme of the proposed approach, designed in Taverna workflow manager, is depicted.

5.3.1 Background Subtraction and Mean Shift Segmentation Settings

The input arguments for the background subtraction and mean shift segmentation operators were set heuristically and the optimal settings are described below.

The background subtraction algorithm provides a clearer view of the skin lesion area, removing the background noise. The rolling ball radius variable was set at 50 pixels, while it was calibrated to process images with light background. Additionally, the optimal value of the spatial radius, of the mean shift algorithm variable is 5, while the optimal value for the color distance variable is 26.

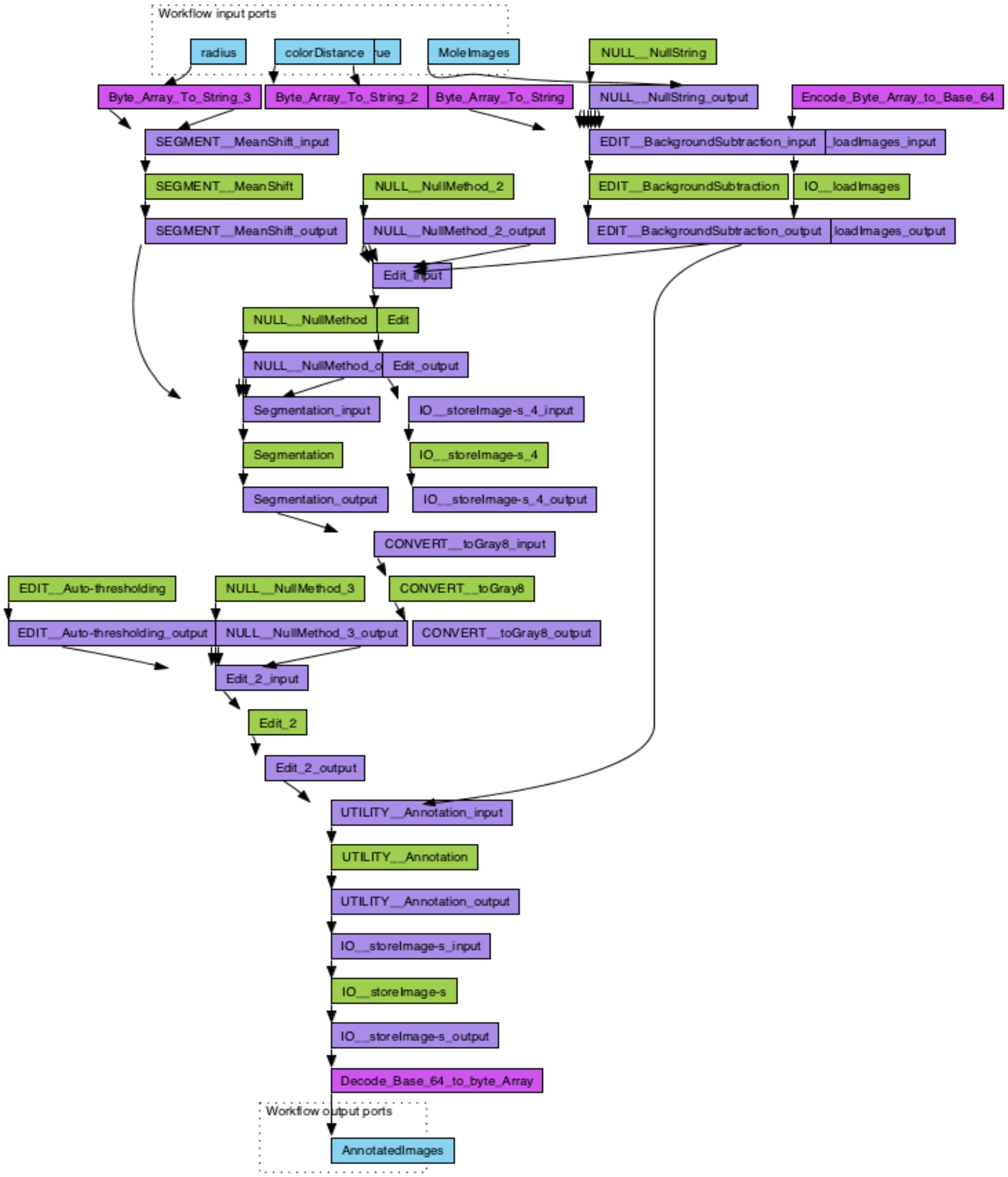


Figure 112. Workflow of the proposed approach for detecting and annotating the lesion areas from skin cancer images.

5.3.2 Evaluation of the proposed workflow scheme (Matthews Correlation Coefficient)

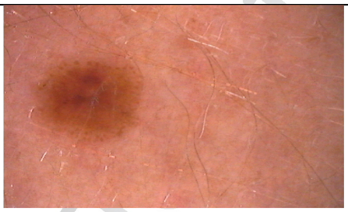
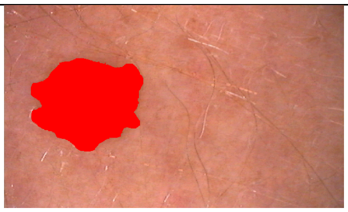
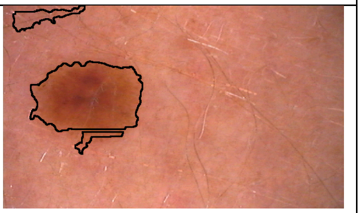
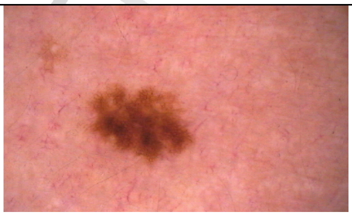
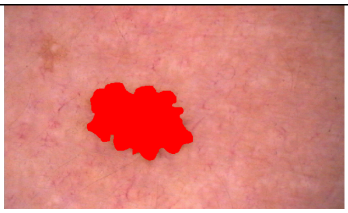
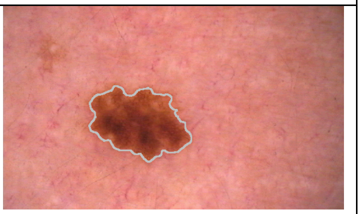
The size of the skin lesion is not defined. Thus, Matthews Correlation Coefficient [12] was utilized, because it is not affected by the possible large amount of true negative pixels.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (7)$$

The metrics of accuracy, sensitivity and specificity were also utilized.

5.3.3 Skin lesion detection and annotation results

The examined image dataset was obtained from (http://www.dermoncology.com/). The dataset contained about 3000 images, 800 of them containing melanoma, while the rest of them containing dysplastic nevus and benign nevi. 20 random images were used from the melanoma set, so as to evaluate the proposed methodology, as data input on the constructed workflow scheme. In Figure 113 are depicted 6 of these images, representing the worst, the best and the average case scenarios.

