Faculdade de Engenharia da Universidade do Porto

Fundamental Topics in Machine Intelligence and their Application to Digital Colposcopy

Kelwin Fernandes

MAPi: Doctoral Programme in Computer Science
Universidade do Minho, Universidade de Aveiro, Universidade do Porto

Supervisor: Jaime S. Cardoso

January 15, 2019

Fundamental Topics in Machine Intelligence and their Application to Digital Colposcopy

Kelwin Fernandes

MAPi: Doctoral Programme in Computer Science
Universidade do Minho, Universidade de Aveiro, Universidade do Porto

January 15, 2019

Resumo

A colposcopia digital é uma tecnologia de baixo custo amplamente utilizada na deteção precoce de cancro cervical e na avaliação forense de violação sexual. Este exame consiste na observação do cólo do útero, vagina e vulva, envolvendo geralmente a aplicação de soluções corantes e filtros de luz para realçar padrões anómalos. A análise manual de coloposcopias digitais é frequentemente subjetiva, sendo a sua eficácia altamente dependente da perícia do especialista. Consequentemente, discrepâncias entre avaliações de vários especialistas resultam num tratamento inadequado dos pacientes com cancro cervical e das vítimas de violação sexual. Neste sentido, apesar de poder ser curado quando detetado numa fase inicial, o cancro cervical continua a ser uma causa de mortalidade significativa nos países em desenvolvimento, onde há escassos recursos materiais e humanos para o seu tratamento. Por outro lado, embora tenham sido descobertos diversos padrões associando lesões genitais a agressão sexual, a sua validade legal é frequentemente questionada.

Deste modo, o desenvolvimento de Sistemas de Apoio à Decisão focados em dados objetivos para a análise de colposcopias digitais torna-se um problema relevante, pois facilitará o processo de decisão e comunicação de resultados procedentes de especialistas com diferentes formações e níveis de especialização. Por se tratar de uma modalidade baseada em imagem com alta variabilidade e de um processo de aquisição não controlado, o desenvolvimento de sistemas deste tipo aplicados a colposcopia requere inequivocamente o recurso a técnicas de Inteligência Computacional, abrangindo a interpretação de imagens através de Visão por Computador e o processo de construção de algoritmos de decisão robustos utilizando Aprendizagem Computacional. Neste trabalho, apresentam-se contribuições de caráter fundamental e aplicado para estas duas áreas, motivadas pela aplicação na colposcopia digital.

A primeira parte desta tese é dedicada a contribuições de caráter fundamental em Inteligência Computacional. Estas contribuições resultaram de discussões com médicos especialistas em colposcopia digital, sendo, no entanto, aplicáveis noutros domínios. Dentre estas, destacam-se: algoritmos de ordenação e a sua aplicação como um paradigma alternativo para resolver problemas nos quais existe desequilíbrio de classes; uma metodologia para transferência de aprendizagem capaz de reproduzir propriedades de alto nível semântico entre modelos; algoritmos de classificação capazes de lidar com dados direcionais e integração de ideias clássicas de visão por computador em algoritmos de aprendizagem profunda, para classificação e segmentação de imagens. Para além da avaliação destas técnicas no domínio da colposcopia digital, foram também conduzidas experiências noutras áreas que reforçam a relevância e eficácia das metodologias propostas. A segunda parte deste trabalho centra-se em contribuições aplicadas especificamente em colposcopia digital e nas suas duas aplicações fundamentais: rastreio de cancro cervical e análise forense de vítimas de violação sexual. Neste contexto, destacam-se: a aquisição e anotação de uma base de dados de colposcopias digitais; o desenvolvimento de um sistema para reconhecimento da

modalidade da colposcopia em vídeos contínuos e não controlados; a análise de risco de pacientes utilizando dados de estilo de vida e histórico clínico; a avaliação da qualidade de colposcopias digitais e um processo totalmente automático para deteção e caracterização de lesões genitais em vítimas de agressão sexual. A relevância das contribuições apresentadas nesta tese foi validada através da realização de diversas experiências.

Abstract

Digital colposcopy is a low-cost technology widely used in the early detection of cervical cancer and in the forensic evaluation of sexual assault. It consists on the observation of the cervix, vagina, and vulva, typically involving application of staining solutions and light filters to highlight abnormal patterns. The manual analysis of digital colposcopies is often subjective, being its effectiveness highly dependent on the expertise of the human evaluator. As a result, inter-expert disagreements and wide range of efficacy derives on improper handling of cervical cancer patients and sexual assault victims. In this sense, despite the possibility of curing cervical cancer when detected on early stages, it remains a significant cause of mortality in low-income countries, where resource availability and trained personal are scarce. Also, while several patterns have been found to correlate genital injuries with sexual assault, the legal validity in courts of these findings is often questioned.

Thus, the development of objective data-driven Decision Support Systems for the analysis of digital colposcopies is a relevant problem, facilitating the decision process and communication of findings of experts with different backgrounds and expertise levels. Being an image-based modality with high variability and open acquisition process, the development of such systems requires unequivocally resorting to Machine Intelligence techniques, from the perception of images using Computer Vision to the process of building robust decision models using Machine Learning. In this work, we propose fundamental and applied contributions to these two areas motivated by digital colposcopy.

The first part of this thesis is devoted to fundamental contributions on Machine Intelligence. These contributions were devised from the discussion with medical experts in the field, aiming to extrapolate the applicability of the proposed methodologies to other areas. Several contributions can be highlighted, including: ranking algorithms and their usage as an alternative paradigm to solve problems such as class imbalance, a transfer learning framework to transfer high-level model properties, classification models that can handle directional data, and the integration of ideas from traditional computer vision and deep learning for image classification and segmentation. Besides the evaluation of these techniques on digital colposcopy tasks, we conduct experiments on other knowledge areas that support the general relevance of the proposed methodologies.

The second part of this work covers applied contributions to digital colposcopy and its two core applications: cervical cancer screening and forensic assessment of sexual assault victims. The main contributions in this direction are: the acquisition and annotation of a digital colposcopy database, the development of a system for the recognition of the colposcopy modality in continuous unconstrained videos, the analysis of patient's risk using behavioral data and medical records, the assessment of the quality of digital colposcopies, and a fully automated end-to-end framework for the detection and characterization of genital lesions in sexual assault victims. By performing several experiments, we validate the relevance of all the contributions presented in this thesis.

Acknowledgments

I would like to start by expressing my most sincere and eternal gratitude to the two main pillars of my formation in this area: Prof. Carolina Chang and Prof. Jaime dos Santos Cardoso. Carolina guided me since the very beginning of my professional career in Venezuela, from symbolic logic to machine learning and computer vision. Carolina inspired me to follow this path and it was worth every mile. Jaime fed my curiosity for science, challenging me with new questions and ideas, and being an outstanding researcher, professor, and role model. You could learn more from an hour talking to him than in an entire college course. That kind of knowledge is scarce, his ability to pass it is unique. Thank you both for your patience and support over the years.

This piece would not be playing if not for the support of the three Universities that support the MAP-i programme (Minho, Aveiro, and Porto). I'm grateful to INESC TEC for all the work conditions and positive environment for research and to *Fundação para a Ciência e a Tecnologia* (Portuguese Foundation for Science and Technology) for the financial support. To all my co-authors, for all the quality work and passion.

I want to thank all my friends, those that I met in Venezuela and those that received me with their arms wide open in Portugal. From Venezuela, I want to thank especially to Aldo, Luis, and Alejandro. I'm proud to call you colleagues; I'm even more proud to call you friends. From Portugal, I want to thank all the VCMI group members; your friendship was priceless upon my arrival. A special thanks to Eduardo, who became one of my closest friends.

I would like to thank my family for all their love and for all the sacrifices they made so I could see further. Your love and devotion are only comparable to your greatness. To my grandparents that taught that opportunities are sometimes an ocean away, that big sacrifices lead to great rewards, and how to love two countries since I was a kid. To my parents, who work tirelessly to succeed in a sea of adversities without stopping smiling.

Finally, I would like to thank my wife, Nohelia. You are my compass and my fuel. For you, all the hard work. To you, all the joy of my life.

Thank you all for being there, thank you with all my heart.

Kelwin Fernandes
Porto, January 15, 2019

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence

ANN Artificial Neural Networks

AUC Area Under the Curve

AW acetowhite

CAD Computer-aided Diagnosis

CE Columnar Epithelium

CEMD Circular Earth's Mover Distance

CIN Cervical Intraepithelial Neoplasia

CLPT Conditional Lexicographic Preference Trees

CNN Convolutional Neural Network

CP Completeness

CR Correctness

CRF Conditional Random Field

CV Computer Vision

DAG Directed Acyclic Graph

DL Deep Learning

dLR directional Logistic Regression

DNN Deep Neural Networks

DP Dynamic Programming

DSS Decision Support System

DT Decision Trees

EMD Earth's Mover Distance

EN Elastic Net

F&H Frank & Hall

GMM Gaussian Mixture Model

GNB Gaussian Naive Bayes

HIV human immunodeficiency virus

HPV Human papillomavirus

HSV Hue-Staturation-Value

HTL Hypothesis Transfer Learning

IR Imbalance Ratio

KNN K-Nearest Neighbors

LBP Local Binary Patterns

LOOCV leave-one-patient-out cross-validation

LR Logistic Regression

LTP Local Ternary Patterns

LxE Lexicographic Ensemble

LxO Lexicographic orders

MAE Mean Absolute Error

MDS Multi Dimensional Scaling

MI Machine Intelligence

MKCLPT Conditional Lexicographic Preference Trees with Multiple Kernel

ML Machine Learning

MLP Multilayer Perceptron

MRF Markov Random Field

NCI National Cancer Institute

nDCG normalized Discounted Cumulative Gain

NIH National Institute of Health

NSR Non-scoring Rankers

OCLPB Over-Complete Local Binary Pattern

PCA Principal Component Analysis

QA Quality Assessment

RBF Radial Basis Function

RF Random Forest

RGB Red-Green-Blue

RNN Recurrent Neural Network

ROC AUC Area Under the Receiver Operating Characteristic curve

ROI Region of Interest

SCJ Squamocolumnar Junction

SE Squamous Epithelium

SpR specular reflections

SRk Scoring Rankers

SVD Singular Value Decomposition

SVM Support Vector Machines

TL Transfer Learning

vMNB von-Mises Naive Bayes

WFA Weighted Finite Automaton

WHO World Health Organization

Chapter 1

Introduction

Cervical cancer remains a significant cause of mortality in low-income countries [147, 174]. It is estimated that over a million women worldwide have cervical cancer [147]. Every year, half a million of new cases are diagnosed, and more than a 250,000 women die in the same period [147]. Just in Portugal, the mortality rate for cervical cancer in 2011 was 3,8 women per 100,000 inhabitants [235]. Despite cervical cancer can often be cured when detected in its early stages [111], the lack of symptoms on the first stages results in carelessness prevention. Moreover, the most vulnerable populations often live in remote regions without access to screening programs.

Digital colposcopy is a low-cost technology involved in the screening and diagnosis of the cervix and the tissues of the vagina and vulva. It is frequently used in the detection of cervical (pre)cancerous lesions [147] and for examination and acquisition of evidence for rape and sexual assault victims [19, 20]. Nowadays, portable and mobile devices have been introduced in the market as an alternative to traditional colposcopes [187, 188, 237], facilitating its scalability and portability to locations with vulnerable populations. Given the high variability of the abnormal patterns observed in digital colposcopy images, the sensitivity of the digital colposcopy varies depending on the human expertise.

Thereby, it is relevant to propose automated Computer-aided Diagnosis (CAD) systems for the decision support of digital colposcopies. Being a medical imaging modality with a partially uncontrolled acquisition and complex decision models, the development of such systems require the use of intelligent systems. Thus, we focus on two sub-areas of Machine Intelligence (MI): Machine Learning (ML) and Computer Vision (CV). In order to facilitate the acceptance by medical teams, the development of automated techniques for the analysis of digital colposcopies raises several challenges to the CAD research community. Properties such as the imbalance distribution of healthy and sick patients, the technical and resource limitations of collecting a vast corpus of medical data, and the reluctance to incorporate uninterpretable decisions on sensitive tasks require rethinking some of the most fundamental ML tasks. Thus, the present work presents fundamental contributions to MI that aim to close the gap between the human decision process and

the automated machine decision process on the analysis of medical data. The first part of this thesis is devoted to this line of work. Part I of this thesis is formed by the following chapters:

- Chapter 2: in this chapter, we present an interpretable model for ranking.
- Chapter 3: the use of ranking as an alternative paradigm to solve classification tasks in unbalanced settings is presented.
- Chapter 4: in this chapter, we propose a new classification setting that aims to facilitate the automation of simple cases. The classification strategy based on ranking is used to tackle this problem.
- Chapter 5: we propose a general framework for transfer learning in this chapter. We illustrate and empirically validated instantiations of the proposed methodology to major problems in machine learning.
- Chapter 6: gynecologists often use directional –periodic– features to describe the cervix region. Thus, we propose a directional classifier that is aware of the periodic nature of the data.
- Chapter 7: in this chapter, we extend Local Binary Patterns, a descriptor typically used to characterize texture information in the cervix, to deep architectures.
- Chapter 8: an alternative approach to tackle image segmentation by quality inference is proposed.

The second part of this work is devoted to applied contributions to the automated analysis of digital colposcopies. We addressed the two core applications of the colposcopy assessment: cervical cancer screening and forensic analysis of sexual assault. Part II covers the following chapters:

- Chapter 9: we do a comparative analysis of the state-of-the-art and reference framework for the main strategies used in the development of CAD systems for digital colposcopies. Also, we acquired, annotated and published a database to validate the main tasks surrounding the automated processing of digital colposcopies.
- Chapter 10: in this chapter, we study the problem of recognizing the colposcopy modality of each frame in a video. Also, we propose a methodology to select relevant excerpts of video for diagnosis.
- Chapter 11: a methodology for image segmentation when the objects hold an ordinal arrangement is proposed and validated in Chapter 11.

- Chapter 12: we instantiate the transfer learning framework proposed in Chapter 5 to the risk estimation of patients in a cervical cancer screening program and to predict the subjective quality of a digital colposcopy.
- Chapter 13: we proposed a framework to assist in the forensic assessment of sexual assault using digital colposcopies. We cover tasks from the modality detection to the detection and characterization of genital injuries.

1.1 Dissemination

1.1.1 Publications

The work developed in the context of this project has been published as listed below.

- **International Journals:**
  - Kelwin Fernandes and Jaime S. Cardoso. Discriminative directional classifiers. *Neurocomputing*, 207:141–149, 2016
  - Kelwin Fernandes and Jaime S. Cardoso. Hypothesis transfer learning based on structural model similarity. *Neural Computing and Applications*, pages 1–14, 2018
  - Kelwin Fernandes, Jaime S. Cardoso, and Birgitte Schmidt Astrup. A deep learning approach for the forensic evaluation of sexual assault. In *Pattern Analysis and Applications*. Springer, 2018
  - Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. Automated methods for the decision support of cervical cancer screening using digital colposcopies. *IEEE Access*, 2018
  - Ricardo Cruz, Kelwin Fernandes, Joaquim F. Pinto Costa, María Pérez-Ortiz, and Jaime S. Cardoso. Binary ranking for ordinal class imbalance. In *Pattern Analysis and Applications*. Springer, 2018
  - Kelwin Fernandes, Davide Chicco, Jaime S. Cardoso, and Jessica Fernandes. Supervised deep learning embeddings for the prediction of cervical cancer diagnosis. *PeerJ Computer Science*, 2018
- **International Conferences:**
  - Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. Temporal segmentation of digital colposcopies. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 262–271. Springer, 2015
  - Kelwin Fernandes, Jaime S. Cardoso, and Hector Palacios. Learning and ensembling lexicographic preference trees with multiple kernels. In *Neural Networks*

(IJCNN), 2016 International Joint Conference on, pages 2140–2147. IEEE, 2016

– Ricardo Cruz, Kelwin Fernandes, Jaime S. Cardoso, and Joaquim F. Pinto Costa. Tackling class imbalance with ranking. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 2182–2187. IEEE, 2016

– María Pérez-Ortiz, Kelwin Fernandes, Ricardo Cruz, Jaime S. Cardoso, Javier Briceño, and César Hervás-Martínez. Fine-to-coarse ranking in ordinal and imbalanced domains: An application to liver transplantation. In *International Work-Conference on Artificial Neural Networks*, pages 525–537. Springer, 2017

– Ricardo Cruz, Kelwin Fernandes, Joaquim F. Pinto Costa, and Jaime S. Cardoso. Constraining Type II Error: Building Intentionally Biased Classifiers. In *International Work-Conference on Artificial Neural Networks*, pages 549–560. Springer, 2017

– Ricardo Cruz, Kelwin Fernandes, Joaquim F. Pinto Costa, María Pérez Ortiz, and Jaime S. Cardoso. Ordinal class imbalance with ranking. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 3–12. Springer, 2017

– Ricardo Cruz, Kelwin Fernandes, Joaquim F. Pinto Costa, María Pérez Ortiz, and Jaime S. Cardoso. Combining ranking with traditional methods for ordinal class imbalance. In *International Work-Conference on Artificial Neural Networks*, pages 538–548. Springer, 2017

– Kelwin Fernandes, Jaime S. Cardoso, and Birgitte Schmidt Astrup. Automated detection and categorization of genital injuries using digital colposcopy. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 251–258. Springer, 2017

– Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. Transfer learning with partial observability applied to cervical cancer screening. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 243–250. Springer, 2017

– Kelwin Fernandes, Ricardo Cruz, and Jaime S. Cardoso. Deep image segmentation by quality inference. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, 2018

– Kelwin Fernandes and Jaime S. Cardoso. Ordinal image segmentation using deep neural networks. In *Neural Networks (IJCNN), 2018 International Joint Conference on*. IEEE, 2018

1.1.2 Collaborations

During the planning and execution of this work, we established several collaborations with medical and technical researchers. Medical partnerships allowed us to set the grounds for the conception of fundamental ideas and the requirements for the execution of applied work. The development of a CAD system for cervical cancer screening was proposed by Dr Jessica Fernandes from *Hospital Universitario de Caracas* and *Universidad Central de Venezuela*, Caracas, Venezuela. Prof. Birgitte Schmidt Astrup from the Institute of Forensic Medicine at the University of Southern Denmark was the medical partner with whom we devised the automated analysis of forensic assault.

Technical collaborations broadened the scope of this project. We worked together with Hector Palacios from *Universitat Pompeu Fabra* in Spain, Maria Perez-Ortiz from *Universidad Loyola Andalucia* in Spain, Davide Chicco from Princess Margaret Cancer Centre and the University of Toronto in Canada, and Ricardo Cruz, Diogo Pernes, Wilson Silva and Ricardo Araujo from our research group at INESC TEC, Portugal.

These collaborations boosted the development and impact of this project, expanding the boundaries of the proposed methodologies to other national and international research teams. The direct outcome of these collaborations can be observed in the publications related to this project.

Part I

Contributions to Fundamental Machine Intelligence

"El que no carga machete
saca la miel con las uñas."

Florentino y el diablo
Alberto Arvelo Torrealba

Context

Biomedical applications are an inexhaustible source of challenges for MI researchers. The high dimensionality and diversity of medical data impose interesting challenges to the community. Medical data range from unstructured corpora from medical records to images and signals from multiple acquisition modalities. The need of ubiquitous and universal access to health-care makes mandatory the inclusion of Decision Support System (DSS) for medical applications in the decision process. However, the assumptions made by the MI experts when modeling these problems, the non-interpretable nature of most decision algorithms, and the intrinsic properties of medical data (i.e. imbalance distribution, privacy, lack of data) arises some difficulty for the acceptance of computational models by the medical community. Thereby, it is relevant to work towards interpretable models for medical applications that resemble the expert's decision process. In the first part of this work, we propose fundamental contributions in MI that attempt to close the gap between computational models and medical teams. Our contributions can be framed in two sub-areas of MI: ML and CV.

Summary of Contributions to Machine Learning

We made contributions in three main areas of ML: ranking (chapter 2) and its application to imbalance classification (chapter 3 and chapter 4), transfer learning (chapter 5), and directional classification (chapter 6). Next, we summarize the motivation and main contribution on each of these branches.

Ranking

Typical predictive tasks within the context of a CAD to support the decision of a physician during a (cancer) screening procedure include: to acquire and to decide if the input data is conclusive [95, 132], to provide a decision about the health state of the patient (with and without a disease) [345], and to suggest the best treatment or procedure [253].

Traditional paradigms to model these problems encompass anomaly detection, classification, regression, and ordinal classification. Anomality detection is usually handled by modeling one of the classes (e.g. healthy patients or patients with a disease) using one-class classifier and generative models [176]. Then, the prediction is done by assigning a score to new observations according to their probability of being a member of that class. While this technique can detect anomalies that were not present in the training data, discarding the data from the reciprocal class is undesirable. Therefore, binary classification settings are the most standard approach, where observations are dichotomized between those with and without the property of interest (e.g. conclusive/non-conclusive data, cancer/non-cancer). Also within classification, problems are often modeled as ordinal classification tasks, by deciding a degree of membership to the property of interest

(e.g. bad/fair/good data, Cervical Intraepithelial Neoplasia (CIN) I/II/III). Classification settings often lead to imbalance distributions, where the large amount of normal observations (e.g. healthy patients) overshadows the scarce cases with the property of interest (e.g. cancer). Consequently, it is common to observe trivial models that always predict the majority class.

In this work, we propose supervised ranking as an alternative to address these tasks. From the extensive set of ranking strategies [212], we focus on pairwise ranking strategies, where observations are ranked by querying by pairs. Pairwise preferences relax the annotation process by ignoring the scale, leading to a higher agreement between experts. We present ranking as an alternative approach to model health-care problems from a ML perspective. We argue that ranking is a general strategy that: 1) is closer to the human decision process than classification/regression approaches, and 2) has a stable learning process that can benefit the learning of traditional models for binary and ordinal classification.

Being interpretability a core trait of ML algorithms for medicine, we propose a class of interpretable ranking models in chapter 2. We proposed an approach to tackle binary imbalance classification tasks with ranking, achieving competitive performance with state-of-the-art techniques used in these scenarios (see chapter 3). We extended the proposed methodology to the ordinal case in [60–62, 264].

Finally, we define the problem of classification by constraining Type II error and show how can we model it using ranking. In this setting, we are interested in classification models that are able to automate as many cases as possible, while keeping a low rate of sensitive errors. To illustrate the problem, let us assume a binary classification setting for cancer detection. It is acceptable for a DSS to assign the cancer status to healthy patients (Type I error) since a human expert can correct this type of error at a later stage of the process. Contrarily, classifying a cancer patient as a healthy one (Type II error) will lead to insufficient attention to patients at risk. Thereby, to promote the acceptance of CAD in the medical community, ML models for medical applications should aim to automate (i.e. remove from the manual evaluation pipeline) as many healthy patients as possible while keeping a low rate of Type II errors. While deciding the right threshold of acceptance for the error rate remains a difficult problem, the human error and resources may drive this decision. Chapter 4 formalizes this problem and provides ranking-based alternatives to solve it.

Transfer Learning

Despite massive access to screening programs, the amount of publicly-available annotated data is often scarce. With a few exceptions [1], medical databases range from a few dozens of observations [3] to a few thousands [151, 167]. The general lack of biomedical data makes difficult the learning of robust models, a problem that is aggravated in complex modalities that involve unstructured data such as text and images.

While data for each problem and acquisition modality is scarce, the wide variety of related biomedical problems that have been tackled using ML drives the need to reuse knowledge acquired from one task to another. The process of reusing knowledge from one task to a new one is known as Transfer Learning (TL). While several works have been proposed in this area, traditional lines of research in this area are not able to transfer knowledge between models and are limited to transfer data among tasks [27, 63, 122, 308]. Hypothesis Transfer Learning (HTL) allows transferring knowledge from one model to another, without revisiting the data [85, 184, 185]. However, the spectrum of models that have been instantiated by current frameworks is narrow.

We formalized a general framework for HTL that allows to transfer high-level structural knowledge between models (see chapter 5). To validate the potential of the framework, we instantiated the framework to several categories of models (i.e. regression, classification, ranking, recommender systems). We show how to transfer knowledge between models of different topologies. Also, we study relevant problems from biomedical applications such as transfer under partial observability. TL under partial observability promotes a high-level transfer of knowledge either from pre-trained models and from human experts. Also, it copes with privacy concerns when working with sensitive data.

Finally, we define an evaluation framework for the assessment of TL techniques that allows a fair assessment of models at different stages of the data acquisition process.

Directional Classification

In different areas of knowledge, phenomena are represented by directional -angular or periodic- data; from wind direction and geographical coordinates to time references like days of the week or months of the calendar. Angular data is often used to reference the human body. Anthropologists use directional coordinates to reference the human skull [298], gynecologists use them to reference the cervix on the colposcopic screening of cervical cancer [369], and forensic doctors use angles to reference rape patterns on female genitalia [16,19,20]. These values are usually represented in a linear scale, and restricted to a given range (e.g. $[0, 2\pi)$), hiding the real nature of this information. Therefore, dealing with directional data requires special methods.

Traditional methods on the design of classifiers for directional variables adopted generative approaches assuming a von Mises distribution of the input data. We proposed discriminative directional classifiers that do not make any assumption on the data distribution. First, we proposed a directional Logistic Regression (dLR) model able to deal with mixed (angular and linear) data.

Summary of Contributions to Computer Vision

The contributions to CV in this work address two problems: image classification (chapter 7) and segmentation (chapter 8). In both cases, we aimed to merge ideas from tradi-

tional and deep learning paradigms in order to reduce the impact of small datasets while allowing access to complex decision spaces induced by deep methodologies.

Deep Local Binary Patterns

As the size of image datasets grow, the area of CV has become an extension of ML by the application of Deep Neural Networks (DNN) than an overlapping area. While in the early years of CV, the main focus was on the development of image processing techniques, robust descriptors, among others, the efforts to solve CV tasks by end-to-end learning of deep models have overshadowed the interest on traditional techniques in the last years.

The history of science, in general, has shown that while an area grows at the expense of the other in some form of unidirectional contribution, in the end, both areas find an equilibrium and discover their niche, joining in a bidirectional – symbiotic – interaction. Such was the case of Computer Science and Mathematics. Now, it is the case of Deep Learning and CV. Namely, deep learning has evolved from the grounds of computer vision techniques such as banks of kernels to detect basic features (e.g. convolutions), pyramid representations and translation invariance (e.g. pooling), illumination invariance (e.g. batch normalization).

Nowadays, research efforts have been devoted to bringing deep learning concepts to traditional methodologies in ML and CV, such was the case of Deep Kernels [48]. Finding a good trade-off between traditional and deep techniques leads to a seamless integration of expert and data-driven knowledge, allowing traditional models to reach new performance boundaries on situations where data is scarce. In this work, we propose a deep extension to Local Binary Patterns (LBP) (see chapter 7), a traditional descriptor in image processing. This contribution attempts to illustrate a path between traditional and deep learning techniques.

Image Segmentation

Traditional methodologies for semantic image segmentation often involve an iterative refinement of the candidate solution by measuring the expected quality of the proposed alternative. That is the case of algorithms such as watershed, region growing, active contours, and level-sets. Conversely, deep learning techniques for image segmentation attempt to segment the image at a single pass through encoder-decoder networks [283]. Despite improvements through cascades of networks, the idea of iterative refinements of solutions is almost lost in deep methodologies.

We propose an alternative deep paradigm of semantic image segmentation (see chapter 8). In the proposed approach, the outcome is achieved by 1) estimating the quality of a segmentation mask given, and by 2) applying local refinements to the input mask to maximize such estimation. In this case, the network is used as an oracle to estimate the

best direction of improvement to manipulate the incremental candidate solution. In chapter 8, we analyze this idea, propose a deep architecture to infer the image-mask quality relation, and use the traditional gradient descent and backpropagation to maximize such estimation.

Chapter 2

Learning and Ensembling Lexicographic Preference Trees with Multiple Kernels

This chapter was published in [97]:

- Kelwin Fernandes, Jaime S. Cardoso, and Hector Palacios. Learning and ensembling lexicographic preference trees with multiple kernels. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 2140–2147. IEEE, 2016

We study the problem of learning lexicographic preferences on multi-attribute domains, and propose *Rankdom Forests* as a compact way to express preferences in learning to rank scenarios. We start generalizing Conditional Lexicographic Preference Trees by introducing multiple kernels in order to handle non-categorical attributes. Then, we define a learning strategy for inferring lexicographic rankers from partial pairwise comparisons between options. Finally, a Lexicographic Ensemble is introduced to handle multiple weak partial rankers, being *Rankdom Forests* one of these ensembles. We tested the performance of the proposed method using several datasets and obtained competitive results when compared with other lexicographic rankers.

2.1 Introduction

Preference learning is an inductive learning task concerned about inferring underlying preference models given partially declared preferences [115]. In Artificial Intelligence (AI), preferences are of critical importance, as they express agent's desires in a declarative fashion [159]. The choices of an agent that acts rationally are driven by an underlying preference model [159]. Therefore, being able to infer models that reflect user's preferences is a key issue of recommending decisions or actions [159, 214].

It is particularly interesting to learn ranks in combinatorial domains, where the options to rank are represented by a set of variables assigned to values. They are also studied in combinatorial preference aggregation [191], a well-studied problem in computational social choice, an area dealing with the computational aspects of social choice problems like voting. The main challenge is to have a language that allows compact representation of a number of options which are exponential in the number of variables. Any chosen language would have pros and cons. More compact languages are usually less expressive and *vice versa*, and there are also complexity issues [35].

Learning to Rank in combinatorial domains [290] has become a trendy topic in recent years due to the growing number of applications involving the prediction of structured preference data instead of traditional classification and regression tasks. For example, search engines [295] receive a query and return a set of objects. Rankings are used to present such objects according to their relevance. Good rankings can be obtained from the query and the previous interaction of the user with the engine's results, that can be translated into their preferences, characterized by a number of attributes. Recommender systems and subjective image quality assessment can benefit as well from learnt ranks [117, 305].

Learning to Rank methods can be divided, according to their input representation, in three types: pointwise, pairwise and listwise [212]. Pointwise methods rely on assigning a numerical (e.g. [0, 1]) or ordinal (e.g. *poor, fair, good* or *excellent*) score for each observation. Pointwise methods reduce the learning to rank problem to traditional regression[56], classification [200], and ordinal regression/classification [57] tasks. In contrast, Pairwise models rely on comparing pair of observations [34, 109, 149, 290]. Finally, the listwise paradigm deals directly with the ordering of the entire set of entities associated to a given query [340]. All these three approaches present advantages and disadvantages, both regarding the accuracy and workload associated with the training set acquisition, and regarding their adequacy to different scenarios.

The scope of this work is limited to the area of pairwise rankers, which can be further subdivided into Scoring Rankers (SRk) and Non-scoring Rankers (NSR). Assuming that the learned ranker can be represented by a binary function $f(a, b)$ that decides the relative order between observations a and b, a SRk decides which observation is better by comparing the score assigned by a pointwise (unary) function s with monotonous outcome, $f(a, b) = s(a) > s(b)$. Indeed, SRk project the options into a linear space. Examples of these models are GBRank [371] and RankSVM [149]. Pointwise rankers and pairwise SRk assume that the underlying preference model can be understood as a utility function. However, whilst it is always possible to transform a utility function into a preference model, it is not the case in the opposite direction [277]. In contrast, NSR cannot be directly obtained from pointwise scoring functions, but instead use the two observations together, enabling more expressive non-linearity.

Lexicographic orders (LxO) compactly express the order between any pair of options

by specifying a particular order of the variables and of their respective values. LxO are widely accepted as a plausible representation of humans preferences [36, 292]. However, their adequacy depends on the assumption of having underlying lexicographic preferences (i.e. learning bias).

Several types of LxO have been proposed in the last decade. Linear LxO, and their learning strategies, have been studied in [109, 292]. Booth et al. [34] introduced Conditional Lexicographic Preference Trees (CLPT). CLPT are an extension of LxO to express conditional local preferences, where the preference value of an attribute depends on the values of previous attributes, and conditional importance, where the ordering of the attributes depends on the values of previous attributes. Liu and Truszczynski [215] extended CLPT to allow partial rankings, focusing on formal properties of the language. CLPT can be understood as general decision trees, where linear LxO is a particular case representable as decision lists. Bräuning and Hüllermeier studied CLPT from a ML perspective and demonstrated their adequacy in different problems [36].

The contributions of this work can be summarized as follows. First, we extend CLPT [34] to handle non-categorical attributes, like real-valued or structured, through learning multiple kernels. Second, we propose a learning strategy to induce CLPT encodings with multiple kernels from a partial pairwise preference set. Third, we propose a lexicographic ensemble strategy to aggregate (conditional) lexicographic rankers, called *Rankdom Forest*. Finally, we validate the advantages of the proposed model from a ML point of view using several datasets.

2.2 Languages for Representing Multi-attribute Lexicographic Preferences

In this section, we formalize three languages for encoding lexicographic preference models. Section 2.2.1 and 2.2.2 define linear and non-linear (i.e. conditional) lexicographic languages. Then, the proposed language is presented in Section 2.2.3.

2.2.1 Lexicographic Orders

As stated in [36], LxO can be defined as follows. Let us define an option $o \in \mathcal{O} = \mathcal{D}(A_1) \times \ldots \times \mathcal{D}(A_n)$, where $A = \{A_1, \ldots, A_n\}$ is the set of attributes and $\mathcal{D}(A_i)$ is the domain of the corresponding attribute A_i. For a given attribute subset $A' = \{A_{i_1}, \ldots, A_{i_k}\} \subseteq A$, $\mathcal{D}(A')$ is defined as $\mathcal{D}(A_{i_1}) \times \ldots \times \mathcal{D}(A_{i_k})$. Also, for an option $o \in \mathcal{O}$ and a subset $A' \subseteq A$, $\pi_{A'}[o]$ denotes the projection of o from $\mathcal{O}$ to $\mathcal{D}(A')$. For the sake of readability, we write o_k instead of $\pi_{\{A_k\}}[o]$ if $A' = \{A_k\}$ is a single attribute.

An LxO on $\mathcal{O}$ is a total order $\succ$ defined in terms of [36]:

- The *attribute importance*, defined as the total order $\sqsupset$ on A.

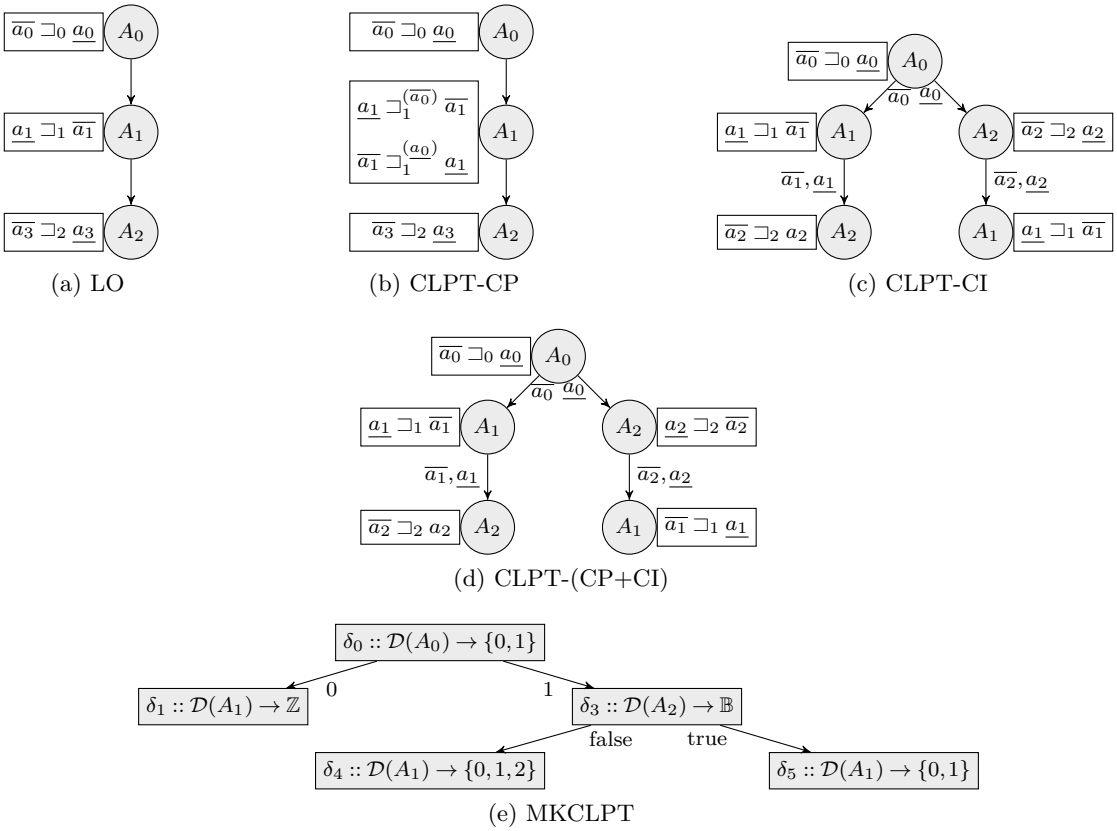


Figure 2.1: Examples of LxO, CLPT with conditional preferences (CP), CLPT with conditional attribute importance (CI), CLPT with CP and CI, and MKCLPT encodings. Attributes (A_i) are assumed to be binary, $\mathcal{D}(A_i) = \{\overline{a_i}, \underline{a_i}\}$.

- The *preferences on attribute values*, defined as a total order $\sqsupset_i$ on each attribute domain $\mathcal{D}(A_i)$.

Figure 2.1a illustrates a LxO with three binary attributes. Hereafter, unless we said otherwise, we assume without loss of generality that $A_1 \sqsupset A_2 \sqsupset \ldots \sqsupset A_n$. Formally, o^* is preferred to o, $o^* \succ o$, if and only if there exists a $k \in \{1, \ldots, n\}$ such that

$$(o_k^* \sqsupset_k o_k) \ \wedge (\forall i | i \in 1 \leq i < k : o_i^* = o_i) \tag{2.1}$$

It is important to notice that we refer to each object of the universe $\mathcal{O}$ as an option to recall the idea of being the preferred alternative in given set (pairs in our case). In other contexts, we might think about them as states or observations.

2.2.2 Conditional Lexicographic Preference Trees

CLPT extend linear LxO by allowing to express conditional preferences on attribute values and conditional attribute importance based on the observed values of previous attributes [34].

2.2.2.1 Conditional Preferences on Attribute Values

Formally, conditional preferences introduce the notion of attribute dependency by refining the preference between attribute values $\sqsupset_k$ into $\sqsupset_k^{(o_1,\ldots,o_{k-1})}$, so it depends on $(o_1,\ldots,o_{k-1})$, the assignments of the attributes more important than k [36].

Then, for conditional preferences on attribute values, $o^* \succ o$ if and only if there exists a $k \in \{1,\ldots,n\}$ such that

$$\left(o_k^* \sqsupset_k^{(o_1,\ldots,o_{k-1})} o_k\right) \wedge (\forall i | 1 \leq i < k : {o_i}^* = o_i) \tag{2.2}$$

Figure 2.1b illustrates how the observed value on A_0 changes the preferred value on A_1. This concept was originally introduced in [34] to handle preferences on single attributes and extended in [36] to consider attribute tuples.

2.2.2.2 Conditional Attribute Importance

Observed values on important attributes define the order in which subsequent attributes are evaluated (their position in the hierarchy) [34]. Formally, the partial assignment $o_{i_1} \times \ldots \times o_{i_k}$ of the most important characteristics $(A_{i_1},\ldots A_{i_k}) \in A^k$ determines the next attribute in the hierarchy $A_{i_{k+1}} \in A \setminus (A_{i_1},\ldots A_{i_k})$. Figure 2.1c illustrates how the observed value on A_0 changes the relative importance of A_1 and A_2. Namely, if the observed value on A_0 is $\overline{a_0}$, A_1 is evaluated before A_2. Otherwise, A_2 is more important than A_1.

CLPT can be graphically understood as a tree, where every node is labeled with an attribute A_k and a total ordering $\sqsupset_k$ of its domain $\mathcal{D}(A_k)$. Each node has an outgoing edge to a descendant node for each possible value $a_k \in \mathcal{D}(A_k)$. A CLPT can be used as a ranker by feeding an option pair $(o^*, o) \in \mathcal{O} \times \mathcal{O}$ to the root of the tree and propagating it through the structure. At a given node v with attribute A_k, the projections o_k^* and o_k are compared. If their scores are different (i.e. $o_k^* \neq o_k$), $\sqsupset_k^{(o_{v_1},\ldots,o_{v_{k-1}})}$ is used to decide the preferred option. Otherwise, the pair is propagated to the proper descendant through the edge o_k. If a pair is propagated to a leaf without a decision, the model rejects the pair (i.e. it decides not to answer). Since the model is allowed to abstain, the preference order induced by it might be partial [46, 215]. Figures 2.1b and 2.1c show two examples of CLPT with conditional preferences on attribute values and conditional attribute importance respectively. Figure 2.1d illustrates an example with combined importance in both, preferences and importance.

2.2.3 Conditional Lexicographic Preference Trees with Multiple Kernels

CLPT have been studied in the literature[34, 36, 215]: 1) assuming discrete attributes and 2) assuming that preferences behave in a strictly lexicographic manner, which might

impose a very restrictive bias. Thereby, we extend the CLPT proposed in [34] to: 1) handle infinite (countable and uncountable) value domains and 2) to reduce the lexicographic bias $A_i \sqsupset A_k$. In order to do this, we introduce the notion of a kernel δ,

$$\delta :: \mathcal{D}(A') \rightarrow \mathbb{T}$$

where $A' \subseteq A$ and $\mathbb{T}$ is a totally ordered set under the relation $\leq$. The kernel function can be considered as a weak SRk, that takes as input the projection of o over A', $\pi_{A'}[o]$, and computes a score $\delta(\pi_{A'}[o])$. For readability, we assume that A' is known and write instead $\delta(o)$.

The language of Conditional Lexicographic Preference Trees with Multiple Kernel (MKCLPT) extends the language of CLPT by introducing a customized kernel to each node. Thereby, instead of defining local preferences (see (2.2)) on attribute values, we consider a linear order of the assigned local scores. In the same way, conditional importance (c.f. Section 2.2.2.2) is defined in terms of the assigned scores of previous attributes instead of the original attribute values. Consequently, MKCLPT defines a parametric language to encode pseudo-lexicographic preferences.

Kernel expressiveness may range from naïve functions that consider a total order of a single attribute domain (traditional CLPT) to very complex multi-attribute SRk (e.g. RankSVM, GBRank) able to learn any ranking function. Although we allow kernels to return an infinite number of values, we encourage the use of kernels with finite ordinal range. The proper trade-off between kernel expressiveness and lexicographic prior depends on the underlying preference model. In this sense, expressive kernels induce shallow tree structures, reducing the lexicographic assumption on the original attribute space.

As done with CLPT, a pairwise ranker can be inferred from a given MKCLPT. The only difference relies on the use of the scores assigned by the node's kernel δ_v when comparing two options instead of the projection on the node's attribute (see Figure 2.1e).

2.3 Learning CLPT with Multiple Kernels

There have been several attempts in the past to learn lexicographic models from partial pairwise annotations. In this sense, the training set $\mathcal{T}$ consists of object pairs $(o^*, o) \in \mathcal{O} \times \mathcal{O}$, being o^* preferred to o.

$$\mathcal{T} = \{(o^*,\ o) \mid o^* \in \mathcal{O} \wedge o \in \mathcal{O} \wedge o^* \succ o\} \subseteq \mathcal{O} \times \mathcal{O}$$

The learning goal is to find lexicographic models with high agreement with the pairwise preferences in $\mathcal{T}$, avoiding overfitting the training data.

Each aforementioned language defines a space of preference models with its proper expressiveness. From an ML perspective, a predictive model is a particular instantiation from the entire search space defined by the language, which is able to generalize the underlying

```
1: procedure FIT-MKCLPT(𝒯, A, k, δ)
2:     if 𝒯 = ∅ ∨ A = ∅ then
3:         return MKCLPT_Dummy()
4:     end if
5:
6:     A* ← Select-k-Features(A, k)
7:     δ_n ← Fit-Kernel(δ, 𝒯, A*)
8:     A_c ← Consumed(δ_n)
9:     children ← {}
10:
11:    for each s in Range(δ_n) do
12:        𝒯_s ← {(o*, o) | (o*, o) ∈ 𝒯 ∧ δ_n(o*) = δ_n(o) = s}
13:        children[s] ← Fit-MKCLPT(𝒯_s, A − A_c, k, δ)
14:    end for
15:
16:    return MKCLPT_Node(δ_n, children)
17: end procedure
```

Figure 2.2: Pseudocode of the proposed algorithm for fitting MKCLPT

preference model. Namely, the preferences induced by the model are highly consistent with the unknown preferences. The resulting model is a computationally tractable way to represent the space of all the possible combinatorial preferences. While from a ML point of view, we are interested in compact models able to generalize, from a social choice perspective, this can be used for combinatorial preference aggregation [35].

For instance, LexRank presents a greedy strategy to learn linear LxO by learning unconditional attribute importance and preferences on attribute values for Boolean domains [109]. On the other hand, CLeRa [36] is a greedy algorithm to induce CLPT for options with discrete domains.

Given that reasoning directly in a lexicographic manner with infinite-valued features returns a ranker with zero probability of non-abstention, leaving the entire decision to the most important attribute, previous efforts on learning CLPT preprocess real-valued features using global discretization techniques (e.g. equal frequency binning) [36]. However, these discretization techniques behave in a global fashion, preventing the model to decide the best discretization technique to handle local preferences.

We propose an extension of the CLeRa algorithm to induce MKCLPT from a training set $\mathcal{T}$ (see Figure 2.2). In our proposal, kernels are learned in a local manner, instead of traditional global discretization schemes learned as a preprocessing step. To increase coherence between our extension and the original CLeRa algorithm, this section follows of [36].

As usually done in Decision Trees (DT) learning, our approach presents a greedy top-down strategy. Thereby, similarly to the information gain measure used in DT, we use a performance estimator that can be used as a heuristic for selecting the best topology

[36, 46].

Thus, (2.3)-(2.4) measure respectively the number of concordant and discordant pairs between the induced preference model and the training set.

$$C(\succ, \mathcal{T}) = |\{(o^*, o) \in \mathcal{T} | o^* \succ o\}| \tag{2.3}$$

$$D(\succ, \mathcal{T}) = |\{(o^*, o) \in \mathcal{T} | o \succ o^*\}| \tag{2.4}$$

Then, motivated by [46], (2.5)-(2.6) define Correctness (CR) and Completeness (CP), that express respectively the agreement degree between two preference models and the degree of abstention.

$$CR(\succ, \mathcal{T}) = \frac{C(\succ, \mathcal{T}) - D(\succ, \mathcal{T})}{C(\succ, \mathcal{T}) + D(\succ, \mathcal{T})} \tag{2.5}$$

$$CP(\succ, \mathcal{T}) = \frac{C(\succ, \mathcal{T}) + D(\succ, \mathcal{T})}{|\mathcal{T}|} \tag{2.6}$$

Finally, we introduce the following performance measure, $0 \leq \mathcal{C}_{\mathcal{RP}} \leq 1$, motivated by the area under the correctness-completeness curve (see (2.7)). Since correctness range from -1 to +1, we introduce a scale transformation in order to be consistent with traditional Area Under the Curve (AUC) interpretations.

$$\mathcal{C}_{\mathcal{RP}}(\succ, \mathcal{T}) = \frac{(CR(\succ, \mathcal{T}) + 1)}{2} \cdot CP(\succ, \mathcal{T}) \tag{2.7}$$

2.3.1 Initialization and Termination

We begin considering the entire training set $\mathcal{T}$ and the complete set of attributes A. The recursion stops when there are no remaining attributes ($A' = \emptyset$) or training pairs ($\mathcal{T}' = \emptyset$). In this case, the function returns a dummy node that always rejects (i.e. full abstention).

2.3.2 Creating a node

In [36], a node is created by enumerating the exhaustive search on the attribute tuples and total orders of attribute values (with some heuristics to reduce computational time). The best node is chosen by maximizing correctness, using completeness to break ties. We generalize this process by learning a kernel that creates a latent attribute that encodes local preferences. At any given time, the kernel function has access to the remaining (non-used) attributes. By using kernels able to handle non-categorical data types, we provide to CLPT nodes a way to reason about complex attributes. Moreover, we may introduce combinations of attributes in a more compact way than the total order of tuples proposed in [36] by using kernels that consider multiple attributes.

The kernel is assumed to be a weak SRk. Thereby, we might use the scores directly in the lexicographic comparison without learning a total order on these values.

Although the decision process regarding the best kernel configuration depends on the kernel's learning strategy, maximizing correctness might encourage overfitted kernels that reject most comparisons when the kernel has high expressiveness. Thereby, in the evaluation of kernels we maximize $\mathcal{C}_{\mathcal{RP}}$, using correctness and completeness to break ties (in that order).

2.3.3 Recursion

After the best kernel has been identified and learned, the training pairs (o^*, o) predicted by the current node are removed from the training set. Then, a child is created for each possible score s in the image of δ. Each child is recursively trained using the training set defined in (2.8) and the attributes $A' \setminus A_\delta$, where A_δ are the attributes consumed by the kernel function.

$$\mathcal{T}_s = \{(o^*, o) \mid (o^*, o) \in \mathcal{T} \ \wedge \ \delta(o^*) = \delta(o) = s\} \tag{2.8}$$

2.3.4 Randomization

In order to introduce variability, we apply the random subspace method in the selection of attributes (features) to be considered by the kernel. This method has been traditionally used in the construction of Random Forests [153]. Thereby, each kernel is trained using $k < |A'|$ randomly chosen features. During the feature selection process, a fitness value is assigned to each attribute by fitting a kernel using it individually. Then, the roulette wheel selection strategy is employed to decide the best subset of features. The probability of a given features to be chosen is defined by

$$p_i = \frac{\mathcal{C}_{\mathcal{RP}}(\succ_{\delta i}, \mathcal{T})}{\sum\limits_{a \in A'} \mathcal{C}_{\mathcal{RP}}(\succ_{\delta a}, \mathcal{T})} \tag{2.9}$$

where $\succ_{\delta a}$ is the preference model induced by fitting a single node using the remaining training pairs and the feature a.

2.4 Lexicographic Ensemble

Given the discrete nature of the comparisons performed by lexicographic preference models, coarse rankings (i.e. partial orders) are produced by them when compared with orders induced by SRk with continuous output. Therefore, it is relevant to study a way to aggregate multiple NSR.

