ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

ENHANCING MEDICAL IMAGING CLASSIFICATION SCHEMES WITH EXPLAINABILITY PROPERTIES

A Dissertation
Submitted to
The Department of Digital Systems,
University of Piraeus in Complete Fulfillment
of the Requirements for the Degree
of Doctor of Philosophy

by

Athanasios Kallipolitis

Piraeus, 6 June 2023

ENHANCING MEDICAL IMAGING CLASSIFICATION SCHEMES WITH EXPLAINABILITY PROPERTIES

Approved by:

Professor Ilias Maglogiannis, Supervisor
School of Information and Communication Studies
Department of Digital Systems
University of Piraeus

Assistant Professor Orestis Telelis,
School of Information and Communication Studies
Department of Digital Systems
University of Piraeus

Professor Konstantinos K. Delimpasis
School of Sciences
Department of Computer Science and
Biomedical Informatics
University of Thessaly

Date Approved: 6 June 2023

ENHANCING MEDICAL IMAGING CLASSIFICATION SCHEMES WITH EXPLAINABILITY PROPERTIES

Accepted by:

Professor Konstantinos K. Delimpasis
School of Sciences
Department of Computer Science and Biomedical Informatics
University of Thessaly

Assistant Professor Orestis Telelis,
School of Information and Communication Studies
Department of Digital Systems
University of Piraeus

Professor Andriana Prentza
School of Information and Communication Studies
Department of Digital Systems
University of Piraeus

Associate Professor Spyretta Golemati
Medical School
Department of Biomedical Engineering
University of Athens

Professor Michael Filippakis
School of Information and Communication Studies
Department of Digital Systems
University of Piraeus

Professor Nikitas-Marinos Sgouros
School of Information and Communication Studies
Department of Digital Systems
University of Piraeus

Date Approved: 6 June 2023

ACKNOWLEDGEMENTS

Upon completion of this personal endeavor, thoughts and emotions grow to their full extent and, therefore, cannot be put into words. Nevertheless, I will make an effort to express my deepest appreciation to my wife who has uncomplainingly supported my efforts and has patiently recovered several shortcomings of me not being a “full-time” husband and father. Several pieces need to be put together in order to visualize the whole puzzle of support, devotion, trust, motivation and patience but the most important pieces, the pieces that form the frame belong to Professor Ilias Maglogiannis. It has been his unselfish and unstoppable desire for tutoring in academic, but life matters as well that boosted my every step and relieved the burden of my shoulders. To anticipate and feel that all commitments and needs will be satisfied, and potential issues will be smoothly solved is the most powerful source of energy that a supervisor can provide to his students.

To put all the pieces together, I would like to manifest my many thanks to all my beloved relatives, including the in-law members, for providing me limitless resources to accomplish my research goals.

Moreover, I would like to communicate my sincere gratitude to all my professors in the Digital Systems Department of the University of Piraeus. Had it not been for their willingness to share their knowledge and experience with me, the writing of this thesis would not be feasible. To Professor Konstantinos. K. Delimpasis and Assistant Professor Orestis Telelis, I would like to express my earnest acknowledgement for their participation in the Advisors committee.

Lastly, a special thanks to all people that doubt and question the ability, strength and desire for self-improvement.

Dedicated to Dione and Lefteris

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AI	*Artificial Intelligence*
AUC	*Area under Curve*
BCC	*Basal Cell Carcinoma*
BOVW	*Bag of Visual Words*
BoW	*Bag of Words*
CBIR	*Content Based Image Retrieval*
CLAHE	*Contrast Limited Adaptive Histogram Equalization*
CNN	*Convolutional Neural Network*
CV	*Computer Vision*
DL	*Deep Learning*
DCNN	*Deep Convolutional Neural Network*
DoG	*Difference of Gaussian*
DPS	*Digital Pathology System*
FV	*Fisher Vector*
GAN	*Generative Adversarial Network*
GLCM	*Greyscale Cooccurrence Matrix*
GMM	*Gaussian Mixture Model*
GradCAM	*Gradient based Class Activation Map*
ICM	*Inner Cell Mass*
LIS	*Laboratory Information System*
LoG	*Laplacian of Gaussian*
MB	*Mobile Inverted Bottleneck*
MRI	*Magnetic Resonance Imaging*
ML	*Machine Learning*
NLP	*Natural Processing Language*
RDBMS	*Relational Database Management System*

ROI	*Region of Interest*
TF	*Term Frequency*
IDF	*Inverse Document Frequency*
TML	*Traditional Machine Learning*
TE	*Trophectoderm*
SCC	*Squamous Cell Carcinoma*
SIFT	*Scale Invariant Feature Transform*
SQL	*Structured Query Language*
SURF	*Speeded up Robust Features*
SVM	*Support Vector Machine*
RCM	*Reflectance Confocal Microscopy*
VGG	*Visual Geometry Group*
VLAD	*Vector Locally Aggregated Descriptors*
VPWL	*Visual Pattern Weighted Localization*
VWWC	*Visual Word Weight Calculation*
WSI	*Whole Slide Images*
XML	*eXtensible Markup Language*

SUMMARY

The understanding of machine learning concepts and techniques provides us the ability to create a virtual tool inventory by which real-life problems can be addressed in an automated, cost and time-efficient manner. At the beginning of a self-improving journey, a machine learning engineer worries about the performance of the machine learning pipeline which is mostly expressed in terms of bias error. The first step of maturation involves the acknowledgment of underfitting and overfitting as part of a machine learning process that involves only a small dataset, as a representative part of the a-priori knowledge, in order to solve a complex and multifactorial problem. Inevitably, a "mature" machine learning practitioner needs to grasp the importance of generalization, the corresponding variance, and irreducible errors, if he is bound to evolve into a "grown-up" machine learning expert. We learn to account for our work by measuring the performance of a proposed methodology. It is a strong requirement that evaluation metrics such as accuracy, balanced accuracy, precision, recall, and others according to the nature of a machine learning task are retained at high values, but there exist equally significant aspects that need to be taken into consideration. Discovering the principal and auxiliary causes upon which a predictive model decides on one class in favor of another can be a powerful source of useful knowledge and, therefore, the light should be shed on the inner mechanisms of decision-making. Naturally, the journey of self-improvement is a never-ending loop, as the realization of acquired knowledge leads to new unanswered questions. For this thesis, the maturation of the author with basic machine learning notions has evolved into a quest for developing machine learning approaches for image classification that, inherently or post-hoc, have the ability to provide meaningful connections between

the predicted outcome and the visual patterns that most influenced it. Since traditional machine-learning approaches remain in the field as efficient solutions for tasks that are covered with little data, the proposed methodologies cover both traditional machine learning and deep learning architectures. The scope of the thesis is limited to medical imaging and the application of explainable machine-learning approaches for the handling of corresponding health-related issues. Medical images are considered one of the richest sources of information concerning health data and the basis upon which experts make decisions for treatment plans and interventions. Creating explainable automated systems that support these decisions can speed up their integration process into everyday clinical workflows since experts will be able to understand and trust the generated predictions through added transparency and causality. Towards the integration of explainable properties in machine learning classification schemes, in this thesis a novel explainability scheme is proposed which is based on the Bag of Visual Words paradigm for the interpretation of image classification results by means of ensemble explainable classifiers. Since Fisher Vectors push the performance of vocabulary-based approaches to higher values, the proposed methodology evolves to support the architecture of generative models, such as Gaussian Mixture Models. Concerning deep learning techniques, a novel modular explainability approach is proposed that exploits the advantages of two well-established approaches, Gradient-Based Class Activation Maps, and Superpixels. The results show that the combined scheme significantly increases the performance of the original Gradient-based approach and the modularity allows for implementation with different explainability approaches

INTRODUCTION

1.1 Scope

1.1.1 Traditional Machine Learning techniques against deep learning for medical images

Starting from the earlier years of machine learning's engagement in decision support and assistive diagnosis related to the healthcare domain, automated systems for the detection and classification of visual patterns have been developed, discussed, and improved by the corresponding researchers [1]. Medical images have been utilized as a part of the incoming data that train ML algorithms and constitute a special chapter in the field of Computer Vision due to their special characteristics in terms of quality, contrast, density, and resolution. These features through various modalities of medical images are the main reason why dedicated algorithms for knowledge extraction from medical images are explored and developed and incorporated into automated systems. In the beginning, these systems were based on the traditional machine learning (TML) paradigm that separates each process of the whole machine learning (ML) task into discrete modules as shown in Figure 1. The described modularity facilitates better management and more transparency of the individual processes in order to intervene and thoroughly comprehend the inner workings in contrast to the all-in-one process that corresponds to the deep learning workflow. The transparency of TML approaches can be exploited for the development of explainability schemes to link the outcome to the visual stimuli of the input data. Although the ground for providing visual explanations related to the predicted outcome is more prosperous when utilizing TML, there are certain shortcomings concerning the ability of

the earlier approaches to achieve high-performance results in terms of classification

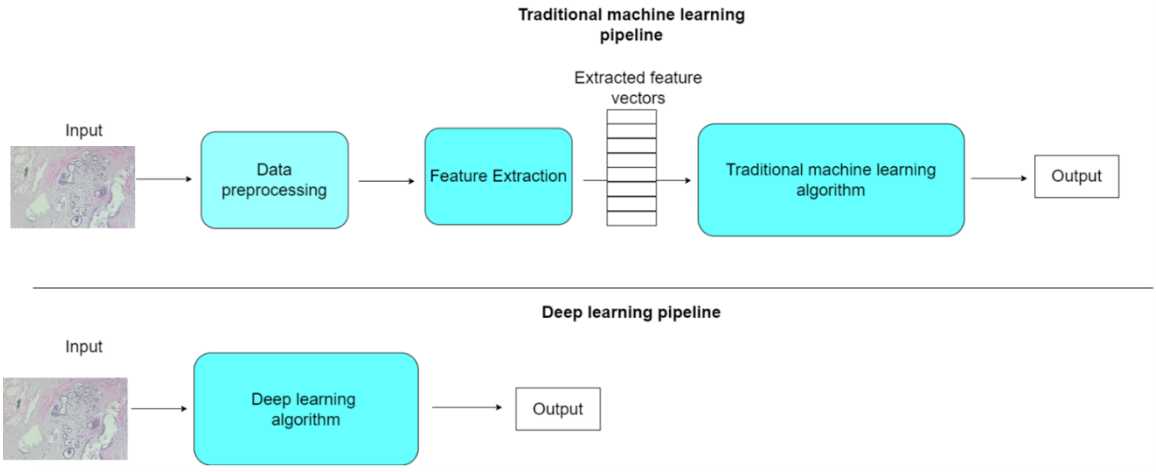


Figure 1 - Traditional machine learning vs deep learning pipeline for image classification tasks.

accuracy. Furthermore, due to the fact that the need for providing explainable results for ML approaches was highlighted after the development and establishment of TML, there are not so many works that address the issue of explainability for these approaches. Another determining factor for the absence of related work is the diffusion of DL. In the modern years, by the exploitation of well-established properties of Convolutional Neural Networks (CNNs) for the extraction of useful knowledge from medical images, researchers have strived for performance in terms of the relevant metrics of accuracy/balanced accuracy, sensitivity, and specificity. Towards this end, CNNs have managed to rival or even outperform human experts in various machine learning tasks such as X-Rays or Optical coherence tomography images [2]. To achieve this goal, CNNs have evolved into complex and deep architectures comprising various layers and structural blocks that have transformed them into black boxes. A key step in the proposed methodologies is the extraction of features from the images. Handcrafted features are designed with the primary purpose of quantifying the visual patterns that experts have identified as important based

on their experience. Local instead of global descriptors are considered the best choice for the extraction of robust and compact features that are in turn combined with a vocabulary-based approach for the transformation of multiple low-level vectors into one. On the other hand, deep learning requires no prior expert knowledge to quantify useful features (learned) in images that can, in turn, be utilized for downstream tasks. When utilized in ML workflows, learned features in many cases go beyond human vision and achieve better performance in terms of accuracy and repeatability [3] in contrast to handcrafted features. The processing of traditional medical imaging materials such as MRIs, X-rays, Ultrasounds, Endoscopy, Thermography, Tomography, Microscopy, and Dermoscopy has been transformed to each digital version, providing numerous benefits in a variety of tasks that were earlier performed manually [4-11]. The abovementioned tasks fall under the umbrella of well-known computer vision tasks, namely, image classification [12]-generation[13]-registration[14, 15], semantic segmentation[16], and object detection[17]. To address these tasks, machine learning techniques with a dedicated emphasis on deep learning methodologies have been applied successfully in the field of health informatics as an assistive tool for the relief of workload that specialized medical personnel need to carry [18] and for educational purposes [19]. The iterative process of continuously evolving the concerned algorithms has brought to light more effective implementations that exceed the human eye's discriminative capability [20-22] and enhance the objectivity criteria by means of visual patterns' quantification.

As shown in Figure 2, TML approaches achieve comparable or better results with DL up to a certain amount of data samples. It is important to mention the superficial equality between TML and DL until the plateau of TML can be confusing for the selection of the

learning strategy. When improved performance is the desired outcome, machine learning engineers often overlook the consisting "ingredients" of the generalization error and choose DL approaches. The generalization error consists of a) bias, b) variance error, and irreducible error. However, the common routine of splitting a small population of data samples in order to present classification results based on the generated errors hides significant pitfalls, such as the neglection of the variance error. Regardless of the amount of data samples that should be a determining factor for the selection of ML against DL techniques, most of the presented approaches that classify medical images in an attempt to provide accurate automated diagnosis apply complex deep convolutional architecture. The utilized networks may achieve low bias error but there is always the issue of overfitting if the dataset's distribution is simpler than the predictive model. In such cases, the models

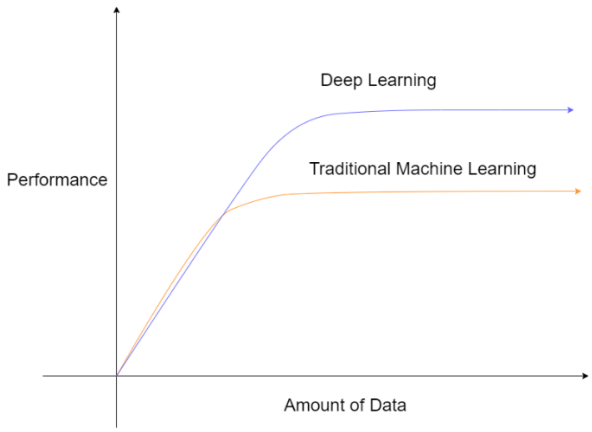


Figure 2 – Performance of deep vs traditional machine learning with respect to amount of data utilized for training.

may focus on existing random noise that favors the reduction of bias error for the specific dataset but will drastically increase the variance error when a dataset including different random noise appears. Selecting by default a deep convolutional architecture should not be considered a "panacea". The utilization of both learning approaches is thoroughly analyzed

in this work for the extraction of useful conclusions concerning the corresponding advantages and drawbacks.

1.1.2 Complexity and Explainability, two contradictory elements.

The extraction of useful knowledge from medical data started with the application of purely statistical methods. Their simplicity and transparent inner workings towards the generation of useful conclusions required no further elaboration for the detection of connections between stimuli and results. In Figure 3, statistical approaches are drawn in a three-dimensional space (complexity, explainability, and performance), assuming a position that corresponds to low complexity, low performance, and high explainability properties. Apart from the fact that the magnitude of complexity and explainability that can be easily deducted from the position in a two-dimensional space, the magnitude of the third dimension (performance) is deducted by the size of the ellipsis and the letter's font. As health-related information increased in size and dimensionality, a strong necessity was born for the development of machine learning techniques that were able to recognize and exploit interesting patterns in multi-dimensional spaces. These techniques derived from the statistics utilized many well-established notions from the relevant field and managed to develop into more complex algorithmic structures according to the transformation of medical data. The complexity added a certain layer of obscurity in the mechanisms of the proposed algorithms, but, under no circumstance, did they arrive at the point where deep learning techniques, nowadays, are considered black boxes.

As it can be easily understood, the rise of complexity in health-related data goes hand-by-hand with the generation of more complex models. This development is inevitable if the requirement of capturing the essence of related data distribution is to be met. Machine

learning engineers translate what is commonly called “capturing the essence …” as the extraction of meaningful features that can, in turn, be utilized for downstream tasks. The quality of the extracted features in terms of repeatability, identifiability, and robustness results in improved performance of the model, which is the third dimension of the graphical representation in Figure 3. While there is a strong and concurring correlation between data complexity, predictive models’ complexity, and predictive models’ performance, the opposite holds for predictive models’ complexity and explainability. In general, as machine learning engineers strive for performance, they are exploiting more complex models to capture important features from multidimensional data. This complexity deprives all stakeholders of the ability to comprehend the inner mechanisms and discover significant reasoning for the decision-making of the predictive model. As a result, explainability has been sacrificed in favor of accuracy. Especially, in the field of medical imaging where computer-aided diagnosis systems are being utilized for decision-making that can significantly influence the quality of life of an individual, it is a strong prerequisite to designing them with explainability capabilities in order to enhance trust, transparency, and verifiability [23]. For medical experts to embrace Artificial Intelligence (AI) in the healthcare domain and integrate it into their daily routine, the generated results should be retraceable and reliable. The efforts of researchers are directed toward the discovery of methods that can highlight the relationships and interactions between the visual patterns included in an input image and the final prediction. ‘Unveiling these connections is of crucial importance since humans demand that health-threatening decisions are thoroughly justified [24]. This rightful requirement has also been published in the Checklist for

Trustworthy Artificial Intelligence [25] as an imperative guideline for autonomous systems that are in line with the European approach to AI development.

The integration of explainability properties in computer-aided diagnosis systems, either by design or as a post-hoc capability, will add significant value to the proposed classification schemes for the following reasons:

- Enhancement of the classification results with trustworthiness, transparency, and verifiability.

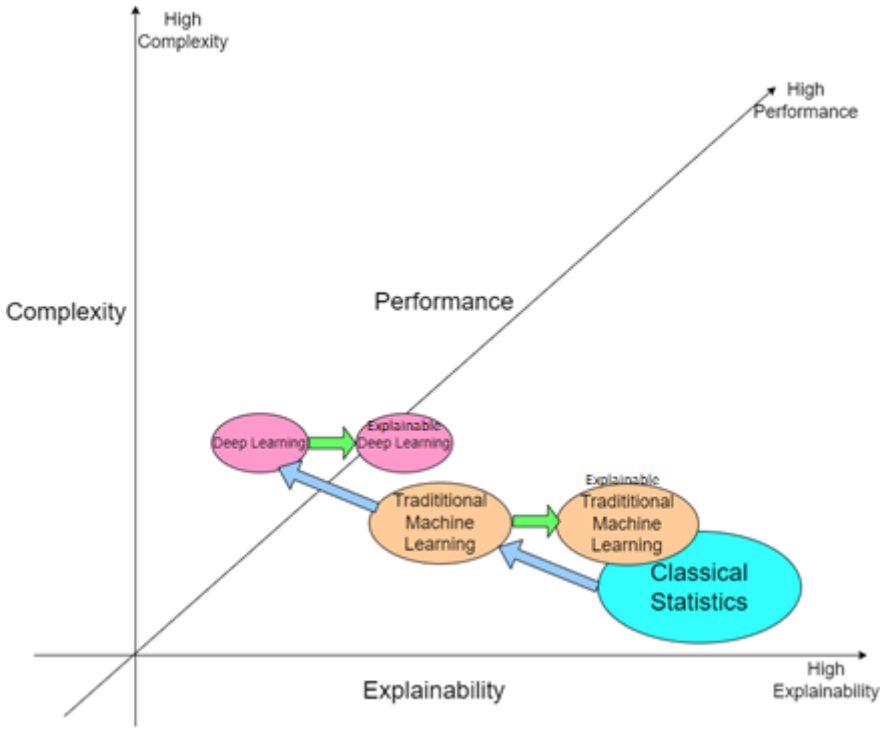


Figure 3 – Graphical representation between Explainability, Complexity and Performance notions. The size of letters and ellipses are representative of the approaches' magnitude in the performance axis. The light blue arrows demonstrate evolution over time and the green arrows the desired level of explainability.

- Increase the engagement of relevant medical experts in the whole process since they will have the ability to interpret the results.

- Improvement of the predictive models' performance since reasoning for erroneous/correct predictions will be retraceable.

- Discovery of new knowledge through the detection of unknown visual patterns to experts yet related to the illness.

This trend to create predictive models that provide plausible explanations of the generated results is highlighted in Figure 1 by the arrows that show Interpretable Traditional Machine Learning or Deep Learning techniques.

1.1.3 Ensemble schemes for improved generalization error and enhanced interpretability properties

The need for improvement concerning the generalization error has been an ever-ending effort starting from the earlier days of ML. This need has motivated scientists to seek and expand their inventory with additional algorithms that can better handle the ingredients of generalization error individually or together. ML tasks with medical images cannot be absent from this pursuit towards better performance, since the corresponding predictive models are responsible for assisting in decision-making in the healthcare setting, a role which is strongly related to the management of clinical interventions and treatment plans, and to a more general perspective, with quality of life. In general, ensemble methods existed before the rise of deep learning and were utilized in machine learning methods with the main purpose to increase the performance of the decision-making mechanisms that they consist of. Starting from ancient Greece and the foundation of Democracy, the idea of ensemble decision-making derives from the human best practice of seeking for opinions of different experts before taking high-risk decisions. The experts' opinion in the domain of

machine learning is represented by the prediction of a classifier. In an ensemble classifier, the input is analyzed by a set of classifiers, each implementing an algorithmic logic, resulting in a set of corresponding predictions that need to be combined in various manners in order to reach a final total prediction. Ensemble models have shown remarkable performance and the capability of correcting the faulty prediction of each included predictive model [26]. Such an example of exploiting the benefits of ensemble classifiers in the field of medical imaging can be manifested in [27], where authors employ a new weighted voting procedure on a self-supervised scheme towards the improved performance of medical X-Ray and computed tomography images' classification tasks. Apart from the advantage of providing a boost to the performance metrics, their simple implementation that lies on different architectural combinations provides the advantage of imposing explainability modules on top of existing architectures.

1.2 Research Questions

Machine learning engineers choose deep convolutional networks for the extraction of features from medical images of all kinds of modalities due to their smooth utilization without the need for prior knowledge for the task in question, their performance in terms of minimum bias error, and the easiness by which one can learn to train a CNN. This frivolous choice should not be made easier if significant factors are to be taken into consideration concerning the generalization and explainability properties of the designed predictive model. Traditional machine learning approaches should be equally considered as potential candidates in accordance with the volume of data and complexity of the ML task in question. Keeping in mind that all ML approaches offer advantages and are

degraded by drawbacks, this research has been structured to address challenges on both sides.

With reference to the volume of data samples, the curation and completion of a labeled dataset for an ML task require the involvement and tedious collaboration of an accountable number of scientists. There are cases where the generation of a large number of samples is hindered, possibly owing to an elaborated management process or the limited availability of required resources. Lacking standardization in the generation process can also be a reason for which data samples are limited. In such cases, the traditional machine learning approaches fit better to the task. On the other hand, the level of complexity is a defining characteristic of the task. There is an inherent difficulty in measuring the degree of complexity of an ML task, but by executing different trials and accounting for the returned evaluation metrics significant feedback can be provided. The analysis and evaluation of the TML and DL approach for medical images across the axes of generalization and interpretability with reference to the volume of data and complexity of the task is the basis of the research questions we attempt to answer. Towards the abovementioned axes, the following research questions are stated:

To what extent can TML techniques efficiently address medical imaging classification tasks? If so, are these techniques accompanied by an interpretability scheme that provides plausible explanations concerning the visual stimuli that determined the classification result?

To what extent can explainability techniques for DL approaches be quantitively evaluated? Do existing explainability techniques tackle efficiently the explainability requirement for medical images?

Are there explainability schemes for ensemble classifiers that can effortlessly be integrated into medical imaging classification models? Are these classifiers improving explainability results as well?

Since explainability relates to designated areas of the medical images, to what extent can DL approaches isolate their explainability results to areas that correspond to human expertise, or should segmentation algorithms be utilized to narrow down the candidate areas?

1.3 Contribution

In this thesis, the focus is concentrated on microscopy images such as histopathology whole slide and reflectance confocal images. Both modalities are strongly related to different cases of cancer and can be utilized for the detection of suspicious patterns in human tissues. Utilizing automated machine learning techniques for the prognosis and diagnosis of pathogenic conditions is vital for the early detection of malignancies in both cases aiming at total healing and avoidance of metastasis [42, 43]. In the case of digital pathology, the existence of public datasets for the provision of a priori knowledge is immense, and reported results of the deep learning techniques are high [44], but the need for explaining the connection between the input and the result is overlooked, yet, compelling. On the other hand, public datasets for RCM images are greatly limited which makes the application of DL techniques extremely difficult. The lack of public datasets for RCM

images inevitably influences the development of corresponding explainable approaches. Especially in the case of predictive models in healthcare information systems where the responsibility for high-stake decisions lies heavy, the need for classification schemes that are followed by explainable results is highly underlined. “In order to build trust in intelligent systems and move towards their meaningful integration into our everyday lives, it is clear that we must build ‘transparent’ models that have the ability to explain why they predict what they predict” [45]. In order to provide improved results with respect to the base classifiers, ensemble classifiers are utilized for the classification tasks when the abundance of samples permits so. Apart from the advantage of providing a boost to the performance metrics, their simple implementation that lies on different architectural combinations provides the advantage of imposing explainability modules on top of existing architectures. This integration is made possible, as well, due to the nature of the well-known Gradient Weighted Class Activation Mapping (Grad-CAM) technique [45] that can be applied effortlessly to the last convolutional layer of existing deep learning schemes without interfering with the functionality of the predictive model. The combination of an ensemble classifier with a Grad-CAM explanation scheme that can highlight the visual patterns which are responsible for each class prediction, while providing promising results is one of the presented contributions. Furthermore, a standalone application that follows the principles of distributed computing is available for online validation and experimentation, providing its functionality (classification and explainability) as a web service.

Although the Grad-CAM technique performs well when dealing with discrete objects in images of general interest, the case is different for complex medical images. With reference

to microscopy images, the assignment of importance to each specific structure at a cellular level is an important requirement to address. The depicted visual patterns are combined in a complex way and with such density that the existing explainability scheme fails to highlight specific structures [46]. There is a need for the provision of explainability results that highlight cellular structures of common visual characteristics, a requirement that goes beyond the rectangular regions provided by the Grad-CAM.

In this context, an explainability scheme that combines the well-established properties of Grad-CAM, while enhancing them with the localization information of specific structures that derive from three different segmentation algorithms is proposed. The results are analyzed and measured in terms of the impact of the designated superpixels on the classification results. The outcome shows a significant improvement when compared to the original Grad-CAM technique in terms of the Area Over Perturbation Curve (AOPC) metric and the correspondence between the highlighted regions and the experts' experience. Specifying regions of the image based on common visual characteristics is of importance in medical images with dense information. Instead of highlighting rectangular regions, the algorithm facilitates the designation of individual structures that are responsible for the classification results. The accurate delineation of cellular entities' boundaries can be beneficial to the explanation process since it provides better visualizations without hiding the important regions, and it restricts the important region to better-defined structures that make sense to the human experience. The proposed technique has been verified against skin cancer confocal and breast cancer histopathology images, but it can be utilized for the explanation of other medical image modalities. Although Grad-CAM has been chosen as a benchmark explainability algorithm, since it can be applied to

various CNN architectures and is considered efficient in terms of resources and implementation, other explainability algorithms can be utilized as backbones for the provision of feature importance.

Since local descriptors have demonstrated their efficiency in capturing identifiable, robust, and compact features in images, they are widely utilized in this thesis as the basis for feature extraction. The need to transform the generated vectors into a single representation for the classification task is addressed by vocabulary-based techniques such as BOVW and FV. In the literature, there is an unexpected lack of explainability approaches that can be embedded in vocabulary-based classification schemes for the designation of important areas that influence the result, mainly due to the fact that the explainability requirement showed up after the establishment of such techniques and the adherence of DL approaches. To make amends for this absence, a methodology for fast classification and interpretation of RCM images is proposed. The proposed methodology consists of the formation of a visual vocabulary based on Speeded Up Robust Features (SURF), a plain "vanilla" neural network classifier, and their interpretation schemes. The proposed methodology innovation lies on two factors: The classification results demonstrate that automated classification of RCM can relieve the time-consuming burden from dermatologists and save patients for the invasive procedure of biopsy, while the interpretation scheme provides useful insight for clinicians concerning the visual patterns mainly responsible for the classification outcome. To the best of our knowledge, it is the first time that explainability indicators are produced via the BOVW technique for the interpretation of an image classification system. It is also the first time that an explainability scheme is being utilized for the interpretation of the predictive result concerning confocal images. The explanation of each result is of great

importance to computer-aided diagnosis systems since it can produce knowledge about domain relationships contained in data [47].

DL techniques will take advantage of every pattern found in medical images for the sake of improving classification results. However, most of the time the areas that are important for decision-making are concentrated on segments of the medical image and not the whole of it. Utilizing patterns outside of these segments may lead to unwanted results since there are not related to the problem and constitute noise which seriously affects the generalization properties of the model in a negative manner. Providing the ground truth for segmentation algorithms is a far more arduous task than for classification algorithms since it is performed pixel by pixel. The proposal of an unsupervised segmentation approach for the segmentation of the different blastocyst segments is proposed as well as part of an explainability scheme based on the FV classification approach. The methodology initially segments the blastocyst from the background without the need for human labeling, provides visual explanations with a novel approach, and cross-examine the explainability results with class activations maps for deep learning techniques to verify the classification outcomes. To the best of our knowledge, this is the first time an explainability scheme is presented for the FV technique and one of the rare attempts to provide visual explanations for the prediction outcome on blastocysts images. To avoid erroneous assessment, it would be particularly useful to utilize machine learning methods that attribute quantitative indicators to the characteristics of the image to classify the blastocysts in one of the corresponding categories.

Apart from the contributions that address the main research questions herein, steps have been taken to propose automated tools for the assistance of human experts in their everyday

routine workflow in labs. Specifically, a content-based image retrieval tool that can import medical books (histopathology atlases) in pdf format is proposed. Inputs to a content-based image retrieval system can be whole slide images, images extracted from digital pathology books (literature), and local storage or digital pathology databases. These inputs, digital images to their whole, are imported into the system and compared to a query image of choice in order to detect the most similar in the literature. This automated process can relieve pathologists from the burden of manually searching in different hard copies for the detection of patterns that are already known to them. The proposed methodology for fast retrieval and classification of such images is based on the creation of visual vocabularies from scale invariable texture features and the notion of Bag of Visual Words (BOVW). Although BOVW is widely utilized for CBIR, the novelty of this work refers to the fusion of different features in the construction of a final vector that can describe ideally the dataset.

1.4 Thesis Structure and Summary per Chapter

The thesis is divided into five (5) main sections, as follows: Introduction, Related Work, Methodology, Experimental Results, Discussion, and the Conclusion.

The first part, Introduction, contains a description of the thesis' scope, the research questions with reference to explainability in the field of medical imaging and the proposed contributions. First, it refers to the application of traditional machine learning and deep learning classification techniques in the field of medical imaging. In the following lines, the notions of complexity and explainability are presented as two contradictory factors and the scope subsection ends with a detailed presentation of the ensemble architectures' role to improving the generalization of the corresponding base classifiers. The second

subsection mentions the research questions that the thesis is dealing with, and the third subsection refers to the proposed contribution in presenting novel explainability approaches for TML and DL classification schemes.

The second part, Related Work, contains existing works in the domain of medical imaging, where machine learning approaches are analyzed, evaluated, and put to comparison in terms of classification performance and explainability properties. The contained subsections refer to traditional machine learning and deep learning classification schemes that are applied for knowledge extraction from various modalities of medical images and further, describe the feature extraction and modelling mechanisms that lead to high-end predictive schemes. The related work section explains in short the inner workings of well-established explainability approaches and superpixel segmentation algorithms, while retaining in the end two smaller subsections with reference to ensemble classifiers and unsupervised segmentation.

The third section, Methodology, is divided into three subsections that contain descriptions and graphical representations about the proposed explainability approaches for traditional machine learning and deep learning classification schemes and the unsupervised segmentation of blastocyst images.

The fourth section, Experimental Results, presents in extensive detail the qualitative and quantitative results of the conducted experiments on medical images of different modalities such as whole slide, reflectance confocal microscopy and blastocyst images. The generated results provide substantial proof of the contributions in this thesis.

A discussion section follows the Experiments Results, where the limitations of the proposed methodologies are presented and directions for future works are set.

The thesis ends with the conclusion section, where a summary of the thesis, the research questions and the presented contributions are commented.

RELATED WORK

1.5 Machine Learning Applications and Explainability for knowledge extraction from medical images

As a means of knowledge extraction from data, machine learning has been exploited for the transformation of visual stimuli in medical images into objective measurable quantities that can, in turn, be utilized for various health-related predictive tasks such as the classification in grades of diseases, the detection and quantification of suspicious structures or the evolution of an illness. Each modality of medical imaging shares a set of distinguishable characteristics related to the density of visual information, the low contrast, the inherent noise, or obstacles that hinder the detailed review of corresponding specialists, the generated artifacts, and the variety in color distributions. This special set-up constitutes an operational field of opportunities and challenges for ML algorithms to delve into and explore.

In the field of digital pathology, where the introduction of Whole Slide Scanners (WSI) has allowed the digitalization of tissue slides and enabled the era of digital pathology, ML applications have shown an unseen bloom owing to the latest advancements in the handling of a large amount of data and the diffusion of several expert-curated public datasets, as depicted in Table 1. Content-Based Image Retrieval (CBIR) and classification in digital pathology images preoccupy several research groups and there are some works already reported in the literature. A part of them exploits global features extracted from the images, as opposed to other systems that utilize local features, while methodologies using both global and local features exist as well. Concerning the extraction of local features, a

Table 1 - Digital Pathology Datasets

Name	Url	Paper	Task
CAMELYON 16	https://pan.baidu.com/s/1UW_HLXXjjw5hUvBIUYPgbA	Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, van der Laak JAWM, and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer. JAMA. 2017;318(22):21992210. doi:10.1001/jama.2017.14585	Lymph node metastases detection
CAMELYON 17	https://pan.baidu.com/s/1mIzSewImtEisclPtTHGSyw#list/path=%2F	Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, Quirine F Manson, Nikolas Stathonikos, Alexi Baidoshvili, Paul van Diest, Carla Wauters, Marcory van Dijk, Jeroen van der Laak. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience, giy065, DOI: 10.1093/gigascience/giy065	Lymph node metastases detection and disease classification
WARWICK QU	https://warwick.ac.uk/fac/sci/dcs/research/tia/glascontest/download/	K. Sirinukunwattana, J. P. W. Pluim, H. Chen, X Qi, P. Heng, Y. Guo, L. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, A. Böhm, O. Ronneberger, B. Ben Cheikh, D. Racoceanu, P. Kainz, M. Pfeiffer, M. Urschler, D. R. J. Snead, N. M. Rajpoot, "Gland Segmentation in Colon Histology Images: The GlaS Challenge Contest" http://arxiv.org/abs/1603.00275 [Preprint]	Gland segmentation
KIMIA 960	http://kimia.uwaterloo.ca/kimia_lab_data_Path960.html	Kumar MD, Babaie M, Zhu S, Kalra S, Tizhoosh HR. A comparative study of CNN, BoVW and LBP for classification of histopathological images. ArXiv171001249 Cs; 2017.	Classification
BREAKHIS	https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/	Spanhol, F., Oliveira, L. S., Petitjean, C., Heutte, L., A Dataset for Breast Cancer Histopathological Image Classification, IEEE Transactions on Biomedical Engineering (TBME), 63(7):1455-1462, 2016.	Disease Classification
MITOS-ATYPIA	https://mitos-atypia-14.grand-challenge.org/Download/	Zakariapour, Sooshiant & Jazayeriy, Hamid & Ezoji, Mehdi. (2017). Mitosis detection in breast cancer histological images based on texture features using adaboost. Journal of Information Systems and Telecommunication. 5. 88-96.	Mitosis detection, nuclear atypia classification
MITOS 2012	http://ludo17.free.fr/mitos_2012/download.html	Roux L, Racoceanu D, Loménie N, Kulikova M, Irshad H, Klossa J, et al. Mitosis detec- tion in breast cancer histological images An ICPR 2012 contest. J Pathol Inform 2013;4:8. https://doi.org/10.4103/2153-3539.112693.	Mitosis detection
BIO-SEGMENTATION	https://bioimage.ucsb.edu/research/bio-segmentation	Gelasca ED, Byun J, Obara B, Manjunath BS. Evaluation and benchmark for biological image segmentation. 2008 15th IEEE Int. Conf. Image Process; 2008. p. 1816–9. https://doi.org/10.1109/ICIP.2008.4712130.	Disease Classification

		Classification and Retrieval of Digital Pathology Scans: A New Dataset	
KIMIA PATH24	http://kimia.uwaterloo.ca/kimia_lab_data_Path24.html	Morteza Babaie, Shivam Kalra, Aditya Sriram, Christopher Mitcheltree, Shujin Zhu, Amin Khatami, Shahryar Rahnamayan, H.R. Tizhoosh. CVMI Workshop @ CVPR 2017	Disease classification and retrieval
TGCA	https://portal.gdc.cancer.gov/	Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet 2013;45:1113–20.	
GTex	https://brd.nci.nih.gov/brd/image-search/%20searchhome	The genotype-tissue expression (GTEx) project. Nat Genet 2013;45:580–5. https://doi.org/10.1038/ng.2653.	
TMAD	https://tma.im/cgi-bin/home.pl	Marinelli RJ, Montgomery K, Liu CL, Shah NH, Prapong W, Nitzberg M, et al. The Stanford tissue microarray database. Nucleic Acids Res 2008;36:D871–7. https://doi.org/10.1093/nar/gkm861.	
TUPAC16	http://tupac.tue-image.nl/node/3		Mitosis detection
BIOIMAGING CHALLENGE 2015	https://rdm.inesctec.pt/dataset/nis-2017-003	Classification of breast cancer histology images using convolutional neural networks. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0177544	Disease classification
DEEP LEARNING DATASETS	http://www.andrewjanowczyk.com/deep-learning/	Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J Pathol Inform 2016;7:29. https:// doi.org/10.4103/2153-3539.186902	Gland segmentation, tumor detection, disease classification, mitosis detection, lymphocyte detection, epithelium segmentation, nuclear segmentation
COLON TMA EGFR	http://fimm.webmicroscope.net/supplements/epistroma	Doyle S., Madabhushi A., Feldman M., Tomaszeweski J. (2006) A Boosting Cascade for Automated Detection of Prostate Cancer from Digitized Histology. In: Larsen R., Nielsen M., Sporring J. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006. MICCAI 2006. Lecture Notes in Computer Science, vol 4191. Springer, Berlin, Heidelberg	

popular selection is the Scale Invariant Feature Transform (SIFT) [48]. In [49], SIFT features are also used to classify histopathological images by means of Support Vector Machines (SVMs). The main advantage of such types of features like SIFT and SURF is that they detect interest points in an unsupervised manner and, therefore, do not require tedious segmentation tasks to locate specific structures. On the other hand, some works have been reported [43] where the segmentation task is an inevitable burden to carry. For instance, in [50], authors propose a methodology to segment nuclei in breast cancer images

and the extraction of Haralick features from gray-level co-occurrence matrices is used to measure pleomorphism and heterogeneity.

Once feature extraction is completed, the task of classifying the images into the corresponding classes takes place. To this cause, one useful technique to decrease the dimensionality of multiple vectors is the Bag of Visual Words which comes from the field of Text Mining/Categorization. Several works that apply this technique have been reported in the last years concerning medical images [51]. Results are promising and in many cases, some systems outperform Deep Neural Network configurations, where the image is digested on its whole, but resized, by a neural network. In configurations of the latter case, the processes of feature extraction are entirely handled by the neural network. Going a step further, Jegou et al. at [52] propose the Vector Locally Aggregated Descriptors (VLAD) for image retrieval and classification that demonstrate better results in comparison to BOVW in big datasets.

As far as histopathology images are concerned, the need to provide feasible semantic and visual explanations of the generated results of the predictive model has been addressed in some studies [53]. In [54], the LIME methodology is utilized to generate explanations on the Patch Camelyon histopathology images [55]. Three different segmentation algorithms are compared for the generation of superpixels that are consequently used for the perturbation of initial images. By exploiting basic principles from the Grad-CAM and Guided Grad-CAM technique the authors in [56] propose an explainability scheme on top of a convolutional neural network. The presented work achieves competitive results in terms of classification while providing accurate cancerous evidence localization. In a different fashion than other schemes, the work in [57] presents a methodology that provides

explanations about the predictive result in a human-friendly manner. Along with a semantic explanation regarding the confidence of each prediction, visualizations of similar and non-similar samples of different classes are provided. The assignment of importance to each specific structure at a cellular level is an important requirement to address.