	Original Image	Ground Truth	Annotation
8.bmp			
13.bmp			

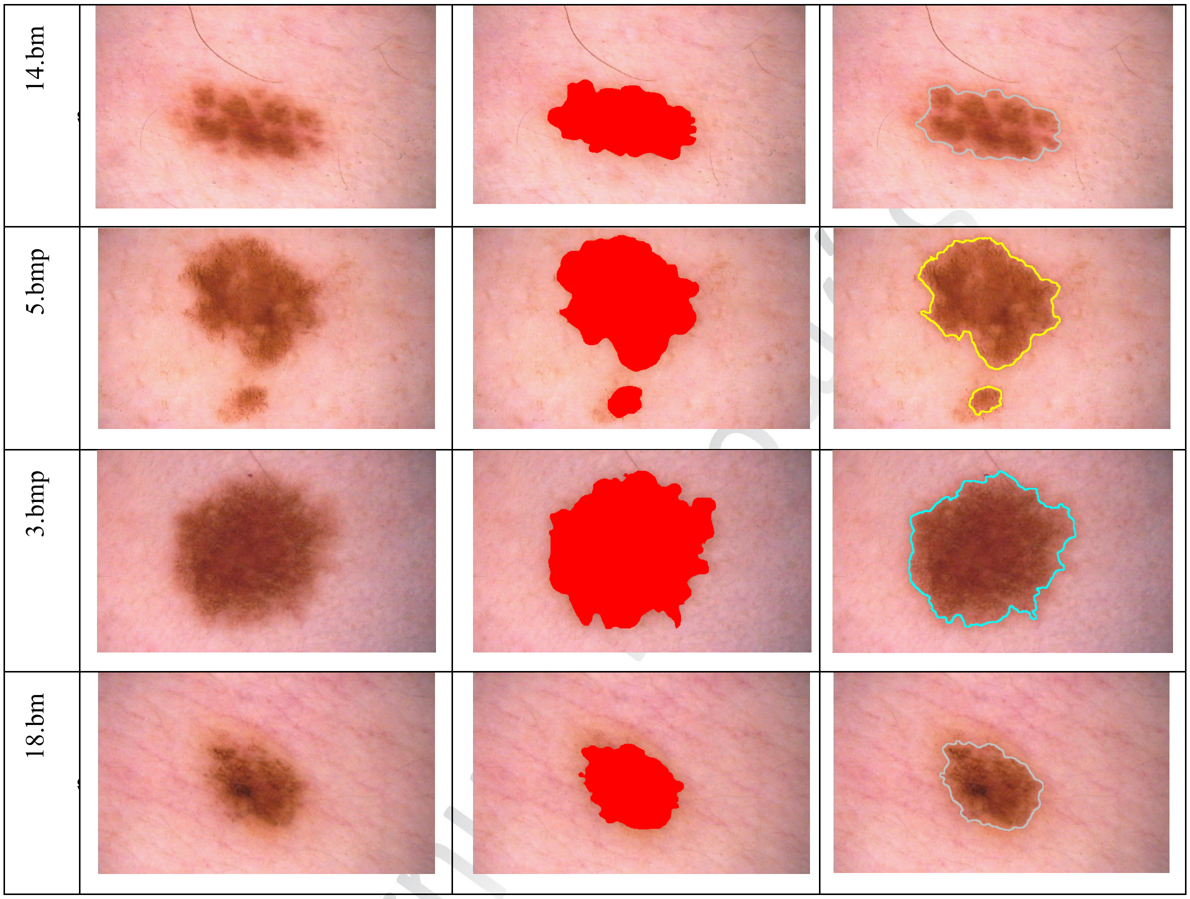


Figure 113. The first column depicts the original images, the second one the ground truth provided by the expert physicians and the third one the annotation achieved by the proposed approach

It can easily be noticed that the algorithm achieved a satisfying performance. The evaluation results for each examined image are presented in Table 20. It can be noticed that even the worst case (Image 8.bmp), achieved satisfying detection results. The most rigorous metric, Matthews Correlation Coefficient, achieved sufficient performance reaching an average percentage of 93.74% (see Table 21). The proposed workflow scheme was proved capable of dealing with such images, filtering artifact objects, such as hair and detecting the exact lesion area.

Table 20. **Evaluation of the proposed approach for each imported image**

Image name	Accuracy (%)	Sensitivity (%)	Specificity (%)	Matthews CC (%)
0.bmp	98.98	93.29	99.91	95.74
1.bmp	97.62	93.14	99.46	94.21
2.bmp	99.38	94.84	99.90	96.60
3.bmp	97.55	91.52	99.88	93.90
4.bmp	98.00	92.56	99.59	94.24
5.bmp	97.96	90.21	99.89	93.56
6.bmp	98.97	95.84	99.36	94.84
7.bmp	97.55	92.67	99.56	94.05
8.bmp	97.17	85.11	98.65	85.23
9.bmp	97.34	89.08	99.86	92.53
10.bmp	97.98	92.40	98.84	91.30
11.bmp	98.13	95.67	98.96	95.02
12.bmp	99.25	93.96	99.87	95.98
13.bmp	99.05	89.85	99.85	93.37
14.bmp	98.65	89.63	99.88	93.49
15.bmp	97.82	88.53	99.97	92.78
16.bmp	98.18	97.14	98.35	92.87
17.bmp	99.05	92.56	99.87	95.16
18.bmp	99.43	97.39	99.64	96.60
19.bmp	97.54	96.42	97.90	93.33

Table 21. **The average performance of the proposed workflow scheme**

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Matthews CC (%)
Average	98.28	92.59	99.46	93.74
Standard Deviation	0.74	3.20	0.60	2.44

5.4 Optimal Scheme proposal through comparison with Ground Truth

In this case scenario, accurate detection and annotation of skin lesion areas will be achieved through the utilization of the proposed framework. The constructed workflow scheme (see Figure 114) will generate multiple combinations of a number of image analysis and segmentation algorithms (see Figure 115). The optimal scheme will be exported after the processing of 113 skin lesion images and the input of the corresponding ground truth dataset.

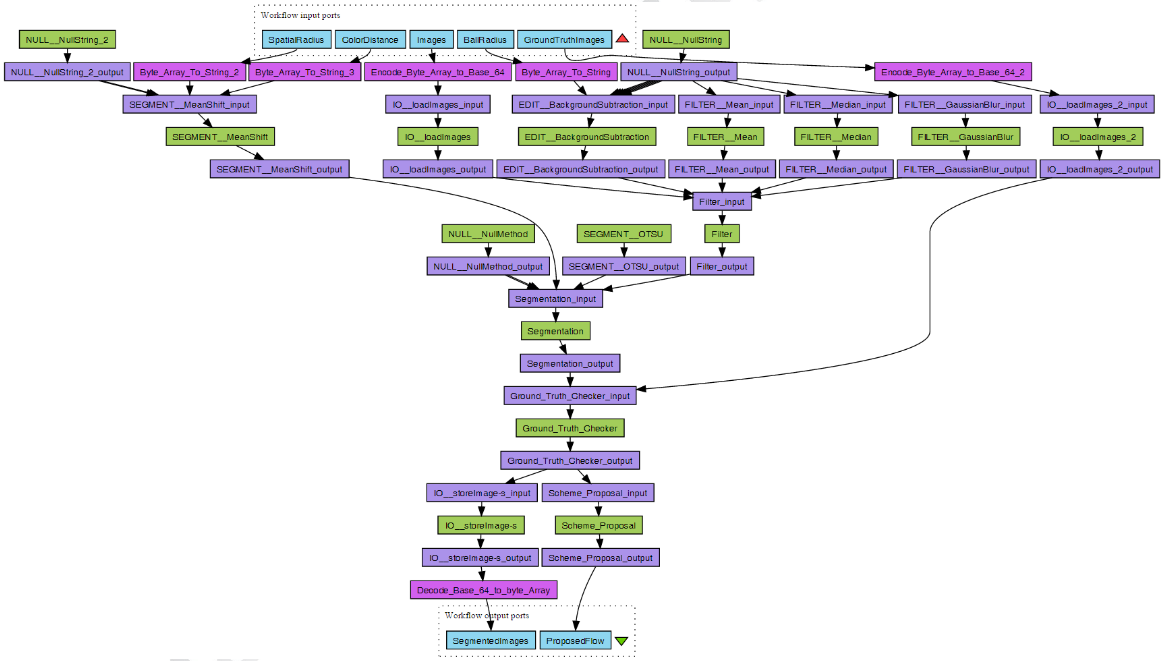


Figure 114. The constructed workflow scheme in TAVERNA workflow manager.