While simple voting schemes (e.g. weighted votes) can be considered in general classification, preserving consistency (i.e. symmetry and transitivity) when combining multiple NSR is not trivial, even if each individual ranker is consistent [46]. Formally, if n rankers r_i with induced preference relations $\succ_i$ are merged, the ensemble R formed by them with

preference relation $\succ$ must preserve (2.10) and (2.11). Namely, the underlying preference models must induce a Directed Acyclic Graph (DAG).

$$a \succ b \equiv b \prec a \tag{2.10}$$

$$a_0 \succ a_1 \ \wedge \ \ldots \ \wedge a_{k-1} \succ a_k \implies a_0 \succ a_k \tag{2.11}$$

A previous attempt on merging lexicographic preference models was proposed in [189] from a computational social choice perspective. Thus, several constraints over the voting scheme were introduced that, in general, turn the problem intractable. However, from an ML point of view, we are interested in building an aggregation, without concerning about social choice properties.

A Lexicographic Ensemble (LxE) comprises a sequence of (weak) lexicographic rankers $R = \{r_i \mid i \in [0 \ldots n)\}$ with a hierarchic voting rule. Namely, the predicted outcome for the pair (a, b) is specified by the estimator (r_i) with highest priority that doesn't abstain $(\neg a \sim_i b)$. Formally, the pair (a, b) is predicted as $\oplus \in \{\prec, \succ\}$ if and only if (2.12) holds, otherwise, $a \sim b$.

$$a \oplus b \equiv \underset{i \in [0 \ldots n)}{\exists} \left(a \oplus_i b \wedge \left(\underset{j \in [0 \ldots i)}{\forall} a \sim_j b \right) \right) \tag{2.12}$$

The proposed voting scheme can be understood as recursively appending a copy of the estimators with lower priority to the leaves (tail for linear LxO) of the estimators with higher priority.

The learning process is reduced to find the best linear arrangement (i.e. permutation) of the base estimators. Although this scheme presents a *dictatorship*, from a ML perspective, this is acceptable since estimators with higher priority are better informed than estimators in the tail of the arrangement. This process can be done by enumerating all the possible permutations and choosing the one that maximizes the metric of interest (e.g. correctness, completeness, $\mathcal{C}_{\mathcal{RP}}$). Since this becomes rapidly intractable for large ensembles, two strategies are proposed to address this task.

The first idea relies on sorting the estimators in decreasing order by performance in a training (or validation) set using either correctness (with completeness to break ties), $\mathcal{C}_{\mathcal{RP}}$ (with correctness and completeness to break ties) or Borda count (number of wins minus number of losses) as evaluation metric. Given that the base estimators are weak and may abstain, only pairs of non-rejected predictions are considered when computing the winner between a pair of rankers in the Borda count strategy.

The second idea consists in modeling the problem as learning an unconditional lexicographic order with fixed preferences (correct prediction over incorrect prediction) of the estimators. Thus, any lexicographic learning strategy like LexRank [109] can be used to optimize the final order.

Table 2.1: Datasets used in the experimental evaluation

Dataset	Options	Features	Pairs
Lenses	24	4	155
Hepatitis	155	19	3,936
Echocardiogram	131	9	5,418
Parkinson [209]	195	22	7,056
SPECT Heart	267	22	17,270
Ionosphere	351	34	28,350
Ecoli	336	7	37,831
Arrhythmia	451	14	68,990
Blood Transfusion [355]	748	4	101,460
Breast Cancer Wisconsin [30]	683	9	106,116
Mammographic Mass [82]	830	5	172,081
Tic-Tac-Toe	958	9	207,832
Banknote Authentication	1372	4	464,820
Car Evaluation	1728	6	682,721
Wine Quality (Red) [54]	1599	11	821,581
Chess	3196	36	2,548,563

2.5 Rankdom Forest

A Rankdom Forest is the aggregation using a Lexicographic Ensemble of weak Conditional Lexicographic Preference Trees. As done by traditional Random Forests [153], our proposal uses feature randomization to introduce variability in the model construction (c.f. Section 2.3.4).

Also, in order to induce weak rankers, the maximum depth for each base estimator may be limited. This parameter configuration can be optimized through cross-validation. Encouraging shallow trees increases the probability of abstention on the top of the ensemble (i.e. those with the highest priority), enabling estimators in the tail of the ensemble to contribute to the construction of preference relation with higher completeness.

2.6 Experiments and Results

In this section we detail the experimental evaluation of the proposed method to induce MKCLPT against two state-of-the-art methods to induce lexicographic rankers: CLeRa [36] algorithm to induce CLPT and LexRank [109] to induce linear LxO. These three methods were assessed using 16 real-life datasets from the UCI [207] repository (c.f. Table 2.1).

We used a 10-fold cross validation assessment strategy on the space of options. In this sense, we consider for the training stage preferences whose two options belong to the training set. Then, we validate the results on two different test sets. The first test set, hereinafter referred as train-test (TR-TS) contains preference pairs with one option in the training test and one in the test set. The second test set, referred as test-test (TS-TS)

contains pairs with both options in the test set. TR and TS sets are important in different contexts and applications. While the TR-TS set measures the performance of the models when comparing new (unseen) options and old (with partial annotations) options, the TS-TS set measures their performance on purely new options. Moreover, all the experiments were repeated 10 times (100 executions in total). To facilitate the reproducibility of the proposed experiments, we have made available the source code, training/test partitions and extended results.

2.6.1 Kernels Used in the MKCLPT Learning Process

For experimental evaluation, we considered the following kernels.

2.6.1.1 Binning

a given feature is discretized using equal-frequency binning. The optimal number of bins ($n \in \{2,\ldots,6\}$) and the feature to be used in the binning are decided at learning time. Then, a score is assigned to each bin according to its quality measured using the Borda count.

2.6.1.2 Clustering

observations are clustered using K-means with the number of centroids ($k \in \{2,3,4\}$) decided at learning time by maximizing $\mathcal{C}_{\mathcal{RP}}$. Centroids are indexed by Borda count. Then, each observation is assigned to its closest centroid.

2.6.1.3 RankSVM

an SRk is trained. Then, scores are discretized using equal-frequency binning (5 bins). Given that the scores assigned by an SRk are already monotonous, the final bins preserve the preference order. Notice that RankSVM [149] can be considered a SRk given that the decision rule $\omega^T(a-b) > 0$ can be transformed into a scoring function since $\omega^T(a-b) > 0 \equiv \omega^T a > \omega^T b \equiv s(a) > s(b)$.

2.6.1.4 Multiple Kernel

in order to merge multiple kernels, a meta-kernel that simultaneously fits a sequence of base kernels and decides the best transformation based on $\mathcal{C}_{\mathcal{RP}}$ is used.

2.6.2 Assessment of the Individual Languages and Learning Strategies

Tables 2.2 and 2.3 show a pairwise comparison of the three methods. MKCLPT obtained better correctness results than CLPT and LxO in most of the datasets. This behavior was consistent in both test sets. Although the proposed method abstains in a higher degree than the others state-of-the-art methods, the improvement in terms of correctness

Table 2.2: Counts of wins for each pair of models in terms of CR and CP. The numbers indicate the number of times the model in the row obtained a better performance than the model in the column.

Train-Test						
Model	**Correctness**			**Completeness**		
	MKCLPT	CLPT	LO	MKCLPT	CLPT	LO
MKCLPT	-	**14**	**15**	-	4	2
CLPT	2	-	**9**	**12**	-	0
LO	1	7	-	**14**	**16**	-
Test-Test						
MKCLPT	-	**13**	**12**	-	4	2
CLPT	3	-	4	**12**	-	0
LO	4	**12**	-	**13**	**16**	-

Table 2.3: Average relative difference (%) for each pair of models in terms of CR and CP. The models in the column are used as reference to measure the gain of the models in the rows.

Train-Test						
Model	**Correctness**			**Completeness**		
	MKCLPT	CLPT	LO	MKCLPT	CLPT	LO
MKCLPT	–	**14.75**	**14.10**	–	−0.82	−1.53
CLPT	−10.24	–	−0.41	**0.90**	–	−0.72
LO	−9.70	**0.84**	–	**1.63**	**0.73**	–
Test-Test						
MKCLPT	–	**15.95**	**11.39**	–	−0.83	−1.97
CLPT	−10.28	–	−3.71	**0.92**	–	−1.14
LO	−6.79	**4.21**	–	**2.11**	**1.18**	–

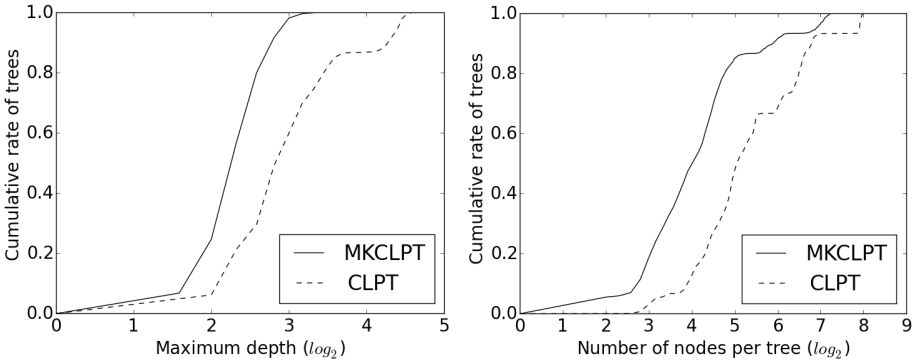
Table 2.4: Average performance in terms of CR and CP

Model	**Train-Test**		**Test-Test**	
	CR	**CP**	**CR**	**CP**
MKCLPT	**0.8165**	0.9773	**0.7764**	0.9730
CLPT	0.7268	0.9859	0.6880	0.9817
LO	0.7309	**0.9931**	0.7114	**0.9931**

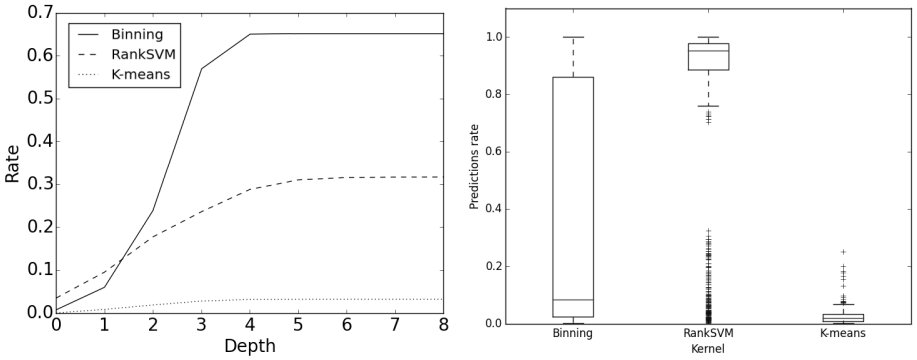
is higher than the decrease in terms of completeness. Table 2.4 shows the average absolute correctness and completeness for each method and test set.

2.6.3 Topology Comparison of the MKCLPT and CLPT Models

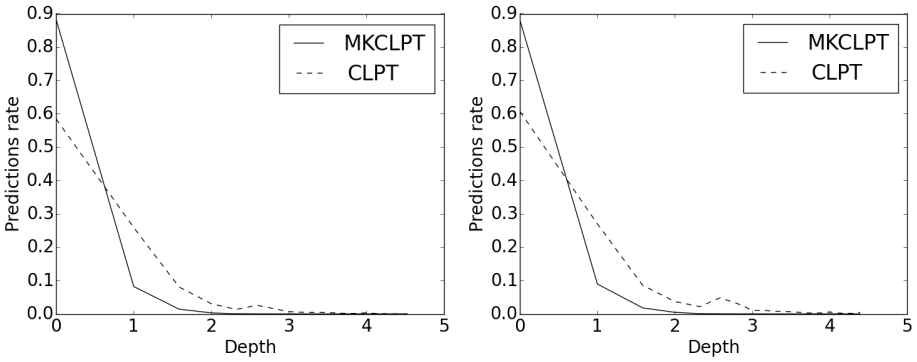
As shown in Figure 2.3a, the proposed MKCLPT language and its training algorithm induces more compact topologies than its non-kernelized counterpart (CLPT and CLeRa). This fact can be observed by the fact that the curve generated by MKCLPT completely dominates the curve induced by CLPT.



(a) Cumulative maximum depth and number of nodes per tree.



(b) Cumulative number of kernels per depth and prediction rate per kernel.



(c) Predictions done by depth. **Left:** Train-test set, **Right:** Test-test set.

Figure 2.3: Topology analysis

Also, the learning strategy detailed in Section 2.3 tends to locate expressive kernels (e.g. RankSVM) near the root of the trees and a higher number of kernels with limited expressiveness (e.g. binning) in the leaves. Therefore, given that the number of nodes increases exponentially on the height of the tree, it is natural to observe a higher number of nodes using Binning than RankSVM (see Figure 2.3b). However, given that MKCLPT follows a Pareto rule, a vast majority of the predictions are performed by the underrepresented RankSVM kernel (see Figures 2.3b and 2.3c). In this sense, MKCLPT is a NSR able to reduce the lexicographic bias of traditional CLPT by inducing shallower topologies

Table 2.5: Average performance of the ensembles in terms of correctness and completeness.

Ensemble	Train-Test		Test-Test	
	CR	**CP**	**CR**	**CP**
$\mathcal{C}_{\mathcal{RP}}$	0.7862	0.9958	0.7424	0.9956
Borda	0.7881	0.9958	0.7437	0.9956
CR	**0.7883**	0.9958	0.7442	0.9956
LexRank	0.7878	0.9958	**0.7459**	0.9956

Table 2.6: Counts of wins and ties for each ensemble models in terms of correctness and completeness.

Train-Test							
Ensemble	**Correctness**				**Completeness**		
	MKCLPT	CLPT	LO	LE	MKCLPT	CLPT	LO
$\mathcal{C}_{\mathcal{RP}}$	4	**16**	**14**	5	**16**	**12**	3
Borda	5	**16**	**15**	4	**16**	**12**	3
CR	5	**16**	**15**	7	**16**	**12**	3
LexRank	5	**16**	**15**	1	**16**	**12**	3
Test-Test							
$\mathcal{C}_{\mathcal{RP}}$	4	**12**	**11**	6	**16**	**12**	3
Borda	5	**14**	**10**	5	**16**	**12**	3
CR	6	**14**	**10**	7	**16**	**12**	3
LexRank	5	**14**	**11**	5	**16**	**12**	3

with expressive kernels near the root.

2.6.4 Assessment of the Lexicographic Ensemble Methods

We validated the proposed LxE by comparing the single MKCLPT model trained in Section 2.6.2 with ensembles consisting of 5 randomized MKCLPT. The number of estimators in the ensemble is relatively low since the MKCLPT produced relatively high completeness values ($> 97\%$). Therefore, it is expected to take a few number of steps in the hierarchy to produce a final answer. We used 75% of the features when creating each node following the randomization rule explained in Section 2.3.4.

Tables 2.5, 2.6 and 2.7 show the behavior of the proposed ensemble method using the four aforementioned heuristics (i.e. $\mathcal{C}_{\mathcal{RP}}$, Borda, correctness and LexRank). All the methods behaved similarly way with slightly different results, being the one based on Borda count the method with the highest correctness. Since the abstention degree does not depend on the order of the base estimators, all the methods obtained the same completeness score.

The proposed aggregation strategy was able to increase the completeness of the base MKCLPT model to 99.58% in the train-test set and 99.56% in the test-test set. Moreover, the attained completeness values are higher than the ones obtained by the rest of the

Table 2.7: Average relative difference (%) between each LE strategy and base lexicographic model in terms of CR and CP.

Train-Test							
Ensemble	**Correctness**				**Completeness**		
	MKCLPT	CLPT	LO	LE	MKCLPT	CLPT	LO
$\mathcal{C}_{\mathcal{RP}}$	-3.2183	9.1312	8.3927	-0.4529	1.9132	1.0712	0.3419
Borda	-2.7968	9.5616	8.8072	-0.0824	1.9132	1.0712	0.3419
CR	**-2.7741**	**9.5883**	**8.8369**	**-0.0367**	1.9132	1.0712	0.3419
LexRank	-2.8385	9.5164	8.7632	-0.1364	1.9132	1.0712	0.3419
Test-Test							
$\mathcal{C}_{\mathcal{RP}}$	-3.7773	9.0306	4.6674	-0.5214	2.3551	1.4845	0.3089
Borda	-3.3340	9.4438	5.0503	-0.2930	2.3551	1.4845	0.3089
CR	-3.2482	9.5393	5.1382	-0.1941	2.3551	1.4845	0.3089
LexRank	**-3.0632**	**9.7629**	**5.3451**	**-0.0123**	2.3551	1.4845	0.3089

models. The best average results in the train-test and test-test sets were achieved by the CR and LexRank rules respectively.

Since increasing completeness requires to introduce more risky predictions, the average correctness decreased when compared with the base MKCLPT model. However, as can be seen in the results, the behavior of the ensemble strategies is better in terms of correctness than the other methods in the literature.

2.7 Conclusions and Future Work

Learning to rank is an inductive learning technique that focuses on creating predictive models able to generalize underlying preference models on combinatorial domains. Being a plausible representation of human preferences [292], conditional and unconditional LxO are languages (and predictive models) for representing compactly and reasoning about multi-attribute preferences. Moreover, since LxO are NSR, they are able to represent preference models that can not be expressed as a utility function [277]. However, LxO may pose a strong bias when dealing with preference functions that do not behave in a lexicographic way. While other Learning to Rank methods can deal with non-categorical features, traditional LxO methods require introducing global preprocessing techniques to transform the attributes domain [36].

We overcome these limitations introducing Conditional Lexicographic Preference Trees with Multiple Kernels, MKCLPT, a new parametric language to reason about lexicographic preferences. MKCLPT is a kernelized version of CLPT, allowing: 2) to decrease the dependence on the assumption of having an underlying lexicographic preference model, and 2) to deal with non-categorical attributes.

We proposed a learning strategy to induce CLPT highly consistent with a partially annotated training set. As shown in the experimental evaluation, the proposed language

and learning strategy performed better in terms of correctness than CLPT (trained using CLeRa algorithm [36]) and LxO (trained using LexRank [109]). However, MKCLPT was more prune to abstain than other rankers. Thereby, we defined an aggregation technique to build ensembles of lexicographic rankers that induces more complete orders. Through the proposed experiments we validated that the aggregation was able to increase the completeness of the induced rankers with a controlled correctness reduction. On the other hand, the computational cost of learning a MKCLPT node is higher than learning a CLPT node, given that kernels are learned locally.

Combining learned randomized MKCLPT and lexicographic aggregation techniques, we defined Rankdom Forest, which can be understood as the NSR analog of Random Forest (RF) used for classification. As we show, Rankdom Forests offer promising results when reasoning about partially annotated preference models, leading to future work in different directions. For instance, we are considering the inclusion of a regularization factor in the allowed completeness of each level to balance the prediction rate through the hierarchy. Also, we intend to explore general voting rules by allowing inconsistencies and introducing optimization techniques to minimize them.

Chapter 3

Ranking for Imbalance Classification

This chapter was published in [58]:

- Ricardo Cruz, Kelwin Fernandes, Jaime S. Cardoso, and Joaquim F. Pinto Costa. Tackling class imbalance with ranking. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 2182–2187. IEEE, 2016

Kelwin Fernandes conceived the proposed methodology and participated in the development of the preliminary versions of the solution. Ricardo Cruz worked in the experimental assessment and paper writing. Extensions of this work to handle imbalance distributions in ordinal classification tasks were published in [60–62, 264]:

- Ricardo Cruz, Kelwin Fernandes, Joaquim F. Pinto Costa, María Pérez Ortiz, and Jaime S. Cardoso. Ordinal class imbalance with ranking. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 3–12. Springer, 2017
- Ricardo Cruz, Kelwin Fernandes, Joaquim F. Pinto Costa, María Pérez Ortiz, and Jaime S. Cardoso. Combining ranking with traditional methods for ordinal class imbalance. In *International Work-Conference on Artificial Neural Networks*, pages 538–548. Springer, 2017
- María Pérez-Ortiz, Kelwin Fernandes, Ricardo Cruz, Jaime S. Cardoso, Javier Briceño, and César Hervás-Martínez. Fine-to-coarse ranking in ordinal and imbalanced domains: An application to liver transplantation. In *International Work-Conference on Artificial Neural Networks*, pages 525–537. Springer, 2017
- Ricardo Cruz, Kelwin Fernandes, Joaquim F. Pinto Costa, María Pérez-Ortiz, and Jaime S. Cardoso. Binary ranking for ordinal class imbalance. In *Pattern Analysis and Applications*. Springer, 2018

In classification, when there is a disproportion in the number of observations in each class, the data is said to be class imbalance. Class imbalance is pervasive in real-world applications of data classification and has been the focus of much research. The minority class contributes too little to the decision boundary because the learning process learns

from each observation in isolation. In this chapter, we discuss the application of learning pairwise rankers as a solution to class imbalance. We compare ranking models to alternatives from the literature.

3.1 Introduction

It is not uncommon in classification problems for data to be class imbalance; that is, the class distribution is not uniform, sometimes dramatically not. This is true in such fields as medicine where more people are cleared as negative in screening than are accused positives by such tests. In such cases, a naive application of learning algorithms will produce uninteresting models that have very good overall accuracy, at the expense of the minority class.

Several approaches have been proposed in tackling this problem, which usually involve:

A. ***Pre-processing*** step changing the class priors by undersampling the majority class and/or creating new synthetic examples of the minority class [43], or even changing class priors by changing class labels themselves (e.g., MetaCost [76]);

B. ***Training with costs*** instead of maximizing accuracy, the training algorithm maximizes weighted accuracy, so that the cost of misclassifying a class is inversely proportional to its frequency;

C. ***Post-processing*** by tweaking the decision boundary by such measures as changing a threshold after which one class is selected, sometimes with the aid of a ROC curve;

D. ***Ensembles*** by which each model within the ensemble is trained with balanced subsets of the data, coupled with the previous preprocessing techniques.

This list is by no means meant to be exhaustive. One-class models are also sometimes used to identify the minority class, though they do not usually produce interesting results [241, see Table 4]. On the other hand, some rule induction models can be made to prioritize one class, and have been found to produce interesting results [182].

In this work, we propose adding pairwise rankers to the repertoire of such techniques. We will first introduce some of the current methods from the literature in more detail before delving into learning pairwise rankers. Previous work had found that rankers can produce better AUC curves [52].

3.1.1 Pre-processing

Stratification is the most popular pre-processing approach. It works either by undersampling from the majority class which has the side benefit of significantly improving training times when class imbalance is severe. Another strategy, sometimes used in conjunction, is oversampling by creating new synthetic samples. A common algorithm is known as

SMOTE [43], where new observations are built in between two existing observations using Euclidean distances. SMOTE has been extended by other algorithms; for example, MSMOTE [154]. These extensions improve SMOTE by adding new heuristics, most notably by identifying outliers and refraining from using them for the oversample, as well as identifying boundary points.

Also worth of notice is MetaCost which works by first assigning a probability to each observation as belonging to each class, by using bagging of models, and secondly by calculating a threshold below which any observation of the majority class is assigned to the minority class [76]. In other words, it balances class priors by actually changing the class of those majority observations for which the underlying estimator is less certain about.

The immediate advantage of pre-processing is that it is model agnostic: class imbalance can then be solved as a separate problem. This is especially important when one is unsure of the most appropriate learner, and would prefer to tackle class imbalance as a separate issue.

3.1.2 Training with costs

Training by explicitly defining costs seems like the most direct approach. Instead of minimizing total misclassification, FP+FN, we minimize the weighted misclassification, $w_{\mathrm{P}}\mathrm{FP} + w_{\mathrm{N}}\mathrm{FN}$, where to the weights w_{P} and w_{N} are assigned the inverse frequency of their class priors (P and N stand for Positives and Negatives, respectively, while TP and FP are True and False Positives, conversely for TN and FN).

Unfortunately, adding cost-sensitivity to the training algorithm is not always straightforward and is sometimes cumbersome. Taking Support Vector Machines (SVM) as an example, suggestions have been made to introduce costs in the feature space transformation by changing the kernel function [335], introducing different penalties for the positive and negative SVM slack variable ξ [23], among other approaches.

Furthermore, cost training sometimes saturates and cannot expand beyond the limits of the data, which pre-processing methods can help. It has been found that pre-processing approaches are in fact oftentimes superior [76].

3.1.3 Post-processing

This step consists in varying the threshold by which the class is chosen in a binary classifier (e.g., Artificial Neural Networks (ANN)), or it could mean changing the bias in a SVM model to adjust the decision boundary [249].

3.1.4 Ensembles

Several ensembles have been proposed recently. Easy Ensemble is a high performing bagging technique where each model is trained from undersampled pools of the data, in which each pool has the same number of positive and negative observations [213]. An

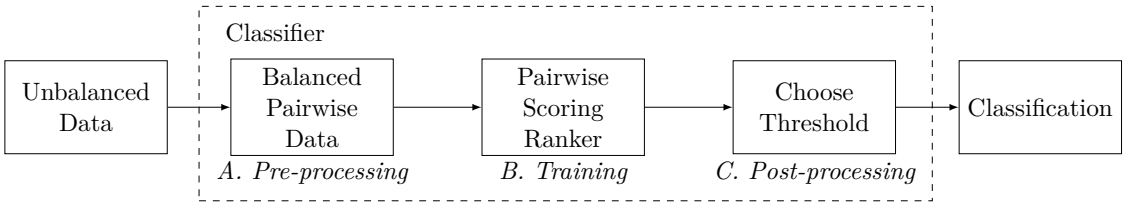


Figure 3.1: Schematic of the pairwise ranking classifier applied to class imbalance data.

extension to this model is Balance Cascade whereby undersampling of the majority class is guided to remove observations that have been correctly classified by the previous model in the ensemble, also [213].

Other methods, such as SMOTE Boost, SMOTE Bagging, IIVOTES or RUSBOOST, on the other hand, create new synthetic observations based on the harder to classify cases of minority observations; this is in opposition to traditional boosting methods which assign a distribution of weights to the observations.

Ensembles tend to triumph in recent literature [241]. It does not seem completely clear however whether what contributes to these gains is the combination of ensemble and stratification, or whether it is simply the ensemble since they are not usually benchmarked against ensembles of the other approaches.

3.2 Ranking for Class Imbalance

One possible family of methods to tackle the class imbalance problem is pairwise ranking algorithms, in particular, pairwise SRk. The term document is typically used in the literature to refer to observation because of its genesis and tight connection to information retrieval techniques [199].

In ranking, document $\mathbf{x}_i$ is compared with another document $\mathbf{x}_j$, and we are interested in predicting whether $\mathbf{x}_i \succ \mathbf{x}_j$, meaning $\mathbf{x}_i$ is "preferred" to $\mathbf{x}_j$. The three big umbrellas of rankers are:

- **Pointwise**, in which each document $\mathbf{x}_i$ is trained individually and a score function, $f(\mathbf{x}_i)$, is given based on its relevance;
- **Pairwise**: each document $\mathbf{x}_i$ is compared against all others $\mathbf{x}_j$, and if $\mathbf{x}_i \succ \mathbf{x}_j$, then we train a function f so that:
  - **pairwise scoring ranker:** if $\mathbf{x}_i \succ \mathbf{x}_j$ then $f(\mathbf{x}_i) > f(\mathbf{x}_j)$, with $f\colon X \to \mathbb{R}$;
  - **pairwise non-scoring ranker:** the decision function is such that it decides which of two documents is preferred, $f\colon X^2 \to X$;
- **Listwise**, where the training loss function is based on all documents and their scores.

We propose to consider pairwise SRk for the class imbalance problem. Ranking algorithms for classification have been found to make highly competitive classifiers [109]. And, as we will see, there is no class imbalance when doing pairwise ranking. Since we are comparing each observation of one class to every observation of the other class, the ensuing training process is necessarily balanced.

In the same spirit of the category of models presented in the introduction, we can see the ranking process as being composed of the following steps: pre-processing, training, and post-processing (see diagram in Figure 3.1).

3.2.1 Pre-processing

In the case of binary classification, pairwise rankers are trained so that for every two observations, $(\mathbf{x}_i, \mathbf{x}_j)$ and respective class labels (y_i, y_j), a transformation f is applied so that $\mathbf{x}_i \succ \mathbf{x}_j$ if $\mathrm{P}(y_i = 1) > \mathrm{P}(y_j = 1)$, and $\mathbf{x}_i \prec \mathbf{x}_j$ otherwise. Here we take 1 as being the minority class, without loss of generality.

In order to illustrate the ranking approach, among many potential pairwise SRk, here three are considered. Any others could be adapted in an analogous manner. These were selected because they are a) pairwise scoring, and b) they encompass major families of rankers:

Table 3.1: Ranking models explored

Family	Ranker	Reference
Linear SVM	RankSVM	[150]
ANN	RankNet	[39]
AdaBoost	RankBoost	[113]

In **RankSVM**, data is transformed into the space of differences, so the original dataset $\mathbf{X}$ becomes $\mathbf{X}'$, where $\mathbf{x'}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ and $\mathbf{x'}_{ji} = \mathbf{x}_j - \mathbf{x}_i$, for all pairs (i, j) such that $y_i \neq y_j$, with $y'_{ij} = y_i$ and $y'_{ji} = y_j$.

In all others, data is transformed so that $\mathbf{X}_i = \{\mathbf{x}_i, y_i\}$ becomes $\mathbf{X}'_{ij} = \{\mathbf{x_i}, y'_{ij}\}$, for all combinations (i, j), (and ditto for the symmetric relation (j, i)), where y'_{ij} denotes a relation of preference between $\mathbf{x}_i$ and $\mathbf{x}_j$ ($y'_{ij} = 1$ if $\mathbf{x}_i \succ \mathbf{x}_j$, or -1 otherwise). In **RankBoost**, y'_{ij} denotes the class of i, $y'_{ij} = y_i$ (and all combinations such that $y_i = y_j$ are omitted), while in **RankNet** y'_{ij} represents the ranking probability we aim to estimate,

$$y'_{ij} = \begin{cases} 0, & \text{if } y_i < y_j \\ 1, & \text{if } y_i > y_j \\ 0.5, & \text{if } y_i = y_j. \end{cases}$$

The data with which the ranking estimator is trained is therefore usually bigger, and so training times tend to be slower than ordinary classification methods. In general, the

transformed dataset $N' \in \mathcal{O}(N^2)$, but in cases like RankSVM or RankBoost which use only pairs of opposite classes, $N' = 2N_0N_1$ and, because of class imbalance, $N_0 \gg N_1$, so $N' \approx 2N_0$.

3.2.2 Training

When it comes to training, **RankSVM** makes use of a linear SVM as a base estimator to classify observations within the space of differences, where the decision rule $\mathbf{w} \cdot (\mathbf{x}_i - \mathbf{x}_j) > 0$ can be transformed into a scoring function since $\mathbf{w} \cdot (\mathbf{x}_i - \mathbf{x}_j) > 0 \equiv \mathbf{w} \cdot \mathbf{x}_i > \mathbf{w} \cdot \mathbf{x}_j \equiv s(\mathbf{x}_i) > s(\mathbf{x}_j)$. In **RankNet**, an ANN is used to estimate y'_{ij} which denotes the probability $\mathrm{P}(\mathbf{x}_i \succ \mathbf{x}_j)$.

RankBoost, like AdaBoost from the same authors, trains a base estimator, at each iteration t, using an underlying distribution of weights for each pair, for which $D_{ij} = D_{ji}$. The difference from AdaBoost is in that α_t is a function of the number of pairs whose order has been correctly or incorrectly estimated: $\alpha_t = \frac{1}{2} \log \frac{W_{-1}}{W_{+1}}$, where $W_{b=\{-1,+1\}} = \sum_{i,j} D_{ij} I_{f_t(\mathbf{x}_i) - f_t(\mathbf{x}_j) = b}$.

Loss functions are therefore hinge loss, logistic loss, and exponential loss for RankSVM, RankNet, and RankBoost, respectively.

3.2.3 Post-processing

As discussed, a pairwise SRk produces a score function $f \colon X \to \mathbb{R}$, with which we predict a class $\{0, 1\}$ using a scoring threshold. We have chosen the threshold T that maximizes the F_1 score, which is more appropriate than accuracy for class imbalance, and is defined as:

$$\mathrm{F}_1 = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}.$$

Using the training data, we have $s_i = f(\mathbf{x}_i)$ which we order, and use each midpoint $s'_i = \frac{s_i + s_{i+1}}{2}$ as possible candidates for threshold T, so that

$$T = \arg\max_{s'_i} \ \mathrm{F}_1(s'_i).$$

For fairness, the F_1 score metric is also used when cross-validating the best parameters of the models we are comparing.

In the following experiments tables, we also make use of these scores to calculate the area under curve (AUC) of ROC curves. The ROC curve is a common measure to evaluate how correctly classified observations would be if the decision threshold T was changed. What should constitute this decision threshold is not always obvious; in the SVM, this could be the distance to the separating hyperplane, as was used, or varying the bias [249]. In ranking models, the ROC can easily be drawn by choosing different scores s_i for T.

Table 3.2: Datasets

Dataset	Minority	N	Features	**IR**	Overlap
sonar	—	208	60	0.466	0.216
breast-cancer	—	699	9	0.345	0.075
german	—	1000	24	0.300	0.563
haberman	—	306	3	0.265	0.605
transfusion	—	748	4	0.238	0.629
vehicle	van	846	18	0.235	0.090
CTG	—	2126	22	0.222	0.172
hepatitis	—	143	14	0.203	0.621
segment	1	2310	19	0.143	0.009
winequality-red	7,8	1599	11	0.136	0.512
vowel	1	990	13	0.091	0.011
abalone	9vs18	731	7	0.057	0.595
glass	6	214	9	0.042	0.556
car	good	1728	6	0.040	0.667
yeast	ME1	1484	8	0.030	0.341

Acknowledgments: Datasets come courtesy of the UCI Machine Learning repository [207]. The breast cancer dataset was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg [224]. The vehicle dataset is originally from the Turing Institute, Glasgow, Scotland. Wine-quality is originally from [54].

3.3 Experiments

Fifteen empirical datasets are considered (see Table 3.2). Most datasets used have N (the number of observations) in the order of thousands, to ensure what is being tested is "relative rarity", and avoid "absolute rarity" issues [144]. Some of these are multinomial classification datasets which were converted to binary classification using the class label mentioned in the "Minority" column. These samples are based on [45]. All others are binary classification datasets. Datasets in all proceeding tables are ordered by Imbalance Ratio (IR), N_1/N.

Overlap is a measure of how intertwined the observations from the two classes are. There is a significant amount of literature on the role of overlapping in class imbalance [274], with some authors arguing these problems are often conflated [72]. Our measure of overlap is defined as the ratio of minority observations whose closest neighbor is an observation of the majority class. We compare models' correlation to IR and overlap in the results.

Experiments are done by cross-validation through a bootstrap process: each sample is randomly split into 40 folds of 80-20% stratified splits of train-test. The average of F_1 and ROC AUC for each dataset is exhibited. The best scores are presented in bold, as well as all statistically identical scores, using a paired difference Student's t-test with a 95% confidence level.

Furthermore, a 5-fold validation is performed for SVM and ANN to find the best parameter: the regularization coefficient C and hidden nodes H, respectively, choosing

Table 3.3: Family: Linear SVM

Linear SVM	F_1						*ROC AUC*					
Sample	Ranking	Baseline	Weights	SMOTE	MSMOTE	MetaCost	Ranking	Baseline	Weights	SMOTE	MSMOTE	MetaCost
sonar	**0.741**	0.723	**0.734**	0.723	0.724	0.725	**0.832**	0.823	**0.826**	0.823	0.823	0.816
breast-cancer-wisconsin	**0.957**	0.953	0.954	**0.955**	0.954	0.953	0.994	**0.994**	0.994	0.994	**0.994**	**0.994**
german	**0.618**	0.568	**0.622**	0.616	**0.617**	0.598	**0.808**	**0.809**	**0.809**	0.806	0.804	**0.808**
haberman	**0.485**	0.188	**0.476**	**0.484**	**0.469**	0.253	**0.685**	**0.689**	**0.690**	0.677	0.668	**0.688**
transfusion	0.516	0.154	**0.528**	0.518	0.522	0.174	**0.758**	**0.758**	**0.757**	0.753	0.752	**0.758**
vehicle-van	**0.937**	**0.940**	0.930	0.932	0.934	0.927	**0.995**	0.995	0.995	0.995	0.994	0.994
CTG	0.925	0.931	0.915	0.917	0.919	**0.934**	**0.993**	**0.993**	**0.993**	0.993	0.992	0.993
hepatitis	**0.634**	**0.606**	**0.651**	**0.630**	**0.632**	**0.648**	**0.882**	**0.877**	**0.884**	**0.882**	**0.871**	**0.881**
segment-1	0.986	**0.991**	0.988	**0.990**	**0.990**	0.990	0.996	**0.998**	0.997	**0.998**	**0.998**	**0.999**
winequality-red-7,8	**0.517**	0.228	0.479	0.482	0.491	0.346	**0.858**	**0.857**	**0.858**	0.856	0.850	0.855
vowel-1	**0.457**	0.180	**0.445**	0.442	0.421	0.273	**0.892**	0.884	**0.893**	0.887	0.850	0.870
abalone-9vs18	**0.652**	0.502	0.473	0.493	0.500	**0.632**	0.948	**0.952**	**0.950**	0.948	0.926	**0.950**
glass-6	**0.695**	0.000	0.234	0.236	0.226	0.010	**0.984**	0.507	0.763	0.752	0.761	0.436
car-good	**0.476**	0.064	0.422	0.438	0.391	0.274	**0.959**	0.958	**0.959**	0.958	0.941	0.955
yeast-ME1	**0.612**	0.523	0.556	0.562	0.564	0.571	**0.986**	**0.986**	**0.986**	0.985	0.984	0.986
Average	0.681	0.503	0.627	0.628	0.624	0.554	0.905	0.872	0.890	0.887	0.881	0.865
Winner	80%	20%	40%	26%	26%	20%	80%	66%	73%	13%	20%	46%

Table 3.4: Family: AdaBoost

AdaBoost	F_1						*ROC AUC*					
Sample	Ranking	Baseline	Weights	SMOTE	MSMOTE	MetaCost	Ranking	Baseline	Weights	SMOTE	MSMOTE	MetaCost
sonar	0.787	**0.824**	**0.824**	**0.824**	**0.824**	0.818	0.891	**0.917**	**0.917**	**0.917**	**0.917**	**0.916**
breast-cancer	0.937	0.932	0.932	0.937	0.934	**0.948**	**0.990**	0.989	0.989	0.990	0.989	**0.991**
german	**0.597**	0.541	0.541	**0.588**	0.583	**0.587**	0.783	**0.794**	**0.794**	**0.792**	**0.793**	**0.796**
haberman	0.419	0.375	0.375	**0.464**	**0.458**	0.441	0.638	0.663	0.663	0.671	0.673	**0.692**
transfusion	**0.502**	0.418	0.418	**0.511**	**0.509**	0.493	0.718	**0.745**	**0.745**	0.737	0.737	**0.740**
vehicle-van	**0.928**	0.901	0.901	0.905	0.909	0.906	**0.993**	0.991	0.991	0.991	0.991	0.989
CTG	0.973	0.972	0.972	0.970	0.971	**0.979**	0.996	**0.997**	**0.997**	0.996	**0.997**	0.996
hepatitis	**0.578**	0.525	0.525	**0.581**	**0.596**	**0.596**	0.822	0.808	0.808	**0.833**	**0.842**	**0.846**
segment-1	**0.993**	0.990	0.990	0.987	0.988	0.988	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	1.000
winequality-red-7,8	**0.520**	0.431	0.431	0.509	0.511	**0.528**	**0.867**	**0.868**	**0.868**	0.864	0.863	**0.869**
vowel-1	**0.692**	0.443	0.443	0.610	0.549	0.633	**0.953**	0.946	0.946	**0.948**	0.930	0.939
abalone-9vs18	**0.369**	0.318	0.318	0.287	0.298	**0.377**	0.803	**0.820**	**0.820**	0.793	0.791	**0.822**
glass-6	**0.801**	0.670	0.670	**0.825**	**0.777**	0.713	**0.996**	**0.998**	**0.998**	**0.993**	0.989	0.982
car-good	**0.573**	0.388	0.388	**0.596**	0.515	0.401	0.974	**0.977**	**0.977**	0.976	0.965	0.919
yeast-ME1	**0.667**	**0.671**	**0.671**	0.654	0.635	**0.698**	0.982	**0.986**	**0.986**	**0.985**	0.982	**0.986**
Average	0.689	0.627	0.627	0.683	0.671	0.674	0.894	0.900	0.900	0.899	0.897	0.899
Winner	73%	13%	13%	46%	33%	46%	40%	66%	66%	46%	33%	60%

Table 3.5: Family: Artificial Neural Networks

Neural Networks	F_1						*ROC AUC*					
Sample	Ranking	Baseline	Weights	SMOTE	MSMOTE	MetaCost	Ranking	Baseline	Weights	SMOTE	MSMOTE	MetaCost
sonar	**0.805**	**0.804**	**0.801**	0.731	0.732	0.725	**0.881**	**0.896**	**0.895**	0.830	0.829	0.817
breast-cancer	0.946	0.942	0.947	**0.955**	**0.954**	**0.955**	0.975	0.989	0.991	0.994	0.993	**0.994**
german	0.513	0.540	0.543	**0.618**	0.609	**0.623**	0.665	0.744	0.721	**0.808**	0.807	**0.809**
haberman	0.444	0.361	0.459	0.435	0.431	**0.497**	0.604	**0.675**	**0.678**	0.676	0.663	**0.688**
transfusion	0.495	0.391	0.506	0.492	0.502	**0.525**	0.549	**0.764**	0.753	0.757	0.755	0.758
vehicle-van	**0.944**	**0.940**	**0.941**	0.593	0.596	0.596	**0.982**	**0.996**	**0.996**	0.985	0.985	0.923
CTG	0.961	**0.965**	**0.963**	0.919	0.919	0.925	0.993	0.997	**0.998**	0.993	0.993	0.993
hepatitis	**0.600**	0.502	0.528	0.520	0.515	**0.611**	**0.828**	0.803	0.801	0.795	0.791	**0.846**
segment-1	**0.989**	**0.990**	**0.975**	0.984	0.984	0.986	**0.999**	**0.999**	**0.999**	0.996	0.996	0.996
winequality-red-7,8	0.477	0.482	**0.508**	0.316	0.317	0.269	0.555	0.823	**0.843**	0.790	0.786	0.655
vowel-1	0.397	**0.946**	0.855	0.480	0.482	0.379	0.516	**0.989**	0.973	0.963	0.954	0.899
abalone-9vs18	**0.511**	0.485	0.362	0.301	0.347	0.358	0.801	**0.917**	0.907	**0.927**	**0.922**	0.889
glass-6	0.024	0.000	0.136	**0.350**	**0.347**	0.290	0.442	0.338	0.556	**0.683**	**0.698**	0.610
car-good	**0.839**	**0.849**	0.737	0.447	0.402	0.392	0.959	**0.996**	0.982	0.966	0.951	0.953
yeast-ME1	**0.653**	0.564	0.528	0.603	0.583	0.597	0.950	**0.986**	0.983	**0.986**	0.984	**0.986**
Average	0.640	0.651	0.653	0.583	0.581	0.582	0.780	0.861	0.872	0.876	0.874	0.854
Winner	46%	40%	33%	20%	13%	33%	26%	60%	40%	26%	13%	33%

between 5 parameters along the range $C \in [0.01, 100]$, and $H \in [F, F^2]$, with F being the number of features. SVM was trained using a linear kernel with liblinear. Stochastic gradient descent was used with learning rate = 1.0, and 1000 as the epochs maximum.

The data was normalized for both. AdaBoost and RankBoost were trained as an ensemble of 50 binary classifiers.

Four variants of the baseline model are provided in each column: loss function with weights using inverse class frequencies, and also together with SMOTE [43] and MSMOTE [154] with number of neighbors $k = 5$ applied to equalize frequencies, as well as with MetaCost [76] using an ensemble of $m = 50$ with the other parameters being $n = N$, $p =$ False and $q =$ True.

3.4 Results

Table 3.6: Correlations: Data Complexity

Spearman's ρ	Ranking	Baseline	Weights	SMOTE	MSMOTE	MetaCost
Linear SVM						
IR	0.312	0.477	0.576	0.555	0.544	0.401
Overlap	-0.185	-0.302	-0.293	-0.285	-0.293	-0.293
AdaBoost						
IR	0.115	0.224	0.224	0.148	0.208	0.238
Overlap	-0.668	-0.609	-0.609	-0.613	-0.601	-0.674
Neural Networks						
IR	0.210	0.135	0.277	0.398	0.407	0.463
Overlap	-0.645	-0.756	-0.791	-0.650	-0.647	-0.584

The ranking models here considered have performed statistically significantly better than their counterparts from the literature, especially with regard to the F_1 score (Table 3.3). In Linear SVM, when ranking won, it won by a much bigger margin than when it lost, $0.068/-0.009$, relative to the second performer and best performer.

Table 3.7: Correlations: Inter-Family

Spearman's ρ	Baseline	Weights	SMOTE	MSMOTE	MetaCost
RankSVM	0.710	0.742	0.751	0.756	0.715
RankBoost	0.842	0.842	0.888	0.861	0.862
RankNet	0.736	0.754	0.610	0.614	0.656

The lower performance from the ROC AUC scores could suggest the threshold selection (section 3.2.3) is partly responsible for the gain.

While not the main point of the work, it is worth noticing that, as other authors have argued [274], data's overlap (from Table 3.2) explains better model's F_1 scores than IR, as measured by Spearman's ρ ($\rho \in [-1, 1]$ with high/low $|\rho|$ meaning high/low correlation), see Table 3.6. More importantly, rankers, when compared to the other models within their family, produce models least correlated to the IR. It was already visible from Table 3.3, which is ordered by IR, that rankers gains are concentrated in the bottom (the more unbalanced). And, while overlap explains scores better than IR, as already stated, no

systemic tendency is apparent, and so all gains from rankers seem to accrue to solving the IR problem.

Table 3.8: Correlations: Intra-Family

Spearman's ρ	RankSVM	RankBoost	RankNet
RankSVM	1.000	0.365	0.235
RankBoost	0.365	1.000	0.555
RankNet	0.235	0.555	1.000

Finally, we compare correlations within and between families of models. This can help in differentiating, on the one hand, whether rankers are competing classifiers or if, on the other hand, rankers are alternative models that learn different data patterns. Correlation is, again, measured by Spearman's ρ. Naturally, the correlation between any two models will be high since we are using datasets that were chosen because they have different IR, and IR is (inversely) correlated to model's performance, and is, therefore, a confounder. We control for IR's correlation using Fisher's partial correlation formula. This does not affect the relative correlations between any two models, but it reduces the overall magnitude of correlations to be more aligned to use cases when random samples from the same population, having the same IR, are used.

Table 3.7 and 3.8 clearly show rankers more closely follow the decision function of their family of models than that of the rest of the rankers. Ranking techniques are therefore an extra technique of tackling class imbalance to try to improve a currently employed solution.

3.5 Discussion

There is a latent benefit when considering rankers as possible classifiers; a latent benefit that has not so far been discussed. Rankers can use extra information about the order of classes. This means that data collection is not as constrained to broad categories such as "healthy" and "sick", or "credit-worthy" and "not credit-worthy". Rankers can make use of extra subtlety in the classification by having a gradient of classes. A tangent point is that in many real-world applications it might make sense to express the data from the get-go in terms of pairwise comparisons. It is often more intuitive for the human classifier to compare observations than to assign labels.

This was not a focus of this discussion, but one inconvenience is that training times are usually higher, possibly insuperably higher for very big datasets. We have however only implemented and experimented with the three major ranking families while ignoring the more recent progress.

Further work is required to more finely tuned ranking solutions, as well as combining rankers with current pre-processing and ensemble solutions. Tackling imbalance in multi-class problems and reducing training times are other problems of interest. The ranking

threshold decision could possibly be solved using a SVM to separate classes, or, more elegantly, while training.

3.6 Conclusion

Almost two hundred papers have been published since 2012, just by searching titles by "class imbalance" as reported by Google Scholar. It is not clear that ranking is a superior solution, but it is a very competitive and promising alternative that we felt was sorely lacking in the literature.

Some classical ranking models were compared with conventional classification models. This work shows promising efficiency improvements from training using pairwise SRk models. These models have been in general superior, and when their performance was worse, it was worse by a smaller margin than when the performance was positive. It is clear class imbalance performance can be improved by combining these models with other approaches from the literature.

It was also found that performance scores of ranking models tend to correlate with those of their underlying models, and so they may be seen as potential improvements on top of traditional classifiers.

Chapter 4

Constraining Type II Error

An extended version of this chapter was published in [59]:

- Ricardo Cruz, Kelwin Fernandes, Joaquim F. Pinto Costa, and Jaime S. Cardoso. Constraining Type II Error: Building Intentionally Biased Classifiers. In *International Work-Conference on Artificial Neural Networks*, pages 549–560. Springer, 2017

The version included in this dissertation is restricted to the contributions done by Kelwin Fernandes. Alternative approaches to the methodology presented in this chapter were proposed by Ricardo Cruz in the aforementioned publication.

In many applications, false positives (type I error) and false negatives (type II) have a different impact. In medicine, it is not considered as bad to falsely diagnosed someone healthy as sick (false positive) as it is to diagnose someone sick as healthy (false negative). But we are also willing to accept some rate of false negatives errors in order to make the classification task possible at all. Where the line is drawn is subjective and prone to controversy. Usually, this compromise is given by a cost matrix where an exchange rate between errors is defined. For many reasons, however, it might not be natural to think of this trade-off in terms of relative costs. We explore novel learning paradigms where this trade-off can be given in the form of the number of false negatives we are willing to tolerate. The classifier then tries to minimize false positives while keeping false negatives within the acceptable bound. Here we consider classifiers based on kernel density estimation, gradient descent modifications and applying a threshold to classifying and ranking scores.

4.1 Introduction

A common problem in medical and financial decision-making is that when classifying an observation, different classification costs must be considered for different classification errors. These costs are real, but expressing them in the form of a number is messy, hard and prone to controversy. For instance, it is not obvious how bad it is to fail to diagnose cancer versus being too careful and subjecting the patient to undue biopsies.

Metrics based on minimizing the expected cost (or maximizing the expected benefit) are not intuitive for human evaluators, such as physicians. Instead, humans tend to instead avoid things like misclassifying positives beyond an acceptable rate of false negatives. It is therefore important to hold our models to the same standard if they are intended to be realistic contenders (or complements) to an existing human evaluator. An ML metric based on a false negative threshold for training and evaluating models can potentially be more intuitive and provide a higher sense of trustworthiness which is important for ML penetration in such fields as medicine.

This situation is further aggravated by the fact that too commonly these problems suffer from class imbalance; that is, there are typically too few observations of the more severe positive class, which further taints training toward the majority class [31].