In dermatology, experts utilize digital dermoscopy to review potential malignancies on human skin. To verify their diagnosis, a sample from the skin is removed and forwarded to the histopathology labs where the nature of the tissue is determined. Histopathology remains the golden standard in experts' routines, but reflectance confocal microscopy is a non-invasive technology that can be utilized as an alternative for the diagnosis of skin cancer. Since skin cancer preoccupies a large portion of humans on a global scale [30], the majority of samples are operated on and sent to the lab in vain (only a few of them are malignant). Reflectance confocal microscopy offers the advantage of being able to review the sample at a cellular scale in vivo and capture the area of interest in a digital copy. However, generated images are of high density and complexity and only a few experts are specialized in evaluating the corresponding images. Due to RCM's recent adaptation in skin cancer assessment workflow and the cost of acquiring the respective hardware, literature in the dedicated domain of computer vision utilizing RCM images is very poor. The sparsity of labeled samples in conjunction with the density of visual patterns depicted in RCM images compose a very challenging scenario for Computer Vision (CV) researchers. Nowadays, most CV research work is focused on the utilization of Deep Convolutional Neural Networks (DCNN). The ability of the latter to absorb and analyze a large number of images and to efficiently execute machine learning tasks on them has led to the general acceptance of related programming frameworks such as Keras [58],

Tensorflow [59], and Deeplearning4j [60]. In [61], a hybrid approach improves the accuracy of a Convolutional Neural Network classification scheme owed to a sparsity of samples. The classification accuracy (51% with the CNN) is improved by 31% with the combination of a CNN and a texton-based unsupervised scheme. With the assistive hand of transfer learning, the issue of samples' sparsity can be tackled, as in [62], where RCM images are classified by a pretrained Resnet network. In the same fashion, but for a different task, a deep learning approach is utilized in [63] for the automated staining of RCM images. On the other hand, traditional machine learning techniques based on handcrafted features do not suffer from the need for a large number of image samples. While such techniques demonstrate promising results in relevant bioimaging domains [64], no proof of research work with the utilization of traditional methods was found in the literature.

An immense sparsity of research work is evident in the field of computer vision regarding confocal image classification models and explainability. Although deep convolutional networks are utilized in confocal images' classification tasks and demonstrated effective performance results, attempts to explain their predictions are not equally appreciated, since the Grad-CAM technique focuses on tiles of the image rather than highlighting specific structures at a cellular level. Although the technique performs well when dealing with discrete objects in images of general interest, the case is different for complex medical images. When referring to all modalities of medical microscopy imaging, the depicted visual patterns are combined in a complex way and with such density that the existing explainability scheme fails to highlight specific structures [46]. Therefore, there is a need for the provision of explainability results that highlight cellular structures of common

visual characteristics, a requirement that goes beyond the rectangular regions provided by the Grad-CAM. Even if the saliency maps are provided in fine granularity by the exploitation of pixel-wise explainability schemes, the need to define a larger area in order to measure the impact of meaningful structures in the image is compelling.

Especially in the field of skin histopathology or confocal images, attempts to interpret the predictions of classification algorithms are rare. In [65], the CAM and Grad-CAM technique is utilized to interpret results from a multiclass (melanoma / intradermal nevus/ compound nevus / junctional nevus) histopathology image classification task employing VGG19[66] and ResNet50[67] DCNNs, whereas in [68] a machine learning approach, based on the Bag of Words technique, is presented to highlight specific visual words on the image by retrieving their influence on the classification outcome.

Moving from microscopy images to a different modality of medical images that have attracted the attention of this research, the advantages of using ML and DL have been exploited by researchers in the case of images depicting blastocysts with the main purpose of predicting the most suitable candidate for implantation. The results are encouraging, as seen in [69] using the Inception - v1 neural network with an Area under Curve (AUC) of 0.987. This study utilizes a large database of 50,000 images in time-lapse format from the Weill Cornell Reproductive Medicine Center. Embryos are classified into three groups, good, fair, and poor quality according to a consensus of multiple embryologists. Regarding the selection of the appropriate embryo for transfer, in [70], the proposed neural network succeeds in predicting the most suitable embryo for transfer with 90% accuracy in 113-hour post-insemination samples. In [71] a deep learning approach demonstrates an AUC of 0.93 in predicting the probability of pregnancy with the fetal heartbeat. The

generalization properties of this methodology are verified through images collected from eight different laboratories.

Most of the studies manage to differentiate blastocysts coarsely (good vs bad quality), but the critical clinical need of distinguishing blastocysts of similar quality remains unsatisfactory. Furthermore, they fail to interpret the internal mechanisms of the utilized predictive models, meaning that no explanation is provided concerning the visual patterns that influenced the prediction. The association and localization of the visual patterns that greatly influence the result of the predictive model are crucial in high-risk predictive models, such as those utilized in medical applications. Explainability contributes to the discovery of reasoning for erroneous predictions or confounding factors [72]. In addition, customers, who intended to invest in such systems, are entitled to explanations that can cross-examine their working experience [73], a legitimate right that is manifested in the General Data Protection Regulation as well. Among several approaches for the generation of visual explanations in medical images, Grad-CAM is utilized broadly since it can be applied to various configurations of DL architectures.

1.6 Local Descriptors for feature extraction from medical images

1.6.1 Scale Invariant Feature Transform

The way humans perceive the surrounding world by means of vision is strongly related to the notion of scale. As quoted in [74], “An inherent property of objects in the world is that they only exist as meaningful entities over certain ranges of scale. If one aims at describing the structure of unknown real-world signals, then a multi-scale representation of data is of crucial importance”. Although the human brain has demonstrated a remarkable

ability to adapt in different scales in its attempt to recognize patterns and objects, the opposite occurs with computer vision. While a human can effortlessly recognize a leaf both when he is holding it in his arm and when he is looking at it from afar, computers struggle with such tasks. It has been from the beginning of the 21st century that scientists grasped the importance of robustness through different scales and occupied themselves with the development of feature extraction algorithms that can deliver the desired result.

A basic means towards the achievement of scale invariance has been the scale-space representation by image pyramids. Image pyramids consist of multiple representations of the same visual content but in different sizes and scales. The effect of resizing and rescaling is generated by the iterative blurring and downsampling of the initial image. Although there exist various techniques in the literature for the above-mentioned process, Gaussian blurring and bilinear interpolation are the chosen algorithms in the SIFT technique for the generation of the scale space representation. Using as an example the histopathology image, shown in Figure 4, a scale-space pyramidical representation is created. The SIFT algorithm [48] uses this set of generated images for the interest point detection procedure. By convolving the second-order derivative of the gaussian distribution (Laplacian of Gaussian-LoG) with each image, blobs are detected efficiently at the point of local extrema in three scales. Each pixel is compared by its eight neighbor pixels and the 18 pixels from the previous and next scales. The pixel with the highest value is selected as a potential key point that corresponds to a blob in the image with a certain r radius. It should be noted that the detection of key points do not take place on the convolution of images with the second other derivative but on a cheaper approximation which is expressed by the Difference of Gaussians (DoGs). SIFT algorithm consists of four steps:

- Keypoint detection.
- Keypoint localization.
- Orientation assignment.
- Keypoint description.

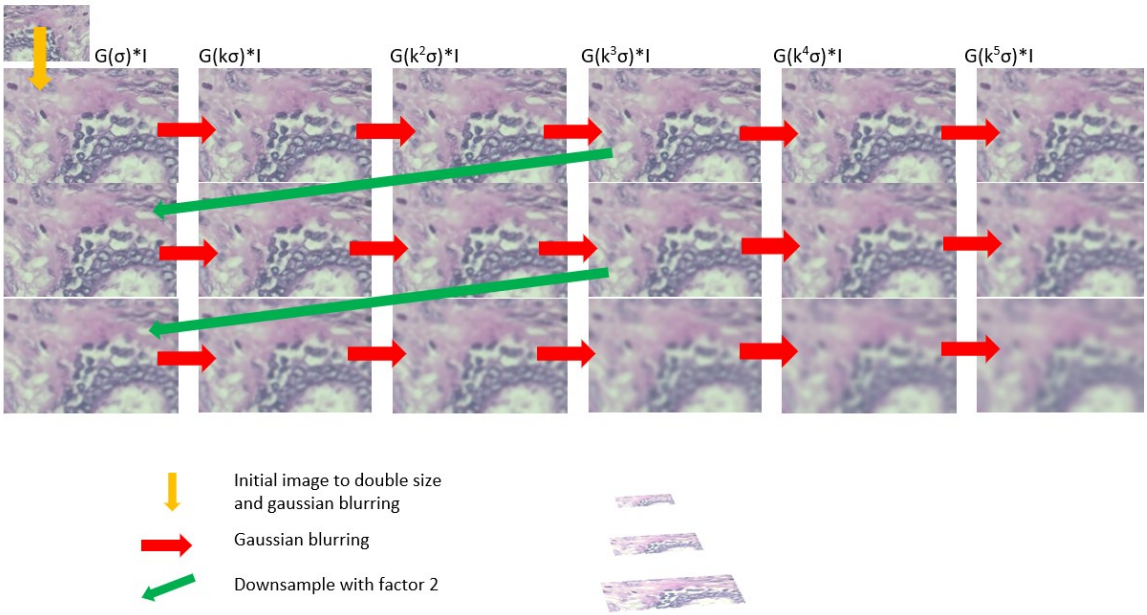


Figure 4 – Scale space representation of histopathology image.

If the histopathology image is represented as I (x, y) where x, y are pixel coordinates on a grayscale image then the gaussian blurring takes place by Equations (1), (2):

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{1}$$

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{2}$$

Moving one step further, we can detect blobs between edges and corners by convolving the image with the Laplacian of Gaussian, without neglecting an important property between convolution and differentiation that allows for conducting calculations according to Equation (3).

$$\frac{d}{dx}(f(x) * g(x)) = \frac{d}{dx}\big(f(x)\big) * g(x) \tag{3}$$

Once we have performed calculations according to Equations (1) (2) and (3), we can refer to the original SIFT paper to recall that they can be bypassed by simply approximating them with the DoGs. Since the DoGs depend on the scale by the inverse 1/σ2, a multiplication with σ2 offers scale invariance. As already mentioned, the detected key points are potential locations of blobs and poorly localized in subpixels due to the effect of the scale-space representation, a Taylor series expansion is utilized to allow for more accurate calculation of extrema by applying a threshold of 0.03 and a Hessian matrix is exploited to further refine key points by removing certain weak responses and outliers. The described operations signal the end of keypoint detection and localization. Once scale invariance is accomplished, the next aim is to transform key points into rotation-invariant information. This is conducted by finding the dominant orientation in a manner similar to the one described in the Histogram of Gradients algorithm [75]. The surrounding area of the key point is divided into subregions and the gradients to both xx' and yy' axis are calculated to provide measures of angle and magnitude. For each key point, a 16x8 vector is provided resulting in a robust representation of each key point of 128 values. 16 neighbor areas around the key point are described by an 8-bin histogram of gradients. Where two

dominant angles are retrieved above the threshold, both are kept as shown in Figure 5, where the radius and orientation of each keypoint are shown by a colored circle and line. It is important to keep in mind that each image is described by several vectors of 128 values. The extracted information is scale and rotation invariant but requires further processing for image classification since the ideal representation would be a single vector for each image. The same is not valid for other downstream tasks such as image registration, stitching, or retrieval where the pipeline can be successfully fulfilled with multiple vectors

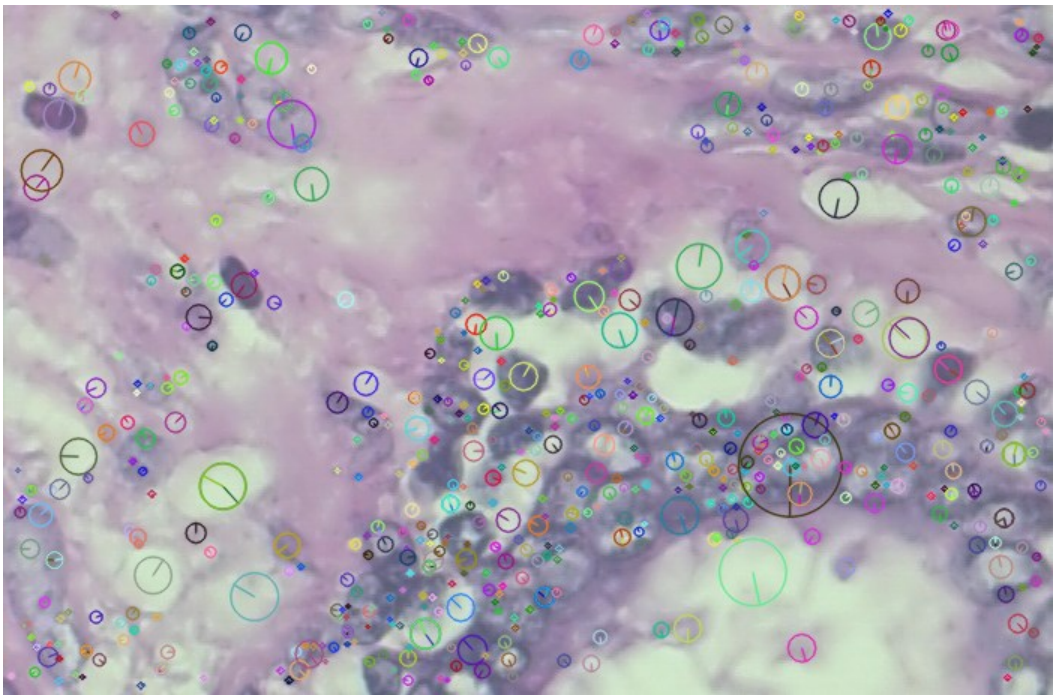


Figure 5 - Keypoint localization and orientation assignment by SIFT algorithm in a histopathology image.

1.6.2 Speeded Up Robust Features

Although the SIFT algorithm has been an important pioneering step for local descriptors that achieve scale, rotation, and translation invariance in a time and resource-saving manner, the SURF algorithm [76] pushed the limits of efficiency one step further by

exploiting integral images and box filters. The combination of these two techniques produces results that outperform the previously proposed local descriptors while maintaining the desired properties of distinctiveness, quantity, and efficiency concerning time.

SURF algorithm consists of the same steps as SIFT but fulfills each step in a different manner. It relies on the advancement of Integral images and the utilization of box filters for the detection of interest points in the image. Another novel idea that was introduced in the original paper was the utilization of the Hessian-Laplacian detector for the detection of interest points. Summing all these newly bred notions leads to a far more efficient computation scheme for the key points detection scheme. Taking things from the start, the SURF paper utilizes the Hessian-Laplacian operator for the detection of key points' location and scale. Although the SIFT algorithm accomplishes a fast computation of the Laplacian of Gaussian second derivative by the approximation of the DoGs, SURF excels in the time-efficiency perspective by approximating the LoG with box filters (Figure 6). By convolving box filters with the image, the algorithm seeks maximum responses. These responses are discovered in the zero crossings of the determinant of the Hessian matrix that consists of box filters instead of LoGs. The original Hessian matrix assumes the form with is shown in Equation (4).

Another important component of the algorithm is the Integral images that derive from the work in [77], otherwise called Summed Area Tables.

$$H = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{bmatrix}, \; H_{approx} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{4}$$

If we consider a 7x7 matrix with values ranging from 0 to 255, then we can calculate the Integral image by creating a new 7x7 matrix, where each value is the sum of values in the containing rectangle. To gain a better understanding of how the algorithm works, Figure 7 provides the original 7x7 image and the derived integral image. By applying this conversion, we can calculate the sum of pixels in a rectangle of the original image by conducting three operations instead of many more. Assuming that our objective is to calculate the sum between values 481, 393, 113, and 136, we can calculate 481-393-113-136 instead of adding all the containing values. The economy of applying fewer

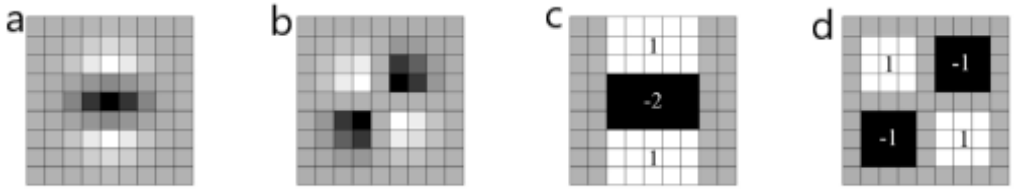


Figure 6 - Representations of approximations of Gaussian second order partial derivatives in y-direction and xy-direction. A) and b) show the original filters whereas c) and the approximated by utilizing box filters [76]

calculations is more evident to larger scale of images. To connect all the dots together, since the convolution of box filters with the image requires the calculation of sums between rectangular areas of the image, integral images play the role of an important accelerator.

The application of the SURF algorithm in the histopathology image in Figure 8 results in the detection and description of several keypoints as shown in Figure 8. The details of comparison between the two algorithms can be witnessed in Table 2. A quick review of

the two copies shows that the keypoints for SURF algorithm are more and detect larger patterns on the image than the SIFT. The examination of the quantitative results of

Original image

0	1	5	8	98	3	87
0	1	11	0	0	120	55
55	56	0	23	12	0	0
0	5	65	0	0	0	2
9	9	0	0	0	0	55
1	1	1	0	1	4	0
9	44	0	34	45	5	1

Integral image

0	1	6	14	112	115	202
0	2	18	26	124	247	389
55	113	129	160	270	393	535
55	118	199	230	340	463	607
64	136	217	248	358	481	680
65	138	220	251	362	489	688
74	191	273	338	494	626	826

Figure 7 – Conversion representation of regular 7x7 grayscale image to integral image.

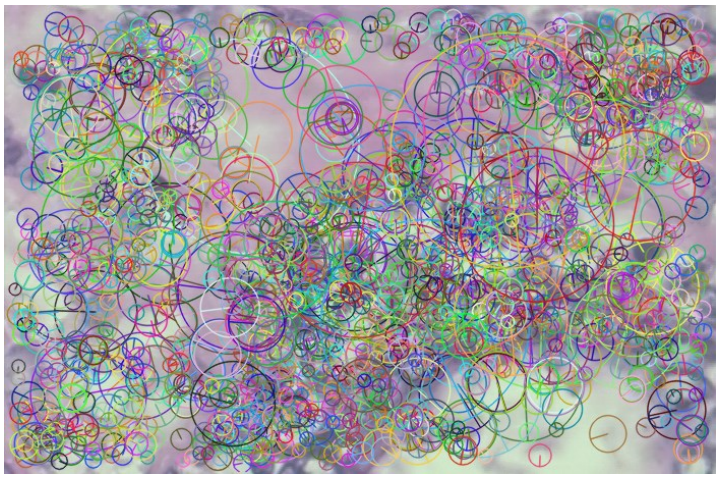


Figure 8 - Keypoint localization and orientation assignment by SURF algorithm in a histopathology image.

the comparison in Table 2 demonstrates the superiority of SURF algorithm in producing robust, compact and plenty features from an image with respect to its elder equivalent. It has been noticed that, in contrast to what is written in the original paper, the SURF

implementation takes longer for an image to accomplish its task by utilizing all known python libraries.

Table 2 - Characteristics of applying SIFT and SURF algorithm on histopathology images.

Algorithm	Feature Length	Time Allotment	Number of keypoints
SIFT	128	0.08min	716
SURF	64	0.33min	1665

1.7 Single vector representations for local descriptors

Local descriptors were invented to provide a more detailed and robust transformation representation of an image with respect to global descriptors. Although when applying global descriptors to images, we are facilitated with a single vector per image that can be easily fed into a classifier of choice, the case is more complex with the application of local descriptors such as SIFT and SURF to images. To proceed with the usual pipeline of inserting a single feature vector into a classifier, the need to encode multiple vectors detected for each key point of an image into a single one is imperative. Several algorithms have been proposed in the literature for the task in question, but the most prominent are considered to be BOVW [78], VLAD [52], and Fisher Vectors (FV) [79]. The BOVW is a direct derivation of the BOW model that was initially proposed for the encoding of words in the Natural Processing Language (NLP) field. In order to encode each image with a vector that is identifiable and descriptive, a visual vocabulary is created by exploiting the properties of well-known clustering techniques. Kmeans [80] and its offspring (J-Means, Kmeans++) is the algorithm of choice for many BOVW implementations due to its simple and effective execution. The first step of the methodology consists of the utilization of the dataset subset for the extraction of local descriptors and their gathering in a bag for the

formation of local descriptors' clusters. The selection of clusters K is a hyperparameter for the BOVW algorithm are depends on the complexity of the task we are attempting to solve. The centroid of each cluster represents a visual word of the vocabulary. In the same fashion that words are the best fit for the semantics of a certain notion, centroids attempt to capture the essence of important patterns in the image. The clustering of several local descriptors of the subset results in the formation of a visual vocabulary of K words (Figure 9). The

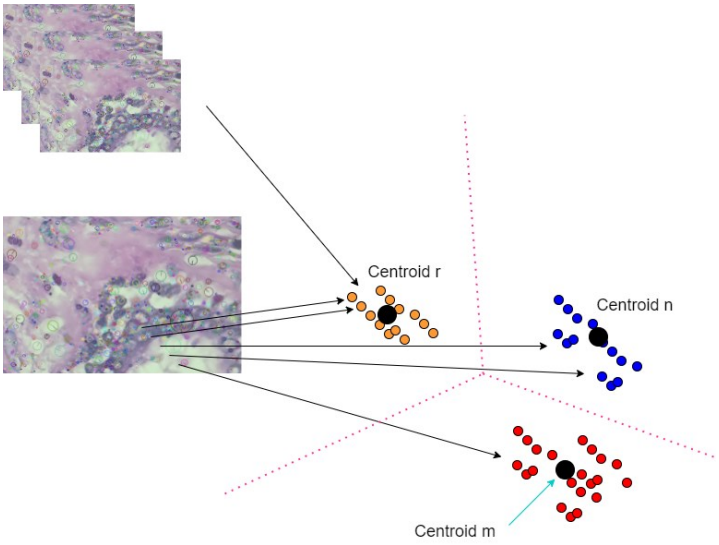


Figure 9 – Formation of visual vocabulary by extracting local descriptors and clustering by Kmeans. Black dots represent visual words of the vocabulary (clustering centroids), while colored dots are local descriptors in a 2-dimensional space.

next step is the representation of each image of the dataset by a single vector. The local descriptors are extracted from the image and each descriptor is assigned to a visual word by means of a distance metric (Euclidean, Mahalanobis). This operation leads to the formation of a histogram with k bins where each bin records the count of descriptors assigned to each visual word, as demonstrated in Figure 10. Consequently, each sample of the dataset can be inserted in a classifier for the corresponding classification task. This simplistic approach to producing a dense representation can be referred to as a hard

assignment technique since it exclusively assigns each descriptor to a visual word. Therefore, it neglects the influence of other visual words on the result. Furthermore, despite the fact that the location of each descriptor is known, no spatial information is encoded in

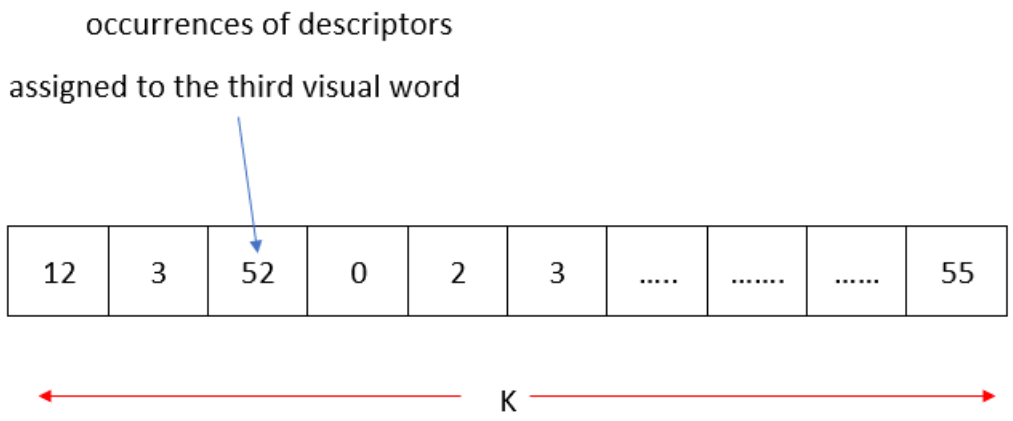


Figure 10 – Representation of histogram formation for each sample of the dataset by assigning contained visual descriptors to a specific visual word by means of Euclidean distance.

the next step is the representation of each image of the dataset by a single vector. The local representation vector, thus leading to a depreciation of useful knowledge. In order to upgrade the amount of useful information that is encoded into a representation from multiple descriptors' vectors, the authors in [52] utilize soft assignment instead of the hard assignment scheme that is exploited in the BOVW. Each image in the Vector Locally Aggregated Descriptors scheme is described by a vector that is a concatenation of vectors. This concatenation refers to the sum of residuals between to visual word of each cluster and each descriptor of the image. The improvement of VLAD against the simple BOVW technique is related to the integration of the distance information from each centroid. The first step of the algorithm refers exactly to the computation of each descriptor's distance from the nearest visual word (residual). The second step is the sum of all residuals for each

visual word, while the last one is their concatenation in a vector of kxd values, where k is the number of clusters and d is the number of the descriptor's values. Although VLAD is a hard assignment technique, the result of the whole operation is the formation of a kxd representation vector that enhances the information encapsulated by the BOVW scheme by means of the intra-cluster distance (Figure 11). However, Fisher

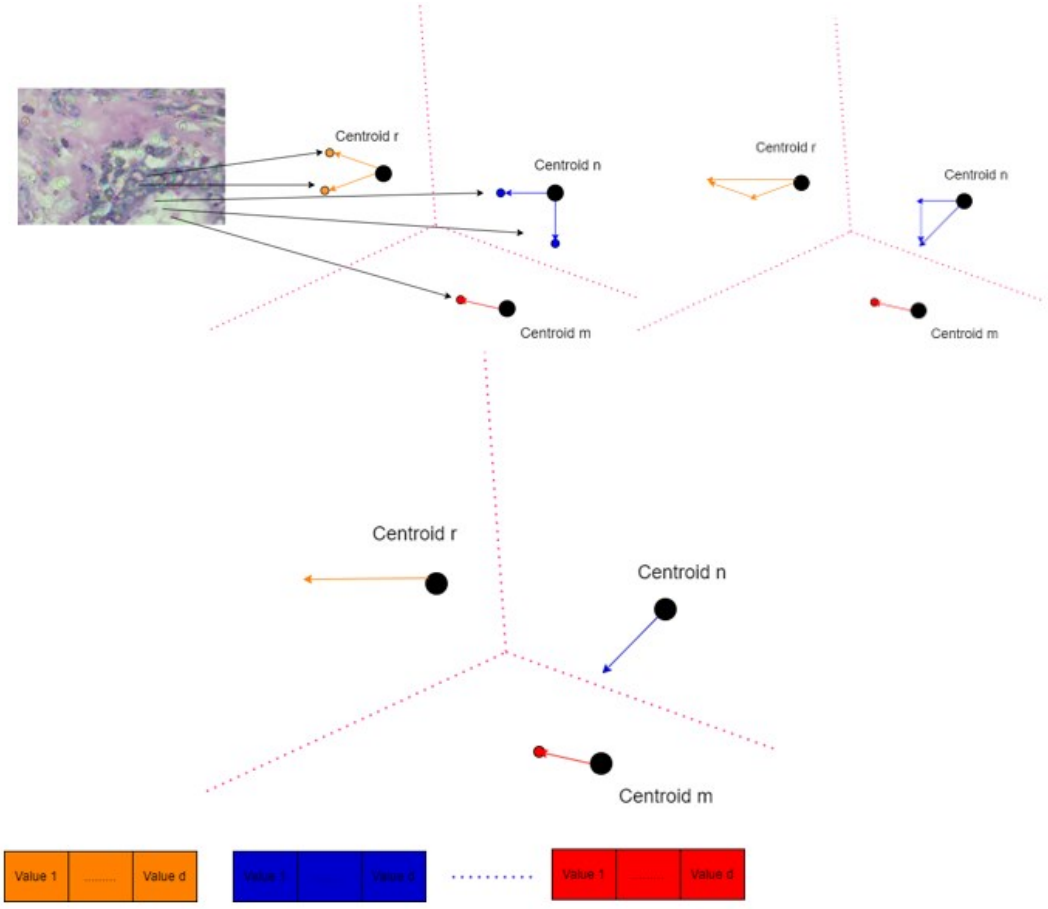


Figure 11 – Steps for VLAD implementation.

Vectors, as they were proposed in [79], present the most elaborate and sophisticated technique among the three mentioned herein. Fisher Vectors technique is the only one that applies soft assignment as each descriptor is not exclusively assigned to one visual word, but it results in a weighted sum of all visual words. The technique is a generalization of

BOVW and while the second employs zero-order statistics for the embedding of low to high-level information, Fisher Vectors incorporate in the proposed representation higher-order statistics. The technique is innovative due to the fact that it attempts to model the generative process of how the descriptors are created. The generative model, which plays the role of a visual vocabulary, is a Gaussian Mixture Model (GMM), where each mean parameter represents a visual word. Suppose n low-level descriptors are extracted from an image, then each image can be summarized in the following set X= [x_r r=1,…n]. The process through which these descriptors are created can be modeled by means of a generative model, a Gaussian Mixture Model (GMM). This Gaussian Mixture Model consists of k Gaussian distributions and their parameters λ are the following λ = [w_i, μ_i, Σ_i, i=1,…k], where w is the weight, μ the mean vector and Σ the covariance matrix of the Gaussian distribution. The probability that a random descriptor x_r is generated by the GMM in question is described by Equation (5). The algorithm attempts to compute the gradient

$$p_r(x_r|\lambda) = \sum_{i=1}^{k} w_i N(x_r; \mu_i \Sigma_i) \quad (5)$$

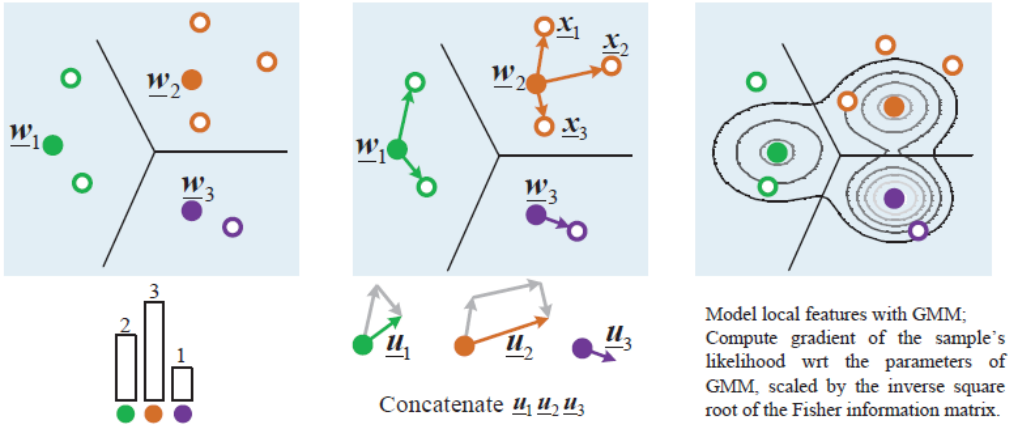
vector of each sample with respect to the model's parameters based on the probability density function p(X|λ) which models the generation process of the descriptors. Although the probability density function is annotated normally by a small p, in Equation (6) a log function is added to the notation due to the easiness of calculations. The logarithm of the pdf that models the descriptors' generative process is equal to the sum of logarithms of the probability for each descriptor x_r. The gradient vector we are seeking consists of the partial derivatives of the L function with respect to w_i, μ_i, and Σ_i. Once the gradients have been

calculated the final vector representation is the one described in Equation (7). It contains K(1+2D), where K is the number of gaussian for the GMM and D is the number of values in the descriptor.

$$L(X|\lambda) = logp(X|\lambda) = \sum_{i=1}^{n} logp(x_r|\lambda) \tag{6}$$

$$G(X|\lambda) = \left\{ \frac{\partial L}{\partial w_1} \dots, \frac{\partial L}{\partial w_k}, \frac{\partial L}{\partial \mu_1}, \dots, \frac{\partial L}{\partial \mu_\kappa}, \frac{\partial L}{\partial \Sigma_1^{-1}}, \dots \cdot, \frac{\partial L}{\partial \Sigma_\kappa^{-1}} \right\} \tag{7}$$

Fisher Vectors, in comparison to the other representation techniques that are presented herein, achieve to encode more information in a single vector for the same number of visual words. In Figure 12, a summary of BOVW, VLAD, and Fisher Vector representations is presented in an attempt to provide a straightforward explanation of how the visual vocabularies are utilized for the encoding of multiple descriptors into one.



(a) BoW (b) VLAD (c) FV

Figure 12 – Summary of representations for Bag of Visual Words, VLAD and Fisher Vector techniques [81]

1.8 Deep learning architectures for image classification

1.8.1 EfficientNets

EfficientNets are a group of deep convolutional networks that achieve and surpass state-of-the-art accuracy in different classification tasks with up to ten times better efficiency, thus the name (smaller and faster). Their main novelty lies in the latest achievement of AutoML, and, specifically, in the intelligent and controlled expansion of the three dimensions (width, depth, resolution) of a neural network by the utilization of a compound coefficient. Throughout years of research, the basic concern has been the growth of a neural network's dimensions in such a way that accuracy is improved with the minimum of operations given certain resource constraints. Even when the minimum of operations is not a basic goal, increasing the dimensions of a neural network greedily does not have the expected results due to the vanishing gradients phenomenon. Efficient Nets address this issue by exploring the relation of the increase in each dimension and applying a grid search under a fixed resources constraint instead of arbitrarily changing these dimensions. The compound scaling method is summarized in the set of Equations (8):

$$\begin{aligned} d &= \alpha^{\varphi} \\ w &= \beta^{\varphi} \\ r &= \gamma^{\varphi} \end{aligned} \tag{8}$$

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$$

$$\alpha \geq 1,\ \beta \geq 1,\ \gamma \geq 1$$

where φ EfficientNets are a group of deep convolutional networks that achieve and surpass state-of-the-art accuracy in different classification tasks where φ is a global scaling factor that controls how many resources are available and α, β, γ determine how to allocate these resources to network depth, width, and resolution respectively. By assigning φ=1 and applying grid search, α, β and γ can be determined for a given convolutional architecture to achieve better accuracy. Once concluding with the definition of α, β and γ, φ can be gradually increased to augment the dimensions of the network towards better accuracy. The scaling method applies to any convolutional architecture that consists of a repeated pattern of layers. However, the authors of the EfficientNets paper proposed a specific architecture where the main building block is the mobile inverted bottleneck convolution (MB Conv), shown in its three basic configurations in Figure 13. The base model of the EfficientNets group is Efficient Net B0 and its architecture is shown in Table 3, consisting mainly of MBConv1 and MBConv6. By utilizing MBConv blocks and increasing the value φ, the Efficient Net group reaches its most complicated form B7. In the heart of these building blocks two important innovations have found ground to act: the depthwise separable convolution[82] that performs the functionality of a normal convolution with fewer resources and the squeeze and excitation unit that enables the network to perform dynamic channel-wise feature recalibration [83]. Concerning depthwise separable convolution, the convolution operation is divided into two parts. First, the convolution is conducted depthwise, meaning that the convolution kernel is applied to each channel

individually in order to learn channel-dependent features, and second, pointwise, meaning that a 1x1 kernel is applied to each point in order to combine the channel-dependent learned features. In reference to the squeeze and excitation unit, the unit consists of two parts. Starting the squeeze part, global average pooling is applied to each channel leading to the formation of a 1x1xC vector (where C are the channels), followed by a fully connected→ ReLU→ fully connected→sigmoid block (excitation part). In this manner, each channel is enhanced with additional information concerning the other channels and captured in between interactions. Finally, the output of the excitation part is multiplied by the original input.

1.8.2 InceptionNet, XceptionNet, ResNet

The above-mentioned building blocks and architectures are learned lessons through months of development and produced experience in the ever-evolving domain of deep learning and encapsulate notions that have been partially tested and evaluated in earlier deep learning architectures such as ResNet[67], XceptionNet, and InceptionNet[84]. These approaches achieved state-of-art results in computer vision tasks because they have incorporated these blocks partially. Once combined in a structured manner by means of a controlled augmentation mechanism such as in the EfficientNets, the performance is further improved.

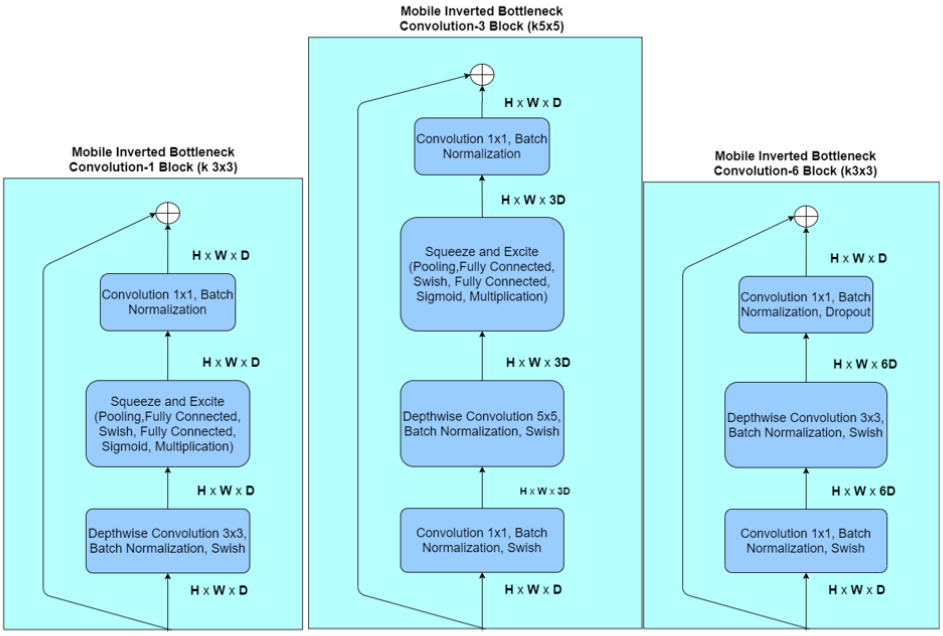


Figure 13- Three main building blocks of Efficient Nets architecture from left to right: Mobile Inverted Bottleneck Convolution-1 Block-MBlock1 (left), Mobile Inverted Bottleneck Convolution-3 Block-MBlock3 (center), Mobile Inverted Bottleneck Convolution-6 Block-MBlock6 (right).

ResNets are driven by the intuitive need for neural networks to grow deeper in order to understand and quantify more complex features and simultaneously compensate for the vanishing gradient issue. The authors discovered that, by adding the identity function between layers, the network can reach deeper architectures and cope with the vanishing gradient issue, since the layers where the gradients diminish rapidly get bypassed. Since its publishing, the idea has spread around fast and is being utilized in different deep CNN architectures including EfficientNets.

Rather than investing in deeper architectures, the authors of InceptionNet prioritized the importance of creating wider approaches, meaning filters with multiple sizes, and

leveraged their options between these two dimensions in order to capture salient patterns in the image that appears in different sizes. The initial version V1 was improved in terms of accuracy and speed by adding an auxiliary classifier during the training process, factorizing convolution operations, and placing them at a wider grid. By further improvement of the initial proposal, the InceptionNet is now transformed into its fourth version. A combined approach of Resnet and Inception is proposed by the enhancement with residual blocks (Inception-ResNet). Moving a step forward, an extreme version of the InceptionNet, called XceptionNet managed to achieve even better results, inspired by the inverse sequence of operation in the depthwise convolution (firstly proposed in Inception Net) and the removal of non-linearity between convolutional layers.

Table 3 - EfficientNet B0 architecture

Stage i	Operator F_i	Resolution H_i x W_i	Channels C_i	Layers L_i
1	Conv 3x3	224 x 224	32	1
2	MBConv1, k3x3	112 x 112	16	1
3	MBConv6, k3x3	112 x 112	24	2
4	MBConv6, k5x5	56 x 56	40	2
5	MBConv6, k3x3	28 x 28	80	3
6	MBConv6, k5x5	14 x 14	112	3
7	MBConv6, k5x5	14 x 14	192	4
8	MBConv6, k3x3	7 x 7	320	1
9	Conv 1x1, Pooling, FC	7 x 7	1280	1

1.9 Superpixel segmentation algorithms

The explainability techniques that will be described in the next section can be also divided into two categories, fine-grained and coarse, depending on the detailed representation of the generated heatmap. Concerning both techniques, the need to determine boundaries between important structures in the image is evident. Human experts assign no importance to isolated pixels nor to rectangular regions that cover mixed content

in the image. The specific requirement is to assign importance to structures that are semantically important for the expert. Superpixel segmentation algorithms can assist in better defining these boundaries according to the depicted structures. Superpixels are groups of pixels that share common low-level characteristics of the image. The grouping is achieved by means of a segmentation algorithm that classifies each pixel as being part of a homogenous and compact cluster. The ability of these algorithms to significantly reduce the number of primitives in dense and complex images has been well-appreciated in several computer vision tasks and plays an important role as submodules of state-of-the-art computer vision algorithms [85]. An overview of the relevant literature highlights SLIC [86], Felzenswalb [87], and Quickshift [88] superpixel algorithms as popular choices. Simple Linear Iterative Clustering or SLIC is characterized by the low computational resources required to achieve an efficient segmentation result of the original image into superpixels.