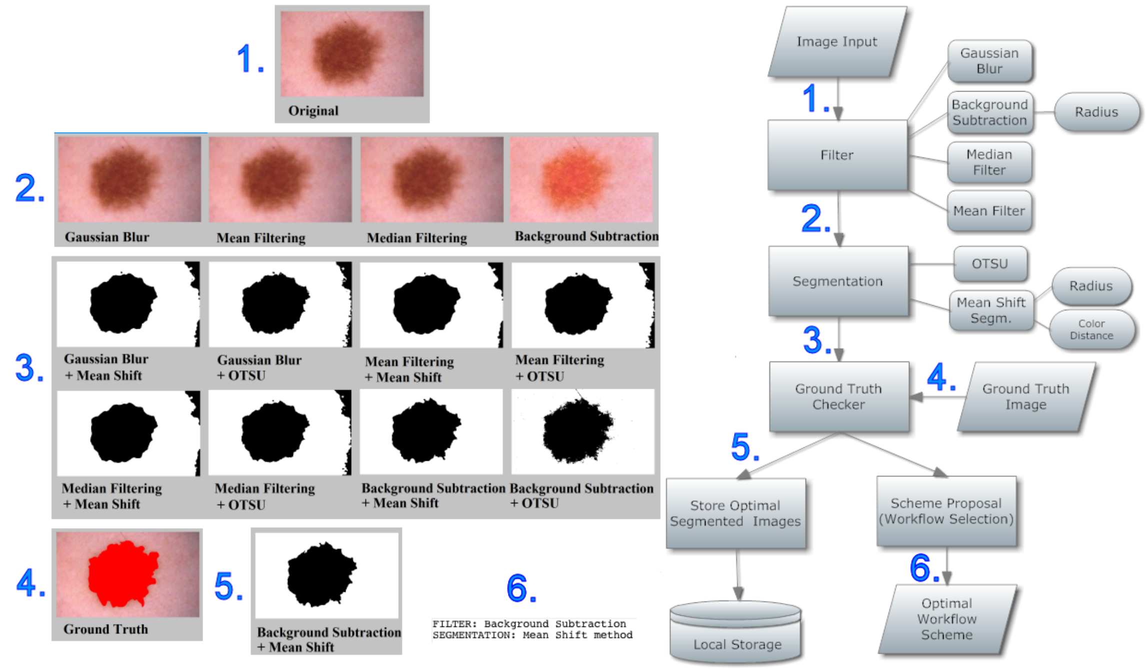


Figure 115. The theoretical architecture of the constructed workflow in Figure 114, containing the image processing outputs of each integrated operator.

5.4.1 Filter

As a first step of the constructed methodology, a pre-processing of the input images is applied. Each of the examined images contains one salient object, in most cases, but their background may also contain some noise. The meaning of the noise term is the uneven illumination of the background, the few uncut hair and some artefact objects (pimples, blains, etc.) of high chromatic similarity with the skin lesion. Thus, makes necessary the removal of the above-mentioned noise, in order to achieve clear segmentation results. In this example, four filtering methods were integrated in the first filtering operator of the developed workflow scheme. These are the Gaussian Blur filter, the Background Subtraction algorithm, the median filter and the mean filter operators. Therefore, the corresponding operators are integrated in the main filtering operator. Each additional operator holds its own arguments. As already defined above, by setting an operator parameter as null, generates the default value of this parameter. The default values are different for each operator. In this case, the only parameter that was customized from these procedures is the pixel radius of the Background Subtraction algorithm. Its value was set at 50 pixels. The background algorithm subtraction method is based on the idea of the ‘rolling ball’ algorithm described in [13].

5.4.2 Segmentation

Segmentation operator receives the output of the filtering operator, which contains four different versions of each image imported to the constructed workflow. In this case, two segmentation method operators will be integrated in the core segmentation operator. These are the well-known Otsu method and the Mean Shift [14] segmentation algorithm. In order to provide a user-friendly framework, the mean shift operator provides an additional option about the result of the segmentation. Although the multicolored output of the mean shift algorithm, the operator contains the option to provide binary results in cases of two-class problems. The current scenario presents a two-class case. The default value for this feature provides the binary version of the mean shift algorithm.

5.4.3 Ground Truth Checker and Scheme proposal operators

The segmented versions of each of the imported images are utilized as input for the Ground Truth Checker operator. The second mandatory input gate requires the corresponding image dataset, containing the ground truth annotation of the salient objects on each of its images. The operator calculates the MCC metric for each version of the segmented images. The image that succeeds the greater MCC value is selected as optimal and the operator process tracks the processes that have been applied at. The Scheme proposal operator exports the workflow that provides the optimal results.

Table 22. The mean MCC performance of each workflow instance

Workflow instance	MCC (mean) %
Gaussian Blur + Mean Shift	72.18
Gaussian Blur + OTSU	69.13
Mean Filtering + Mean Shift	72.03
Mean Filtering + OTSU	69.8
Median Filtering + Mean Shift	72.07
Median Filtering + OTSU	68.36
Background Subtraction + Mean Shift	93.74
Background Subtraction + OTSU	90.14

5.4.4 Workflow Result

In this constructed workflow scheme, there are 2 tiers of image analysis techniques. This implies that the four Filtering operators and the two segmentation methods generate 8 different workflow instances. Each instance corresponds to a different combination of the above operators. Following the two tiers of data processing, the Ground Truth Checker exports the images with the optimal MCC results (see Table 22). The optimal workflow scheme is the one that utilizes the background subtraction algorithm and the mean shift segmentation method. The corresponding images followed that workflow instance are also exported and stored locally.

5.5 Workflow Evaluation through Clustering Log-Likelihood Distance

In this case scenario, the customized scheme is able to detect cancer and apoptotic cells from a breast cancer image dataset. This dataset was obtained from the National Cancer Institute tumor repository (Frederick, MD) [33].

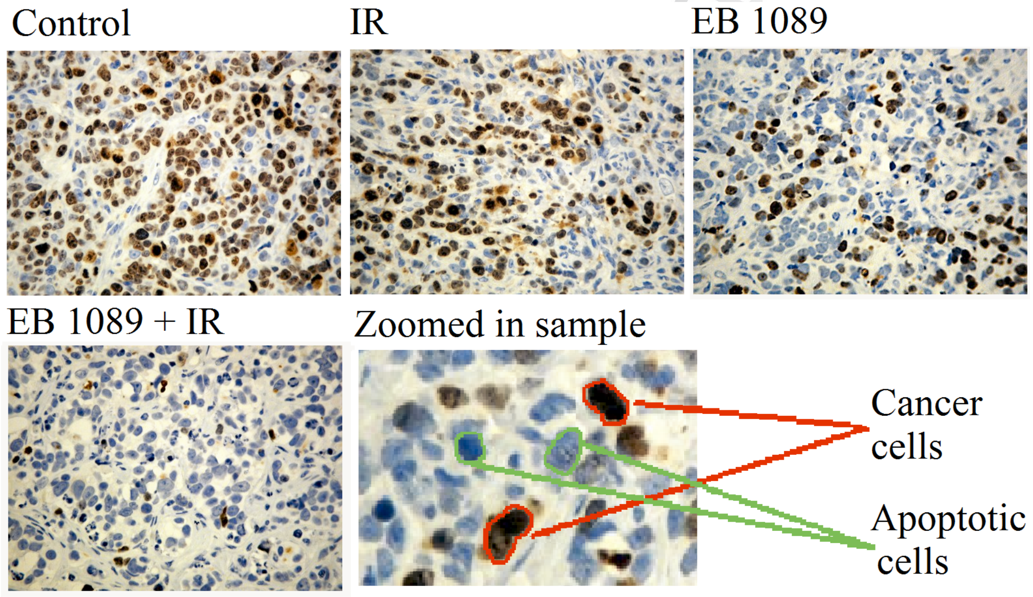


Figure 116. Four treatment groups of cancer cell images and indication of Cancer and Apoptotic Cells

The image dataset is consist of four groups of datasets, Control, IR (Radiation), EB 1089 (Drug) and EB 1089 + IR combined (Described in Section 4.4). As also depicted in Figure 116, cancer and apoptotic cells have a circular form [34] and sometimes are merged, making harder their detection.

Due to the huge number of salient objects in each image (300+), the clustering operator of the proposed framework will be utilized, in order to evaluate a number of workflow instances and propose the optimal one. When ground truth dataset is missing or its creation becomes impossible, the distinction of the most accurately segmented version of an image, via the human eye, is a quite difficult task.