The problem we aim to tackle is best illustrated in broad strokes by the following optimization problem:

$$\begin{aligned} &\text{Maximize TNR} \\ &\text{subject to FNR} \leq \rho. \end{aligned}$$

Here, TNR refers to the true negatives rate, also known as specificity, and FNR is the false negatives rate, and is equal to $1-$sensitivity. Throughout the manuscript, we will be using the user-defined parameter ρ to represent the user-acceptable FNR, and $\widehat{\rho}$ for the empirical FNR, estimated from the data sample.

As is common in the literature, and without loss of generality, we take the positive class $(+)$ to be the minority class, and assume this is the class whose classification errors we aim to control ($\text{FNR} \leq \rho$).

4.2 State of the Art

Imputing costs via a cost matrix is the *de facto* approach for tackling false classification trade-offs. In the most common case, when the correct decision has null cost, then the cost matrix has only one degree of freedom,

$$\begin{pmatrix} 0 & c_p/c_n \\ 1 & 0 \end{pmatrix}$$

These costs are then taken into account by the model through:

- **pre-processing** by changing the priors: either by stratification techniques (over-sampling and undersampling) or by synthetically creating new observations [43] or even changing the class labels as done by MetaCost [76];

- **the training algorithm** may in some cases be made sensible to the misclassification costs as well. However, adding cost-sensitivity to the training algorithm is not always straight-forward and is sometimes cumbersome. Taking SVMs as an example,

suggestions have been made to introduce costs in the feature space transformation by changing the kernel function [335], introducing different penalties for the positive and negative SVM slack variable ξ [23], among other approaches. Furthermore, such penalties might saturate the decision boundary as shown in Figure 4.1;

- **post-processing** techniques involve using class posterior probabilities or the distance to the decision boundary, usually in the context of a Area Under the Receiver Operating Characteristic curve (ROC AUC) curve.

All these solutions, however, are based on a relative trade-off between FNR and FPR, false negative and positive rates respectively. None of the approaches offers a means to define an absolute trade-off.

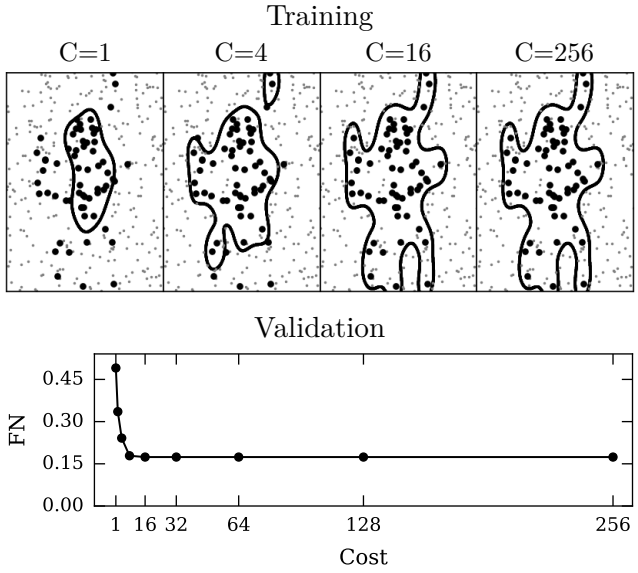
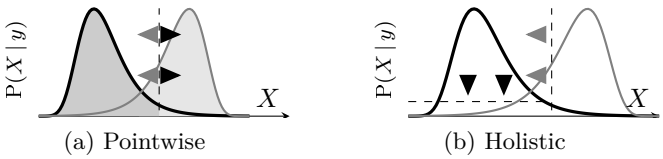


Figure 4.1: SVM trained with several costs in a noisy synthetic sample. After a while, there are no gains in $\hat{\rho}$ in the validation sample.

4.3 Proposal

Several proposals are explored in the following sections. The goal is to minimize true negatives while keeping false negatives under a user-specified threshold.



(a) Pointwise (b) Holistic

Figure 4.2: Comparing the current pointwise methodology to the proposed one.

More concretely, if y is the endogenous label, we want to ensure $\mathrm{P}(\hat{y}{=}-\,|\,y{=}+) \leq \rho$ (or, equivalently, $\mathrm{P}(\hat{y}{=}+\,|\,y{=}+) \geq 1-\rho$) in order to keep this type of error within a reasonable bound while maximizing specificity, $\mathrm{P}(\hat{y}{=}-\,|\,y{=}-)$. The user-defined parameter ρ corresponds to an absolute trade-off in terms of false negatives, in contrast to the orthodox approach of using relative trade-offs between false negatives and false positives.

One way to consider this change in methodology is to consider that current approaches expand the decision boundary until $c_p\mathrm{P}(\hat{y}{=}+|y{=}-) = c_n\mathrm{P}(\hat{y}{=}-|y{=}+)$, see Figure 4.2a, while our approach considers expanding the decision boundary of one class until the total error rate in the other class is controlled, see Figure 4.2b.

Most of the approaches here presented can be described within the following general framework:

(a) a function $f\colon \mathbb{R}^n \to \mathbb{R}$ ranks how confident we are in that an observation $\mathbf{x}$ is positive, usually in the form of a probability $\mathrm{P}(y{=}+\,|\,\mathbf{x})$;

(b) apply a threshold t such that $\mathrm{P}(f(x) < t) = \rho$ or, in other words, find the ρ quantile of $f(x)$ for the training data.

4.4 Scoring Threshold

A model is trained to produce scores representing the likelihood that the observation belongs to either class. A threshold is then used to find a quantile such that the FNR is contained. More concretely:

1. train the model, a model such that $f\colon \mathbb{R}^n \to \mathbb{R}$,
2. get the scores for the positive training data, $s_i = f(\mathbf{X}_i)$, $\forall i\ y_i = 1$,
3. choose a threshold t in the desired quantile, $\mathrm{P}(s_i < t) = \rho$.

A scoring function can be estimated for such models as SVM by using the distance to the hyperplane. More elegantly, rankers may also be built on top of SVM and other models so that the output is a score representing order. There is a vast literature on this: we will here consider RankSVM, and compare it against SVM with linear kernel. The penalty term used was $C = 1$.

Using ranking for classification has already been used in [58] with good results for class imbalance. This is essentially the same approach, using a different threshold function.

4.4.1 Ranking Threshold

One possible family of methods to tackle this problem is pairwise scoring ranking models. In contrast to classifiers, which not always make it easy to apply a threshold, this family of methods ranks the observations making a threshold straightforward to apply.

Table 4.1: Datasets used for the experiments.

Dataset	Minority	N	Features	IR
breast-cancer	wisconsin	699	9	0.345
car	good	1728	6	0.040
german	—	1000	24	0.300
haberman	—	306	3	0.265
heart	—	270	13	0.444
sonar	—	208	60	0.466
transfusion	—	748	4	0.238
vehicle	van	846	18	0.235
vowel	1	990	13	0.091
winequality-red	7,8	1599	11	0.136

Acknowledgments: Datasets come courtesy of the UCI Machine Learning repository [207]. The breast cancer dataset was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg [224]. The vehicle dataset is originally from the Turing Institute, Glasgow, Scotland. Wine-quality is originally from [54].

In ranking, observation $\mathbf{x}_i$ is compared with another observation $\mathbf{x}_j$, and we are interested in predicting whether $\mathbf{x}_i \succ \mathbf{x}_j$, meaning $\mathbf{x}_i$ is "preferred" to $\mathbf{x}_j$. In the particular case of pairwise scoring ranking, each observation $\mathbf{x}_i$ is compared against all others $\mathbf{x}_j$, and if $\mathbf{x}_i \succ \mathbf{x}_j$, then we train a function f so that $f(\mathbf{x}_i) > f(\mathbf{x}_j)$, with $f\colon X \to \mathbb{R}$ [199].

In this work, we have compared linear SVM with RankSVM [149]. In RankSVM, data is transformed into the space of differences, so the original dataset $\mathbf{X}$ becomes $\mathbf{X}'$, where $\mathbf{x'}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ and $\mathbf{x'}_{ji} = \mathbf{x}_j - \mathbf{x}_i$, for all pairs (i, j) such that $y_i \neq y_j$, with $y'_{ij} = y_i$ and $y'_{ji} = y_j$. When it comes to training, RankSVM makes use of a linear SVM as a base estimator to classify observations within the space of differences, where the decision rule $\mathbf{w} \cdot (\mathbf{x}_i - \mathbf{x}_j) > 0$ can be transformed into a scoring function since $\mathbf{w} \cdot (\mathbf{x}_i - \mathbf{x}_j) > 0 \Leftrightarrow \mathbf{w} \cdot \mathbf{x}_i > \mathbf{w} \cdot \mathbf{x}_j \Leftrightarrow s(\mathbf{x}_i) > s(\mathbf{x}_j)$.

4.5 Experiments

The datasets used are summarized in Table 4.1. IR is the imbalance ratio, $\mathrm{IR} = \mathrm{N}_+/\mathrm{N}$, where N_+ and N are the number of positive and total observations. All empirical tests are targeting $\rho = 0.05$. The *Minority* column shows which class is being considered positive, if the dataset is multi-class. All data and code used in the elaboration of this Chapter are available from: `http://vcmi.inescporto.pt/reproducible_research/iwann2017/Type2Error/`. Implementations used Python, SciPy, scikit-learn, and TensorFlow.

The results provided in this work correspond to a stratified k-fold assessment technique with k=5. For each dataset, we show scores for the training and testing set using the previously defined metrics. Table 4.2 summarizes the results using traditional and rank-based SVM.

4.6 Discussion and Future Work

One major difficulty shows to be keeping the training in tandem with the validation results. Solutions could encompass (a) aggressive regularization strategies, (b) evaluating $\widehat{\rho}$ in a different sample while training, and (c) using a smaller desired TNR value $\rho' = \eta\rho$ with $0 < \eta < 1$ and obtained by cross-validation to ensure desired TNR is controlled.

A simple post-processing threshold method has proven simple and maybe acceptable, especially when used in tandeum with a scoring pairwise ranker.

Possibly, rule induction could provide fruitful models for FNR-constraining. Models such as PNrule offer an interesting framework [7]. Positive and Negative rules are constructed, each with an associated level of PN or FN rate, which are then applied in sequence. Firstly, Positive rules are applied to reject examples; only then Negative rules are applied on top of the rejected examples, optimizing recall and precision in separate.

In multiclass scenarios, this new formulation to classification would be highly specific to the application. Suggestions would be to use a different ρ for each minority class or aggregate minority classes and then subclassify among them.

Table 4.2: Performance of threshold approach using SVM-based models ($\rho = 0.05$). The first and second lines of each dataset corresponds to the model performance on the training and test set respectively.

	Linear SVM		**RankSVM**	
	FNR	TNR	FNR	TNR
breast-cancer-wisconsin	0.05	0.98	0.05	0.98
	0.09	0.98	0.07	0.98
car-good	0.04	0.90	0.04	0.91
	0.11	0.80	0.17	0.83
german	0.05	0.39	0.05	0.41
	0.09	0.38	0.08	0.40
haberman	0.05	0.12	0.05	0.13
	0.08	0.11	0.08	0.12
heart	0.04	0.60	0.04	0.64
	0.06	0.59	0.10	0.59
sonar	0.04	0.99	0.04	0.77
	0.42	0.69	0.34	0.51
transfusion	0.05	0.26	0.05	0.27
	0.09	0.24	0.09	0.23
vehicle-van	0.05	0.99	0.05	0.99
	0.07	0.98	0.07	0.98
vowel-1	0.04	0.68	0.04	0.76
	0.26	0.68	0.27	0.76
winequality-red-7,8	0.04	0.56	0.04	0.57
	0.07	0.54	0.07	0.56

4.7 Conclusion

We have started by proposing a new learning problem: instead of defining a relative trade-off using cost matrices, we suggest it might be useful in some cases for learning algorithms to allow defining an absolute trade-off in the form of a false negative threshold. In the terminology we have used, we try to maximize specificity (true negatives) while keeping Type II errors, false negatives, within a certain bound.

We suggest a post-processing technique that, combined with ranking models, have shown to be simple and effective. Two problems and conclusions arrive. (I) It is not easy to improve on a simple scoring threshold when considering specificity. (II) A big difficulty arises in keeping FNR at bay when using the estimated model on validation data.

Chapter 5

Transfer Learning

This chapter was published in [89]:

- Kelwin Fernandes and Jaime S. Cardoso. Hypothesis transfer learning based on structural model similarity. *Neural Computing and Applications*, pages 1–14, 2018

TL focuses on building better predictive models by exploiting knowledge gained in previous related tasks, being able to soften the traditional supervised learning assumption of having identical train-test distributions. Most efforts on TL consider revisiting the data from the source tasks or rely on transferring knowledge for specific models. In this chapter, a general framework is proposed for transferring knowledge by including a regularization factor based on the structural model similarity between related tasks. The proposed approach is instantiated to different models for regression, classification, ranking and recommender systems, obtaining competitive results in all of them. Also, we explore high-level concepts in TL like sparse transfer, partially-observable transfer and cross-model transfer.

5.1 Introduction

Traditionally, supervised learning focuses on building models able to generalize from labeled training instances to test instances drawn from the same distribution [257]. However, since we are living in a data-driven world that is constantly changing, domain distributions change quickly in real applications, and concepts that were valid in training time may no longer hold. Moreover, requirements, understood as the predictive task, may have changed. Thereby, classical approaches require to collect and to annotate new data and to build new models from scratch. Since the repetitive data collection and model fitting process may become rapidly intractable in real-world applications [257], it would be advantageous to transfer knowledge obtained from related problems to our target problem.

TL aims to extract knowledge from at least one source task and use it when learning a predictive model for a target task [257]. The intuition behind this idea is that learning a new task from related tasks should be easier (faster or with better solutions) than learning the target task in isolation. In this work, we focus on inductive TL, where both domains are represented by the same feature space and where the source and target tasks are different but related [257].

Pan and Yang [257] categorized previous efforts on inductive transfer into four groups depending on what is being transferred: instances, feature-representation, model parameters and relational knowledge [71]. Instance transfer consists in using data from the source task when learning the target problem [27, 63, 122, 308], usually by means of assigning different weights to the observations. Feature-representation transfer concerns on finding a shared low-dimensional feature representation that is suitable for learning the target task [12, 84, 218, 287]. We group these two approaches under the umbrella of data-driven transfer, where source data is re-used to train the target task. Although these approaches may seem appealing, the vast amount of training data in the source task turns the process prohibitively expensive. Analogously to Nearest Neighbors techniques in traditional ML, deferring the entire learning to the target-learning stage may be understood as lazy learning, i.e., deferring the actual learning until the query (target task) is made to the system. From a human-inspired point of view, this would be analogous to revisiting basic arithmetic problems when learning differential calculus or re-learning to walk when learning to run.

Thereby, TL techniques (and its community) should be focused on adapting knowledge instead of data. This idea is handled by parameter transfer approaches, which rely on the idea that individual models for related tasks should share some structure (parameters or hyperparameters) [257]. In this sense, the knowledge generated from a source task is understood as the parameters that define a given model: the coefficients of a regression, the weights of an ANN, the feature hierarchy of a DT. A few methods have been proposed on this line [13, 33, 85, 163, 197], most of them for transferring parameters for specific models: Gaussian Processes [33], SVM [85, 197], ANN [357] and ensembles [164]. Also, initialization-based models [169, 270] can be included in this group, which use the source model as an initialization for the target task optimization process. This is frequently done in ANN to promote convergence to local optima near the source model [169, 306]. This behavior can also be achieved by applying a small number of iterations in the optimization process [270, 306] or by fixing certain parameters from the source model [169, 254]. However, this scheme does not guarantee that knowledge is preserved during optimization.

HTL is a generalization of parameter transfer that has gained traction in the last few years [28, 77, 85, 163, 184, 185, 205, 267, 323, 363]. HTL assumes that knowledge is transferred directly from the source hypotheses. Experimental assessment [22, 323] as well as theoretical properties regarding the stability of these models have been addressed by several authors in the past [28, 184, 267]. However, these works assumed that transfer was

done between generalized linear models by regularizing the difference between source and target coefficients. In a more recent work [185], the problem of transferring knowledge from multiple source hypotheses with fast convergence using Regularized Empirical Risk Minimization was addressed.

In this work, we generalize the HTL framework to be able to include other learning models and types of transfer. Thereby, we propose a unified structure-transfer approach that aims to transfer knowledge by regularizing the structural distance between the target and the source model. In order to illustrate the potential and flexibility of the proposed framework, we instantiate the proposed framework to four learning tasks: regression, classification, learning to rank and recommender systems (Sect. 5.3). Also, we explore three high-level concepts in the TL area: sparse, partially observable, and cross-model transfer.

The motivation for sparse transfer relies on using almost equivalent decision processes for related tasks by sparsely updating minor details in the model. For example, when an *English Checkers* player tries to play *International Checkers*, most of the decision rules learned for playing the former version are still valid for the second version. Thereby, the effort devoted to transfer the knowledge from one game to the other is spent in learning the few new rules instead of learning slight variations of the entire set of rules. Further details about this type of transfer are presented in Sect. 5.3.1.

On the other hand, partial transfer can be understood as having limited observability of the source model. This partial observability can be defined as restricting the set of assumed parameters by the source model that are observable when fitting the target model or by limiting the source model properties that are accessible during transfer. Being able to reuse knowledge in this context allows transfer in environments where privacy and security are important. Also, by transferring high-level properties instead of low-level parameters, we can cover a wider spectrum of related tasks. We illustrate this concept in Sect. 5.3.2 and 5.3.3.

Finally, we explore in Sect. 5.3.3.2 an additional capability of the proposed framework, which relies on transferring knowledge between different types of models (e.g., Logistic Regression and DT, SVM and AdaBoost, etc.).

5.2 Transfer Learning using Structural Model Similarity

We consider the following scenario in this work. We have two learning tasks denoted by *source* and *target*. Without loss of generality, we assume that both tasks share the same feature space $X \subset \mathbb{R}^d$ and output type $Y \subset \mathbb{T}$ (e.g., regression, classification). Although this notation is an oversimplification that can be extended to other specific tasks like ranking and recommender systems, we adopt this simplistic scenario to present the method. For

a given task $T \in \{source, target\}$, we have the training data $D^T \subseteq X^T \times Y^T$. Thus, the learning objective, Eq. (5.1), is to find the best model M^* given D^T

$$M^* = \arg\max_M \Big(P(M|D^T) \Big) \tag{5.1}$$

, where M is an instance belonging to the space of models. Applying the Bayes theorem and a monotonous logarithmic transformation, Eq. (5.1) can be transformed to Eq. (5.2) with the same solution.

$$M^* = \arg\max_M \Big(log(P(D^T|M)) + log(P(M)) \Big) \tag{5.2}$$

In this sense, Eq. (5.2) can be understood as finding the model that maximizes the (weighted) tradeoff between fitting the data (*dataFitness*) and having a desired structure (*modelFitness*).

$$M^* = \arg\max_M \Big(dataFitness(M, X^T) + \lambda \; modelFitness(M) \Big), \; \lambda \geq 0 \tag{5.3}$$

In a TL context, *dataFitness* is only associated to the model performance on the target data. While in classical learning settings the *modelFitness* term gives priority to simple models, we propose to prioritize models with high similarity with the model obtained using the source data only:

$$M^* = \arg\max_M \Big(dataFitness(M, X^{target}) + \lambda \; similarity(M, M^{source}) \Big), \; \lambda \geq 0 \tag{5.4}$$

Eq. (5.4) presents a unified framework for hypothesis-transfer that can be instantiated to several predictive models given:

- A function that defines the similarity between the knowledge synthesized in the target model and the one in the source model.

- An optimization framework that allows introducing the regularization term using the structural similarity function.

The analogous minimization problem can be defined using a data-driven loss function and a model-driven dissimilarity function.

As defined in Eq. (5.5), this framework can be extended to support transfer from multiple sources $S = \{s_1, s_2, \ldots, s_n\}$ in a straightforward manner, where λ_j denotes the regularization level associated to the source task j.

$$M^* = \arg\max_{M} \; dataFitness(M, X^{target}) + \sum_{j=1}^{n} \lambda_j \; similarity(M, M^{s_j}) \qquad (5.5)$$
$$\text{where } \lambda_j \geq 0, \; \forall j \in \{1, \ldots, n\}$$

Thereby, instead of transferring data from the source task as done by previous methods in the literature, knowledge is transferred through the model structure. Since a predictive model is a succinct representation of the data, the proposed approach is an efficient way to introduce knowledge obtained from the source task without resorting to the source data. Therefore, the proposed approach is also useful in scenarios where source data is unavailable at transfer time and in online learning settings.

5.3 Instantiations and Experimental Evaluation

In this section, several instantiations of the proposed framework to different models are presented. These models explore general learning tasks usually studied in the literature: regression, classification, learning to rank and recommender systems. Moreover, we validate high-level transfer concepts in each one of them in order to prove the flexibility of the proposed framework. For instance, concepts like sparsity, partial observability of the source model and cross-model transfer are analyzed.

For readability, the experimental evaluation is presented along with the model instantiation. Also, the following baselines [69, 197] are used for comparison purposes:

- **Target-only:** the target model is learned using the target data only. This baseline is analogous to ignoring the source task and building a target model from scratch.
- **Weighting (W):** the target model is learned using a weighted combination of source and target data. The weight associated to each class is trained using nested cross-validation.
- **Extended (Ext):** the target model is learned using the target data extended with the prediction obtained by the source model. For classification tasks, the estimated probability is considered instead of the final class.

In order to avoid overfitting to the training data in these settings, all the baselines are regularized using their corresponding penalty terms (e.g., L_1, L_2).

In the experimental evaluation, data was split using a stratified training-test partition (80-20). Then, in order to validate the model performance on different stages of the data acquisition process, the training set was randomly subsampled in 10 nested subsets with several sizes $(10\%, 20\%, 30\%, \ldots, 100\%)$. Each experiment was repeated 30 times varying the test partition. For reproducibility purposes, source code and training-test

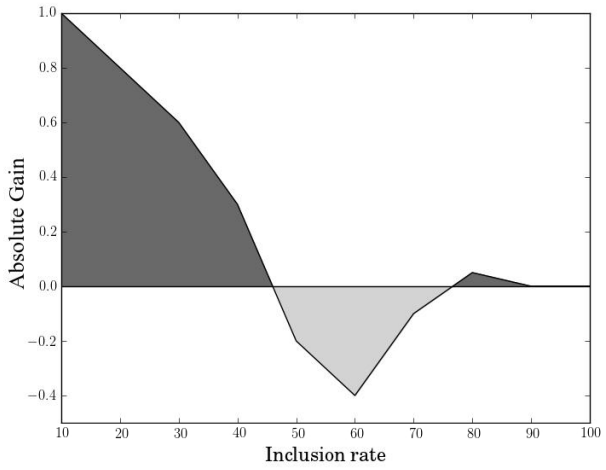


Figure 5.1: The Signed Area under the Gain Curve (sAUC) is the sum of the area of all positive transfer regions (dark areas) minus the area of the negative transfer regions (light areas).

partitions are made available[1]. The regularization factor and all the remaining intrinsic meta-parameters were learned using nested Stratified K-fold cross-validation (K = 3) over the training set. The same parameter fine-tuning scheme was conducted for all the baselines and proposed methods.

For each method, the absolute gain is measured when compared with the **Target-only**. Thus, positive gain reflects positive transfer and, analogously, negative gain reflects negative transfer. Figure 5.1 illustrates this concept, where dark regions represent positive transfer and light regions negative transfer. Many papers in the literature confine the results to a predefined training-test partition [63, 85, 164], restricting the comparison of the methods to specific stages of the data acquisition process. Other methods enumerate the performance when varying training set sizes [122, 169, 197]. In order to provide useful feedback about the actual performance of the method through the entire spectrum of data acquisition, the normalized Discounted Cumulative Gain (nDCG) was considered. nDCG is frequently used in *learning to rank* tasks to compare different rankers and, to the best of our knowledge, has not been used for assessing TL. Its adequacy to TL stands as follows. If we consider a sequence of nested training sets, the main focus of TL is to increase the performance especially on the smallest sets [352], where data is scarce. Thereby, considering the aforementioned nested training subsets, the gain obtained by considering the i-th training subset is analogous to the relevance of the item ranked at position i in a ranking setting. Wang et al. [334] show that nDCG can decide consistently the best ranker in every pair of substantially different ranking functions. Eq. (5.7) defines a continuous version of the nDCG $\in (-\infty, 100]$ over the space of percentage inclusion of training data, where $BE(x)$ and $ME(x)$ are the error of the baseline strategy and of the model of interest

[1]https://github.com/kelwinfc/transfer-learning

when considering $x\%$ of the data. ME^* is the zero constant representing the error of the best model assuming a noiseless training set. Given that it would be computationally intractable to build all possible training sets, we considered an approximation of Eq. (5.7) using the trapezoidal rule of the aforementioned partitions.

$$\mathrm{DCG}(BE, ME) = \int_0^{100} \frac{BE(x) - ME(x)}{log_2(x+1)+1} dx \tag{5.6}$$

$$\mathrm{nDCG}(BE, ME) = 100 \frac{DCG(BE, ME)}{DCG(BE, ME^*)} \tag{5.7}$$

In order to simplify the assessment of the proposed methodologies, we validate the performance of the proposed methodologies with single source-target settings.

5.3.1 Regression

In this section, we instantiate the proposed framework to the Linear Regression model. In Eq. (5.8) we adopt the well known Elastic Net (EN) loss function, where ω^s and ω^t stands for the source and target coefficients respectively and $\| \dot{\|}_p$ is the p-norm of the coefficients. In this case, the model similarity is instantiated as the distance between the target and source coefficients. In order to allow concept drift, the independent term is not regularized.

$$J_{X,y}(\theta) = \sum_{i \in N} \left(y_i - X_i^\top \cdot \omega^t \right)^2 + \lambda \left(\alpha \parallel \omega^t - \omega^s \parallel_1^1 + (1-\alpha) \parallel \omega^t - \omega^s \parallel_2^2 \right) \tag{5.8}$$

The target model ω^t can be defined in terms of the source model as $\omega^t = \omega^s + \Delta$ and, considering the residuals of the source model on the target task, $\epsilon_i = y_i - X_i^\top \cdot w^s$, the optimization objective defined in Eq. (5.8) can be rewritten as stated in Eq. (5.9). Thereby, the optimization objective is equivalent to fitting a classical regularized linear regression to the residuals.

$$J'_{X,\epsilon}(\Delta) = \sum_{i \in N} \left(\epsilon_i - X_i^\top \cdot \Delta \right)^2 + \lambda \left(\alpha \parallel \Delta \parallel_1^1 + (1-\alpha) \parallel \Delta \parallel_2^2 \right) \tag{5.9}$$

Sparse transfer is an interesting concept that can be achieved using this framework and the proper regularizer. The intuition behind this idea is that an intelligent agent should be able to reuse a decision strategy obtained from a related source task by changing a small number of details instead of updating the entire model. In this specific instantiation, such property can be obtained by using an L_0 or L_1 regularizer. Since Eq. (5.9) is agnostic about the source coefficients distribution, encouraging sparsity in the transfer stage induces sparse differences between the source and target model instead of sparse coefficients per se.

Table 5.1: Comparison of Regression models using different transfer strategies: Ridge (L_2), Lasso (L_1) and ElasticNet (EN). Performance is measured using Mean Absolute Error.

Dataset[207]	**W**	**Ext**	**Proposed**		
			L_2	**L_1**	**EN**
Automobile Gas/Diesel	6.75	3.07	**22.01**	**19.31**	**17.74**
Solar Flare M/C	**11.83**	0.14	1.67	-0.29	0.47
Parkinson Men/Women[326]	5.32	-0.42	**5.74**	-7.60	-7.48
Students P1/M1 [55]	0.30	0.29	**2.06**	0.09	0.25
Students P1/P2 [55]	**4.08**	1.00	3.83	**4.30**	**4.23**
Wine Red/White [54]	-0.60	-0.06	**0.15**	-2.50	-0.92
Wine White/Red [54]	**0.42**	-0.16	**0.52**	-0.88	-0.21

In the experimental assessment, three regularizers for the transfer step were used: Ridge ($\alpha = 0$), Lasso ($\alpha = 1$) and the general Elastic Net ($0 \leq \alpha \leq 1$).

Table 5.1 shows the results obtained in several datasets. Gain was measured in terms of decrease in the Mean Absolute Error (MAE). Hereafter, the best scores are presented in bold, as well as all statistically identical scores, using a paired difference Student's t-test with a 90% confidence level. The best results were obtained by at least one of the proposed regularization schemes on most datasets (see Table 5.1). As can be seen in Figure 5.2, the proposed strategy using L_2 normalization dominates the other curves, especially in the smallest partitions where the larger gains are achieved. As expected, while all the models achieved positive transfer on the first partitions, as we move towards the full inclusion of the training data, the gains become negative.

It is well known that, when evaluated using only prediction quality, Ridge tends to be superior to Lasso (and Elastic Net). Thus, results are aligned with this. An interesting behavior can be observed by studying the results obtained in the *Students Performance* [55] dataset, where we explored predicting the students' grades on maths (M) and Portuguese (P). In the case that knowledge was transferred between different courses in the same academic period (P1/M1) the L_2 regularizer achieved the best results. On the other hand, when knowledge was transferred between the same course but using different periods (P1/P2), a sparse transfer strategy obtained the best results. These examples validated the motivation behind sparse transfer, which focuses on changing a small subset (sparse) of properties of the model when the tasks are strongly related.

5.3.2 Classification

The proposed TL framework is instantiated to Linear SVM and to the AdaBoost classifier in this section. Although other classifiers can be adapted to this framework, these models are suitable to explore the idea concisely. For example, ANN may be regularized using the coefficients difference and DT by considering the edit distance between the source and target trees.

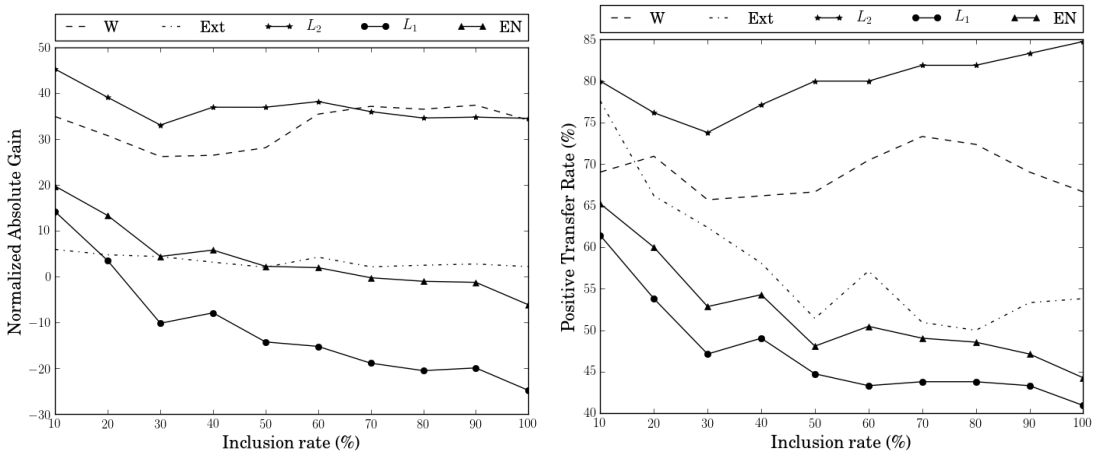


Figure 5.2: Average gains (left) and positive transfer rates (right) with nested training sets on regression tasks

Also, we explore in this section the concept of partial transfer, allowing to selectively transfer knowledge from the source model. Partial transfer can be understood as improving the model performance on the target task by using a partially observable source model. This can be done by considering regularization schemes that explore high-level properties of the model instead of its actual state (i.e., assumed values). This capability is especially important in some scenarios, where unlimited access to the model parameters is not possible due to privacy and security concerns (e.g., health and biometrics applications). In these cases just high-level properties of the model are available. Also, regularizing high-level properties of the models allows transfer between less similar tasks. Thereby, even when the source model is fully observable, it could be interesting to study partial transfer mechanisms.

5.3.2.1 Support Vector Machines

Similarly to the Linear Regression, the proposed framework can be instantiated to linear SVM considering the difference between the source and target coefficients. This idea was previously explored for Structural SVM by Lee and Jang [197] and in a multitask learning setting by Evgeniou and Pontil [84]. In both cases, the dual formulation is used. Instead, we use the soft-margin primal formulation with hinge loss (cf. Eq. 5.10) using stochastic subgradient descent [300]. Also, some authors explored this problem from a theoretical point of view to show its stability [184, 267].

$$\arg\min_{\omega^t} \frac{1}{N}\sum_{i=1}^{N}\max\left(0, 1 - y_i X_i^\top \cdot \omega^t\right) + \lambda \parallel \omega^t - \omega^s \parallel_2^2 \tag{5.10}$$

In order to validate the concept of partial transfer, we explore the idea of transferring

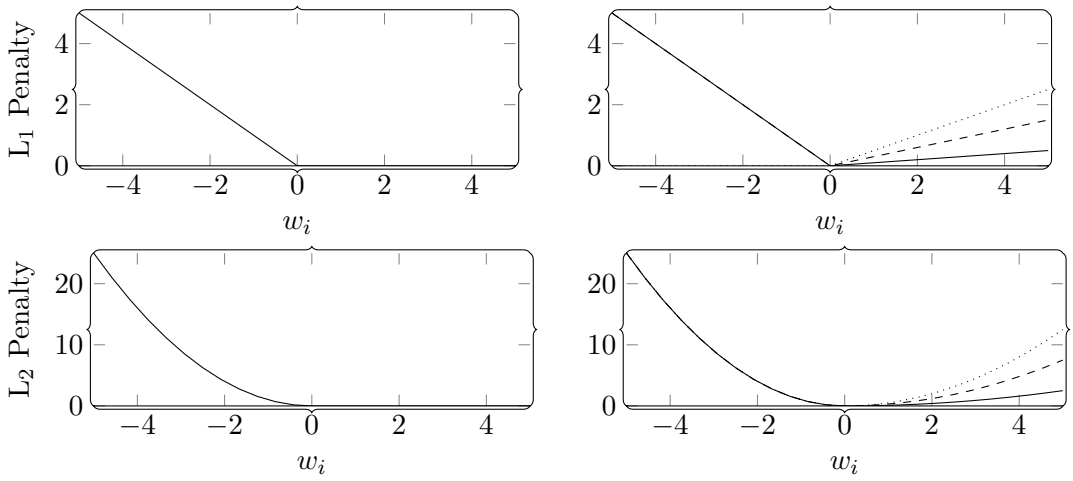


Figure 5.3: Sign regularization factors assuming $w_i^s > 0$. First row illustrates the penalization using L_1 regularizers ($p = 1$) with same-sign uncontrolled penalty on the left and with different α values on the right (0.9 - solid, 0.7 - dashed, 0.5 - dotted). Second row is analogous to the first row but using L_2 penalty ($p = 2$).

the contribution direction of each feature (i.e., coefficient sign) instead of its importance in the source task (i.e., coefficient magnitude). This type of transfer is not only pertinent in partially observable settings but also allows positive transfer between tasks that are only slightly related. Eq. (5.11) defines a way to regularize the coefficient sign [95].

$$\delta_p(\omega^t, \omega^s) = \sum_{i=1}^{d} \max(0, -\omega_i^t \cdot \text{sign}(\omega_i^s))^p \tag{5.11}$$

Although this regularizer is able to control the sign change between the source and target task, it does not establish any control on models with large coefficients with the same sign. Thereby, we include the classical Tikhonov regularization (see Eq. (5.12)). Figure 5.3 illustrates the behavior of two particular instances of the proposed regularizer with $p = 1$ and $p = 2$.

$$\Delta_{p,\alpha}(\omega_i^t, \omega_i^s) = \alpha\delta_p(\omega_i^t, \omega_i^s) + (1-\alpha) \parallel \omega^t \parallel_p^p,\ 0 \leq \alpha \leq 1 \tag{5.12}$$

The proposed regularizer is based on the Hinge loss traditionally used in the optimization of SVM. In this sense, the particular case when $p = 2$ is a smooth version that allows gradient computation on its entire domain. Thereby, it does not introduce further complexity to the loss function defined in Eq. (5.10).

On the other hand, when $p = 1$, the derivative at $\omega_i = 0$ is non-deterministic. However, the subgradient at $\omega_i = 0$ can be computed, inducing a subgradient descent optimization strategy. This type of regularization would also induce sparse transfer; a concept previously studied in Section 5.3.1. In this work, we only present results for the smooth version of the proposed regularizer.

Table 5.2: Comparison of classifiers using different transfer strategies: SVM with Structural regularization (SVM), SVM with Structural Sign regularization (S-SVM), SVM with Structural Sign-mixed regularization (αS-SVM). Performance is measured using accuracy.

Dataset [207]	**W**	**Ext**	**Proposed**		
			SVM	**S-SVM**	**αS-SVM**
Echocardiogram (fluid)	0.16	-3.76	3.77	5.22	**9.21**
Glass (RI high)	**13.48**	0.37	**2.08**	4.07	**12.86**
Hepatitis (No Histology)	9.05	-1.89	**17.54**	8.84	9.62
Car Evaluation high/med	-1.00	-1.29	1.99	**2.40**	**3.51**
Pima Indian (old)	-9.91	**2.83**	**3.32**	**6.17**	**3.58**
Contraceptive (Working)	-7.94	0.78	5.75	**9.13**	**9.84**
Ionosphere Ft. 29 (High)	**15.96**	2.29	10.02	5.57	4.87
Wine (White/Red)	-10.96	1.21	-0.46	11.76	**18.74**

Table 5.2 shows the results for SVM. The proposed schemes achieved the best results in most datasets using the proposed structural similarity transfer. Moreover, the *sign* regularization scheme obtained better results than the difference-based in several datasets. In this sense, structural regularization TL offers a competitive and efficient framework for transferring knowledge from SVM. As was validated in the experimental evaluation, considering partial-transfer schemes improves the transfer gains between more dissimilar tasks. For example, comparing the gain obtained in the *Wine* dataset by the partial transfer strategy was higher than in the *Hepatitis* dataset. This suggests that predicting the life expectancy of patients with and without histology is more related than predicting the quality of different types of wine. Thereby, using a strong regularization with full observability achieves the best performance in the latter while using a more flexible regularization (partial observability) achieves the best performance in the former.

For this instantiation, the gain achieved by the models as we collect more data does not decrease. In general, we may observe that the gains achieved by the models with partial observability dominate the other curves (see Figure 5.4). Thereby, it was validated the relevance of transferring partial knowledge instead of promoting low-level similarity between source and target models.

5.3.2.2 AdaBoost

In this case, we instantiate the proposed framework to the Discrete AdaBoost model [114]. As typical, we used unidimensional decision thresholds as weak learners. However, the concepts explored in this section can be easily extended to other types of estimators.

In the AdaBoost model, two type of concepts can be transferred from a source model: the weak estimators and their associated importance. We regularized the weak estimators by encouraging similar decision thresholds, considering that the target model can probabilistically choose a learner from the pool of source weak learners or can create a new estimator from scratch. On the other hand, in order to regularize the relative importance

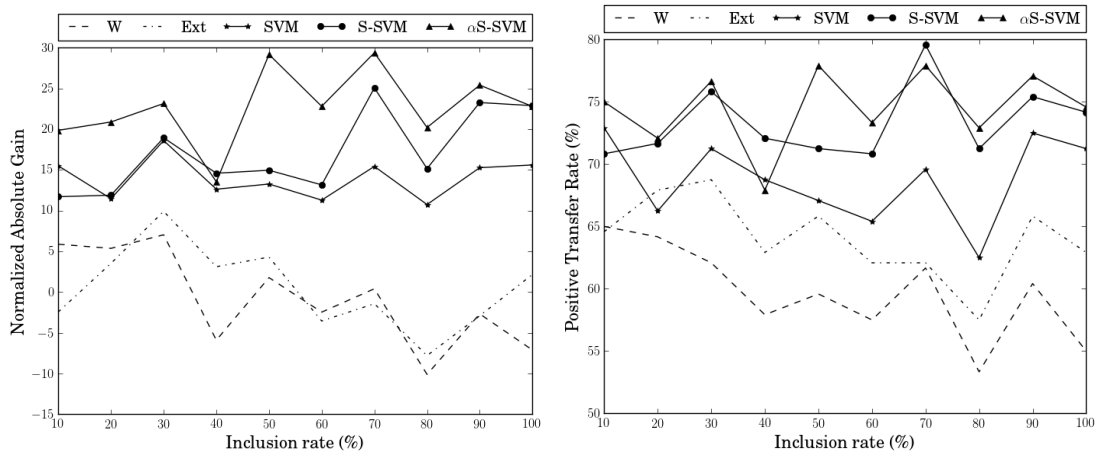


Figure 5.4: Average gains (left) and positive transfer rates (right) with nested training sets on classification tasks using SVM.

of each estimator, we encourage closeness between the iteration at which each estimator was chosen in the source and target tasks. This type of regularization has the secondary advantage of promoting similar updates to the weight distribution associated with the training set in both, the source and the target task.

In this sense, at each iteration of the AdaBoost training algorithm, we select the estimator (f,t,d,i) that minimizes the tradeoff between the exponential loss, traditionally used in AdaBoost, and the regularizer defined in Eq. (5.13), where f is the feature of interest, t is the threshold value, $d \in \{-1,+1\}$ is the estimator output when the thresholding condition is satisfied, i is the iteration where the estimator was included in the ensemble and, N is the maximum number of estimators.

$$D(e,\text{pool}) = \underset{p \in \text{pool}}{\arg\min} \left(\alpha D_{thrs}(e,p) + (1-\alpha) D_{order}(e,p)\right) \tag{5.13}$$

$$D_{thrs}((f^t,t^t,d^t,i^t),(f^s,t^s,d^s,i^s)) = \begin{cases} |t^t - t^s| & , \text{ if } f^t = f^s \ \wedge \ d^t = d^s \\ 1 & , \text{ otherwise} \end{cases}$$

$$D_{order}((f^t,t^t,d^t,i^t),(f^s,t^s,d^s,i^s)) = \begin{cases} \frac{|i^t - i^s|}{\text{N}} & , \text{ if } f^t = f^s \ \wedge \ d^t = d^s \\ 1 & , \text{ otherwise} \end{cases}$$

Given that D_{thrs} denotes the similarity between the decision thresholds in the source and target hypotheses and D_{order} denotes the similarity between the feature relevances, the α parameter controls the model observability. By setting $\alpha = 1$, we will observe the decision thresholds and ignore the importance. Conversely, using $\alpha = 0$ will ignore the thresholds but will encourage the target model to choose the features in a similar order. In the experimental assessment, we considered models with 50 estimators. Table 5.3 shows the results for the proposed regularizers. While the proposed strategy achieved positive

Table 5.3: Comparison of classifiers using different transfer strategies: AdaBoost with observable thresholds and order (Full), AdaBoost with observable thresholds (Thres) and Adaboost with observable order (Order). Performance is measured using accuracy.

Dataset [207]	**W**	**Ext**	**Proposed**		
			Full	**Thres**	**Order**
Echocardiogram (fluid)	**10.55**	-0.28	-0.07	2.38	-4.53
Glass (RI high)	**4.04**	**3.60**	**2.63**	**2.14**	**-5.56**
Hepatitis (No Histology)	-12.66	**-2.76**	**-4.36**	**-6.70**	-10.96
Car Evaluation high/med	-37.81	**17.83**	2.51	**19.61**	17.56
Pima Indian (old)	**0.32**	-4.03	**-0.06**	**-1.77**	-2.87
Contraceptive (Working)	**14.90**	0.63	1.00	4.74	2.72
Ionosphere Ft. 29 (High)	**28.66**	12.45	5.13	16.47	13.13
Wine (White/Red)	-0.38	-0.54	**0.30**	-0.74	-0.26

transfer in most cases, the weighting strategy achieved the larger gains in the smallest partitions on average (see Figure 5.5).

Partial observability of the decision thresholds obtained better performance than transferring the selection order of the weak estimators.

5.3.3 Learning to Rank

Learning to Rank in combinatorial domains has become a trendy topic in recent years due to the growing number of applications involving the prediction of structured preference data. Examples of applications where predicting rankings is crucial are found in information retrieval (e.g., search engines) and recommender systems. Learning to rank strategies can be categorized according to their input type into pointwise, pairwise and listwise techniques. In this section, we consider pairwise rankers, which rely on deciding which observation, if any, is better in a given pair.

5.3.3.1 Lexicographic Orders

Here, we instantiate the proposed TL framework to lexicographic orders [109], which compactly express the order between any pair of observations. Instantiating other ranking models like RankSVM [150] is very straightforward using the techniques explored in previous sections. In order to simplify the presentation of the structural similarity function between lexicographic orders, we limit the scope of this work to unconditional/linear LxO – (e.g., LexRank [109]) with binary features. LxO can be understood as a total order of the attributes and of their respective values. Thereby, given a ranking task with D binary attributes, a LxO model M can be understood as a pair $M =< A, V >$ where $A : \mathbb{N}^{\leq D} \rightarrow \mathbb{N}^{\leq D}$ is a bijective function that indicates the relevance of each feature and $V : \mathbb{N}^{\leq D} \rightarrow \mathbb{B}$ is a function that defines the preferred value for a given feature.

Figure 5.6 illustrates an instance of a linear Lexicographic Ranker with three features: main course, drink and dessert. The attribute domains are {meat, fish}, {wine, water}

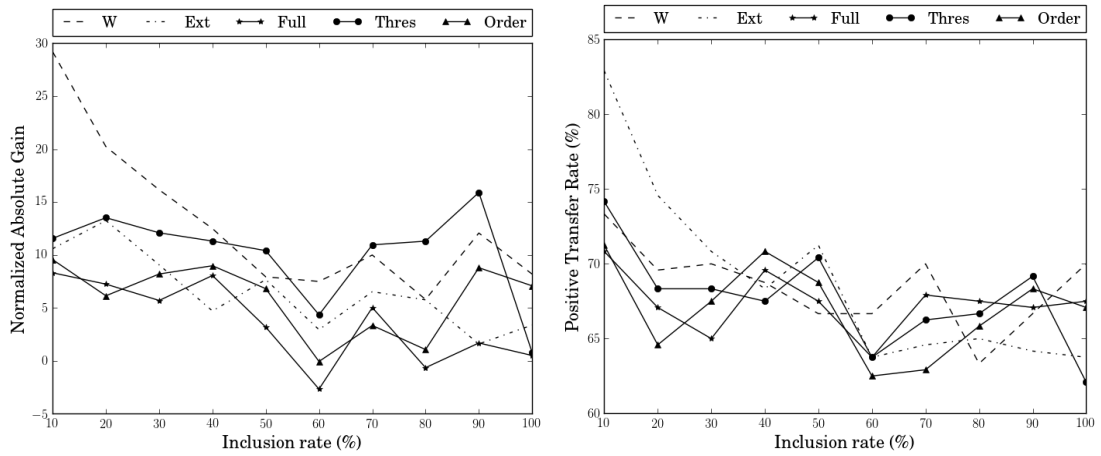


Figure 5.5: Average gains (left) and positive transfer rates (right) with nested training sets on classification tasks using AdaBoost

and {cake, pie} respectively. To predict the ordering of two options using such model, the two observations are compared through the model in a cascade manner (using the feature relevance), until they differ in a given feature. The order direction is dictated by the preferred value for that feature. For instance, using the model illustrated in Figure 5.6, the following is a valid ordering of options:

$$(\text{meat}, \text{wine}, \text{cake}) \sqsupset (\text{meat}, \text{wine}, \text{pie}) \sqsupset (\text{meat}, \text{water}, \text{pie}) \sqsupset (\text{fish}, \text{wine}, \text{cake})$$

Linear LxO are of interest due to their high interpretability. Despite the existence of lexicographic rankers with higher expressiveness, we limit the scope of this work to this type of ranker to simplify the regularizer definition. The ideas explored in this section can be extended to conditional LxO [97].

We define the distance between two LxO as the weighted sum of the normalized Kendall tau distance between the attribute ordering and the number of attributes with different preferred values in Eqs. (5.14)-(5.16). This distance can be extended to conditional LxO [34, 98] by considering the edit distance between trees instead of the Kendall tau distance.

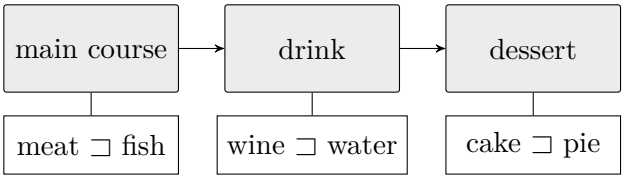


Figure 5.6: Illustration of a unconditional Lexicographic Ranker with three attributes

$$dist_\alpha(\langle A^s, V^s\rangle, \langle A^t, V^t\rangle) = \alpha K(A^s, A^t) + (1-\alpha)P(V^s, V^t) \tag{5.14}$$

$$K(A^s, A^t) = \binom{D}{2}^{-1} \sum_{1\le i<j\le D} [A^s(i) < A^s(j) \not\equiv A^t(i) < A^t(j)] \tag{5.15}$$

$$P(V^s, V^t) = \frac{1}{D}\sum_{i=1}^{D} [V^s(i) \not\equiv V^t(i)] \tag{5.16}$$

Given the discrete nature of LxO, greedy algorithms have been used in the literature to obtain models fitted to data [109]. In our experimental evaluation, the regularization term is introduced as part of the objective function in a local search strategy. The neighborhood is defined by all possible swaps of consecutive attribute pairs and by changing the preferred value of each feature. A first-best approach was conducted for choosing the next neighbor to be expanded. Table 5.4 shows the results obtained for this task. Since local search rapidly converges to local optima, two independent runs were executed starting from different initial solutions. These solutions were generated using the greedy LexRank algorithm proposed by Flach and Matsubara [109] on the source and target data separately. Besides the instantiation with full-knowledge transfer, which was denoted in Table 5.4 as Comb ($\alpha = 0.5$), two instances with partial observability of the model structure were considered: Priorities ($\alpha = 1$) and Preferences ($\alpha = 0$). Performance is measured in terms of correctness [47] (see Eq. (5.17)), which considers the balance between concordant (C) and discordant (D) predicted pairs. As was observed with the classification models, using partial transfer improved the model performance in most datasets. Moreover, as can be seen in Figure 5.7 the gains achieved by the proposed strategies are consistently higher than the ones achieved by the other methods in the literature, being able to achieve positive transfer in more than 80% of the cases.

$$CR(\sqsupset, \sqsupset_*) = \frac{|C| - |D|}{|C| + |D|} \tag{5.17}$$

5.3.3.2 Cross-model Transfer: from RankSVM to Lexicographic Orders

In this section we explore another capability of the proposed transfer framework: transferring knowledge between models with different nature. In order to do this, we can use a regularizer that relies on high-level structural properties of the model instead of model specific parameters. We explored some intuitions behind this idea in the sign regularization for the SVM. In this section, we will transfer information from the RankSVM model [150] to LxO. We can use linear SVM in the context of rankings by transforming the decision function $f(a) < f(b)$ into $g(a-b) > 0$. In this sense, the final linear SVM model will induce a decision boundary defined by $\omega^\top(a-b) > 0$. Then, for binary variables, the absolute-valued magnitude of each coefficient can be understood as the feature relevance

Table 5.4: Comparison of Ranking models using different transfer strategies: Priorities (Prior), Preferences (Pref) and Combined (Comb). Performance is measured using correctness.