As its name declares, SLIC performs a local clustering of pixels in a 5-dimensional space that is composed of three coefficients of the CIELAB color space and the coordinates of the pixel. The clustering is based on a distance metric that is specifically designed to balance the outweighing between two elements, the pixel distances that refer to spatial features against the pixel distances that refer to color features. An advanced version of the SLIC superpixel algorithm, namely SLICO, adaptively decides on the compactness parameter for each superpixel in order to alleviate the issue of choosing a unique value for all superpixels regardless of the content's texture (SLIC version).

Felzenswalb superpixel algorithm is another popular choice for creating superpixels by uniting regions with common visual characteristics. The algorithm achieves the

segmentation based on specific principles of the graph theory. Each image is represented as an undirected graph G= (V, E), where pixels correspond to vertices of the graph (V) and a non-negative measure of dissimilarity is the weight of each edge (E) between two neighbor pixels. The generation of superpixels is equivalent to the partition of set V to components so that each component corresponds to a connected component in graph G' = (V, E') where E' is a subset of E. The goal is to achieve low dissimilarity measurements on edges between pixels of the same component and high elsewhere. A minimum tree-spanning clustering is exploited for the partitioning of the graph into groups of pixels. The decision for partitioning pixels is based on a certain predicate that measures the dissimilarity between pixels along the boundaries of two neighbor components relative to the dissimilarity of pixels within each component.

Lastly, the proposed methodology exploits the properties of the Quickshift algorithm for the division of images in superpixels. Quickshift is based on the same principles of the mean shift clustering algorithm for mode seeking. The algorithm utilizes a kernel for the estimation of density in an attempt to discover places in the feature space with higher density. The procedure is described as a hill-climbing process to discover the tops of highest densities (modes). In Figure 15, the segmentation results of two images that are selected from histopathology and confocal image datasets for the three superpixel algorithms are depicted. The segmentation results can vary based on selected tuning parameters and in the case of the depicted examples the baseline parameters are proposed by the relevant python libraries.

1.10 Explainability approaches

Two basic categories of approaches can be found in literature concerning the explainability task for image classification through Convolutional Neural Networks, the gradient-based and the axiomatic approaches. When utilizing gradient-based approaches to distinguish the attribution of visual patterns to the prediction result, the gradients report the degree of influence for slight variations on images' local patterns to the prediction. The responses of these methods denote the highly influential parts of the image with a color that is representative of a high influence value and vice-versa. On the other hand, axiomatic approaches are focused on defining important properties that the algorithm should satisfy in order to provide reliable explainability results. These properties are called axioms and are formally expressed in such a manner to depict the notion of relevance between the stimulus and the result. Consequently, these algorithms are based on measurable indicators that satisfy the corresponding axioms. Representative methodologies of the first type are Visualizing Gradients[89], SmoothGRAD[90], Deconvolution Networks[91], and Guided Backpropagation[45], whereas Layer-wise Relevance Propagation[92] and DeepLIFT[93] are typical examples of the second type. Nevertheless, there exists an approach, namely Integrated Gradients[94], that, despite their gradient-based role, follows required axioms, namely, Sensitivity and Implementation Invariance. Starting from the intuitive need to establish a comparable baseline to measure the effect of a visual pattern's absence on the prediction result, the sensitivity axiom is defined as follows: 'For every input and baseline that differ in one feature but have different predictions then the differing feature should be given a non-zero attribution' [94]. On the other hand, the Implementation Invariance axiom focuses on the requirement that functionally equivalent predictive models should produce identical results concerning the attribution of visual stimuli to the prediction. Within the

class of gradient-based methodologies, there is a subclass based on the class activation maps. The initial idea [45] was restrictive due to the application limitation on specific neural network architecture requirements (the need for global average pooling) but the Grad-CAM and Guided Grad-CAM explainability schemes extended it to make it applicable to a wider range of architectures. The refined limitation in these improved approaches is that the architecture between the feature maps and the softmax layer needs to be differentiable. A different taxonomy can be created by dividing techniques into those that utilize perturbed versions of the initial inputs to measure the changes and those that do not. LIME is a popular technique that can be employed in any machine learning model with even fewer restrictions, regardless of the architecture. It is based on the interpretation of perturbations of the input data since the variations in the feature values influence the predictive result to a different degree[95].

Overall, it can be stated that perturbation approaches are the most intuitive[91, 96, 97] since they ground their reasoning on the same principles as humans do. To further explain, humans attempt to find causal connections by answering the 'what if' counterfactual question which can be imitated in images by creating perturbations on the original visual content. However, there are certain drawbacks such as computational efficiency as the need to rerun forward pass for each perturbation cannot be surpassed and the inefficiency to define objectively neutral images for perturbations.

1.10.1 Local Interpretable Model-agnostic Explanations (LIME)

An interesting model-agnostic explainability technique, called Local Interpretable Model-agnostic Explanations (LIME) is proposed in [95]. The main novelty, presented in

the paper, is the utilization of a simple interpretable surrogate model in the neighborhood of the sample under investigation for the explanations of more complex predictive models. The approach can, among others, be categorized in the perturbation-based approaches since the image in question is perturbed several times to produce multiple neighbor samples, a vital step of the whole procedure.

The technique starts with the creation of a neighborhood of data samples (perturbed images) in the proximity of the image, upon which a prediction demands an explanation. Suppose that a complex predictive model generates predictions from medical images. By choosing a random image from the dataset, a new dataset is created by applying a superpixel algorithm such as Felzenswalb [87], Slic [86], Quickshift [88], or Watershed [98]. The superpixels algorithm divides the whole image into smaller regions (superpixels) to include pixels of common characteristics. By choosing arbitrarily and alternating the content of superpixels, new images are created that can be considered to be located in the vicinity of the original image. In order to better capture the process of dataset generation, Figures 14 and 15 are provided. In Figure 15, a binary classification task is depicted by coloring the spaces of the negative classes purple and the positive class yellow, while the original image is represented by a fat red cross. In the original paper, this image is denoted as $x \epsilon R^d$. Examples of such images are shown in Figure 14a where a reflectance confocal microscopy image with basal cell cancer is depicted and in Figure 14b, where a histopathology image including typical fibroadenoma patterns is presented. Around the fat red cross in Figure 15, there are several thinner crosses of various sizes that correspond to the new images, generated as a result of the superpixel algorithm's application. The outcome of the superpixel and perturbation operation is shown in Figures 14c and d for the

confocal and histopathology image respectively. For each original image, random

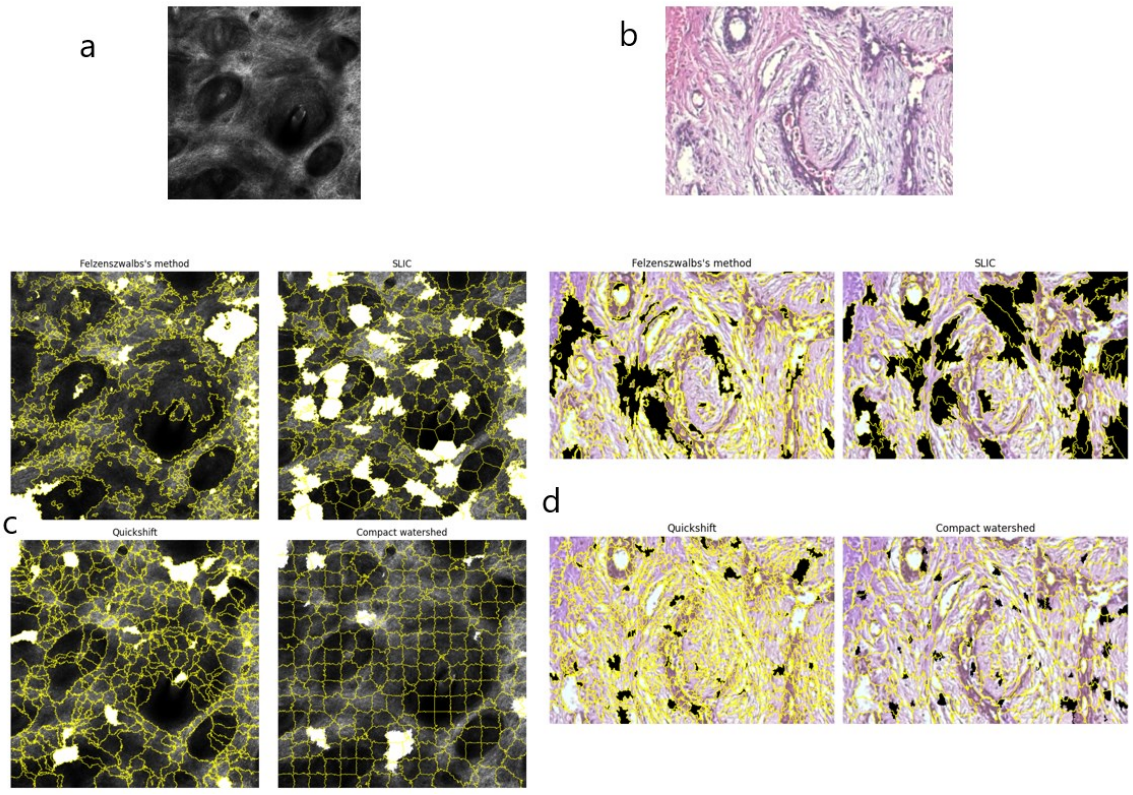


Figure 14 - Examples of perturbed microscopy images, as neighbor samples for LIME explainability approach. (a) Original reflectance confocal microscopy image that depicts a basal cell cancer pattern is presented, (b) Original histopathology image that depicts a benign fibroadenoma pattern is presented, (c) Perturbed confocal image by applying Felzenswalb, Slic, Quickshift and Watershed superpixel algorithm. (d) Perturbed histopathology image by applying Felzenswalb, Slic, Quickshift and Watershed superpixel algorithm.

superpixels are selected and colored black or white to provide slight deformations that can be quantified in the space as a Euclidean distance from the original image. The choice of color for the perturbation remains a research question, but in the paper, the black color is suggested as a standard for the MNIST dataset considering the black background with reference to the color of numbers. The algorithm requires a binary encoding for each of the perturbed images, meaning that a vector of dimension d′ is created, where d′ the number

of superpixels, and each place in the vector receives a value of 0 or 1, according to the absence or presence of the corresponding patch. In this fashion, each perturbed image is defined as $x' \epsilon [0,1]^{d'}$. The dotted pink line represents a simple linear regression model g that is utilized for its explainability properties. Two additional measures are proposed in the paper a) the measure of complexity Ω(g) to account for each model's weakness to interpret the generated predictions, and b) the measure of local fidelity, πx(z) to quantify the proximity of a newly generated data sample z near x. The whole essence of the LIME algorithm is captured in Equation (9), where f is the function, representative of the complex predictive model.

$$\xi(x) = \ argmin_g L(f, g, \pi_x) + \Omega(g) \qquad (9)$$

The explanation of sample x is the simple explainable local surrogate model (e.g., a linear regression model) that minimizes a loss function (e.g., mean squared error). This error represents the difference between the prediction of the original sample and the prediction of the generated sample that is weighted by the in-between distance. For the interpretation of each sample, a local model is created that is only accurate in the defined neighborhood instead of being locally accurate.

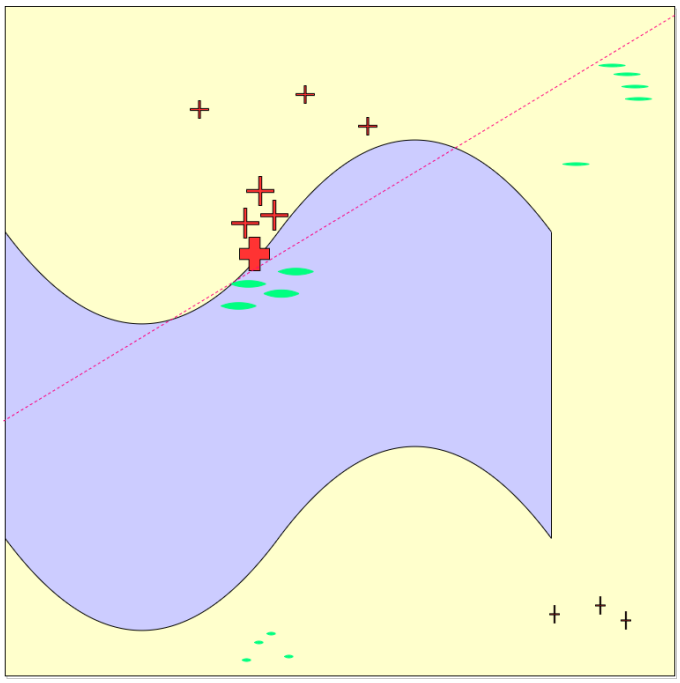


Figure 15 - Representation of LIME technique

1.10.2 Integrated gradients

The Integrated Gradients [94] algorithm is an axiomatic gradient-based approach for explainability on deep neural networks. It operates as a post-hoc mechanism on an already trained network and poses no additional modification to it. Although it shares many common points with the work presented in [99] that is followed [100], such as the use of a baseline input, and will be discussed in the following lines, the main difference is the introduction of two central axioms upon which the whole functionality and evaluation of the proposed methodology are built. As the name suggests, Integrated Gradients (IG) follow the paradigm of measuring the gradient of inputs with respect to the output of a model on a backward pass, but this calculation is conducted by selecting and cumulating

all possible values between the baseline input and the input, resulting in a resource-demanding procedure. To better comprehend the basis of the idea behind IG and the issuance of a baseline input, it would be suggested to think about how people interpret the significance of an influencing factor, meaning that they measure the effect in its absence. The 'what if' counterfactual process is fundamental in the causal analysis domain and a natural choice for extracting useful knowledge from image downstream tasks. In the IG algorithm, this absence is formally represented by the baseline input.

Integrated gradients focus on resolving specific issues that other gradient-based approaches fail to address. Suppose there is a G(x) function that is representative of the predictive models' functionality. If the input values x are selected in such a way, that their G response is constant or nearly constant, the gradient defined as dG(x)/dx will be equal to zero or very small resulting in a zero x*dG(x)/dx product, which is a measure of the input x's importance, as it is thoroughly discussed in [99] as well. According to the paper, relying simply on the gradients of the model for assigning contribution values to the inputs fails because it is incoherent to the Sensitivity axiom. The axiom dictates that "a) for every input and baseline that differ in one feature but have different predictions then the differing feature should be given a non-zero attribution and b) if the function implemented by the deep network does not depend (mathematically) on some variable, then the attribution to that variable is always zero". The second axiom, upon which the algorithm is designed is the Implementation Invariance. According to the Implementation Invariance axiom, "Two networks are functionally equivalent if their outputs are equal for all inputs, despite having very different implementations." [94]. In accordance with the property presented in [100, 101], the completeness axiom is also highlighted as an important characteristic of

explainability approaches. As such, the IG algorithm satisfies the dictated condition that the sum of importance values assigned to inputs is equal to the target output value subtracted by the baseline output value.

Following the formal definition that is presented in the IG paper, we will consider an input $X \epsilon R^n$ and a baseline input $X' \epsilon R^n$. In order to compute the value of the integrated gradient for the input X at the pixel or feature i, since we refer to images, Equation (10) is utilized. In the Equation, the fraction $\vartheta G(x)/ \vartheta x_i$ is the gradient of the predictive model's representative function G along the i_{th} pixel or feature of input X with reference to baseline input X′.

$$Integrated\ Gradients_i(x) = (x - x') \times \int_{a=0}^{1} \frac{\partial G(x' + a \times (x - x'))}{\partial x_i} da \quad (10)$$

The alpha value is the interpolation constant between 0 and 1 that is used to move from the baseline input X′to the input X respectively. By shifting from the baseline input to input we see to discover how the network change decision from assigning zero confidence to a specific class at the baseline to reaching its highest value and the pixels that played the most important role in the increase. The process of computing the integral is approximated by the utilization of the Riemann sum, as described in Equation (11).

$$Integrated\ Gradients_i^{approximated}(x) = \sum_{k=1}^{m} \frac{\partial G(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m} \quad (11)$$

By spending time understanding IG and DeepLIFT, it is verified that there is a civilized competition between to two techniques. While both rely on a baseline input, the first argues on its superiority based on the full coverage of the proposed axioms in contrast to DeepLIFT fails to satisfy implementation variance due to the use of discrete gradients. The latter showcases specific types of predictive models ("AND"/min scenarios) that IG cannot respond to efficiently. The use of a baseline input is considered an open research question as the straightforward selection of a black image or random noise cannot be deemed a neutral input for predictive models. The dependence of the approach on a baseline input (the discovery of a neutral input is not a straightforward task) and the computationally intensive burden seem to deprive of its value.

1.10.3 Deep Learning Important Features (DeepLIFT)

Among the various explainability techniques that are presented herein, DeepLIFT [100] authors argue to provide novelty in the form of example-specific explanations that are generated by means of a back-propagated manner and based on a reference input or more precisely, a difference to reference input. As a method that follows the path of back-propagation, DeepLIFT offers an efficient process to assign an importance score to each specific input for the predicted output, since a single pass is required for the computation of all inputs' importance scores whereas, in the case of the perturbation-based techniques, a single pass for each input is needed. Apart from the straightforward advantage of computational efficiency, DeepLIFT is coined with the successful handling of special cases of predictive models where gradient-based and perturbation-based approaches are deemed to fail. Such failure cases are represented in Figures 16a, b. In Figure 16a, a simple model

is shown, consisting of two inputs i1, i2 and a function that maps

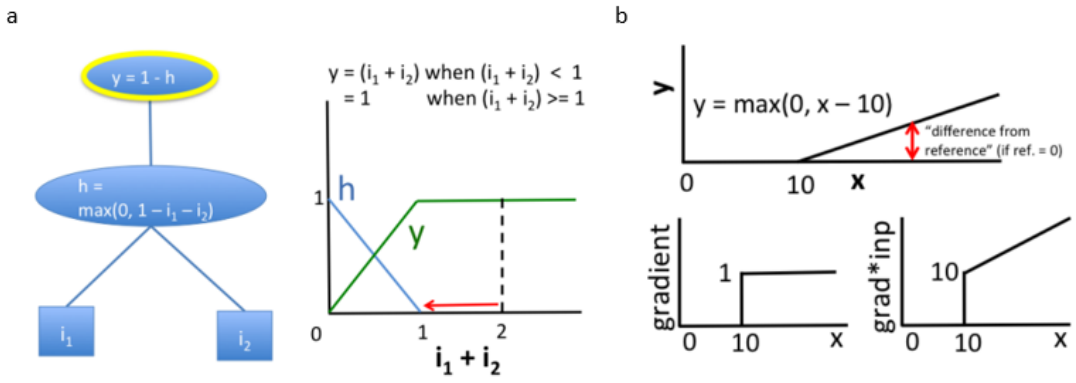


Figure 16 - Failures cases in gradient-based and perturbation-based approaches for explaining predictive models[100].

the inputs to y. If we were to select $i_1 = i_2 = 1$, as input values, then a change of either i_1 or i_2 to 0 would generate zero difference, resulting in the model's saturation failure. The same failure is witnessed for gradient-based approaches when values $i_1 + i_2$ are greater than 1. A different failure mode can be observed in Figure 16b. For the function that maps x to y by thresholding x at value 10, the gradient is 1 for all values of x over 10. A solution would be to multiply with the difference of inputs, resulting in a linear increase of the gradient with respect to increasing the input value, but in this case, the sudden increase of the gradient from 0 to 10 is misleading. In the paper, a third failure case is presented for functions of minimum [$y=\min(i_1, i_2)$], where the predictive model is trying to represent "AND" relationships.

The basis of DeepLIFT's contribution is the introduction of difference to reference notion. To fully comprehend this notion, a series of terms need to be explained. Starting from a target neuron t, a set of input neurons $I= [x_0, ...x_j]$ from layer X, t_i the result of

inserting any given input I at layer X and t0 the result of inserting the reference input at layer X, Δt, the difference from reference, is equal to ti-t0. At this point, it is clear that the "difference to reference" term corresponds to the input neurons, as Δx_n, and the target neurons, as Δt. Having defined these terms, the next step is to demonstrate that Δt is attributed to all Δx_j of the input neurons through Equation (12), which is known as the summation to delta property.

$$\sum_{n=1}^{j} C_{x_n \Delta t} = \Delta t \tag{12}$$

By assigning contributions relative to a baseline, DeepLIFT manages to alleviate the limitations, as discussed in Figures 16a, b and refer to a) the absence of x_n neuron contribution at the occurrence of $\partial x_n/\partial t$ zero gradient and b) the biases from the importance attribution due to gradient discontinuities.

Except for the "reference to difference" term, the need to connect the various layers of a predictive model in a backpropagation pass has led to the definition of "multipliers". The multiplier $m_{\Delta x \Delta t}$ is defined as the fraction of the contribution of Δx to Δt divided by Δx as shown in Equation (13). The formula is a strong reminder of the partial derivative definition except for calculating the latter on infinitesimal differences instead of finite ones.

$$m_{\Delta x \Delta t} = \frac{C_{\Delta x \Delta t}}{\Delta x} \tag{13}$$

In the same fashion that partial derivatives are subject to the chain rule, multipliers follow it as well. Given an intermediate layer R of the predictive model between layer X and the output t and a single input n of layer X, the multipliers of involved layers are connected by means of Equation (14), the multiplier's chain rule. The rule is utilized to move backward, starting from the output to concluding to the input layers in an attempt to assign contributions to each neuron.

$$m_{\Delta x_n \Delta t} = \sum_r m_{\Delta x_n \Delta y_r} m_{\Delta y_r \Delta t} \tag{14}$$

The DeepLIFT algorithm is designed for applications on both linear on non-linear functions. When referring to linear functions, such as convolutional and dense layers a simple linear rule for the propagation of contributions is applied to each step. However, that is not the case for non-linear functions, such as ReLUs, sigmoids, and tanhs. For non-linearities, the authors proposed the rescale and reveal cancel rules that can handle some or all special cases presented earlier with the respective drawbacks as shown in Table 4.

Table 4 - Properties of DeepLIFT rules on linear and non-linear functions.

Rule	Functions	Treats	Drawbacks
Linear	Linear	-	Does not handle non-linearities
Rescale	Non-Linear	Saturation, Thresholding	Does not handle min/AND
Reveal-Cancel	Non-Linear	Saturation, Thresholding, min/AND	Subject to noise

It is vital to mention that for the treatment of different failure cases, the DeepLIFT algorithm is introducing the notion of negative vs positive contributions. In the Rescale

rule Δy^+ and Δy^-, the positive and negative part of the "difference to reference" output, are considered to be proportional to the Δx^+ and Δx^-, the positive and negative part of the "difference to reference" input, whereas, in the reveal-cancel rule, they are calculated as an improved approximation of Shapely values.

The DeepLIFT approach has been proposed to improve the shortcomings of prior explainability approaches and to a great extent it achieves the goal. However, its dependence on the reference input for the difference to reference scheme requires thorough analysis for which little work has been accomplished in the relevant literature. The authors provide some feedback on their considerations and experiments with different reference inputs for specific datasets (for MNIST[102] dataset all zeros à to measure differences against the background), but this is far from a detailed and documented report on factors and limitations for the choice of reference inputs.

1.10.4 Grad-CAM and Guided Backpropagation Grad-CAM

One way to achieve the goal of extracting localization information of important visual patterns for decision-making is the construction of class activation maps [103]. Class activation mapping is a method that indicates the discriminative regions of an image that influenced the predictive model in reaching its final decision. Initially, the predictive model needed to follow a certain architecture for the technique to provide plausible results, meaning that the output of the convolutional layers should be directed to a global average pooling layer and then directly to the SoftMax activation function. This architecture, as discussed earlier, demands retraining of the predictive model and sacrifices complexity (added by the insertion of fully connected layers) for explainability. A generalization of

this method (Grad-CAM) is proposed in [45]. In the same paper, the combination of Grad-CAMs with the guided-back propagation technique is proposed to provide fine-grained pixel-to-pixel visualizations. This approach fits better with the visual characteristics of digital pathology images, where the patterns correspond to small cellular structures as opposed to larger structures. By computing the gradients for the score of each class with respect to the feature maps from the last convolutional layer and performing global average pooling on them the importance weights for each feature map are obtained. In this fashion, the architecture of the predictive model remains intact. When utilizing the Grad-CAM technique in a single classifier environment the feature maps of the last convolutional layers and the gradients for the score of each class with respect to the feature maps are necessary to produce a heatmap with the explainability visualizations. As explained in [45], the technique can be divided into three steps. The first step refers to the calculation of the gradient G (Equation 2), where Y^c is the raw output of the CNN before applying softmax to turn it into a probability and A_k is the generated feature map activations. c is the class indicator for which the heatmap is generated since the technique is class-dependent and k reflects the number of utilized convolution filters. An important requirement that needs to be addressed for the technique to be valid is that the layers after the final convolutional layer up to the softmax should be differentiable (Figure 17). The second step is the calculation of alpha values (Equation 3). This operation is performed by applying global average pooling on the gradients G. Z parameter registers the number of pixels in the feature map. To provide an intuition of the technique, it is important to note that the technique utilizes the information of the gradient's value flowing into the last convolutional layer of the network to assign importance values to each neuron for a particular decision

of interest. The third step rests on the application of ReLU on the product of each feature map with the corresponding alpha value (Equation 4).

$$G = \frac{dY^c}{dA^k} \tag{2}$$

$$a_k^c = \frac{1}{Z} \sum_{i=1}^{v} \sum_{i=1}^{u} \frac{dY^c}{dA_{ij}^k} \tag{3}$$

$$L_{Grad-CAM}^c = ReLU(\sum_K a_k^c A^k) \tag{4}$$

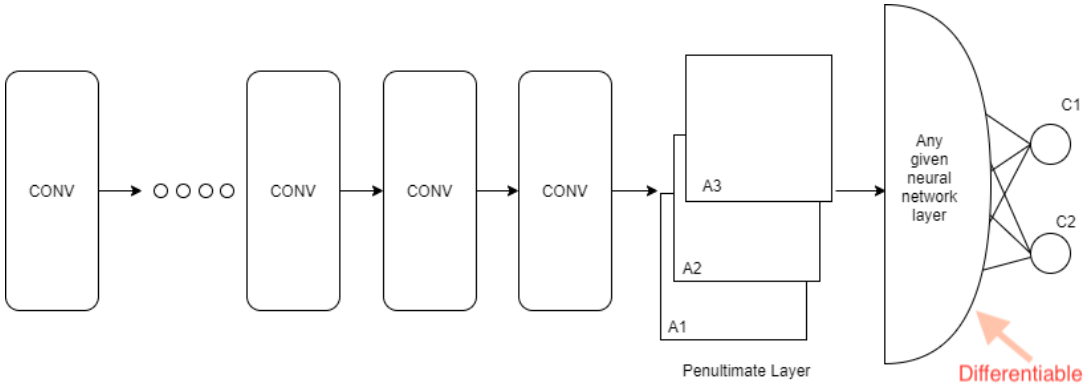


Figure 17 - Architecture of a CNN for the Grad-CAM to be applicable. The number of feature maps is set to three for visualization purposes.

Apart from the calculation of Grad-CAMs, an independent procedure is conducted in parallel, namely guided backpropagation. Guided backpropagation is the combination of two distinct operations. The first is the backpropagation at ReLU activation functions. This backward pass ensures that values being greater than zero during the forward pass in the -1 filter are passed as is one step backward. The second operation is deconvolution at ReLU. Values greater than zero in the current filter are passed as one step backward. To reach the final heatmap, the results of guided backpropagation and Grad-CAM are multiplied. In

contrast to [104], where the authors presented a weighted patch ensemble method that requires the modification of the ensemble classifier for the integration of the explainability scheme, the proposed methodology maintains the classification scheme as-is. This is an important feature to consider since the alteration of (removal or addition) layers may significantly influence the performance of the classifier. Therefore, leaving the neural network intact when integrating an explainability scheme is an important advantage.

Apart from the initial implementation of Grad-CAM, several other modifications have been proposed based on the basic principle of calculating gradients: HiResCAM [105] which is like Grad-CAM but the activations with the gradients are multiplied in an element-wise manner, Grad-CAM++ [106] that uses second order gradients, XGrad-CAM [107] that scales the gradients by the normalized activations, AblationCAM [108] that zeros out activations and measure how the output changes, EigenCAM [109], LayerCAM [110] and so many others.

1.11 Ensemble classifiers

The ensemble classifiers notion lies on the founding principles of democracy as it was first established in ancient Greece. The Greeks did not need much to realize that the best decision is reached only when many opinions (the opinions of people) are heard and processed. This simple yet efficient idea has become for modern humans merely an intuitive action since on the verge of taking an important decision, they demand the opinion of several experts. But if we were to let alone the empirical and intuitive evidence, literature in the health informatics domain proves in a placid way that classifiers produce more accurate results when they have gathered together and their predictions - opinions are

combined in different ways to reach a final result [39-43]. The manner utilized for the combination of different base classifiers is one of the basic criteria for characterizing ensemble classifiers. The basic classification of ensemble classifiers consists of the following three major categories, bagging, boosting, and stacking. The name of the first class, Bagging, derives from the words Bootstrap and Aggregation. Bootstrapping is a well-known sampling process according to which samples are iteratively selected from a population with replacement, while aggregation refers to the addition of all partial results into an outcome by means of a deterministic function. The method dictates the division of the initial dataset into subsets and their utilization for the training of base classifiers. This process follows the divide and conquer paradigm as the base classifiers are trained to learn simpler tasks and are, in turn, combined to reach a final solution (Figure 18). In total, bagging is based on a parallel and independent learning procedure of base classifiers that are in turn combined as dictated by a deterministic averaging process, while boosting corresponds to a sequential adaptive learning method that adaptively modifies the distribution of the training set based on the performance accuracy of previously trained classifiers [42]. Boosting models can be presented as sequential models in contrast to the parallel workflow of bagging schemes. In this case, each following base classifier assumes the load of decreasing the error of the previous classifier. This can be accomplished by assigning larger weights to prior erroneous predictions or utilizing the gradient descent to direct the scheme toward lower error (Figure 19). Stacking refers to a parallel learning algorithm that results to a training of a meta-model. This meta-model is responsible for the combination of base learners' predictions. Another aspect of categorizing the different

types of ensembling methods is related to the input patterns. Utilizing different classifiers,

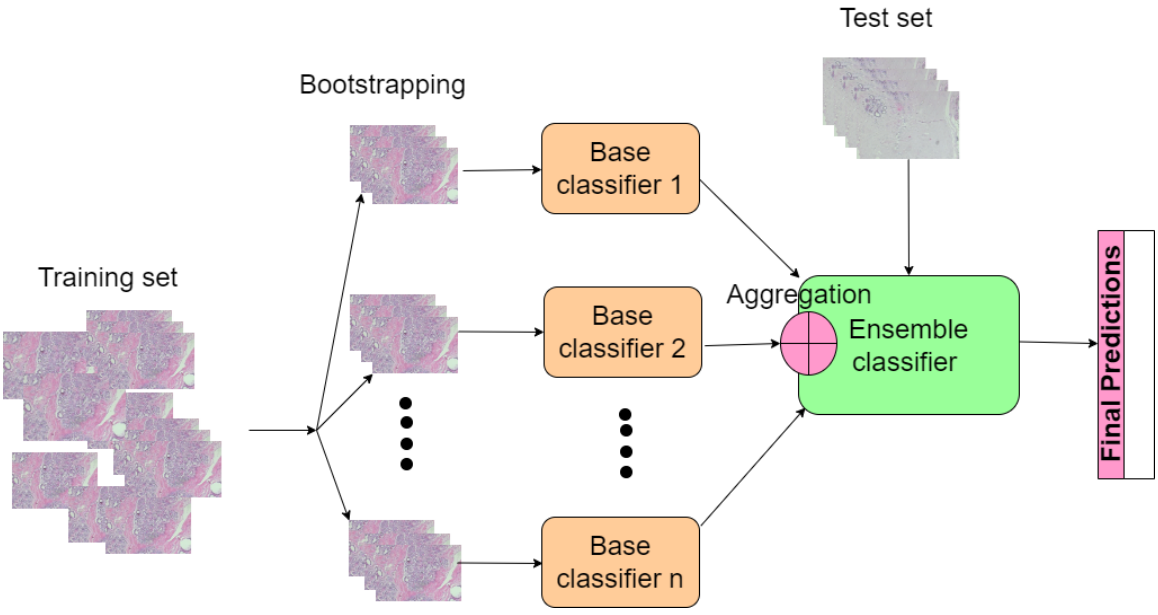


Figure 18– Basic workflow for bagging predictive models.

whereas one is trained with the original input and others with modified input versions is common practice [43]. Another aspect categorizes ensemble classifiers as those that utilize different classifiers to solve the same task and those that break the original task into subtasks and employ a different classifier for each decomposed problem [44]. Moving further to distinguish ensemble classifiers by means of the manner between base classifiers achieves diversity. There exist randomized methods to populate an ensemble classifier with other classifiers and metrics-based techniques with a main concern to increase diversity to a certain extent that does not harm performance [45, 46]. To remember our basic instincts with reference to machine learning and what our goals are, a very important property that

our model should have the ability to generalize. Ensemble models are created to enhance

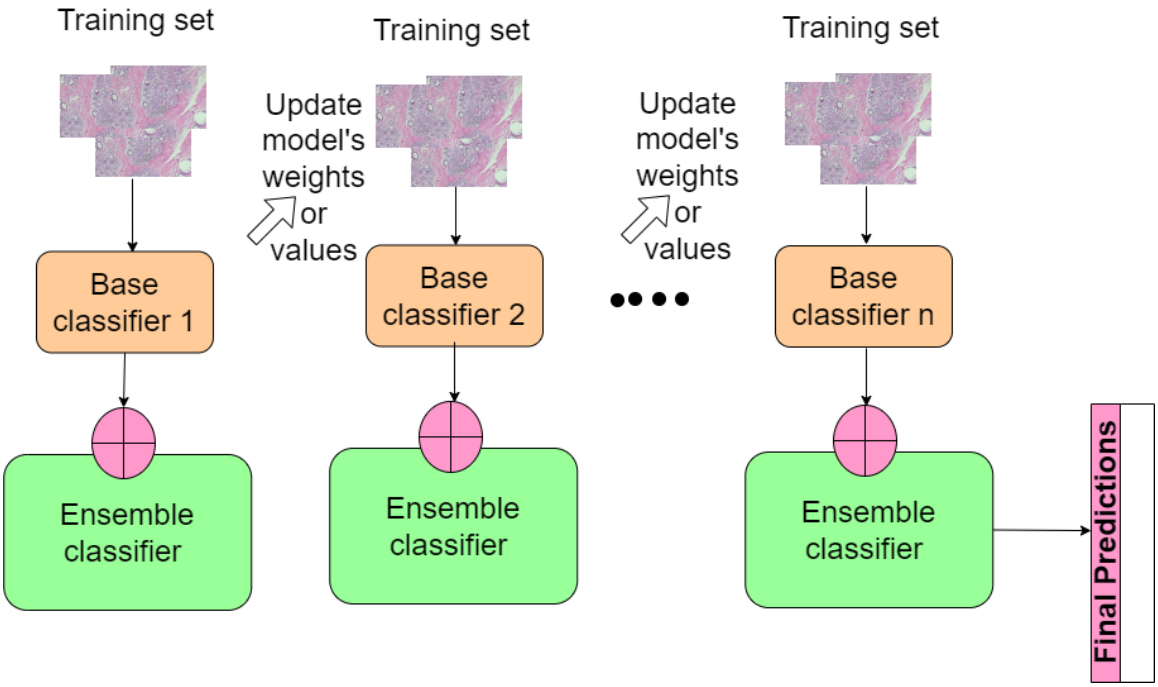


Figure 19 - Basic workflow for boosting predictive models.

the generalization properties of weaker base models. Bagging models are very effective against the variance error while boosting models address the bias error.

While ensemble models have been widely utilized for knowledge extraction from medical images few references address the explainability issue for these schemes. In [111], the authors apply a novel stacked ensemble model that utilizes the combination of traditional machine learning and deep learning techniques for the extraction of useful features from chest X-Rays. Global features such as Histogram of Gradients and GIST along with SIFT local features and learned features for pretrained DCNNs are jointly inserted in a Support Vector Machine classifier and its results are, in turn, inserted into a linear regression model. The evaluation shows improved results in terms of the binary classification task, but no interpretation scheme is provided for the generated results. The authors in [112] proposed

the EMS-Net for the classification of breast histopathology images by resizing the images and creating inputs at different scales for DCNNs to digest. The results are promising as expected, however, the lack of explanations concerning the visual stimuli that led to these results is evident. By reviewing the relevant literature, several proposed methodologies address and improve prior results while neglecting the significance of explainability in contrast to the bloom that is witnessed regarding the enhancement of plain deep learning models with explainable properties [113]. However, some works focus on the implementation of deep learning architectures with explainable results based on class activation maps [114, 115] or Integrated Gradients and SmoothGrad [116].

1.12 Unsupervised segmentation

Far from the modality of microscopy images where the region of interest usually occupies all given pixels, there are other medical imaging modalities such as dermoscopy images or blastocyst images where the region of interest (ROI) is concentrated in only one part of the image. The requirement to isolate the pixels in question is an arduous task to address, mainly owing to the difficulty of manually labeling ROIs in an attempt to provide significant ground truth for the supervised algorithm. A key step in the proposed methodologies is the extraction of features from the images. Handcrafted features are designed with the primary purpose of quantifying the visual patterns that experts (embryologists) have identified as important based on their experience. When utilizing handcrafted features based on local descriptors the isolation of ROIs is an effortless task to accomplish since it requires no further modification of the image and the localization information as part of the local descriptor algorithm. On the other hand, recent developments in Artificial Intelligence show that in many cases DL approaches go beyond

human vision and achieve better performance in terms of accuracy and repeatability [9]. However, in the case of learned features the isolation of features that make part of the ROI cannot be accomplished since the engineer cannot intervene with the feature extraction mechanism. Therefore, techniques that involve alteration of the original image are employed for the masking of the areas that are not to be included in the analysis. Another drawback of deep learning techniques is the fact that they cannot digest images greater than 500x500x3 pixels or else they demand ridiculously large computational power. For that reason, original images are often rescaled to smaller ones in order to comply with the size requirement.

In the use case of blastocyst images, image segmentation is deemed necessary as in almost every image the ROI is contained in a specific area excluding other surrounding visual stimuli. For the segmentation of medical imaging, often is the case where U-Nets are involved in the classification of pixels into the background and foreground classes. Their application in the relevant field of medical imaging is numerous and is often improved by the utilization of conditional random fields (CRFs). One of the papers that are considered a breakthrough in terms of achieving a high rate of performance was presented in [117] with the proposal of a symmetric encoder-decoder architecture that allows for the passing of information between opposing contracting and expanding sides by means of a residual connection (Figure 20). The application of equivalent architectures or more elaborate variants [67, 118, 119] cannot be the exception to the rule for blastocyst segmentation such as in the case of [120]. In [121], a shallower convolutional architecture than the state-of-the-art is utilized for the segmentation of different blastocyst components by means of a labeled dataset. The basic structuring block of the encoder is a sprint convolutional block

that utilizes asymmetric kernel convolutions in combination with depth-wise separable convolutions to achieve a decrease in the number of utilized learning parameters. Many existing DCNN architectures that are utilized in the medical imaging field (although rarely found for blastocyst images) tend to explore the technique of conditional random fields for the improvement of the generated results. These architectures for image segmentation present certain inconsistencies such as isolated areas of a different label in contrast to their surroundings. This negative phenomenon can be addressed by the utilization of prior knowledge that can be encapsulated in the form of regularization terms that can be learned from the dataset. The whole architecture can be trained as an end-to-end deep learning model that combines the cost function of the initial classifier (unary potential) with the penalty of two adjacent pixels being classified differently (pairwise potential).

Although the presented results are more than promising, all the above-mentioned techniques suffer from the ever-laborious process of manually classifying pixel-by-pixel large image datasets. This restriction has led many researchers to search for and development of segmentation techniques that require no supervision.

Most of the studies manage to differentiate blastocysts coarsely (good vs bad quality), but the critical clinical need of distinguishing blastocysts of similar quality remains unsatisfied.