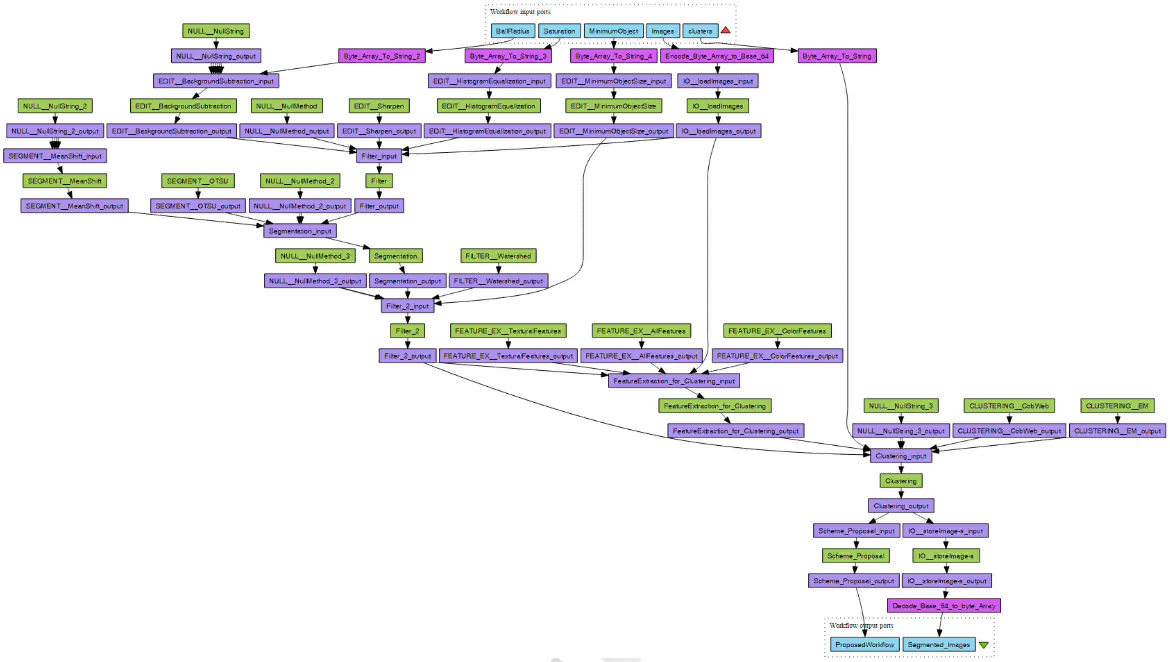
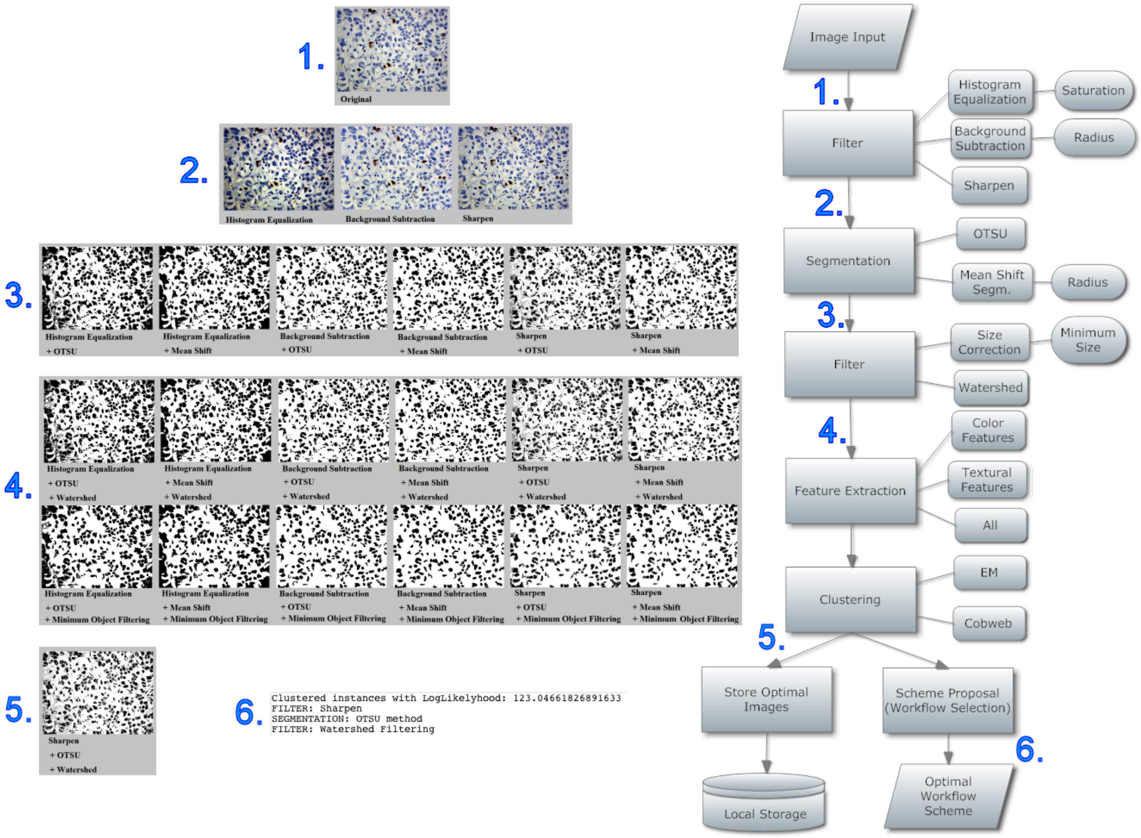


Figure 117. The constructed workflow scheme in TAVERNA workflow manager.

5.5.1 Filtering process

The first step of dealing with this problem is to perform a filtering process. The provided dataset is almost clear from noise, except for some lightning and background issues. So as to acquire clear segmentation results, the salient objects need to be distinguishable. In this case, three segmentation methods for the salient object enhancement were utilized. Histogram equalization, sharpen and background subtraction operators were integrated to the core filtering operator (see Figure 117). Histogram equalization saturation parameter was set at 2, which implies to 2% saturated pixels in the contrast enhancement procedure. Background subtraction spatial radius parameter was set at 50. The sharpen operator contains no arguments or setting, so it required no input.


1.
Original
2.
Histogram Equalization
Background Subtraction
Sharpen
3.
Histogram Equalization + OTSU
Histogram Equalization + Mean Shift
Background Subtraction + OTSU
Background Subtraction + Mean Shift
Sharpen + OTSU
Sharpen + Mean Shift
4.
Histogram Equalization + OTSU + Watershed
Histogram Equalization + Mean Shift + Watershed
Background Subtraction + OTSU + Watershed
Background Subtraction + Mean Shift + Watershed
Sharpen + OTSU + Watershed
Sharpen + Mean Shift + Watershed
Histogram Equalization + OTSU + Minimum Object Filtering
Histogram Equalization + Mean Shift + Minimum Object Filtering
Background Subtraction + OTSU + Minimum Object Filtering
Background Subtraction + Mean Shift + Minimum Object Filtering
Sharpen + OTSU + Minimum Object Filtering
Sharpen + Mean Shift + Minimum Object Filtering
5.
Sharpen + OTSU + Watershed
6.
Image Input
Filter
Histogram Equalization
Saturation
Background Subtraction
Radius
Sharpen
Segmentation
OTSU
Mean Shift Segm.
Radius
Filter
Size Correction
Minimum Size
Watershed
Feature Extraction
Color Features
Textural Features
All
Clustering
EM
Cobweb
Store Optimal Images
Local Storage
Scheme Proposal (Workflow Selection)
Optimal Workflow Scheme

```
Clustered instances with LogLikelyhood: 123.04661826891633
FILTER: Sharpen
SEGMENTATION: OTSU method
FILTER: Watershed Filtering
```

Figure 118. The theoretical architecture of the constructed workflow in Figure 108, containing the image processing results from each operator.

5.5.2 Segmentation

Following the salient object enhancement, the segmentation procedure takes place. Due to the similarity of the two case scenarios, the same segmentation algorithms will be applied. Mean shift and Otsu segmentation operators were integrated to the core segmentation operator.