Dataset[207]	**W**	**Ext**	**Proposed**					
			LexRank-LexRank			**RankSVM-LexRank**		
			Prior	**Pref**	**Comb**	**Prior**	**Pref**	**Comb**
Lenses (Hyper./Myope)	**37.38**	33.87	**42.71**	33.71	32.24	11.37	12.39	-2.24
T.A Regular/Summer	11.77	10.81	**17.53**	12.68	7.40	11.73	8.30	9.01
Acute Infl Urin./Renal	28.19	12.32	28.08	28.08	**31.08**	28.08	28.08	**30.94**
Servo A/C	1.83	-4.96	13.97	11.39	**20.49**	7.67	-5.27	7.56
Mammographic (Old)	3.44	1.49	1.70	0.19	1.67	**9.34**	7.91	**9.18**
Contraceptive/Std. Liv	-7.50	-1.26	**-0.17**	-0.45	-0.48	-0.91	-1.02	-1.07

and the coefficient sign as the preferred value of each feature in the lexicographic orders. Thereby, we can use the regularizer formalized in Eq. (5.20) to transfer knowledge from RankSVM to linear Lexicographic Rankers.

$$dist_\alpha(\omega^s, \langle A^t, V^t \rangle) = \alpha K(\omega^s, A^t) + (1-\alpha) P(\omega^s, V^t) \tag{5.18}$$

$$K(\omega^s, A^t) = \binom{D}{2}^{-1} \sum_{1 \le i < j \le D} [|\omega_i^s| > |\omega_j^s| \not\equiv A^t(i) < A^t(j)] \tag{5.19}$$

$$P(\omega^s, V^t) = \frac{1}{D} \sum_{i=1}^{D} [(\text{sign}(\omega_i^s)) > 0) \not\equiv V^t(i)] \tag{5.20}$$

Table 5.4 shows the behavior of the cross-model transfer between these two models. As can be seen in the results, the proposed framework was able to achieve competitive results, obtaining correctness values similar to other traditional techniques and even better results in some datasets. Although the performance gain is not transversal to the entire set of problems used for validation, it was shown that using regularization on high-level structural properties of the models were able to transfer knowledge even between highly dissimilar learning paradigms.

This idea can also be explored in other predictive tasks (e.g., classification, regression) and between other models. For example, we can transfer the thresholds decided by a DT as the weak estimators used in AdaBoost, the features chosen by a sparse generalized linear model to the probabilities of including each feature in a RF, etc.

5.3.4 Recommender Systems

Collaborative filtering is a frequent paradigm in Recommender Systems based on the idea of using preferences from many users to guide predictions about a given user's preferences, conversely, for items. Given N users and M items, Matrix Factorization is a type of collaborative filtering technique that approximates the preference matrix $R \in \mathbb{R}^{N \times M}$

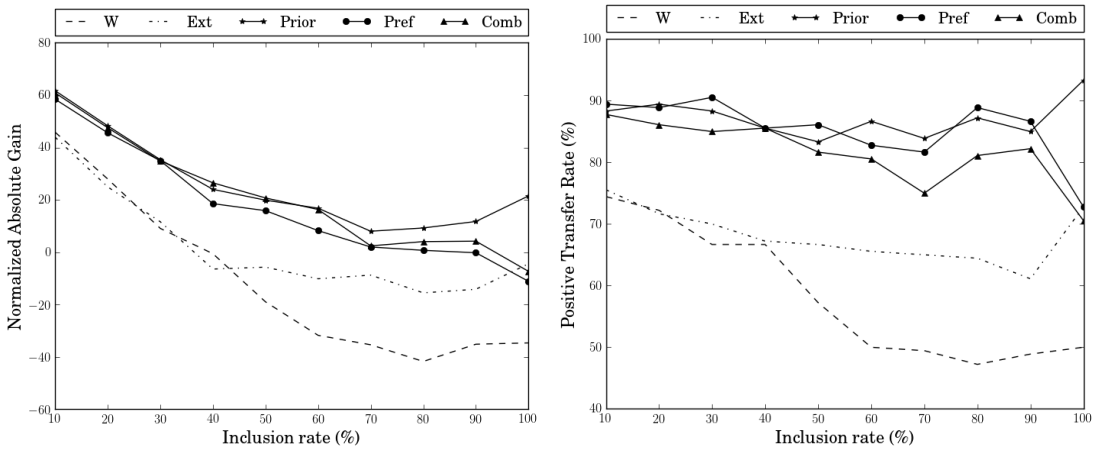


Figure 5.7: Average gains (left) and positive transfer rates (right) with nested training sets on ranking tasks using LexRank

by combining two matrices $U \in \mathbb{R}^{N \times D}$, $V \in \mathbb{R}^{M \times D}$, where D is a small number of unobserved factors that model user and items preferences, U and V respectively [236]. As typical, we consider the combination $R = U \cdot V^{\top}$. Salakhutdinov and Mnih [236] proposed Probabilistic Matrix Factorization, a method for fitting these latent factors by means of minimizing the regularized sum-of-squared-errors (see Eq. (5.21)), where $\| \cdot \|^2_{Fro}$ denotes the Frobenius norm and I_{ij} equals 1 if user i rated item j and equals 0 otherwise.

$$J(U,V) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{M} I_{ij} (R_{ij} - U_i V_j^{\top})^2 + \frac{\lambda_U}{2} \sum_{i=1}^{N} \|U_i\|^2_{Fro} + \frac{\lambda_V}{2} \sum_{j=1}^{M} \|V_i\|^2_{Fro} \tag{5.21}$$

A local minimum of J can be found using gradient descent. A well-known problem in Recommender Systems is the cold-start problem [208], which can be understood as the impossibility of producing accurate predictions for users (or items) with scarce information. This problem has been tackled in the past by introducing content information, using some priors when initializing the latent features of an entity, among others. In general, this problem can be understood as transferring knowledge from existing users to new users. In this work, we instantiate the proposed TL framework for solving the cold-start problem. Given k new users, the fitted unobserved factors for a given user U_i are regularized in order to be similar to its most similar previously fitted user U_i^* (see Eq. (5.22)).

$$J'(U) = \frac{1}{2} \sum_{i=N+1}^{N+k} \sum_{j=1}^{M} I_{ij} (R_{ij} - U_i V_j^{\top})^2 + \frac{\lambda_U}{2} \sum_{i=N+1}^{N+k} \|U_i - U_i^*\|^2_{Fro} \tag{5.22}$$

In order to simplify the computation of the most similar user, two variations of the proposed idea were considered: a subset of candidate users obtained using K-means (K=10) and a unique central user with the averaged latent features of the previously trained users.

Table 5.5: Comparison of Recommender Systems using different transfer strategies: Structural with a unique central user (Global) and Structural with a subset of candidate users (Subset). Performance is measured using Mean Absolute Error.

Dataset	**W**	**Ext**	**Proposed**	
			Global	**Subset**
Movielens100k [142]	9.08	4.52	12.06	**12.59**
Amazon Instant Video [229]	5.67	-0.11	6.08	**6.64**
Amazon Musical Instruments [229]	2.45	**6.30**	3.88	5.12
Amazon Videogames [229]	9.27	0.22	10.12	**11.05**
Jester2+ [124]	2.92	-1.72	12.46	**19.40**

Applying the same ideas explored in Linear Regression (cf. Section (5.3.1)), the problem can be formulated in terms of the target residuals (Eqs. (5.23)).

$$J''(U) = \frac{1}{2}\sum_{i=N+1}^{N+k}\sum_{j=1}^{M} I_{ij}(\hat{R}_{ij} - \Delta_i V_j^\top)^2 + \frac{\lambda_U}{2}\sum_{i=N+1}^{N+k}\|\Delta_i\|_{Fro}^2 \tag{5.23}$$
$$\text{where } \hat{R}_{ij} = R_{ij} - U_i^* V_j^\top$$

In the experimental evaluation, the Extended baseline was modeled by interpolating the average ratings for the specified item and the predicted ratings. All experiments were executed using $D = 50$ latent factors and $\lambda_U = \lambda_V \in [10^{-3}, \ldots, 10^3]$. The users considered for transfer were the top 100 users with more votes in order to validate the widest spectrum of known ratings. As can be seen in Table 5.5, the proposed transfer schemes obtained the best results in most cases. An interesting property on the results that wasn't observed in previous cases is that gains achieved by our model increase through most of the spectrum of inclusion rates while the gains achieved by the other strategies saturate and decrease drastically after a given point (60% of inclusion rate). Moreover, the rate of cases with positive transfer using the proposed strategy is close to 100% (see Figure 5.8).

5.3.5 Discussion

The proposed generic strategies achieved good performance when compared to traditional transfer strategies (see Table 5.6). For example, in more than 76% of the cases, the HTL techniques achieved better performance than their literature counterparts in at least half of the datasets. In general, at least one of the HTL-based strategies performed better than the alternative approaches from the literature. Moreover, the proposed methodologies tend to dominate the other approaches when data is scarce which is one of them main goals of TL (see Figures 5.2, 5.4, 5.5, 5.7 and 5.8).

The optimal performance of the gain curves should be a monotonically decreasing curve, where the gains achieved by using TL are high when data is scarce and tend to zero as we add data to the training set. However, given that we measure the performance of

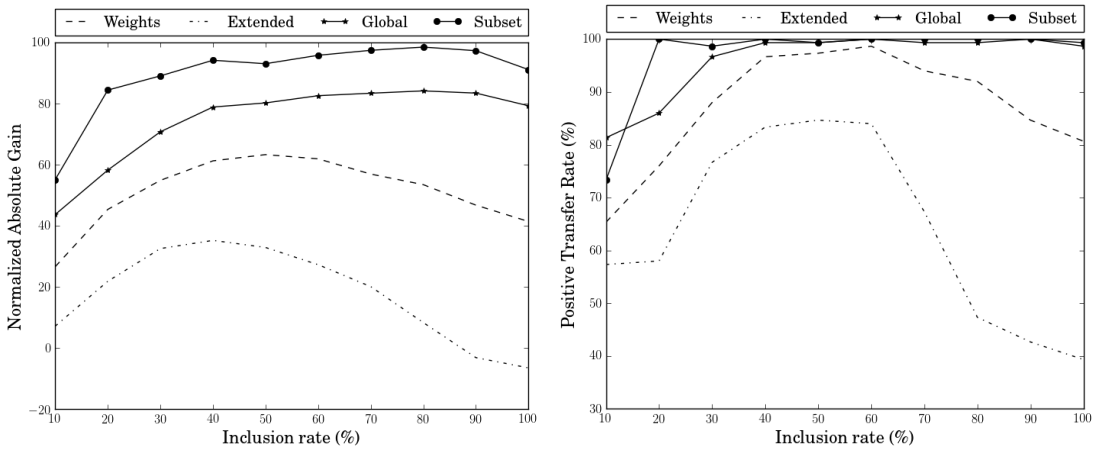


Figure 5.8: Average gains (left) and positive transfer rates (right) with nested training sets on Recommender Systems

the model on a small subset of partitions (30 runs), it is expectable to observe an irregular behavior.

We focused on providing a general framework that may be instantiated to achieve good performance in a wide diversity of scenarios. As in traditional ML settings where the best model is unknown a priori, finding the best regularizer, its observability and the regularization strength (λ) are application-dependent problems which can be solved – in general – using cross-validation. Moreover, application knowledge can be used to conduct this selection.

5.4 Conclusions

In this work, we presented a new TL framework based on structural model regularization. In contrast to most TL techniques, which either transfer data or are designed for specific models, the proposed framework addresses the problem of transferring knowledge in a general way. Namely, knowledge is transferred by including a regularization term that measures the structural similarity between source and target models. Thereby, the proposed method is able to reuse knowledge gained from the source task without revisiting source data, which might be prohibitively large or even unavailable at transfer time. In order to show its flexibility, the proposed framework was instantiated to several learning tasks: regression, classification, learning to rank and recommender systems. Positive results were obtained in most experiments, being competitive with other methods in the literature both, in terms of predictive performance and in terms of computational cost. Furthermore, key problems like sparse, partial and cross-model transfer were analyzed and assessed, showing their adequacy in several scenarios. The proposed method relies on defining a good relatedness measure between models, which may allow the integration of application-specific knowledge.

Table 5.6: Overview of the performance of the proposed strategies. The table summarizes the number of datasets (%) where each proposed strategy achieved an average behavior better than the literature baselines. The cases where the proposed techniques performed better than the baselines are presented in bold.

Task	Model	Type	W	Ext
Regression	**L_2**	Full	**71**	**100**
	L_1	Full, Sparse	29	29
	EN	Full	29	43
Classification	**SVM**	Full	**75**	**88**
	S-SVM	Partial	**62**	**100**
	αS-SVM	Partial	**75**	**100**
	AdaBoost-Full	Full	38	**50**
	AdaBoost-Thres	Partial	25	**62**
	AdaBoost-Order	Partial	38	**50**
Ranking	**LexRank-LexRank-Prior**	Partial	**67**	**83**
	LexRank-LexRank-Pref	Partial	**50**	**67**
	LexRank-LexRank-Comb	Full	**50**	**67**
	RankSVM-LexRank-Prior	Cross-model, Partial	**50**	**83**
	RankSVM-LexRank-Pref	Cross-model, Partial	33	**50**
	RankSVM-LexRank-Comb	Cross-model, Full	**67**	**67**
RecSys	**Global**	Partial	**100**	**80**
	Subset	Partial	**100**	**80**

As future work, it is relevant to evaluate the performance of the proposed methodology with multiple source tasks and with multiple similarity functions, enabling the user to specify several alternatives to embed the desired knowledge in the learning process. While this could be done in a straightforward manner using weighted regularization terms, the empirical study of this problem is relevant.

TL research line should move towards a deep understanding of how models encode knowledge and how to transfer this knowledge in a general and unified way. Through this chapter, we explored how this can be done efficiently. Emerging regularization schemes that favor this kind of transfer are feasible paths to explore, as well as similarity learning techniques able to infer the actual relatedness between models for a specific task.

Chapter 6

Directional Classification

This chapter was published in [88]:

- Kelwin Fernandes and Jaime S. Cardoso. Discriminative directional classifiers. *Neurocomputing*, 207:141–149, 2016

In different areas of knowledge, phenomena are represented by directional -angular or periodic- data; from wind direction and geographical coordinates to time references like days of the week or months of the calendar. These values are usually represented in a linear scale, and restricted to a given range (e.g. $[0, 2\pi)$), hiding the real nature of this information. Therefore, dealing with directional data requires special methods. So far, the design of classifiers for periodic variables adopts a generative approach based on the usage of the von Mises distribution or variants. Since for nonperiodic variables state of the art approaches are based on nongenerative methods, it is pertinent to investigate the suitability of other approaches for periodic variables. We propose a discriminative dLR model able to deal with angular data, which does not make any assumption on the data distribution. Also, we study the expressiveness of this model for any number of features. Finally, we validate our model against the previously proposed directional naïve Bayes approach and against an SVM with a directional Radial Basis Function (RBF) kernel with synthetic and real data obtaining competitive results.

6.1 Introduction

Several phenomena and concepts in real life applications are represented by angular data or, as is referred in the literature, directional data. Some examples of directional information are the wind direction as analyzed by meteorologists, magnetic fields in rocks studied by geologists, geographic coordinates, among others [222]. Also, some entities are usually referenced in an angular manner; gynecologists denote the location to perform a biopsy, when performing a colposcopic screening, using the angle formed by the vertical axis of the

cervix. Another example can be found in the area of CV, where color is often defined in cylindrical spaces like the Hue-Staturation-Value (HSV) color space. However, directional information is not constrained to scientific contexts; on a daily basis, we naturally use angular variables. For example, time is usually represented by hours, days of the week, the day of the month, season, etc. This reference system is cyclic by nature.

Directional variables are usually encoded as a periodic value in a given range (e.g., $[0, 2\pi)$, $[0°, 360°)$). This work focuses merely in this representation of directionality, where an angular variable is a real-valued number with periodicity defined by a range. However, directional data can also be found in other representations, such as discrete categorical values ordered by a circular relation [73]. Also, some literature makes use of histograms which lie in a circular space instead of the linear one.

Working effectively with directional data requires dealing with techniques that are aware of the angular nature of the information [222]. For example, 0 and 2π are indeed the same angle, and their average is not π but 0. In this sense, directional statistics concerns the problems derived from using traditional linear statistics with this type of data [222]. Even visualization of this type of data requires different representations to illustrate its periodic behavior (e.g., rose diagrams and circular histograms). In order to formalize the definition of a directional function, consider the predicate *dir* defined in the Eq. (6.1), where $\mathbb{N}$ is the set of integers and $\mathbb{B} = \{\text{true}, \text{false}\}$.

$$\begin{aligned} &dir : \mathbb{N} \longrightarrow \mathbb{B} \\ &dir(i) = \text{true}, \quad \text{iff the } i\text{-th feature is directional} \end{aligned} \tag{6.1}$$

We will say that the function f, with domain in $\mathbb{R}^n$, is directional with period $\vec{P}$ (i.e. the feature in the position i has period $\vec{P}_i$), if and only if the Eq. (6.2) holds, where non-directional features are assumed to have infinite period (i.e. $\neg dir(i) \Rightarrow \vec{P}_i = \infty^+$).

$$f(\vec{\theta}) = f(\vec{\theta} + \vec{k} \circ \vec{P}), \qquad \vec{k} \in \mathbb{Z}^n \tag{6.2}$$

Here on, we will restrict the periodicity of the directional values to $P_i = 1$, without loss of generality.

Supervised learning can be understood as the process of learning a function f based on so-called training data that comprises examples of the input vectors and their corresponding target values [32]. In this work, we are interested in the learning task known as classification, where the target can take a finite number of values. These values are usually denoted as classes or labels, and the input vector defines a set of features that describe objects in the domain of the function. As the result of a supervised classification task, we obtain a classifier, which is used to assign a class to an object that has not been seen at

the training stage. The ability to correctly label new instances is known as generalization [32]. Traditional models that do not take into account directionality may suffer a drop of generalization in areas near to the period of the function. Furthermore, the function may return different decisions for different $\Delta + \vec{k} \circ \vec{P}$, $\vec{k} \in \mathbb{Z}^n$, and a fixed $\Delta \in \mathbb{R}^n$, despite all of them semantically represent the same angle.

In this work we propose a binary classifier aware of the directional constraint. The rest of this chapter is organized as follows. Section 6.2 describes related work in the area of directional statistics and learning. Sections 6.3, 6.4 and 6.5 detail the proposed model, its expressiveness and the optimization strategy, respectively. Section 6.6 summarizes the performed experiments to assess the relevance of the proposed model and, finally, Section 6.7 summarizes some conclusions and future work.

6.2 Related work

Most different types of problems and approaches in ML can be broadly defined as a classification, regression or clustering tasks. Classification and Regression are the most common supervised learning tasks. On the other hand, clustering is probably the best known unsupervised learning task, where the objective is to group data into non-predefined categories based on some similarity criterion.

Previous attempts to address learning tasks with directional data have been carried out in each of the aforementioned areas. Most of them take advantage of circular distributions (such as von Mises and von Mises-Fisher). For instance, Banarjee et al. [25] proposed a generative mixture-model approach for clustering directional data using the von Mises-Fisher distribution. Moreover, they conclude that the Spherical k-means is a particular case of the mixture of von Mises-Fisher model. Fitting mixtures of angular distributions have been separately studied by Mooney et al. [238] and Mardia et al. [225].

Regression scenarios with directional data have been studied in several contexts [108, 173, 339]. Xu and Schoenberg [339] proposed a kernel regression method based on the von Mises distribution. Their method was used to discover the relationship between a single directional explanatory variable (wind direction) and a real-valued linear response variable (total area burned per day in wildfires). Fisher and Lee [108] studied the regression problem where the predictive variables are linear and the model outcome is directional. Their work also assumes that angular observations follow von Mises distributions and focuses on the estimation of the distribution parameters. Finally, Kato et al. [173] addressed the circular-circular problem, wherein both, predictive and target observations, have a circular nature.

Circular ordinal regression is an intermediate problem in this area, which lies between regression and classification. It considers a discrete number of labels which preserve a certain circular order. Devlaminck et al. [73] proposed two methods to solve this problem. The first one is an SVM variation, and the second method transforms the circular ordinal

regression problem into multiclass classification. However, the directionality concerns in [73] are focused on the model outcome rather than on the feature space.

In the area of directional classification, different approaches have been considered: from Discriminant Analysis [105, 106] to generative models [221, 222, 361]. SenGupta and Roy [297] proposed a distance-based classification rule using the chord-length between two points on the circle to classify unidimensional data. In more recent work, SenGupta and Ugwuowo [298] developed a multidimensional method for binary classification using directional data; they studied data on a torus (two directional variables) and cylinder (one linear variable and one directional variable). Their approach has the limitation that it assumes as known the probabilities of misclassification [298].

Kirby and Miranda [179] proposed a variation on the classic feed-forward ANN by including the notion of a circular node, able to store and transmit angular information. In fact, their node is an abstraction for the combination of a pair of coupled nodes, whose combined values are constrained to lie on the unit circle. However, their solution is not invariant to the same inputs at different periods. Namely, a pair of coupled nodes may return different responses to the same angular input. Furthermore, their model requires defining the hybrid architecture manually.

Finally, adaptations to generative models were studied in the past. First, Zemel et al. [361] extended the Boltzmann machine to consider cyclic units. On the other hand, López et al. proposed a directional naïve Bayes formulation [221, 222]. Their contribution involves using the von Mises and von Mises-Fisher distributions for the directional variables instead of the classic Gaussian distribution. The effectiveness of this method relies on the independence assumption of the features and the adequacy of the von Mises distribution to model the behavior of the directional features.

In this work, we propose a dLR, the discriminative counterpart to the Naïve Bayes model, which does not make assumptions on the distribution of the input data.

6.3 Directional Logistic Regression

Generative classifiers aim to model the joint probability $p(x, y)$, where x and y respectively denote the input and output variables. Traditional generative models would then make their predictions by choosing the label y that maximizes $p(x, y)$, computed using Bayes rules [244]. Instead, discriminative classifiers model the posterior probability $p(y|x)$. This computation is done in a direct manner or by learning a map from inputs x to the class labels [244].

As we have shown in Section 6.2, previous attempts to design classifiers for periodic data adopted a generative approach based on the von Mises distribution or variants [222]. Since state-of-the-art approaches are based on non-generative methods for non-periodic variables [219], in this work we propose a discriminant approach to classify directional data. Our contribution stands as a directional-aware version of the Logistic Regression

(LR) [230], which is the discriminant counterpart of the naïve Bayes classifier, previously used to address this problem. This relation is known as a Generative-Discriminative pair [244].

Eq. (6.3) defines the dLR model. This model can be understood as a LR with a mapping from the original angular space to a linear one. As we show in the Section 6.5, this mapping is learned simultaneously with the feature coefficients. Hereinafter, the two possible labels belong to $\{0,1\}$, and n is the number of features.

$$
\begin{aligned}
f(\theta) &= \frac{1}{1+e^{-k\cdot h(\theta)}} \\
h(\theta) &= \omega_0 + \sum_{i=1}^{n} \omega_i g_i(\theta_i) \\
g_i(\theta_i) &= \begin{cases} \sin(2\pi(\theta_i+\varphi_i)) & ,\text{if } dir(i) \\ \theta_i & ,\text{otherwise.} \end{cases}
\end{aligned}
\tag{6.3}
$$

This model is a hybrid approach to LR for modeling linear and directional data, whereby a mapping from angular variables to linear space is learned. The number of parameters involved in the proposed model is

$$n+1+(\#i \in \mathbb{N}^+ |\ i \leq n : dir(i))$$

If all the variables are linear, the model is reduced to the traditional LR with $n+1$ parameters. Also, we have included an extra k parameter that defines the slope of the sigmoid function, which does not change the predicted label but softens the decision boundary. Given the properties of the sine function, the model holds the directional condition.

6.4 Expressiveness of the Model

In this section we analyze the model's expressiveness by studying the induced boundaries, as was done by López et al. [222] for the von Mises naïve Bayes model. We start with the scenario where the feature space is constrained to one directional feature (Section 6.4.1). Section 6.4.2 presents the most general scenario with an unconstrained number of directional and linear features. As previously mentioned, when all variables are linear, the model becomes a classical LR and the subsequent decision surface is a hyperplane in the $\mathbb{R}^n$ space. Therefore, we are interested in settings where at least one variable is directional.

6.4.1 One-dimensional feature space with one angular variable

In this section we show the expressiveness of the dLR for the trivial case of unidimensional problems with a single angular variable. As we show below, it is easier to reason about

the expressiveness of the model in the equivalent space where each variable is transformed into a pair of coordinates in a $(0,0)$-centered unit 2-sphere, where $x_i = \cos(2\pi\theta_i)$ and $y_i = \sin(2\pi\theta_i)$. This space will hereafter be referred to as the *transformed space* or *extended space*, while the original data representation will be denoted as the *original space*.

Without loss of generality, we assume that the model classifies an instance as positive if its outcome is larger than 0.5, thus leaving the final decision to the sign of the h function.

Theorem 1. *The dLR classifier with one predictive directional variable induces a separation boundary equivalent to a two dimensional line in the transformed space. Moreover, the set of induced decision lines is complete in the space of two dimensional lines.*

Proof.

$$
\begin{aligned}
& h(\theta) = 0 \\
\equiv\ & \langle \text{Definition of } h \rangle \\
& \omega_0 + \omega_1 \sin\left(2\pi(\theta_1 + \varphi_1)\right) = 0 \\
\equiv\ & \langle \text{Sum of two angles} \rangle \\
& \omega_0 + \omega_1 \left(\sin(2\pi\theta_1)\cos(2\pi\varphi_1) + \cos(2\pi\theta_1)\sin(2\pi\varphi_1)\right) = 0 \\
\equiv\ & \langle x_1 = \cos(2\pi\theta_1),\ y_1 = \sin(2\pi\theta_1) \rangle \\
& \omega_0 + \omega_1 \left(y_1\cos(2\pi\varphi_1) + x_1\sin(2\pi\varphi_1)\right) = 0 \\
\equiv\ & \langle \text{Arithmetic} \rangle \\
& \omega_1\cos(2\pi\varphi_1)y_1 = -\omega_1\sin(2\pi\varphi_1)x_1 - \omega_0 \\
\equiv\ & \langle \text{Arithmetic} \rangle \\
& y_1 = -\tan(2\pi\varphi_1)x_1 - \frac{\omega_0}{\omega_1\cos(2\pi\varphi_1)}
\end{aligned}
$$

Then, given that the range of the tangent function is $\mathbb{R}$, the decision boundary can be rewritten as the two dimensional line equation $y = mx + b$, with any possible slope $m = -\tan(2\pi\varphi_1)$ and y-intercept $b = \frac{\omega_0}{\omega_1\cos(2\pi\varphi_1)}$. □

Theorem 1 shows that the expressiveness of the dLR for unidimensional problems with one predictive directional variable in the transformed space is defined by the entire set of two dimensional lines. However, the decision boundary in the original space is not linear; it is translated as two decision angular-thresholds, ϕ and ϕ', such that, if the angular distance between them is Δ, one of the possible induced models in the original space is represented by the parameter configuration:

$$\varphi_1 = \pm\frac{1}{2\pi}\arcsin\left(\sqrt{\frac{1+\cos(2\pi\Delta)}{2}}\right) - \phi$$
$$\omega_0 = \mp\sqrt{\frac{(1+\cos(2\pi\Delta))}{2}}$$
$$\omega_1 = 1$$

where ω_0 takes the positive version of the equation if the distance between both thresholds is greater than half of the period (ϕ_1 the negative side) and vice versa. Notice that there is an infinite number of models with the same decision boundary, since we can scale ω by any non-zero factor and obtain the same predictions. This property is also true for the standard LR. An example of the model expressiveness for this trivial case is illustrated in Figure 6.1.

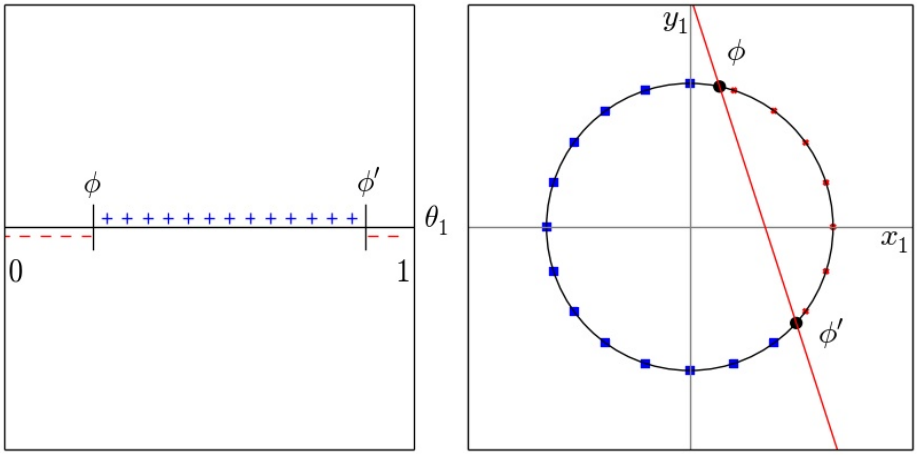


Figure 6.1: Decision boundary for a problem with one directional variable. **Left:** decision boundary in the original space represented by two decision thresholds. **Right:** decision boundary in the extended space represented by the 2-dimensional line.

6.4.2 N-dimensional feature space with K angular variables

We now analyze the general scenario where the feature space has an unrestricted number of directional and non-directional variables. For the sake of simplicity, we assume that the first K features are directional and the remaining linear (referred to as hypothesis H_0 in the proof of the Theorem 2. This assumption does not suppose a loss of generality given that the model is invariant to the arrangement of the features. As we did before, we analyze the expressiveness of the model in the transformed space.

Theorem 2. *The dLR classifier with N predictive variables, being $K \leq N$ of them directional, induces a separation boundary equivalent to a $(N+K)$-dimensional hyperplane in the transformed space.*

Proof.

$$
\begin{aligned}
& \omega_0 + \sum_{i=1}^{N} \omega_i g_i(\theta_i) = 0 \\
\equiv \;& \langle \text{Range Split} \rangle \\
& \omega_0 + \sum_{i=1}^{K} \omega_i g_i(\theta_i) + \sum_{i=K+1}^{N} \omega_i g_i(\theta_i) = 0 \\
\equiv \;& \langle H_0, \text{ Definition of } g \rangle \\
& \omega_0 + \sum_{i=1}^{K} \omega_i \sin(2\pi(\theta_i + \varphi_i)) + \sum_{i=K+1}^{N} \omega_i \theta_i = 0 \\
\equiv \;& \langle \text{Sum of two angles, Arithmetic} \rangle \\
& \omega_0 + \sum_{i=K+1}^{N} \omega_i \theta_i \; + \\
& \sum_{i=1}^{K} \omega_i \left(\sin(2\pi\theta_i)\cos(2\pi\varphi_i) + \cos(2\pi\theta_i)\sin(2\pi\varphi_i)\right) = 0 \\
\equiv \;& \langle x_i = cos(2\pi\theta_i),\; y_i = sin(2\pi\theta_i),\; \text{Arithmetic} \rangle \\
& \omega_0 + \sum_{i=K+1}^{N} \omega_i \theta_i \; + \sum_{i=1}^{K} (\omega_i \sin(2\pi\varphi_i) x_i + \omega_i \cos(2\pi\varphi_i) y_i) = 0
\end{aligned}
$$

□

An interesting and usual two dimensional scenario arises when angular measurements are accompanied by a scale factor or magnitude (e.g. wind direction and speed, forces, etc), thereby inducing a cylinder as the geometric space where input vectors lie. Figure 6.2 shows an example of the decision region in both, the original $\mathbb{R}^2$ space and the transformed $\mathbb{R}^3$ space, where one variable is directional.

6.5 Optimization Strategy

For the purpose of this work, the traditional gradient descent learning strategy from the LR was adapted to the proposed directional version. Let us assume we have a set of labeled input data S, where each instance $<\theta, y> \; \in S \subseteq \mathbb{R}^n \times \{0,1\}$, is a pair of an input vector θ and its corresponding label y.

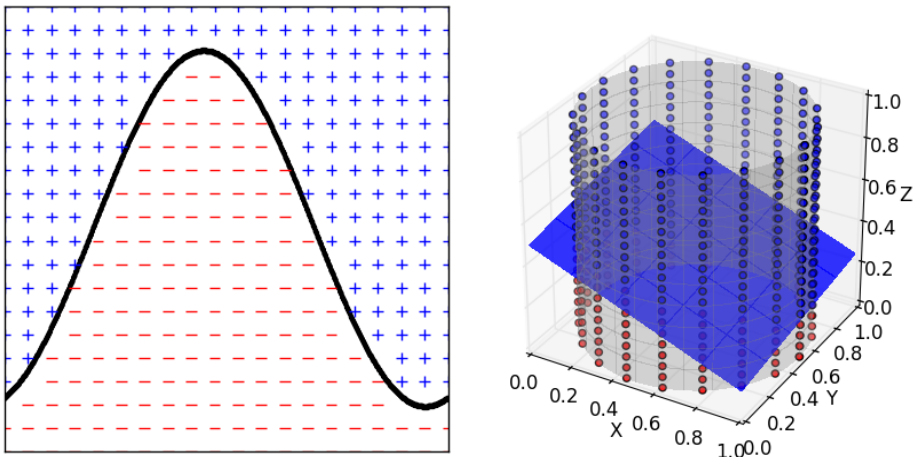


Figure 6.2: Decision boundary for a mixed problem in $\mathbb{R}^2$. **Left:** non-linear decision boundary in the original space. **Right:** decision boundary in the extended space represented by a three dimensional plane.

From this scenario, we consider the traditional regularized Logistic loss function (Log loss) used in (multinomial) LR (c.f. Eq. (6.4)).

$$J(\omega,\varphi) = -\frac{1}{|S|}\sum_{<\theta,y>\in S} cost(y,\theta) + \frac{\lambda}{2n}\sum_{i=1}^{n}\omega_i^2 \tag{6.4}$$

$$cost(y,\theta) = y\log(f(\theta)) + (1-y)\log(1-f(\theta)) \tag{6.5}$$

This function can be enhanced in order to include different misclassification costs by considering the weighted sum of the errors. In order to fit the model, the goal of our optimization task is to find the best parameter configuration ω,φ such that:

$$\arg\min_{\omega,\varphi} \; J(\omega,\varphi)$$

Using a gradient descent strategy requires the computation of the partial derivatives of the goal function J with respect to each model parameter. The corresponding derivatives

are shown below in the Eqs. (6.6a) - (6.6d).

$$\frac{\partial}{\partial \omega_0} J(\omega, \varphi) = \frac{k}{|S|} \sum_{<\theta, y> \in S} (f(\theta) - y) \tag{6.6a}$$

$$\frac{\partial}{\partial \omega_{i>0}} J(\omega, \varphi) = \frac{k}{|S|} \sum_{<\theta, y> \in S} (f(\theta) - y) \cdot g_i(\theta_i) + \frac{\lambda}{n} \omega_i \tag{6.6b}$$

$$\frac{\partial}{\partial \varphi_i} J(\omega, \varphi) = \frac{k \cdot \omega_i}{|S|} \sum_{<\theta, y> \in S} (f(\theta) - y) \frac{\partial}{\partial \varphi_i} g_i(\theta_i) \tag{6.6c}$$

$$\frac{\partial}{\partial \varphi_i} g_i(\theta_i) = \begin{cases} 2\pi \cos(2\pi(\theta_i + \varphi_i)) & , \text{ if } dir(i) \\ 0 & , \text{ otherwise} \end{cases} \tag{6.6d}$$

Algorithm 1 Gradient descent with variable sigmoid's slope

```
1:  function GRADIENT DESCENT(samples, labels)
2:      ω, φ ← initialize_model()
3:      ω*, φ* ← ω, φ
4:      J* ← J(ω, φ)
5:      k, ε ← 1, ε_init
6:
7:      for i ← 1 to max_iterations do
8:          ω, φ ← ω − α · ∂/∂ω J(ω, φ), φ − α · ∂/∂φ J(ω, φ)
9:          J_next ← J(ω, φ)
10:         if k < k_max ∧ |J_next − J*| < ε then
11:             k, ε ← k + 1, ε · ε_Δ
12:         end if
13:         if J_next < J* then
14:             ω*, φ* ← ω, φ
15:             J* ← J_next
16:         else
17:             α ← α · decaying_rate
18:         end if
19:     end for
20:
21:     return ω*, φ*
22: end function
```

Then, we can use a gradient-based optimization strategy to fit the model. In our case, we have used a Gradient Descent variation with decaying learning rate and increasing slope of the sigmoid function to boost the algorithm's convergence (see Algorithm 1). In order to avoid having to change the learning rate as the slope of the sigmoid function changes, we removed the constant k from the derivatives, which preserves the direction of the gradient but simplifies parameter tuning. In gradient descent optimization techniques,

monotonously decreasing the learning rate towards zero guarantees the convergence of the iterative process. In our setting, given that the search space is not convex, the method may converge to a local minimum. However, as will be shown in the experimental evaluation, the proposed algorithm is able to reach competitive results.

6.6 Experiments

In this section, we detail the experimental evaluation of the proposed dLR classifier and its non-directional version LR against their generative counterparts von Mises naïve Bayes and Gaussian naïve Bayes classifiers [222]. These methods can be summarized as follows:

1. Gaussian Naive Bayes (GNB): Gaussian NB classifier that models continuous variables using Gaussian distributions.

2. von-Mises Naive Bayes (vMNB): NB classifier that models linear variables using Gaussian distributions and directional variables using von Mises distributions.

Furthermore, López et al. [222] validated a feature selection strategy proposed by Langley and Sage [196] as a wrapper of their NB approach. Also, they evaluated the performance of the NB classifier by discretizing all the continuous variables. However, given that the goal of this section is to validate the performance of the proposed discriminative method against its generative counterpart, we considered the plain GNB and vMNB methods. The study of feature selection and discretization strategies are out of the scope of this work and might improve the results shown below. In the following experiments, the κ parameter of the von Mises distribution was approximated by 100 iterations (a much larger number of iterations than required to have good convergence values) of the Newton's method proposed by Sra [314].

Also, we compare our model with an SVM [53] using a directional squared exponential (i.e. Gaussian RBF) kernel [293]. This kernel considers the distance between a given pair of points, wherein the distance between two directional variables is considered in an angular manner instead of the traditional Euclidean distance. The regularization parameter (C) and the γ parameter of the squared exponential kernel was chosen by cross-validation among seven different values in the logarithmic scale between 10^{-2} and 10^{2}.

On the other hand, both LR variants had an initial learning rate value (α) of 0.1 and a maximum number of $20,000$ iterations for the synthetic data and $10,000$ iterations for real data, but most datasets required much less iterations to converge. The model was initialized using small random values ($\omega_i \in [-0.05, 0.05]$ and $\varphi_i \in [-0.05, 0.05]$). The regularization constant $C = \lambda^{-1}$ was chosen following the same strategy used in the training of the SVM.

6.6.1 Experiments with Synthetic Data

We evaluated the performance of the classifiers using one directional predictive variable and two possible responses (binary classification), under three different statistical distributions (e.g., uniform, triangular and von Mises). Then, for each possible distribution, we randomly generated 75 synthetic binary datasets with 100 samples (50 samples per class). Afterward, we validated the accuracy of each model using a training and test validation assessment using the classic 70-30 partition. We compared the two aforementioned naïve Bayes versions with the two versions of the LR. Also, we assessed the proposed strategy by comparing the results with a brute force search that compares each possible pair of thresholds (by maximizing the margin between two observations belonging to different classes) and minimizes the training error (g-dLR), which represents the best value that could be achieved by optimizing the model according to its training classification error. It should be clear that the brute force optimization is not an option in practice when several features are used.

Table 6.1: Average classification error per model with unidimensional synthetic datasets.

Distr.	GNB	vMNB	LR	dLR	g-dLR
Uniform	91.25 ± 4.85	92.56 ± 6.15	82.99 ± 7.63	**93.44** ± 3.02	*96.07* ± *2.00*
Triangular	93.56 ± 4.14	94.82 ± 2.62	86.99 ± 8.53	**95.34** ± 2.43	*96.78* ± *1.89*
von Mises	95.25 ± 2.56	**96.25** ± 2.42	87.34 ± 9.46	95.56 ± 2.52	*96.47* ± *2.25*

Table 6.1 summarizes the accuracy results for these experiments. In general, the Grid Search strategy obtained the best results for each possible distribution. As expected, the dLR classifier trained with the gradient descent algorithm outperforms both generative models for all the distribution but the von Mises distribution. Furthermore, the difference between the gradient-based and the grid strategy suggests that there is still room for improving the optimization stage, although the optimization is doing a good job. The worst results were obtained by the LR as it does not have enough expressiveness to discriminate these directional datasets.

6.6.2 Experiments with Real Data

Then, we validated the advantages of the proposed approach using thirteen real datasets. For this purpose, we compared the two naïve Bayes variations and the SVM with directional RBF kernel against the classic LR and the directional version proposed in this work. For computational reasons we only validated the gradient-based optimization strategy, given that the Grid-Search approach, used in the previous experiments, would be computationally intractable. Table 6.2 summarizes the dimensionality of the evaluated datasets (e.g. number of variables, class values and instances).

Table 6.2: Summary of the main characteristics of the datasets used in this work. Including number of features per type (i.e. Directional - Dir, Linear - Lin, Discrete - Disc) and number of samples per dataset (#).

Dataset	**Number of variables**			**Class**	#
	Dir	**Lin**	**Disc**	**values**	
Colposcopy	3	6	0	3	150
Behavior	140	426	20	4	261
Arrhythmia	4	191	66	2	430
eBay	1	2	0	11	528
Megaspores	1	0	0	2	960
Characters	5	31	0	10	1,000
OnlineNews	1	12	0	2	1,000
Continents	2	0	0	5	3,481
Wall	6	6	0	4	5,456
Temperature0	1	1	1	3	8,764
Temperature1	2	1	0	3	8,764
Temperature2	5	1	0	3	8,764
MAGIC	1	10	0	2	19,020

Multiclass instances were handled using a one-versus-one approach for both versions of the LR. All the experiments detailed below were executed with a stratified 5-fold cross-validation technique (by preserving the percentage of samples for each class), and results of 40 different runs were averaged. Results of these experiments are summarized in Table 6.3, exhibiting average accuracy and standard deviation for forty independent runs. The best model for each dataset is represented bold.

When comparing generative models, we obtained similar results to those obtained by López et al., namely vMNB achieves similar or better results than the GNB in most datasets [222]. The directional version of the LR classifier reports a broad and significant advantage when compared with the non-directional approach, achieving up to 22% more percentage points in the `eBay` dataset than the traditional LR.

In general, the best results in the entire set of problems were achieved either by the dLR (6 datasets) or by the SVM model (7 datasets). As can be seen in Table 6.3, dLR obtained better results than the SVM model mainly in the datasets with fewer instances. Given that the RBF kernel can be understood as a projection on a feature space with an infinite number of dimensions, the SVM model can generate highly nonlinear decision regions in contrast with the $(N+K)$-hyperplanes generated by dLR. Thereby, dLR offers a much more succinct representation to reason about directional data without compromising accuracy. Also, in some contexts, it is preferred to use simpler (linear) models, especially when computational resources are limited or when there are interpretability requirements.

Moreover, dLR achieved better results than its non-directional and generative counterparts in almost all datasets, being only surpassed in the `Temperature0` dataset. Furthermore, when combining the best descriptors for the basic `Temperature` dataset, considering

Table 6.3: Average accuracy per model using 5-fold cross-validation.

Dataset	GNB	vMNB	LR	dLR	SVM
Colposcopy	74.71 ± 7.08	70.93 ± 7.83	73.66 ± 6.97	**80.61** ± 6.49	80.39 ± 7.51
Behavior	47.21 ± 9.43	49.26 ± 9.20	82.46 ± 3.48	**82.68** ± 3.56	82.63 ± 3.71
Arrhythmia	67.06 ± 4.03	67.05 ± 4.07	78.31 ± 3.99	78.38 ± 4.04	**78.66** ± 3.87
eBay	77.45 ± 3.37	83.88 ± 3.75	62.33 ± 4.42	**84.86** ± 3.21	84.11 ± 3.21
Megaspores	76.72 ± 2.54	76.61 ± 2.71	62.50 ± 0.00	**76.78** ± 2.58	76.32 ± 2.72
Characters	70.94 ± 2.62	73.40 ± 2.99	94.99 ± 1.59	**95.77** ± 1.35	95.75 ± 1.27
OnlineNews	55.37 ± 2.12	55.29 ± 2.03	56.25 ± 2.94	**56.26** ± 2.95	52.80 ± 0.10
Continents	94.66 ± 0.72	94.90 ± 1.08	94.79 ± 0.74	95.87 ± 0.72	**97.72** ± 0.48
Wall	45.69 ± 2.01	51.07 ± 2.79	58.06 ± 1.39	66.53 ± 1.29	**86.41** ± 0.94
Temperature0	68.56 ± 0.83	69.99 ± 1.80	59.15 ± 0.78	56.14 ± 0.92	**72.76** ± 0.82
Temperature1	64.44 ± 0.83	65.04 ± 1.49	59.15 ± 0.90	70.28 ± 0.89	**71.21** ± 0.91
Temperature2	12.84 ± 0.09	67.70 ± 1.66	59.65 ± 0.70	79.21 ± 0.87	**82.28** ± 0.79
MAGIC	72.68 ± 0.53	73.01 ± 0.52	79.08 ± 0.50	80.77 ± 0.49	**87.35** ± 0.46

the season as a nominal value encoded as an integer for the vMNB and, as a directional variable for the dLR, the dLR classifier achieves the best performance. On average, the proposed model achieved accuracy values 8.15% higher than the vMNB.

The main disadvantage of the proposed model, when compared with its generative counterpart, is the computational time required for the training stage. While naïve Bayes approaches require basic fitting of statistical distributions, dLR is learned by means of an iterative procedure, with asymptotic complexity $\mathcal{O}(I \times |S| \times N)$, where I is the maximum number of iterations, $|S|$ is the number of samples in the training set and N is the number of features. However, once trained, dLR is computationally competitive as it has linear complexity on the number of features – $\mathcal{O}(N)$.

6.7 Conclusions

Different concepts in real life applications are represented by directional variables. These concepts are not restricted to the scientific domain but can be easily found in daily routines, such as representing of time in a periodic repetitive calendar (e.g., hour, the day of the week, month, etc.). Traditional classifiers, which are unaware of the angular nature of these variables, might not properly model the data. Thereby, some directional classifiers

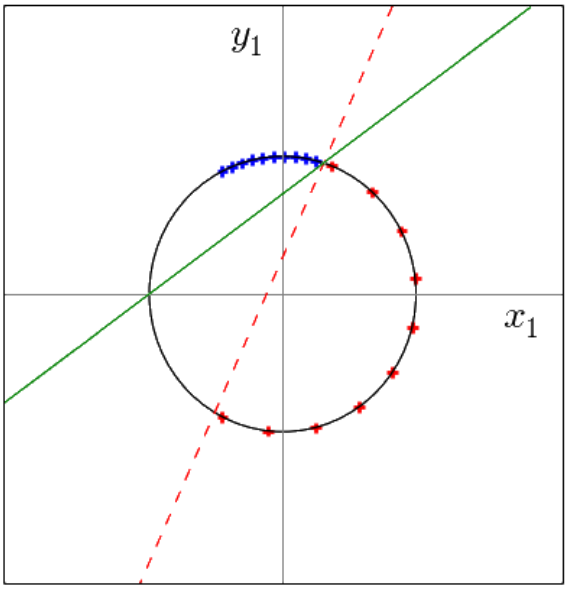


Figure 6.3: Decision boundary for a linear SVM in the extended space (dashed line) and in the original space (solid line).

have been proposed in the past, most of them using generative approaches [222, 361] and the directional von Mises distribution [105, 222].

In this work, we proposed a discriminative binary classifier that is able to receive mixed data (directional and linear). This classifier adds to the classic LR awareness about the angular nature of the data. As we demonstrated in the experimental assessment of the proposed model with both synthetic and real data, it can achieve competitive results when compared against traditional non-directional LR, previous generative approaches, and SVM using a directional RBF kernel. Other advantages of the dLR model accruing from retaining the access to the posterior probabilities include risk minimization, reject option, compensating for class priors, combining models, etc. Non-probabilistic methods, like the SVM, need to involve an intermediate step where a map from the decision regions to the actual probability is estimated [32]. Therefore, the dLR classifier offers promising results when dealing with directional data, and there is room for future improvement.

We plan to adapt our dLR model to be intended as a directional perceptron within an ANN. The opportunity of studying the effect of dynamic frequency regimes in this classifier is an open problem, as the dLR classifier was defined in a way that it is able to encode only a single period of the directional variables. Finally, there is room for exploring more advanced optimization techniques that may improve the performance of this model.

Extending the proposed work to Support Vector Machines is relevant since the decision region obtained in the extended space does not correspond to the right margin that maximizes the distance between the support vectors and the decision boundary in the original space. Figure 6.3 shows the different regions obtained by maximizing the margin

in the extended space (dashed line) and in the original space (solid line). For visualization purposes, the decision boundary computed in the original space is transformed into its equivalent line in the extended space. This artifact is a result of comparing the distance between the support vector and the decision boundary in the two-dimensional Euclidean space instead of using the angular distance between the support vector and the boundary-sphere intersection points. Thus, we proposed several instantiations of directional Support Vector Machines covering kernels, including their kernelized counterparts.

Chapter 7

Deep Local Binary Patterns

LBP is a traditional descriptor for texture analysis that gained attention in the last decade. Being robust to several properties such as invariance to illumination translation and scaling, LBP achieved state-of-the-art results in several applications. However, LBP are not able to capture high-level features from the image, merely encoding features with low abstraction levels. In this work, we propose Deep LBP, which borrow ideas from the deep learning community to improve LBP expressiveness. By using parametrized data-driven LBP, we enable successive applications of the LBP operators with increasing abstraction levels. We validate the relevance of the proposed idea in several datasets from a wide range of applications. Deep LBP improved the performance of traditional and multiscale LBP in all cases.