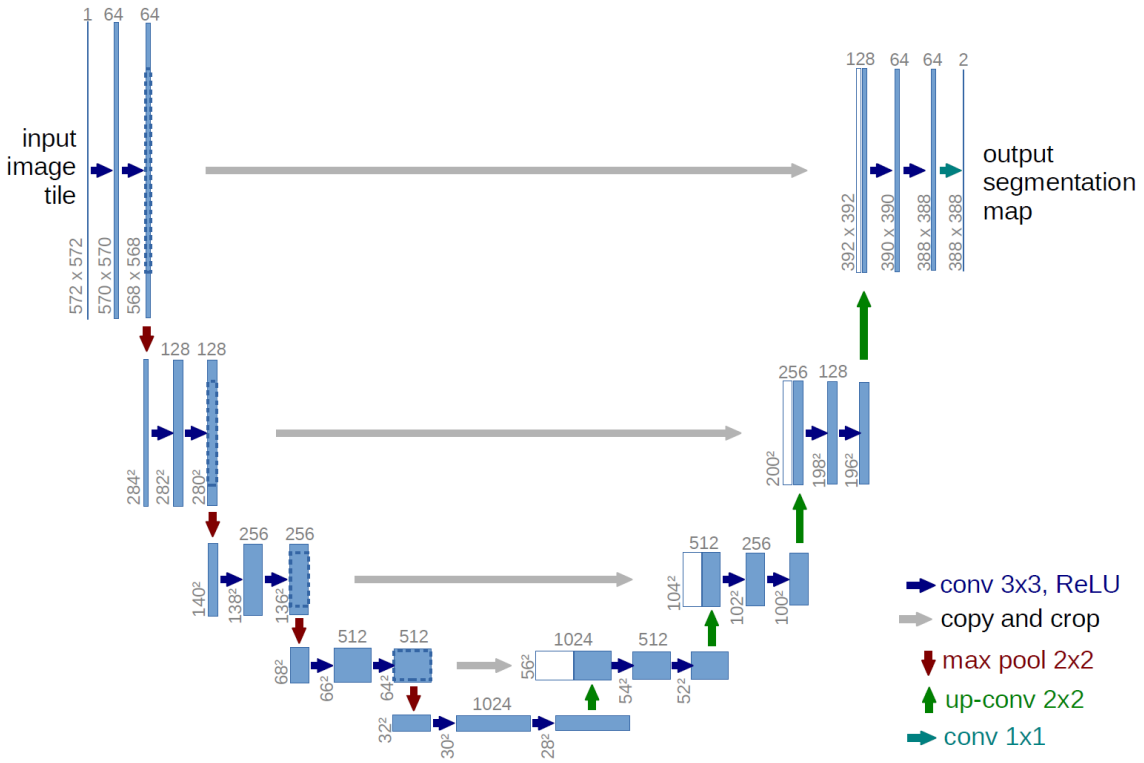


Figure 20 - U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations [117].

Furthermore, they fail to interpret the internal mechanisms of the utilized predictive models, meaning that no explanation is provided concerning the visual patterns that influenced the prediction. The association and localization of the visual patterns that greatly influence the result of the predictive model are crucial in high-risk predictive models, such as those utilized in medical applications. Explainability contributes to the discovery of reasoning for erroneous predictions or confounding factors [13]. In addition, customers, who intended to invest in such systems, are entitled to explanations that can cross-examine their working experience [14], a legitimate right that is manifested in the General Data Protection Regulation as well. Among several approaches for the generation of visual

explanations in medical images, Gradient Weighted Class Activation Mapping (Grad-CAM) [15] is utilized broadly since it can be applied to various configurations of DL architectures.

METHODOLOGY

1.13 Classification Techniques

1.13.1 Traditional Machine Learning for Medical Image Classification

In all the proposed techniques the contribution lies in the ability to incorporate explainable properties on existing well-defined classification schemes. In the following lines, the detailed description of each classification scheme offers a broader view of the inner mechanisms of feature extraction and classification processes that are exploited for the explainability scheme and are therefore deemed necessary. As far as the traditional machine learning approaches are concerned, a selected group of local descriptors is utilized alone or in combination for the feature extraction process to provide a repeatable and robust quantification of the visual content. The handcrafted features of choice are the following:

- SURF. Local SURF descriptors are utilized for the generation of multiple vectors of 64 values. The utilization of SURF is relieved the burden of unnormalized color, which is an unsolved challenge for histopathology images, deriving from different labs, and offers significant robustness for scale variations.
- Haralick features. Histopathology and confocal images are characterized by rich information expressed in the form of texture. The grayscale cooccurrence matrix offers a compact summary of the texture-based information in medical images by means of applying higher-order statistics to it.
- Color Moments. Since color information is not retrieved by the earlier feature extractors color moments are utilized as a means of color quantification as a three-valued vector.

Global descriptors are applied to the image as a whole and therefore no explainability information is gained from them. On the contrary, local descriptors return several interest points along with their coordinates and can be utilized for explainability purposes, assigning an importance value to its blob of interest which is described not only by its coordinates and the SURF descriptor but their scale as well. This scale is indicative of the radius of each interest point and can be utilized for visualizing the area of visual stimuli that influenced the final result. In total, the application of handcrafted descriptors results in the description of medical images with multiple localized vectors deriving from local descriptors and a global vector for each global descriptor.

For classification purposes, the need to create a final representation vector for each image is addressed by encoding the multiple local vectors into one and, in turn, fusing this vector with the two global ones. The process of transforming multiple vectors into one is managed by means of the BOVW and its more advanced alternatives, namely VLAD and Fisher Vectors. The contribution of this work is focused on the enhancement of this procedure with explainability properties and is based on the idea of extracting useful information from the clustering mechanism that is utilized to create the visual vocabulary. The classification scheme follows the pipeline of a vocabulary-based vector-embedding mechanism which is highly influenced by the number of visual words. This value is the only hyperparameter of the system and can be assessed by utilizing techniques such as the Silhouette index[122], the Variance Information Criterion[123], or the Bayesian Information Criterion[124], providing significant intuition towards the selection of this parameter. The final representation vectors for each medical image can be utilized for image retrieval purposes

as well by the assignment of a resemblance value according to the distance between image vectors.

An interesting case of the proposed ML pipeline is showcased in the medical imaging modality of histopathology by creating visual vocabularies for retrieval and classification [31]. A general overview of the proposed system is shown in Figure 21. Whole Slide Images, images extracted from digital atlases (literature), and images locally stored or digital pathology databases can be imported into the system, which is separated into two subsystems, the content-based image retrieval component, and the image classification component. These inputs are processed, and the output of this process is either a set of images that bear the greatest resemblance to the query image, or a set of images with the corresponding malignant/benign labels.

The method to address the image retrieval and classification problems consist of five (5) stages, namely:

- Image preprocessing. Pre-processing of the images consists of transforming the various inputs into the query image-annotated images dataset and it is performed in the case of digital atlases and Whole Slide Images with the utilization of PDFBox and Openslide [32] libraries accordingly. Consequently, all images are transformed into appropriate java objects for their manipulation by the ImageJ library [125].
- Image analysis. By means of a fast Hessian Detector, a set of interest points are detected in all images (query dataset from database/training-test set).
- Feature extraction. From each interest point, detected in the previous step, a vector with 64 values is extracted which uniquely describes the interest point. Nevertheless, a vast variety of features can be utilized locally or globally in order to combine the invariant

robust characteristics of the SURF algorithm with advantages provided by other known descriptors such as Haralick features and Color Moments.

- Creation of Visual Vocabulary. Given a collection of r images, an algorithm that extracts local features is utilized to create the visual vocabulary (Visual Vocabulary). In our case the Speeded Up Robust Features (SURF) algorithm extracts vectors (64 or 128-valued) where n is the interest points which are automatically detected by using a Fast

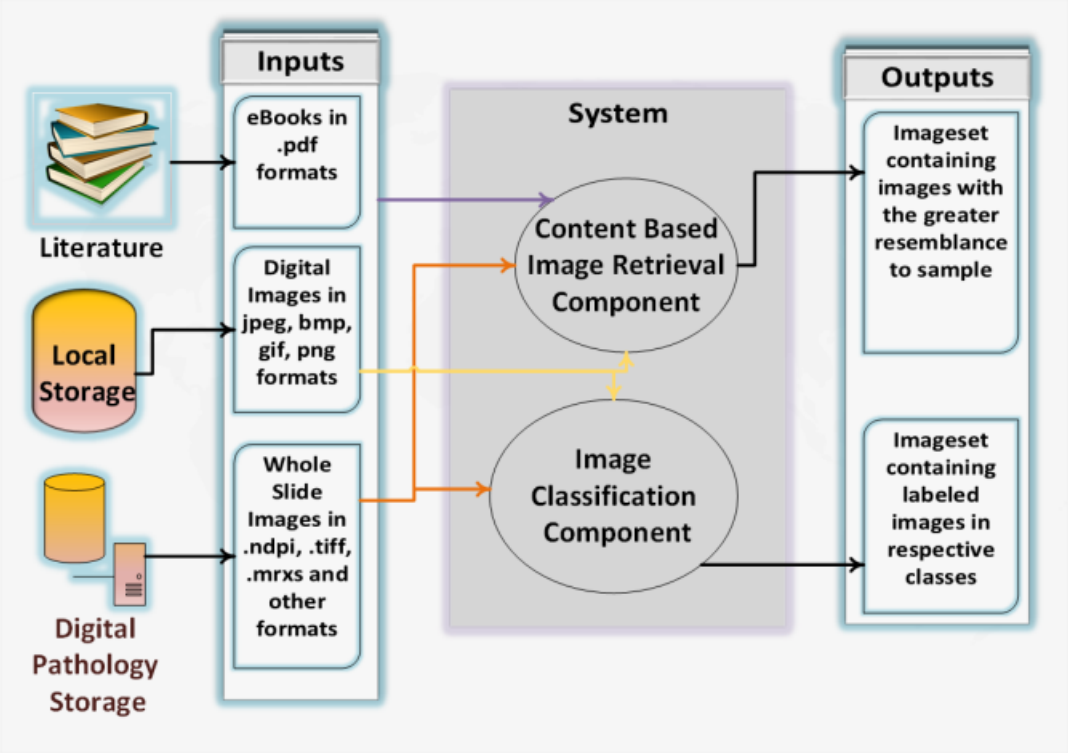


Figure 21- Overall system architecture.

- Hessian Matrix (SURF Descriptor) in each of the r images. Upon completion of the feature extraction process from the r images, a collection of r x n 64-value vectors is formed. This collection is grouped using a clustering algorithm (here k means is utilized) in k groups. The centroid of each group represents the visual word, resulting in the formation of a visual vocabulary of k visual words as shown in Figure 22.

Image retrieval/Image Classification. Image. Each image is represented by a vector computed by the number of occurrences of the image's interest points to each visual word. In order to retrieve the images with a greater resemblance to the query image, a metric is utilized (Euclidean or Mahalanobis distance) as depicted in Figure 23. The distance is calculated between the vector of the query image and the vectors of all other images included in the dataset with the sample images. Sample images with a vector closer to the vector of the query image are the retrieved ones. In order to classify a collection of test images in c classes, a model is created by exploiting the a priori knowledge provided by a training set (Figure 24). In this model, each image is defined by a duple of one characteristic vector and a label. The label of each image corresponds to one of the c classes. The model is formed using the training set as follows: The training set consists of f images. To each of the f images, BOVW is applied with a predefined visual vocabulary of k words, resulting in the extraction of a k-value descriptor of the image. In order to classify a collection of test images in c classes, a model is created by exploiting the a priori knowledge provided by a training set. In this model, each image is defined by a duple of one characteristic vector and a label. The label of each image corresponds to one of the c classes. The model is formed using the training set as follows: The training set consists of f images. To each of the f images, BOVW is applied with a predefined visual vocabulary of k words, resulting in the extraction of a k-valued descriptor of the image. The utilized classifiers are the K-Nearest Neighbor classifier and the Random Forest.

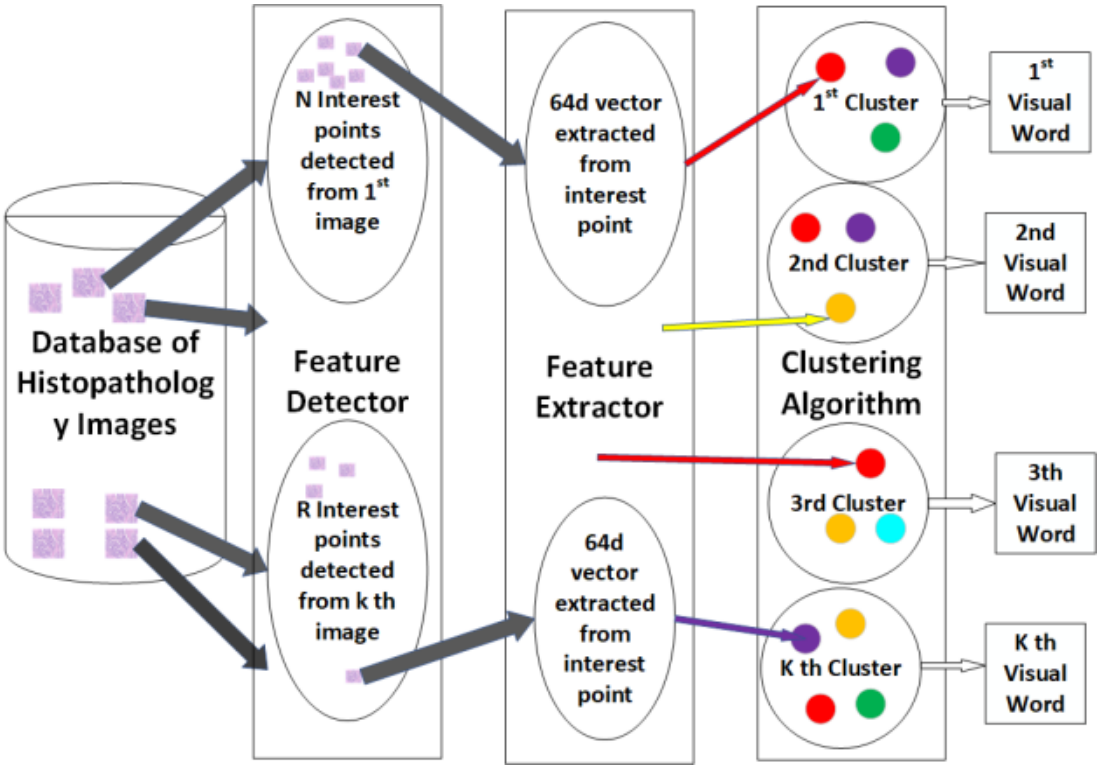


Figure 22- Creation of Visual Vocabulary.

The BOVW measures only the number of correspondences of each interest point to each visual (Term Frequency-TF). Thus, it does not consider the significance of each visual word in the categorization of each image. In order to measure the significance, Inverse Document Frequency (IDF) is calculated, which leads to the enhancement of the provided information of each visual word (Equation 18). N is the number of images in the training set and N_{con} is the number of images in the training containing a specific visual word. The TFxIDF product (Equation 19) is the equivalent weight (W_{vw}) attached to each visual word (VW).

$$IDF_{vw} = log\left(\frac{N}{N_{con}}\right) \tag{18}$$

$$W_{vw} = TF \times IDF \tag{19}$$

In the proposed system Vector Locally Aggregated Descriptors (VLADs) are also utilized. Contrary to BOVW, where the final image vector is created by the correspondences of the interest points of the image to the visual words, in VLAD the image vector is created by the sum of differences between the interest point descriptor and the visual word.

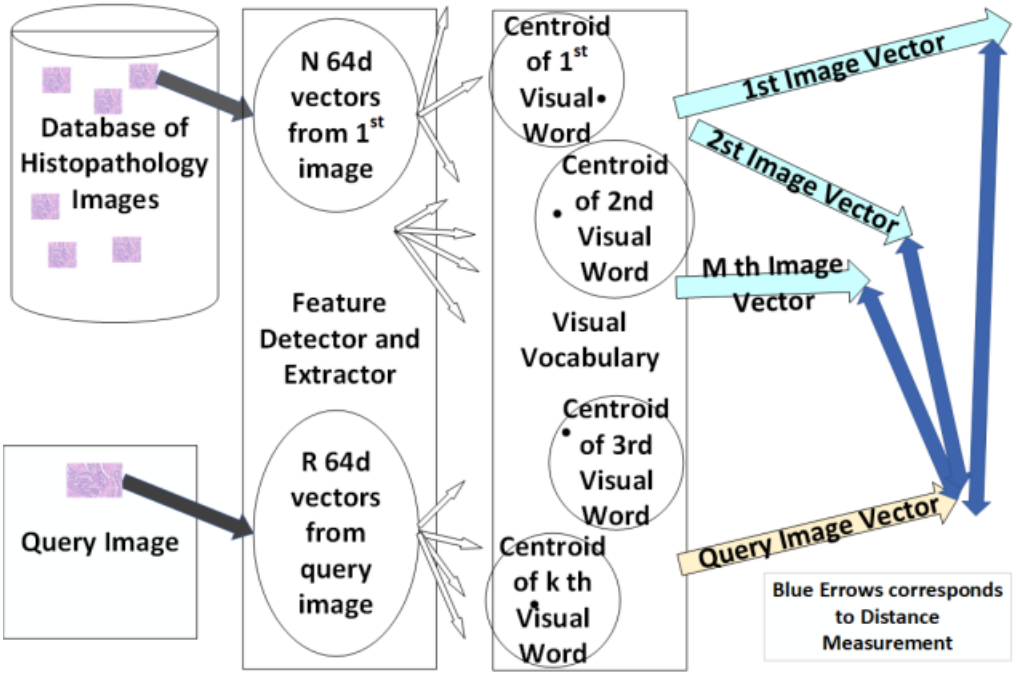


Figure 23 - BOVW Image Retrieval.

BOVW implementation with the utilization of SURF features leads to results of high accuracy. Nevertheless, the final vector can be improved by adding information such as the texture Haralick features extracted using gray-level cooccurrence matrices (GLCM). In this way, it is possible to exploit the information concerning the texture of the represented structures, which takes the form of a vector of 14 statistical characteristics (Angular Second Moment, Contrast, Correlation, Sum of Squares, Inverse Difference Moment, Sum Average, Sum Variance, Sum Entropy, Difference Variance, Difference Entropy, Information Measures of Correlation, Maximal Correlation Coefficient). Since both

algorithms, SURF and Haralick are applied to grayscale images, the color information is not exploited. To further enrich the generated vector, Color Moments are extracted from each image. To calculate Color Moments four low-

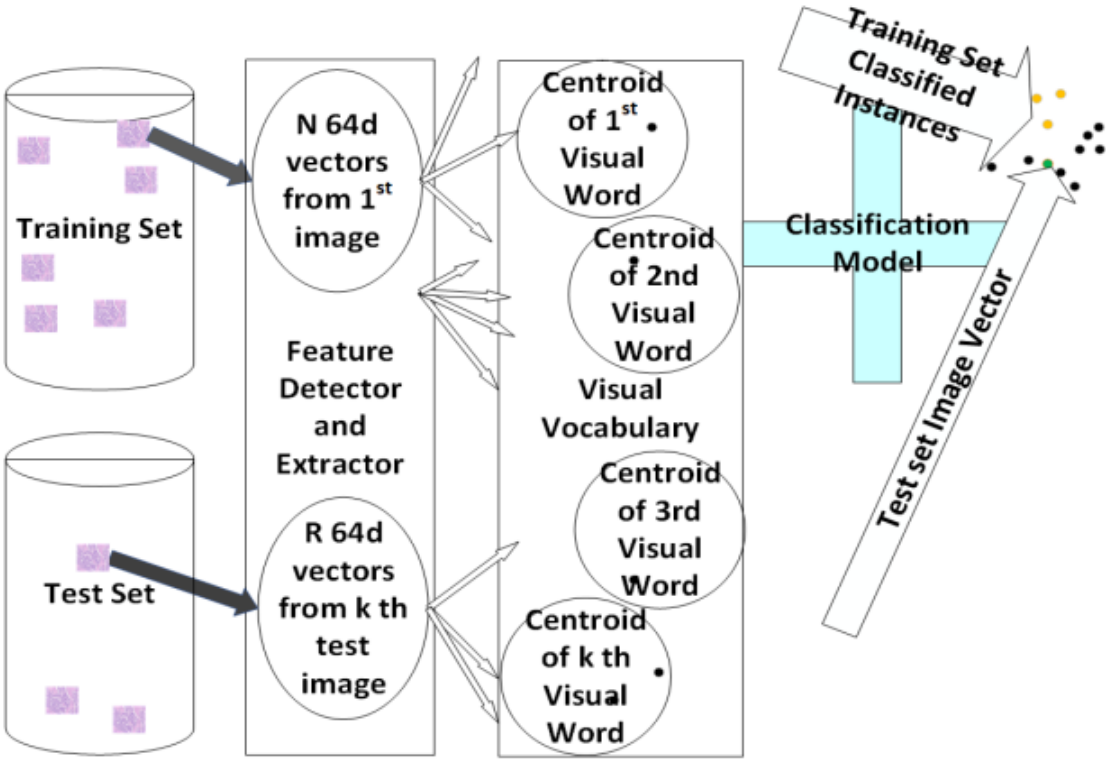


Figure 24- BOVW Image Classification

order statistical measures (Mean, Standard Deviation, Skewness, and Kurtosis) are extracted globally from each image. The vector representing the image can be further improved by adding weights to each visual word in accordance with the term frequency definition explained in the field of text categorization. Given a visual vocabulary that contains visual words, each interest point of an image is correlated with a visual word. The number of interest points correlated to each visual word corresponds to the term frequency.

The same ML pipeline is utilized in [68] for the classification of skin cancer confocal images in an explainable manner (Figure 25). A minor modification can be witnessed in

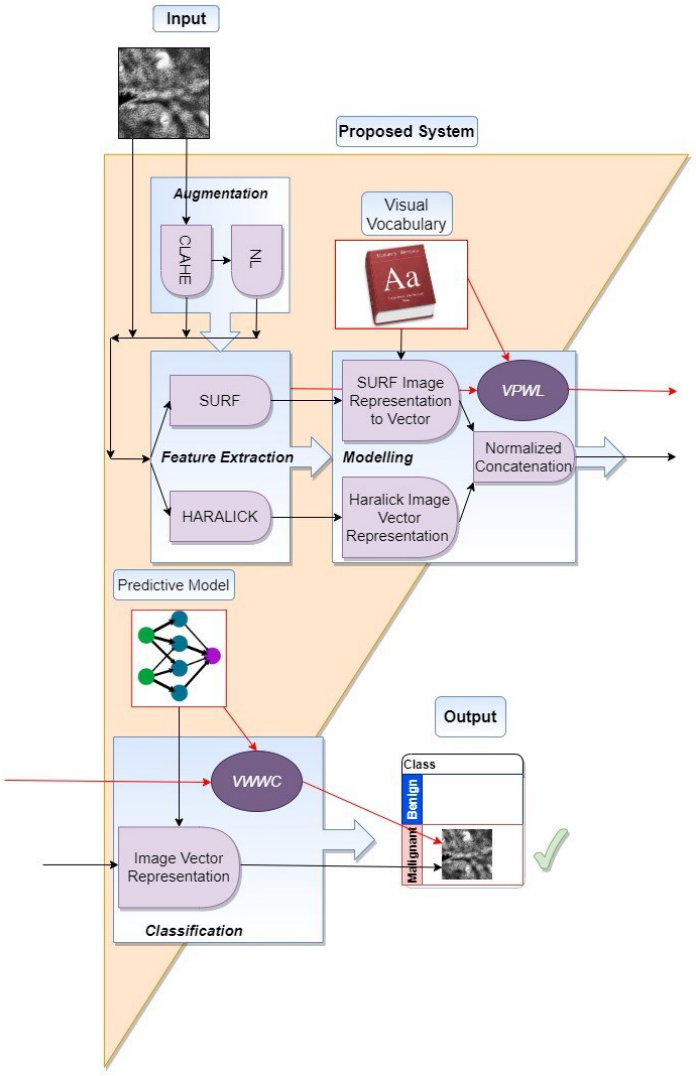


Figure 25 - Overall system architecture. Black lines and light purple shapes depict the classification task's workflow, whereas red lines and dark purple shapes depict the interpretation task workflow.

the preprocessing phase where the initial dataset consisting of 136 RCM is augmented by the utilization of two transformations to mitigate the issue of the sparsity of samples. Although it is often observed that data augmentation takes place by simple alterations of the original images (rotation, flip, etc.), the methodology follows a different path by

selecting a contrast enhancement and denoising algorithm to reach its goal. The choice is based on experiments that demonstrated the improved performance of the classification algorithm in images that were initially imposed to contrast enhancement and denoising afterward. In order to get the first set of images, Contrast Limited Adaptive Histogram Equalization (CLAHE) is performed. CLAHE[126] is an Adaptive Histogram Equalization algorithm; therefore, it generates localized image histograms corresponding to each area that displays different brightness levels from another, and through them increases the intensity value at the points where edges are located. For the generation of the second set of images, a Non-Local Means Denoising algorithm [127] is applied to the contrast enhanced image. The NL Means Denoising algorithm is utilized to reduce noise through non-local means. This algorithm works as a convolutional filter calculating the mean from the values of all the pixels in the image (instead of only the adjacent pixels) with added weight on each pixel.

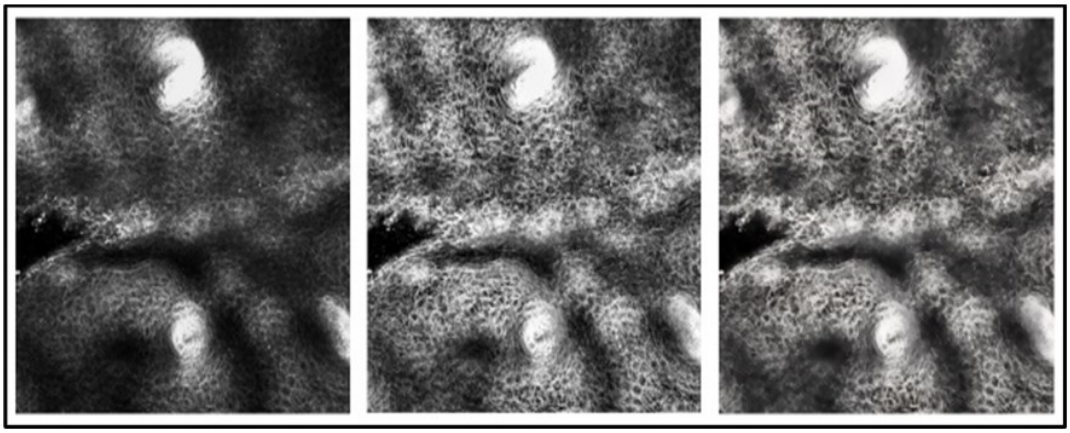


Figure 26- Samples of transformed image after augmentation. (a) is the original image, (b) the enhanced image and (c) the enhanced-denoised image.

The data augmentation procedure results in the triplication of the dataset size, which is essential for training the neural network in the predictive model. In Figure 26, the initial

RCM image shows an Acral Nevus and two synthetic copies produced by the augmentation procedure. As a result, the augmented dataset includes 408 images. Another difference is the utilized classifier. The binary classification task is performed by a simple vanilla neural network consisting of k+14 (input) and 2 (output) neurons. The basic hyperparameters of the plain 'vanilla' neural network are depicted in Table 5.

Table 5 - Basic neural network configuration

Hyperparameters	Values
Activation function	Tanh
Weight initialization	Xavier
Learning Rate	5 for Adam optimizer
Output layer activation function	Softmax
Output layer loss function	Cosine proximity
Epochs	10000

1.13.2 Deep Learning for Medical Image Classification

The manipulation of handcrafted feature extraction techniques for the quantification of visual content is challenging since it requires prior expertise and knowledge of the problem in question but offers the ability to the inner mechanisms' transparency, a vital characteristic when explainability capabilities are in discussion. On the other hand, the utilization of learned features is relieved from the expertise requirement and has demonstrated great efficiency in terms of classification performance while keeping the inner workings hidden from their makers. As already mentioned, the aim of this work is not directed to the improvement of existing classification schemes and therefore only a few descriptive lines will be spent on behalf of such details. Most of the utilized learned feature extractors are pretrained well-established DCNNs that exceed in terms of classification evaluation metrics. Transfer learning is a prosperous technique that saves engineers a lot of time and thinking, since the training of DCNNs on a general classification

problem such as the task for ImageNet[128] can be drifted to more specific classification tasks by applying fine-tuning on a subgroup of the last convolutional layers without the time and resource consuming task of training very deep architectures. Pretrained DCNNs digest a large number of medical images and extract a significant number of features as a result of a Global Max Pooling operation. These features can, in turn, be inserted in a fully connected network classifier or any other kind of ML classifier to provide classification predictions. However, rather than exploiting the discriminative power of a single DCNN in our proposed methodology we utilize ensembles of DCNNs in an attempt to reduce the bias error.

In [24] an ensemble classifier consisting of three different pretrained implementations of the EfficientNets group is employed in a parallel configuration that results in the concatenation of three different groups of feature maps. The pretrained models are trained by means of the ImageNet dataset [128]. A modification is applied after the global max pooling layer of each base feature extractor in order to provide a unified vector to the final classifier. This modification assumes the form of a concatenation layer before the classifier (Figure 27). The training set is augmented to 3 times the initial size by the utilization of three randomized operations, flip, rotation, and zoom. The final concatenated set of features is driven into a fully connected layer that acts as a classifier following typical best practices of deep CNNs. For the selection of the pretrained models, a preliminary examination of the individual performance on the two datasets led to the selection of the best-performing models in terms of accuracy. The best individually performing deep CNNs are the Inception Net, XceptionNet, and the EfficientNets group. Consequently, an ablation study is conducted between these selections in groups of three to determine the best

selection. Upon removing a CNN, the influence of this removal is measured in terms of a

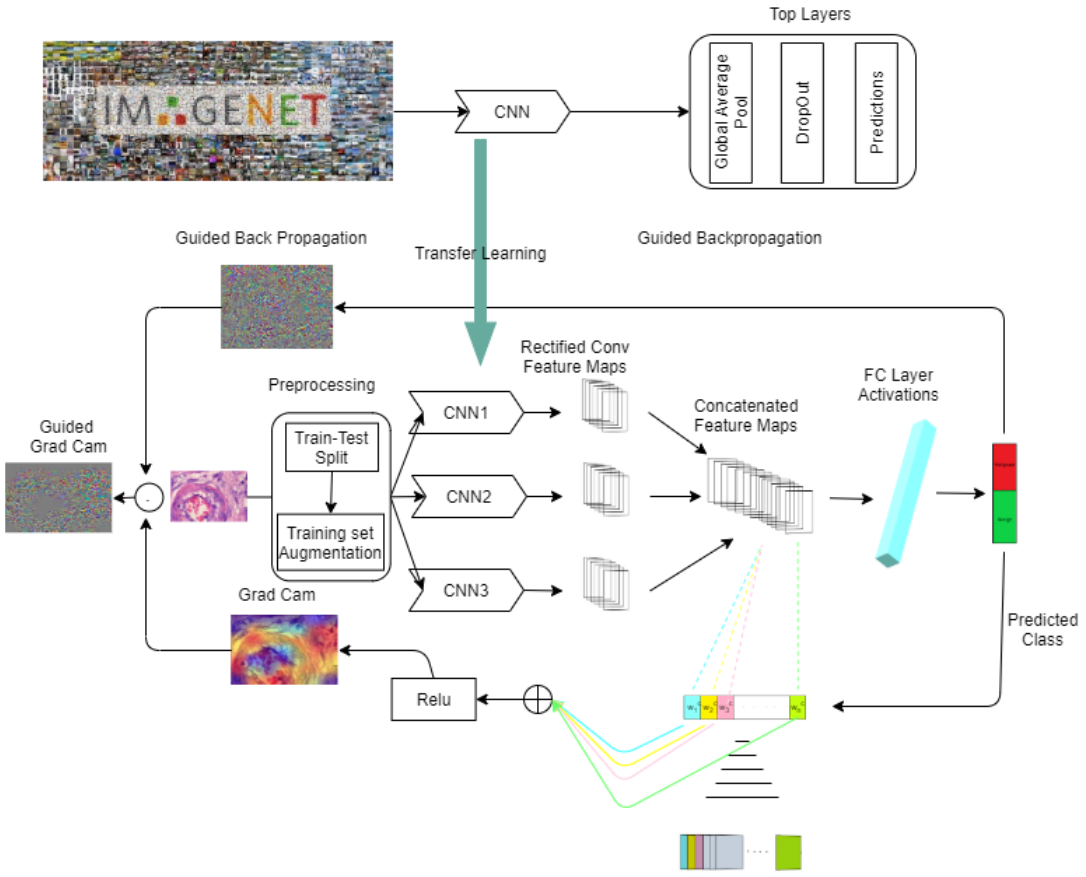


Figure 27– Interpretable ensemble network's architecture

difference in accuracy. The final selection results in the EfficientNets B2, B3, and B4. Although the basic building blocks for the three networks are the same, the required diversity in the basic classifiers of the ensemble classifier is achieved by different values provided by the compound scaling method.

The system is developed with two main purposes:

- Image classification.
- Interpretability.

Two integrated subsystems in the whole architecture interact seamlessly and are responsible for the fulfillment of each purpose.

1.14 Explainability Techniques

Most of the contribution presented in this thesis is focused on the proposal of novel techniques that can provide a straightforward link between the visual patterns in the medical image that are responsible for the classification result by utilizing traditional machine learning classification schemes. The first contribution starts with the proposal of an explainability scheme that is based on the BOVW multiple vector representation technique and expanded to the Fisher Vector technique which is a far more compact and elaborate approach for squeezing multiple vectors into one. In the following lines, the description of the proposed explainability schemes is provided, initially for the BOVW and secondly for the Fisher vector technique. Apart from the proposal of a novel approach to provide a visual explanation of the visual stimuli that were most influential by means of TML, an enhancement of the original Grad-CAM technique is proposed to refine the delineation of semantically important patterns in microscopy images in contrast to the gross grained visualization of the Grad-CAM approach. This method can be proven beneficial even if the explainability approach is performed in a pixel-wise manner since explanations that are attributed to sole pixels have no semantic value and therefore a meaningful grouping of important pixels is required.

1.14.1 Traditional Machine Learning for Medical Image Explainability

With reference to TML, the explainability approach that is based on the BOVW scheme is presented as an integrated mechanism of the classification approach presented

in [68]. As shown in Figure 25, the interpretation scheme is comprised of two stages, namely Visual Pattern Weighted Localization (VPWL), and Visual Word Weight Calculation (VWWC). The functionality of the VPWL stage takes place in the modeling stage taking advantage of the spatial information including the visual vocabulary due to the utilization of the SURF feature local extractor, while the functionality of the VWWC stage takes place in the classification stage. Although the Visual Vocabulary Classification method is criticized in the literature [129] for its inability to encode localized information, the case is contradictory for the interpretation task. The utilization of local descriptors (SURF) for the automated detection of visual patterns of interest is proven beneficial since it encapsulates the coordinates and scale of the interest point in the descriptors' vector. The coordinate is a token of the interest point's center and the scale is representative of its radius. Given this information, a well understanding of the visual stimulus in question is provided and can be utilized for visualization purposes. To get feedback concerning the influence of the interest point, the distance of each interest point from the assigned visual word (center of the cluster) is obtained when K-Means clustering is performed for the creation of the visual vocabulary. This distance (D) is treated by the proposed interpretation scheme as a measure of influence (I) of the specific interest point to the final decision. The rule is quite simple: The closest the vector of the interest point to the assigned visual word is, the bigger the influence becomes. The last 14 elements of the input vector (Haralick features) contain no localized information and describe the image globally; thus, they are left intact. Since all the necessary information is extracted from the feature extraction and visual vocabulary mechanisms the remaining information to connect the stimuli to the output concern the explainable classifier. Any kind of base classifier from a Decision Tree

or ensemble of Decision Trees such as Random Forest or XGBoost can be utilized to provide values of importance between the prediction and the feature inputs as represented by visual words. The feature importance mechanism is based on the measure upon which the classifier separates the dataset in order to improve its performance. The metric may be based on data impurity, Gini impurity is one choice or on the information gain (information entropy theory). In the case of base classifiers such as Decision Trees, the response is straightforward whereas in ensemble schemes these metrics are combined in the same fashion that ensemble classifiers reach a decision (majority voting, weighted sum, etc.). The first k inputs of the classifier are the correspondences of the interest points to each visual word detected in a query image (part of the Image Representation Vector). Each visual word shares a weighted connection with the classification output. The weight (W) of each connection adds to the influence of each visual word. Therefore, the final equation that demonstrates the influence of each interest point on the prediction outcome is provided by Equation (20), where D_{ip} is the Euclidean distance of the vector of the interest point from the corresponding centroid and W_{vw} is the weight of the connection of the visual word to the classification outcome. outcome.

$$I_{ip} = \frac{W_{vw}}{D_{ip}} \tag{20}$$

The parallel flow of the explainability task extends the idea of weighing visual words to the Fisher Vector. The formula presented in Equation (20) for the attribution of an importance value I_{ip} to each visual word is transformed into Equation (21):

$$I_{ip} = W_{GMM} \frac{W_{cl}}{D_{ip}} \quad (21)$$

The D_{ip} is the Euclidean distance of the interest point's descriptor from the nearest Gaussian Mixture Model mean value, W_{GMM} is the weight of the assigned gaussian distribution in the GMM and W_{cl} is the feature importance derived from the classification process. Local descriptors inform us about the coordinates and size of the visual stimuli in concern. The GMM paradigm, exploited for modeling the generative process of descriptors, provides the details of the distances from each gaussian distribution and the corresponding weights, while an ensemble classifier (AdaBoost, XGBoost, Random Forest) provides the importance of each feature to the final classification outcome.

Both techniques result in assigning an importance value to each interest point. These are normalized between zero and one and return a color of importance to each interest point. Therefore, the user can be informed of the visual stimuli on the image that are influenced the classification result. In order to provide a smoother visualization, the values of importance decrease as the pixels increase their distance from the center of the interest point following a Gaussian distribution.

1.14.2 Deep Learning for Medical Image Explainability

Starting with the idea of enabling explainability properties in ensemble schemes, part of the thesis' contribution refers to expanding the application of Grad-CAM to ensemble deep CNNs. In the ensemble environment, all the necessary information regarding the calculation of the Grad-CAMs exists but needs the addition of a concatenation layer so as

to bring together all extracted feature maps. This concatenation layer takes place after the last convolutional layer of each base classifier. This minor modification enables the integration of the Grad-CAM explanation module into the ensemble classifier.

Grad-CAM is an explainability scheme that is more than often utilized for the unveiling of connections between stimuli and predictions, especially in classification tasks that address the determination of the class between distinct objects in an image. However, certain categories of medical imaging such as confocal and histopathology images contain rich and dense information that differs from the cat vs dog paradigm. To further improve the performance of the Grad-CAM technique and the generated visualizations, we propose a segmentation-based explainability scheme that focuses on the common visual characteristics of each segment in an image to provide enhanced visualizations instead of highlighting rectangular regions. This proposed methodology can be applied to ensemble schemes as well. While Grad-CAM is utilized as a proof-of-concept paradigm, the technique can be applied to any explainability technique that returns importance attribution maps. The architecture of the presented methodology and workflow consists of four stages which are shown in Figure 28. The explainability pipeline consists of the neural network where the Grad-CAM technique is applied, a segmentation algorithm to further improve explainability results based on common visual properties of the generated superpixels, and a visualization module. As already mentioned, the basic advantage regarding the selected approach is that the classification scheme is left intact in comparison to other methods that apply to convolutional networks and require some modification to provide explanations. These modifications have a certain amount of influence on the performance of the

classifier, usually for the worst. Therefore, in Figure 28, any CNN or ensemble of CNNs

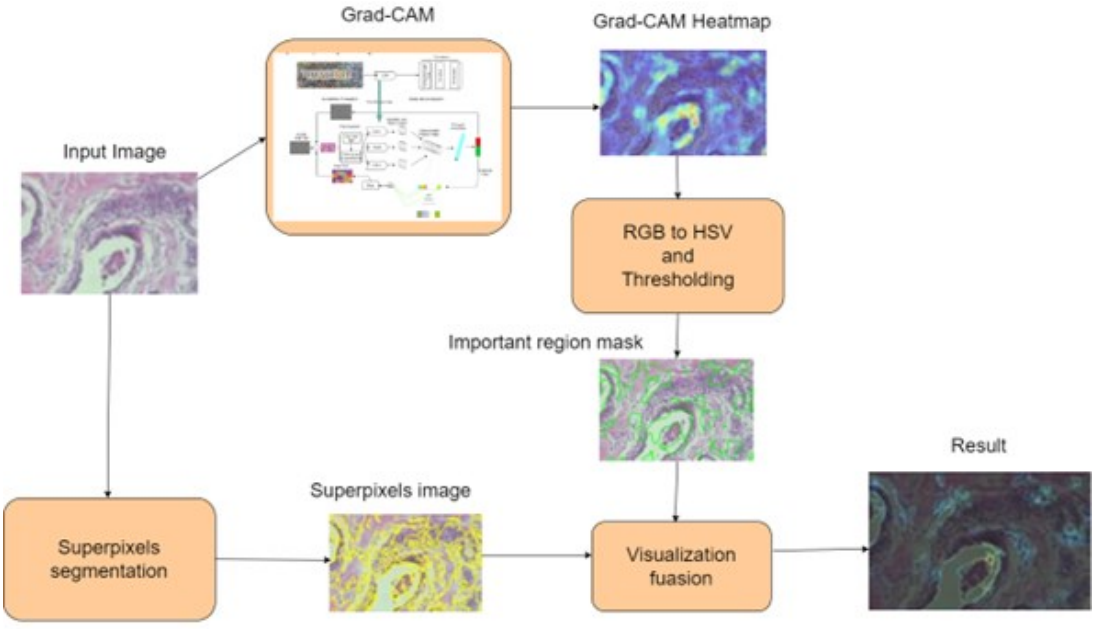


Figure 28 – Combined Grad-CAM-Superpixel system's architecture and workflow

can be employed for the classification task on the condition that the part of the architecture between the last convolutional and the softmax layer is differentiable. The main characteristics of the approach are that a) it is class-dependent, as the generated heatmaps differ from one class to another and b) belongs to the post-hoc attention approaches since the network is first trained to adjust its weights and the explanation scheme is applied afterward. Grad-CAM consists of three steps resulting in the outcome of Equation (4). The initial image enters the neural network, and a vector is generated with each value of the vector corresponding to the probability of this sample belonging to a class. Along with the prediction the respective Grad-CAM heatmap (Figure 29c) is generated and blended with the initial image.