5.5.3 After segmentation filtering

Even the clear segmentation results from the above processes, some salient objects are merged to each other. In order to avoid the miss-clustering issues of the following operator, additional filtering is required. The watershed filtering operator is applied on all the segmented images, in order to accurately cut the merged particles and provide clear images.

Size is another major factor, except for the color, that characterizes a cancer or apoptotic cell. For that reason, an analyse particles method was developed - similar to a component labelling algorithm [15]. More specifically, it removes each object that is equal or smaller than a specific value (in pixels). Based on the assistance of our expert pathologists, the size of a true cancer cell is interpreted as an approximately 100±15 pixels area on the image. Thus, the minimum size of a valid object is set at 75 pixels.

5.5.4 Feature Extraction, Clustering and Scheme Proposal operators

The feature extraction operator receives the clear segmented images of all the possible combinations of the above operators. In this case, based on the specific operator integration, the core operator will export the textural, the color and the combination of textural and color features. These combinations will feed the clustering operator, in order to evaluate the produced clusters and detect the optimal workflow scheme between all the above combinations.

5.5.5 Workflow Result

Considering the three tiers of the image processing techniques, 12 possible workflow instances can be generated. The evaluation process contains also the ability to generate additional workflow instances, in order to extract an objective performance of each workflow. The workflow that achieved the greatest performance - which corresponds to 123.05 log-likelihood distance (see Table 23) - is the instance that utilizes the sharpen algorithm in the first tier, the Otsu segmentation method in the second and the watershed filtering in the third.

Table 23. The mean log-likelihood distance of each workflow's clustered segmented objects

Workflow Instance	log-likelyhood (mean) distance
Histogram Equlization + OTSU + Watershed	86.01
Histogram Equlization + Mean Shift + Watershed	98.33
Background Subtraction + OTSU + Watershed	113.45
Background Subtraction + Mean Shift + Watershed	114.67
Sharpen + OTSU + Watershed	123.05
Sharpen + Mean Shift + Watershed	100.14
Histogram Equlization + OTSU + Minimum Object Filtering	88.9
Histogram Equlization + Mean Shift + Minimum Object Filtering	111.02
Background Subtraction + OTSU + Minimum Object Filtering	107.55
Background Subtraction + Mean Shift + Minimum Object Filtering	108.89
Sharpen + OTSU + Minimum Object Filtering	115.69
Sharpen + Mean Shift + Minimum Object Filtering	103.22

5.6 Block Detection and Quantification of Fibrotic Areas in Microscopy Images of Obstructive Nephropathy (via scoring system)

In order to evaluate the efficiency of the proposed application tool we tested unknown images of our dataset, containing images from each day of pathogenesis. The segmentation's block width had been selected for the demonstration runs at 200x200, because it provides better resolution results. More specifically, the classification algorithm, which achieved the highest accuracy, rates in both approaches and in all segmentation blocks widths, was the Random Forest (see Table 24).

Table 24. Accuracy of the Classification model for the proposed application

	Accuracy (%)	
segmentation block width	Semgentation Majority Voting approach	Overall Image approach
1000	84.64%	92.50%
600	82.56%	97.50%
200	83.42%	87.50%

So as to estimate the scoring system, we tested it on the images of the dataset in three segmentation block widths. As it can be visualized from Figure 119, segmentation block widths achieved satisfying results. 200x200 block based on segmentation offered clearer and more discrete results, due to the lower standard deviation of the mean values.

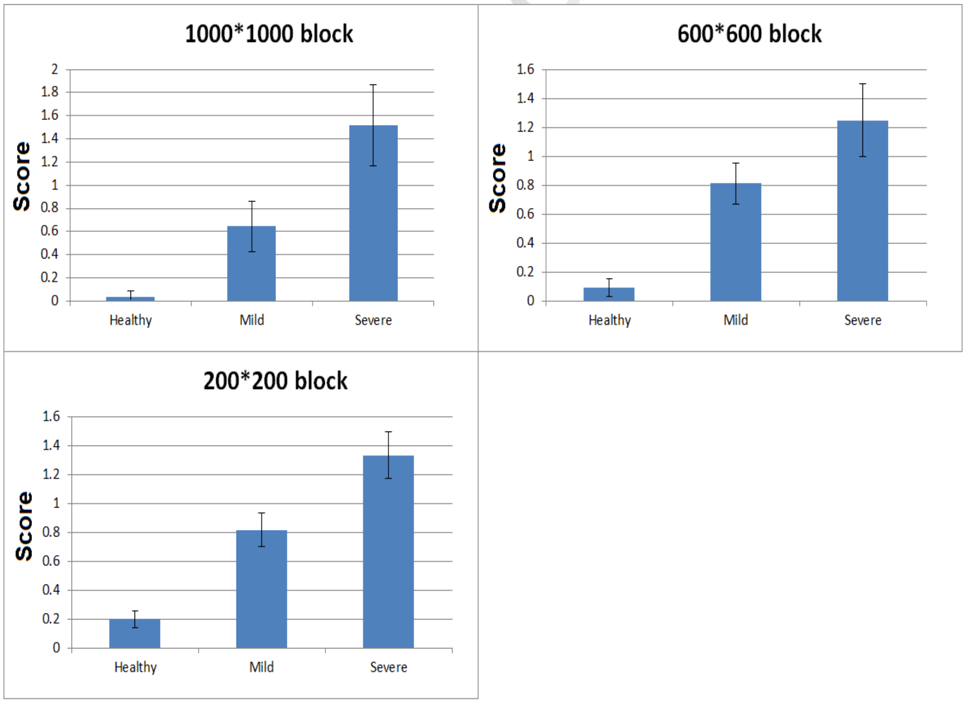


Figure 119. Mean score values for each pathogenesis and each segmentation block width