7.1 Introduction

In recent years, CV community has moved towards the usage of Deep Learning (DL) strategies to solve a wide variety of traditional problems, from image enhancement [337] to scene recognition [372]. DL concepts emerged from traditional shallow concepts from the early years of CV (e.g. filters, convolution, pooling, thresholding, etc.).

Although these techniques have achieved *state-of-the-art* performance in several of tasks, the DL hype has overshadowed research on other fundamental ideas. Narrowing the spectrum of methods to a single class will eventually saturate, creating a monotonous environment where the same basic idea is being replicated over and over, and missing the opportunity to develop other paradigms with the potential to lead to complementary solutions.

As deep strategies have benefited from traditional -shallow- methods in the past, some classical methods started to take advantage of key DL concepts. That is the case of deep Kernels [48], which explores the successive application of nonlinear mappings within the kernel umbrella. In this work, we incorporate deep concepts into LBP [251, 252], a

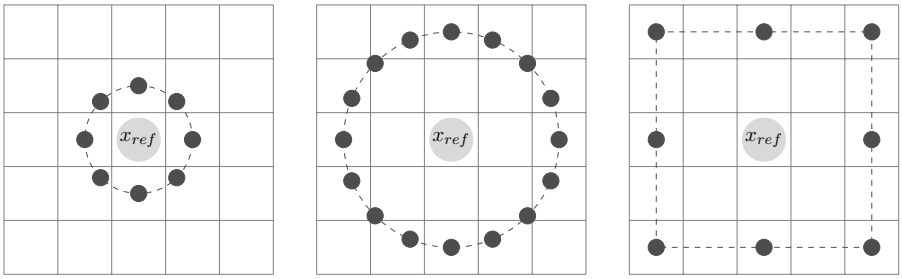
traditional descriptor for texture analysis. LBP is a robust descriptor that briefly summarizes texture information, being invariant to illumination translation and scaling. LBP has been successfully used in a wide variety of applications, including texture classification [134, 135, 211, 368], face/gender recognition [8, 155, 281, 301, 362, 367], among others [242, 342, 354].

LBP has two main ingredients:

- The neighborhood ($\mathcal{N}$), usually defined by an angular resolution (typically 8 sampling angles) and radius r of the neighborhoods. Figure 7.1 illustrates several possible neighborhoods.
- The binarization function $b(x_{ref}, x_i) \in \{0, 1\}$, which allows the comparison between the reference point (central pixel) and each one of the points x_i in the neighborhood. Classical LBP is applicable when x_{ref} (and x_i) are in an ordered set (e.g., $\mathbb{R}$ and $\mathbb{Z}$), with $b(x_{ref}, x_i)$ defined as

$$b(x_{ref}, x_i) = (x_{ref} \prec x_i), \tag{7.1}$$

  where $\prec$ is the order relation on the set (interpolation is used to compute x_i when a neighbor location does not coincide with the center of a pixel).



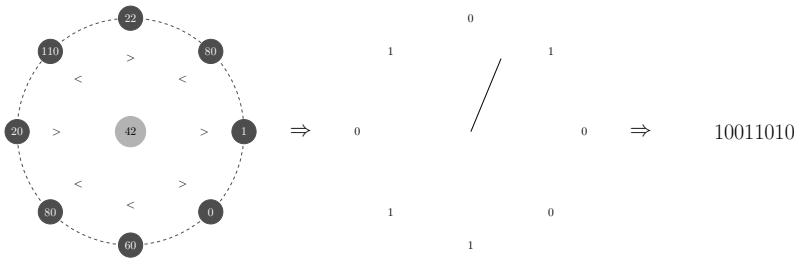
(a) r=1, n=8 (b) r=2, n=16 (c) r=2, n=8

Figure 7.1: LBP neighborhoods with radius (r) and angular resolution (n). The first two cases use Euclidean distance to define the neighborhood, the last case use Manhattan distance.

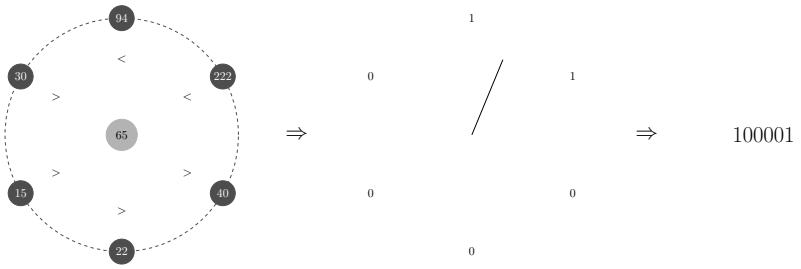
The output of the LBP at each position ref is the code resulting from the comparison (binarization function) of the value x_{ref} with each of the x_i in the neighborhood, with $i \in \mathcal{N}(ref)$, see Figure 7.2. The LBP codes can be represented by their numerical value as formally defined in (7.2).

$$LBP(x_{ref}) = \sum_{i \in |\mathcal{N}(ref)|} 2^i \cdot b(x_{ref}, x_i) \tag{7.2}$$

LBP codes can take $2^{|\mathcal{N}|}$ different values. In predictive tasks, for typical choices of angular resolution, LBP codes are compactly summarized into a histogram with $2^{|\mathcal{N}|}$ bins,



(a) LBP code with 8 neighbors



(b) LBP code with 6 neighbors

Figure 7.2: Cylinder and linear representation of the codes at some pixel positions. Encodings are built in a clockwise manner from the starting point indicated in the middle section of both figures.

being this the feature vector representing the object/region/image (see Figure 7.3). Also, it is typical to compute the histograms in sub-regions and to build a predictive model by using as features the concatenation of the region histograms, being non-overlapping and overlapping [26] blocks traditional choices (see Figure 7.4).

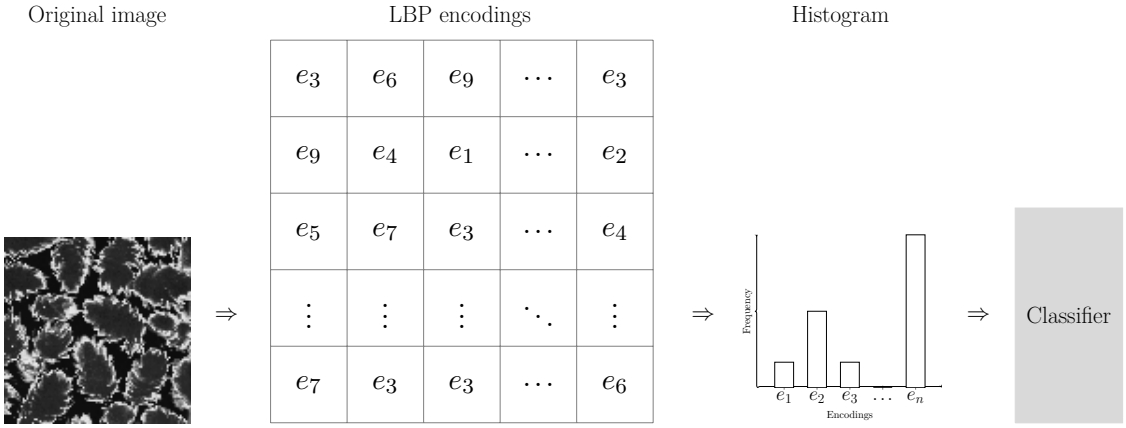


Figure 7.3: Traditional pipeline for image classification using LBP.

In the last decade, several variations of the LBP have been proposed to attain different properties. The two best-known variations were proposed by Ojala et al., the original authors of the LBP methodology: rotation invariant and uniform LBP [252].

Rotation invariance can be achieved by assigning a unique identifier to all the patterns that can be obtained by applying circular shifts. The new encoding is defined in (7.3),

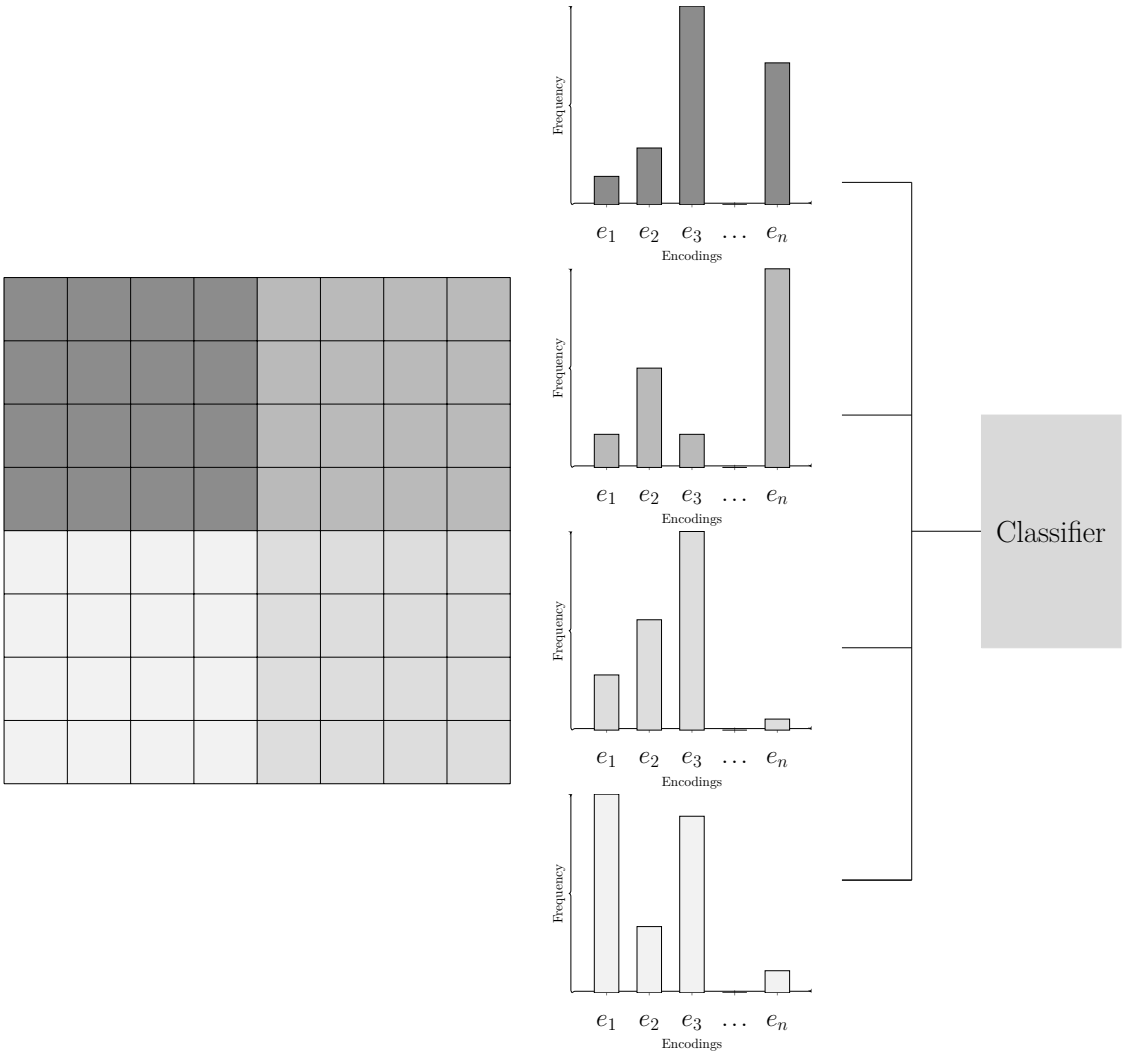


Figure 7.4: Multi-block LBP with 2×2 non-overlapping blocks.

where $ROR(e, i)$ applies a circular-wrapped bit-wise shift of i positions to the encoding e.

$$LBP(x_{ref}) = \min\{ROR(LBP(x_{ref}), i) \mid i \in [0, \dots |\mathcal{N}|)\} \tag{7.3}$$

In the same work, Ojala et al. [252] identified that uniform patterns (i.e., patterns with two or less circular transitions) are responsible for the vast majority of the histogram frequencies, leaving low discriminative power to the remaining ones. Then, the relatively small proportion of non-uniform patterns limits the reliability of their probabilities, all the non-uniform patterns are assigned to a single bin in the histogram construction, while uniform patterns are assigned to individual bins. [252].

Heikkila et al. proposed Center-Symmetric LBP [148], which increases robustness on flat image areas and improves computational efficiency. This is achieved by comparing pairs of neighbors located in centered symmetric directions instead of comparing each neighbor with the reference pixel. Thus, an encoding with four bits is generated from

a neighborhood of 8 pixels. Also, the binarization function incorporates an activation tolerance given by $b(x_i, x_j) = x_i > x_j + T$. Further extensions of this idea can be found in [307, 325, 346].

Local Ternary Patterns (LTP) are an extension of LBP that use a 3-valued mapping instead of a binarization function. The function used by LTP is formalized at (7.4). LTP are less sensitive to noise in uniform regions at the cost of losing invariance to illumination scaling. Moreover, LTP induce an additional complexity in the number of encodings, producing histograms with up to $3^{\mathcal{N}}$ bins.

$$b_T(x_{ref}, x_i) = \begin{cases} -1 & x_i < x_{ref} - T \\ 0 & x_{ref} - T \leq x_i \leq x_{ref} + T \\ +1 & x_{ref} + T < x_i \end{cases} \tag{7.4}$$

So far, we have mainly described methods that rely on redefining the construction of the LBP encodings. A different line of research focuses on improving LBP by modifying the texture summarization when building the frequency histograms. Two examples of this idea were presented in this work: uniform LBP [252] and multi-block LBP (see Figure 7.4).

Since different scales may bring complementary information, one can concatenate the histograms of LBP values at different scales. Berkan et al. proposed this idea in the Over-Complete Local Binary Pattern (OCLPB)[26]. Besides computing the encoding at multiple scales, OCLPB computes the histograms on several spatially overlapped blocks. An alternative to this way of modeling multiscale patterns is to, at each point, compute the LBP code at different scales, concatenate the codes and summarize (i.e., compute the histogram) of the concatenated feature vector. This latter option has difficulties concerning the dimensionality of the data (potentially tackled with a bag of words approach) and the nature of the codes (making unsuitable standard k-means to find the bins-centers for a bag of words approach).

Multi-channel data (e.g., Red-Green-Blue (RGB)) has been handled in a similar way, by 1) computing and summarizing the LBP codes in each channel independently and then concatenating the histograms [49] and by 2) computing a joint code for the three channels [373].

As LBP have been successfully used to describe spatial relations between pixels, some works explored embedding temporal information on LBP for object detection and background removal [70, 347, 351, 356, 365, 367].

Finally, LBP Network (LBPNet) was introduced by Xi et al. [336] as a preliminary attempt to embed DL concepts in LBP. Their proposal consists in using a pyramidal approach on the neighborhood scales and histogram sampling. Then, Principal Component Analysis (PCA) is used on the frequency histograms to reduce the dimensionality of the feature space. Xi et al. analogize the pyramidal construction of LBP neighborhoods and

histogram sampling as a convolutional layer, where multiple filters operate at different resolutions, and the dimensionality reduction as a Pooling layer. However, LBPNets aren't capable of aggregating information from a single resolution into higher levels of abstraction which is the main advantage of DNN.

In the next sections, we will bring some ideas from the DL community to traditional LBP. In this sense, we intend to build LBP blocks that can be applied recursively to build features with higher-level of abstraction.

7.2 Deep Local Binary Patterns

The ability to build features with increasing abstraction level using a recursive combination of relatively simple processing modules is one of the reasons that made Convolutional Neural Network (CNN) – and in general ANN – so successful. In this work, we propose to represent "higher order" information about texture by applying LBP recursively, i.e., cascading LBP computing blocks one after the other (see Figure 7.5). In this sense, while an LBP encoding describes the local texture, a second order LBP encoding describes the texture of textures.

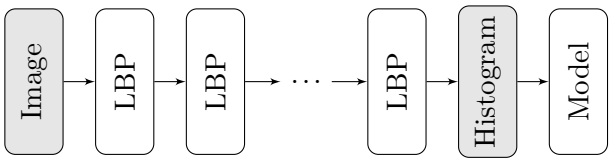


Figure 7.5: Recursive application of LBP.

However, while it is trivial to recursively apply convolutions – and many other filters – in a cascade fashion, traditional LBP are not able to *digest* their own output.

Traditional LBP rely on receiving as input an image with a domain in an ordered set (e.g., grayscale intensities). However, LBP codes are not in an ordered set, dismissing the direct recursive application of standard LBP. As such we will first generalize the main operations supporting LBP and discuss next how to assemble a deep/recursive LBP feature extractor. We start the discussion with the binarization function $b(x_{ref}, x_i)$.

It is instructive to think, in the conventional LBP, if non-trivial alternative functions exist to the adopted one, Eq. (7.1). What is(are) the main property(ies) required by the binarization function? Does it need to make use of a (potentially implicit) order relationship? A main property of the binarization function is to be independent of scaling and translation of the input data, that is,

$$b(k_1 x_{ref} + k_2,\ k_1 x_i + k_2) = b(x_{ref}, x_i),\ k_1 > 0. \tag{7.5}$$

It is trivial to prove that the only options for the binarization function that hold Eq. (7.5) are the constant functions (always zeros or always ones), the one given by Eq. (7.1) and its reciprocal.

Proof. Assume $x_i, x_j > x_{ref}$ and $b(x_{ref}, x_i) \neq b(x_{ref}, x_j)$. Under the independence to translation and scaling (Eq. (7.5)), $b(x_{ref}, x_i) = b(x_{ref}, x_j)$ as shown below, which is a contradiction.

$$
\begin{aligned}
& b(x_{ref}, x_i) \\
=& \langle \text{ Identity of multiplication } \rangle \\
& b\left(\frac{x_j - x_{ref}}{x_j - x_{ref}} x_{ref}, \frac{x_j - x_{ref}}{x_j - x_{ref}} x_i\right) \\
=& \langle \text{ Identity of addition } \rangle \\
& b\left(\frac{x_j - x_{ref} + (x_i - x_i)}{x_j - x_{ref}} x_{ref}, \frac{x_j - x_{ref}}{x_j - x_{ref}} x_i + \left(\frac{x_{ref} x_j - x_{ref} x_j}{x_j - x_{ref}}\right)\right) \\
=& \langle \text{Arithmetic} \rangle \\
& b\left(\frac{x_i - x_{ref}}{x_j - x_{ref}} x_{ref} + x_{ref} \frac{x_j - x_i}{x_j - x_{ref}}, \frac{x_i - x_{ref}}{x_j - x_{ref}} x_j + x_{ref} \frac{x_j - x_i}{x_j - x_{ref}}\right) \\
=& \left\langle \text{ Eq. (7.5), where } k_1 = \frac{x_i - x_{ref}}{x_j - x_{ref}}, k_2 = x_{ref} \frac{x_j - x_i}{x_j - x_{ref}} \right\rangle \\
& b(x_{ref}, x_j)
\end{aligned}
$$

Therefore, $b(x_{ref}, x_i)$ must be equal to all x_i above x_{ref}. Similarly, $b(x_{ref}, x_i)$ must be equal to all x_i below x_{ref}. □

Among our options, the constant binarization function is not a viable option, since the information (in information theory perspective) in the output is zero. Since the recursive application of functions can be understood as a composition, invariance to scaling and translation is trivially ensured by using a traditional LBP in the first transformation.

Given that transitivity is a relevant property held by the natural ordering of real numbers, we argue that such property should be guaranteed by our binarization function. In this sense, we will focus on strict partial orders of encodings. Following, we show how to build such binarization functions for the i-*th* application of the LBP operator, where $i > 1$. We will consider both predefined/expert-driven solutions and data-driven solutions (and therefore, application specific).

Hereafter, we will refer to the binarization function as the partial ordering of LBP encodings. Despite the existence of other types of functions may be of general interest, narrowing the search space to those that can be represented as a partial ordering induce efficient learning mechanisms.

7.2.1 Preliminaries

Let us formalize the deep binarization function as the order relation $b^+ \in \mathcal{P}(E_\mathcal{N} \times E_\mathcal{N})$, where $E_\mathcal{N}$ is the set of encodings induced by the neighborhood $\mathcal{N}$.

Let Φ be an oracle $\Phi :: \mathcal{P}(E_\mathcal{N} \times E_\mathcal{N}) \to \mathbb{R}$ that assesses the performance of a binarization function. For example, among other options, the oracle can be defined as the performance of the traditional LBP pipeline (see Figure 7.3) on a dataset for a given predictive task.

7.2.2 Deep Binarization Function

From the entire space of binarization functions, we restrict our analysis to those induced by strict partial orders. Within this context, it is easy to see that learning the best binarization function by exhaustive exploration is intractable since the number of combinations equals the number of DAG with $2^{|\mathcal{N}|} = |E_\mathcal{N}|$ nodes. The DAG counting problem was studied by Robinson [282] and is given by the recurrence relation Eqs. (7.6)-(7.7).

$$a_0 = 1 \tag{7.6}$$

$$a_{n>1} = \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} 2^{k(n-k)} a_{n-k} \tag{7.7}$$

Table 7.1: Lower bound of the number of combinations for deciding the best LBP binarization function as *partial orders*.

# Neighbors	Rotational Inv.	Uniform	Traditional
2	$2 \cdot 10^1$	$2 \cdot 10^4$	$5 \cdot 10^2$
3	$5 \cdot 10^2$	$1 \cdot 10^{15}$	$7 \cdot 10^{11}$
4	$3 \cdot 10^6$	$2 \cdot 10^{41}$	$8 \cdot 10^{46}$
5	$7 \cdot 10^{11}$	$6 \cdot 10^{94}$	$2 \cdot 10^{179}$
6	$1 \cdot 10^{36}$	$2 \cdot 10^{190}$	$1 \cdot 10^{685}$
7	$2 \cdot 10^{72}$	$3 \cdot 10^{346}$	$3 \cdot 10^{2640}$
8	$1 \cdot 10^{225}$	$1 \cdot 10^{585}$	$3 \cdot 10^{10288}$

Table 7.1 illustrates the size of the search space for several numbers of neighbors. For instance, for the traditional setting with 8 neighbors, the number of combinations has more than 10,000 digits. Thereby, a heuristic approximation must be carried out.

7.2.2.1 Learning b^+ from a User-defined dissimilarity function

The definition of a dissimilarity function between two codes seems reasonably accessible. For instance, an immediate option is to adopt the Hamming distance between codes, d_H. With rotation invariance in mind, one can adopt the minimum hamming distance between

all circularly shifted versions of x_{ref} and x_i, d_H^{ri}. The circular invariant hamming distance between x_{ref} and x_i can be computed as

$$d_H^{ri} = \min_{s \in 0,\cdots,\mathcal{N}-1} d_H(ROR(x_{ref}, s), x_i) \tag{7.8}$$

Having defined such a dissimilarity function between pairs of codes, one can now proceed with the definition of the binarization function.

Given the dissimilarity function, we can learn a mapping of the codes to an ordered set. Resorting to Spectral Embedding [245], one can obtain such a mapping. The conventional binarization function, Eq. (7.1), can then be applied. Other alternatives for building the mappings can be found in the manifold learning literature: Isomaps [321], Multi Dimensional Scaling (MDS) [181], among others. In this case, the oracle function can be defined as the intrinsic loss functions used in the optimization process of such algorithms.

Preserving a desired property P such as rotational invariance and sign invariance (i.e., interchangeability between ones and zeros) can be achieved by considering P-aware dissimilarities.

7.2.2.2 Learning b^+ from a High-dimensional Space

A second option is to map the code space to a new (higher-dimensional) space that characterizes LBP encodings. Then, an ordering or preference relationship can be learned in the extended space, for instance resorting to preference learning algorithms [97, 109, 165].

Some examples of properties that characterize LBP encodings are:

- Number of transitions of size 1 (e.g., 101, 010).
- Number of groups/transitions.
- Size of the smallest/largest group.
- Diversity on the group sizes.
- Number of ones.

Techniques to learn the final ordering based on the new high-dimensional space include:

- Dimensionality reduction techniques, including Spectral embeddings, PCA, MDS and other manifold techniques.
- Preference learning strategies for learning rankings [97, 109, 165].

A case of special interest that will be used in the experimental section of this work are Lexicographic Rankers (LR) [97, 109]. In this work, we will focus on the simplest type of LR, linear LR. Let us assume that features in the new high-dimensional space are SRk (e.g., monotonic functions) on the texture complexity of the codes. Thus, for each code

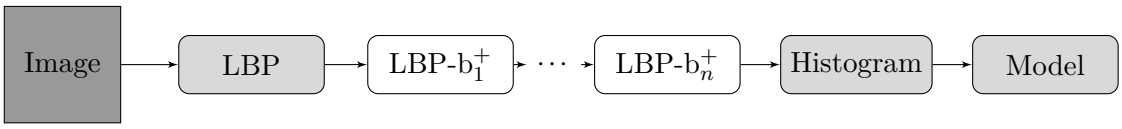


Figure 7.6: Deep LBP.

e_i and feature s_j, the complexity associated to e_i by s_j is denoted as $s_j(e_i)$. We assume $s_j(e_i)$ to lie in a discrete domain with a well-known order relation.

Thus, each feature is grouping the codes into equivalence classes. For example, the codes with zero transitions (i.e., flat textures), two transitions (i.e., uniform textures) and so on.

If we concatenate the output of the SRk in a linear manner $(s_0(e_i), s_1(e_i), \cdots, s_n(e_i))$, a natural arrangement is their lexicographic order (see Eq. (7.9)), where each $s_j(e_i)$ is subordering the equivalence class obtained by the previous prefix of rankers $(s_0(e_i), \cdots, s_{j-1}(e_i))$.

$$\text{LexRank}(a,b) = \begin{cases} a = b & ,|a| = 0 \ \vee |b| = 0 \\ a \prec b & ,a_0 \prec b_0 \\ a \succ b & ,a_0 \succ b_0 \\ LexRank(t(a), t(b)) & ,a_0 = b_0 \end{cases} \tag{7.9}$$

where $t(a)$ returns the tail of the sequence. Namely, the order between two encodings is decided by the first SRk in the hierarchy that assigns different values to the encodings.

Therefore, the learning process is reduced to find the best feature arrangement. A heuristic approximation to this problem can be achieved by iteratively appending to the sequence of features the one that maximizes the performance of the oracle Φ.

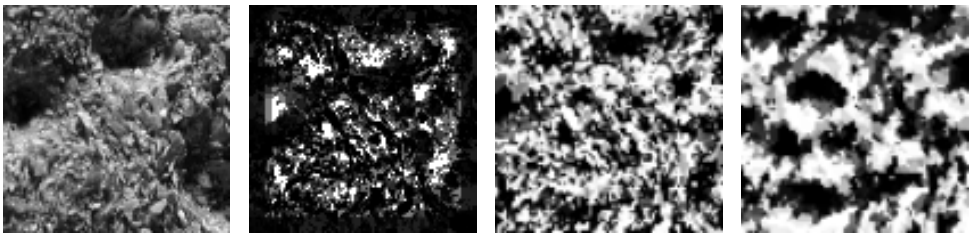
Similarly to property-aware dissimilarity functions, if the features in the new feature vector $\mathcal{V}(x)$ are invariant to P, the P-invariance of the learned binarization function is automatically guaranteed.

7.3 Deep Architectures

Given the closed form of the LBP with deep binarization functions, their recursive combination seems feasible. In this section, several alternatives for the aggregation of deep LBP operators are proposed.

7.3.1 Deep LBP (DLBP)

The simplest way of aggregating Deep LBP operators is by applying them recursively and computing the final encoding histograms. Figure 7.6 shows this architecture. The first transformation is done by a traditional shallow LBP while the remaining transformations are performed using deep binarization functions.



(a) Input image (b) LBP (c) DeepLBP(2) (d) DeepLBP(4)

Figure 7.7: Visualization of LBP encodings from a Brodatz database [37] image. The results obtained by applying n layers of Deep LBP operators are denoted as DeepLBP(n). A neighborhood of size 8, radius 10 and Euclidean distance was used. The grayscale intensity is defined by the order of the equivalence classes.

Figure 7.7 illustrates the patterns detected by several deep levels on a cracker image from the Brodatz database. In this case, the ordering between LBP encodings was learned by using a lexicographic ordering of encodings on the number of groups, the size of the largest group and IR between 0's and 1's. We can observe that the initial layers of the architecture extract information from local textures while the later layers have higher levels of abstraction.

7.3.2 Multi-Deep LBP (MDLBP)

Although it may be a good idea to extract higher-order information from images, for the usual applications of LBP, it is important to be able to detect features at different levels of abstraction. For instance, if the image has textures with several complexity levels, it may be relevant to keep the patterns extracted at different abstraction levels. Resorting to the techniques employed in the analysis of multimodal data [171], we can aggregate this information in two ways: feature and decision-level fusion.

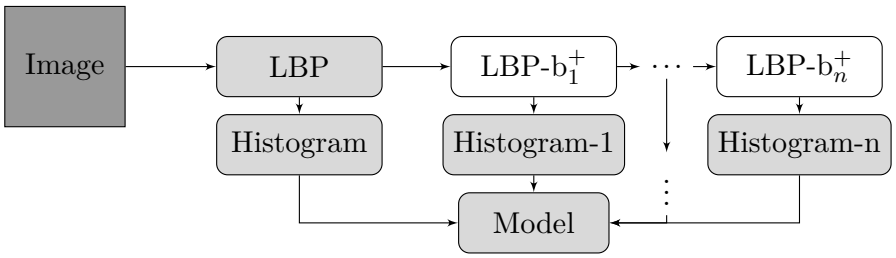
7.3.2.1 Feature-level fusion

one histogram is computed at each layer and the model is built using the concatenation of all the histograms as features.

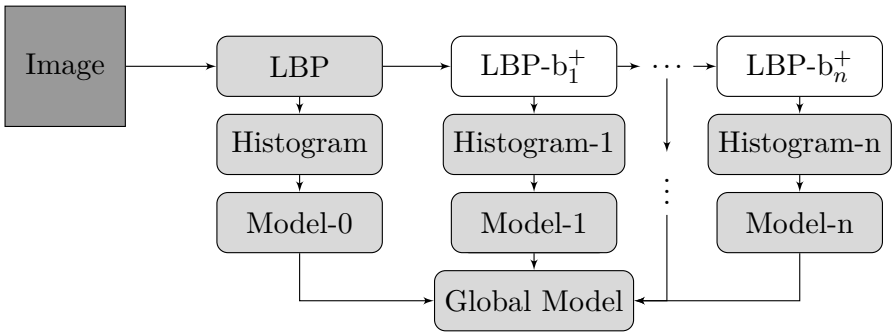
7.3.2.2 Decision-level fusion

one histogram and decision model is computed at each layer. The final model uses the probabilities estimated by each model to produce the final decision.

Figures 7.8a and 7.8b show Multi-Deep LBP architectures with feature-level and decision-level fusion respectively. In our experimental setting, feature-level fusion was used.



(a) Multi-Deep LBP with feature-level fusion.



(b) Multi-Deep LBP with decision-level fusion.

Figure 7.8: Deep LBP architectures.

7.3.3 Multiscale Deep LBP (Multiscale DLBP)

In the last few years, DL approaches have benefited from multi-scale architectures that can aggregate information from different image scales [79, 243, 333]. Despite being able to induce higher abstraction levels, deep networks are restrained to the size of the individual operators. Thereby, aggregating multi-scale information in deep architectures may exploit their capability to detect traits that appear at different scales in the images in addition to turning the decision process scale invariant.

In this work, we consider the stacking of deep independent architectures at several scales. The final decision is made by concatenating the individual information produced at each scale factor (cf. Figure 7.9). Depending on the fusion approach, the final model operates in different spaces (i.e., feature or decision level). In an LBP context, we can define the scale operator of an image by resizing the image or by increasing the neighborhood radius.

7.4 Experiments

In this section, we compare the performance of the proposed deep LBP architectures against *shallow* LBP versions. Several datasets were chosen from the LBP literature covering a wide range of applications, from texture categorization to object recognition. Table 7.2 summarizes the datasets used in this work.

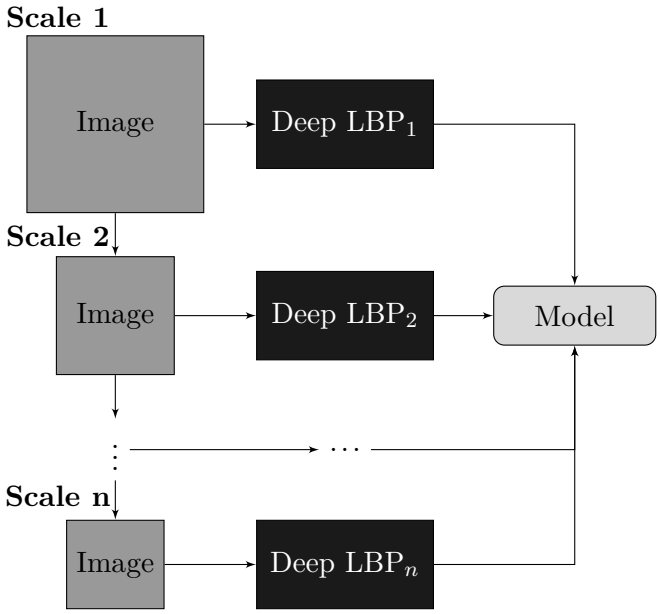


Figure 7.9: Multi-scale Deep LBP.

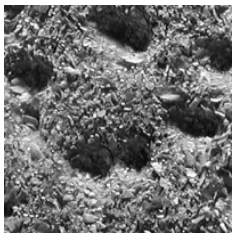
Table 7.2: Summary of the datasets used in the experiments

Dataset	Reference	Task	Images	Classes
KTH TIPS	[143]	Texture	810	10
FMD	[303]	Texture	1000	10
Virus	[288]	Texture	1500	15
Brodatz*	[37]	Texture	1776	111
Kylberg	[186]	Texture	4480	28
102 Flowers	[246]	Object	8189	102
Caltech 101	[87]	Object	9144	102

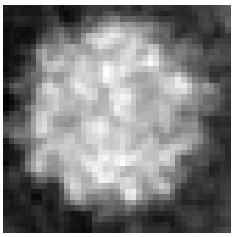
We used a 10-fold stratified cross-validation strategy and the average performance of each method was measured in terms of:

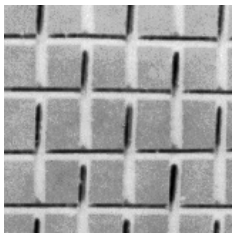
- Accuracy.
- Class rank: Position (%) of the ground truth label in the ranking of classes ordered by confidence. The ranking was induced using probabilistic classifiers.

While high values are preferred when using accuracy, low values are preferred for class rank. All the images were resized to have a maximum size of 100×100 and the neighborhood used for all the LBP operators was formed by 8 neighbors on a radius of size 3, which proved to be a good configuration for the baseline LBP. The final features were built using a global histogram, without resorting to image blocks. Further improvements in each application can be achieved by fine-tuning the LBP neighborhoods and by using other spatial sampling techniques on the histogram construction. Since the objective of this work was to compare the performance of each strategy objectively, we decided to


(a) KTH TIPS


(b) FMD


(c) Virus


(d) Brodatz


(e) Kylberg


(f) 102 Flowers


(g) Caltech 101

Figure 7.10: Sample images from each dataset

Table 7.3: Class rank (%) of the ground-truth label and accuracy with single-scale strategies

Dataset	**Strategy**	**Class Rank Layers**					**Accuracy Layers**				
		1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
KTH TIPS	LBP	1.55	-	-	-	-	89.22	-	-	-	-
	Similarity	-	1.60	1.68	1.82	1.86	-	88.96	88.57	86.99	87.97
	High Dim	-	0.94	**0.91**	1.09	1.11	-	92.96	**93.58**	92.72	92.36
FMD	LBP	26.96	-	-	-	-	29.20	-	-	-	-
	Similarity	-	25.77	25.79	25.61	25.59	-	28.90	30.00	30.90	30.80
	High Dim	-	**23.20**	23.36	23.30	23.50	-	**33.40**	32.60	33.30	33.00
Virus	LBP	8.08	-	-	-	-	56.80	-	-	-	-
	Similarity	-	6.61	6.78	6.72	6.73	-	61.00	61.33	60.93	61.93
	High Dim	-	6.65	**6.50**	6.53	6.55	-	61.53	61.27	61.47	**62.27**
Brodatz	LBP	0.25	-	-	-	-	89.23	-	-	-	-
	Similarity	-	0.22	0.22	0.23	0.25	-	89.73	90.23	90.50	90.36
	High Dim	-	**0.21**	**0.21**	0.22	0.23	-	**90.72**	**90.72**	90.09	89.59
Kylberg	LBP	0.23	-	-	-	-	95.29	-	-	-	-
	Similarity	-	0.18	0.16	0.14	0.14	-	96.14	96.52	96.72	96.81
	High Dim	-	**0.07**	**0.07**	**0.07**	**0.07**	-	**98.37**	98.35	98.26	98.24
102 Flowers	LBP	13.46	-	-	-	-	23.18	-	-	-	-
	Similarity	-	13.34	13.56	13.99	14.29	-	**25.59**	24.46	24.92	24.58
	High Dim	-	13.10	**12.99**	13.15	13.32	-	24.56	23.76	22.81	22.36
Caltech 101	LBP	13.05	-	-	-	-	39.71	-	-	-	-
	Similarity	-	12.37	12.23	12.32	12.38	-	40.35	40.07	39.74	39.81
	High Dim	-	**11.98**	12.19	12.16	12.34	-	**41.45**	40.78	40.56	40.43

fix these parameters. The final decision model is a RF with 1000 trees. In the last two datasets, which contain more than 100 classes, the maximum depth of the decision trees was bounded to 20 in order to limit the required memory.

In all our experiments, training data was augmented by including vertical and horizontal flips.

7.4.1 Single-scale

First, we validated the performance of the proposed deep architectures on single scale settings with increasing number of deep layers. Information from each layer was merged at a feature level by concatenating the layerwise histogram (c.f. Section 7.3.2). Table 7.3 summarizes the results of this setting. In all the datasets, the proposed models surpassed the results achieved by traditional LBP.

Furthermore, even when the accuracy gains are small, the large gains in terms of class rank suggest that the deep architectures induce more stable models, which assign a high probability on the ground-truth level, even on misclassified cases. For instance, in the Kylberg dataset, a small relative accuracy gain of 3.23% was achieved by the High Dimensional rule, the relative gain on the class rank was 69.56%.

With a few exceptions, the data-driven deep operator based on a high dimensional projection achieved the best performance. Despite the possibility to induce encoding orderings using user-defined similarity functions, the final orderings are static and domain independent. In this sense, more flexible data-driven approaches as the one suggested in Section 7.2.2.2 are able to take advantage of the dataset-specific properties.

Despite the capability of the proposed deep architectures to achieve large gain margins, the deep LBP operators saturate rapidly. For instance, most of the best results were found on architectures with up to three deep layers. Further research on aggregation techniques to achieve higher levels of abstraction should be conducted. For instance, it would be interesting to explore efficient non-parametric approaches for building encoding orderings that allow more flexible data-driven optimization.

7.4.2 Multi-Scale

A relevant question in this context is if the observed gains are due to the higher abstraction levels of the deep LBP encodings or to the aggregation of information from larger neighborhoods. Namely, when applying a second-order operator, the neighbors of the reference pixel include information from their own neighborhood which was initially out of the scope of the local LBP operator. Thereby, we compare the performance of the Deep LBP and multiscale LBP.

To simplify the model assessment, we fixed the number of layers to 3 in the deep architectures. A scaling factor of 0.5 was used on each layer of the pyramidal multiscale operator. Guided by the results achieved in the single-scale experiments, the deep operator based on the lexicographic sorting of the high-dimensional feature space was used in all cases.

Table 7.4 summarizes the results on the multiscale settings. In most cases, all the deep LBP architectures surpassed the performance of the best multiscale shallow architecture. Thereby, the aggregation level achieved by deep LBP operators goes beyond a multiscale analysis, being able to address meta-texture information. Furthermore, when combined with a multiscale approach, deep LBP achieved the best results in all the cases.

Table 7.4: Performance of multi-scale strategies

Dataset	**Strategy**	**Class Rank Scales**			**Accuracy Scales**		
		1	**2**	**3**	**1**	**2**	**3**
KTH TIPS	Shallow	1.55	1.17	1.22	89.22	90.94	90.93
	Deep	0.91	0.79	**0.62**	93.58	94.21	**94.96**
FMD	Shallow	26.96	26.31	26.32	29.20	29.60	29.80
	Deep	**23.36**	23.54	23.77	32.60	**33.20**	33.00
Virus	Shallow	8.08	7.51	7.97	56.80	60.60	58.60
	Deep	6.50	**5.92**	6.04	61.27	**66.13**	64.87
Brodatz	Shallow	0.25	0.20	0.23	89.23	90.77	90.00
	Deep	0.21	**0.13**	**0.13**	90.72	92.97	**93.11**
Kylberg	Shallow	0.23	0.13	0.12	95.29	97.34	97.57
	Deep	0.07	0.05	**0.04**	98.35	98.84	**98.95**
102 Flowers	Shallow	13.46	13.10	12.79	23.18	25.10	26.40
	Deep	12.99	**12.68**	12.71	23.76	26.02	**26.87**
Caltech 101	Shallow	12.92	12.46	12.28	40.07	40.84	41.03
	Deep	12.21	11.74	**11.60**	40.68	41.67	**42.01**

7.4.3 LBPNet

Finally, we compare the performance of our deep LBP architecture against the state-of-the-art LBPNet [336]. As referred in the introduction, LBPNet uses LBP encodings at different neighborhood radius and histogram sampling in order to simulate the process of learning a bag of convolutional filters in deep networks. Then, the dimensionality of the descriptors is reduced by means of PCA, resorting to the idea of pooling layers from CNN. However, the output of a LBPNet cannot be used by itself in successive calls of the same function. Thereby, it is incapable of building features with higher expressiveness than the individual operators.

In our experiments, we considered the best LBPNet with up to three scales and histogram computations with non-overlapping histograms that divide the image into 1×1, 2×2 and 3×3 blocks. The number of components kept in the PCA transformation was chosen to retain 95% of the variance for most datasets except for 102 Flowers and Caltech, where a value of 99% was chosen due to poor performance of the previous value. A global histogram was used in our deep LBP architecture

Table 7.5 summarizes the results obtained by multiscale LBP (shallow), LBPNet and our proposed deep LBP. In order to understand if the gains achieved by the LBPNet are

due to the overcomplete sampling or to the PCA transformation preceding the final classifier, we validated the performance of our deep architecture with a PCA transformation on the global descriptor before applying the RF classifier. Despite being able to surpass the performance of our deep LBP without dimensionality reduction, LBPNet did not improve the results obtained by our deep architecture with PCA in most cases. In this sense, even without resorting to local descriptors on the histogram sampling, our model was able to achieve the best results within the family of LBP methods. The only exception was observed in the 102 Flowers dataset (see Figure 7.10f), where the spatial information can be relevant. It is important to note that our model can also benefit from using spatial sampling of the LBP activations. Moreover, deep learning concepts such as dropout and pooling layers can be introduced within the Deep LBP architectures in a straightforward manner.

Table 7.5: Comparison with LBPNet

Dataset	**Strategy**	**Class Rank**	**Accuracy**
KTH TIPS	Shallow	1.17	90.94
	LBPNet	0.43	96.29
	Deep LBP	0.62	94.96
	Deep LBP (PCA)	**0.16**	**98.39**
FMD	Shallow	26.31	29.80
	LBPNet	25.71	30.00
	Deep LBP	23.36	**33.20**
	Deep LBP (PCA)	**23.20**	32.30
Virus	Shallow	7.51	60.60
	LBPNet	7.18	60.73
	Deep LBP	5.92	**66.13**
	Deep LBP (PCA)	**5.91**	65.60
Brodatz	Shallow	0.20	90.77
	LBPNet	0.20	91.49
	Deep LBP	0.13	93.11
	Deep LBP (PCA)	**0.12**	**94.46**
Kylberg	Shallow	0.12	97.57
	LBPNet	0.19	95.80
	Deep LBP	0.04	98.95
	Deep LBP (PCA)	**0.02**	**99.55**
102 Flowers	Shallow	12.79	26.40
	LBPNet	**9.61**	**35.56**
	Deep LBP	12.68	26.87
	Deep LBP (PCA)	22.30	8.80
Caltech 101	Shallow	12.46	41.03
	LBPNet	12.11	42.69
	Deep LBP	11.60	42.01
	Deep LBP (PCA)	**10.87**	**45.14**

7.5 Conclusions

LBP have achieved competitive performance in several CV tasks, being a robust and easy to compute descriptor with high discriminative power on a wide spectrum of tasks. In this work, we proposed Deep LBP, an extension of the traditional LBP that allow successive applications of the operator. By applying LBP in a recursive way, features with a higher level of abstraction are computed that improve the descriptor discriminability.

The key aspect of our proposal is the introduction of flexible binarization rules that define an order relation between LBP encodings. This was achieved with two main learning paradigms. First, learning the ordering based on a user-defined encoding similarity metric. Second, allowing the user to describe LBP encodings on a high-dimensional space and learning the ordering on the extended space directly. Both ideas improved the performance of traditional LBP in a diverse set of datasets, covering various applications such as face analysis, texture categorization and object detection. As expected, the paradigm based on a projection to a high-dimensional space achieved the best performance, given its capability of using application-specific knowledge efficiently. The proposed deep LBP can aggregate information from local neighborhoods into higher abstraction levels, being able to surpass the performance obtained by multiscale LBP as well.

While the advantages of the proposed approach were demonstrated in the experimental section, further research can be conducted on several areas. For instance, it would be interesting to find the minimal properties of interest that should be guaranteed by the binarization function. In this work, since we are dealing with intensity-based image, we restricted our analysis to partial orderings. However, under the presence of other types of data such as directional (i.e., periodic, angular) data, cycling or local orderings could be more suitable. In the most extreme case, the binarization function may be arbitrarily complex without being restricted to strict orders.

On the other hand, constraining the shape of the binarization function allows more efficient ways to find suitable candidates. In this sense, it is relevant to explore ways to improve the performance of the similarity-based deep LBP. Two possible options would be to refine the final embedding by using training data and allowing the user to specify incomplete similarity information.

In this work, each layer was learned in a local fashion, without space for further refinement. While this idea was commonly used in the DL community when training stacked networks, later improvements take advantage of refining locally trained architectures [248]. Therefore, we plan to explore global optimization techniques to refine the layerwise binarization functions.

DL imposed a new era in CV and ML, achieving outstanding results on applications where previous state-of-the-art methods performed poorly. While the foundations of DL rely on very simple image processing operators, relevant properties held by traditional methods, such as illumination and rotational invariance, are not guaranteed. Moreover,

the amount of data required to learn competitive deep models from scratch is usually prohibitive. Thereby, it is relevant to explore the path to a unification of traditional and DL concepts. In this work, we explored this idea within the context of LBP. The extension of deep concepts to other traditional methods is of great interest in order to rekindle the most fundamental concepts of CV to the research community.

Chapter 8

Image Segmentation by Quality Inference

This chapter was published in [101]:

- Kelwin Fernandes, Ricardo Cruz, and Jaime S. Cardoso. Image segmentation by quality inference. In *Neural Networks (IJCNN), 2018 International Joint Conference on*. IEEE, 2018

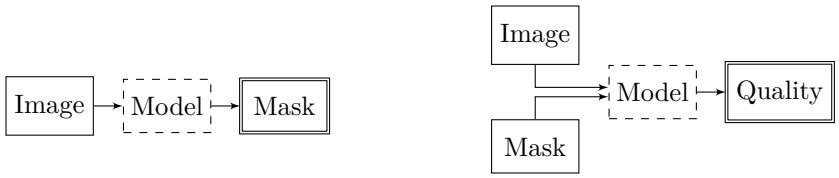
This work was done in collaboration with Ricardo Cruz, who designed and implemented the artificial mask generation procedure (see Section 8.3.2.2).

Traditionally, CNN are trained for semantic segmentation by having an image given as input and the segmented mask as output. In this work, we propose an ANN trained by being given an image and mask pair, with the output being the quality of that pairing. The segmentation is then created afterwards through backpropagation on the mask. This allows enriching training with semi-supervised synthetic variations on the ground-truth. The proposed iterative segmentation technique allows improving an existing segmentation or creating one from scratch. We compare the performance of the proposed methodology with state-of-the-art deep architectures for image segmentation and achieve competitive results, being able to improve their segmentation.

8.1 Introduction

Segmentation of images into its constituent parts is a decades-old problem. Traditional methods range from the usage of a color threshold to clustering, and iterative methods such as region growing and active contours. However, all these methods require strong human supervision and tuning to find the right parameters.

The advent of ML, in particular CNN like SegNet [24], has allowed semantic segmentation – where the parameters of the model are optimized automatically in a supervised



(a) Traditional approach (b) Proposal

Figure 8.1: Diagram representing segmentation flows.

manner on the object of interest. These new methods lack the iterative nature of previous techniques. The downside of such methods is the vast amount of data required for training. Furthermore, applying these models to slightly different contexts, without re-training or fine-tuning, proves problematic.

Opposite to how ML algorithms are trained, as humans, we do not have a single ground-truth solution on our daily tasks, but a spectrum of choices that can fulfill our goal to a certain degree. From economics and social choice perspective, this decision process usually involves a utility function that reflects our satisfaction degree about a solution [78,107,190]. Based on such utility function, we can apply local (and non-local) updates to fulfill the requirements. In this work, we propose a novel segmentation paradigm: the CNN is trained to learn the quality of an (image, segmentation) pair. For a given dataset of images, multiple possible (synthetically-created) segmentations of varying qualities are used in the training process. The model not only has more information, but the problem complexity is reduced. The output, instead of having size $2^{\text{width}\times\text{height}}$ (the segmentation probability mask), is a single number (the quality of the given segmentation). This is represented in a diagram in Figure 8.1.

Once training is performed, the segmentation process is no longer done in a single forward-pass like in the traditional approach. In our proposal, the segmentation process also makes use of the network as an oracle of the current segmentation quality to refine the mask iteratively. In order to do this, we rely on the backpropagation algorithm. This iterative process is inspired by previous segmentation techniques such as region growing, and the human visual system; human design evolves steps of "anticipated emergence" – sketching, in particular, involves seeing-moving-seeing steps [255]. The proposed model is an iterative process that can, not only produce segmentations from scratch, but also improve on those provided by an existing model.

8.2 State-of-the-art

Many traditional CV techniques have involved iterative processes. This is the case, for example, of region growing and active contours (also known as snakes).