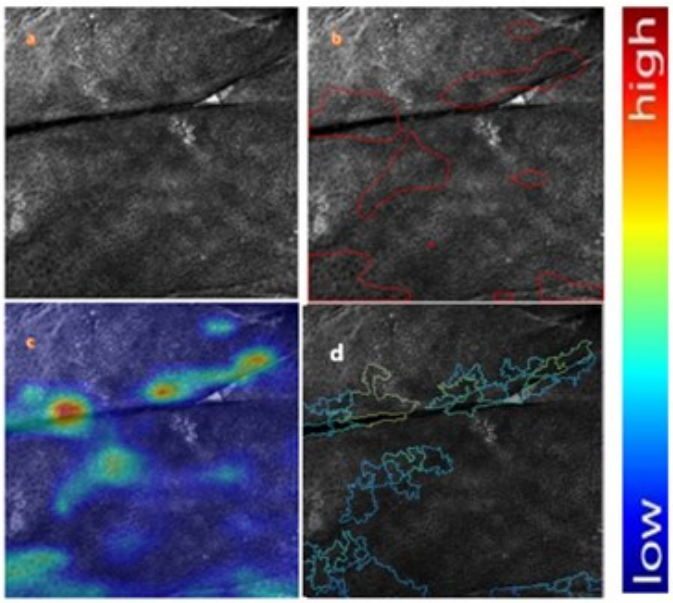


Figure 29 - Sample of the (a) initial skin confocal image depicting a nevus, and (b) the generated image by thresholding important regions/pixels as specified at; c) the Grad-CAM heatmap/initial image blend and d) the proposed heatmap/initial image blend.

The next step of the proposed methodology is the detection of important regions. This process is conducted by creating a mask on the heatmap that includes only regions/pixels of high importance (Figure 29b). The mask is generated by transforming the heatmap image to the corresponding HSV version and applying a threshold that separates the important pixels from the unimportant ones in terms of explainability. The designated regions are delineated with green color. In parallel to this process, the initial image is segmented into visually uniform regions called superpixels. The utilization of a segmentation algorithm intends to narrow down the Grad-CAM importance region to the level of cellular structures. The final step of the proposed methodology is called visualization fusion and aims to combine the generated visualization (Importance region/Grad-CAM and Segmentation) into a more compact one. During this process, the

algorithm checks the number of pixels in each superpixel that are considered important. For each one of these pixels, the algorithm assigns an importance value as dictated by the Grad-CAM algorithm. The average importance is then calculated by summing the importance values of all pixels contained in the superpixel divided by the number of pixels in the superpixel. In Fig. 27d, the outcome of the proposed technique is shown as the result of segmenting the initial image by the Felzenswalb superpixel algorithm.

Regarding the interpretability task, the concerning modules are attached to the architecture of the classification scheme while providing feedback for the localization of important visual patterns that influence the outcome of the classifier without interfering with its functionality. When utilizing the Grad-CAM technique in single classifier environments the feature maps of the last convolutional layers and the gradients for the score of each class with respect to the feature maps are necessary to produce a heatmap with the interpretability visualizations. In the ensemble environment, all the necessary information regarding the calculation of the Grad-CAMs exists but needs the addition of a concatenation layer in order to bring together all extracted feature maps. This concatenation layer takes place after the last convolutional layer of each base classifier. This minor modification enables the integration of the Grad-CAM interpretation module into the ensemble classifier.

An explainability scheme that combines the well-established properties of Grad-CAM, while enhancing them with the localization information of specific structures that derive from three different segmentation algorithms is introduced herein. The question is the following: As already known, there are several variations of CAM algorithms. The proposed technique can support the different implementations of the Grad-CAM algorithm

since they result in saliency maps. A minor difference is the fact that some of the alternatives produce pixel-wise explanations. The utilization of the technique in this scenario is prosperous as well because humans have the habit of assigning importance values to structures that stand on an area of pixels and therefore the assignment of importance to a single pixel makes no sense. Maybe you have to discuss it as future research in an effort to generalize your research. However, in this thesis, the specific Grad-CAM version of the algorithm is utilized as a pilot approach.

1.15 Unsupervised segmentation

Far from handling microscopy images, part of the contribution described in this thesis refers to the proposal of an unsupervised segmentation algorithm for blastocyst images. In the proposed methodology the manual labeling is initially bypassed by forcing a superpixel algorithm to return two superpixels, one for the foreground and one for the background. All initial images are pre-processed with contrast-limited adaptive histogram equalization (CLAHE) [130] to enhance visual patterns' contrast. Although this method provides efficient segmentation results, as experts visually inspect images, there are occasions where it fails to work efficiently (Figure 30). In order to avoid these failures, the segmentation task is conducted by utilizing only the successful segmentation results as masks to train a U-Net. This U-Net is further extended to include Conditional Random Fields (CRF) to achieve the final segmentation mask without the need for human intervention. To separate the inner cell mass (ICM) from the trophectoderm (TE) region we fit an ellipse in the inner part of the segmentation mask, with a 1/5 proportion to the mask's size (Figure 31). This ellipse provides a coarse separation between ICM and TE visual patterns. The Felzenswalb superpixel algorithm [87] with extreme values of the scale

parameter of 1500 and the minimum size parameter of 3500 is utilized to enforce the

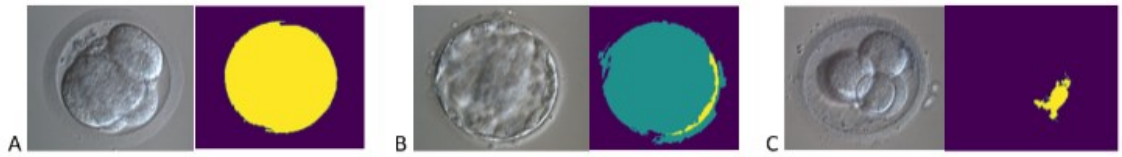


Figure 30 - Examples of applying the superpixel technique on blastocyst images. A. Successful segmentation of the blastocyst pixels from the background. B, C. Failure to segment the image in blastocyst and background pixels.

generation of two superpixels. For the U-Net, we employ a simple 4 2d-convolution encoding layers, followed by 4 2d-deconvolution decoding layers that are autocorrected by

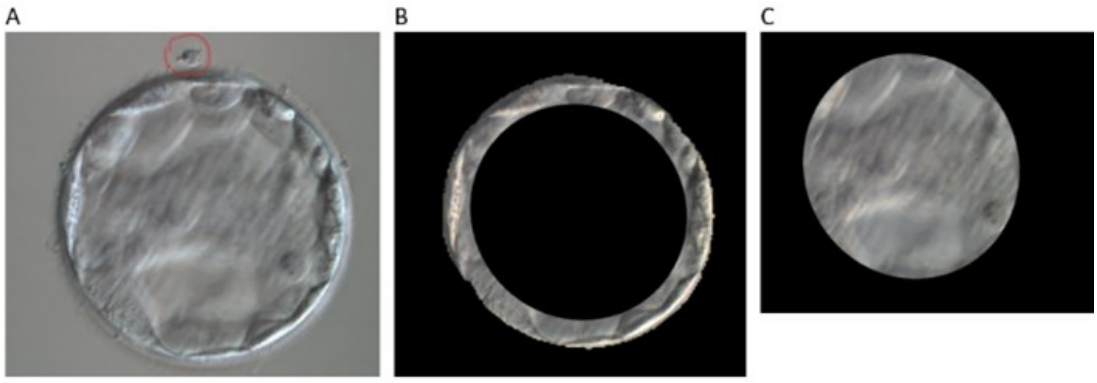


Figure 31 - A. Initial blastocyst image, with structure outside from the blastocyst annotated in red circle. B. Segmentation of the trophectoderm region by applying the proposed method. C. Inner cell mass region segmented from the rest of the image.

skip layers from the encoder. The loss function is categorical cross-entropy and the Adam optimizer with a learning rate of 0.0001 is utilized. The CRF model is proposed in [131] where the pairwise edge potentials on all pairs of pixels in the image are defined by a linear combination of Gaussian kernels.

1.15.1 Classification

The proposed methodology for the classification of blastocyst images is divided into the following five main steps and one preprocessing step: the contrast enhancement, the unsupervised segmentation, the feature extraction, the image-to-vector representation, and the classification and explainability process. A separate workflow manages the explanation task. The explanation is provided by exploiting the mechanisms of the image to vector and classification steps since both provide useful insight concerning the connection between the visual stimuli and the predictions. The basic workflow information is presented in Figure 32. The modules of the workflow that are generated during the training phase, namely the segmentation model, the classification model, and the generative model are depicted as yellow polygons, and preprocessing step is colored pink.

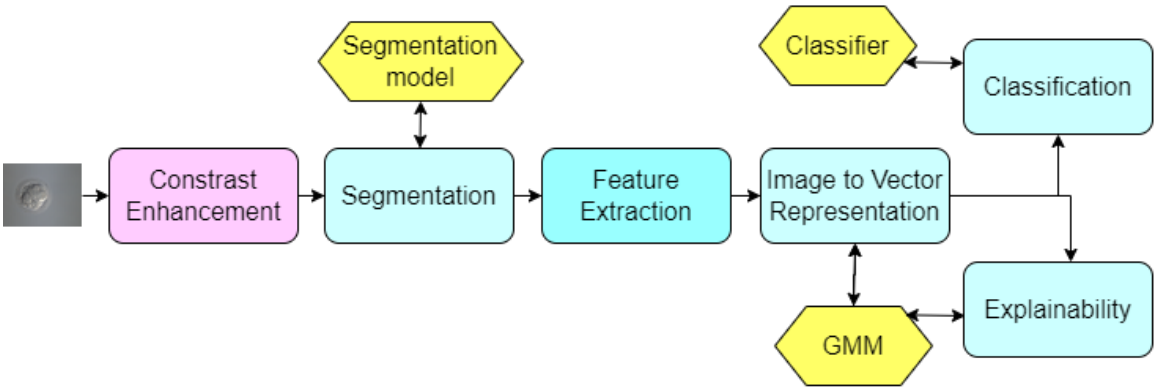


Figure 32 - Explainable model workflow for blastocyst images. The models generated in the training phase as depicted as yellow polygons. The preprocessing step is colored pink.

The classification tasks are the following: a) Degree of Expansion (DE), which refers to morphological findings in the whole blastocyst and the quality of the blastocyst's expansion, b) ICM which refers to the shape and number of cells in the ICM and c) TE, that refers to the number and shape of TE cells. The set of initial descriptors is filtered to

contain blobs that belong to the blastocyst region. Since two out of the three classification tasks refer to the specific regions (ICM or TE) of the blastocyst, we further select interest points from the corresponding regions in each task to exclude irrelevant information from the machine learning process. Descriptors are selected from the corresponding region to form the equivalent Fisher vector that will be utilized as input to the classifier of choice. Although various classifiers were benchmarked for the tasks, the best result was achieved by an XGBoost classifier. The XGBoost classifier is an efficient ensemble classification scheme that bases its success on the exploitation of the gradient boosting technique in combination with decision trees.

EXPERIMENTAL RESULTS

1.16 Medical use case scenarios for the proposed methodologies and systems

In this thesis, the proposed methodologies and systems are evaluated by the utilization of images that derive from various medical fields and modalities. Starting from the domain of Pathology, the Digital "alter-ego" was meant to solve many problems that physicians had faced in the earlier years, mainly associated with the management and preservation of tissue samples, the inability of conducting telemedical consultations, and the lack of advanced computer-based systems for diagnosis, analysis, and education. The documented ability of deep convolutional networks to identify visual patterns beyond the human perspective is gaining popularity in the field of digital pathology as well. Driven by the rise of digital scanners that produce whole slide images, the assessment of human tissue in histopathology images can be conducted by means of a virtual microscope. A whole slide image, containing on average 4 GB, can satisfy the needs of data-hungry deep convolutional networks and alleviate issues concerning the creation, handling, and preservation of glass slides. In this framework, patches, extracted from whole slide images, are inserted as inputs in deep convolution networks in a supervised or unsupervised manner, exploiting the benefits of the latest developments in the field of deep learning such as transfer learning with pretrained models and the unlabeled training via auto-encoders or Generative Adversarial Networks (GANs)[28, 29]. Apart from deep learning techniques, traditional machine learning algorithms have been utilized in the field of digital pathology for content-based image retrieval and classification of histopathology images. While first introduced for text classification, the Bag of Words (BoW) technique is utilized in[30, 31]

for the description of dense imagery content and its exploitation on designated tasks. However, whole slide imaging is introduced to the scientific community with a new breed set of challenges that need to be addressed, mainly related to the polymorphism of the data formats, the big data management, the standardization of staining, and the transparency and explainability of predictions. A vast amount of data is created every second as the digital image is produced by a glass slide along with its metadata. Analyzing large images in order to recognize patterns and similarities against images found in medical books and atlases has been proven tedious and time-consuming task for pathologists. Furthermore, the variety of whole slide scanners vendors led to the building of a new "Babel" tower, where each Digital Pathology System (DPS) speaks a different language as far as hardware, operating systems, formats of digital images, and communicating protocols are concerned[32].

A special case of histopathology images is the ones that depict glandular structures. Glands are considered of great importance due to their fundamental operations of removing, altering, or concentrating substances from the blood and then releasing them, using them, or eliminating them. The normal function of glands is, often, affected negatively by malignant tumors which arise from their epithelium, known as adenocarcinomas. The examination of their morphology by appropriate scientists through the microscope has been an everyday routine workflow to determine the existence of malignancy, their extent, and the following treatment if needed. As already mentioned, a common practice of pathologists refers to the comparison of visual patterns detected in the microscope against scientific atlases and books. The need for automated and fast literature, database, and

storage systems screening and efficient classification of the suspect structures into categories of malignancy is more evident than ever.

The basic purpose of capturing images at very high resolution from biopsies conducted on different parts of the human body is to analyze and classify the depicted visual patterns into cancer types of malignancies or benign tissues. Skin cancer is one of the deadliest forms and can be divided into two main categories: Melanomas and non-Melanomas. Melanomas, the most lethal form of skin cancer, refer to the uncontrolled multiplication of melanocytes and the creation of malignant tumors either outwards or inwards to the dermis of human skin[33]. Non-Melanomas are, in turn, divided into two categories: Basal Cell Carcinomas (BCCs) and Squamous Cell Carcinomas. BCCs are the most common form of skin cancer corresponding to 80% of Non-Melanoma Skin Cancers and they are related to the abnormal growth of basal cells at the top of the epidermis[34]. SCCs, which are the second most common form[35], refer to the formation of malignant tumors at the outer layer of the skin (squamous cells). A glance at the numbers about skin cancer is indicative of its severe impact on human health. In the United States, more than 9,500 people are diagnosed with skin cancer every day[36].

In dermatology, experts utilize digital dermoscopy to review potential malignancies on human skin. To verify their diagnosis, a sample from the skin is removed and forwarded to the histopathology labs where the nature of the tissue is determined. Histopathology remains the golden standard in experts' routines, but reflectance confocal microscopy is a non-invasive technology that can be utilized as an alternative for the diagnosis of skin cancer. Since skin cancer affects a large portion of humans on a global scale, the majority of samples are operated on and sent to the lab in vain (only a few of them are malignant).

Reflectance confocal microscopy offers the advantage of being able to review the sample at a cellular scale in vivo. However, generated images are of high density and complexity and only a few experts are specialized in evaluating the corresponding images. In the past decade (2009 – 2019), the number of new invasive melanoma cases diagnosed annually increased by 54 percent [36, 37]. Nevertheless, early detection through routine screening and treatment of skin cancer has been proven to reduce drastically mortality rates[38]. Apart from the non-invasive method of dermoscopy for screening which is most of the times accompanied by a biopsy, Reflectance Confocal Microscopy (RCM) offers an improved alternative due to its capacity to review horizontal sections of human skin at a cellular level and in vivo. The main advantage of this technique is the ability of a more accurate diagnosis without removing a sample from the patient. However, the evaluation of RCM images from experts is a tedious and time-consuming task that depends on human acuity. Training new dermatologists into experts that can interpret confocal images is a demanding task as well. The automation of this procedure via computer vision and machine learning techniques can be beneficial in terms of reproducibility and time efficiency.

Far from the field of medical images exploited for cancer diagnosis purposes, specialized cameras are utilized to capture the shape and morphology of embryo blastocysts for their transfer to the candidate mother's uterus. Infertility significantly affects the life quality of people on social and psychological levels and is estimated to expand in the coming years, contributing to the reduction of the fertility rate from 2.5 live births in 2019 to 2.2 in 2050 and 1.9 in 2100 [39]. At the same time, it deprives the fulfillment of the basic instinctive desire to have a descendant[40]. In charge of the embryos' evaluation process and based on their experience, embryologists classify images of fifth-day blastocysts, in an attempt

to select the most suitable candidate. For the evaluation of these blastocysts, most laboratories rely on the system proposed by Gardner and Schoolcraft with some modifications added later [41]. According to this system, the characteristics that determine the quality of the blastocyst are the degree of blastocyst expansion (rate of expansion), the number and shape of cells of the inner cell mass (ICM), and the number and shape of trophectoderm cells (TE).

The evaluation process is time-consuming and arduous, it requires specialized training and many years of experience, and relies strongly on the subjectivity of each evaluator in the absence of objective criteria.

1.17 Datasets

The performance of each predictive algorithm is evaluated for the performance of the classification and explainability scheme by utilizing image datasets that are representatives of the classes and problem in question. Five labeled datasets are presented in the following lines to verify the extent to which each methodology fulfills the designated purposes.

1.17.1 Reflectance confocal microscopy dataset

The dataset contains RCM images that are provided by the Andreas Syggros Hospital of Cutaneous and Venereal Diseases in Athens. The images are captured by a Mavig Vivascope 3000 that operates at 830nm, resulting in a depth of 200μm. The dataset is composed of 133 benign samples and 127 malignant ones. The benign samples are divided into the following types: Seborrheic Keratosis (SK), Solar Lentigo (SL), and Nevus

(N). The malignant samples consist of the following types of skin cancer: Spitz, Basal Cell Carcinoma (BCC), Actinic Keratosis (AK), Lentigo Maligna-Lentigo Maligna Melanoma (LM-MM), and their class distribution is described in Table 6. The fact that the malignant samples are composed of Melanoma and non-Melanoma types composes a challenging scenario for the classification task. Samples of the dataset are depicted in Figure 33.

Table 6 - Class Distribution for Reflectance Confocal Microscopy Images in benign and malignant subclasses.

Class	Subclasses	Number of Samples	Total
Benign	Benign Keratosis	61	133
	Melanocytic Nevus	68	
	Solar Lentigo	4	
Malignant	Actinic Keratosis	34	127
	Basal Cell Carcinoma	52	
	Lentigo Malignant Melanoma	41	
Total		260	

1.17.2 Breast Cancer Histopathology (BreakHis) dataset

The dataset, named Break Histological Image Classification (BreakHis), consists of 7,909 microscopic, breast tumor tissue images that are collected from 82 patients using different magnifying factors [132]. The images are:

- Divided into 2,480 benign and 5,429 malignant samples.

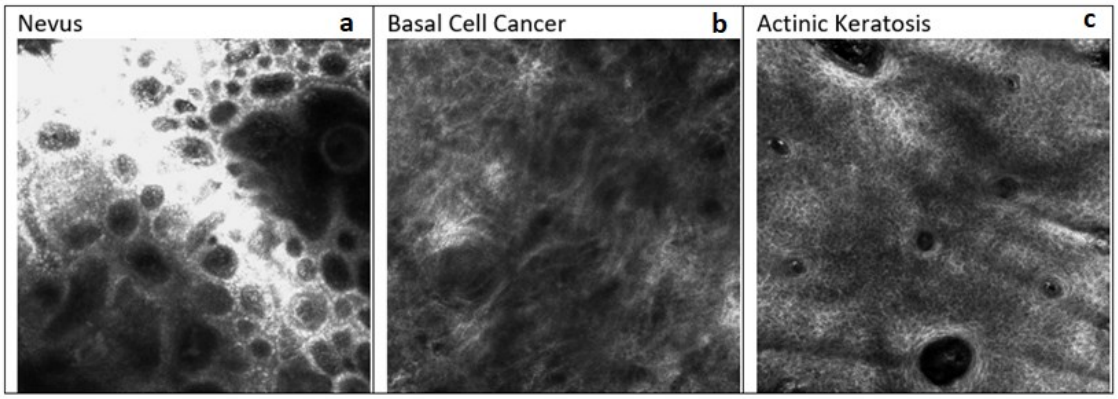


Figure 33 - This is an overview of the RCM dataset. (a) depicts a benign sample is depicted falling in the Nevus category, whereas in (b) a basal cell cancer sample is presented. (c) shows another malignant sample that belongs in the Actinic keratosis category.

- 3-channel RGB (8 bits in each channel).
- In PNG format.
- In four different magnifying factors (40x, 100x, 200x,400x).
- Contain 700x460 pixels.

Table 7 - Class distribution of the BreakHis dataset

Class	Subclasses	Magnification Factors				Total
		40x	100x	200x	400x	
Benign	Adenosis	114	113	111	106	444
	Fibroadenoma	253	260	264	237	1014
	Tubular Adenoma	109	121	108	115	453
	Phyllodes Tumor	149	150	140	130	569
Malignant	Ductal Carcinoma	864	903	896	788	3451
	Lobular Carcinoma	156	170	163	137	626
	Mucinous Carcinoma	205	222	196	169	792
	Papillary Carcinoma	145	142	135	138	560
Total		1995	2081	2013	1820	7909

Separation of benign images in the following four distinct histological types is provided in the BreakHis dataset: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA). Four malignant tumor types are provided as well: Ductal carcinoma (DC),

lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC). Samples of the BreakHis dataset are shown in Figure 34 and the class distribution of the dataset is depicted in Table 7. For the classification task, the datasets are divided into 70-30% training-test splits.

1.17.3 Breast Cancer histology (Bach) dataset

The dataset [133] consists of 500 images that are divided into 80/20 percent split for the training and test set. The images are equally divided into four classes, benign, in situ, invasive and normal. A Leica DM 2000 LED microscope and a Leica ICC50 HD camera are utilized to collect the images that correspond to patients from the Porto and Castelo Branco regions (Portugal). The annotation was performed by two medical experts. Where there was disagreement between the Normal and Benign classes, images were discarded. The remaining doubtful cases were confirmed via immunohistochemical analysis. The provided images are in RGB .tiff format and have a size of 2048 × 1536 pixels and a pixel scale of 0.42 μm × 0.42 μm. In order to ensure an unbiased evaluation process, participants were provided with a partial patient-wise distribution of the images of the training set, and the test data was collected from a completely different set of patients (Figure 35).

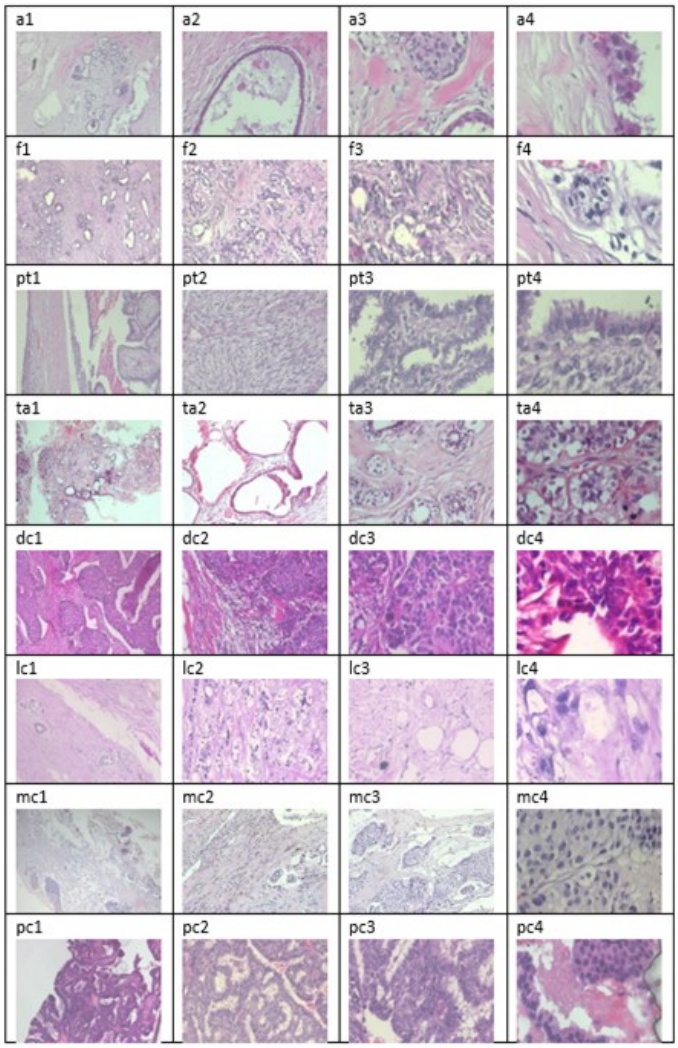


Figure 34- This is an overview of BreakHis dataset. Each row depicts a specific tissue type.: Adenosis is indicated as (a), fibroadenoma as (f), phyllodes tumor as (pt), and tubular adenoma as (ta), ductal carcinoma as (dc), lobular carcinoma as (lc), mucinous carcinoma as (mc) and papillary carcinoma as (pc). Each number

stands for a specific magnification factor: 1 for 40x, 2 for 100x, 3 for 200x and 4 for 400x (i.e., pc2 image depicts a papillary carcinoma in 100x magnification).

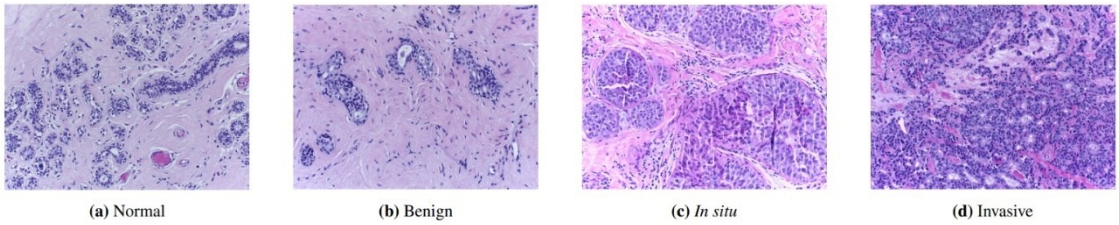


Figure 35 – Samples of the four containing classes of BACH dataset[134].

1.17.4 Colorectal cancer histopathology dataset

The dataset is a set of 100,000 non-overlapping image patches from hematoxylin & eosin (H&E) stained histological images of human colorectal cancer (CRC) and normal tissue. All images are 224x224 pixels at 0.5 microns per pixel (MPP). Tissue classes are Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM) [24]. Samples of the dataset are shown in Figure 36 and the class distribution is provided in Table 8.

Table 8 - Class distribution of the colorectal dataset

Class	Number of samples	Percentage (%)
ADI	10407	10,4
BACK	10566	10,56
DEB	11513	11,51
LYM	11556	11,56
MUC	8896	8,9
MUS	13537	13,54
STR	8763	8,76
NORM	10446	10,45
TUM	14316	14,32

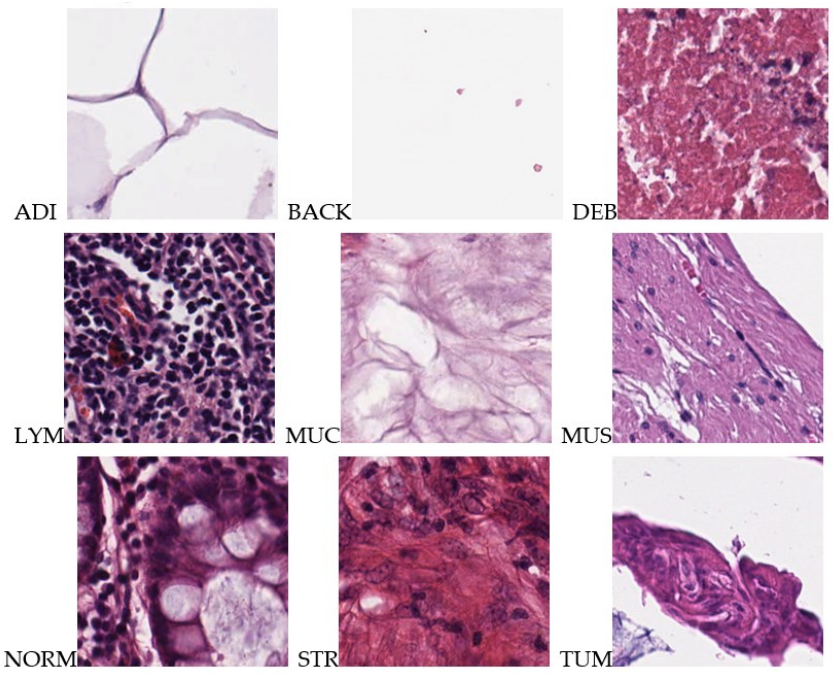


Figure 36 - This is an overview of colon cancer dataset. Each image depicts a specific tissue type.: Adipose is indicated as (ADI), background as (BACK), debris as (DEB), and lymphocytes as (LYM), mucus as (MUC), smooth muscle as (MUS), normal colon mucosa as (NORM), cancer associated stroma as (STROMA) and colorectal adenocarcinoma epithelium as (TUM).

1.17.5 Warwick-QU dataset

In order to evaluate the presented methodologies, the Warwick-QU dataset [135] is utilized. This dataset was acquired by a team of pathologists at the University Hospital Coventry and Warwickshire, UK, and comprises 165 images of .bmp format depicting types of colorectal cancer (Figure 37). The dataset derives from 16 H&E (Hematoxylin and Eosin) whole slide images that are further divided and graded as malignant or benign according to their overall glandular architecture at the time of the division by expert pathologists. Images are captured by a Zeiss MIRAX MIDI Whole Slide Scanner in resolution 20X (0.62005μm/pixel).

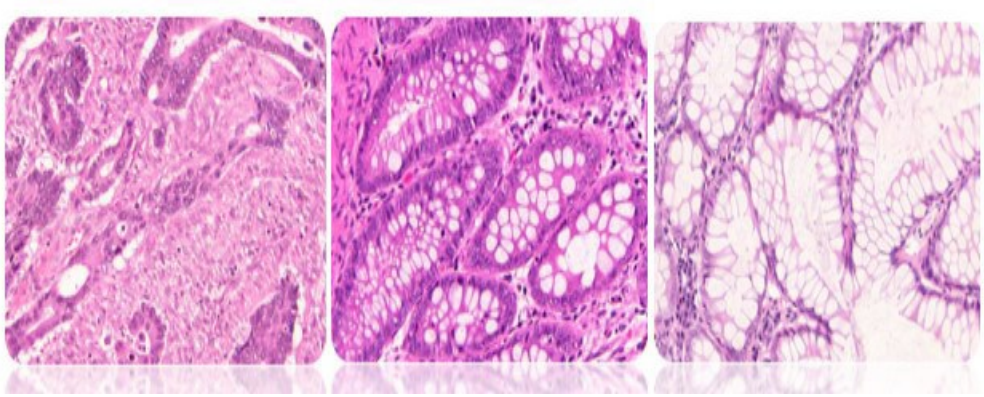


Figure 37 - Representative images of Warwick dataset. The left image depicts malignant glands, whereas the remaining images show benign glands.

1.17.6 In-vitro fertilization blastocyst image dataset

The dataset was provided by the REA Fertility and IVF Unit and consists of 1057 day 3, and 1036 day 5 blastocyst images along with their metadata including significant health-related information from both genders and demographics. All images correspond to a total of 267 patients, are in .bmp format and their dimensions are 1280x1024 at 96 dpi. The dataset is labeled by expert embryologists by means of the Gardner and Schoolcraft system. The classification tasks are the following: a) Degree of Expansion (DE), which refers to morphological findings in the whole blastocyst and the quality of the blastocyst's expansion, b) ICM which refers to the shape and number of cells in the ICM and c) TE, that refers to the number and shape of TE cells. The dataset contains 8 labels for the DE task, 6 labels for the ICM task, and 6 labels for the TE task, and class distribution is described in Table 9. The system is considered a quality standard for the evaluation of blastocysts and focuses on the morphological features and condition of the inner cell mass, the trophectoderm, and the overall blastocyst. There is an ongoing process that is expected to augment the size of the dataset in the forthcoming years, providing the ability to further

exploit deeper machine learning architectures for the extraction of useful knowledge. Samples of the IVF dataset are presented in Figure 38.

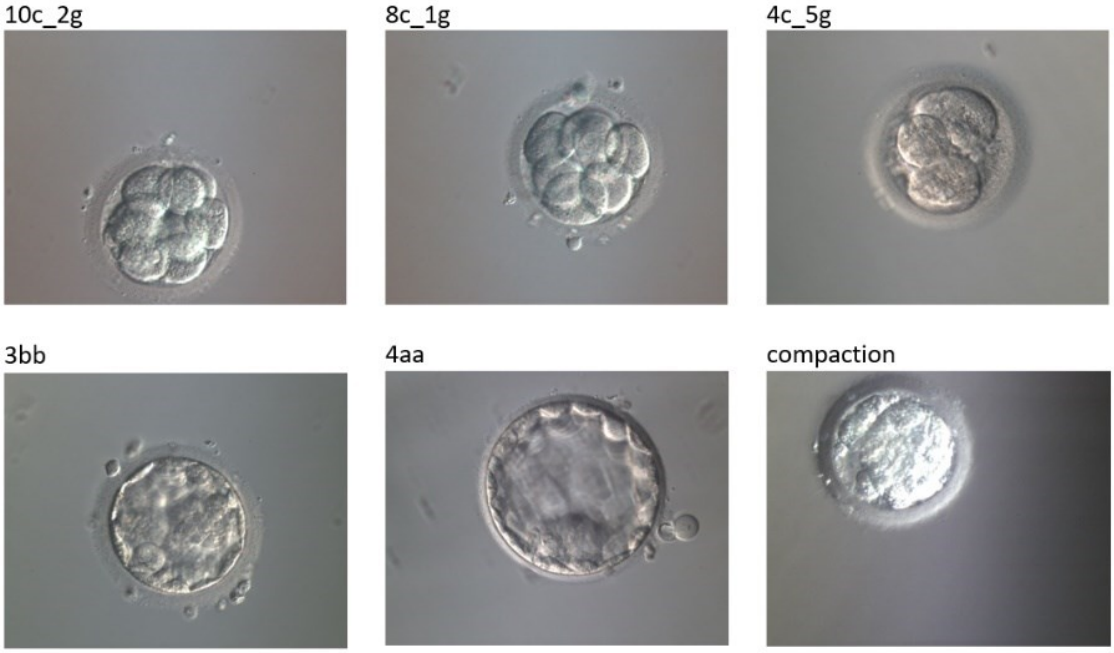


Figure 38 - The upper line of images depicts 3-day blastocyst images, whereas the lower line 5-day blastocyst images.

Table 9 - Class distribution for three classification tasks, Degree of Expansion (DE), Inner Cell Mass (ICM) and Trophectoderm (TE) on IVF blastocyst images.

Class DE DAY5	Number of samples	Percentage	Class ICM DAY5	Number of samples	Percentage	Class TE DAY5	Number of samples	Percentage
1	39	3.8%						
2	95	9.2%						
3	161	15.5%	A	354	9.7%	A	278	26.8
4	239	23.1%	B	124	28.2%	B	177	17
5	41	4.0%	C	97	6.4%	C	120	11.6
Blastocyst	68	6.6%	Blastocyst	68	34.3%	Blastocyst	68	6.6
Cells	292	28.2%	Cells	292	12.0%	Cells	292	28.2
Compaction	101	9.7%	Compaction	101	9.3%	Compaction	101	9.8

1.18 Evaluation Metrics

The utilized performance metrics for the binary classification tasks are described hereafter:

The accuracy metric is defined as the fraction of the correctly classified instances divided by the total number of instances, as shown in Equation (22).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

Correctly classified instances are analyzed in true positives (TP) and true negatives (TN), where TP are the instances predicted as positive and truly are positives (ground truth) and TN are the instances predicted as negative and truly are negative. The total number of instances consists of TP, TN, false positive (FP), and false negative (FN) instances. FP are the instances that are predicted as positive by the classifier but are negative in reality, whereas FN are the instances predicted as negative but are positive. The precision metric is defined as the fraction of the true positives divided by the true positives and false positives as shown in Equation (23):

$$Precision = \frac{TP}{TP + FP} \tag{23}$$

The recall (Sensitivity) metric is defined as the fraction of the true positives divided by the true positives and false negatives as shown in Equation (24):

$$Recall = \frac{TP}{TP + FN} \tag{24}$$

Specificity is another popular metric for the evaluation of classification tasks, defined as the fraction of TN divided by the sum of TN and FP (Equation 25), which in this work is utilized for the measurement of balanced accuracy.

$$Specificity = \frac{TN}{TN + FP} \tag{25}$$

In cases where the dataset shows imbalanced properties, the metric of balanced accuracy is utilized to provide a better insight into the model's performance. Balanced accuracy is defined as the sum of sensitivity and specificity divided by 2 as shown in Equation (26).

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \tag{26}$$

The area under Curve (AUC) metric is defined as the area under the receiver operating curve. The receiver operating curve is drawn by plotting the true positive rate (TPR) versus the false positive rate (FPR) at different classification thresholds. TPR is another word for recall whereas FPR is the fraction of the false positives divided by the true negatives and false positives as shown in Equation (27):

$$FPR = \frac{FP}{TN + FP} \tag{27}$$

For the multiclass classification tasks, some of the above metrics require minor changes to support their role. The recall will be changed into micro averaging leading to a new fraction (Equation 28) which is defined as the sum of all true positives through all classes divided by the sum of all the TP and FN through all classes. Such a division assigns equal weight to each sample. In this fashion the metrics for Precision and Specificity are computed.

$$Micro\ Recall = \frac{Sum\ (TP)}{Sum(TP) + Sum(FN)} \quad (28)$$

However, the macro averaging metrics are also useful when imbalances in classes are evident. The metric is calculated for each class separately and then the arithmetic mean returns the macro-averaging value. In other cases, the strict criterion of predicting exactly the right class is loosened to provide feedback for the validity of the ground truth by means of the top-k accuracy.

Concerning the evaluation of the explainability technique, the Most Relevant First (MoRF) curve is exploited in combination with the Area Over the MoRF Perturbation (AOPC) Curve. The first metric is computed by arranging tiles of the input images in a sequence starting with the most important tile. The importance of each tile is measured by the generated heatmap of the explainability technique that assigns an importance value to each tile. Once the sequence is generated, we perform random noise perturbations on the image tiles starting from the most important tile and witness the decrease in the predicted probability of the dominating class. The intuition is that the more relevant the patch, the more it will affect the classification output, so we expect a steeper decrease in the initial stages of the perturbation process and a lower slope of the curve from there onward. On the other hand, the Area Over the MoRF Curve should increase as we modify larger areas of importance of the initial image. The AOPC curve is defined in Equation (29). For the multiclass classification tasks, some of the above metrics require minor changes to support their role. The recall will be changed into micro averaging leading to a new fraction (Equation 28) which is defined as the sum of all true positives through all classes divided

by the sum of all the TP and FN through all classes. Such a division assigns equal weight to each sample. In this fashion, the metrics for Precision and Specificity are computed.

$$AOPC = \frac{1}{L+1}\langle \sum_{k=0}^{L} f\left(x_{MoRF}^{(0)}\right) - f\left(x_{MoRF}^{(k)}\right) \rangle_{p(x)} \quad (29)$$

In Equation (29), the L parameter is the number of tiles, <.>p(x) is the average over all images in the dataset, and f(x) is the function that measures the importance of each tile in the classification result.

1.19 Classification Results

Although the main contribution of this thesis lies in the proposal of explainability schemes on top of existing classification techniques, the presentation of the classification results herein is indicative of the need to accomplish high-performance numbers in terms of both explainability and classification. It is demonstrated that the proposed schemes not only achieve high accuracy metrics but can also return useful visual explanations for the generated predictions that in some cases are reported to exceed state-of-the-art results.

1.19.1 Applying visual vocabulary schemes on colorectal cancer histopathology images

The BOVW measures only the number of correspondences of each interest point to each visual word. This is referred to as Term Frequency-TF in the relevant literature. Therefore, it treats each visual word equally and does not consider the significance of each visual word individually in the categorization of each image. In order to measure the specific significance, Inverse Document Frequency (IDF) is calculated, which leads to the enhancement of the provided information (Equation 18). N is the number of images in the training set and N_{con} is the number of images in the training containing a specific visual

word. The TF x IDF product (Equation 19) is the equivalent weight (W_{vw}) attached to each visual word (VW).

$$IDF_{vw} = log\left(\frac{N}{N_{con}}\right) \tag{27}$$

$$W_{vw} = TF \times IDF \tag{28}$$

In the proposed system, Vector Locally Aggregated Descriptors (VLADs) are also utilized. Contrary to BOVW, where the final image vector is created by the correspondences of the interest points of the image to the visual words, in VLAD, the image vector is created by the sum of differences between the interest point descriptor and the visual word.