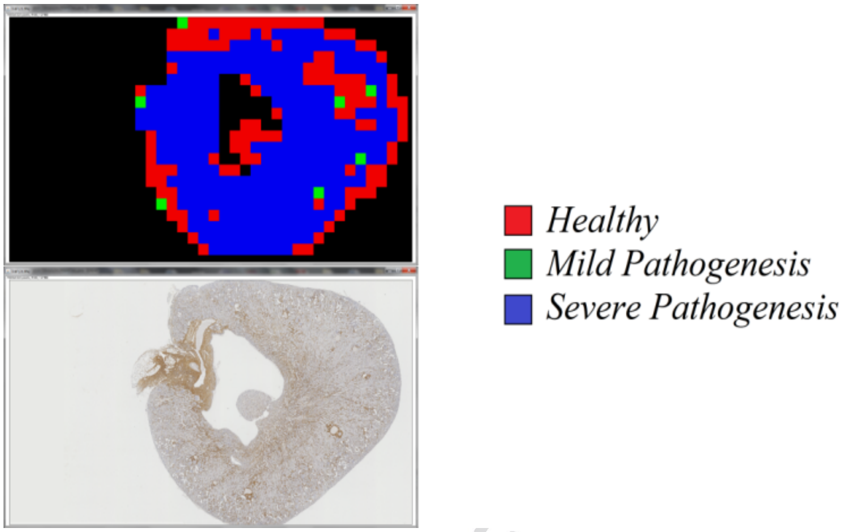
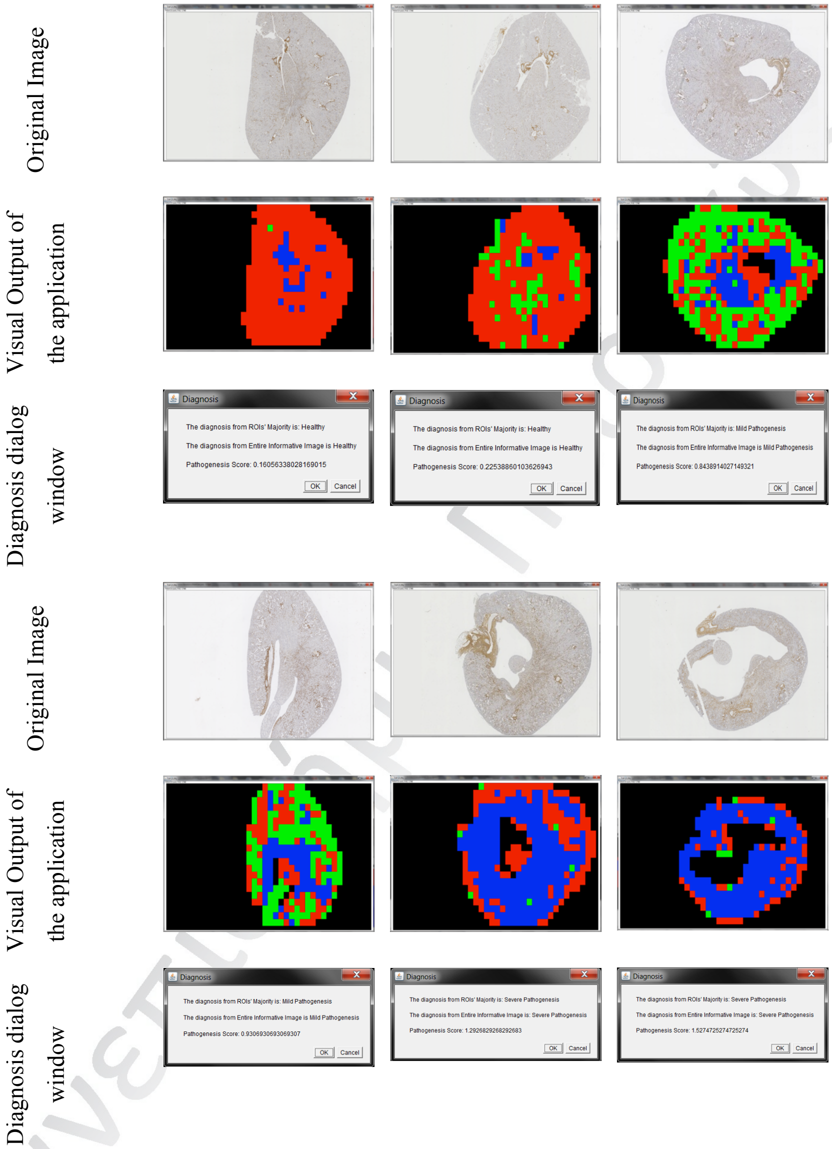


Figure 120. Color captions of the visual results (red is healthy tissue, green is mild and blue is severe pathogenesis)

Table 25. Practical accuracy of the proposed application

	1000*1000 block size	
	Characterization based on the majority of the Informative Segments	Characterization based on Entire Informative Image
Practical Accuracy (%)	80	92.5
	600*600 block size	
	Characterization based on the majority of the Informative Segments	Characterization based on Entire Informative Image
Practical Accuracy (%)	82.5	92.5
	200*200 block size	
	Characterization based on the majority of the Informative Segments	Characterization based on Entire Informative Image
Practical Accuracy (%)	82.5	85

The visual results of the proposed application are presented in Figure 121, where it can be noticed that are quite satisfying. Each segment is classified to a specific class (see Figure 120). Each class has a representative color (Red for Healthy, Green for Mild and Blue for Severe pathogenesis). The tool not only characterizes the whole image, but also illustrates the extension of the disease.



The diagnosis from ROIs' Majority is: Healthy
The diagnosis from Entire Informative Image is Healthy
Pathogenesis Score: 0.16056338028169015

The diagnosis from ROIs' Majority is: Healthy
The diagnosis from Entire Informative Image is Healthy
Pathogenesis Score: 0.22538860103626943

The diagnosis from ROIs' Majority is: Mild Pathogenesis
The diagnosis from Entire Informative Image is Mild Pathogenesis
Pathogenesis Score: 0.8438914027149321

The diagnosis from ROIs' Majority is: Mild Pathogenesis
The diagnosis from Entire Informative Image is Mild Pathogenesis
Pathogenesis Score: 0.9306930693069307

The diagnosis from ROIs' Majority is: Severe Pathogenesis
The diagnosis from Entire Informative Image is: Severe Pathogenesis
Pathogenesis Score: 1.2926829268292683

The diagnosis from ROIs' Majority is: Severe Pathogenesis
The diagnosis from Entire Informative Image is: Severe Pathogenesis
Pathogenesis Score: 1.5274725274725274

Figure 121. Experimental Results of the Proposed Application

The accuracy of the developed tool, achieves high rates in all segmentation block widths and in both approaches for each of them (see Table 25). The time required for the proposed tool to complete the above procedures and generate the visual results for an unknown biopsy image (on 7680x4320 pixels) is about 2 seconds for the 1000*1000 segmentation block on an Intel ® Core™ i7 CPU Q720 at 1.60GHz with 8 GB RAM installed, while for the 200*200 (which requires more processing time) segmentation block it requires 7 seconds for the characterization of the kidney biopsy image.

5.7 References

[1] G. Izbiki, M. J. Segel, T. G. Christensen, M. W. Conner, and B. R., "Time course of bleomycin-induced lung fibrosis," *Int J. Exp. Path*, vol. 83, pp. 111-119, 2002.

[2] C. Doukas, I. Maglogiannis, A. Chatziioannou, “Computer Supported Angiogenesis Quantification Using Image Analysis and Statistical Averaging” IEEE Transactions on Information Technology in Biomedicine 12 (2008), pp. 650-658.

[3] Haralick M et al, (1973) Textural Features for Image Classification. IEEE Transactions on systems man and cybernetics Vol. SMC-3 pp. 610-621

[4] Cortes C, Vapnik V. (1995) Support-Vector Networks, Machine Learning, 20, pp.273-297.

[5] Friedman N, Geiger D, Moises et al. (1997) Bayesian Network Classifiers, Machine Learning, pp. 131-163.

[6] Roussopoulos N, Kelley S, Vincent F. (1995) Nearest Neighbor Queries, SIGMOD '95 Proceedings of the 1995 ACM SIGMOD international conference on Management of data ISBN:0-89791-731-6, p71-79.

[7] T. Mitchell, "Decision Tree Learning", in T. Mitchell, Machine Learning, The McGraw-Hill Companies, Inc., 1997, pp. 52-78.

[8] Kohavi R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence P.1137—1143.

[9] Balazs Harangi, Rashid Jalal Qureshi, Adrienne Csutak, Tünde Petö, András Hajdu. Automatic detection of the optic disc using majority voting in a collection of optic disc detectors. In Proceedings of ISBI'2010. pp.1329-1332

[10] Tasoulis SK, Tasoulis DK, Plagianakos VP (2010) Enhancing principal direction divisive clustering, Pattern Recognition, Volume 43, Issue 10, October 2010, Pages 3391-3411.

[11] Sundaram S, Sea A, Feldman S, Strawbridge R, Hoopes P, Demidenko E, Binderup L, Gewirtz A (2003) The Combination of a Potent Vitamin D3 Analog, EB 1089, with Ionizing Radiation Reduces Tumor Growth and Induces Apoptosis

of MCF-7 Breast Tumor Xenografts in Nude Mice1. Clinical Cancer Research, vol. 9, pp. 2350-2356 June 2003.

[12] Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000, 16, 412–424.

[13] Stanley Sternberg, "Biomedical Image Processing", IEEE Computer, Volume 16, Issue 1 January 1983, pp 22-34.

[14] Comaniciu, D. and Meer, P. Mean shift analysis and applications. Computer Vision. The Proceedings of the Seventh IEEE International Conference on (Volume:2), 1999. Pp 1197-1203.

[15] K. Suzuki, I. Horiba, N. Sugie, "Linear-time connected-component labeling based on sequential local operations." *Computer Vision and Image Understanding*, vol. 89 Issue 1, pp.1-23, January 2003.

[16] Bangor, A., Kortum, P.T. and Miller, J.A. (2008) An empirical evaluation of the System Usability Scale (SUS). International Journal of Human-Computer Interaction 24(6). 574–594.