In **region growing** [86], the segmentation is initialized from a seed point $R(0)$ at time 0, and then grows to include its neighbor pixels $\mathbf{N}$, $R(t+1) = \bigcup_j R(t) \cup N_j$ according to a

user-provided logical predicate $P(R \cup N_j)$. In **active contours** [172], a curve composed of discrete points $\mathbf{v}(s) = \{(x(s), y(s))\}$, indexed by $s \in [0, 1]$, is found by minimizing an energy function $E_{\text{snake}} = \int_0^1 [E_{\text{internal}}(\mathbf{x}(s)) + E_{\text{external}}(\mathbf{x}(s))]\, ds$. The E_{internal} acts as a regularizer punishing many oscillations in the curve, while E_{external} is a function of the intensity or gradients within the image and can be both negative (repellent) or positive (attractor).

These traditional techniques have recently been surpassed by CNN, which are capable of "semantic segmentation"; i.e., the ground-truth is used to guide the learning process.

The most widely used architectures are based on an **encoder-decoder** two-phased ANN; the image is first compressed into a smaller semantic representation, usually using convolutions and pooling (the encoding phase), and then decompressed into the final segmentation, usually using convolutions transposes (the decoding phase). The first example of this was SegNet [24].

A big problem in the encoding-decoding strategy is in avoiding the so-called checkerboard problem. Some detail is lost during the encoding step, which prevents the decoding step of doing as good a job as it could in refining the segmentation. This can result in segmentations with a checkerboard effect. Since the encoding-decoding phases of the ANN are symmetric, U-Net [283] created so-called "skip-layers" where each decoding layer ℓ does not only receive as input the activation output of the previous layer $\ell - 1$, but also of the symmetric layers $L - \ell$ from the encoding phase, where L is the number of layers and $\ell > \frac{L}{2}$. In summary, each encoding layer computes the usual function $a^{(\ell)} = f(a^{(\ell-1)})$ and each decoding layer computes the function $a^{(\ell)} = f(a^{(\ell-1)}, a^{(L-\ell)})$ which is also using information from the encoding phase. It should be pointed out that U-Net was the best performing model in the ISBI 2016 Skin Lesion Analysis Towards Melanoma Detection Challenge [141], which supports its choice as one of the baseline models in this work.

Another important landmark in avoiding the checkerboard effect are **dilated kernels** (originally known as atrous convolutions). DeepLab [44], which makes use of such kernels, ranks first place in many benchmarks, including PASCAL VOC. There are no distinct encoding and decoding phases which produce activation maps of varying size. In this model, the activation maps remain the same size across the network. Filters are interconnected to the layers in a way, that each weight is shared across the same activation, so that the activation produced can have the same dimension along the network. Such a model is also used as a baseline in this work.

Also, worth mentioning is that iterative segmentation already exists in the form of Recurrent Neural Network (RNN) adapted for segmentation [370]. The current work is innovative in that it is far simpler than any previous approach, since it most resembles traditional architectures used for segmentation.

8.3 Deep Segmentation by Quality Inference

The main idea of this chapter can be summarized as follows:

1. learn to evaluate the quality of a certain segmentation mask for a given image;

2. use the model learnt in 1) to find a (local) optimal segmentation by walking in the space of segmentation masks.

The proposed idea is illustrated in Figure 8.2, where the image-mask is iteratively fed to the quality oracle in order to estimate the correspondence between them. Then, a search procedure is used to discover potential improvements to the input mask. This process is repeated until a convergence criterion (e.g., desired quality, improvement tolerance, number of iterations) is met.

We propose to achieve 1) by using Deep CNN and 2) by using gradient ascent (back-propagation) over the input mask. We argue that it is more robust to learn to evaluate the quality of a given image/segmentation pair than to learn how to segment the image. Also, the quality concept has the potential to be more generic and easily transferred between tasks.

8.3.1 Quality Inference as Deep Similarity Learning

In this section, we address several aspects of the construction of a model capable of predicting the quality of a semantic image segmentation mask. How to express a utility of an entity (e.g., commodity, good, segmentation mask) is an open problem in economics and social choice [190]. The main choices for modeling utilities are pairwise preferences [97] and cardinal/ordinal functions [81]. These paradigms can be mapped to similarity learning as regression and ranking similarity learning respectively. To simplify the learning process (i.e., decision models and optimization), we model the utility (quality) of a segmentation mask as a cardinal function. Thereby, we are interested in regression models where pairs of objects – image i and mask m – are given to the model, being the quality $\hat{q}$ the outcome of the model. In our case, the utility is a measure of quality or correspondence between the mask and the image. This can be learnt by minimizing the regularized loss function

$$\min_{\theta} \sum_{k} \mathcal{L}_{\theta}\left(f(i_k, m_k), q_k\right) + \lambda \mathcal{R}(\theta), \tag{8.1}$$

where $\mathcal{L}$ can be instantiated to the squared error of the estimated quality and the corresponding ground-truth quality and $\mathcal{R}$ is a regularizer of the model complexity (e.g., L_2). In this approach, we assume that, during training, a quality function for each image-mask pair is available.

The most straightforward strategy to solve this problem would be to use a traditional CNN where the mask is appended to the image as an additional channel (see Figure 8.3). These two data sources (i.e., image and mask) belong to different categories (real-valued and binary) but are handled by the same operation (i.e., convolution) which may difficult the learning process.

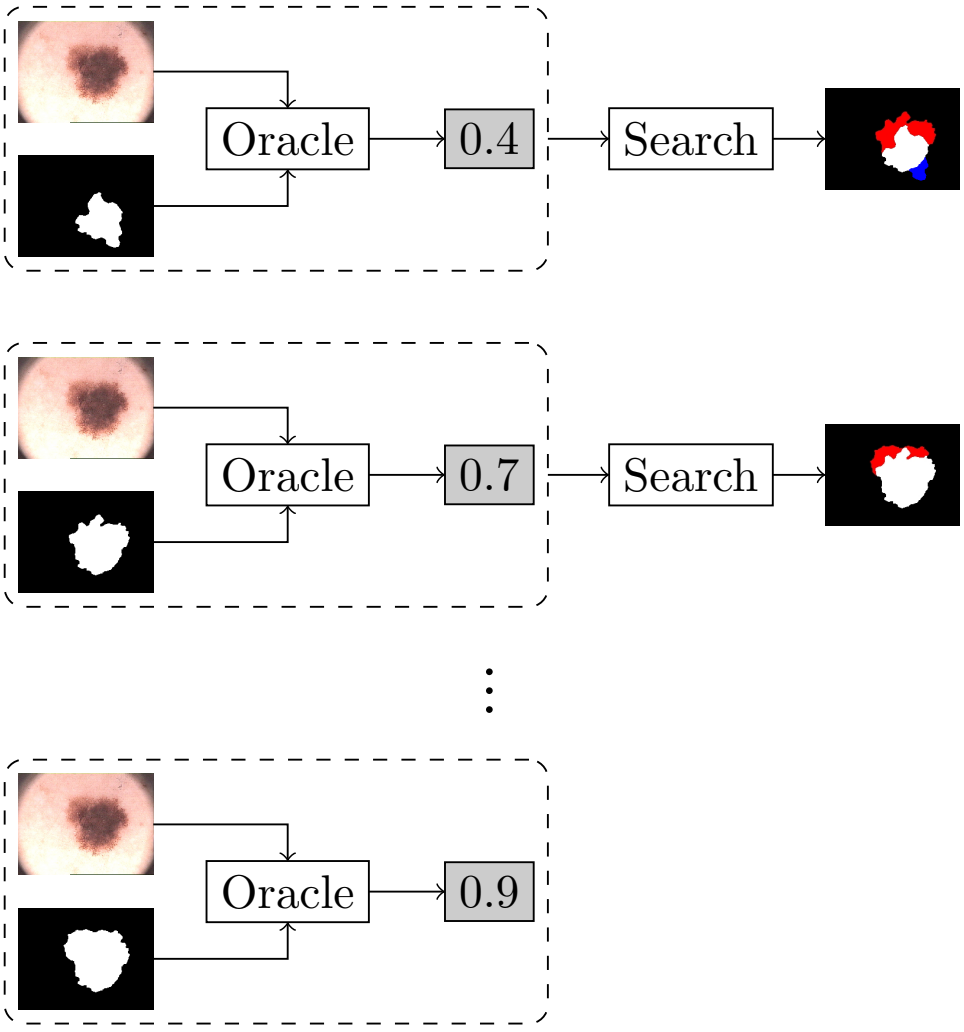


Figure 8.2: Illustration of the iterative process of quality estimation and improvement. The search procedure indicates that red/blue regions should be added/removed from the input mask to improve the quality estimated by the oracle.

An alternative approach would be to have separate streams for the input image and masks (see Figure 8.4), being merged in the final dense blocks by concatenating their latent representation. The main drawback of this model would be that as we move deeper through the network, the intrinsic loss of resolution would limit the analysis of low-level patterns. Moreover, since each stream works with a different type of data, it is not clear how similar would be their latent representation.

In this work, we propose a deep architecture to tackle this problem, allowing an early integration of the information from image and masks. The main intuition behind this architecture is (i) having two streams that attempt to model the regions defined as foreground and background respectively by the input mask; (ii) streams communicate – "gossips" – to each other in order to increase/decrease their confidence on the recognition of their corresponding regions. In the rest of this section, we formalize the proposed architecture

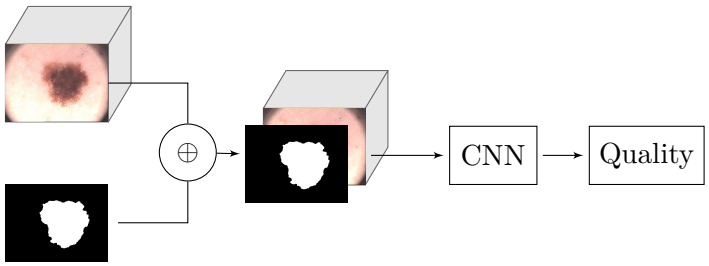


Figure 8.3: Diagram representing a potential single-mixed stream approach to the problem.

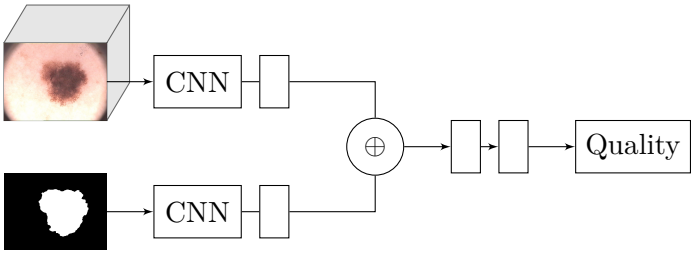


Figure 8.4: Diagram representing a potential dual stream approach to the problem.

and its training procedure.

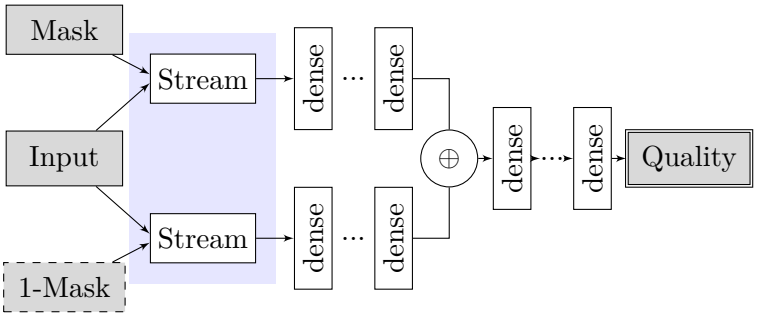


Figure 8.5: Diagram of the general Gossip network.

8.3.1.1 Gossip Networks

Gossip networks are structured in such a way that the foreground and background representation is modeled by a pair of streams. This architecture is best described in three degrees of scale:

1. The **general architecture** is composed of the initial split into background and foreground streams, followed by dense layers, which produce the final quality score (see Figure 8.5).

2. Within each **stream**, gossip blocks are consecutively intertwined with traditional convolutions; pooling is applied at the end of the stream to feature maps and masks (see Figure 8.6).

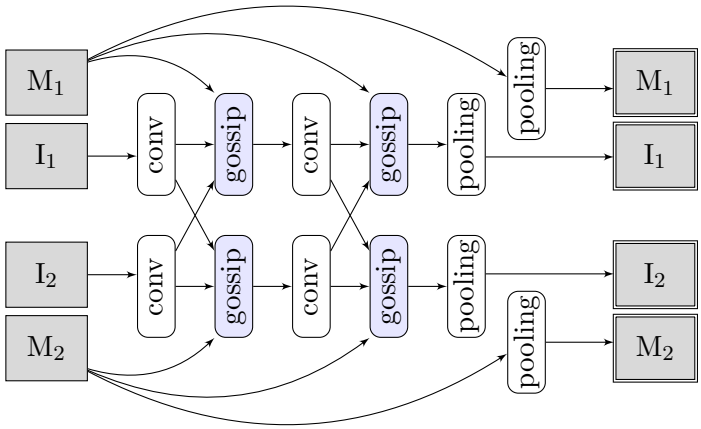


Figure 8.6: Diagram of the two streams containing the gossip blocks.

3. At the lower-level, the **gossip block** combines the information from the two streams (see Figure 8.7). The gossip block receives the feature maps obtained by 2D convolutional layers for the corresponding stream S and reciprocal layer $\hat{S}$. Then, for the Region of Interest (ROI) of each channel, we penalize the activations where the reciprocal channel has stronger activations than the current one. We set the non-linearity term of these layers to the penalty term to favor the propagation of gradients to the original source pixels at inference. Namely, we avoid the problem of dead units where gradients are zero [145].

This type of double-helix connection between streams seen in the diagram was used to ensure an early interaction between both streams to reinforce/penalize their assumptions on each resolution-level.

The propagation of gradients through max-pooling blocks is sparse, leading to an unstable refinement of the segmentation masks (see section 8.3.3). Thereby, we decided to use average pooling that ensures gradients are propagated through all the pixels in the block. Also, we restrain the activations of the convolutions to the valid regions to avoid unbalanced magnitude of the gradients in the edges of the image.

8.3.2 Training

Here, we describe how to efficiently train the proposed architecture in order to cover the space of potential segmentation masks in an efficient manner.

8.3.2.1 Similarity Metric

The similarity metric used in this work was the Sørensen-Dice coefficient D, often referred to as simply the Dice Coefficient. This index is given by the intersection over the union

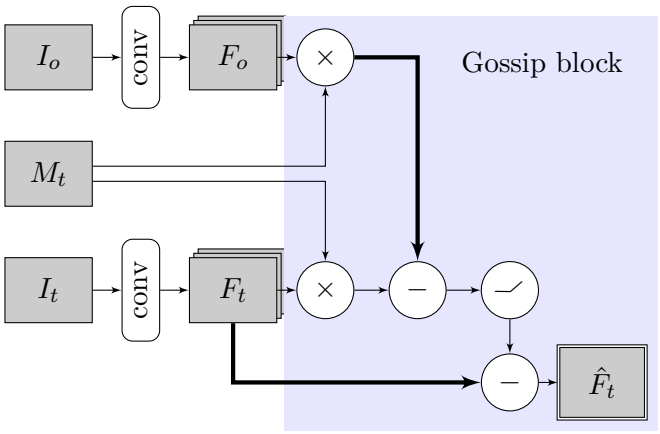


Figure 8.7: Diagram of the gossip block. Thick arrows define the first argument of the operations that are not commutative.

of the true and predicted masks,

$$D(Y,\hat{Y}) = \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|}. \tag{8.2}$$

The index may be seen as a kind of F_1 score, hereby ensuring both positives and negatives (mis)classifications are captured equally in the metric.

8.3.2.2 Transformations

Two levels of transformations were applied: at the level of the ground-truth image and mask pairs, and then further transformations were applied to synthetically generate different segmentations of varying degrees of quality similarities – the ability to perform this latter data augmentation process is one of the key features that makes the Gossip architecture stand out from the current state-of-the-art.

Traditional data augmentation was applied to the **ground-truth** image and mask pairs. These transformations encompass random horizontal and vertical flips, horizontal and vertical shifts, random zoom scales, as well as shear and contrast stretching deformations. These transformations were customized for each trained dataset.

Figure 8.8: Examples of synthetically created segmentations.

Furthermore, the Gossip architecture can be trained for new segmentation masks created **synthetically** – with their corresponding similarity metric. The following transformations have been used. Notice all of them have one or multiple parameters, which are listed in brackets.

- Elastic deformation (α, σ, α'), which is a type of local, affine distortion [309]
- Morphological erosion and dilation (size)
- Random pixel switching (#pixels)
- Rotations (angle)
- Flip transformations (horizontal and/or vertical), and horizontal and vertical mask shifts (xoffset, yoffset).

These transformations are illustrated in Figure 8.8. The parameters of these transformations were transversed in order to provide a balanced range of qualities of Dice. There is an inverse relation between the magnitude of each one of these parameters and the similarity index, but quantifying this relation is not straightforward. For this reason, the following procedure was applied.

First, the impact of each combination of transformations and respective parameters had on the similarity index was computed empirically. For each transformation, parameters were drawn by grid-search, and a similarity index $D(Y,Y')$ was computed between the ground-truth mask Y and the synthetically created Y'. Dice was discretized into B bins (in our case, $B=8$). A frequency distribution p_{ib} was then found, representing the number of times p that the parameter combination i resulted in Dice b.

A couple of these transformations are stochastic (elastic and random pixel switch), therefore these two transformations were repeated 10 times for each ground-truth mask.

After this initial distribution was empirically computed, a second distribution was then computed from which we could sample parameters in order to ensure the similarity index was being drawn equitably across all bins (so that dice was evenly represented). This was performed by finding $\boldsymbol{\beta}$ which minimized the system composed of B equations, representing each bin as

$$\begin{cases} \boldsymbol{\beta} \cdot \mathbf{p}_1 = \frac{1}{B} \\ \quad \vdots \\ \boldsymbol{\beta} \cdot \mathbf{p}_B = \frac{1}{B}. \end{cases} \tag{8.3}$$

This was solved as a non-negative linear square problem, in order to ensure that each β represented a probability, which was then used to draw the parameters.

Notice there is a different $\boldsymbol{\beta}$ for each dataset and transformation. The reason why each transformation had their own $\boldsymbol{\beta}$ was to ensure that each transformation was used uniformly. Otherwise, some transformations might have overshadowed others. Notice

that horizontal/vertical flips and mask shift offsets transformations were combined into a single transformation (see the previous bullet list) because it was not possible to create an equitable distribution from which to sample mask flips alone.

8.3.3 Improving Segmentation by Backpropagation

During training, the Gossip network uses traditional gradient descent by computing the gradients of predicted quality relative to each weight w_i within the network, $\frac{\partial \mathcal{L}}{w_i}$. Each weight is then updated in the opposite direction of the gradient, $w_i \leftarrow w_i - \alpha \frac{\partial \mathcal{L}}{w_i}$, using a learning rate α. This step is known as backpropagation.

On segmentation inference, inspired by the literature on generating adversarial examples [126], we propose improving a given segmentation by performing backpropagation on the predicted segmentation mask $\hat{m}_{ij}$ by minimizing $\frac{d\mathcal{L}(\hat{q},q)}{d\hat{m}_{ij}}$, after the model has been trained. Here, $\hat{q} = f(i,m)$ and q is a constant for the best possible similarity.

We have found, as reported below, that this technique works well both for improving existing segmentations, as well as creating segmentations from scratch, by starting with a black mask. An initial mask $\hat{m}_{ij}$ is then updated using the following rule,

$$\hat{m}_{ij} \leftarrow \hat{m}_{ij} - \alpha \frac{\partial \mathcal{L}}{m_{ij}}. \tag{8.4}$$

Some architectural design choices were based on allowing this use-case. To avoid coarse gradients, average pooling has been used, rather than the more traditional maximum pooling approach. The derivative of average pooling is the averaging constant, while the derivative of maximum pooling is 1 for one pixel (the activation pixel) and 0 for all others. Empirically, this resulted in a very coarse segmentation.

For better convergence, gradients are normalized ($\frac{\mathcal{L}}{m_{ij}} \in [-1,1]$) per mask and a sigmoid smoothness is applied ($S(x) = \frac{1}{1+\exp -kx}$). Both this k and the number of backpropagation iterations were found empirically for each dataset using the validation set, and a fixed learning rate α was used.

An alternative to backpropagation would have been using an exhaustive or heuristic exploration of the space of segmentation masks using the network as a fitness oracle. While these techniques would be able to discover non-local improvements, backpropagation stands as an efficient exploration strategy when the decision function is known and C^1 (differentiable).

8.4 Experiments

In this section, we provide a detailed analysis of the experiments and results. First, we describe the datasets and baselines used in the validation of the proposed strategy. Then, we validate the performance of the proposed strategy on these datasets and on cross-database applications, where models are trained and validated on different datasets.

8.4.1 Data

Table 8.1: Summary of the datasets used in this work. FS denotes the average relative foreground size.

Dataset	Ref.	# Imgs.	% FS
SmartSkins	[331]	80	37.5
PH2	[232]	200	49.1
ISBI 2017	[141]	2750	9.3
Teeth-UCV	[102]	100	23.7
Breast-Aesthetics	[40]	120	19.1
Cervix-HUC	[95]	287	5.8
Cervix-MobileODT	[167]	1613	17.1
Mobbio	[299]	2164	5.1

We validated the performance of the proposed architectures on six real-life biomedical datasets. The datasets cover applications on the segmentation of melanoma, teeth, breast, cervix, and iris. Further details and sample images are shown in Table 8.1 and Figure 8.9 respectively. These datasets were chosen to provide a range of segmentation of diverse complexity used in real clinical applications.

The goal of the first three datasets (i.e., SmartSkins, PH2, and ISBI 2017) is to segment skin lesions in dermoscopic images. The task on the Teeth-UCV, Breast-Aesthetics and Mobbio databases is to segment teeth, breasts, and iris from the background on natural RGB images. Finally, the object of interest in Cervix-HUC and Cervix-MobileODT datasets is the cervix, being the images acquired using digital colposcopy with several modalities (i.e., Hinselmann, Schiller and Green filter [94]).

We divided all the datasets into training, validation and test sets following the standard 60-20-20 partitioning. All images were first resized to 128×128 for easy comparison.

8.4.2 Models

To validate the performance of the proposed technique, we compared our results with the state-of-the-art U-Net [283] and U-Net with Dilated Convolutions (DilatedNet) [44]. For each model, we choose the best number of blocks (i.e., two convolutional layers and one pooling layer) on the validation set within the interval 2, 3 and 4. We use 32 filters on the first convolutional layer and double the value on each level as typically done in the literature [283]. The loss function for the Gossip Network was the mean squared error, and the networks were trained with 2, 3 and 4 gossip-gossip-pooling blocks, with one and two dense layers per each stream and in the final common section. ReLU activations were used on the intermediate dense layers and a sigmoid activation on the final layer to predict a quality value between 0 and 1.

(a) SmartSkins (b) PH2 (c) ISBI 2017 (d) Teeth-UCV

(e) Breast Aesthetics (f) Cervix-HUC (g) Cervix-MobileODT (h) Mobbio

Figure 8.9: Sample images and masks from the several datasets used for training.

We trained the models using Adam [178] for a maximum number of 500 iterations. To avoid overfitting, early-stopping was used after 50 iterations without improvement, and the best validation model was used.

8.4.3 Results

First, we explore the performance of the model in the most extreme scenario, where the initial segmentation is empty (i.e., black mask). Figure 8.10 shows the performance of

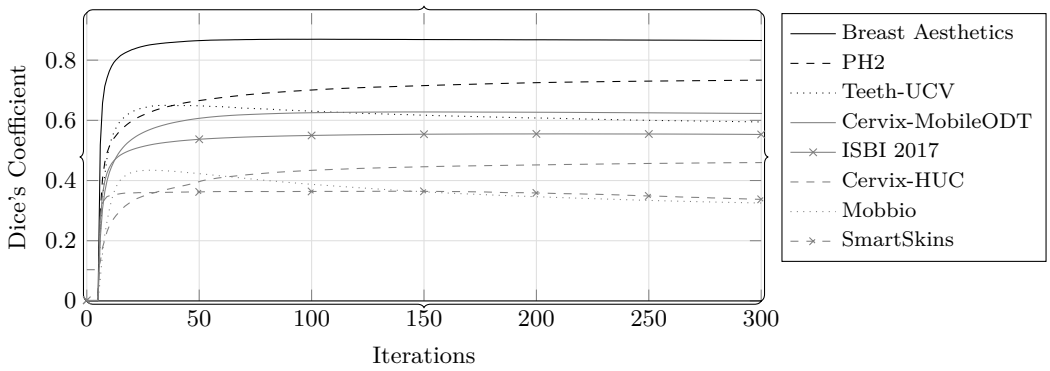
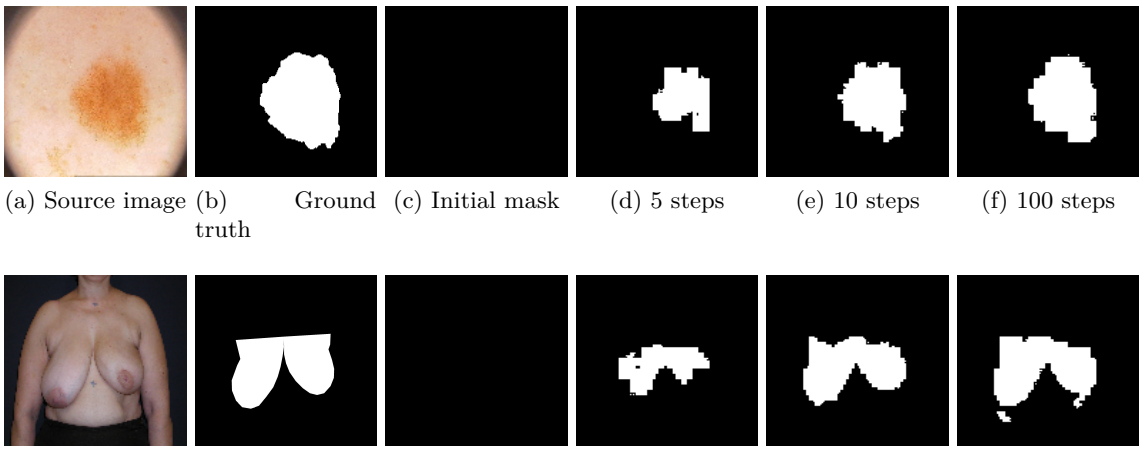


Figure 8.10: Average Gossip Network performance after N iterations of refinement starting from empty masks.

the network after N iterations of refinement. As can be seen in the Figure 8.10, the network converges to a good solution on about 20 iterations. The remaining steps of the optimization focus on minor details with little impact on the overall performance. Some degenerated cases were observed where the network performance decayed after some iterations. This effect is the result of miss-estimations of the quality function, where the network was not able to learn the right direction for improvement. Figure 8.11 illustrates the network progress at several stages of the refinement. As can be seen in the figure, the proposed approach emulates the traditional region-growing strategy [86], where the mask is progressively extended.



(a) Source image (b) Ground truth (c) Initial mask (d) 5 steps (e) 10 steps (f) 100 steps

(g) Source image (h) Ground truth (i) Initial mask (j) 5 steps (k) 10 steps (l) 100 steps

Figure 8.11: Iterative refinement of images from PH2 and Breast Aesthetics datasets, respectively, using Gossip Networks. Initial masks are completely void.

The second scenario we explored was the iterative refinement of base segmentation done by the UNet and DilatedNet architectures. In this case, we choose the best number of refinement steps on the validation set. As can be seen in Table 8.2, the proposed strategy

improved the performance of the UNet and DilatedNet architectures in all databases.

Finally, we validate the performance of the proposed model on cross-database scenarios, where the model was trained for a given task and validated on a different dataset. This is common in applications where training data is synthetic due to field restrictions (e.g., aerospace) and cross-sensor applications. Results of this experiment are presented in Table 8.3. In the first two cases, we use datasets for melanoma segmentation. We can observe large gains achieved by the Gossip network being initialized by the U-Net and Dilated-Net masks. For the validation of cervix segmentation (i.e., Cervix-MobileODT to Cervix-HUC), we observe a drop in the model performance when compared to training on the Cervix-HUC dataset directly. However, the Gossip Network achieves better performance than its counterparts. The last case covers cross-domain transitions, from melanoma to cervix segmentation. We observe a gain of about 6% when comparing the U-Net and Gossip Networks. The intuition behind these gains is that the notion of segmentation quality is more robust to changes in the data distribution. Namely, some concepts associated to predicting the quality of a segmentation such as the alignment between edges, the difference of contrast between foreground and background can be easily transferred among tasks.

Table 8.2: Model performance in terms of Dice's coefficient. Best results per database are presented in bold.

Dataset	**U-Net**		**Dilated-Net**	
	Original	**Gossip**	**Original**	**Gossip**
SmartSkins	76.62	79.45	76.35	**83.36**
PH2	83.70	84.09	85.52	**86.41**
ISBI 2017	71.35	**76.52**	72.06	76.11
Teeth-UCV	85.85	85.91	86.03	**86.14**
Breast Aesthetics	93.08	93.31	94.03	**94.15**
Cervix-HUC	77.25	**77.26**	75.37	75.37
Cervix-MobileODT	88.24	**88.25**	86.38	**88.25**
Mobbio	67.91	68.23	69.90	**70.11**

Table 8.3: Cross-database Model performance in terms of Dice's coefficient

Source	**Target**	**U-Net**		**Dilated-Net**	
		Original	**Gossip**	**Original**	**Gossip**
PH2	**SmartSkins**	76.87	81.21	75.71	**81.60**
PH2	**ISBI 2017**	64.44	67.02	66.13	**72.10**
Cervix-MobileODT	**Cervix-HUC**	57.94	**57.99**	32.18	36.00
PH2	**Cervix-HUC**	44.44	50.28	60.42	**60.62**

8.5 Conclusion

This chapter addresses the problem of semantic image segmentation with DNN. We propose a new paradigm based on similarity learning techniques that tries to learn a quality

function that maps an image-mask pair to the corresponding segmentation quality. Using the proposed architecture and, in combination with backpropagation, the proposed strategy is able to improve segmentation masks by maximizing the expected quality. By framing the problem as a regression task, we reduce the output complexity. Moreover, we are able to exploit the dataset size by learning from a synthetically generated large number of candidate segmentation masks with their corresponding quality values.

We validated the proposed strategy in several biomedical applications and achieved the best results when compared with the state-of-the-art U-net and Dilated-Net architectures. Also, we validated the proposed approach on cross-database scenarios and achieved promising results.

As future work, we intend to explore pairwise approaches based on the triplet loss [294], where the learning process is driven by comparing the outcome of pairs of masks for a single image. The proposed network could also be used for dynamic ensembles of models from which to produce the final segmentation. Namely, the importance of each model on the decision can rely on the quality predicted by the proposed network.

Smaller details that could be improved are the fixed learning rate and the fixed number of iterations used on the segmentation by backpropagation part of the work.

Part II

Automated Processing of Digital Colposcopies

"Valeu a pena? Tudo vale a pena
Se a alma não é pequena."

Mar Português
Fernando Pessoa

Summary of the Contributions

The second part of this work is devoted to the analysis of applied contributions to the automated analysis of digital colposcopies for the analysis of cervical cancer and forensic assessment of sexual assault.

Literature Review and Database

Despite the vast amount of work in the automated analysis of digital colposcopies, there is no unified framework for discussion and assessment of DSS in this area. Thus, the main open challenges, areas of actions and previous contributions in the field are scattered. In Chapter 9, we define and characterize the main areas of research surrounding the automated analysis of digital colposcopies. We study the main problems that have been tackled in the past by presenting a comparative review of the literature in the area, from the segmentation of the image in the main constituent anatomical parts of the cervix to the final diagnosis support. Also, we define the main open challenges in ML and CV for the automated analysis of digital colposcopies. Besides the analysis of the literature, we acquired, annotated and published a video database that serves as a common ground for the evaluation of the main tasks in the area.

Temporal Segmentation

In order to promote the acceptance of DSS by healthcare practitioners, the data acquisition protocols should be close to the daily routine of medical teams. Typical approaches in the development of CAD systems constrain the acquisition settings to unrealistic scenarios, introducing additional complexity to the already complex analysis of human patients. However, an unconstrained acquisition setting induces new sources of noise and uncertainty, including the appearance of irrelevant scenes in the colposcopy video. Thus, in Chapter 10, we propose a framework to remove such noisy scenes and to identify the stage of the protocol that is being executed at each frame of the video. This system can be used as a preprocessing step to facilitate the work of decision models tailored to each acquisition modality.

Cervix Segmentation

The segmentation of the anatomical parts of the cervix is a challenging problem due to the high semantic level involved in the characterization of each region and to the lack of clear boundaries between each part, limiting the applicability of traditional unsupervised methodologies used in CV. In Chapter 8, we proposed a deep methodology for image segmentation, finding satisfactory results in the segmentation of the cervix.

In several applications, including the analysis of digital colposcopies, objects of interest hold a spatial arrangement that is known *a priori*. The main anatomical parts of the

objects observed in a typical colposcopy follow an ordinal arrangement, being nested one inside the other. For instance, the external orifice, the transformation zone, the cervix, the vaginal walls and the colposcope are arranged one inside the other. In Chapter 11, we propose a DL approach for image segmentation that promotes ordinal arrangements between objects in the image. Besides the evaluation of colposcopic images, we instantiate the problem to other biomedical tasks where ordinal arrangements are naturally found, including the segmentation of the iris, teeth, and breasts.

Risk Prediction and Quality Assessment

In developing countries, resources are insufficient and patients usually have poor adherence to routine screening due to low problem awareness. Consequently, the prediction of the individual patient's risk becomes a fundamental problem in the proper management of cervical cancer screening programs. On the other hand, most of these screening methods highly depend on the physician expertise and subjective comfort on the decision process, being a key aspect to improve data acquisition using the physician preferences.

Thus, the multi-modal and multi-expert nature of these problems imposes critical challenges to ML methodologies, requiring to learn robust models with scarce data from each modality and expert. To tackle this problem, we instantiate the HTL framework based on structural similarity that was proposed in Chapter 5 to these two tasks: cross-modality individual risk and cross-expert subjective quality assessment of colposcopic images for different modalities. More specifically, we validate the performance of transferring the sign of the coefficients in linear models.

Forensic Assessment of Sexual Assaults

Finally, we study the application of forensic assessment of sexual assaults using digital colposcopies in Chapter 13. We propose an end-to-end framework to support the forensic decision covering several computational tasks that were mentioned in the previous parts of this work, from modality recognition to the segmentation and characterization of lesions, to the final classification of each case according to the forensic assessment. We validate the performance of both, traditional methodologies in CV and DL strategies. Also, we validate the relevance of the ranking-based strategy for class imbalance proposed in Chapter 3, extending the proposed methodology to DL.

Chapter 9

Automated Methods for the Decision Support of Cervical Cancer Screening using Digital Colposcopies

This chapter was published in [96]:

- Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. Automated methods for the decision support of cervical cancer screening using digital colposcopies. *IEEE Access*, 2018

Cervical cancer remains a significant cause of mortality in low-income countries. However, it can often be cured by removing the affected tissues when detected in early stages. Thereby, it is relevant to provide universal and efficient access to cervical screening programs, being digital colposcopy an inexpensive technique with high potential of scalability. The development of CAD systems for the automated processing of digital colposcopies has gained the attention of the CV and ML communities in the last decade, giving origin to a wide diversity of tasks and computational solutions. However, there is a lack of a unified framework to discuss the main tasks and to assess their performance. Thus, in this work, we studied the core research lines surrounding the automated analysis of digital colposcopies and built a topology of problems and techniques, including their key properties, advantages, and limitations. Also, we discussed the open challenges in the area and released a database that serves as a common basis to evaluate such systems.

9.1 Introduction

Cervical cancer remains a significant cause of mortality in low-income countries [174]. Despite the possibility of prevention with regular cytological screening, cervical cancer is the cause of more than 500,000 cases per year, and kills more than 250,000 patients in the same period, on world basis [94].

However, cervical cancer can be prevented by means of the Human papillomavirus (HPV) infection vaccine, and regular low-cost screening programs (e.g., cytology, digital colposcopy) [111]. Furthermore, cervical cancer can often be cured by removing the affected tissues when identified in early stages [94, 111]. The development of cervical cancer is usually slow and preceded by changes in the cervix (dysplasia). Despite the presence of symptoms on its later stages (e.g., postcoital bleeding, bleeding between periods, increased vaginal discharge, and pelvic pain), the absence of early-stage symptoms might incur in carelessness prevention. Additionally, in developing countries, resources to perform screening programs with universal access are scarce and insufficient. Also, patients usually have poor adherence to routine screening due to low problem awareness.

While improving the resection of lesions in the first visits has a direct impact on patients that attend the screening programs, the most vulnerable populations have difficult access to such programs' information and medical centers. Consequently, the individual risk estimation has a crucial role in this context in order to optimize the effectiveness of these programs. Identifying patients with the highest risk of developing cervical cancer can improve the targeting efficacy of cervical cancer screening programs. Thus, recent attempts to address the predictive analysis of this problem have been proposed [95], including a competition sponsored by Genentech and Symphony Health Solutions [166].

During the cervical cancer examination, cervical cancer screening programs cover the following stages:

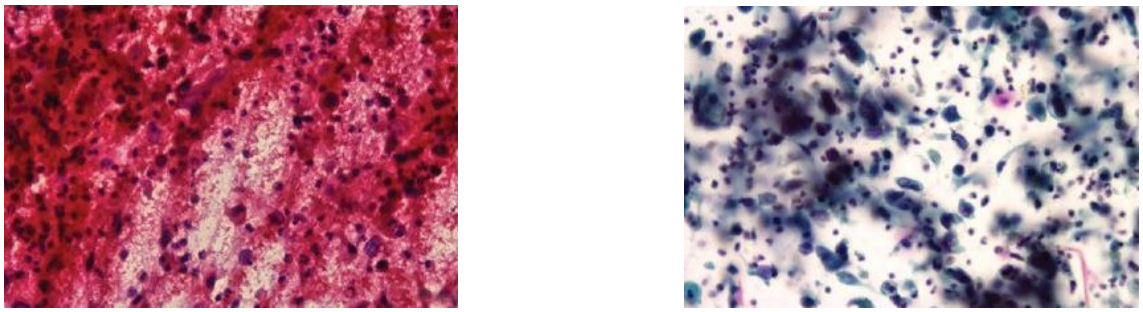


Figure 9.1: Samples of cytological screening [311]. **Left:** conventional cytology. **Right:** liquid-based cytology

- Cytology, either conventional or liquid (see Figure 9.1).
- Colposcopy, covering several modalities (see Figure 9.2).
- Biopsy.

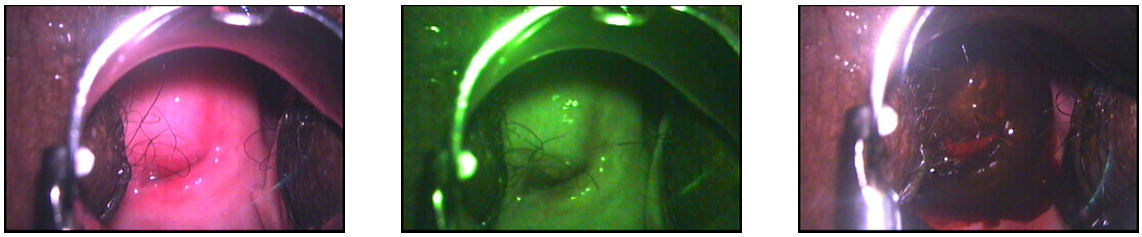


Figure 9.2: Modalities of the colposcopy examination. **From left to right:** Hinselmann, Green-filter, Schiller

These stages are often done in a cascade fashion, by moving towards the succeeding steps with the discovery of relevant indicators on the preliminary ones. Both cytology and colposcopy are image-based screening processes. The former focuses on the examination of vaginal and cervical cells under the microscope and the latter on the macroscopic examination with the naked eye (or with a magnifier lens).

The conventional cytological screening involves manual smearing and staining [29]. The complexity of the acquisition process for conventional cytology requires mobilizing expert teams to the field. Even when the acquisition is properly made, the uneven distribution of cells may induce dense regions where light cannot penetrate and empty regions of the slide [29]. Other artifacts such as blood may harm the effectiveness of this screening modality. To overcome these difficulties, liquid-based cytology (LBC) preparations have been delved. Liquid preparations help to standardize the distribution of cells and to dilute the presence of external factors. Some common techniques for the preparation of LBC can be found in [269, 322]. However, the cost increase (e.g., about 5 to 10 times higher [29]) and technical difficulties to make equipment available in remote locations appease the use of this technique in low-income countries.

On the other hand, digital colposcopy is a low-cost alternative to cytology. Nowadays, portable and mobile devices have been introduced in the market as an alternative to traditional colposcopes [187,188,237], facilitating its scalability and portability to locations with vulnerable populations. The main drawback of the digital colposcopy is the high sensitivity variability when carried out by experts with different levels of expertise. Plenty efforts have been devoted in the last two decades to automate the analysis of colposcopy images in order to support the medical decision process and to provide a data-driven channel for communication of findings. These efforts aim to objectify the analysis of this modality.

The automated analysis of digital colposcopies using ML and CV techniques has grown over the last years. Figure 9.3 shows the number of published papers per year reported by Google Scholar in this area. In addition to the aforementioned competition on the analysis of vulnerable population, Intel and MobileODT organized a constet for the automatic analysis of digital colposcopies in 2017 [167]. This increasing interest has resulted in a stable community with well identified problems that range, from the Quality Assessment (QA) [95] and enhancement [130, 204] of digital colposcopies, to the segmentation of the

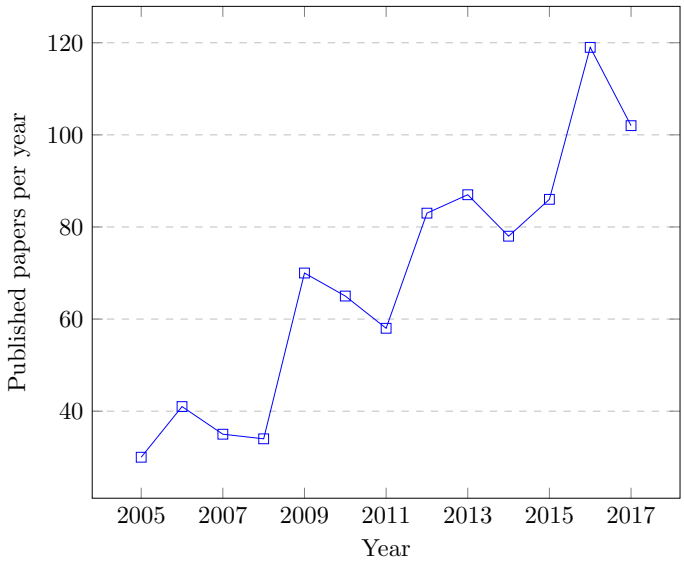


Figure 9.3: Number of papers reported by Google Scholar for the query `(‘computer vision‘ OR ‘image processing‘ OR ‘machine learning‘) AND (‘colposcopy‘ OR ‘cervigram‘)`, not including patents nor citations

anatomical parts of the cervix [65, 206], to the final diagnosis [289, 344, 345]. While the vast majority of databases that were used in the development of these papers are closed, as a result of these competitions, new public databases of considerable size and with new challenges were released [151, 167]. We are facing a possible turning point in the area, with the driving interest of governments and companies involved in the area, and the new advent of DL techniques that has been permeating all the areas of CV.

Therefore, it is relevant at this point to formalize the basis of the area, providing a comparative analysis of the main tasks involved in the area and the solutions that have been proposed in the last years. In this work, we aim to provide such foundations. Also, we release a database that will be continuously updated with transverse annotations. Finally, we enumerate the open problems and challenges in the area.

The rest of the chapter is organized as follows. Section 9.2 defines background medical concepts involved in the analysis of digital colposcopies. Section 9.3 describes the main tasks involved in colposcopic image processing and the solutions that have been proposed in the literature to tackle each one of them. Section 9.5 describes the available databases and challenges associated with each one of them. Section 9.6 describes the database and annotations provided in this work. Finally, section 9.7 concludes the work and discuss the main open challenges in the area.

9.2 Preliminary Concepts

9.2.1 Cervix Anatomy

The main regions of interest in the analysis of colposcopy images include: the external orifice (external os), the area of ectopy, Squamocolumnar Junction (SCJ), the transformation zone and the area of Squamous Epithelium (SE), also known as the exocervix. Figure 9.4 identifies the location of these regions.

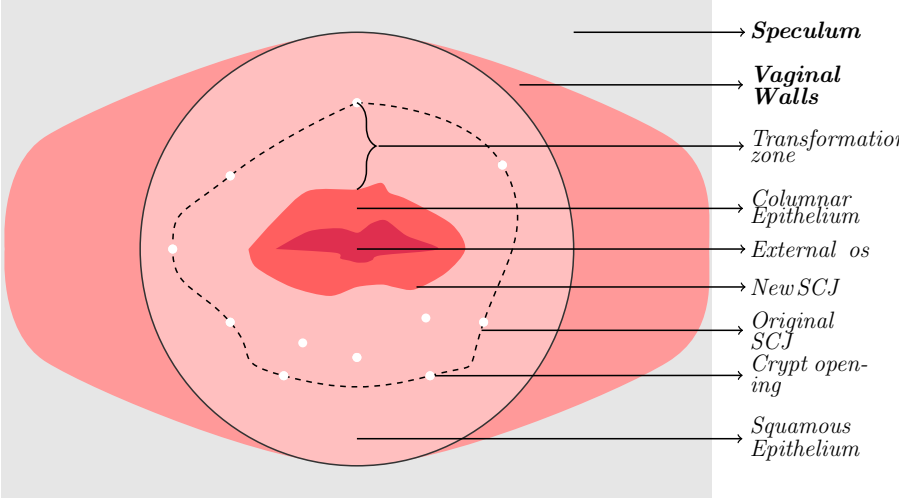


Figure 9.4: Relevant parts of the Cervix Anatomy and external objects (in bold).

Overall, the epithelium covers the superficial cells of the cervix. The low environmental aggression in the internal orifice of the cervix makes the cells in that region of columnar type. Thereby, it is relatively easy to observe the vascularity in this region. Conversely, the aggressive environment in the external region, caused by external factors such as the acid pH levels and trauma during intercourse, makes the external cells of squamous type. In some cases, the Columnar Epithelium (CE) extends outside the external orifice and gets exposed. Being exposed to external stimuli, CE turns into SE, originating the transformation zone (see Figure 9.4). The intersection of these two regions is the SCJ.

9.2.2 Colposcopy Examination

The observation of the cervix following the recommended protocol for digital colposcopies covers four main stages [296].

First, observation of the SE and CE with a magnifier lens is performed after application of a normal saline solution. During this step, the SE is observed to define landmarks of the transformation zone. The SE is typically smooth with a pink tone. The main landmarks of interest constitute crypt openings and nabothian follicles. These artifacts define the external boundary of the transformation zone. The inner border is determined by the SCJ.

The entire observation of the regions of interest is often unachievable from a single image since the SCJ may recede into the canal as the woman ages. Also, the CE is observed at this stage. The common appearance of the CE is dark red with complex patterns such as grape-like or sea-anemone tentacles-like or villous appearance [296].

To improve the visualization of the vasculature, a green filter is used on the colposcope to enhance the contrast of the vessels. The two most common vascular patterns observed in the SE are reticular and hairpin-shaped capillaries [296]. These patterns are typically found on specific regions of the cervix.

The third stage of the colposcopy examination consists in the observation of the cervix tissues after application of 5% acetic acid solution. This step is known as Hinselmann. In this step, SE and CE should be observed again. The change of appearance of these tissues after the application of acetic acid improves the discriminability of these regions by a human expert. Precancerous lesions can be observed in this phase.

Finally, the physician applies Lugol's iodine solution to the cervix, a step that is known as the Schiller's test. The normal vaginal and cervical SE stain and become mahogany brown or black [296], the immature squamous metaplastic epithelium does not stain or partially stain. Some abnormal patterns such as cervical polyps do not stain with iodine [296]. Thereby, the Schiller's test improves the discriminability of normal and abnormal regions in the transformation zone.

Cervical cancer is characterized by the abnormal growth of cells on the cervix. The wide spectrum of abnormal features associated with CIN may difficult the labor of a medical examiner. The high variance of appearance between women may difficult an objective assessment from unskilled examiners. Thereby, the characterization of these patterns and the identification of abnormal features in each part of the cervix anatomy have a direct impact on the expert decision.

9.3 Main Tasks

The applications surrounding the development of CAD systems for digital colposcopies cover a wide spectrum of tasks, from the analysis of the image quality, to the semantic segmentation of the image on its constituents parts, to the final diagnosis of the patients. Thus, CV and ML researchers have gathered around these tasks in the last decades.

The main source of data for this analysis are static color images directly captured from digital colposcopes. However, given that in some cases it is not possible to observe all the structures of the cervix in a single frame as well as its response to the acetic acid solution, some lines of work focused on the analysis of multiple views and even continuous videos.

In this section, we organize the literature into five main areas:

- QA and enhancement of digital colposcopies (section 9.3.1).
- Segmentation of cervix tissues (section 9.3.2).

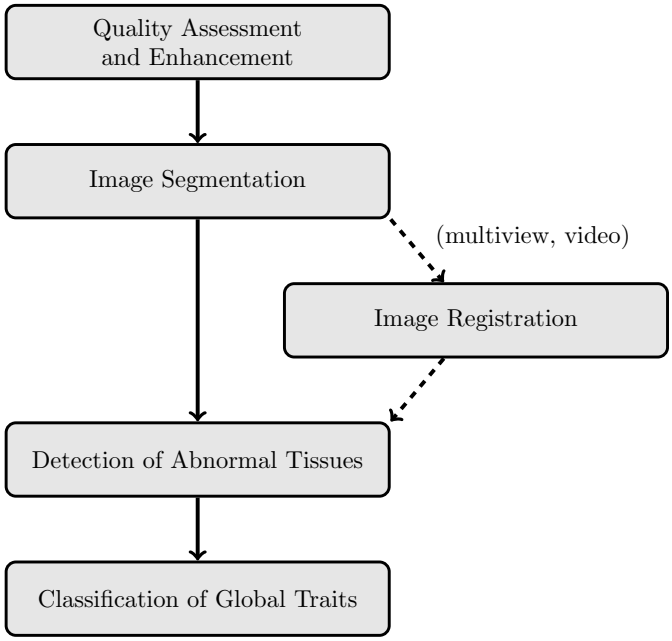


Figure 9.5: Pipeline of the main steps in the development of CAD systems for the automation of digital colposcopy analysis.

- Image Registration (section 9.3.3).
- Detection and characterization of abnormal tissues (section 9.3.4).
- Classification of patient traits (section 9.3.5).

These tasks are typically applied in a cascade fashion as illustrated in Figure 9.5. However, some methods may ignore parts of the pipeline or even include additional dependencies between them. For instance, methods focusing on static data would ignore the image registration step and some strategies to address the image quality require to segment the cervix tissues. Thus, this pipeline serves as a general overview of the main tasks but can be adapted to the intrinsic properties of each automation strategy.