BOVW implementation with the utilization of SURF features leads to results of 91% accuracy. Nevertheless, the final vector can be improved by adding information such as the texture Haralick features extracted using GLCM matrices. In this way, it is possible to exploit the information concerning the texture of the represented structures, which takes the form of a vector containing 14 statistical characteristics (Angular Second Moment, Contrast, Correlation, Sum of Squares, Inverse Difference Moment, Sum Average, Sum Variance, Sum Entropy, Difference Variance, Difference Entropy, Information Measures of Correlation, Maximal Correlation Coefficient). Since both algorithms, SURF and Haralick are applied to grayscale images, the color information is not exploited. In order to further enrich the generated vector, Color Moments are extracted from each image. In order to calculate Color Moments four low-order statistical measures (Mean, Standard Deviation, Skewness, and Kurtosis) are extracted globally from each image. The vector representing the image can be further improved by adding weights to each visual word in

accordance with the term frequency-inverse document frequency definition explained earlier.

By enhancing the BOVW vector (created from multiple SURF descriptors) with Haralick features and Color Moments, we achieved very high accuracy in classifying the Warwick Dataset into two classes (benign and malignant), comparable to the state of the art in [136], which is based on a deep convolutional network. After calculating the average silhouette index of consecutive clustering approaches (KMeans, KMeans++, canopy, and farther first) and the number of clusters (0-350), the number of 15 visual words and KMeans were chosen for the production of visual vocabulary in addition to a Random Forest classifier for classification purposes. This configuration yields the best results from a time-efficiency perspective. In order to further investigate and improve the results, additional implementations are evaluated. The first one incorporates the VLAD technique and creates a vector of 960 values from each image, the second one imposes weights on each visual word according to the TDIDF methodology and the third extracts local Haralick + Color Moments features in the area of interest detected by SURF (in this case a 10x10 pixel area is chosen) and integrates them in a VLAD vector. It is deducted through the evaluation process that the best performance refers to the BOVW Global Haralick-Color Moments Implementation, followed by the BOVW TDIDF variation of the proposed system.

Table 10 - (A)ccuracy, (S)ensitivity, (P)recision, (SP)ecificity of classification results of Warwick dataset in two classes.

	Classifiers		
Implementation scheme	MLP/ Logistic (for VLAD)	Naïve Bayes	Random Forest

VLAD	A:0,97 S:0,97 P:0,97 Sp:0,96	A:0,95 S:0,91 P:1 Sp:1	A:0,95 S:0,91 P:1 Sp:1
VLAD + Local Haralick + Color Moments	A:0,97 S:0,97 P:0,97 Sp:0,96	A:0,95 S:0,91 P:1 Sp:1	A:0,95 S:0,91 P:1 Sp:1
BOVW TDIDF	A:0,95 S:0,97 P:1 Sp:01	A:0,75 S:0,82 P:0,92 Sp:0,77	A:0,98 S:0,97 P:0,1 Sp:01
BOVW	A:0,9 S:0,81 P:1 Sp:1	A:0,9 S:0,94 P:0,96 Sp:0,96	A;0,91 S:0,94 P:0,91 Sp:0,88
BOVW + Global Haralick	A:0,88 S:0,87 P:1 Sp:1	A;0,95 S:0,91 P:1 Sp:1	A:0,96 S:0,97 P:0,91 Sp:0,96
BOVW + Global Haralick + Color Moments	A:0,8 S:0,88 P:0,83 Sp:0,77	A;0,95 S:0,91 P:1 Sp:1	A:1 S:1 P:1 Sp:1

However, a thorough examination of the metrics shows a good performance of VLAD implementation for all utilized classifiers in comparison to the BOVW + Global Haralick -Color Moment's implementation where the best results reflect only the deployment of Random Forest. The results are presented in Table 10. Additional experimentation resulted in adding up all 165 images of the dataset in order to perform 10-fold cross-validation. This experiment demonstrated 96.2% accuracy for the implementation of the VLAD vector with the utilization of a Logistic classifier.

1.19.2 Applying the BOVW scheme on RCM images

Experimentation with the BOVW scheme in practice is conducted with a dataset of RCM images as well, provided by the Andreas Syggros Hospital of Cutaneous and Venereal Diseases in Athens. The fact that the malignant samples are composed of Melanoma and non-Melanoma types composes a challenging scenario for the classification

task. The initial dataset is augmented to triple its size creating a new set that includes 408 RCM images. For the classification task, 10-fold cross-validation is applied to avoid overfitting the predictive model. Experiments are conducted with two variations of K-Means clustering (K-Means Plus-Plus [137] and Canopy [138]), with the exclusion of Haralick features and with a different number of K clusters in order to reach the highest accuracy level. Silhouette index [122] measures the clustering performance. Concerning the classifier, which is a neural network, the epochs for training have been determined by the early stopping technique, while tests with different optimizers, activation, and loss functions resulted in the configurations depicted in Table 5. The absence of many layers in the ANN permits the utilization of a Desktop computer for our experiments with an Intel

Table 11 - Classification results of BOVW technique on RCM images

Clustering Method	Extracted features	Loss Function	Test set accuracy
K-Means	SURF	Negative Log Likelihood	89.23
		Cosine Proximity	89.96
	Haralick	Negative Log Likelihood	69.94
		Cosine Proximity	69.45
	SURF + Haralick	Negative Log Likelihood	89.23
		Cosine Proximity	89.47
K-Means ++	SURF	Negative Log Likelihood	88.49
		Cosine Proximity	88.98
	Haralick	Negative Log Likelihood	69.94
		Cosine Proximity	69.45
	SURF + Haralick	Negative Log Likelihood	89.23
		Cosine Proximity	89.22
K-Means Canopy	SURF	Negative Log Likelihood	89.95
		Cosine Proximity	91.16
	Haralick	Negative Log Likelihood	69.94
		Cosine Proximity	69.45
	SURF + Haralick	Negative Log Likelihood	90.2
		Cosine Proximity	91.17

Core i5 CPU processor at 1.6HZ and 8GB RAM. The results of accuracy for the test set of the described experiment are presented in Table 11. Classification results demonstrate that adding Haralick global texture information to the final vector slightly improves the

performance of the classifier. The highest accuracy is achieved when utilizing K-means canopy clustering and the cosine proximity loss function. To our knowledge, the classification results are the highest reported in the literature for a multiclass classification task concerning RCM images.

1.19.3 Applying ensemble schemes of DCNNs on breast and colorectal histopathology images

As we switch to the deep learning domain, a comparison between base and ensemble DCNN architectures is provided for the evaluation of the improvement the latter introduces. We choose from the pool of the TensorFlow 2.3 API (https://www.tensorflow.org/) the following well-established architectures:

- EfficientNets B0-B7
- InceptionNet V3
- ExceptionNet
- VGG19
- ResNet152V2
- Inception-ResNetV2.

The hyperparameters for the deep convolutional architectures were set after experimentation to the values shown in Table 12.

Table 12 - Hyperparameters settings for the utilized deep CNN architectures.

Hyperparameters	Values
Epochs	10
Optimizer	Adam
Learning Rate	Custom

Regularizer	L2
Batch size	8

To determine which pretrained deep convolutional neural networks are better performing in the specific datasets, a preliminary experiment is conducted for each individual classifier. To further improve the performance of each classification scheme, experiments are conducted with different custom learning rate schedulers that result in the learning rate scheduler which is expressed by Equation (29):

$$Lr(epochs) = Lr_{start} + (Lr_{max} - Lr_{start}) \times k \times epoch \quad (29)$$

where Lr defines a function that depends on epochs, Lrmax is set to 0,00005, and Lrstart to 0,0001. The difference in accuracy increases by 1.6% in the case of EfficientNet B0 when utilizing the above learning rate scheduler in contrast to using a plain Adam optimizer and k a hyperparameter that is computed by heuristic methods. In Table 13 the corresponding results for the binary (benign vs malignant) breast cancer and the multiclass colon cancer classification task (adipose vs background vs debris vs lymphocytes vs mucus vs smooth muscle vs normal colon mucosa vs cancer-associated stroma vs colorectal adenocarcinoma epithelium) are depicted. By forming different groups of three baseline classifiers and removing each turn one, two ensemble architectures were formed. Each architecture contains the baseline implementation that had the greater impact on performance metrics when removed. The two qualified architectures are the EfficientNet group consisting of B0, B1, and B2 and the group consisting of B1, B2, and B3. In order to evaluate the effect of utilizing ensemble architectures against the baselines, Table 14

demonstrates the performance metrics for each configuration. The performance of the baseline architecture leaves small space for improvement even when the dataset is split in a 60-40% ratio. Even so, the Efficient B0-B2 ensemble method is on par with the colon cancer dataset.

Table 13 - Performance metrics for the breast and colon cancer dataset for baseline architectures.

	Breast Cancer		Colon Cancer	
Architecture	**Accuracy**	**AUC**	**Accuracy**	**AUC**
EfficientNetB0	0.9766	0.9945	0.9946	0.9993
EfficientNetB1	0.9798	0.9964	0.9898	0.9984
EfficientNet B2	0.9817	0.9982	0.9920	0.9988
EfficientNet B3	0.9855	0.9988	0.9897	0.9984
EfficientNet B4	0.9858	0.9980	0.9910	0.9982
EfficientNet B5	0.9804	0.9975	0.9924	0.9982
EfficientNet B6	0.9728	0.9953	0.9894	0.9986
ExceptionNet	0.9785	0.9942	0.9909	0.9985
InceptionNetV3	0.8868	0.9430	0.9844	0.9981
VGG16	0.9320	0.9769	0.9795	0.9969
ResNet152V2	0.8720	0.9431	0.9564	0.9913

Table 14 - Performance metrics for the breast and colon cancer dataset for ensemble architectures

	Breast Cancer		Colon Cancer	
Architecture	**Accuracy**	**AUC**	**Accuracy**	**AUC**
EfficientNetB0-2	0.9925	0.9985	0.9946	0.9991
EfficientNetB1-3	0.9855	0.9984	0.9856	0.9989

Even when splitting the dataset in 60-40% the ensemble architecture managed a minor improvement in some cases. Nevertheless, in the worst-case scenario, the proposed ensemble architectures are on par with the baseline implementations. However, the task of classification is made more difficult by splitting the dataset 40-60%(training-test) and 30-70% and bootstrapping the splits 10 times to enhance randomness. In Table 15, the results from these two extreme splits are demonstrated. The difference in performance metrics is greater as the problem of classification becomes more difficult.

Table 15 - Performance metrics for the breast and colon cancer dataset for ensemble and plain architectures for 40-60% and 30-70% splits.

		Breast Cancer		Colon Cancer	
Split	**Architecture**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
40-60%	EfficientNetB0	0.9789	0.9974	0.9645	0.9874
	EfficientNetB1	0.9778	0.9974	0.9688	0.9899
	EfficientNetB2	0.9824	0.9986	0.9764	0.9906
	EfficientNetB0-2	0.9835	0.9989	0.9822	0.9934
30-70%	EfficientNetB0	0.9712	0.9962	0.9618	0.9822
	EfficientNetB1	0.9737	0.9972	0.9666	0.9831
	EfficientNetB2	0.9751	0.9968	0.9703	0.9852
	EfficientNetB0-2	0.9785	0.9979	0.9782	0.9925

1.19.4 Applying TML and DL classification techniques on blastocyst images

Both TML and DL classification schemes are utilized to evaluate the proposed methodology and verify the performance improvement in terms of accuracy with the segmented images. An image dataset from the REA Fertility and IVF Unit is utilized. The classification tasks are the following: a) Degree of Expansion (DE), which refers to morphological findings in the whole blastocyst and the quality of the blastocyst's expansion, b) ICM which refers to the shape and number of cells in the ICM and c) TE, that refers to the number and shape of TE cells. For each classification task, the dataset is labeled individually by experienced embryologists and contains 8 labels for the DE task, 6 labels for the ICM task, and 6 labels for the TE task. It is split into a training and test set with a proportion of 66.6 - 33.3% and a three-fold cross-validation scheme is applied. The evaluation scenarios include the three classification tasks in the following four cases: a) without applying the segmentation b) with the proposed segmentation methodology, c) with a pretrained ResNet-101 (RN-101) [67], d) and a pretrained EfficientNet-B1 (EN-B1) [139] on the ImageNet. Before passing the images through the networks, they are resized according to each model's default input size and normalized with the mean and standard

deviation of the ImageNet dataset. During the training of 20 epochs in total, a cosine annealing schedule with three warm-up epochs is utilized. The utilized metrics for the classification performance include accuracy, balanced accuracy, and top-2 accuracy. Results for all the above-mentioned scenarios are depicted in Table 16. The review of the presented results demonstrates that the proposed methodology in some cases presents comparable results with the state-of-the-art deep learning configurations. The best results are presented in bold. It is worth noticing that top-2 accuracy shows a dramatic increase in all cases.

Table 16 - Classification results. Accuracy is labelled as ac, Balanced Accuracy as Bacc and top-2 accuracy as top-2.

	DE			ICM			TE		
Method									
	Acc	Bacc	Top-2Acc	Acc	Bacc	Top-2 Acc	Acc	Bacc	Top-2Acc
Unsegmented ML	0.55	0.32	0.75	0.60	0.38	0.80	0.54	0.39	0.78
Proposed	0.57	0.36	0.77	0.62	0.40	0.82	**0.62**	0.46	0.82
Rn-101	0.52	0.51	0.76	0.52	0.42	0.79	0.55	0.46	0.79
EN-B1	**0.63**	**0.52**	**0.87**	**0.65**	**0.59**	**0.87**	0.61	**0.57**	**0.87**

1.20 Explainability Results

1.20.1 Qualitative evaluation on proposed BOVW explainability method

Efficient classification results for the related techniques are a strong prerequisite for the integration of explainability schemes upon them. Our main interest lies in the explanation of the results and their relationships with the visual patterns that led to the classification outcome. To this end, in this section, a detailed presentation of the qualitative explainability feedback is provided for BOVW explainability methods. Since the explainability scheme for the BOVW technique is applied to RCM images, samples of the

proposed method on confocal images are illustrated in Figure 39. The influence [Iip-calculated as shown in Equation (20)] of each interest point is represented by a different color starting from the lowest values that appear with dark green to the highest which is shown with light red. Influence values are normalized between 0 and 1. The size scale of the interest point is indicated by the size of the circle. In Figure 39a the original RCM image depicts an actinic keratosis skin cancer (malignant). 2297 interest points are detected by the SURF detector from which the 100 painted red in Figure 39b has the greatest impact on the classifier. In Figures 39b, and c, 54 interest points are depicted in total with a diameter between 100-180 pixels and greater than 180 pixels, respectively. Interest points at these scales have little contribution to the result. In Figure 39d the interest point with a weak impact (enclosed in a red circle) on the classifier's result is represented. This interest point is assigned to the visual word No.165. Visual word No.165 contributes very little (weight of connection to a node that represents malignancy class: 0.0008, the distance between the vector representing the interest point and the visual word is measured at 2.08) to the decision of the classifier. The visualization results are validated qualitatively by our collaborating doctors. In Figure 39a, the explainability module has depicted 100 key points with the highest grade. These key points correspond, mostly, to irregular keratinocytes. The irregularity is found in the size of cells and in the thickness of honeycomb patterns found in the image, which, when they appear massively in an image, are indicative of normal skin tissue. By skimming through the visualization results of the explainability module, it was made clear that the 'skin fold' visual pattern depicted in Figure 40 can be rather misleading for the algorithm. In Figure 40 a typical honeycombed pattern is shown (indicative of benign tissue), which should not be mistaken for malignancy. However, the

algorithm confuses the fold for a different pattern, thus, providing an erroneous prediction. This is an obvious advantage of utilizing integrated explainability modules in classification algorithms, due to the fact that insight is provided for the error detection and redesign of the classification system.

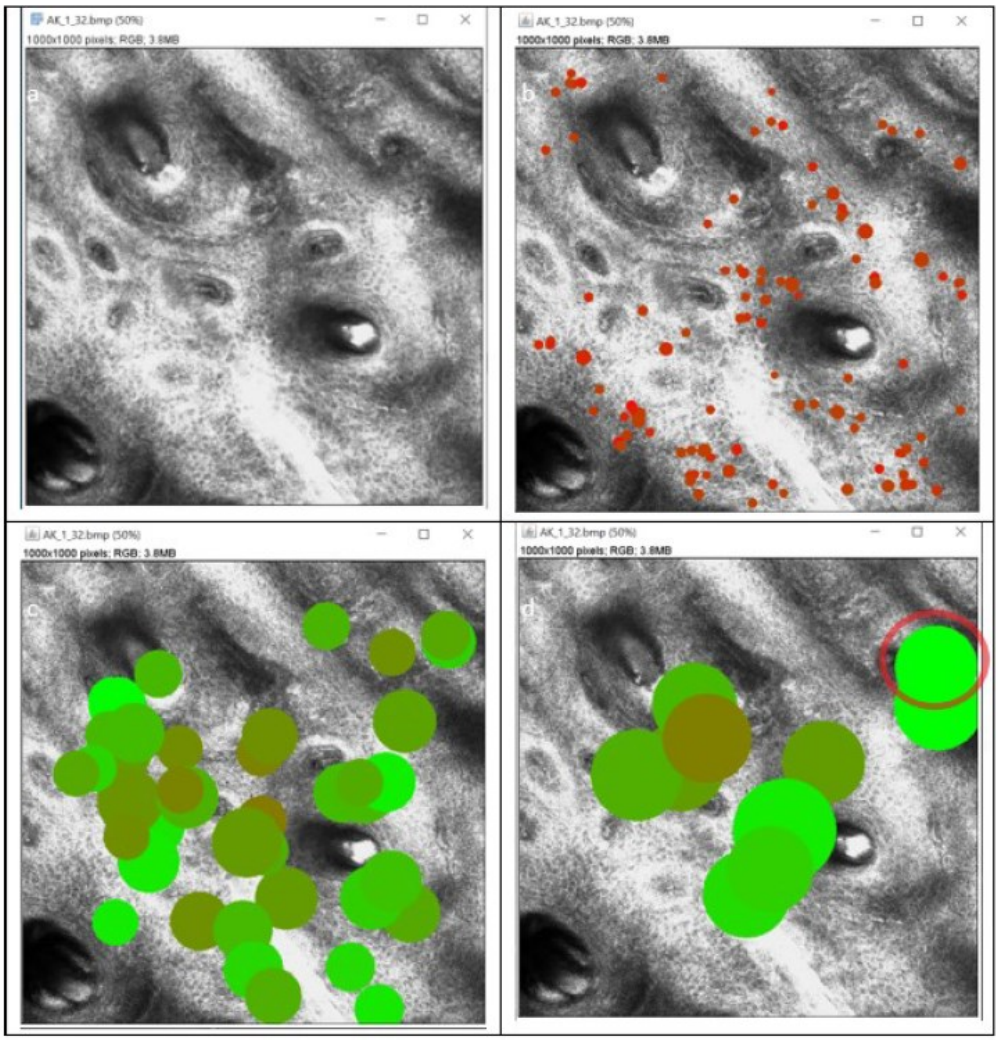


Figure 39 - a. Original Actinic Keratosis Confocal Image, b. Top 100 interest points with influence value equal to 1. c. 44 interest points with scale corresponding to diameter between 100-180 pixels, d. 10 interest points with scale corresponding to diameter greater than 180 pixels(a) is the original image, (b) the enhanced image and (c) the enhanced-denoised image.

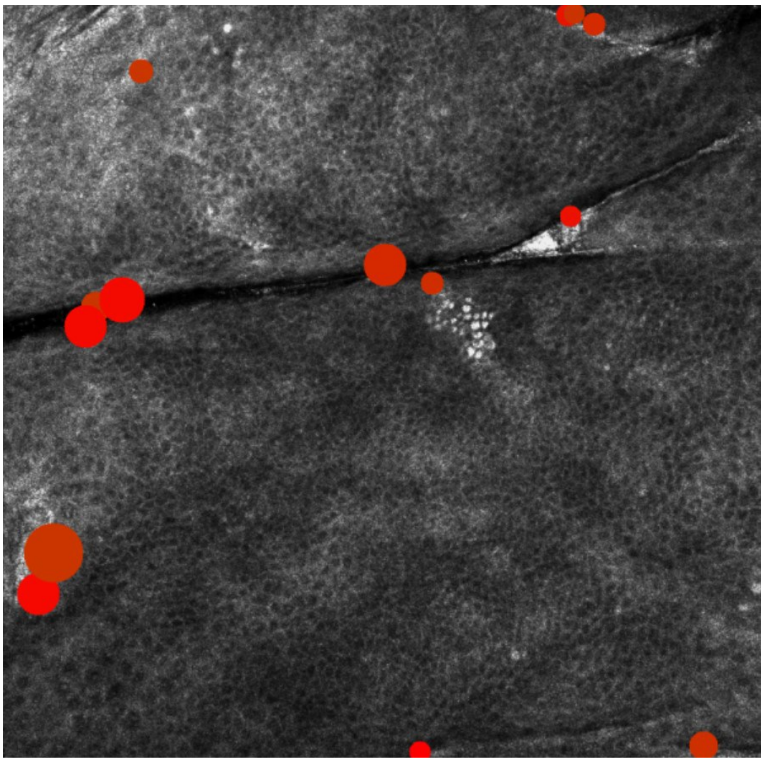


Figure 40 - Erroneous classification of benign representation (NEVUS) due to skin fold.

1.20.2 Qualitative evaluation on proposed Fisher Vector explainability method

The BOVW technique is a well-established method to reduce the dimensionality of multiple vectors into a single representation, but the approach is rather simplistic in the way that this reduction preserves useful information. The Fisher Vector technique is far more elaborate since it employs higher-order statistics for a soft assignment procedure that maintains a significant amount of useful knowledge with reference to the generative process of the descriptors. Therefore, it is deemed beneficial to extend the proposed explainability scheme to the Fisher Vector paradigm by applying minor modifications. To evaluate the results of the proposed methodology in terms of explainability, a qualitative analysis of the generated visual explanations is provided. In Figure 41, the upper line depicts the most influential patches of the image for different classification tasks in the

case of the deep learning scenario with the Grad-CAM technique. After a thorough examination of most

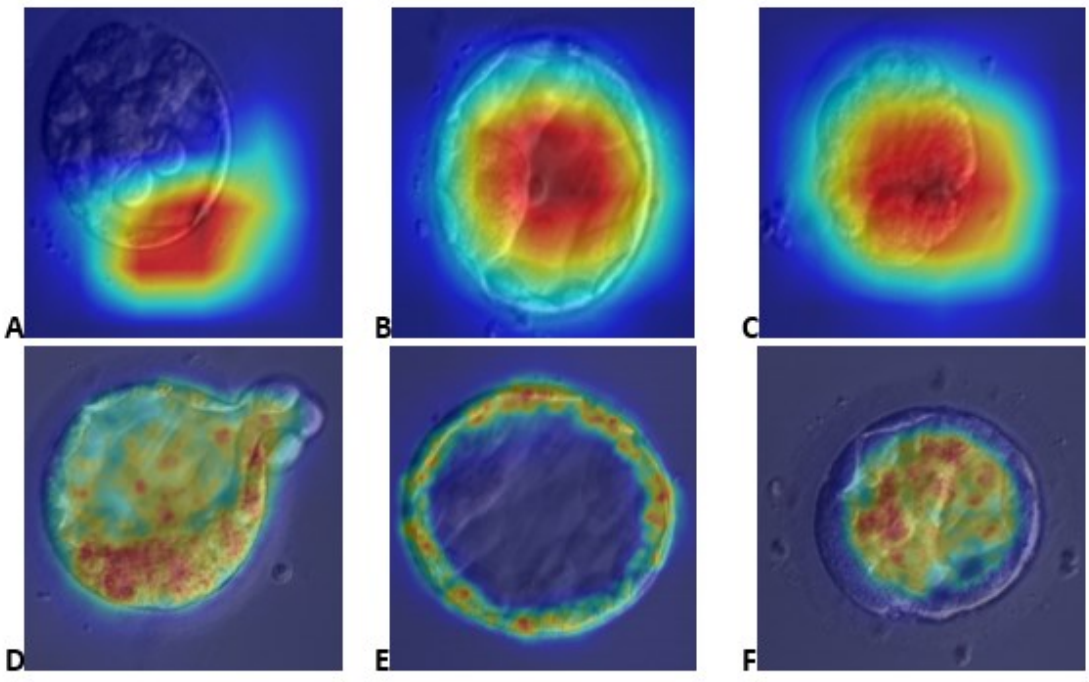


Figure 41 - A. Visual explanations of blastocyst images for the EfficientNet B1 pretrained model are depicted at upper line while for the proposed method at the lower line. A and D images refer to ICM task, B and E images to TE task, C and F images refer to DE task.

images in the test set, we conclude that these patches are located both in the inner cell mass of the blastocyst and in the trophectoderm region for the ICM classification task. The same pattern is observed in the TE classification task as well. Another important observation is that in some images where visual patterns are found outside the blastocyst, the Grad-CAM erroneously focus its attention on them. All the above discrepancies are excluded from the proposed methodology due to the selection of interest points that are in the respective regions. For the proposed methodology the visual explanations are provided in the lower line of Figure 41.

1.20.3 Qualitative evaluation on proposed ensemble explainability method

Apart from the contribution of TML explainability techniques, an improvement of the explainability technique, Grad-CAM is proposed herein. The first part of the contribution is focused on the application of this technique on ensemble models, while the second part combines the well-established properties of Grad-CAM with superpixels for more fine-grained and more efficient visual explanations on microscopy images. Regarding the interpretability module of the proposed methodology's first part, a test bench application is developed for visual inspection and verification of interpretability results by specialized medical personnel. The web interface (Figure 42) is utilized for the insertion of a test histopathology image. The sample is sent to the back end where the best-performing ensemble architecture analyzes it to return an accurate classification result along with the generation of a heatmap of the original image. The visual patterns of the image that are characterized as highly related to the result are painted red, whereas those irrelevant with blue. The specialists inspect the highly related visual patterns and assess the results according to their prior experience in histopathology image-based diagnosis. The initial qualitative results show significant accordance concerning the areas responsible for the characterization of results between specialists and the ensemble classifier. The images are selected from the test set of BreakHis [132] and Bach's dataset [134] randomly and processed by both Grad-Cam and Guided Grad-Cam interpretability techniques. The visualization and classification results are analyzed by specialized personnel and commented on in terms of their opinion concerning the classification in benign or malignant class and the localization of important visual patterns that are responsible for the classification result. In Figure 43, a benign adenosis is depicted in x400 magnification. The

ensemble classifier classifies the image as probably benign but not being representative with high confidence in contrast to the experienced physician that refers to this image as not being totally representative of the benign class in terms of morphological

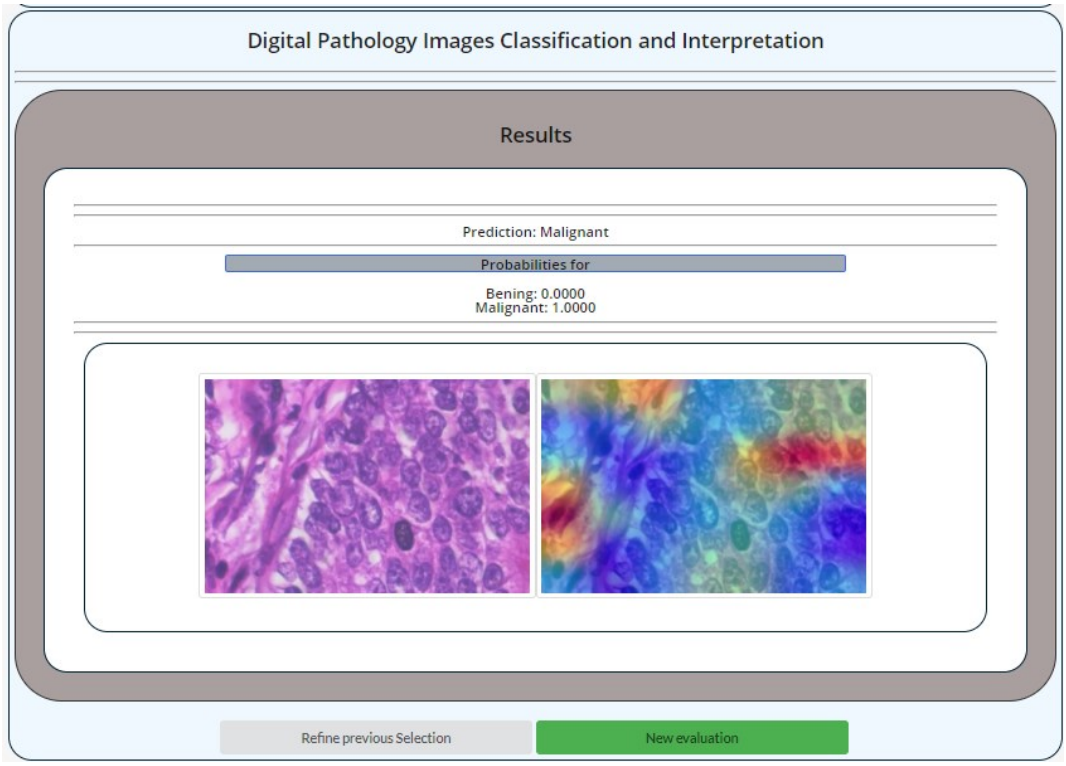


Figure 42 - Overview of the standalone application for the classification and interpretation of histopathology images.

patterns. The red highlighted regions are localized on epithelial tissue, though not totally. Humans tend to point their attention to the specific kind of tissue because carcinomas are malignant neoplasms of epithelial tissue. On the other hand, nearby stromal and epithelial areas are colored yellow as they are in the vicinity of the most important regions. Concerning the Guided Grad-Cam algorithm the coloring of respective areas is fuzzier.

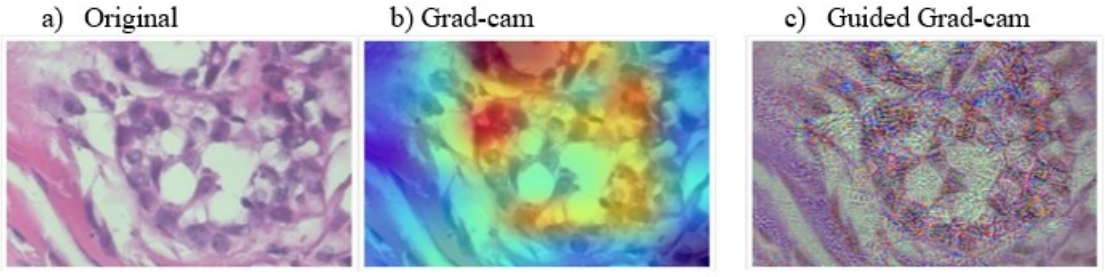


Figure 43 - Application of b) Grad-Cam and c) Guided Grad-Cam interpretability techniques on a) a benign adenosis sample from the BreakHis dataset.

Moving on to the next image presented in Figure 44 which is taken from Bach's dataset and depicts an in-situ carcinoma, the depicted patterns are visually representative of the malignant class. The classifier correctly predicts the class with high confidence and manages to generalize well on an unknown dataset with several variances owing to different production and staining procedures. Concerning the Grad-CAM technique, highly important regions colored red correspond to epithelial cells, whereas in the Guided Grad-CAM case, the coloring of respective regions is fuzzy. Some yellow-painted regions are considered of less importance to the classifier and highlighted due to the vicinity to the most important regions and other yellow regions are colored with no obvious reason to experienced physicians. In other cases, both algorithms fail to highlight the regions which are considered significant by experienced physicians. In Figure 45, drafted from the BreakHis dataset, a benign fibroadenoma is depicted. Fibroadenomas are benign tumors of the epithelial and stromal tissue. The Grad-CAM algorithm highlights mostly epithelial and stromal

regions and ignores epithelial tissue on the lower left part of the image which

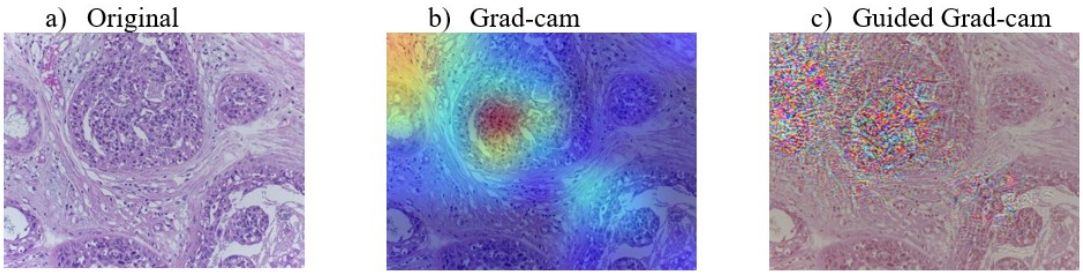


Figure 44 - Application of b) Grad-CAM and c) Guided Grad-CAM interpretability techniques on a) an in-situ carcinoma sample from the Bachs dataset.

is also indicative of the disease. Nevertheless, in terms of morphology, the depicted patterns are not highly indicative of the disease as physicians state.

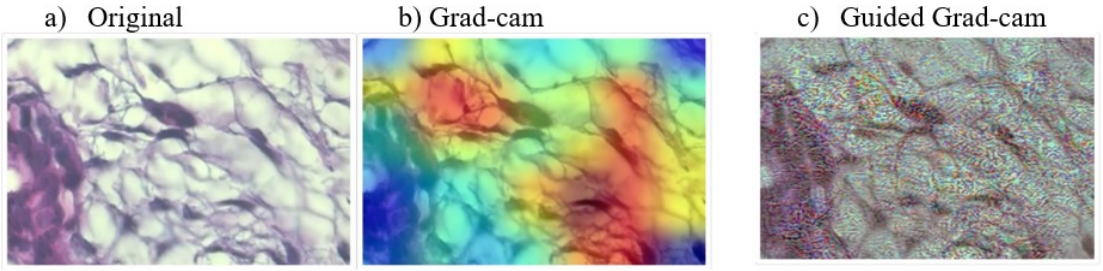


Figure 45 - Application of b) Grad-Cam and c) Guided Grad-Cam interpretability techniques on a) a benign fibroadenoma sample from the BreakHis dataset.

A different case which is characteristic of the interpretability algorithm's deficiency to decode significant regions concerning their influence on classification's result are im-ages that depict uniform patterns as Figure 46.

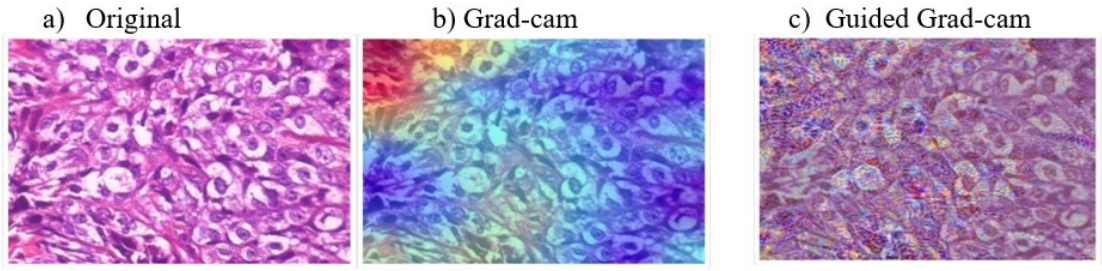


Figure 46 - Application of b) Grad-Cam and c) Guided Grad-Cam interpretability techniques on a) a malignant ductal carcinoma sample from the BreakHis dataset.

1.20.4 Quantitative evaluation on proposed DL explainability method

The second part of the proposed explainability scheme on DL refers to the combination of Grad-CAM with superpixels. This is the only explainability scheme that is evaluated quantitatively by a corresponding metric that requires no prior knowledge of the imaging morphological characteristics. To test the proposed methodology in practice, an application was developed for the visual inspection and verification of the generated results, whereas the technique presented in [140] is utilized for the assessment of explainability performance for the proposed methodology against other counterparts. The utilization of the Grad-CAM explainability technique allows for the utilization of different neural network classification schemes provided that the condition of differentiability exists. Therefore, the proposed technique is applied to various pretrained neural networks and segmentation techniques. Two datasets from different modalities of medical imaging are examined to verify the efficiency of the proposed methodology. Six well-established pretrained neural networks VGG16, VGG19, ResNet50, Resnet101, MobileNet, and EfficientNet B0 are utilized for the classification of images in two classes: benign and malignant. The networks are pretrained on ImageNet [128] and utilized without any

modification as feature extractors. For each network, the metrics of accuracy, precision, recall, and Area under Curve (AUC) metrics are reported for the different image datasets. Consequently, all six pretrained models were utilized for the explainability pipeline. In the search for the most efficient combination for both accuracy and explainability metrics, tests are conducted with the segmentation algorithms Felzenswalb, Quickshift, and Slic. The MoRF and AOPC metrics are utilized to detect the pipeline with the deepest decrease in the MoRF curve and the biggest AOPC area. For this proposed methodology the classification results are presented in parallel with the explainability results in an attempt to draw useful conclusions about the relationship between classification and explainability performance if any exists. The classification results for each of the six pretrained deep convolutional network models are shown in Table 17. The results demonstrate the superiority of the ResNet101 model concerning the classification of the RCM (accuracy score of 0.8667) and BreakHis (accuracy score of 0.9432) images. From Figures 47 to 52, a comparison of each segmentation algorithm for a dedicated neural network is shown. In Figure 47a, and 47c the graphical representations of AOPC and MORF scores concerning the EfficientNet B0 model is presented. The combination of three superpixel segmentation algorithms with the original Grad-CAM approach is presented with different color lines and the area under the perturbation curve is the area bordered by the color line and the x-axis. The AOPC value for the best-performing combination in the case of the EfficientNet B0 is provided by the Quickshift superpixel algorithm for the RCM dataset (AOPC score of 10.8) and the BreakHis dataset as well (AOPC score of 23.5). For a more detailed description of the AOPC scores, Table 18 is provided. The best-performing implementation is the VGG16+Felzenswalb combination for the RCM dataset (AOPC score of 18.9), while

the ResNet101 + Slic implementation takes the first place concerning the BreakHis dataset (AOPC score of 23.9). In Figures 53 and 54, the graphical representations of the combinations with the best explainability results for each classifier are depicted. It is important to mention that when utilizing a segmentation algorithm, the average number of superpixels approximates the number of tiles of the heatmap that is generated by the Grad-CAM technique. The restriction prevents the generation of a bias that larger tiles introduce in favor of one methodology against the other.

Motivated by the work in [24] we applied the presented explainability approaches on an ensemble classifier that is composed of an EfficientNetB0, a VGG19, and a ResNet101 neural network, pretrained on ImageNet. Results of the ensemble classifier performance in classification terms for both datasets are shown in Table 19. Concerning the RCM dataset,

Table 17 - Performance metrics for the classification of RCM and Breast histopathology images in two classes by means of various pretrained neural networks models (best scores in bold).

Configuration	Accuracy		Precision		Recall		AUC	
	RCM	BreakHis	RCM	BreakHis	RCM	BreakHis	RCM	BreakHis
VGG16	0.7600	0.8581	0.7600	0.8581	0.7600	0.8581	0.8268	0.9362
VGG19	0.7633	0.8596	0.7333	0.8596	0.7333	0.8596	0.8233	0.9375
MobileNet	0.7867	0.8915	0.7867	0.8915	0.7867	0.8915	0.8658	0.9624
ResNet50	0.8533	0.9366	0.8533	0.9366	0.8533	0.9366	0.9240	0.9840
ResNet101	**0.8667**	**0.9432**	**0.8667**	**0.9432**	**0.8667**	**0.9432**	**0.9500**	**0.9867**
EfficientNet B0	0.7985	0.8982	0.7945	0.8982	0.7945	0.8982	0.8565	0.9631

the ensemble classifier exceeds the performance of the Resnet101, which is the most efficient base classifier, while the ensemble classifier for the BreakHis dataset is by far the most efficient classifier. In Figures 55 and 56, the graphical representations of the MoRF and AOPC metrics for the ensemble classifier are depicted.