CHAPTER 6
Conclusions

6.1 General conclusion

The initial target of this thesis was to build a tool to assist the creation of efficient image mining workflows. The developed framework enables the generation of multiple workflow versions of the core workflow scheme by combining all the operators with each possible way and auto-selecting the optimal one. In this work, an application, exploiting web services and applying ontological modelling is presented, allowing the intelligent creation of image mining workflows and the optimal workflow scheme selection for each of them. The choice of the optimal workflow is based either to the comparison with the ground truth, or via the log-likelihood distance of the clustered salient objects.

This biomedical image analysis framework may require advanced knowledge of data mining and image analysis theory, but it requires only fundamental programming skills for the development of an intermediate level workflow scheme. It can be directly integrated to TAVERNA [1] or similar workflow management platforms.

Furthermore, additional methodologies and techniques can be integrated into this open web based image mining framework and enhance its current performance.

6.2 Conclusion from the Case Studies Experimentation

In this subsection, the conclusions extracted from the utilization of the proposed framework are presented.

6.2.1 Kidney Biopsies and Obstructive Nephropathy Detection

The proposed framework may support the knowledge discovery cycle downstream of image mining tasks, i.e., from exploratory image preprocessing and transformation to pattern discovery, hypothesis induction and image recognition. Tools and workflows to support the different phases of the image mining process are accessible through simple semantic inquiries in web repositories instrumentally linked and supporting open source workflow tools, such as the Taverna Workflow Manager and Rapid Miner. The above functionalities are novel in comparison with existing image-mining tools. The operators provided by the developed image-mining framework can be individually integrated into workflows and/or combined with other image analysis/biological analysis operators for creating workflows, provided by both environments.

The framework was utilized to rapidly construct a novel methodology tool for segmenting and characterizing microscopic kidney biopsies. The specific image type was selected as a showcase, due to the difficulties it presents for automated image analysis as well as the importance that the correct characterization bears, regarding the efficient disease management. The achieved results are promising. The proposed work-plan has proven effective and useful for automated detection and quantification of pathogenesis in kidney biopsy images, despite some misclassifications, which are expected that through further elaboration and optimization of the workflows would be minimized if not vanished.

In cases where the available resources do not match the requirements of the task as apprehended and specified by the expert, the proposed approach could facilitate and minimize the time required for forging up new solutions. By reusing and combining available computational avenues and programming solutions, set for seemingly different image analysis tasks, in a novel, ad-hoc ensuing, fashion, non-expert image processing users (i.e biologists) could resolve satisfactorily and automatically, image assessment tasks. The gross time reduction from design till the operative deployment of the workflows represents a critical, as well as, evident and technologically innovative component of the proposed approach. The work of the expert for a specific, image processing task can be assessed by the relevant image processing community, while outstanding results attained by these workflows can be refactored and tested for their efficacy on other image types. The results presented here, support the technical feasibility of the methodology proposed in this work. This solution can be extended to accommodate other types of biomedical data (i.e. clinical or medical record data), thus, strengthening knowledge discovery and computer-aided intelligent diagnosis. In this way, translational clinical research is greatly empowered through advanced data mining techniques, paving the way for their introduction in future electronic healthcare systems.

We presented a novel tool for segmenting and characterizing microscopic kidney biopsies. The achieved accuracy results are quite satisfying. Although some misclassifications arise that should be treated in future work, the proposed methodology has been proved effective and useful for automated detection and quantification of pathogenesis in Kidney biopsy images.

6.2.2 Breast Cancer Cells and Skin Lesion Quantification

The framework contains basic image analysis techniques and some advanced segmentation methods. All these methods are accessible through the framework's operators-entities. This allows users to build their own workflows and apply any available method of their choice, depending on the problem they have to deal with. These workflows have the ability to adjust to specific image data mining tasks. More specifically, the proposed framework allows the parallel image processing of large amount of images through the user's developed workflow scheme, combined with the feature of multiple instance generation of the same workflow, each of them is being initialized with different settings (classifiers, variables, etc.), but maintaining the initial structure.

The above functionalities are novel in comparison with existing image-mining tools. The operators provided by the developed image-mining frameworks can be individually integrated into new user created workflows and/or combined with other image analysis operators for creating efficient workflows. In cases where the available resources do not match, the requirements of the task at a researcher's mind, the proposed approach could make far easier the development of new solutions, by reusing and combining available partial solutions. An ultimate goal of the proposed approach is to reduce time from design to deployment. This will follow directly from the natural selection process mentioned above. The work of the researcher for a specific image-processing task will be assessed by the image processing community, and outstanding research results will be picked up and tested on other type of images.

The proposed framework was utilized to rapidly construct a novel tool for segmenting and characterizing breast cancer biopsies and skin lesion images, and can easily adapt to other types of biomedical microscopy images. The proposed framework was evaluated through the two presented use cases of breast and skin cancer pathologies.

In the first case, the corresponding workflow scheme was applied on breast cancer biopsy images and achieved 95.31% accuracy in the classification of cancer and apoptotic cells respectively, providing a ratio of cancer cells, over apoptotic cells. Rapid characterization was achieved through the feature extraction operator, providing the average feature values for each class of cells. In this case, an additional methodology – based on mean shift segmentation – was evaluated. The latter methodology provided less accurate, but still descent results. The strong part of the second approach is the speed performance of the mean shift algorithm.

In the second case, the corresponding workflow scheme was applied on melanoma case images and achieved accurate annotation of the salient objects and 93,7% accurate characterization of the skin lesion, based on the Matthews Correlation

Coefficient metric. Images from both case studies contain noisy background, different lighting states and in some cases, chromatic similarities between the different classes of the salient objects that harden the characterization task. However, both approaches achieved satisfying results, proving the versatility and usability of the proposed framework. The proposed framework was developed in Java, using the Apache MAVEN3 toolkit [2] and it is currently hosted on a Virtual Machine, powered by GRNET-Okeanos.

6.2.3 Scoring System for the fibrotic areas of the complex Kidney Biopsy Image dataset

A tool for the fast and accurate characterization of the complex kidney biopsies images along with correct pathogenesis detection and quantification was proposed. The use of the Random Forest algorithm proved quite sufficient in characterizing an unknown kidney biopsy with the two-way approach, since it achieved correct block-based and entire image classification reaching 84.64% and 97.5% accuracy respectively. In this work, a scoring system for additional image characterization was also proposed. The recommended tool offers fast and reproducible results to the expert pathologists, who have to deal with a high amount of kidney biopsies. The tool has been developed in Java. This provides the versatility of the easy integration to other system and especially into application servers through the utilization of web services. Feature work includes the salient object characterization contained in kidney biopsy image, such as glomerulus, tubulus, vessels, etc.

6.3 Workflow Evaluation and Optimal Scheme Selection

The proposed framework provides two robust features. The first one enables the ability to create multiple workflow instances by integrating additional operators into its core operators. Each of the generated workflows is initialized with different settings and contains different combinations of the constructed operators (filters, processes, variables, etc.), without changing the initial workflow architecture. Thus, it is not necessary to perform multiple test runs with different combinations. The second feature evaluates each generated workflow instance, based to the provided ground truth or the log-likelihood distance of the clustered salient objects, and auto-selects the optimal one. The combination of the above two features facilitates the work of the researchers, the expert physicians and image analysis experts by providing the optimal solution for a specific approach to an image mining task. This application practically removes the most time consuming processes of a complex image mining task.

The above functionalities are novel in comparison with existing image-mining tools and workflow frameworks. The operators provided by the developed image-mining frameworks can be individually integrated into new user created workflows and/or combined with other image analysis operators for creating efficient workflows. In cases where the available resources do not match, the requirements of the task at a researcher's mind, the proposed methodology could make far easier the development of new solutions, by reusing and combining available partial solutions. An ultimate goal of the proposed solution is to reduce time from design to deployment. The work of the researcher for a specific image-processing task will be assessed by the image processing community and outstanding research results will be picked up and tested on other type of images.