In the rest of this section, we do a comparative analysis of the main methodologies that have been applied to each problem and we discuss their advantages and limitations.

9.3.1 Quality Assessment and Enhancement

The concept of quality has attained a significant interest in the CV research community. Traditional methodologies focus on a low-level notion of quality, measuring distortions of the image at a signal level [116,170]. In medical imaging, the notion of quality goes beyond low-level aspects of the images to semantic concepts such as visibility of the anatomical body parts, patient's pose, absence of external artifacts, among others.

Therefore, methodologies to address the assessment and enhancement of medical image quality are often application-specific and require extensive domain knowledge. In this section, we cover the main lines of research in this area for colposcopic image processing.

9.3.1.1 Quality Assessment

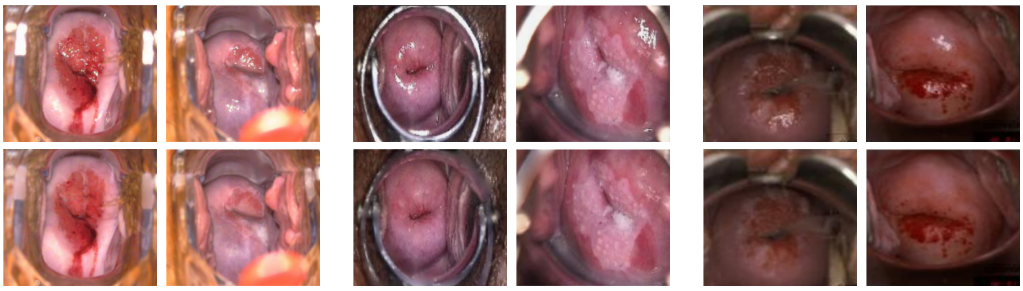
In the area of QA, Gu and Li [132] proposed a framework to validate the quality of uterine cervical imagery in an online scenario, so the physician may perform corrections to improve the acquisition of data in real time. In [132], the QA problem is modeled as a binary task where the program is required to decide if the image is good enough or not. Six types of problems related to colposcopic images were handled: zoom, position, foreign objects, contrast, blur and contamination. These traits were quantified using different models and, using a thresholding operator, it is decided if there are features with inadequate quality. The main disadvantage of this approach is the simplicity of the quality decision model (i.e., thresholding operators). Also, no quantitative assessment of the methodology is presented.

Fernandes et al. [95] proposed a learning methodology to tackle this problem. First, several features related to the image quality are extracted, including the area of the main parts of the cervix, the presence of specular reflections, observability of the entire cervix, and color statistics. Then a SVM is used to learn the quality decision model on a set of images, covering several modalities (e.g., Hinselmann, Green light, and Schiller) and inter-expert annotations. Fernandes et al. proposed a TL approach to improve the robustness of the learning process, where the knowledge acquired from the other modalities and experts is reused when a model for a new modality/physician is learned.

9.3.1.2 Quality Enhancement

Several works have been proposed in the area of quality enhancement of colposcopic images [64, 65, 130, 183, 193, 202, 204, 285], most of them focusing on the removal of specular reflections (SpR) [64, 65, 130, 183, 193, 202]. The remaining works, proposed by Li et al. [204] and Rouhbakhsh et al. [285] focused on the enhancement of images by means of color and contrast normalization. It is relevant to highlight that image enhancement can be done with two goals in mind, which may lead to different techniques and evaluation settings. First, this process may be done to boost the performance of automatic image processing algorithms. Second, image enhancement can be done for human visualization purposes. This can be done by several means, such as: highlighting relevant patterns of the image that are indistinguishable by the human eye, recovering damaged regions of the image, among others. All the aforementioned papers focused on the improvement for further automatic image analysis.

Removal of Specular Reflections Specular reflections or glares raise challenging problems in medical image analysis, as it degrades (partially or entirely) the information in the affected pixels [193]. Moreover, it can introduce artifacts in feature extraction algorithms [193] . The acquisition conditions and involved tissues in the colposcopic assessment are prone to generate this phenomenon.



(a) Lange et al. [193] (b) Das et al. [64, 65] (c) Gordon et al. [130]

Figure 9.6: Illustration of the results for SpR removal proposed in [65, 130, 193]. **Top:** original images. **Bottom:** corrected images.

Lange [193] proposed a method to remove this type of reflection using the green channel of the RGB color space, which classifies the types of glares that can be found in these images in two types: large saturated regions (detected with adaptive thresholds), and small high contrast regions (detected with morphological operators and thresholding). Once these regions are identified, missing information is filled by means of interpolation using Laplace's equation and modifying the intensity component of the HSI color space in the transformed image. The method is validated using qualitative subjective inspection. Similarly, Das et al. [64,65] proposed a similar approach to manage SpR. First, the affected regions are detected using the intersection of a thresholding operator on the three RGB channels independently. Then, Laplace's equation is used to select the smoothest possible interpolant.

Gordon et al. [130] proposed a different approach in both, detection and removal of SpR. In the detection subtask, fixed thresholds are used to detect high brightness and low color saturation areas. Then, pixels located in neighborhoods with high gradients are selected as SpR candidates. These pixels are mapped to the Saturation-Value space from the HSV color space and a mixture of two Gaussians is fitted. In the results, one of the Gaussians represent pixels with color information and the other contains merely white pixels. The pixels that belong to the second Gaussian are considered as damaged and are removed from the original image. To fill the damaged regions, a simple inpainting technique that propagates the color of the surrounding pixels is executed. This process is done under the assumption that the color underneath the SpR regions is almost constant and similar to the neighboring pixels.

Although none of the aforementioned papers show objective assessment of their methods, visual inspection suggests similar results in all of them. Figure 9.6 shows sample

images presented by each author. The methodology proposed by Das et al. shows some undercorrected areas, where residual specular reflections are observable. Also, despite the additional number of manually-defined parameters, the unsupervised learning stage proposed by Gordon et al. makes it more adaptable to new settings and datasets.

Image Normalization Li et al. [204] propose a color calibration system to map the color appearance of different colposcopes into one standard color space with normalized illumination. The process involves a preliminary calibration system where the physician presents a target color palette to the colposcope. The main disadvantage of this method is that it should be done before the acquisition of the images, which limits its applicability to already acquired datasets. Also, with the advent of mobile colposcopes, the acquisition conditions can vary quickly, requiring continuous calibration.

Other attempts, such as the one proposed by Rouhbakhsh et al. [285], perform simple normalization by means of brightness and contrast equalization.

The actual impact of this step in the final pipeline will depend on the type of assumptions made by the following steps of the automatic analysis. The types of invariance (e.g., pose, illumination, etc.) that can be ensured at this stage will facilitate the job of the following methods. However, as we introduce additional constraints, the applicability of automatic methodologies for the analysis of digital colposcopies will be confined, especially on remote settings with inexperienced staff. In counterpart, a new trend in DL to induce robust models is augmenting the data by introducing simulated perturbations (e.g., rotations, flips, contrast stretching, etc.).

9.3.2 Semantic Image Segmentation

Most efforts on the line of semantic image segmentation focused on the Hinselmann stage of the colposcopy protocol. Also, it is assumed that specular reflections have been removed from the image either during acquisition or as a preprocessing step.

The main trend in segmentation of the different regions of the cervix focus on the segmentation of the cervix from the outer parts (i.e., vaginal walls and speculum) and the segmentation of the acetowhite (AW) regions. Typical methodologies on this line belong to the class of unsupervised methods (e.g., clustering). The most common models are K-means [64, 131, 256, 262, 272, 324, 348], Gaussian Mixture Model (GMM) [130, 131, 202, 228, 272, 279, 315, 348, 349, 374], and Mean shift [202]. Regarding the feature space, most methodologies use raw color information on different color spaces, being the Lab color space the most widely used [64, 130, 131, 152, 183, 272, 279, 315, 348, 349, 374, 374], followed by RGB [228, 262], CIE Luv [202] and K-L color spaces [202]). Some additional features such as color ratios [262], texture information [130, 228, 374], and spatial information (i.e., distance to the image center) [272, 279, 315, 348, 349, 374] are used.

Clustering algorithms at a pixelwise level do not guarantee spacial consistency of the segmented regions, even when spatial features are considered. Thereby, post-processing

step was carried out in these works to decide the final segments that represent the areas of interest. Das et al. [64,65] and Traversi et al. [324] use the largest contour as the cervix representative, Gordon et al. [130] select the cluster with the lowest mean distance to the image center and highest mean redness level as the cervix region, using size to solve ties. Gu and Li [132] used morphological operators to fill small holes in the final segmentation.

Since the core cervix structures have smooth and almost indistinguishable contours, the performance of these methodologies is limited, not being able to differentiate the cervix from other structures such as the vaginal walls. For instance, Figure 9.7 shows the results of the method proposed by Das et al., where an oversegmentation of the cervix is done.

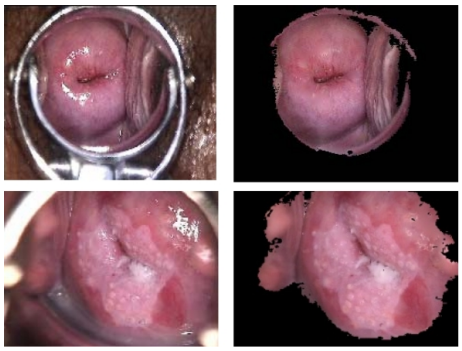


Figure 9.7: Das et al. [64] - input images (left), cervix segmentation (right).

In order to counteract oversegmentations of the vaginal walls, some authors applied domain knowledge on the expected shape of the cervix. For instance, some works used active contours [131, 223, 275, 374] solely or as a post-processing technique. Lotenberg et al. [223] include shape-priors (e.g., circles and ellipses) to encourage this behavior. Van Raad and Bradley [275] applied iterative multi-scale active contours by sequentially using the previous contour to reduce initialization impact.

For the segmentation of other regions, such as the CE, Gordon et al. [130] used a cascade of GMM, where the first level segments the cervix from the background using the redness level of the Lab color space, and the second level segments the CE from the rest of the cervix using texture and contrast features. Li et al. [202] used a cascade of GMM on the K-L color space and Mean shift on CIE-Luv to segment the cervix from the background and the external orifice from the cervix respectively.

A different unsupervised approach was proposed by Lange [192] based on the watershed algorithm. First, cervix and vagina are segmented using a hue color classifier. Then, the watershed algorithm is applied to detect the low-intensity border around the cervix. Finally, they extract a feature related to the AW response consisting in the product of the green channel in the RGB color space and the saturation value in the HSI color space. The watershed algorithm is applied iteratively on the gradient of the AW feature to segment the cervix into a disjoint subset of coherent regions in terms of AW response. This step addresses the separation of CE and SE, such that the CE is identified as the

resulting regions from the AW watershed segmentation that have lower feature values than the surrounding regions (i.e., valleys). The same idea is applied over the gradient of the red channel to detect the external orifice by identifying the valleys. Figure 9.8 shows segmentations obtained by the methodology proposed by Lange [192]. While the core regions of the cervix are properly identified, some artifacts are observed such as the recognition of external objects as AW regions and the disconnected appearance of the external orifice (external os).

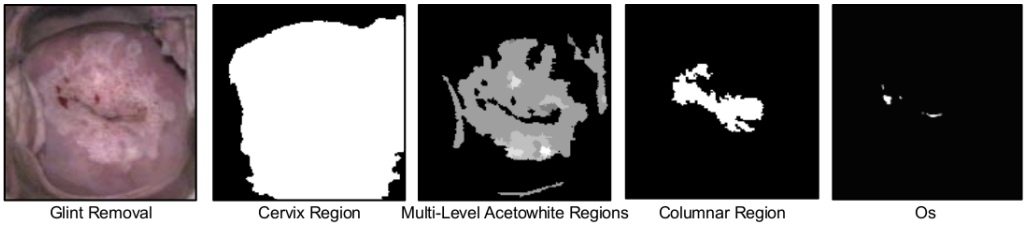


Figure 9.8: Lange and Ferris [192, 195] - cervix segmentation.

Some authors refine the segmentation task by detecting the external orifice [131, 206, 349, 374]. This process is done by gradient analysis in order to find the largest concave region in the image.

In general, these works do not present any objective assessment of the attained performance in terms of segmentation quality, providing in some cases a subjective notion of expert satisfaction [64, 65, 130, 132, 192, 195, 202].

The main drawback of these unsupervised strategies is the low semantic level at the decision process, working at a pixel or neighborhood level. Thereby, spatial coherence is unattained in most cases. Moreover, the lack of contours difficults the separability of the main regions without a global image representation. A common assumption of these techniques is that the cervix covers a significant portion of the image and that external objects (e.g., colposcope, gloves, swabs, etc.) are not present. Thereby, their robustness to unconstrained settings is limited.

In order to overcome the limitations of the unsupervised segmentation algorithms, several supervised methodologies have been proposed using traditional segmentation-by-classification pipelines consisting in feature extraction and modeling with SVM [158, 206, 349]. These techniques rely on color [158, 349] and texture information [349]. Huang et al. performed the recognition on superpixels resulting of a preliminary unsupervised clustering step [158]. Then, they use a one-vs-one SVM to classify the regions as AW, CE and SE. While they present results for the pixelwise classification accuracy of cervix and non-cervix tissues, they do not show any quantitative results of the final multiclass segmentation.

Recent advances on semantic segmentation of digital colposcopies using DL techniques can be found in [90, 91, 93]. The work of Fernandes and Cardoso [90] tackles the joint segmentation of several objects (i.e., colposcope, vaginal walls, cervix, transformation

zone, and external orifice) in digital colposcopies. The proposed methodology extends the U-net deep architecture to improve the spatial ordinal consistency between objects. Namely, they induce segmentations where the objects of interest appear nested one inside the other. They validated the performance of their model on two databases covering all the colposcopy modalities and achieved a macro-average Dice's coefficient of 51.24% and 66.98% on the databases [95] and [167] respectively. Besides the capability of segmenting the entire set of objects globally, using DNN enables more semantic segmentations, where the segmentation of cervix tissues without edges is achieved by considering feature spaces with a high level of abstraction.

9.3.3 Image Registration

According to Shapiro and Stockman [302], image registration defines the process whereby points of two images from similar viewpoints of essentially the same scene are geometrically transformed in such a way that corresponding points of the two images have the same coordinates after transformation. The definition of Shapiro and Stockman might be relaxed when considering multimodal image registration by accepting a broad definition of *similar viewpoints of essentially the same scene*. Medical image registration is a challenging process, the intrinsic properties of each modality may distort the visual aspect of the objects in the image. We can think about medical image registration even in extreme cases where the images to align represent the external (e.g., RGB or depth image of the body) and internal structures (e.g., X-rays, ultrasound, etc.). The registration of body parts is complex, given the elastic deformations that occur in the body. For instance, the cervix is distorted in a non-rigid manner due to the patient breathing, muscular movements, etc. Even more, the modalities involved in the colposcopy may reveal and hide structures. For instance, the green light enhances vascularities, Hinselmann shows AW regions and Schiller's test strongly dichotomizes the cervix into normal and abnormal regions.

In the literature, there are several works that have targeted the registration of colposcopies [2, 5, 6, 15, 118–121, 139, 140, 194, 201, 206, 233, 234, 261]. Three main lines of work have been proposed: global (either rigid or elastic), landmark-based and segmentation-based registration.

Since most of these works dealt with images from the same phase (typically Hinselmann), they were able to use standard (normalized) cross-correlation techniques [2, 5, 139, 140], commonly used when images belong to the same modality. In order to overcome the natural variations of the cervix, some works refine the rigid registration using local elastic registration techniques [119, 121, 194, 201].

Acosta-Mesa and his collaborators have a line of work in this area [2–6, 139, 140], either as the core focus of their work or as a preliminary step for the final classification of patients. Thus, Acosta et al. [6] proposed a two-stage method to deal with local deformations. First, a phase correlation is applied to remove global translation difference between images. This method has some advantages when dealing with different contrast and brightness and

with some simple intra-modal changes (i.e., AW response), as can be observed in the AW response [6]. Then, local deformations are removed by means of locally normalized cross-correlation. To accelerate the registration process, they proposed a method to register cervical images in grayscale [234], which performs a search of small local regions of the image in consecutive frames. The main challenge of colposcopy registration is the lack of distinctive landmarks in almost the entire cervix anatomy. In this sense, Acosta-Mesa et al. [2] proposed to use a manual stain landmark (at acquisition time) using Lugol solution and to use this landmark to simplify the registration process. While it is true that using such landmarks reduces the complexity of automatic methods for image registration, it adds complexity to the physician labor and could occlude relevant regions of the image with abnormal tissues.

Garcia-Arteaga et al. proposed several methods for colposcopic image registration [119–121]. In [121], an elastic registration algorithm was proposed, representing the problem as an optimization over a set of continuous deformation vector fields. Regularization was modeled by describing equilibrium in an elastic material using a linearized 2D elasticity operator (also used by Li et al. in [201]). The registration method is done in a multiscale fashion in order to speed up the process. No objective results are provided, but a simple visual inspection. Similarly to the two-stage approach used by Acosta-Mesa et al., Garcia-Arteaga and collaborators [119] applied rigid registration with cross-correlation followed by elastic registration.

Given the challenges involved in global registration techniques, several attempts to address the problem as a landmark detection have been proposed [15, 110, 118, 206, 233]. These techniques take advantage of interest points such as Harris corner detector [15, 110, 233] that can be used to register images over time. Then, local descriptors such as SIFT [233], cross-correlation and distance [15] are used to identify matches. In posterior work, Garcia-Arteaga et al. [118] introduce geometric information about feasible deformations in order to remove false positives.

An alternative line of work use pre-segmented regions of the cervix to conduct registration [206, 261]. This kind of segmentation produces very coarse results, especially when the reference objects are of limited size such as the external orifice [261].

9.3.4 Abnormal Tissue Detection and Characterization

In the area of abnormal tissue detection and characterization, several methods have been proposed, some of them included hyper-spectral imaging [74, 103, 104, 136, 137, 329]. Since the current challenge in digital colposcopy is the scalability to remote health-care centers with low resources, we will discuss methods that are able to work with traditional digital colposcopy that can be ported to current mobile devices. Namely, we focus on image processing techniques that handle RGB color images (and video).

In this section, we discuss works that tackled the localized recognition of these abnormalities. This could be considered as a midpoint between the previous section that

tackled the pixelwise segmentation of the anatomic part and the next section that will cover the detection of relevant traits at a patient level (i.e., medical records, demographic data, etc.). In this sense, the following strategies address the problem of identifying and characterizing abnormal tissues at specific regions of the cervix. The main assumption of the works in this area is that the image constitutes the cervix regions (either by detection and cropping or by constrained acquisition) and that relevant anatomic parts have been segmented in a previous stage. Also, most works assume specular reflections (see 9.3.1.2) have been removed. This last assumption is especially relevant since these artifacts could be easily recognized as positive AW lesions. We can study these works from three different perspectives: lesion of interest, learning paradigm and type of data. Table 9.1 presents the main alternatives in these lines.

Table 9.1: Summary of the main categories of work on the detection of abnormal tissues.

Topic	**Alternatives**
Lesions	• AW lesions. • Vessels and mosaicism.
Learning paradigm	• Unsupervised. • Supervised.
Data	• Image-based. • Sequence of images (temporal).

Two main types of abnormal traits have been addressed in the literature: AW lesions and abnormal vascularities/mosaicism.

For the detection of vascularities and mosaicism, most works relied on simple image processing techniques on static images. The main lines of research involve morphological operators and template matching [67,161,203,315–317,327], being the former of unsupervised nature and the latter of supervised nature with lazy learning (i.e., neighbor-based). Thereby, these techniques are highly sensitive to changes to the image resolution, scaling and illumination. This area is almost unexplored and has space for more robust techniques, able to cope with complex vessel patterns and with unconstrained settings.

As for segmentation, several works in the detection of AW lesions applied unsupervised techniques, ranging from K-means [156,204], GMM [66,68,129,130,202,315,318,328,328], Mean Shift [201] and Watershed analysis [127,128,193,202,338] to adaptive thresholding [51,318]. The goal of watershed analysis techniques was mainly to over-segment the cervix

according to the AW response of the features [128]. Also, some works used deterministic annealing [327] and active contours [80] to detect lesions. The most widely used features for static images include color [127, 129, 130, 156, 193, 201, 202, 204, 273, 315, 316] texture [129, 204], edges [240] and spatial information [156, 315], also used by other supervised methods that will be explained below. The main limitation of unsupervised strategies is their low discriminative power to differentiate abnormal AW regions from SE since they have similar colors [130]. Other problems such as high number of false negatives on regions with shadows and false positives on the vaginal walls are also typical [130]. In Figure 9.9, Gordon et al. illustrate these problems in the resulting images. This effect is present in general for methods that make predictions using local information (i.e., pixelwise data) without considering a global representation of the image.

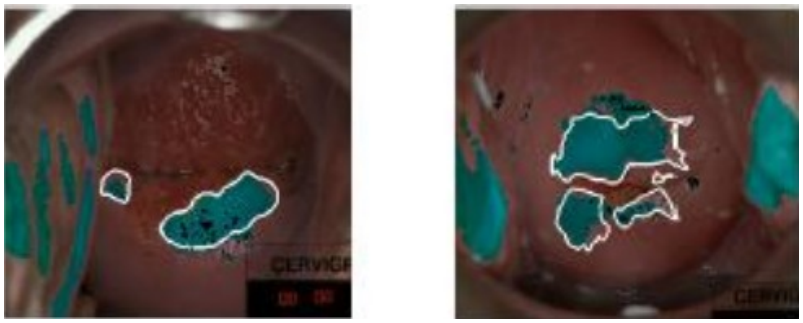


Figure 9.9: Gordon et al. [130] - AW detected regions (green), manual annotations contours(white)

Then, several methods addressed the problem from a supervised learning perspective. In general, this was done by extracting features from individual pixels, overlapping and non-overlapping tiles or by super-pixels obtained by segmentation techniques and applying a learning mechanism on the corresponding space. For classification, the most used method was K-Nearest Neighbors (KNN) [2, 3, 5, 177, 226, 231, 271, 272, 278, 280, 285, 343], followed by SVM [14,177,206,271,343], naive Bayes [2,6,271,279] and Multilayer Perceptron (MLP) [285, 310, 343]. Other authors used Adaboost [342, 343], Conditional Random Field (CRF) [228, 262], among others [261, 285, 291, 343, 358]. The most common features for static images were color histograms at different scales [177, 206, 231, 271, 278, 343], oriented color gradients [177, 285, 342, 343], other color-based features [14, 231, 261, 262, 279, 285, 342, 358, 366], edges and texture [177, 206, 342, 343, 350], discrete wavelet transform [226, 272, 280, 350], and the amount of punctuation and vessels [228, 260, 262]. For sequence-based recognition, the features involve changes on the temporal AW response either in a two image sequence (i.e., before and after application of acetic acid) [119, 119, 120, 201, 210, 260, 261] or at a fine-grained resolution level [2–6, 139, 291].

While these efforts addressed the problem from a pixelwise perspective, Alush et al. [9, 10] modeled the problem from a boundary-based approach, by classifying edges of superpixels as a lesion or not. In this sense, a more global concept of the image is built.

Superpixels are built using the watershed algorithm. The classification is performed by learning a dictionary of visual words and the problem is modeled using Markov Random Field (MRF), where each superpixel corresponds to a binary random variable indicating whether the region is part of the lesion. The final detection is done using belief propagation. Another dictionary-based algorithm was proposed by Zhang et al. [364], who used the K-Singular Value Decomposition (SVD) method to create positive and negative dictionaries of sparse representations. Finally, reconstructive errors of the sparse coefficients from the test images are calculated and compared for classification purposes.

A recent trend in the lesion detection combines different modalities for improving the final performance. In this sense, we have the work of Xu et al. [341] that combines text and image features in a late fusion and the work of Song et al. [313] that combines results from several modalities (e.g., cytology, HPV, colposcopy) and demographics (e.g., age) to train their model. Results on this direction seem promising.

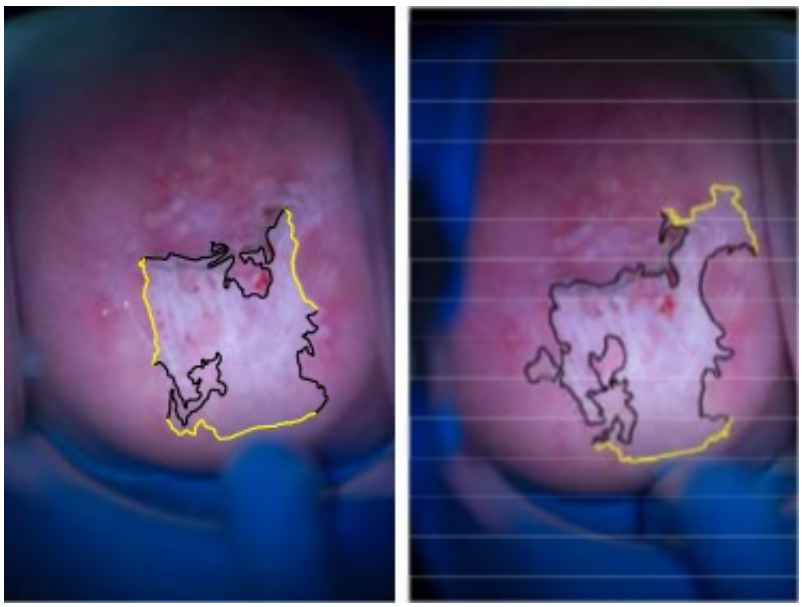


Figure 9.10: Van Raad et al. [330] - yellow segments are characterized as smooth contours and black segments as irregular.

Van Raad and her collaborators [330] proposed an automatic characterization of the lesions borders. After segmenting AW regions using GMM, contours are characterized by detecting smoothness and irregularities. This type of characterization is relevant for medical teams as a way to introduce explanatory predictions for the decision support. Figure 9.10 shows the detected AW regions and the segment characterized as smooth (green) and irregular (black).

For multi-image – temporal – approaches, where a sequence of images is presented to the model, the main lines of work were presented by the teams of Li [201], Park [261], Liu [210], Acosta-Mesa [2, 5, 6], and Garcia-Arteaga [119, 120]. In these cases, the change in color before and after the application of acetic acid is used. The works conducted by Li

et al. [201], Liu et al. [210], and Park et al. [261] focused on pairs of images (i.e., pre-acetic and post-acetic). The first two approaches used Mean Shift clustering and level sets respectively. The last approach, proposed by Park et al. [261] validated the performance of ensembles of supervised classification algorithms.

Acosta-Mesa and his team worked at a more fine-grained level [2, 5, 6] by extracting information from continuous frames at a pixelwise level to measure the AW response. Their preliminary approach modeled the response using a parabola, which parameters are then used as features to classify the tissues using the naïve Bayes classifier. In successive works [2, 5], they explored discretization schemes to encode time series information, being able to surpass the human-level performance by 3% in terms of accuracy at a dataset with about 50 patients (76% and 73% respectively). This study was replicated for the assessment under green light in [139, 140]. In a more recent work [227], they studied the performance of several classifiers on temporal data, achieving the best results with ANN (89% of accuracy). Active contours were used as a post-processing step to identify good candidates for biopsy in [227, 278].

Finally, Garcia-Arteaga et al. [119,120] also considered time series analysis on the AW response of the pixels. They focused on differentiating abnormal from normal tissues as a first task, achieving considerable performance (79.3% accuracy and 85% ROC AUC). Also, they present results for the classification of low-grade and high-grade lesion classification, achieving an accuracy of 92% and ROC AUC of 87%. While these results are satisfactory, the datasets are very limited in terms of the number of patients (3 and 10).

As a side application, Fernandes et al. [91] tackled the detection and characterization of lesions on the vagina using DNN. While the images of study are from digital colposcopies, the application of interest is the forensic evaluation of sexual assault.

9.3.5 Classification of Global Traits in Colposcopies

Typical CAD systems involve the detection of global traits observed at images, from low-level tasks such as the modality recognition [94] to more semantic tasks like the identification of the cervix type [168, 175] and cancer detection [289, 344, 345].

Fernandes et al. [94] proposed a framework to recognize the acquisition modality of each frame in a video sequence. They propose a supervised learning scheme using color information and KNN. Global consistency between the predicted modalities and the colposcopy protocol is enforced using weighted finite automata. Also, a preprocessing step to filter noisy frames where the physician manipulates the cervix region is proposed.

Several works have addressed the problem of classifying a cervigram as cancer or non-cancer, being the line of research proposed by Huang and her team the most prominent [177, 313, 341–343, 345]. The standard scale to grade CIN consists on three ordinal grades [345]: CIN1 (mild), CIN2 (moderate), and CIN3(severe). However, most works address the predictive task as a binary classification one by considering the classification of CIN1 from CIN2/3 or cancer (CIN2/3+) [345]. After some preprocessing steps that cover removal of

specular reflections and identification of the ROI containing the cervix, these works extracted image features in a pyramidal fashion, including color histograms (typically on the Lab color space) [342,343,345], histogram of gradients [342,343,345], and LBP [343,345]. Several classifiers were used, including tree ensembles (RF, Gradient Boosting, AdaBoost), ANN, LR, SVM and KNN. RF achieved a top performance of 84% ROC AUC in a dataset collected by the National Cancer Institute (NCI) in the Guanacaste project [151] with +1000 patients. In a more recent work, Xu et al. [345] compared the performance of deep features and the aforementioned pipeline based on traditional methodologies. In this sense, they extracted the features from the last dense layers from CaffeNet [162] trained on ImageNet and fine-tuned the last layer. While they achieved higher performance by using handcrafted features, further gains may be observed by training the network end-to-end instead of the final layer.

Xu et al. [344] proposed a multimodal approach to predict cervical cancer by merging deep features from AlexNet [180] and high-level information from medical records (e.g., age, HPV status, etc.). They were able to improve the performance obtained with image data from 88.77% ROC AUC to 94%.

Sato et al. [289] used CNN trained from scratch to predict cervical cancer on colposcopies with Hinselmann and Green filter modalities. As in the works mentioned above [344,345], the architecture is considerably shallow, with 3 groups of convolutional-pooling layers and a couple of densely connected layers. They trained the architecture on a dataset with 485 images achieving 50% accuracy in recognizing 3 balanced classes. Further investigation of state-of-the-art architectures and regularization techniques (e.g., TL, data augmentation) should be conducted in order to assess the actual capabilities of deep methodologies in this area.

In a recent competition about the categorization of cervix based on their transformation zones, DL methodologies achieved the best performance. The task was to characterize the cervix into three types depending on the transformation zone tissues type and observability [167]. The database consists of more than 1800 images from several modalities (Hinselmann, Green, and Schiller). The acquisition setting was unconstrained, having images with bad quality and images where the cervix was considerably small. Main pipelines to solve this problem involve the segmentation of the cervix and its transformation zone using the U-net architecture [175] and an ensemble of deep architectures to classify the images [168].

9.4 Summary

The research ecosystem on ML and CV techniques for the decision support of digital colposcopies has reached a sound point, with well-identified problems and paradigms to tackle them. Here, we will summarize the main traits of the aforementioned tasks and solutions. Figure 9.11 gathers the main features and works in each of these areas.

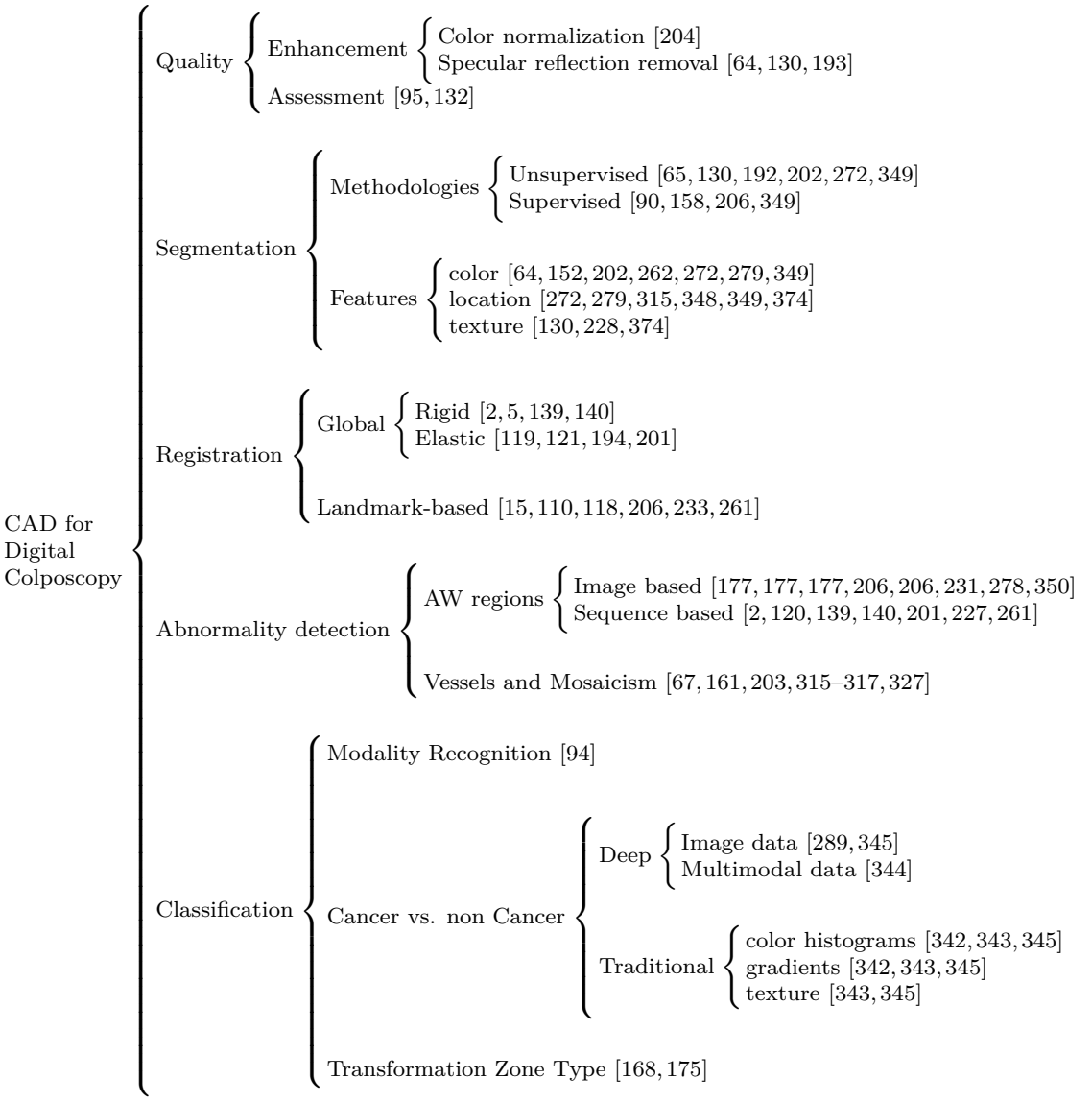


Figure 9.11: Summary of the main research topics and selected works in the area

In the area of quality enhancement, the removal of specular reflections and the standardization of the color space are the main tasks of interest. The former has been tackled by a detection-inpainting scheme while the second one has been solved using camera calibration and simple image processing techniques. For QA, the main features of interest are the entire observability of the cervix and the absence of disturbing artifacts such as specular reflections, bleeding, and external objects.

The semantic segmentation of the cervix tissues has been one of the areas that perceived more attention from the research community. The vast majority of works addressed the problem using unsupervised clustering techniques. However, the hard assumptions on these works and the smooth boundaries between the cervix tissues demand more expressive models with high semantic power. Therefore, some supervised methodologies have

appeared, including traditional ML and DL pipelines.

The registration of colposcopies was studied for unimodal settings. On the one hand, global image registration techniques that attempt to find a good global alignment of the images have been proposed. These techniques typically involve a rigid alignment of the main structures of the cervix, followed by an elastic registration to address eventual deformations of the body parts. On the other hand, landmark-based registration aims to detect and track points of interest.

For the spatial location and characterization of lesions, basic image processing techniques were used to detect vessels and mosaicism, including morphology operators and template matching. The recognition of AW lesions received more attention, with methods covering both unsupervised and supervised techniques, and static and continuous acquisitions.

The final step of any CAD system is the diagnosis support. Therefore, providing a global decision per patient has been widely studied using ML techniques. Traditional methodologies include color and texture information while novel techniques attempt to learn relevant features using deep methodologies. Some works have addressed the aggregation of multimodal data (e.g., medical records) achieving the best results in the literature.

9.5 Databases

As important as the methods delved to solve the aforementioned tasks are the databases used to validate their findings. Thus, the actual impact of any data-driven system relies on the similarity between the test database and the organic data acquired on a daily basis on medical facilities. Also, the diversity of acquisition settings and abnormalities is relevant. In this sense, we summarize the main aspects of the available datasets in the area (see Table 9.2).

Table 9.2: Summary of the datasets available databases

Author	**Multimodal**	**Multiview**	**Size**		**Annotations**	
			Images	**# Patients**	**Spatial**	**Global**
Acosta-Mesa et al. [3]	No	Yes	10 videos	10	Yes*	No
Fernandes et al. [95]	Yes	No	287	100	Yes	Yes
Guanacaste Project (NCI/NIH)	Yes	Yes	+2k	387	No	Yes
Intel and MobileODT [167]	Yes	No	2k	-	Yes*	Yes

* Provided by us as part of this project.

9.5.1 Acosta-Mesa et al.

Acosta-Mesa and his team made available ten videos with digital colposcopies of 10 patients after application of acetic acid [3]. The database does not contain manual annotations and the acquisition was very controlled. The duration of the sequences is 30 seconds

(311 frames). The images have high quality and allow to study small patterns with high temporal and spatial resolution. This dataset can be used to validate (elastic) registration techniques and detection of AW regions.

We made available as part of this project, manual annotations of 10 landmarks per video to validate the performance of image registration techniques [1].

9.5.2 Fernandes et al.

The dataset was acquired by Fernandes et al. [95] in collaboration with *Hospital Universitario de Caracas* from Venezuela. The number of images is 287, including three modalities (Hinselmann, Green light, and Schiller). Several features were extracted from the dataset for the QA task. The original subjective quality annotations were performed by six experts. The dataset also contains manual segmentation masks of the colposcope, vaginal walls, cervix, external orifice, and artifacts. It can be used to validate the performance of QA methodologies and semantic image segmentation algorithms. The dataset can be accessed in the UCI Machine Learning repository [2].

9.5.3 Guanacaste Project (NCI/NIH)

The dataset is made available by the NCI/National Institute of Health (NIH). The dataset was collected by the NCI in the Guanacaste project [151]. It contains 2120 images from 387. Each patient has 1-2 images per visit, with a maximum of 20 images over the images for a single patient. Besides the image data, medical records are made available, including age of the patient, HPV test, histology results. The patients are annotated with the corresponding CIN progression level (i.e., normal, CIN1, CIN2, CIN3, and cancer). The presence of multiple images per patient in combination with other sources of data encourages the development of multi-view and multi-modal algorithms.

The dataset only contains global information about the patient. Therefore, the dataset can be directly used to evaluate automatic methods for the detection of cervical intraepithelial neoplasia and cancer. It is also sensible to be used for assessment of semantic segmentation and registration techniques but it would require further annotations.

9.5.4 Intel & MobileODT

Intel and MobileODT made available a database with about 2000 static images, covering the main modalities of the digital colposcopy. The dataset was released under the scope of a competition to identify the type of cervix among three types according to the location of the transformation zone [167]:

1. **Type I:** completely ectocervical, fully visible, small or large.

[1] `https://github.com/kelwinfc/cervical-cancer-screening`
[2] `https://archive.ics.uci.edu/ml/datasets/Quality+Assessment+of+Digital+Colposcopies`

2. **Type II:** has endocervical component, fully visible, may have ectocervical component which may be small or large.

3. **Type III:** has endocervical component, is not fully visible, may have ectocervical component which may be small or large.

The dataset contains 1481 training images with annotations about the cervix type. The images distribution is unbalanced with 17%, 53%, and 30% respectively. All the cervix images in this dataset are considered normal (not cancerous) but the identification of the cervix type may require further testing [167]. The dataset has a large number of images that have not been curated but that can be used for the development of semi-supervised approaches.

As part of this project, we provide manual segmentation masks for this database, including the cervix region, transformation zone and external orifice.

9.6 DCDB: Digital Colposcopy Database

As was discussed in the previous section, several datasets have been acquired and made available by the research community. However, given the lack of a dataset that can be used on the assessment of all the aforementioned tasks, we collected a database with 129 digital colposcopies in video format. The videos were acquired between 2013 and 2015 at *Hospital Universitario de Caracas* in Caracas, Venezuela. The dataset covers the entire examination, including the main modalities of the colposcopy examination and intervals where the physician manipulates the cervix region. Thus, the dataset raises several challenges, from the multimodal and time-based integration of the decision to the identification of the proper frames to apply the models. In this sense, this dataset is close to a real-life scenario for the assessment of automated techniques. Figure 9.12 shows sample images from the database. Figure 9.13 summarizes some statistics about the videos.

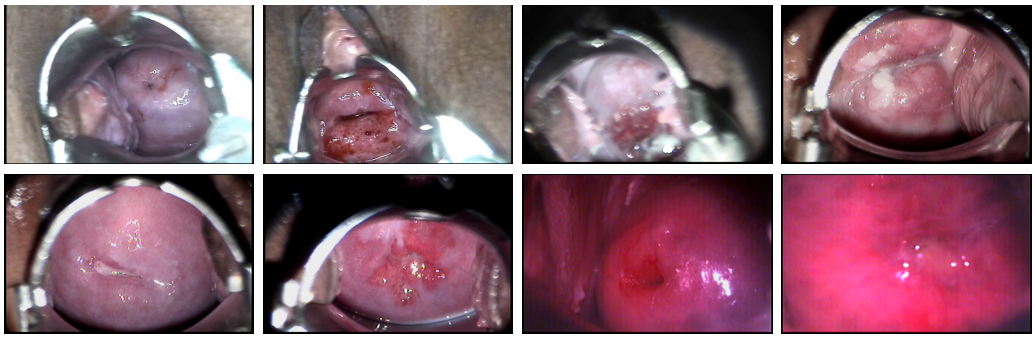
Also, we make available the following annotations:

- **Modality Detection:** temporal annotations for the videos of the start and ending points of the modalities per frame. Also, we include annotations of the transition and noisy frames.

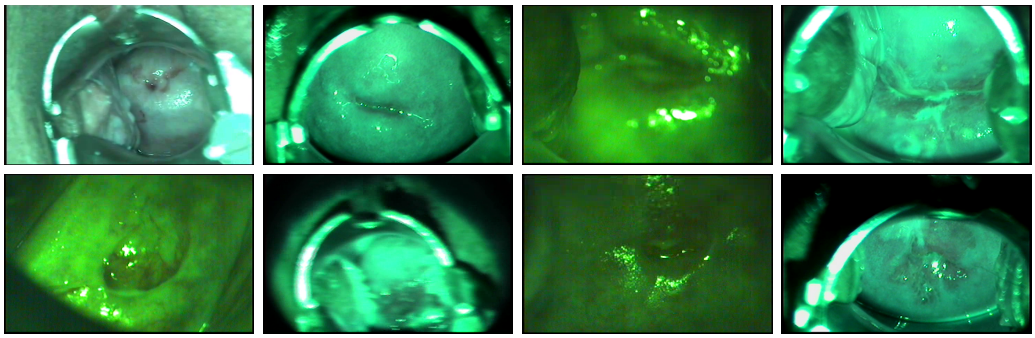
- **Quality Assessment:** annotations from 6 experts in an ordinal scale (i.e., poor, fair, good, excellent) for 287 images.

- **Semantic Segmentation:** annotations for 287 images of the colposcope, vaginal walls, cervix, external orifice, and artifacts.

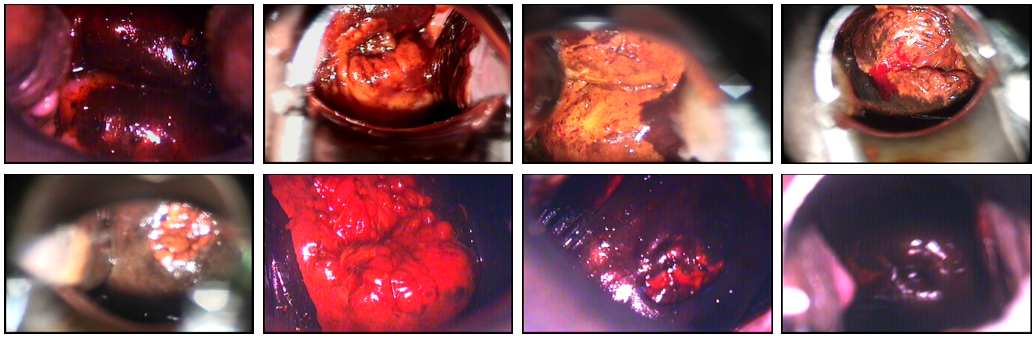
- **Image Registration:** landmark annotations for image registration, including Y points per video annotated every ten frames.



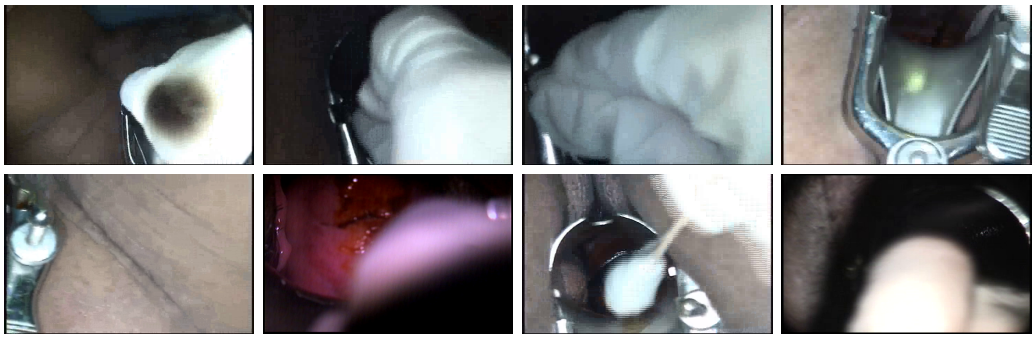
(a) Hinselmann



(b) Green filter



(c) Schiller



(d) Noisy frames where the physician manipulates the colposcope and the cervix region

Figure 9.12: Sample images from the DCDB database.

Minimum duration (seconds)	7
Maximum duration (seconds)	658
Average duration (seconds)	98
Median duration (seconds)	74

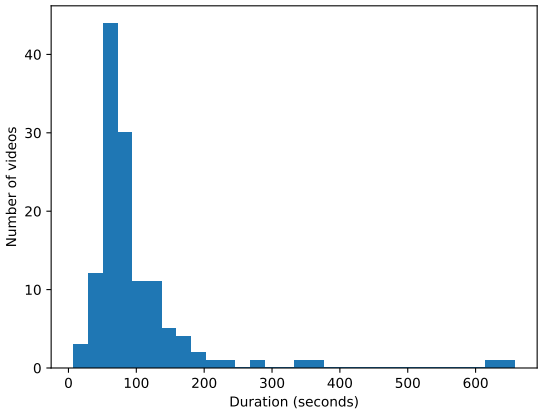


Figure 9.13: Summary of the database statistics and distribution of the video durations.

- **Abnormalities:** annotations about the lesions and abnormalities present in the image.

The videos and annotations will be continuously updated and improved. The database can be accessed online in [3]. Further details about the dataset, training/test partitions can be found on the project website.

9.7 Conclusions and Challenges

The automated analysis of digital colposcopies has attained significant attention from the ML and CV research communities. We studied in this chapter the main lines of research that have been conducted in this field and built a topology of tasks and approaches that encompass the area. While the field reached a certain level of maturity, the recent investment of companies and governments in the area and the recent publication of large databases open the possibility to include more advanced techniques in the development of CAD systems for digital colposcopies.

The main contributions of this work can be summarized as follows:

- We performed a review of the literature in the area.
- We established a common ground for the analysis of CAD systems for digital colposcopies.
- We released a video database with partial annotations that covers the main areas that were identified.

[3] `https://github.com/kelwinfc/cervical-cancer-screening`

- We provide annotations for databases from third parties.
- We released source code and benchmarks for comparison on these databases.

Finally, despite the huge efforts that have been devoted to this area, several open challenges were identified. Below, we enumerate the main open problems in the area.

Quality Assessment and Enhancement:

The notion of quality is a very subjective concept. Therefore, using a binary scale (i.e., bad, good) to define the quality of a digital colposcopy is too reductionist. Thus, a fully automated system should be able to identify, for each expert, the expected image quality in order to 1) suggest improvements during acquisition in real-time, and 2) retrieve the best frame to the human expert to maximize the confidence of his decision. While there is space on the normalization and enhancement of images without constrained acquisition settings, the appearance of DL techniques that are robust to such variability may reduce the impact of these techniques.

Segmentation of Cervix Tissues:

In the area of semantic segmentation, the main limitation of the current strategies is the lack of adaptability to unconstrained settings. Due to the low semantic level of the techniques proposed in the literature, they are not able to segment objects with smooth transitions such as the cervix and the vaginal walls or the SCJ. Thus, most of the published works focused on the segmentation of three entities: background, cervix and the external orifice. Moreover, current techniques are not able to cope with several modalities. Also, it is relevant to explore methods to promote spatial consistency among the detected objects.

The development of DL architectures for semantic segmentation may be able to circumvent these problems, being able to represent global semantic properties of the image.

Image Registration

The main open challenge on the registration of digital colposcopies lies in the elastic registration of several modalities. Given the different signal statistics and disjoint observability of certain structures on the modalities, traditional registration techniques would not be able to cope with multimodal registration. Using segmented regions identified on each modality may drive a coarse alignment of the main cervix structures. However, a deformable alignment of the inner structures of the cervix would require additional complexity.

Lesion Detection and Characterization

Besides the multimodal and temporal analysis that is intrinsic to all tasks, learning to detect and characterize lesions with cost-effective ways of annotations is a relevant problem. Traditional methodologies require a large amount of manual labeling, including

spatial localization of the lesions at an image level. Learning to detect lesions from weakly supervised annotations, where the expert identifies the presence of lesions in the video without explicitly identifying their boundaries, would directly impact the scalability of these frameworks.

Classification of Global Traits

Being the final stage of any CAD system, the amount of open problems in this area is prominent. We should look towards holistic frameworks able to extract knowledge from each modality (image and non-image data). Also, the inclusion of information from multiple visits from the same patient over the years should be addressed in order to identify long-term changes in the cervix.

While current strategies have simplified the prediction task to binary settings, developing predictive systems that are able to identify the progression of the lesions following the CIN ordinal scale would accelerate the acceptance of these systems.