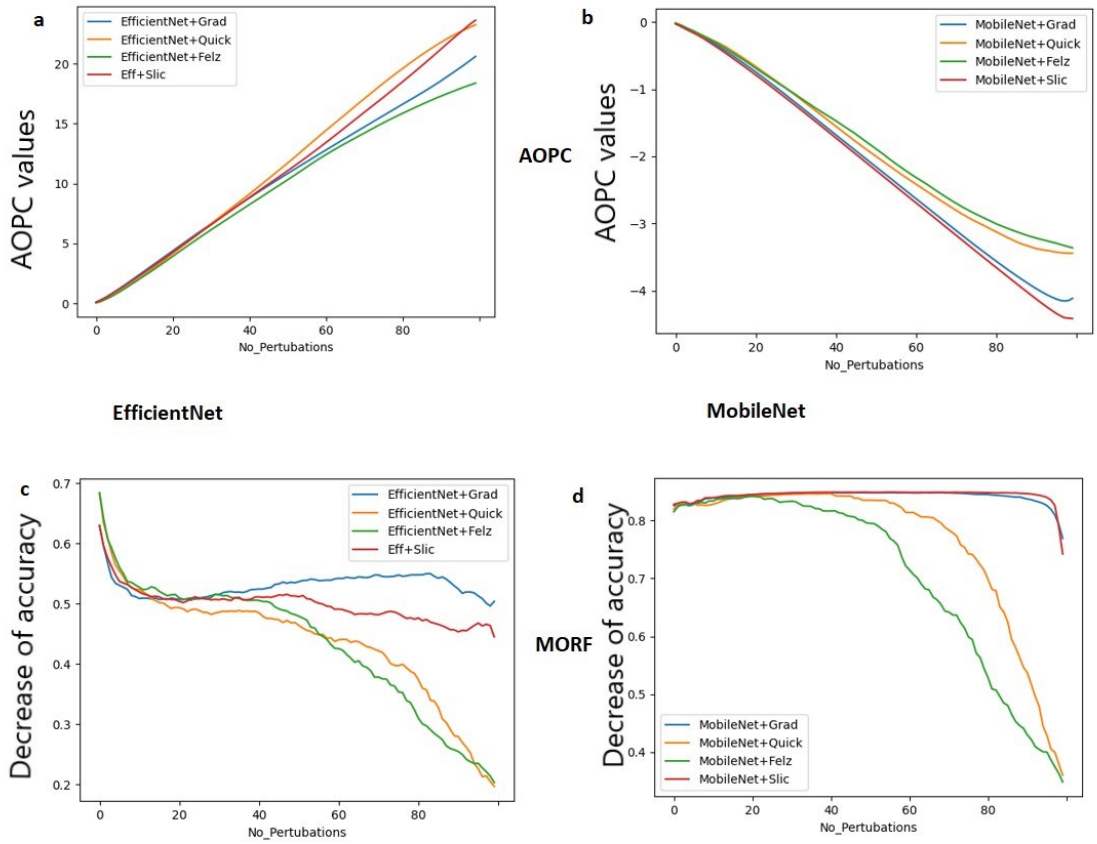


Figure 47 - Results of the AOPC values (a) for the EfficientNet pretrained network; (b) the MobileNet pretrained network, the MORF values (c) for the EfficientNet pretrained network and (d) the MobileNet pretrained network for the Grad-CAM (blue color), Grad-CAM + Quickshift (orange color), Grad-CAM + Felzenswalb (green color) and Grad-CAM + Slic (red color) implementations utilizing the RCM dataset.

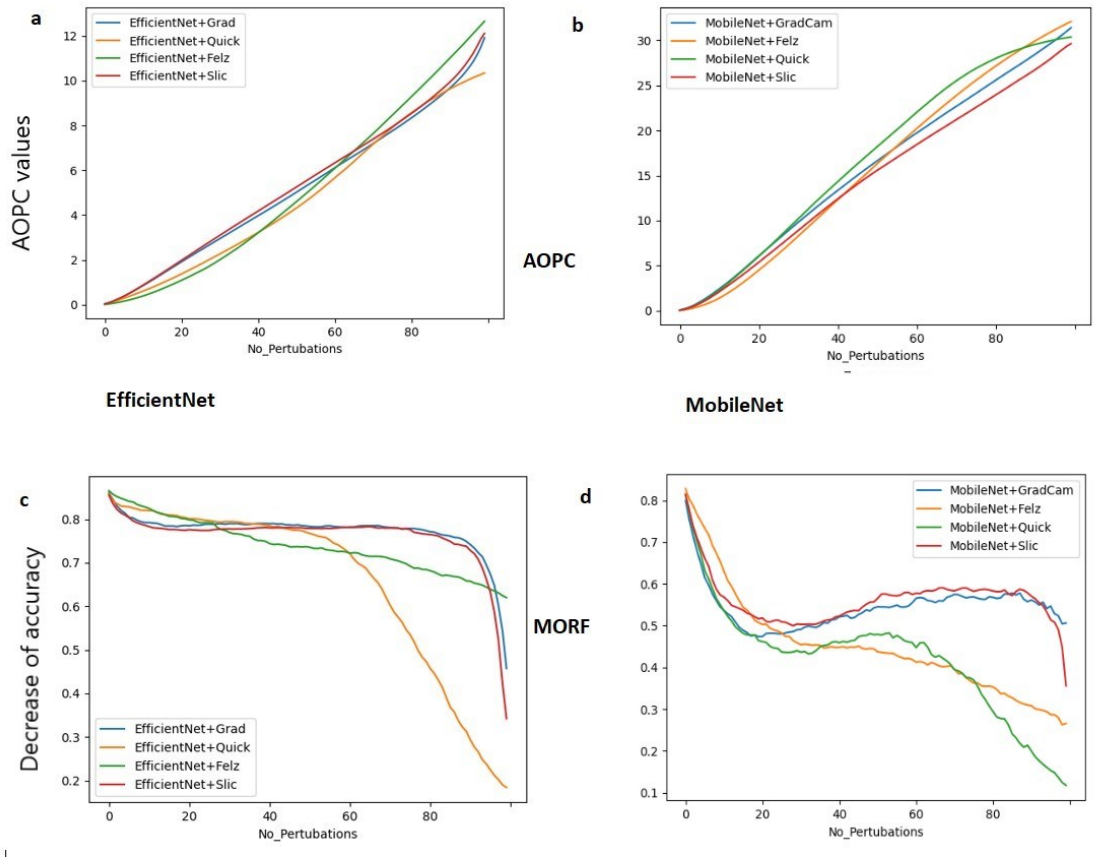


Figure 48 - Results of the AOPC values (a) for the EfficientNet pretrained network, (b) the MobileNet pretrained network, the MORF values (c) for the EfficientNet pretrained network and (d) the MobileNet pretrained network for the Grad-CAM (blue color), Grad-CAM + Quickshift (orange color), Grad-CAM + Felzenswalb (green color) and Grad-CAM + Slic (red color) implementations utilizing the BreakHis dataset.

Table 18 - APOC scores for all combinations of the Grad-CAM + superpixel segmentation algorithm for the RCM and BreakHis dataset.

Configuration	Grad-CAM		+Felzenswalb		+Slic		+Quickshift	
	RCM	BreakHis	RCM	BreakHis	RCM	BreakHis	RCM	BreakHis
VGG16	12.4	10.9	**18.9**	9.9	12.8	11.5	15.9	12
VGG19	**16.2**	17.2	15.1	19.8	**17**	21.5	**16.1**	**23.5**
ResNet50	14.2	19.7	16.5	**23.2**	15.8	22.9	14.9	23
ResNet101	11.7	**21.2**	9.5	20.8	12	**23.9**	11.2	22.2
MobileNet	-2.6	15.8	-1.9	15.2	-2.7	14.5	-2.2	17
EfficientNet B0	10	16.1	9.4	15.9	10.4	16.2	10.8	15.8

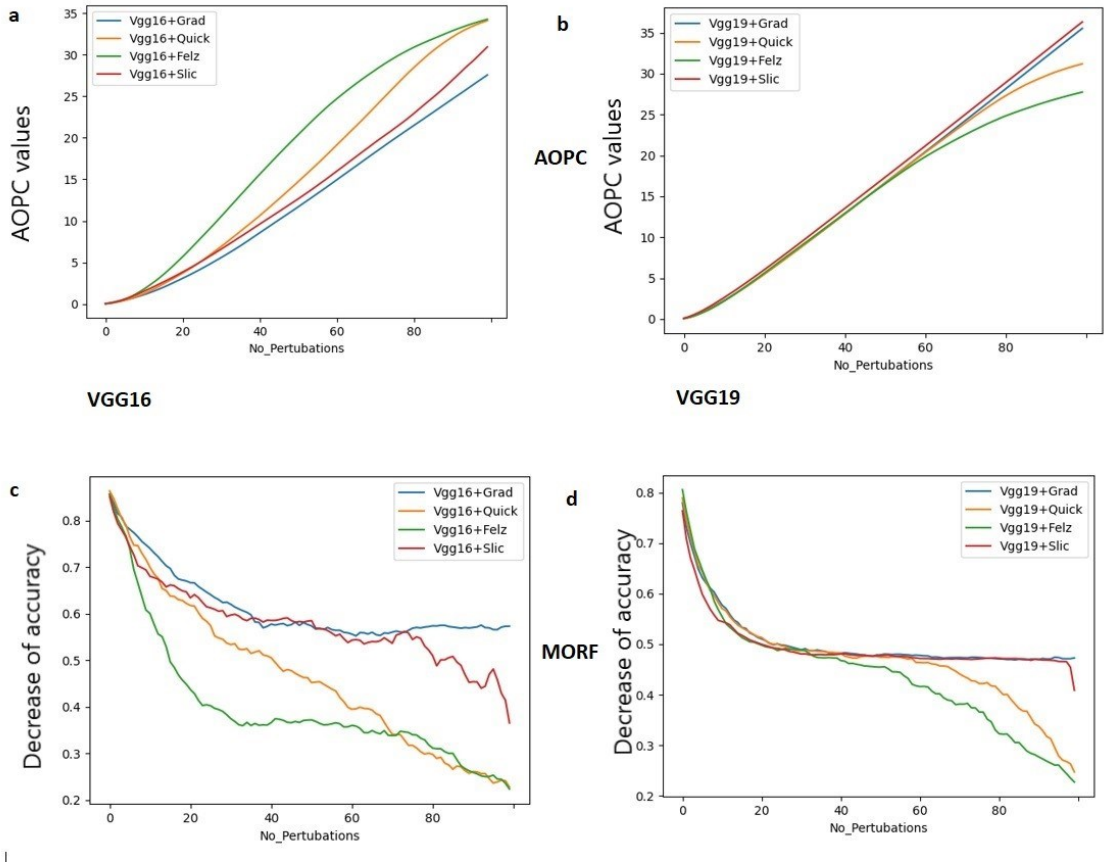


Figure 49 - Results of the AOPC values (a) for the VGG16 pretrained network, (b) the VGG19 pretrained network, the MORF values, (c) for the VGG16 pretrained network and (d) the VGG19 pretrained network for the Grad-CAM (blue color), Grad-CAM + Quickshift (orange color), Grad-CAM + Felzenswalb (green color) and Grad-CAM + Slic (red color) implementations utilizing the RCM dataset.

Table 19 - Performance metrics for the classification of RCM and Breast histopathology images in two classes by means of various pretrained neural networks models.

Configuration	Accuracy		Precision		Recall		AUC	
	RCM	BreakHis	RCM	BreakHis	RCM	BreakHis	RCM	BreakHis
Ensemble Network	0.9333	0.9850	0.9333	0.9850	0.9333	0.9850	0.9484	0.9971

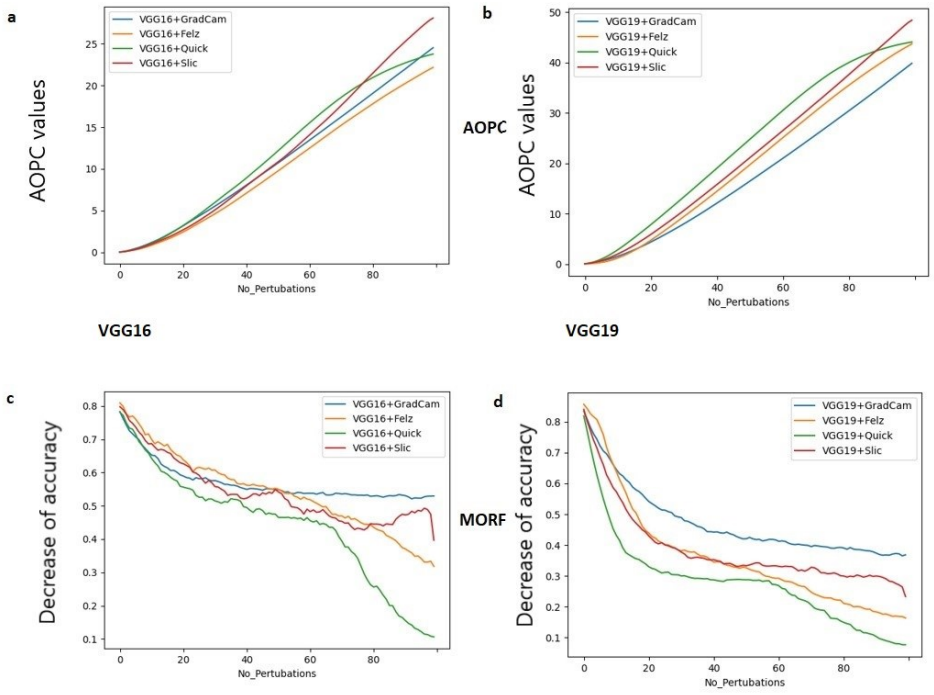


Figure 50 - Results of the AOPC values (a) for the VGG16 pretrained network, (b) the VGG19 pretrained network, the MORF values (c) for the VGG16 pretrained network and (d) the VGG19 pretrained network for the Grad-CAM (blue color), Grad-CAM + Quickshift (orange color), Grad-CAM + Felzenswalb (green color) and Grad-CAM + Slic (red color) implementations utilizing the BreakHis dataset.

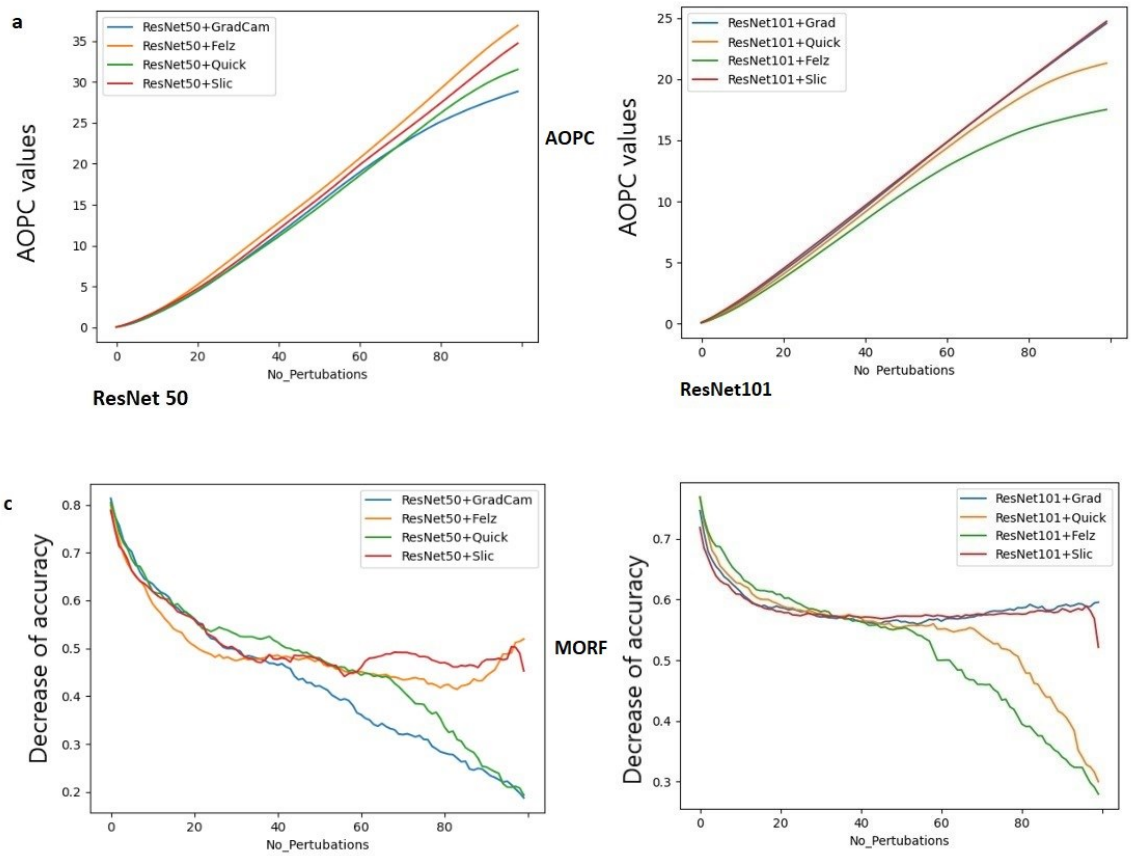


Figure 51 Results of the AOPC values (a) for the ResNet50 pretrained network, (b) the ResNet101 pretrained network, the MORF values (c) for the ResNet50 pretrained network and (d) the ResNet101 pretrained network for the Grad-CAM (blue color), Grad-CAM + Quickshift (orange color), Grad-CAM + Felzenswalb (green color) and Grad-CAM + Slic (red color) implementations utilizing the RCM dataset.

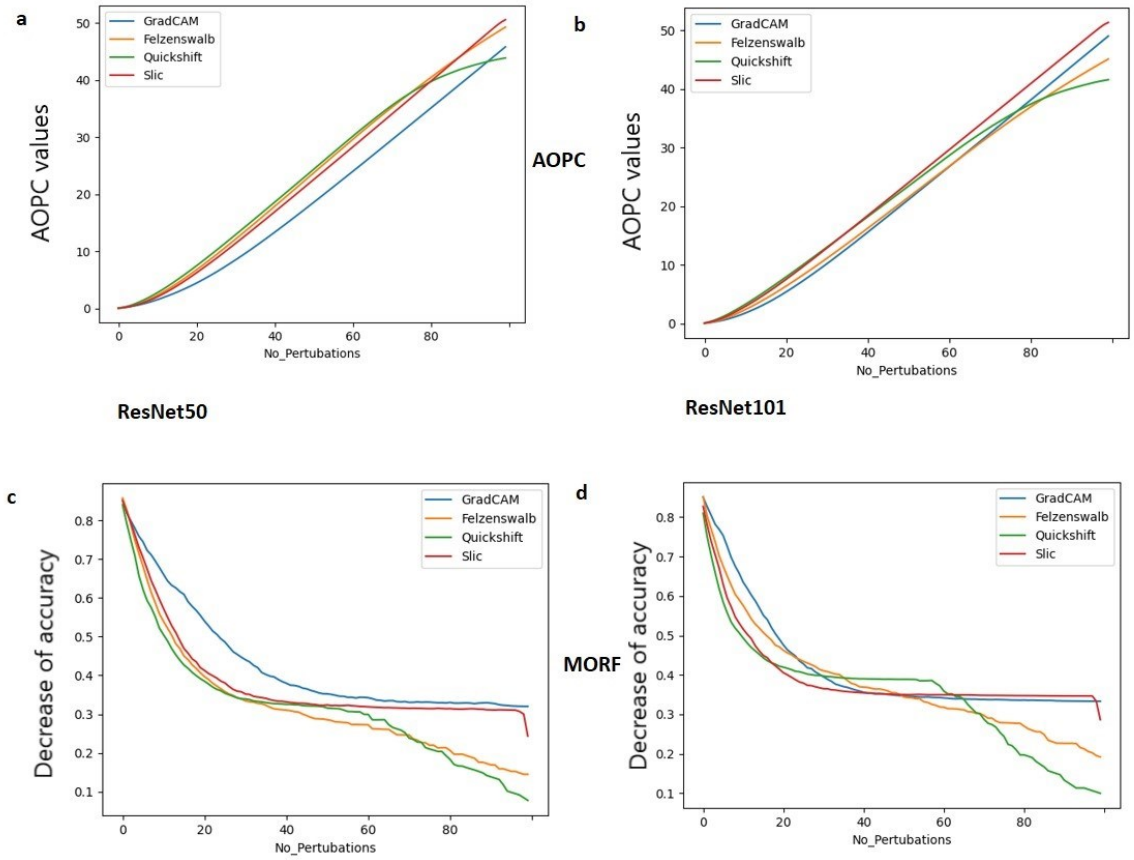


Figure 52 - Results of the AOPC values (a) for the ResNet50 pretrained network, (b) the ResNet101 pretrained network, the MORF values (c) for the ResNet50 pretrained network and (d) the ResNet101 pretrained network for the Grad-CAM (blue color), Grad-CAM + Quickshift (orange color), Grad-CAM + Felzenswalb (green color) and Grad-CAM + Slic (red color) implementations utilizing the BreakHis dataset.

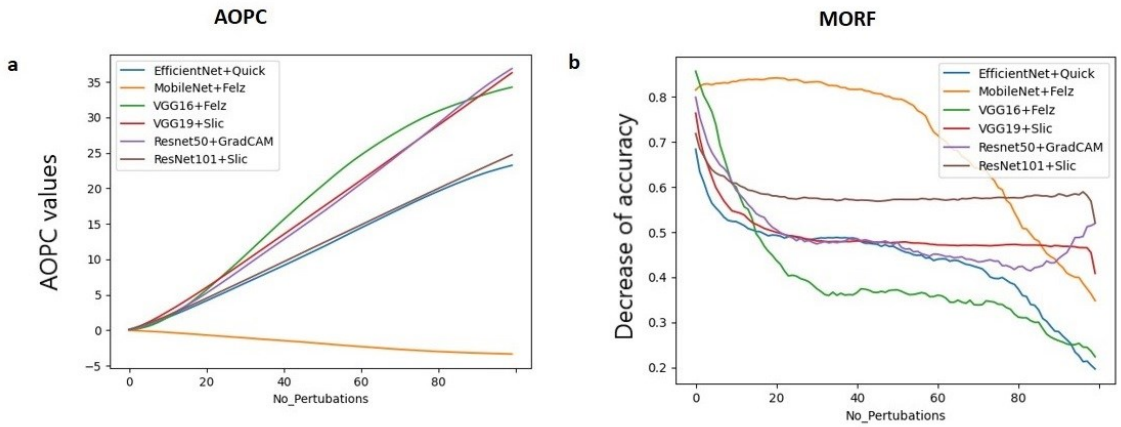


Figure 53 - Results of the best performing combinations [EfficientNet + Quickshift (blue color), MobileNet + Felzenswalb (orange color), Vgg16+Felzenswalb (green color), VGG19+Slic (red color), Resnet50+GradCAM (purple color), Resnet101+Slic (brown color)] by utilizing (a) the AOPC and (b) the MORF metrics for the RCM dataset.

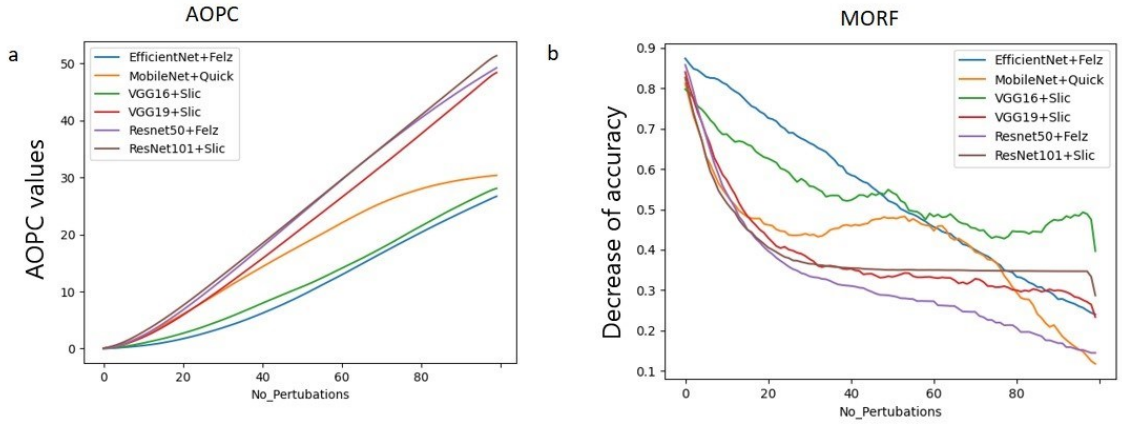


Figure 54 - Results of the best performing combinations [EfficientNet + Felzenswalb (blue color), MobileNet + Quickshift (orange color), Vgg16+Slic (green color), VGG19+Slic (red color), Resnet50+Felz (purple color), Resnet101+Slic (brown color)] by utilizing (a) the AOPC and (b) MORF metrics for the BreakHis dataset.

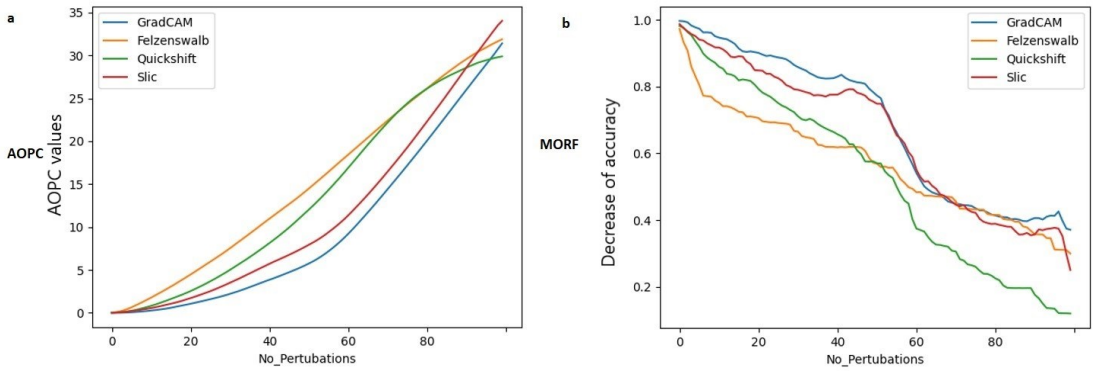


Figure 55 - Results of the Grad-CAM (blue color), Grad-CAM + Quickshift (green color), Grad-CAM + Felzenswalb (orange color) and Grad-CAM + Slic (red color) when applied on an ensemble pretrained convolutional network utilizing (a) the AOPC and (b) MORF metrics for the BreakHis dataset.

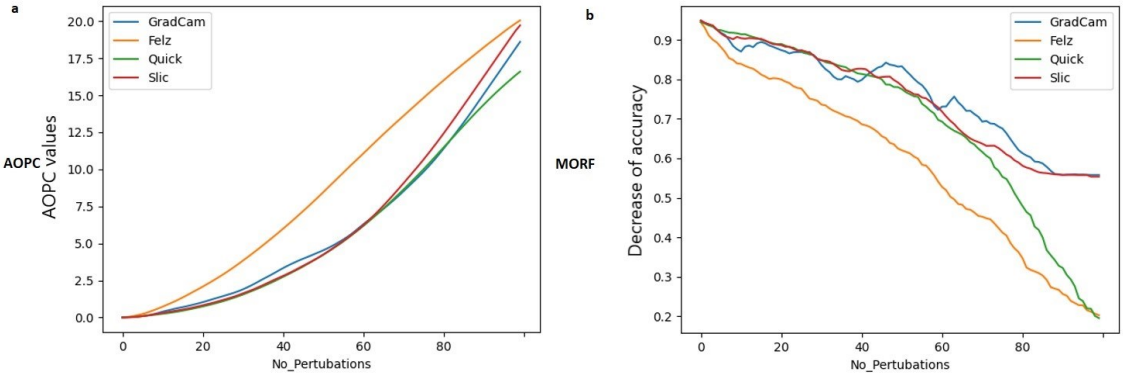


Figure 56 - Results of the Grad-CAM (blue color), Grad-CAM + Quickshift (green color), Grad-CAM + Felzenswalb (orange color) and Grad-CAM + Slic (red color) when applied on an ensemble pretrained convolutional network utilizing (a) the AOPC and (b) MORF metrics for the RCM dataset.

1.21 Histopathology image retrieval system in practice

An application was developed in java programming language utilizing the libraries Openslide (www.openslide.org), ImageJ Surf (www.labun.com/imagej-surf), and PDFBOX (www.pdfbox.apache.org). The GUI of the application, as depicted in Figure 57,

is divided into two panels, the screen panel, and the control panel. Four basics buttons are provided on the upper section of the main menu, as follows:

• Simple Image/WSI.

• Open Image.

• Convert PDF to a file.

• Close App.

The functionality of each button is briefly described below:

- Simple Image/WSI. By pressing this button, the type of query image is selected. The user can choose between two options: A simple digital image or a whole slide image.
- Open Image. The selected query image is opened in order to be viewed and processed (for WSI).
- Convert PDF to a file. A digital pathology atlas can be selected and converted into a folder of digital images.
- Close App. Self-explained.

Once a query image is selected, the requested image is visible in the screen panel section, as shown in Figure 57. In this case, the query image is a whole slide image produced by a Trestle whole slide scanner and its format is single-file pyramidal tiled TIFF (tagged image file format). In the control panel additional functionality appears by means of two new panels, SURF parameters and properties, and three buttons: compare the image to the handbook, compare the image to the folder, and reset points. By using the SURF parameters panels the user can specify the parameters of the SURF algorithm for the detection and description of interest points. In the properties field, the user can view the

meta-data that is stored in

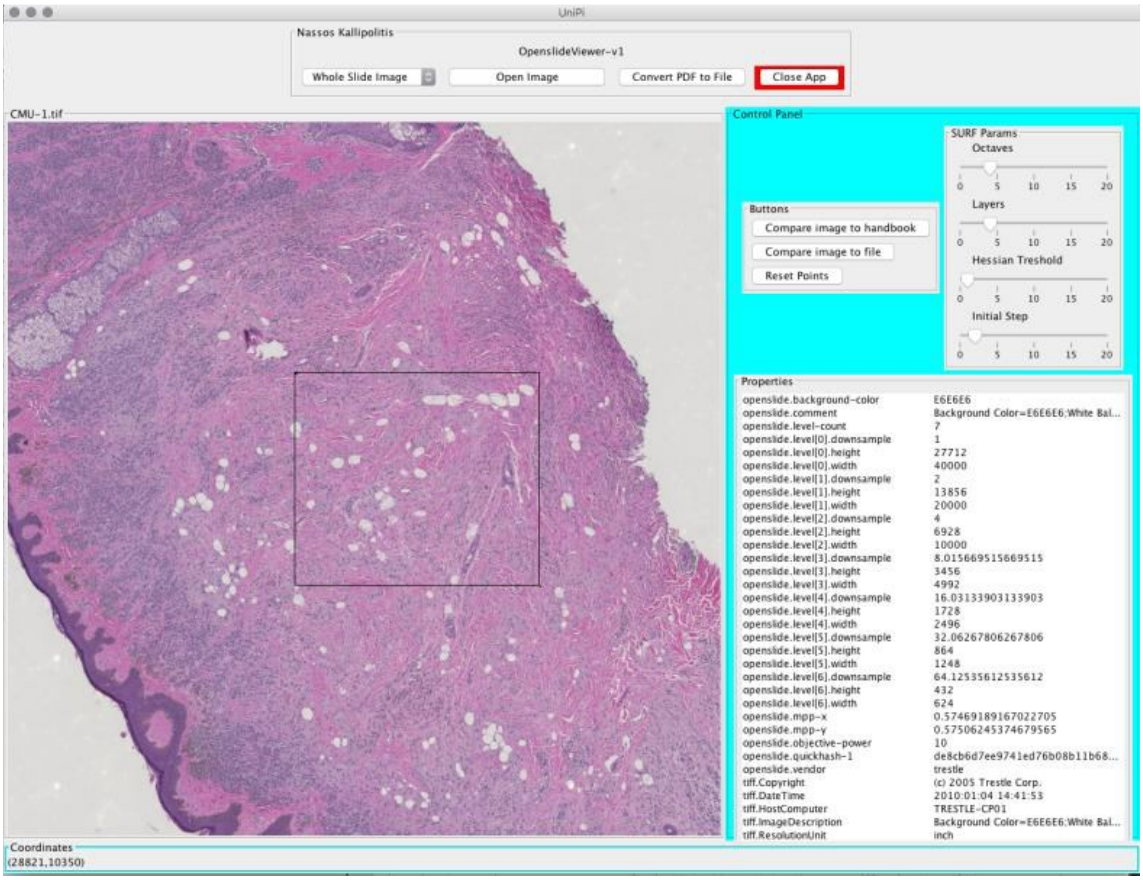


Figure 57 - Main menu of the histopathology image retrieval application.

a whole slide image in reference to the image attributes. The buttons serve the purpose of choosing the dataset of digital images that will be compared to the query image. By choosing a dataset of images and query images, the process of image retrieval begins and, in the end, the result appears on the screen and the image/s with the most similarities to the query image is/are shown. A series of experiments are conducted by comparing a query image with another sample image. The images can be simple (.jpg, .png, .bmp), whole slide images, or images extracted from medical atlases in .pdf format. The query image is altered by applying different brightness, rotation, and scale transformations. Each time the SURF algorithm is applied and a number of matches between the query image and the image

dataset are detected. The basic criteria for understanding the level of influence posed by the transformations are the number of interest points detected in each transformed image (test image) and the number of matches found between the query image and each test image. The first set of conducted experiments uses the query image and a set of six

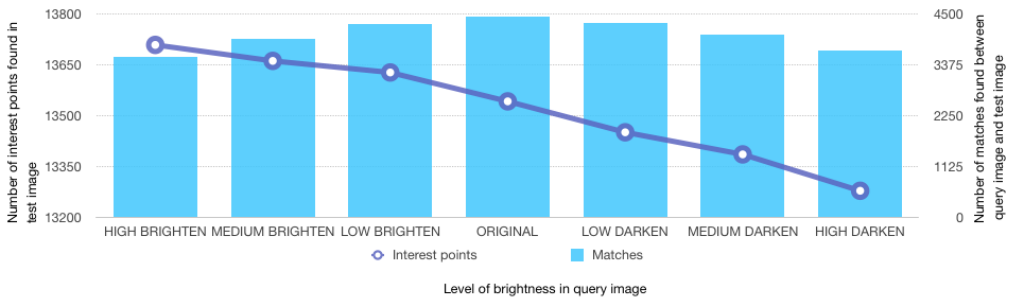


Figure 58 – Graphical representation of matches found between query image and test image and interest points found in test image as the brightness changes.

variations of the query image from the brightest to the darkest one. The results are shown in Figure 58. The number of interest points increases as the image gets brighter and decreases as the image gets darkened. However, the increase in the number of interest points does not necessarily mean an increase in the number of matches between the query and test image. The influence of the variations in brightness is slightly stronger when the pathology image gets darkened (22% fewer matches) with respect to the brightest image (19% fewer matches). The second set of experiments refers to rotation transformations. Results are illustrated in Figure 59.

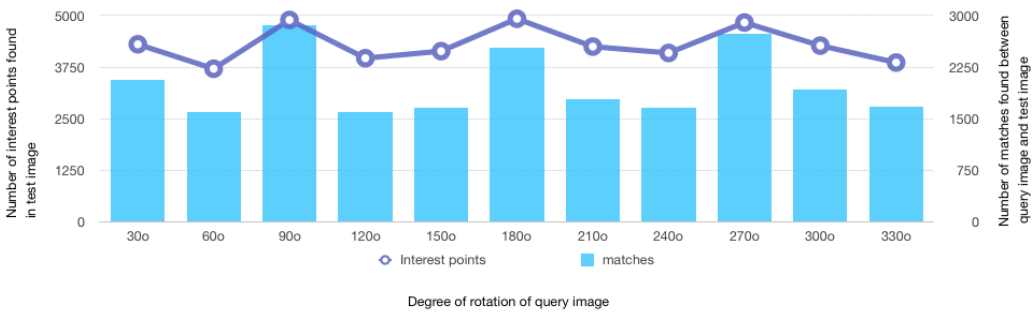


Figure 59 – Graphical representation of matches found between query image and test image and interest points found in test image as the rotation changes.

The worst case for the rotation transformations occurs at 120 degrees (37% fewer matches), which is more intense that the worst case for brightness transformations. Transformations related to the scale up and down of the query image are also tried out to check the effect of these transformations on the function of the SURF algorithm. The results are shown in Figure 60. As it is projected by the graphical representation (Figure 60) the effect of minimizing the query image is devastating more than any other transformation performed (97% fewer matches).

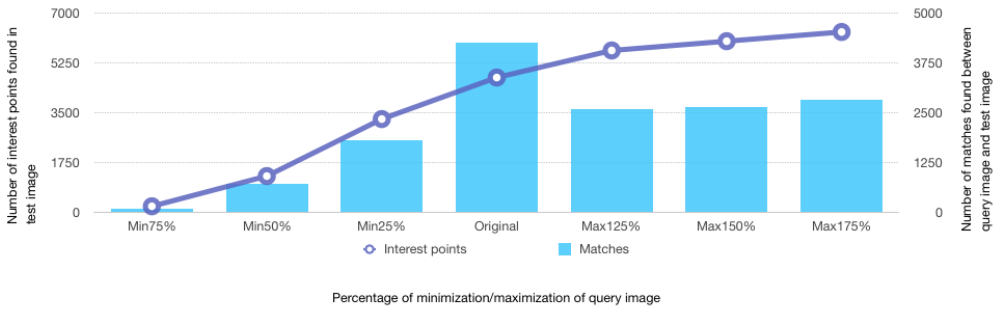


Figure 60 – Graphical representation of matches found between query image and test image and interest points found in test image as the scale changes.

Maximizing the image has a smoother impact on the algorithm with a 33% decrease in matches. Apart from the experiments performed with reference to the transformations of

the query image, tests were conducted with the different values of the following parameters octaves, and the hessian threshold explained earlier, is assigned to check the influence of these variations. The results for variations of the octaves and the threshold parameter are shown in Figures 61 and 62.

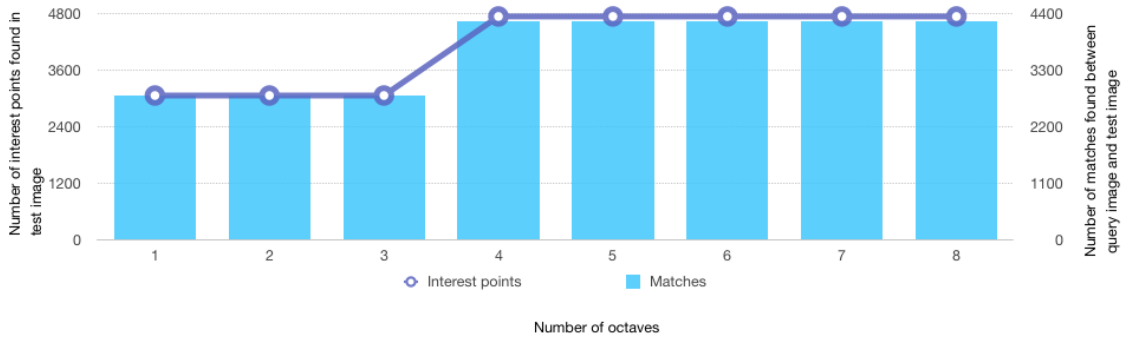


Figure 61 – Graphical representation of matches found between query image and test image and interest points found in the query image as the octave parameter changes.

Figure 62- Graphical representation of matches found between query image and test image and interest points found in the query image as the threshold parameter changes.

In Figure 61 it is highlighted that the most matches are provided for four octaves and more, whereas in Figure 62 the best results are given for a 0.00009 value of the parameter threshold since this value ensures the most matches for the fewer interest points found.

One last experiment is conducted on 4 groups of similar images, which depict the effect of three different treatments (drug, radiation, drug, and radiation) on cancer cells. The dataset, which is comprised of .jpg images, is described in [10]. This experiment proves the effectiveness of the SURF algorithm in retrieving the most similar image/s even when the examined dataset is comprised of images that bear a great resemblance. Using the

Table 20 - Confusion matrix of classification of 24 images in four classes (control, drug, radiation, drug and radiation)

	Actual class			
Predicted class	**Control**	**Drug**	**Radiation**	**Drug and Radiation**
Control	5	0	0	1
Drug	0	4	1	1
Radiation	0	2	3	1
Drug and radiation	0	1	1	4

application, the query image, selected from one of the groups (control, drug, radiation, drug, and radiation), is compared to each image of the dataset. Once the image with the greater similarity (best image) to the query image has been found, the query image is classified in the group of origin of the best image with success 67%, as illustrated in the confusion matrix below (Table 20).

DISCUSSION

1.22 TML Explainability techniques

Having discussed the evaluation of the classification schemes in the lines earlier, as it is necessary to provide proof of the predictive model's ability to distinguish and classify the visual patterns into corresponding categories, in this part of our work, we focus on the visual explanations of the predictive models. The need for classification evaluation is inevitable since the visual explanations of a model that has poor performance are useless. Starting from the BOVW explainability technique, more samples need to be utilized in order to shape a solid opinion about the performance of the classifier and exclude the possibility of overfitting despite the promising classification accuracy for the binary classification task. Towards this direction, the addition of new RCM images in the training procedure is pending. With the extension of the dataset, data demanding deep learning techniques will be applied as well. It should be taken into account that the related classification scheme may have poor performance if the number of samples is higher. However, in ML this is not always the case and, therefore, an additional explainability tool in our attempt to understand the decision-making process of a predictive model can only be beneficial. The functionality of the interpretation scheme provides a plausible explanation concerning the visual patterns that determined the classification result. However, future work should be directed toward the determination of weights concerning the impact of each influencer (W_{vw}, D_{ip}) on the final influence indicator. Furthermore, the elimination of skin fold patterns in confocal images has been proven to be an important factor in more accurate classification results. As described in the case of a confocal image

depicting benign patterns in Figure 40, the presence of an artifact can have devastating results on the predictive outcome. Having the ability to visualize this trend and avoid it is a very important tool in an attempt to improve the predictive algorithm. Furthermore, the scheme can provide useful insight into the correspondence between the decision-making of the model and human expertise.

Moving forward to the use case of blastocyst images and the proposal of an explainability scheme upon the FV technique, the extended explainability approach improves our knowledge of the model's inner mechanisms. The approach provides fine-grained visual explanations that are in accordance with the experts' experience. On the other hand, the visual explanations that are provided by the Grad-CAM technique focus on patches of the images that constitute a confounding factor. In the ICM classification task, highly influential patches are erroneously detected in the TE region and the TE classification task important patches are found in the ICM region as well. The decision of the neural network is misguided by confounding factors that lead to the assumption that a good quality of the inner cell mass region corresponds to a good quality of the trophectoderm region and vice versa. Concerning the classification results, the proposed technique provides comparable results with many of the deep learning schemes. Regarding misclassifications, in most cases, the classifier fails to identify the correct class by categorizing it to the nearest next. This points out the need for labeling images by a consensus of multiple embryologists. Future work is focused on the quantitative results of the unsupervised segmentation and explainability techniques and the discovery of generalization properties of the presented methodology by the utilization of datasets provided by different laboratories.