Two case scenarios were presented, demonstrating the main functionalities of the proposed framework. Both of them exploited the robust features of the framework, which are the multiple workflow instances generation, the parallel running of the instances and the evaluative process of the selection of the optimal workflow instance. In the first case, 8 workflow instances – that were generated from a two-tier image mining procedure – were evaluated through the ground truth checking operator. The operator truly selected the workflow instance that achieved the greatest performance. In the second case, where the salient objects were too many and their manual annotation was impossible, the evaluation of the 12 generated instances was achieved through the clustering of the salient objects and the extraction of log-likelihood distance. The proposed workflow instance was the one that achieved the higher log-likelihood distance, indicating the segmentation, which provided the clustering with the greater distance between the clusters.

In both cases, a lot of time was saved, considering the number of generated workflow instances and the evaluation processes that selected the best workflow scheme. Images from both cases contain noisy background and different lighting states that harden the segmentation task. However, both scenarios achieved satisfying results, proving the functionality and usability of the proposed framework. There are image datasets, far more complex than the presented datasets. The more difficult the image mining, the greater the timesavings provided by the proposed framework. The proposed web-based framework was developed in Java, using the Apache MAVEN3 toolkit and is currently hosted on a Virtual Machine powered by GRNET-Okeanos.

6.4 Future Work

Additional or newly developed image mining techniques can be integrated into this open web-based image mining developed framework, since it is written in Java. The architecture of the framework makes easy the further upgrade of its ontology-based features.

Feature work may include video and audio analysis procedures, which would support the processing and characterization of the provided datasets (i.e. the analysis and characterization of snoring audio recordings, recognizing potential pathologies, the characterization of colonoscopy videos through and enhanced version of the web-based framework).

Due to the recent rise up of smart phones and tablets, a potential feature work could be a mobile version of this framework that will be able to handle data received from the camera of the corresponding device. Sound or motion type data could be exploited, in order to characterize the health status of the device's owner.

Concluding this work, a potential feature work may include improvements in the validation of the Workflow and the optimal workflow selection process.

6.5 References

[1] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R. Pocock, Anil Wipat and Peter Li, “Taverna: a tool for the composition and enactment of bioinformatics workflows”, Bioinformatics, vol. 20, no. 17, pp. 3045-3054, June 2004.

[2] http://maven.apache.org/ “Apache Maven”.

CHAPTER 7

Publications

7.1 Journals

1. T. Goudas, C. Doukas, A. Chatziioannou, I. Maglogiannis. "A Collaborative Biomedical Image-Mining Framework: Application on the Image Analysis of Microscopic Kidney Biopsies". IEEE Journal of Biomedical and Health Informatics, Volume 17 Issue 1, January 2013, pp 82-91

2. Iakovidis DK, Goudas T, Smailis C, Maglogiannis I. Ratsnake: a versatile image annotation tool with application to computer-aided diagnosis. The Scientific World Journal. Volume 2014 (2014), Article ID 286856, 12 pages, 2014. http://dx.doi.org/10.1155/2014/286856 (doi:10.1155/2014/286856)

3. Goudas T, Maglogiannis I. An advanced image analysis tool for the quantification and characterization of breast cancer in microscopy images. Journal of Medical Systems. Volume X, Issue X, pp x-x [Accepted with Minor Revision]

4. Goudas T, Maglogiannis I. Image Mining Framework: Enabling Multiple Parallel Workflow Instances Generation and Optimal Workflow Selection. XX-XX [Pending]

7.2 Conferences

1. T. Goudas, C. Doukas, I. Maglogiannis, A. Chatziioannou. Salient Regions Detection in Microscopic Kidney Biopsies Utilizing Image Analysis Techniques. In: Proceedings of the 12th Mediterranean Conference on Medical and Biological Engineering and Computing– MEDICON 2010, May 27-30, 2010, Chalkidiki, Greece.
2. C. Doukas, T. Goudas, S. Fischer, I. Mierswa, A. Chatziioannou and I. Maglogiannis. An Open Data Mining Framework for the Analysis of Medical Images: Application on Obstructive Nephropathy Microscopy Images. In: Proc. of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2010, Buenos Aires, Argentina, pp. 4108-4111.
3. Theodosios Goudas, Ilias Maglogiannis. Advanced Cancer Cell Characterization and Quantification of Microscopy Images. Artificial Intelligence: Theories and Applications. Lecture Notes in Computer Science Volume 7297, 2012, pp. 315-322.
4. Goudas, T., Doukas, C., Chatziioannou, A., Maglogiannis, I. Advanced characterization of microscopic Kidney biopsies utilizing image analysis techniques (2012) Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, art. no. 6346945, pp. 4414-4417.
5. Goudas, T., Maglogiannis, I. Cancer cells detection and pathology quantification utilizing image analysis techniques. (2012) Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, 2012, pp. 4418-4421.
6. Goudas, T., Maglogiannis, I., Chatziioannou, A. Advanced block detection and quantification of fibrotic areas in microscopy images of obstructive nephropathy (2012) Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, 1, art. no. 6495144, pp. 928-932.

APPENDIX A

Using SUS

The SU scale is generally used after the respondent has had an opportunity to use the system being evaluated, but before any debriefing or discussion takes place. Respondents should be asked to record their immediate response to each item, rather than thinking about items for a long time.

All items should be checked. If a respondent feels that they cannot respond to a particular item, they should mark the centre point of the scale.

Scoring SUS

SUS yields a single number representing a composite measure of the overall usability of the system being studied. Note that scores for individual items are not meaningful on their own.

To calculate the SUS score, first sum the score contributions from each item. Each item's score contribution will range from 0 to 4. For items 1,3,5,7,and 9 the score contribution is the scale position minus 1. For items 2,4,6,8 and 10, the contribution is 5 minus the scale position. Multiply the sum of the scores by 2.5 to obtain the overall value of SU.

SUS scores have a range of 0 to 100.

The following section gives an example of a scored SU scale.

System Usability Scale

	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	□ 1	□ 2	□ 3	□ 4	□ 5
2. I found the system unnecessarily complex	□ 1	□ 2	□ 3	□ 4	□ 5
3. I thought the system was easy to use	□ 1	□ 2	□ 3	□ 4	□ 5
4. I think that I would need the support of a technical person to be able to use this system	□ 1	□ 2	□ 3	□ 4	□ 5
5. I found the various functions in this system were well integrated	□ 1	□ 2	□ 3	□ 4	□ 5
6. I thought there was too much inconsistency in this system	□ 1	□ 2	□ 3	□ 4	□ 5
7. I would imagine that most people would learn to use this system very quickly	□ 1	□ 2	□ 3	□ 4	□ 5
8. I found the system very cumbersome to use	□ 1	□ 2	□ 3	□ 4	□ 5
9. I felt very confident using the system	□ 1	□ 2	□ 3	□ 4	□ 5
10. I needed to learn a lot of things before I could get going with this system	□ 1	□ 2	□ 3	□ 4	□ 5

System Usability Scale

© Digital Equipment Corporation, 1986.

	Strongly disagree				Strongly agree	
1. I think that I would like to use this system frequently	□ 1	□ 2	□ 3	□ 4	□ 5	4
2. I found the system unnecessarily complex	□ 1	□ 2	□ 3	□ 4	□ 5	1
3. I thought the system was easy to use	□ 1	□ 2	□ 3	□ 4	□ 5	1
4. I think that I would need the support of a technical person to be able to use this system	□ 1	□ 2	□ 3	□ 4	□ 5	4
5. I found the various functions in this system were well integrated	□ 1	□ 2	□ 3	□ 4	□ 5	1
6. I thought there was too much inconsistency in this system	□ 1	□ 2	□ 3	□ 4	□ 5	2
7. I would imagine that most people would learn to use this system very quickly	□ 1	□ 2	□ 3	□ 4	□ 5	1
8. I found the system very cumbersome to use	□ 1	□ 2	□ 3	□ 4	□ 5	1
9. I felt very confident using the system	□ 1	□ 2	□ 3	□ 4	□ 5	4
10. I needed to learn a lot of things before I could get going with this system	□ 1	□ 2	□ 3	□ 4	□ 5	3

Total score = 22

SUS Score = 22 *2.5 = 55