An important note to keep in mind is the fact that these vocabulary-based predictive models are utilized for many different ML applications spanning from NLP and knowledge extraction from health data. Therefore, the proposed explainability schemes can be effortlessly utilized for explaining the predictive results in various domains. In [141], the BOVW explainability scheme is applied to image sequences depicting people with psychological disorders. The objective of this work is to put in numbers the degree of symptom severity based on the social behavior and cognitive functioning of mental patients when conducting a routine conversation with their attending doctor. In the work, the technical details of the implementations of a video classification methodology for the prediction of schizophrenia symptoms' severity are described, the BOVW explainability approach for the interpretation of video classification results is introduced and initial results are presented where it is demonstrated that the proposed automated techniques can classify to a certain extend specific indicators for the extent of the mental disease.

1.23 DL Explainability techniques

Moving to the field of DL-based explainability approaches, the experimental results are produced by the application of the proposed methodology on histopathology and RCM datasets. The utilization of different datasets enables the evaluation of the classification and interpretation scheme in terms of performance results concerning images belonging to the same dataset and images from different datasets (exploring generalization properties), and in terms of localization-importance quality. Evaluating the classification accuracy with the utilization of images belonging to the same dataset shows that the task is trivial even for the plain architectures (not ensemble ones). The EfficientNets series supersedes other well-established architectures (VGG, InceptionNet, ResNet, ExceptionNet) and achieve

higher performance in both accuracy and AUC metrics for breast and colon datasets even when the training-test split is 60-40%. The results leave a small space for improvement in the case of applying the ensemble architecture. However, in some cases, such improvement occurs. The signs of better performance are more evident when splitting the datasets in a 40-60% or 30-70% ratio. These extreme setups make it more difficult for the plain architectures to perform as well as the ensemble configurations and, therefore, stress the fact that the added complexity of ensemble classifiers is useful in further improving accuracy.

Utilizing ensemble architectures in order to achieve better results, hinders the effort of interpretability due to the added complexity. However, that is not the case for the Grad-CAM and Guided Grad-CAM techniques which are seamlessly integrated into the network's architecture. The quality of highlighting and detecting correctly the most important regions concerning the final prediction is evaluated by experienced physicians. The interpretability module manages to highlight with red color (highly significant) regions of the images that are indicative of the presence or absence of the respective pathology in most cases concerning images of the same dataset. The red highlighted regions are usually epithelial cells, and in the case of malignancies usually are atypical cells with hyperchromatic (dark-colored) nuclei, which is in accordance with the common practice of the physicians. However, the highlighting is not performed for all similar regions in an image which would be desirable, and, in some cases, it is localized in dark-colored artifacts. Therefore, the implementation of an artifact removal methodology would further enhance the generated results. Yellow-colored regions (less important regions) are generated by the interpretability module of the Grad-CAM technique in regions in the vicinity of red-

highlighted regions. A positive aspect of the method, as shown in Figure 44, as a representative sample of cases deriving from Bach's dataset, is the fact that it generalizes well on unseen data. An important drawback of the proposed interpretability methodology is the failure to highlight important regions when the morphological characteristics of the disease are uniform. To a certain extent, it is acceptable since there is no particular region that excels to highlight, and the granularity of the proposed methodology is coarse. Although the Guided Grad-cam technique was intended to solve the issue of granularity, the provided visualizations are fuzzier than the ones presented by Grad-CAM.

By reviewing the generated results of the proposed explainability scheme with the superpixels enhancement module, it is clearly shown that the presented versions of the Grad-CAM technique are more efficient than the original version apart from the EfficientNet B0 and the ResNet50 models with reference to the BreakHis dataset. In general, the segmented tiles demonstrate higher capability in decreasing the confidence of the predictive model when compared to the original tiles that are provided by the Grad-CAM technique. This is made evident by the steeper descent of the MoRF curve and the bigger AOPC areas that characterize the enhanced versions in comparison to the plain Grad-CAM. Theoretically, the designation of specific structures that share the same visual characteristics, thus corresponding to cellular structures would have a stronger effect, a belief which is verified in practice by the AOPC values. While all configurations provide plausible explainability results meaning that the removal of most relevant regions leads to the decrease of the importance function, the MobileNet results for the RCM dataset do not follow the same rule. In the respective graphical representation, it is shown that there are marginal negative responses instead of positive ones as we would have expected. When

comparing the AOPC curves to the classification results, there is no clear pattern that can be verified between explainability and accuracy. Although in the case of the BreakHis dataset, a close relationship between AOPC values and classification accuracy is registered, especially if VGG19 is excluded, the case is not the same for the RCM dataset. In both cases, the EfficientNet classifier is affected to the minimum by the enhancement of various superpixels approaches, since all results are concentrated in a small area. If we were to distinguish a superpixel segmentation scheme that outperforms all plain Grad-CAM implementations for each dataset that would be Slic for the RCM dataset and Quickshift for the BreakHis dataset. Moreover, the graphical representation (Figure 63) indicates that classifiers with the best accuracy are among the best performing in terms of explainability results, although there exist low-performing classifiers that show high AOPC values.

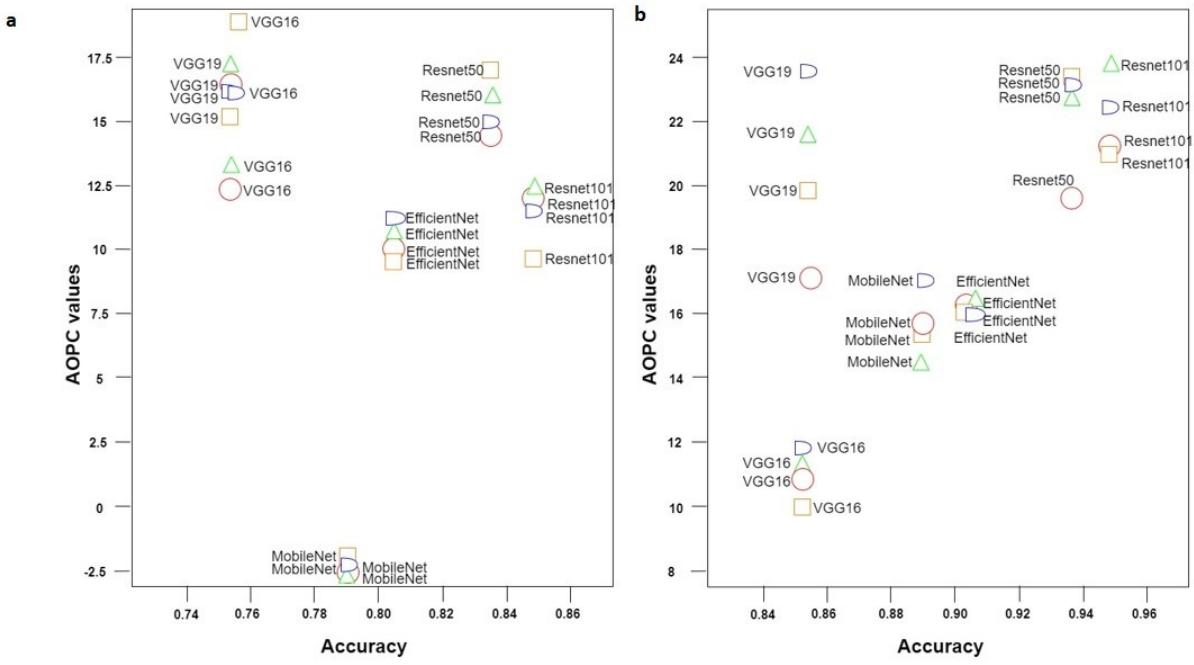


Figure 63 - Results of AOPC values with respect to performance of all configurations for (a) the RCM dataset and (b) the BreakHis dataset. Plain Grad-CAM technique is shown with red color (or circle), Grad-CAM + Slic with green

color (or triangle), Grad-CAM + Quickshift with blue color (or half ellipse) and Grad-CAM + Felzenswalb with orange color (or square).

Concerning the ensemble scheme, the neural networks with the Felzenswalb segmentation algorithm perform better than the other configurations in terms of explainability whereas the plain Grad-CAM approach shows a smaller AOPC curve. These results show that the proposed technique can be applied with a positive effect on ensemble schemes as well. This is due to the properties of the Grad-CAM approach that allows for integration in any CNN with minimum restrictions. In Figure 64, examples of the application of the proposed methodology in samples of both datasets are provided in comparison to the plain Grad-CAM explainability scheme. By the inspection of these samples, it is deduced that the delineation of boundaries with the color of importance gives a clearer view of the designated area, the boundaries are finer and specified to the cellular structures that are depicted. In general, the Slic superpixel algorithm provides boundaries that are closer to the Grad-CAM rectangular paradigm, and can, therefore, be utilized when other superpixels fail to detect extreme patterns of cellular structures. On the contrary, Felzenswalb and Quickshift algorithms demonstrate better flexibility in creating boundaries far from the rectangular shape. In all cases of superpixel algorithms, this flexibility can be adjusted. Focusing on the benign fibroadenoma sample, the proposed methodology provides small evidence of important regions. The result is in accordance with the human experience that classifies the sample as benign due to the absence of abnormal patterns and judging by the depicted uniformity. In terms of visualizations results, the delineation of important areas with a color that is representative of their importance provides an easier preview and assessment of the generated

results for physicians in contrast to the shadowing of a rectangular region (Grad-CAM). Cellular

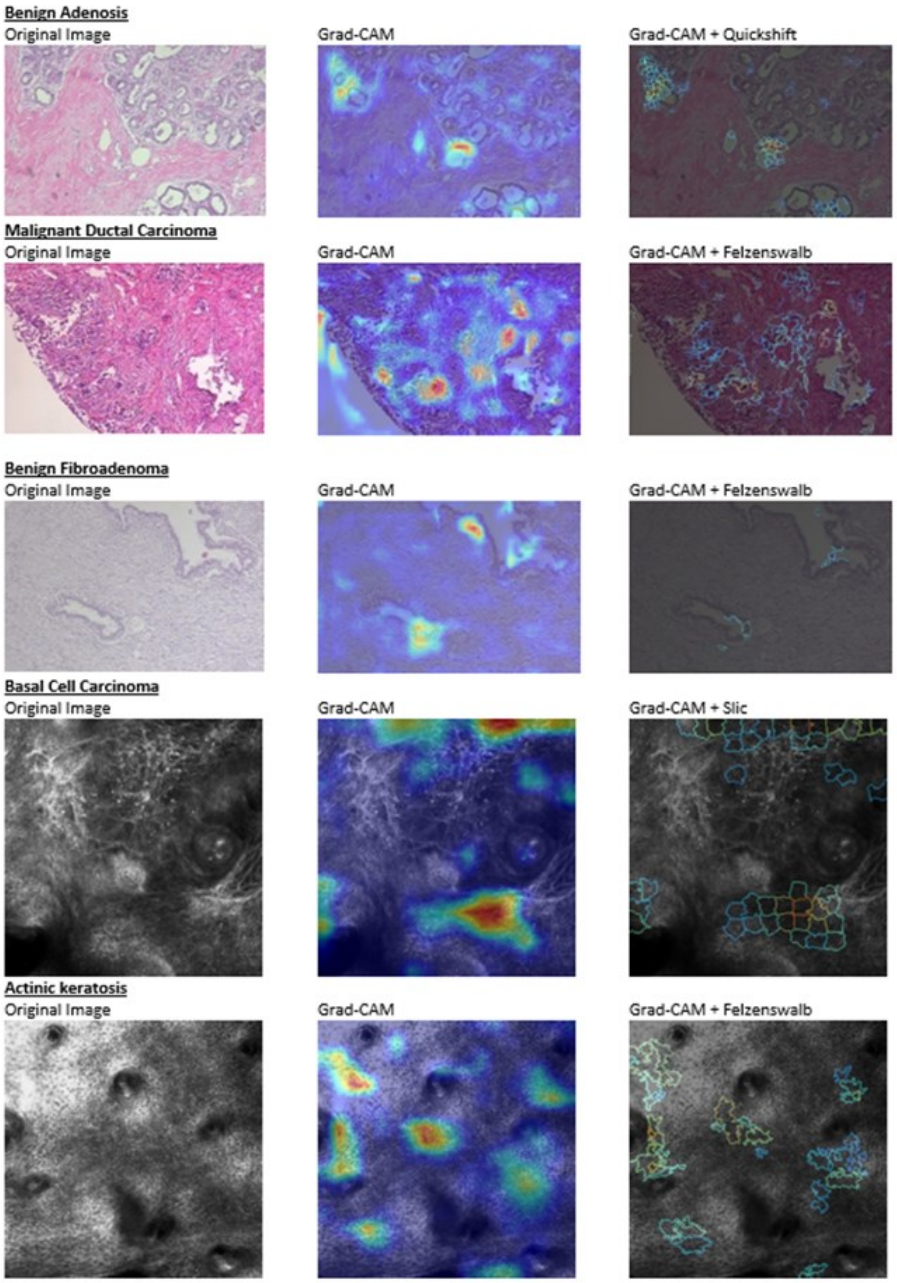


Figure 64 - Examples of results by utilizing data sets' images at random.

structures with common visual characteristics are highlighted in detail and their content is not shadowed by a different color. Earlier works that utilize the Grad-CAM technique

for the provision of visual explanations on microscopy images fail to receive well-defined regions of interest [24, 142]. Towards the delivery of more fine-grained results, the Guided Backpropagation [143] or Guided Grad-CAM [45] is utilized in [144, 145]. However, as presented [146], only Grad-CAM passes the sanity checks for the corresponding saliency maps and therefore can establish a solid base upon which we can build better-defined explanations. Moreover, to the best of our knowledge, most of the presented work requires the existence of experts for a qualitative evaluation concerning the provided saliency maps. The exploitation of the AOPC metric provides a quantitative measure of the performance of the explainability technique that requires no prior knowledge.

It should be taken into consideration that the proposed methodologies have been testified only for a few medical imaging datasets and further exploration of their special characteristics will certainly verify their validity. The quantitative analysis that took place in the case of the combined Grad-CAM superpixel approach needs to be fulfilled for the other schemes as well, while the enhancement of generated heatmaps from other explainability methods will provide a better understanding of the ways that visual explanations should be presented to experts.

CONCLUSION

The thesis deals with the exploration of explainability techniques on existing ML approaches for the classification of medical microscopy images and other modalities. The main contribution presented herein is the proposal of two new explainability schemes for TML classification techniques, namely BOVW and FV, and the improvement of the existing Grad-CAM approach with superpixels in order to better refine the provided coarse explanations by taking into account common visual characteristics on medical images. With reference to the TML approaches, the proposed methodologies can be effortlessly applied to classification tasks other than medical images and present a straightforward manner toward the transparency and trustworthiness of TML predictive models. The DL-presented approach is directed to medical imaging since its special properties require the formation of a more fine-grained solution that returns semantic feedback. Apart from the qualitative analysis of presented visual explanations, the utilization of the AOPC metric has proven to provide objective feedback on the explainability scheme's performance without the need for human expertise. In this thesis, the following methodologies with reference to the visual explanations of medical images are presented, thoroughly explained, and evaluated qualitatively or quantitively:

- A method for the interpretation of visual patterns in skin cancer confocal images. The classification model upon which the interpretation scheme is based forms a visual vocabulary from Speeded up Robust Features (SURF) and utilizes a simple shallow artificial neural network with fully connected layers. Interpretability of the predictive models is an important task since it improves their reliability, accountability, and

transparency and provides useful insight into how to evolve the predictive model towards better performance.

- A novel unsupervised segmentation scheme for the separation of trophectoderm and inner cell mass area provides a significant boost to the performance of traditional machine learning techniques.
- An explainability technique that is based on the information retrieved by the Fisher Vector's generative model provides the necessary connection between the visual stimuli and the predicted results. The classification results of the proposed methodology are comparable with state-of-the-art deep learning techniques and are accompanied by corresponding visual explanations that reveal the inner workings of each model and provide useful insight concerning the predictions' validity.
- An explainability scheme is applied to ensemble classifiers while providing satisfactory classification results of histopathology breast and colon cancer images in terms of accuracy. The results can be interpreted by the hidden layers' activation of the included subnetworks and provide more accurate results than single network implementations. Despite the earlier belief related to deep convolutional networks being treated as black boxes, important steps for the interpretation of such predictive models have been also proposed recently. However, this trend is not fully unveiled for the ensemble models. The interpretation of the predictive model takes place inside the model (in-model) in contrast to other schemes that utilize secondary models offline to interpret the mechanisms of the primary model. The discovery of new correlations between the cause and the result can lead to new findings concerning visual patterns that were previously not considered important or the opposite. Highlighting important visual patterns in medical images of such

density can aid the training of inexperienced personnel. The visual explanations by the proposed methodology are straightforward and provide an influence metric for each corresponding descriptor. This influence metric is visualized by means of a colormap which is indicative of this influence.

- A segmentation-based explainability scheme that focuses on the common visual characteristics of each segment in an image to provide enhanced visualizations instead of highlighting rectangular regions is proposed to further improve the performance of the Gradient - Weighted Class Activation Mapping technique and the generated visualizations. The explainability performance was quantified by applying random noise perturbations on microscopy images. The Area over Perturbation Curve is utilized to demonstrate the improvement of the proposed methodology when utilizing the Slic superpixel algorithm against the Grad-CAM technique by an average of 4% for the confocal dataset and 9% for the histopathology dataset. The results show that the generated visualizations are more comprehensible to humans than the initial heatmaps and demonstrate improved performance against the original Grad-CAM technique. Apart from the fact that the improved visualizations provide better explanations of the generated classification results adding to its transparency and trustworthiness, the proposed methodology can be utilized for the accurate delineation of malignant regions on medical images. Future work will be focused on the utilization of more advanced segmentation techniques and the application of different explainability approaches. Combining superpixels with conditional random fields is reported to improve segmentation results [147] in comparison to the proposed superpixels algorithms that occasionally result in over-segmentation. Furthermore, it should be mentioned that it improves the explainability performance in most cases and

visualizations with respect to the initial Grad-CAM approach and can be easily employed for different neural network configurations with the restriction of differentiability. The initial hypothesis that a segmentation algorithm can improve the results of the Grad-CAM technique is verified and the modularity of the proposed methodology allows for the implementation of different classification, explainability, and segmentation schemes. No specific combination between classifiers and segmentation approaches stands out as best-performing, therefore the selection of a particular superpixel algorithm is data dependent.

Apart from the contribution to the field of explainable artificial intelligence, an image retrieval automated tool for the detection of similar histopathology images in medical books is presented. The fast and consistent properties of the Speeded-Up Robust Features (SURF) algorithm are analyzed in order to search in the content of a digital pathology image, and detect and find similarities for content-based image retrieval. An important aspect of this work is the diversity of Whole Slide Scanners. The proposed methodology that involves the process of the comparison of digital pathology images, mostly WSI, with the use of the SURF algorithm was proved robust to various condition changes.

REFERENCES

[1] L. Adlung, Y. Cohen, U. Mor, and E. Elinav, "Machine learning in clinical decision making," *Med,* vol. 2, no. 6, pp. 642-665, 2021/06/11/ 2021, doi: https://doi.org/10.1016/j.medj.2021.04.006.

[2] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *Journal of Big Data,* vol. 6, pp. 1-18, 2019.

[3] X. Liu *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The Lancet. Digital health,* vol. 1 6, pp. e271-e297, 2019.

[4] J.-Z. Cheng *et al.*, "Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans," *Scientific Reports,* vol. 6, 2016.

[5] T. Brosch and R. C. Tam, "Manifold Learning of Brain MRIs by Deep Learning," *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention,* vol. 16 Pt 2, pp. 633-40, 2013.

[6] P. Rajpurkar *et al.*, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *ArXiv,* vol. abs/1711.05225, 2017.

[7] W. Cong, Y. Xi, P. Fitzgerald, B. D. Man, and G. Wang, "Virtual Monoenergetic CT Imaging via Deep Learning," *Patterns,* vol. 1, 2020.

[8] S. Soffer *et al.*, "Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis," *Gastrointestinal endoscopy,* 2020.

[9] J. Yang, W. Wang, G. Lin, Q. Li, Y. Sun, and Y. Sun, "Infrared Thermal Imaging-Based Crack Detection Using Deep Learning," *IEEE Access,* vol. 7, pp. 182060-182077, 2019.

[10] I. Maglogiannis and K. Delibasis, "Enhancing classification accuracy utilizing globules and dots features in digital dermoscopy," *Computer methods and programs in biomedicine,* vol. 118 2, pp. 124-33, 2015.

[11] I. Maglogiannis, H. Sarimveis, C. T. Kiranoudis, A. A. Chatziioannou, N. Oikonomou, and 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,* vol. 12, pp. 42-54, 2008.

[12] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Annals of Translational Medicine,* vol. 8, 2020.

[13] C. Han *et al.*, "GAN-based synthetic brain MR image generation," *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018),* pp. 734-738, 2018.

[14] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications,* vol. 31, pp. 1-18, 2019.

[15] P. Lukashevich, B. A. Zalesky, and S. V. Ablameyko, "Medical image registration based on SURF detector," *Pattern Recognition and Image Analysis,* vol. 21, pp. 519-521, 2011.

[16] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. B. Dam, and M. Nielsen, "Deep Feature Learning for Knee Cartilage Segmentation Using a Triplanar Convolutional Neural Network," *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention,* vol. 16 Pt 2, pp. 246-53, 2013.

[17] H. Höfener, A. Homeyer, N. Weiss, J. Molin, C. F. Lundström, and H. K. Hahn, "Deep learning nuclei detection: A simple approach can deliver state-of-the-art results," *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society,* vol. 70, pp. 43-52, 2018.

[18] M. K. Santos, J. R. Ferreira Júnior, D. T. Wada, A. P. M. Tenório, M. H. N. Barbosa, and P. M. A. Marques, "Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in

imaging towards to precision medicine," *Radiologia Brasileira,* vol. 52, pp. 387 - 396, 2019.

[19] K. S. Chan and N. Zary, "Applications and Challenges of Implementing Artificial Intelligence in Medical Education: Integrative Review," *JMIR Medical Education,* vol. 5, 2019.

[20] A. Hekler *et al.*, "Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images," *European journal of cancer,* vol. 118, pp. 91-96, 2019.

[21] J. M. Morrow and M. P. Sormani, "Machine learning outperforms human experts in MRI pattern analysis of muscular dystrophies," *Neurology,* vol. 94, pp. 421 - 422, 2020.

[22] D. Ardila *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Medicine,* vol. 25, pp. 954-961, 2019.

[23] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence,* vol. 1, pp. 206-215, 2019.

[24] A. Kallipolitis, K. Revelos, and I. Maglogiannis, "Ensembling EfficientNets for the Classification and Interpretation of Histopathology Images," *Algorithms,* vol. 14, p. 278, 2021.

[25] "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment | Shaping Europe’s digital future." (accessed Oct. 11, 2022.

[26] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Classification of Histopathological Biopsy Images Using Ensemble of Deep Learning Networks," *ArXiv,* vol. abs/1909.11870, 2019.

[27] I. E. Livieris, A. Kanavos, V. Tampakas, and P. E. Pintelas, "A Weighted Voting Ensemble Self-Labeled Algorithm for the Detection of Lung Abnormalities from X-Rays," *Algorithms,* vol. 12, p. 64, 2019.

[28] D. Kucharski, P. Kleczek, J. Jaworek-Korjakowska, G. Dyduch, and M. Gorgon, "Semi-Supervised Nests of Melanocytes Segmentation

Method Using Convolutional Autoencoders," *Sensors (Basel, Switzerland),* vol. 20, 2020.

[29] M. E. Tschuchnig, G. J. Oostingh, and M. Gadermayr, "Generative Adversarial Networks in Digital Pathology: A Survey on Trends and Future Potential," *Patterns,* vol. 1, 2020.

[30] A. Kallipolitis, A. Stratigos, A. Zarras, and I. G. Maglogiannis, "Fully Connected Visual Words for the Classification of Skin Cancer Confocal Images," in *VISIGRAPP*, 2020.

[31] A. Kallipolitis and I. Maglogiannis, "Creating Visual Vocabularies for The Retrieval And Classification of Histopathology Images," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* pp. 7036-7039, 2019.

[32] A. Goode, B. Gilbert, J. Harkes, D. Jukic, and M. Satyanarayanan, "OpenSlide: A vendor-neutral software foundation for digital pathology," *Journal of Pathology Informatics,* vol. 4, 2013.

[33] M. Kostaki *et al.*, "Trends in epidemiology of melanoma in situ in Greece: data from a melanoma reference centre during the period 2000–2018," *British Journal of Dermatology,* vol. 182, 2019.

[34] V. N. Sehgal, K. Chatterjee, D. Pandhi, and A. Khurana, "Basal cell carcinoma: pathophysiology," *Skinmed,* vol. 12 3, pp. 176-81, 2014.

[35] F. F. Gellrich *et al.*, "Medical treatment of advanced cutaneous squamous-cell carcinoma," *Journal of the European Academy of Dermatology and Venereology,* vol. 33, 2019.

[36] "Cancer Facts and Figures 2008." American Cancer Society. (accessed 3 May, 2018).

[37] "Cancer Facts and Figures 2019." Cancer Society

(accessed 14 January, 2019).

[38] A. S. Farberg and D. S. Rigel, "The Importance of Early Recognition of Skin Cancer," *Dermatologic clinics,* vol. 35 4, pp. xv-xvi, 2017.

[39] "2019 Revision of World Population Prospects <p class="references" style="margin-left:17.7pt;text-indent:-17.7pt;mso-list:

l0 level1 lfo1">." United Nations. (accessed 27 October, 2020).

[40] k. Bakhtiyar *et al.*, "An investigation of the effects of infertility on Women's quality of life: a case-control study," *BMC Women's Health,* vol. 19, 2019.

[41] B. Balaban *et al.*, "The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting," *Human reproduction,* vol. 26 6, pp. 1270-83, 2011.

[42] I. Anagnostopoulos and I. Maglogiannis, "Neural network-based diagnostic and prognostic estimations in breast cancer microscopic instances," *Medical and Biological Engineering and Computing,* vol. 44, pp. 773-784, 2006.

[43] T. Goudas and I. Maglogiannis, "An Advanced Image Analysis Tool for the Quantification and Characterization of Breast Cancer in Microscopy Images," *Journal of Medical Systems,* vol. 39, pp. 1-13, 2015.

[44] S. Alinsaif and J. Lang, "Histological Image Classification using Deep Features and Transfer Learning," *2020 17th Conference on Computer and Robot Vision (CRV),* pp. 101-108, 2020.

[45] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision,* vol. 128, pp. 336-359, 2017.

[46] M. Graziani, I. P. d. Sousa, M. Vellasco, E. C. d. Silva, H. Müller, and V. Andrearczyk, "Sharpening Local Interpretable Model-Agnostic Explanations for Histopathology: Improved Understandability and Reliability," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.

[47] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences,* vol. 116, pp. 22071 - 22080, 2019.

[48] G. LoweDavid, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision,* 2004.

[49] J. C. Caicedo, A. Cruz-Roa, and F. A. González, "Histopathology Image Classification Using Bag of Features and Kernel Functions," in *Conference on Artificial Intelligence in Medicine in Europe*, 2009.

[50] A. Saito *et al.*, "A novel method for morphological pleomorphism and heterogeneity quantitative measurement: Named cell feature level co-occurrence matrix," *Journal of Pathology Informatics,* vol. 7, 2016.

[51] F. Zhang *et al.*, "Dictionary pruning with visual word significance for medical image retrieval," *Neurocomputing,* vol. 177, pp. 75-88, 2016.

[52] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* pp. 3304-3311, 2010.

[53] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2016.

[54] M. Poceviciute, G. Eilertsen, and C. F. Lundström, "Survey of XAI in Digital Pathology," *ArXiv,* vol. abs/2008.06353, 2020.

[55] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation Equivariant CNNs for Digital Pathology," *ArXiv,* vol. abs/1806.03962, 2018.

[56] Y. Huang and A. C. S. Chung, "CELNet: Evidence Localization for Pathology Images using Weakly Supervised Learning," *ArXiv,* vol. abs/1909.07097, 2019.

[57] P. Sabol, P. J. Sinčák, K. Ogawa, and P. Hartono, "Explainable Classifier Supporting Decision-making for Breast Cancer Diagnosis from Histopathological Images," *2019 International Joint Conference on Neural Networks (IJCNN),* pp. 1-8, 2019.

[58] F. C. e. al. "Keras." (accessed.

[59] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," *ArXiv,* vol. abs/1605.08695, 2016.

[60] G. Adam *et al.*, "Deeplearning4j: Distributed, open-source deep learning for Java and Scala on Hadoop and Spark," 2016.

[61] P. Kaur, K. J. Dana, G. O. Cula, and M. C. Mack, "Hybrid deep learning for Reflectance Confocal Microscopy skin images," *2016 23rd International Conference on Pattern Recognition (ICPR),* pp. 1466-1471, 2016.

[62] M. Wodzinski, A. Skalski, A. Witkowski, G. Pellacani, and J. Ludzik, "Convolutional Neural Network Approach to Classify Skin Lesions Using Reflectance Confocal Microscopy," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* pp. 4754-4757, 2019.

[63] M. Combalia *et al.*, "Digitally Stained Confocal Microscopy through Deep Learning," in *MIDL*, 2018.

[64] L. Nanni, S. Brahnam, S. Ghidoni, and A. Lumini, "Bioimage Classification with Handcrafted and Learned Features," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 16, pp. 874-885, 2019.

[65] P. Xie, K. Zuo, Y. Zhang, F. Li, M. Yin, and K. Lu, "Interpretable Classification from Skin Cancer Histology Slides Using Deep Learning: A Retrospective Multicenter Study," *ArXiv,* vol. abs/1904.06156, 2019.

[66] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *CoRR,* vol. abs/1409.1556, 2015.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 770-778, 2016.

[68] A. Kallipolitis, A. Stratigos, A. Zarras, and I. Maglogiannis, "Explainable Fully Connected Visual Words for the Classification of Skin Cancer Confocal Images: Interpreting the influence of visual words in classifying benign vs malignant pattern," *11th Hellenic Conference on Artificial Intelligence,* 2020.

[69] P. Khosravi *et al.*, "Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization," *NPJ Digital Medicine,* vol. 2, 2019.

[70] C. L. Bormann *et al.*, "Performance of a deep learning based neural network in the selection of human blastocysts for implantation," *eLife,* vol. 9, 2020.

[71] D. Tran, S. Cooke, P. J. Illingworth, and D. K. Gardner, "Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer," *Human Reproduction (Oxford, England),* vol. 34, pp. 1011 - 1018, 2019.

[72] Q. Xue and M. C. Chuah, "Explainable deep learning based medical diagnostic system," *Smart Health,* 2019.

[73] M. A. M. Afnan *et al.*, "Interpretable, not black-box, artificial intelligence should be used for embryo selection," *Human Reproduction Open,* vol. 2021, 2021.

[74] T. Lindeberg, "Scale-Space Theory : A Basic Tool for Analysing Structures at Different Scales," *Journal of Applied Statistics,* vol. 21, pp. 225-270, 1994.

[75] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05),* vol. 1, pp. 886-893 vol. 1, 2005.

[76] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-Up Robust Features (SURF)," *Comput. Vis. Image Underst.,* vol. 110, pp. 346-359, 2008.

[77] F. C. Crow, "Summed-area tables for texture mapping," *Proceedings of the 11th annual conference on Computer graphics and interactive techniques,* 1984.

[78] M. Xu, F. Liu, Q. Zhang, and S.-i. Kamata, "Beyond Bag of Features : Adaptive Hilbert Scan Based Tree for Image Retrieval," 2016.

[79] F. Perronnin and C. R. Dance, "Fisher Kernels on Visual Vocabularies for Image Categorization," *2007 IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1-8, 2007.

[80] J. MacQueen, "Some methods for classification and analysis of multivariate observations," 1967.

[81] L. Liu, J. Chen, P. W. Fieguth, G. Zhao, R. Chellappa, and M. Pietikäinen, "A Survey of Recent Advances in Texture Representation," *ArXiv,* vol. abs/1801.10324, 2018.

[82] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 1800-1807, 2017.

[83] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 42, pp. 2011-2023, 2020.

[84] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *ArXiv,* vol. abs/1602.07261, 2016.

[85] D. Stutz, A. Hermans, and B. Leibe, "Superpixels: An evaluation of the state-of-the-art," *Comput. Vis. Image Underst.,* vol. 166, pp. 1-27, 2016.

[86] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. V. Fua, and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 2274-2282, 2012.

[87] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient Graph-Based Image Segmentation," *International Journal of Computer Vision,* vol. 59, pp. 167-181, 2004.

[88] M. Salem, A. Ibrahim, and H. A. Ali, "Automatic quick-shift method for color image segmentation," *2013 8th International Conference on Computer Engineering & Systems (ICCES),* pp. 245-251, 2013.

[89] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *CoRR,* vol. abs/1312.6034, 2014.

[90] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "SmoothGrad: removing noise by adding noise," *ArXiv,* vol. abs/1706.03825, 2017.

[91] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *ECCV*, 2014.

[92] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLoS ONE,* vol. 10, 2015.

[93] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," in *International Conference on Machine Learning*, 2017.

[94] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," *ArXiv,* vol. abs/1703.01365, 2017.

[95] I. Palatnik de Sousa, M. Maria Bernardes Rebuzzi Vellasco, and E. Costa da Silva, "Local Interpretable Model-Agnostic Explanations for Classification of Lymph Node Metastases," *Sensors (Basel, Switzerland),* vol. 19, 2019.

[96] J. Zhou *et al.*, "Whole-genome deep learning analysis reveals causal role of noncoding mutations in autism," *bioRxiv,* 2018.

[97] L. M. Zintgraf, T. Cohen, T. Adel, and M. Welling, "Visualizing Deep Neural Network Decisions: Prediction Difference Analysis," *ArXiv,* vol. abs/1702.04595, 2017.

[98] A. Kaur, "Image Segmentation Using Watershed Transform," 2014.

[99] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences," *ArXiv,* vol. abs/1605.01713, 2016.

[100] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," in *ICML*, 2017.

[101] A. E. Roth and L. S. Shapley, "The Shapley value : essays in honor of Lloyd S. Shapley," *Economica,* vol. 101, p. 123, 1991.

[102] Y. LeCun, C. Cortes, and C. J. C. Burges, "The mnist database of handwritten digits," in *http://yann.lecun.com/exdb/mnist/*, ed, 1999.

[103] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2921-2929, 2016.

[104] X. Shi *et al.*, "Loss-Based Attention for Interpreting Image-Level Prediction of Convolutional Neural Networks," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society,* vol. 30, pp. 1662 - 1675, 2021.

[105] R. L. Draelos and L. Carin, "Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks," 2020.

[106] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV),* pp. 839-847, 2017.

[107] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, "Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs," *ArXiv,* vol. abs/2008.02312, 2020.

[108] S. S. Desai and H. G. Ramaswamy, "Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization," *2020 IEEE Winter Conference on Applications of Computer Vision (WACV),* pp. 972-980, 2020.

[109] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class Activation Map using Principal Components," *2020 International Joint Conference on Neural Networks (IJCNN),* pp. 1-7, 2020.

[110] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "LayerCAM: Exploring Hierarchical Class Activation Maps for Localization," *IEEE Transactions on Image Processing,* vol. 30, pp. 5875-5888, 2021.

[111] S. Rajaraman *et al.*, "A novel stacked generalization of models for improved TB detection in chest radiographs," *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* pp. 718-721, 2018.

[112] Z. Yang, L. Ran, S. Zhang, Y. Xia, and Y. Zhang, "EMS-Net: Ensemble of Multiscale Convolutional Neural Networks for Classification of Breast Cancer Histology Images," *Neurocomputing,* vol. 366, pp. 46-53, 2019.

[113] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable Deep Learning Models in Medical Image Analysis," *Journal of Imaging,* vol. 6, 2020.

[114] H. Jiang, K. Yang, M. Gao, D. Zhang, H. Ma, and W. Qian, "An Interpretable Ensemble Deep Learning Model for Diabetic Retinopathy Disease Classification," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* pp. 2045-2048, 2019.

[115] H. Lee *et al.*, "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets," *Nature Biomedical Engineering,* vol. 3, pp. 173-182, 2018.

[116] Z. Papanastasopoulos *et al.*, "Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI," in *Medical Imaging*, 2020.

[117] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *ArXiv,* vol. abs/1505.04597, 2015.

[118] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," *CoRR,* vol. abs/1511.07122, 2015.

[119] R. M. Rad, P. Saeedi, J. Au, and J. Havelock, "Blastomere Cell Counting and Centroid Localization in Microscopic Images of Human Embryo," *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP),* pp. 1-6, 2018.

[120] M. Y. Harun *et al.*, "Image Segmentation of Zona-Ablated Human Blastocysts," *2019 IEEE 13th International Conference on*

Nano/Molecular Medicine & Engineering (NANOMED), pp. 208-213, 2019.

[121] M. Arsalan, A. Haider, J. Choi, and K. R. Park, "Detecting Blastocyst Components by Artificial Intelligence for Human Embryological Analysis to Improve Success Rate of In Vitro Fertilization," *Journal of Personalized Medicine,* vol. 12, 2022.

[122] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics,* vol. 20, pp. 53-65, 1987.

[123] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods,* vol. 3, pp. 1-27, 1974.

[124] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181),* vol. 2, pp. 645-648 vol.2, 1998.

[125] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "NIH Image to ImageJ: 25 years of image analysis," *Nature Methods,* vol. 9, pp. 671-675, 2012.

[126] K. J. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in *Graphics gems*, 1994.

[127] A. Buades, B. Coll, and J.-M. Morel, "Non-Local Means Denoising," *Image Process. Line,* vol. 1, 2011.

[128] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision,* vol. 115, pp. 211-252, 2015.

[129] E. Zhang and M. Mayo, "Improving Bag-of-Words model with spatial information," *2010 25th International Conference of Image and Vision Computing New Zealand,* pp. 1-8, 2010.

[130] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas, and K. E. Muller, "Contrast-limited adaptive histogram equalization: speed and

effectiveness," *[1990] Proceedings of the First Conference on Visualization in Biomedical Computing,* pp. 337-345, 1990.

[131] P. Krähenbühl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," in *NIPS*, 2011.

[132] F. A. Spanhol, L. Oliveira, C. Petitjean, and L. Heutte, "A Dataset for Breast Cancer Histopathological Image Classification," *IEEE Transactions on Biomedical Engineering,* vol. 63, pp. 1455-1462, 2016.

[133] G. Aresta *et al.*, "BACH: Grand Challenge on Breast Cancer Histology Images," *Medical image analysis,* vol. 56, pp. 122-139, 2018.

[134] G. Aresta *et al.*, "BACH: Grand Challenge on Breast Cancer Histology Images," *Medical image analysis,* vol. 56, pp. 122-139, 2019.

[135] K. Sirinukunwattana *et al.*, "Gland segmentation in colon histology images: The glas challenge contest," *Medical Image Analysis,* vol. 35, pp. 489–502, 2016.

[136] P. Kainz, M. Pfeiffer, and M. Urschler, "Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization," *PeerJ,* vol. 5, 2017.

[137] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[138] A. McCallum, K. Nigam, and L. H. Ungar, "Efficient clustering of high-dimensional data sets with application to reference matching," in *Knowledge Discovery and Data Mining*, 2000.

[139] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *ArXiv,* vol. abs/1905.11946, 2019.

[140] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the Visualization of What a Deep Neural Network Has Learned," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 28, pp. 2660-2673, 2017.

[141] M. Tziomaka, A. Kallipolitis, P. Tsanakas, and I. G. Maglogiannis, "Evaluating Mental Patients Utilizing Video Analysis of Facial Expressions," in *AIAI Workshops*, 2021.

[142] A.-C. Woerl *et al.*, "Deep Learning Predicts Molecular Subtype of Muscle-invasive Bladder Cancer from Conventional Histopathological Slides," *European urology,* 2020.

[143] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for Simplicity: The All Convolutional Net," *CoRR,* vol. abs/1412.6806, 2015.

[144] J. Kubach *et al.*, "Same same but different: A Web-based deep learning application revealed classifying features for the histopathologic distinction of cortical malformations," *Epilepsia,* vol. 61, pp. 421 - 432, 2020.

[145] X. Wang *et al.*, "Decoding and mapping task states of the human brain via deep learning," *Human Brain Mapping,* vol. 41, pp. 1505 - 1519, 2019.

[146] J. Adebayo, J. Gilmer, M. Muelly, I. J. Goodfellow, M. Hardt, and B. Kim, "Sanity Checks for Saliency Maps," in *NeurIPS*, 2018.

[147] K. Zormpas-Petridis, H. Failmezger, S.-e.-A. Raza, I. Roxanis, Y. Jamin, and Y. Yuan, "Superpixel-Based Conditional Random Fields (SuperCRF): Incorporating Global and Local Context for Enhanced Deep Learning in Melanoma Histopathology," *Frontiers in Oncology,* vol. 9, 2019.