Current systems work on a disjoint fashion by applying the aforementioned tasks in a cascade fashion. In this sense, the knowledge acquired from one task such as segmentation is not used when learning to solve another task such as QA or cancer prediction. Thus, it is relevant to study TL and multitask learning approaches in order to induce more robust and holistic decisions.

The final challenge –which is ubiquitous in all ML tasks for medicine– is the construction of interpretable and explanatory models. The proper support to the human expert must go beyond a simple categorical label. In order to facilitate and improve the work of the physicians and in order to have a tangible impact in the fight against the disease, CAD systems should be able to illustrate the human expert with similar examples from the past, to identify the factors that influenced the decision, to suggest treatment options with potential pros/cons for each case, among others.

Chapter 10

Temporal Segmentation of Digital Colposcopies

An extended version of this chapter was published in [94]:

- Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. Temporal segmentation of digital colposcopies. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 262–271. Springer, 2015

Cervical cancer remains a significant cause of mortality in low-income countries. Digital colposcopy is a promising and inexpensive technology for the detection of cervical intraepithelial neoplasia. However, diagnostic sensitivity varies widely depending on the doctor expertise. Therefore, automation of this process is needed in both, detection and visualization. Colposcopies cover four steps: macroscopic view with magnifier white light, observation under green light, Hinselmann and Schiller. Also, there are transition intervals where the specialist manipulates the observed area. In this chapter, we focus on the temporal segmentation of the video in these steps. Using our solution, physicians may focus on the step of interest and lesion detection tools can determine the interval to diagnose. We solved the temporal segmentation problem using Weighted Automata. Images were described by their chromacity histograms and labeled using a KNN classifier with a precision of 97%. Transition frames were recognized with a precision of 91%.

10.1 Introduction

As was mentioned in previous chapters, the protocol recommended by the World Health Organization (WHO) [250] for the colposcopic screening covers the following steps (see Figure 10.1): macroscopic view with magnifier white light, followed by observation under green light for diagnosis of aberrant vascularization and then evaluate the cervical characteristics after exposure to acetic acid solution (Hinselmann) and potassium iodine

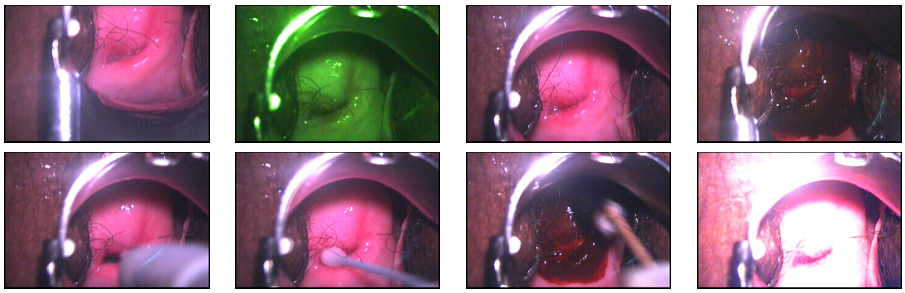


Figure 10.1: **Top:** Diagnosis steps. From left to right: macroscopic observation, green filter, Hinselmann and Schiller. **Bottom:** Transition frames. The first three frames have occlusions of the cervix area and the last one presents a strong illumination difference after removing the green filter.

(Schiller) [250]. Although Hinselmann and macroscopic observation cannot be differentiated on healthy patients, these two steps can be distinguished using contextual information. Throughout the procedure, the expert disturbs the cervix area to achieve better focus, to move from one step to the next, to clean the cervix area, etc. Figure 10.1 shows four transition frames. These scenes do not bring useful information for the diagnosis and should not be considered in the detection of lesions. In this chapter, we propose a methodology to recognize each of these modalities in a continuous video, serving as a preliminary stage for the diagnosis of cervical lesions. Also, we remove irrelevant frames from the video, where the physician is manipulating the cervix region, allowing automated techniques to retrieve information from relevant excerpts of the video. The recognition of the acquisition modality is relevant on the succeeding steps of any multimodal CAD system in order to handle each source of data properly. While several techniques have been proposed for the detection of lesions, this work constitutes the first attempt to automate the modality recognition of digital colposcopies from uncontrolled acquisition settings.

10.2 System Overview

An automated system is proposed in this chapter to segment the different steps of the colposcopic assessment. Our system can be split into three stages: transition removal, screening modality recognition (frame labeling) and temporal segmentation. Figure 10.2 illustrates this process.

In general, the temporal segmentation problem implies finding the segments and the labels simultaneously. It is a hard problem which can be addressed sequentially. In this

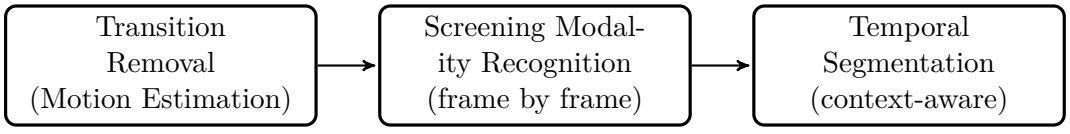


Figure 10.2: Flow chart describing the proposed framework.

work, we take advantage of the knowledge domain by labeling the frames without considering the context and then, minimizing the temporal inconsistencies using the WHO protocol definition. The labeling is done by classification using templates from previous colposcopies and the temporal segmentation is done by translating the colposcopic procedure to a non-deterministic weighted automaton, which can be implemented using Dynamic Programming (DP). The temporal boundaries optimization tries to reach maximal consistency with the preassigned labels. The remainder of this section details the proposed system.

10.2.1 Transition Removal

In order to remove transition scenes, we adopted a motion-based approach. We assume that transitions correspond to frames with high motion. First, we apply a Gaussian blur to attenuate noisy pixels. Then, motion is estimated using the Euclidean pixel-wise distance between a frame and its neighborhood (W frames before and after the given frame). Finally, we apply a thresholding operator to differentiate between transition and non-transition frames. Eq. (10.1) shows the formula that determines if a given frame belongs to a transition. Therein, I_i stands for the $i-th$ frame of the sequence I. Although more advanced approaches could be implemented, this standard procedure attained already good performance.

$$Transition(i) = \left(\frac{1}{2W} \sum_{w \in [-W..W]} \parallel \mathcal{G}(I_{i+w}) - \mathcal{G}(I_i) \parallel_2 > \ threshold \right) \tag{10.1}$$

10.2.2 Screening Modality Recognition

Each colposcopic image is represented by its one-dimensional hue histogram and saturation histogram. To efficiently reduce the presence of noisy objects in the boundaries of the image, we masked the ROI by removing everything outside an image-centered circle (with diameter equal to 0.75 of the image side). This approach considers that the cervix region occupies more than half of the cervigram image [64] and that it is approximately centered.

The recognition of each modality is done on a per-frame basis. We propose a classification method based on KNN. The similarity between two images is defined by the average distance between their histograms. We compared three histogram distances. The bin-to-bin Minkowski distance of order 1 (L_1), which is equivalent to the Histogram Intersection distance [286] and the cross-bin distances: Earth's Mover Distance (EMD) [286] and Circular Earth's Mover Distance (CEMD) [276]. Given the huge amount of images and the low intra-variance between images within the same video, we indexed an equally spaced subset of images in the KNN knowledge base. Each video contains the same number of

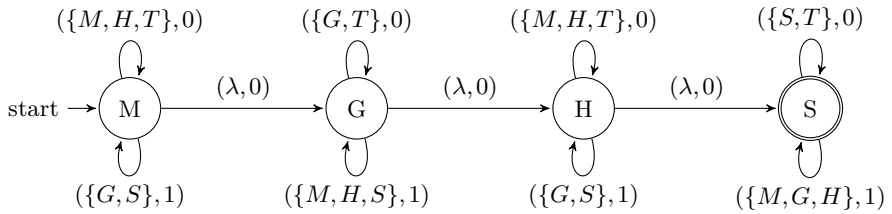


Figure 10.3: Weighted Finite Automaton that recognizes the temporal segmentation of colposcopies (Transition - T, Macroscopic view - M, Green - G, Hinselmann - H and Schiller - S.

images per phase to avoid oversampling and bias. We smooth the labels by selecting the mode of a local window.

10.2.3 Temporal Segmentation

Finally, we have to decide the temporal boundaries between the diagnosis steps. For this purpose, let's generalize the problem of temporal segmentation as the problem of recognizing a word (sequence of predicted labels) in a Weighted Finite Automaton (WFA) with minimal accumulated value. The WFA is derived from the domain-dependant protocol. Furthermore, the transition weights are related to the presence of mislabeling. If any transition in our policy either consumes an input character or moves "forward" to another state in a directed acyclic graph, the recognition problem holds the conditions to formulate a DP implementation. Figure 10.3 shows a graphical representation of the automaton. We denote each phase by its first letter. The automaton represented in Figure 10.3 is formally defined as $A =< \Sigma, Q, \Delta, c, \{M\}, \{S\}, 0, v >$, where

- $\Sigma = \{T, M, G, H, S, \lambda\}$.
- $Q = \{M, G, H, S\}$.
- $\Delta \subseteq Q \times \Sigma \times Q$ is the transition relation defined below, together with the cost function $c : \Delta \longrightarrow \{0, 1\}$. The accepted labels of each state are defined by the top-loop transitions shown in Figure 10.3.
  - $(s, q, s) \longrightarrow 0$, if $q \in accepted\_labels(s)$.
  - $(s, q, s) \longrightarrow 1$, if $q \notin accepted\_labels(s)$.
  - $(s, \lambda, s') \longrightarrow 0$, if $s' \neq s$ and s' follows s in the protocol.
  - $v \in \mathbb{N}$, the minimal threshold that accepts the word.

Using the same reasoning we could instantiate any other policy in a straightforward manner. As we said before, given that after each transition the recognition problem is smaller, we can implement this automaton using the DP function defined in the Eq. (10.2), where *seq* stands for the sequence of labels predicted by the step classifier, next(s)

returns the protocol step that follows s and has_next$(s) \equiv (s \neq S)$. It is assumed that the preconditions are evaluated in the same order they are shown. The optimal boundaries can be retrieved from the DP matrix. Since the number of steps in the colposcopic procedure is constant, the performance of the algorithm is linear in the length of the sequence.

$$B[i,s] = \begin{cases} 0, & \text{if } i = length(seq) \\ B[i+1,s], & \text{if } seq[i] \in \{\text{transition, s}\} \\ min(B[i+1,M], B[i,H]), & \text{if } s = M \ \wedge seq[i] = H \\ B[i+1,H], & \text{if } s = H \ \wedge seq[i] = M \\ min(B[i,\text{next}(s)], 1 + B[i+1,s]), & \text{if has\_next(s)} \\ 1 + B[i+1,s]), & \text{otherwise} \end{cases} \tag{10.2}$$

10.3 Experiments

In this section, we describe the experiments that we performed in this work. We gathered a dataset of 56 colposcopies from different patients that cover a total of 143640 colposcopic images, with every image resized to 64×64 pixels. Sequences were manually annotated by a specialist. The videos and annotations are public on request. Table 10.1 shows the number of frames per video in each phase. In order to avoid biased results due to differences in the length of the procedures, every patient was equally weighted in the compilation of the results.

Table 10.1: Statistics of the class distribution per video

	Number of Frames			**Percentage**	
Class	**Min**	**Max**	**Avg.**	**Max**	**Avg.**
Transition	0	6488	1071	59.76	39.53
Macroscopic	66	1380	313	100.00	13.88
Green	0	767	187	31.32	7.72
Hinselmann	0	2104	688	52.60	28.45
Schiller	0	2752	304	32.33	10.42
Video	200	11998	2565	–	–

For the assessment of every step of the proposed framework, we used a leave-one-patient-out cross-validation (LOOCV). In this sense, each colposcopy is entirely new for the system at the evaluation stage.

For the classification of the transition frames, we performed several experiments varying the number of neighbors. This parameter is internally learned using LOOCV. Figure 10.4 shows the performance of the algorithm at different values of W. Table 10.2 shows the classification results for this stage. The average accuracy of the transition recognition is

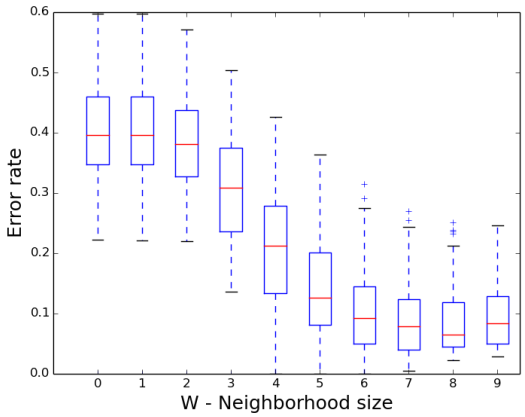


Figure 10.4: Error rate of the transition removal method using different neighborhood sizes

Table 10.2: Transition Classifier Results

Class	Precision	Recall	F-measure
non-transition	0.9325	0.9129	0.9206
transition	0.8610	0.8820	0.8672
Weighted Avg.	0.9146	0.9087	0.9087

90.86%. Furthermore, 93.25% of the frames that pass to the next stage (colposcopic-step classification) belong to a non-transition interval. There is room for human error in the decision of the boundaries of the transition intervals, i.e., it is difficult for a trained human to decide where is the beginning of a transition interval and where it ends. This artifact is also common between the different steps of the colposcopic evaluation. Therefore, the errors shown in these experiments are prone to small human inaccuracies.

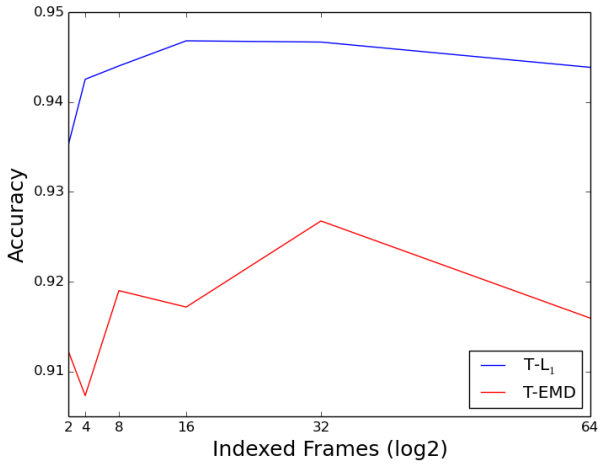


Figure 10.5: Colposcopic Step accuracy varying the number of indexed frames

For the assessment of the step classification, the number of neighbors in the KNN

Table 10.3: Average classification metrics per class: Macroscopic, Green, Hinselmann and Schiller. Results with 16 indexed frames per video. The results denoted by T-d, where d is the similarity distance, include the temporal segmentation step.

Phase	**Distance**	**Transition**				**Non-Transition**			
		Acc.	**Prec.**	**Rec.**	**F**	**Acc.**	**Prec.**	**Rec.**	**F**
Macro	L_1	0.82	0.36	0.28	0.52	0.70	0.38	0.31	0.52
	T-L_1	**0.96**	**0.99**	**0.78**	**0.84**	**0.98**	**1.00**	**0.95**	**0.95**
	EMD	0.80	0.32	0.28	0.48	0.65	0.33	0.31	0.49
	T-EMD	0.95	**0.99**	0.74	0.80	0.96	**1.00**	0.89	0.89
Green	L_1	**0.97**	0.97	**0.67**	**0.75**	**1.00**	**1.00**	**0.98**	**0.98**
	T-L_1	**0.97**	**0.98**	0.66	0.74	0.99	**1.00**	0.96	0.96
	EMD	**0.97**	0.96	**0.67**	**0.75**	**1.00**	**1.00**	**0.98**	**0.98**
	T-EMD	**0.97**	0.97	0.63	0.70	0.99	0.99	0.91	0.90
Hins	L_1	0.80	0.75	0.55	0.56	0.67	0.76	0.62	0.60
	T-L_1	**0.92**	**0.96**	**0.79**	**0.81**	**0.92**	**0.98**	**0.89**	**0.88**
	EMD	0.80	0.76	0.47	0.54	0.65	0.76	0.53	0.58
	T-EMD	0.91	0.93	0.76	0.77	0.89	0.95	0.86	0.83
Sch	L_1	0.90	0.74	0.60	0.60	0.88	0.79	**0.93**	0.79
	T-L_1	**0.91**	**0.83**	**0.61**	**0.65**	**0.89**	**0.89**	**0.93**	**0.82**
	EMD	0.88	0.67	0.54	0.52	0.82	0.70	0.82	0.62
	T-EMD	0.89	0.77	0.55	0.55	0.84	0.84	0.83	0.71
Avg.	L_1	0.87	0.70	0.52	0.61	0.81	0.73	0.71	0.72
	T-L_1	**0.94**	**0.94**	**0.71**	**0.76**	**0.95**	**0.97**	**0.93**	**0.90**
	EMD	0.86	0.68	0.49	0.57	0.78	0.70	0.66	0.67
	T-EMD	0.93	0.91	0.67	0.70	0.92	0.94	0.87	0.83

was set to 5, and the hue and saturation histograms had 180 and 256 bins respectively. We performed experiments varying the number of indexed frames in the KNN database. The results of this experiments can be seen in Figure 10.5. The highest accuracy was achieved with 16 indexed frames per phase per video. These results include the temporal segmentation.

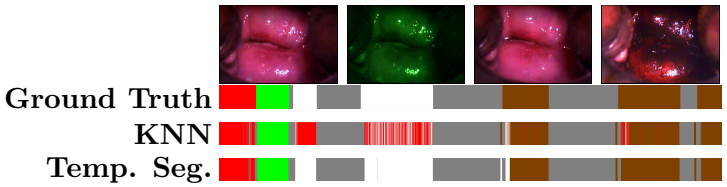


Figure 10.6: Timeline with the steps represented by colors: Transition (gray), Macroscopic View (red), Green (green), Hinselmann (white) and Schiller (brown).

Table 10.3 shows the classification metrics for each colposcopic step using two distance functions: L_1 (equivalent to Histogram Intersection) and EMD. We compare each distance before and after temporal segmentation. Contrary to what we thought, CEMD did not improve the accuracy but obtained a significant performance impact. Therefore, we only show the results related to the first two distances. As can be seen in the results, the selection of the decision boundaries using the proposed DP algorithm improves the detection of almost every stage. On average, the temporal segmentation algorithm improved the accuracy in 14% and 28% in the Macroscopic view phase. In general, the L_1 distance

achieved better performance than the EMD. Figure 10.6 shows an example of the step detection results before and after the temporal decision.

10.4 Conclusions

In this work, we provided a framework to temporarily segment a colposcopic assessment according to its different steps. To assess the quality of the proposed framework we gathered and annotated an open dataset of 56 colposcopies. The proposed framework achieved a precision of 91.46% in the transition detection using an efficient threshold on motion estimation. Using chromacity information (hue and saturation histograms), we achieved a precision of 96.65% in the step classification. As we observed in the experiments, for this problem the L_1 distance behaved better than the EMD, because histograms from different stages are near, and noisy pixels have a high weight in the resulting EMD. Contextual information provided valuable information to smooth and improve the classification results obtained by the per-frame KNN classifier.

Chapter 11

Ordinal Segmentation

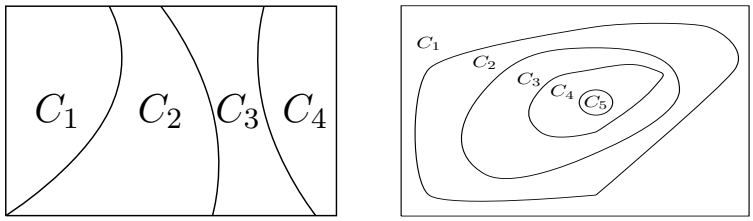
This chapter was published in [90]:

- Kelwin Fernandes and Jaime S. Cardoso. Ordinal image segmentation using deep neural networks. In *Neural Networks (IJCNN), 2018 International Joint Conference on*. IEEE, 2018

Ordinal arrangement of objects is a common property in biomedical images. Traditional methods to deal with semantic image segmentation in this setting are *ad-hoc* and application specific. In this chapter, we propose ordinal-aware deep learning architectures for image segmentation that enforce pixelwise consistency by construction. We validated the proposed architectures on several real-life biomedical datasets and achieved competitive results in all cases.

11.1 Introduction

Semantic image segmentation has attracted attention in the research community in the last decades. From its early stages with handcrafted features [157, 268, 304] to more recent approaches based on DNN[24, 44, 216, 247], semantic segmentation has captured a significant portion of the research efforts on image processing. While the focus of image segmentation is to partition an image into a subset of disjoint coherent regions, in semantic image segmentation, a label is assigned to each pixel corresponding to a concept of interest in the corresponding application [247]. Semantic image segmentation has been successfully used in a wide spectrum of scenarios, from biomedical applications where the acquisition setting is often well controlled and the number of classes is relatively small [157] to more challenging scenarios with unconstrained images acquired *in the wild* like scene parsing and recognition for autonomous driving, where a large open set of classes may appear [24]. The vast amount of unconstrained image databases that have been created in the last few years allowed DNN to gain space in this field [24, 44, 216, 247]. However, in



(a) Linear (b) Nested

Figure 11.1: Ordinal arrangements

some scenarios such as biomedical imaging, where data is scarce, prior knowledge should be imposed to overcome the difficulty of learning robust data representations. Thereby, it is normal to observe even nowadays a significant proportion of handcrafted pipelines with ad-hoc approximations to the semantic segmentation of objects in medical applications [133, 210].

In this work, we address the problem of ordinal semantic segmentation, where the objects of interest hold a spatial order relation. This setting is frequently observed in medical imaging, where the disposition of organs or body structures is known *apriori*. For example, in mammograms, the pectoral muscle, breast and remaining background are always found in a linear arrangement; in cervigrams, the external orifice, SCJ, cervix, and speculum hold a nested arrangement (i.e., one inside the other). Figure 11.1 illustrates typical cases of ordinal segmentation tasks with linear and nested ordinal transitions. While the problem of ordinal classification [138] has been widely studied in the past [41, 112, 138], promoting class coherence in images has not been systematically studied. Thus, most attempts to handle this kind of property focus on cascade methods where the ROI is progressively narrowed class by class or *ad-hoc* strategies. Needless to say that these approaches work in a local fashion not being able to recover from errors in previous stages of the ROI estimation.

11.2 Related work

In this section, we summarize the current research landscape on the two main topics of this work: semantic image segmentation and ordinal classification.

11.2.1 Semantic Image Segmentation

Traditional techniques to tackle image segmentation span over a large set of techniques, including basic thresholding, graph-based methodologies, active shape models, and clustering [268]. These methods are usually combined with handcrafted feature extraction processes and traditional classifiers to assign a semantic level to each pixel (or superpixel).

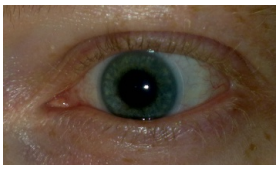
The same pixelwise classification approach can be extended to deep learning by using well known encoder-decoder architectures [24, 247, 283], where the network aims to predict the entire segmentation mask. In order to alleviate the large number of parameters in dense layers, several architectures have been proposed in the past [283]. U-net is a fully-convolutional architecture for semantic image segmentation that has achieved state-of-the-art results on medical segmentation tasks [283]. U-net incorporates skip connections between layers at the same level of the encoder and decoder blocks in order to improve the propagation of the gradients and to improve the resolution of the reconstruction process. An alternative idea to overcome data scarcity is using pre-trained networks on large classification databases such as ImageNet [24, 44, 247].

Since pixelwise approaches tend to produce incoherent spatial labeling, these techniques are often combined with post-processing strategies such as Conditional Random Fields to propagate information to a global level [44]. Finally, all these architectures are frequently combined using a cascade of decision networks that allow further refinement of the segmentation masks.

11.2.2 Ordinal Classification

Ordinal classification, also known as ordinal regression [41,112,138,265,266], can be understood as the supervised learning task of classifying observations into a finite set of classes $C_i, i \in [1, k]$ such that classes are ordered $C_1 \prec C_2 \prec \ldots \prec C_k$.

Despite traditional techniques for nominal multiclass settings can be used for learning ordinal problems, learning stable classifiers that reduce eventual inconsistencies in the decision space are desirable. Also, regression techniques and multipartite ranking strategies can be used with further post-processing at the expense of suboptimal results. Therefore, ordinal-aware classifiers have been proposed in the past to tackle this problem [41, 112, 265, 266]. We can broadly categorize these ideas according to the type of decision region they induce. On the one hand, we have *hard-ordinal* methodologies that enforce the decision boundaries to be parallel (either on the feature space or on a high-dimensional space). This can be achieved by pre-processing the dataset using the data replication method [41] or by imposing such parallelism directly in the learning process [50]. In general, this idea can be understood as having a single decision hyperplane that projects the observations into a linear space and a set of thresholds that dichotomize the decision space into contiguous non-overlapping regions [50]. On the other hand, we have *soft-ordinal* methodologies that do not force the boundaries to be parallel. A well-known representative of this paradigm is the method proposed by Frank & Hall (F&H) [112] where $k-1$ classifiers $\{D_1, D_2, \ldots, D_{k-1}\}$ are learned independently such that D_i classifies an observations as belonging to the first i classes or to the last $k-i$ classes. Then, the final prediction is done by aggregating the decisions using a cascade or voting rule.



(a) Image

(b) Nominal masks

(c) Ordinal masks

Figure 11.2: Visualization of the ground-truth masks for the segmentation of sclera, pupil and iris using the nominal and ordinal representations.

11.3 Ordinal Segmentation using Deep Neural Networks

Let us denote the neighborhood of a given pixel p as $\mathcal{N}(p)$. We assume the neighborhood is formed by the surrounding pixels in a k-connected adjacency and the central pixel itself. We will denote the class of pixel p as $C^{(p)}$. We say a semantic segmentation is strictly consistent with an ordinal segmentation setting if, in the neighborhood of every pixel p in the image $\mathcal{I}$, at most two consecutive classes are present.

$$\bigwedge_{p \in \mathcal{I}} \left(\max_{q \in \mathcal{N}(p)} C^{(q)} - \min_{q \in \mathcal{N}(p)} C^{(q)} \right) \leq 1 \tag{11.1}$$

Such assumption is often too hard in real-life settings, where the object pose may generate occlusions that induce non-adjacent transitions. However, as we will show in this chapter, assuming that prior may be beneficial for segmentation algorithms when most pixels in the image are consistent with an ordinal segmentation.

11.3.1 Ordinal Class Encoding

We propose to use the class ensemble proposed by F&H [112] for ordinal classification. Thus, for a segmentation task with k ordinal classes, we train $k-1$ models such that each model D_i estimates the binary segmentation mask of the classes $\mathcal{Y}_0 = \bigcup_{j=1}^{i} C_j$ and $\mathcal{Y}_1 = \bigcup_{j=i+1}^{k} C_j$. Thus, while the expected segmentation masks for the nominal case would follow

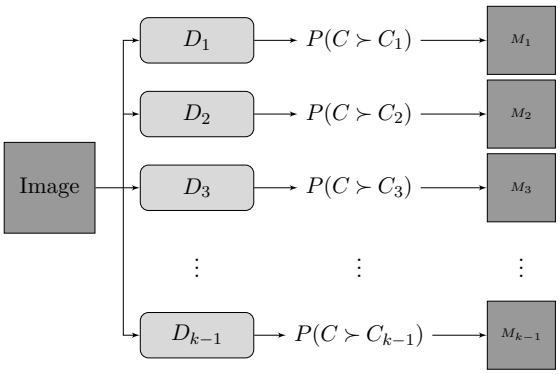


Figure 11.3: Ordinal ensemble based on the Frank & Hall approach.

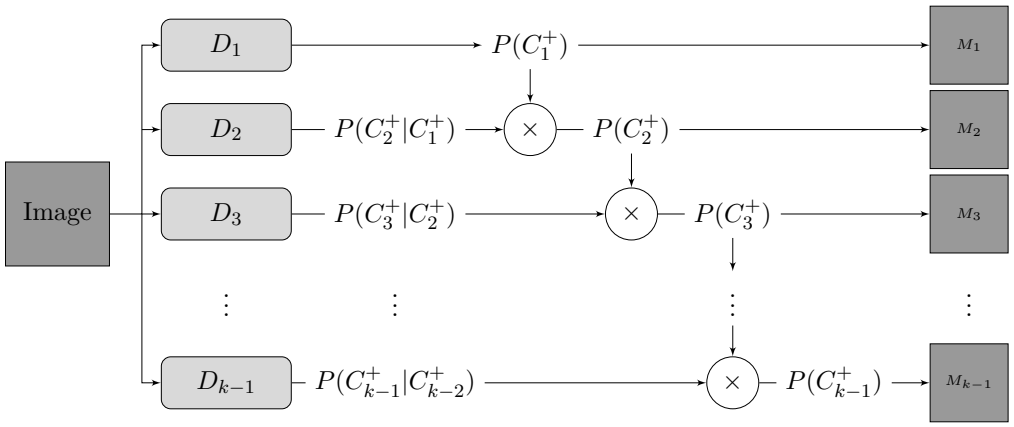


Figure 11.4: Ordinal consistent network based on the Frank & Hall approach.

the ones illustrated in Figure 11.2b, the segmentation masks of the proposed approach approximate the masks from Figure 11.2c. Figure 11.3 illustrates the ensemble. The main intuition behind this idea is that the constituent parts of an object also belong to the object. Thereby, the data representation of an object as the union of its parts should be easier to represent and more coherent than the individual representation of its components. Another advantage of this formulation is the reduction of inter-class transition boundaries, which are the most critic parts on image segmentation. This property can be easily proved by the fact that the set of transitions in the F&H ordinal formulation is a proper subset of the nominal transitions.

In the original formulation of the F&H approach, each model in the ensemble is learned independently. However, using DNN allows to seamlessly integrate the optimization process of all the models in the ensemble by learning a common feature representation. Then, the final class for each pixel is the highest class with a probability higher than 0.5. Here, we are assuming monotonicity on the predicted probabilities. In the following section, we design a deep architecture that holds such property.

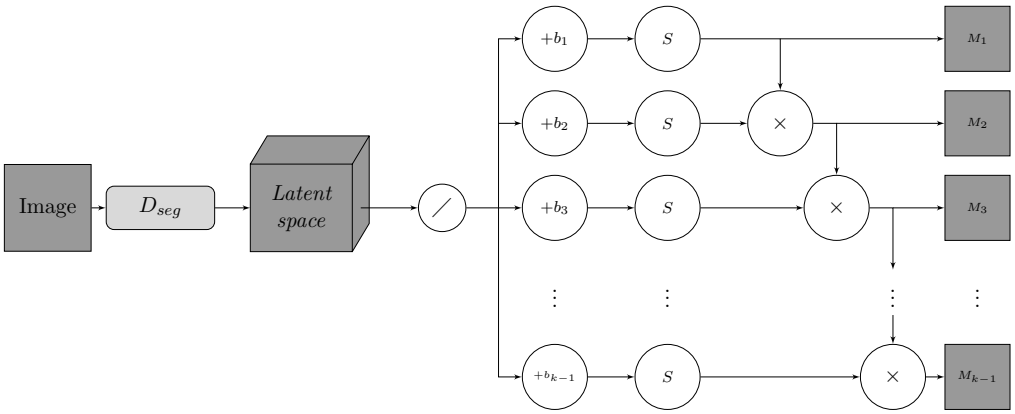


Figure 11.5: Ordinal consistent network with parallel decision boundaries. $\diagup$ denotes the linear model on the pointwise latent space, $+b_i$ denotes the addition of the class-specific bias term and s is the sigmoid function.

11.3.2 Pixelwise consistency

Learning a DNN – or any other standard classifier – with the aforementioned class encoding does not guarantee that the output probabilities are consistent (i.e., monotonous). For instance, the probability assigned to a pixel by the estimator D_i might be lower than the probability assigned by D_{i+1}. Therefore, in order to facilitate the learning process, we aim to force consistency between the class probabilities assigned to each pixel by construction. Namely, we want to ensure that, for each pixel p, $P(D_i^p) > P(D_{i+1}^p)$. Hereafter, we assume that the segmentation models return a probabilistic output $[0, 1]$.

Applying the Bayes theorem, we obtain the following

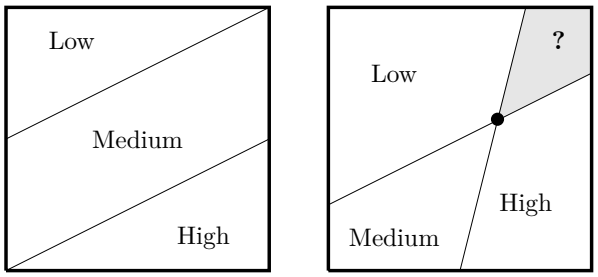
$$P(C_{i+1}^+) = \frac{P(C_{i+1}^+|C_i^+)P(C_i^+)}{P(C_i^+|C_{i+1}^+)},$$

where C_i^+ stands for the transitive closure of class i and $P(C_i^+|C_{i+1}^+) = 1$ by definition. Therefore, we can interpret the output of each model in the ensemble as the conditional probability $P(C_{i+1}^+|C_i^+)$. Pixelwise consistency can be achieved as:

$$P(C_{i+1}^+) = P(C_{i+1}^+|C_i^+)P(C_i^+),$$

where $P(C_{i+1}^+|C_i^+)$ is the output of the $(i+1)$-*th* model and $P(C_i^+)$ is the corrected probability of the class i. The base case of the recursion is the first model $P(C_1^+)$. Figure 11.4 illustrates the architecture of the pixelwise ordinal consistent model. A relevant property of the proposed model is that estimators from lower classes do not need to learn the inner inter-class boundaries and can focus on learning the global concepts while estimators from upper classes can focus on learning to recognize the inner transitions.

In traditional segmentation techniques, this kind of consistency is achieved by reducing the ROI from the base classes to the top ones. The ROI operation can be understood as the intersection of both masks. From fuzzy set theory, an alternative approach to compute a pixelwise ordinal-consistent segmentation is to use the minimum class membership between D_i and D_{i+1} (i.e. replacing the multiplication of the probabilities by the minimum). A main disadvantage of the minimum operator is its non-differentiability at the transition between the two branches of the operator. Moreover, in degenerated cases, gradients only propagate by a single branch of the network during training.



(a) Strong ordinal model with parallel hyperplanes.

(b) Degraded ordinal model with non-parallel hyperplanes.

Figure 11.6: The boundary intersection problem in ordinal classification

11.3.3 Parameter sharing and Decision Boundary Parallelism

So far, we have defined how to build ordinal pixelwise-consistent deep segmentation networks. However, current models do not necessarily hold spatial-consistency constraints. Namely, two neighboring pixels may be predicted as non-contiguous classes. Two proposals led us to mitigate such problem: 1) the F&H encoding reduces the critic regions by diminishing the perimeter of object transitions, and 2) the architecture proposed in Section 11.3.2 guarantees that outputs are locally consistent.

In most semantic segmentation tasks, neighboring pixels in the image reflect similar properties. Therefore, it is expected to observe a similar latent representation of neighboring pixels at a given layer in the network. Thus, we can promote spatial consistency by forcing the final decision function at a given latent space to be well-behaved in terms of ordinal constraints. A common assumption in ordinal classification is the parallelism of the decision boundaries, which strictly avoids the non-ordinal class transitions and intersection problem (see Figure 11.6). While this assumption is often too restrictive for real-life datasets with linear models, imposing such behavior in a latent high-dimensional space may induce robust models, especially when data is scarce.

So far, we described the proposed framework for ordinal segmentation as having an independent model per output mask. However, multiclass ANN are typically learned by using a single model that outputs a vector with the probability of each class. Let us define

the latent pixelwise feature map computed by a DNN for image segmentation D_{seg}. We can induce parallel decision boundaries by considering a linear model with common slope coefficients and individual bias terms. Namely,

$$D^p_{par,i} = b_i + \omega^T \cdot D^p_{seg} > 0,\ i = 1, \ldots, k. \tag{11.2}$$

The probabilistic estimation of $D^p_{par,i}$ can be computed as the sigmoid activation of the $b_i + \omega^T \cdot D^p_{seg}$ linear function. Thus, the final model proposed in Eq. (11.2) and schematized in Figure 11.5 promotes spatial consistency by removing the intersection between the decision hyperplanes at a given latent space.

Contrary to pixelwise consistency, we do not achieve spatial consistency by construction. Namely, the combination of the proposed techniques promotes such behavior on properly regularized models but do not ensure strict ordinal transitions. It is relevant to highlight that strict spatial consistency is not necessarily ideal, especially when object's pose may induce small non-ordinal transitions.

11.3.4 Generalization to Domains with Arbitrary Partial Orders

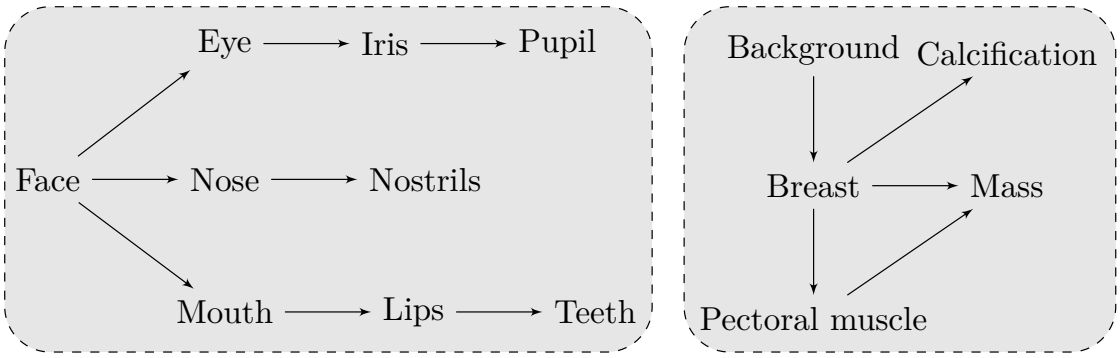


Figure 11.7: Domains with spatial partial orders

Let us define a classification task where the classes belong to a partially ordered set. Namely, we extend the total order of classes by allowing pairs of classes that are not comparable. Examples of this setting can be found in a wide diversity of applications, Figure 11.7 illustrates several segmentation problems that belong to this kind. The partial ordering defined between the classes induces a directed acyclic graph of precedence. The proposed methodology can be extended in a straightforward manner to this scenario by:

- Using the Frank and Hall encoding on the space of predecessors and successors. Namely, the model D_i is a binary segmentation model trained to predict:

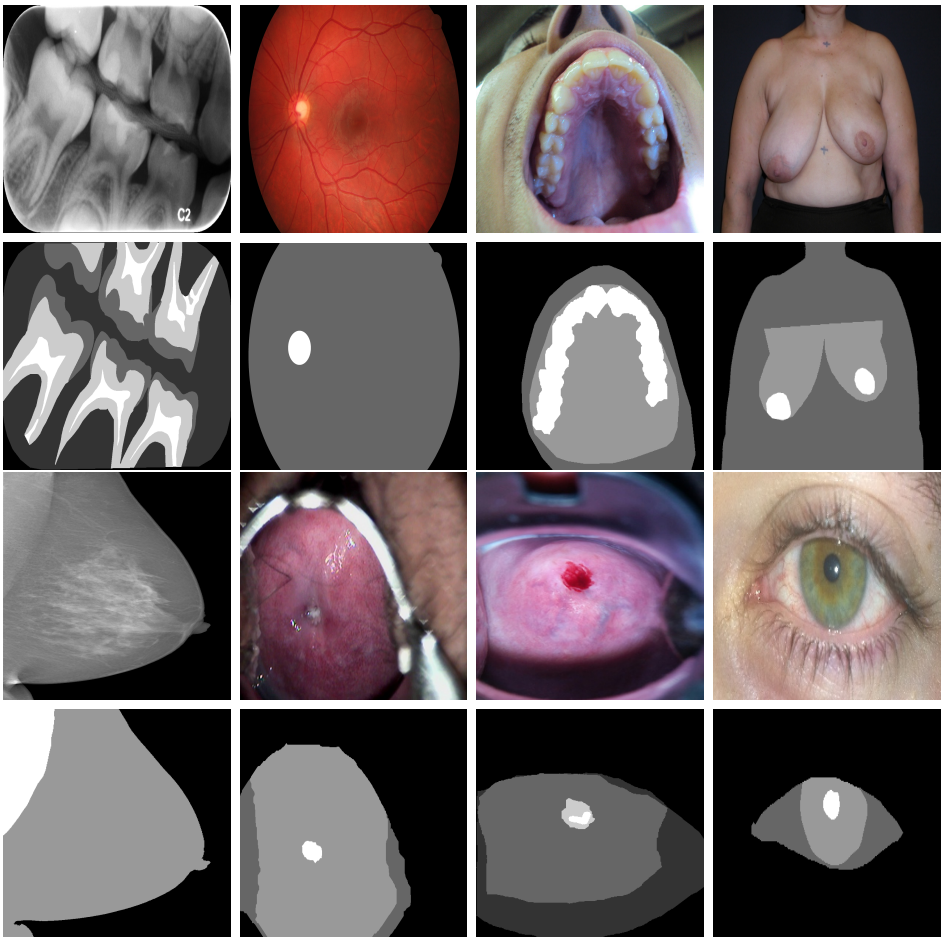


Figure 11.8: Sample images and their corresponding ordinal mask from each dataset. Datasets are ordered by appearance on Table 11.1. The intensity of the classes resembles the order used for the ordinal labels, being the black and white objects from the first and last classes respectively.

$$\mathcal{Y}_0 = \{C_j | 1 \leq j \leq k, C_j \preceq^+ C_i\} \tag{11.3}$$

$$\mathcal{Y}_1 = \{C_j | 1 \leq j \leq k, C_i \prec^+ C_j\} \tag{11.4}$$

where $\prec^+$ is the ordering induced by the transitive closure of the class poset.

- Applying the consistency operator to the outputs of all the predecessor models.

We did not consider this generic setting in the experimental assessment in order to simplify the analysis of the results.

Table 11.1: Summary of the datasets.

Dataset	**Ref.**	**# Imgs**	**# Classes**
Teeth-ISBI	[332]	40	5
HRF	[38]	45	3
Teeth-UCV	[102]	100	4
Breast Aesthetics	[40]	120	4
InBreast	[239]	194	3
Cervix-HUC	[94][95]	287	4
Cervix-MobileODT	[167]	1480	5
Mobbio	[299]	1817	4

11.4 Experiments

We validated the performance of the proposed architectures on eight real-life biomedical datasets. The datasets cover a wide spectrum of applications, acquisition modalities, and difficulties. Further details and sample images are shown in Table 11.1 and Figure 11.8 respectively. We used a 5-fold cross-validation strategy with 3 folds for training, 1 fold for validation and 1 for testing. Model performance is measured in terms of Hausdorff distance [160], Dice coefficient [75] and Average MAE. The best model parametrization was chosen using each target metric on the validation set.

We compare the performance of our approach using the U-net architecture [283] as the base model. The network depth ranges from 2 to 4 groups of convolution blocks (i.e., two convolution layers and one pooling layer) for the encoder and decoder sections. The number of filters for the first layer is 32 and doubles after each block (e.g., 32, 64, 128). Data augmentation is applied in all datasets, covering horizontal/vertical flips, scaling and contrast stretching when possible. Networks are optimized for a maximum number of 500 iterations with Adadelta optimizer [359] and batch-size 16. Early stopping with patience of 100 iterations was used to control overfitting.

As loss function to train the models, we chose the product of the Dice's coefficient [75] per class (see Eq. (11.5)). While we considered other loss functions such as cross-entropy, the imbalanced distribution of the classes led to naive models that always predict the same class. Conversely, the Dice's coefficient, also known as F1-score in the classification community, is a robust metric to assess models on imbalance settings [58].

$$\prod_{i=1}^{k-1} \text{Dice}(M_i, \hat{M}_i) \tag{11.5}$$

$$\text{Dice}(X, Y) = \frac{2X \bigcap Y}{|X| + |Y|} \tag{11.6}$$

It is important to highlight that Eq. (11.5) is not symmetric. Thereby, the results will depend on the class ordering (i.e., $A \prec B \prec C$ or $C \prec B \prec A$). However, in traditional

Table 11.2: Average model performance where – denotes models with ordinal encoding (section 11.3.1) and **Cons** denotes models with pixelwise consistency (section 11.3.2). The best result for each dataset and metric is presented in bold. The number of datasets where each model achieves the best results is shown at the bottom of each table.

(a) Hausdorff distance

Dataset	**U-net**	**Ordinal**		**Ord-Parallel**	
		–	**Cons**	–	**Cons**
Teeth-ISBI	10.52	**9.18**	9.90	9.25	13.59
HRF	5.09	3.22	2.58	3.27	**1.26**
Teeth-UCV	5.70	**5.48**	5.67	6.50	8.51
Breast Aesthetics	2.38	2.30	**2.23**	2.36	2.46
InBreast	22.79	2.79	2.73	**2.67**	3.07
Cervix-HUC	18.20	12.50	12.31	**12.21**	12.36
Cervix-MobileODT	15.68	6.96	6.54	6.45	**6.02**
Mobbio	5.65	4.26	6.78	**4.23**	4.86
Best	0	2	1	3	2

(b) Dice coefficient

Dataset	**U-net**	**Ordinal**		**Ord-Parallel**	
		–	**Cons**	–	**Cons**
Teeth-ISBI	28.41	51.27	44.46	**52.11**	27.83
HRF	66.18	74.27	84.85	78.66	**94.34**
Teeth-UCV	87.37	**87.49**	87.40	85.33	74.00
Breast Aesthetics	93.19	93.35	**93.93**	93.67	92.86
InBreast	65.25	97.31	97.18	**97.26**	96.89
Cervix-HUC	38.64	48.69	48.78	50.08	**51.24**
Cervix-MobileODT	39.67	64.40	66.15	63.78	**66.98**
Mobbio	20.06	40.16	35.19	39.96	**40.54**
Best	0	1	1	2	4

(c) Average MAE

Dataset	**U-net**	**Ordinal**		**Ord-Parallel**	
		–	**Cons**	–	**Cons**
Teeth-ISBI	0.9555	0.7579	0.8711	**0.7566**	1.3113
HRF	0.0187	0.0409	0.0156	0.0450	**0.0096**
Teeth-UCV	0.1197	**0.1150**	0.1179	0.1440	0.2109
Breast Aesthetics	0.0341	0.0343	**0.0308**	0.0337	0.0341
InBreast	0.3085	0.0155	0.0162	**0.0153**	0.0189
Cervix-HUC	0.4971	0.3389	0.3447	**0.3257**	0.3295
Cervix-MobileODT	0.5432	0.1547	0.1440	0.1360	**0.1342**
Mobbio	0.6125	0.2064	0.2552	**0.2059**	0.2248
Best	0	2	1	4	2

semantic segmentation tasks, we are more interested in some categories (e.g., foreground objects) than in the others (e.g., background), so this ordering appears naturally. In other cases, the best direction of the ordering may be defined by cross-validation.

As can be seen in the results (see Table 11.2), the U-net architecture is surpassed by at least one of the proposed models in all databases. In general, using parallel decision boundaries achieved better results than soft-ordinal models in terms of Dice coefficient and MAE. While hard parallelism among hyperplanes might be too restrictive for linear models, we validated that DNN can learn feature representations where such constraint is beneficial. All the proposed alternatives (i.e., with/without pixelwise consistency, soft/hard ordinal) achieved top performance in several datasets. Therefore, the best model among all possible combinations is task dependent.

11.5 Conclusions

This chapter addresses the problem of ordinal semantic segmentation, where objects hold an ordinal spatial arrangement, either in a nested or linear ordering. We propose several DL alternatives to tackle the problem from the two major perspectives on ordinal modeling: parallel and non-parallel decision boundaries. We validated the proposed strategy in eight biomedical datasets and achieved the best results when compared with the state-of-the-art U-net architecture.

Chapter 12

Risk Prediction and Quality Assessment of Digital Colposcopies

An extended version of this chapter was published in [95]:

- Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. Transfer learning with partial observability applied to cervical cancer screening. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 243–250. Springer, 2017

Additional work on the prediction of cervical cancer risk was published in [99]:

- Kelwin Fernandes, Davide Chicco, Jaime S. Cardoso, and Jessica Fernandes. Supervised deep learning embeddings for the prediction of cervical cancer diagnosis. *PeerJ Computer Science*, 2018

Cervical cancer remains a significant cause of mortality in low-income countries. As in many other diseases, the existence of several screening/diagnosis methods and subjective physician preferences creates a complex ecosystem for automated methods. In order to diminish the amount of labeled data from each modality/expert, we propose a regularization-based TL strategy that encourages source and target models to share the same coefficient signs. We instantiated the proposed framework to predict cross-modality individual risk and cross-expert subjective quality assessment of colposcopic images for different modalities. Thus, we are able to transfer knowledge gained from one expert/modality to another.

12.1 Introduction

Despite the possibility of prevention with regular cytological screening, cervical cancer remains a significant cause of mortality in low-income countries. This being the cause of

more than half a million cases per year, and killing more than a quarter of a million in the same period [94]. As in many other diseases, the existence of several screening and diagnosis methods creates a complex ecosystem from a CAD system point of view. For instance, in the detection of pre-cancerous cervical lesions, screening strategies include cytology, colposcopy (covering its several modalities [94]) and the gold-standard biopsy. In developing countries, resources are scarce and patients usually have poor adherence to routine screening due to low problem awareness. Consequently, the prediction of the individual patient's risk and the best screening strategy during her diagnosis becomes a fundamental problem. Most of these screening methods highly depend on the physician expertise and subjective comfort in the decision process, being a key aspect to improve data acquisition using the physician preferences.

Thereby, from a technical point of view, all these predictive tasks are immersed in a multi-modal and multi-expert setting. Traditionally, supervised learning techniques would require to collect a vast amount of data from each source (i.e., modalities and experts) and to build predictive models separately for each task. To overcome the data scarcity problem, we propose to use TL. As was referred in chapter 5, TL aims to extract knowledge from at least one source task and use it when learning a predictive model for a new target task [257]. The intuition behind this idea is that learning a new task from related tasks should be easier (faster, with better solutions or with less amount of labeled data) than learning the target task in isolation.

In this work, we focus on the aforementioned HTL framework based on structural model similarity. More specifically, we validate the performance of transferring the sign of the coefficients in linear models. In order to prove its adequacy to different problems, we instantiated this idea to two different problems: cross-modal individual risk prediction, and cross-modal and cross-expert QA of digital colposcopies.

12.2 Methodology and Validation Strategy

In this work we focus on linear predictive models for regression (e.g. Linear Regression) and classification (e.g. LR, SVM). Thereby, we assume that our model can be defined by a vector of coefficients $\omega \in \mathbb{R}^{d+1}$, which includes the bias term ω_0. Here, we are interested in transferring the contribution direction of each feature (i.e. coefficient sign) instead of its importance in the source task (i.e. coefficient magnitude). Eq. (5.11) defines a dissimilarity regularizer that encourages sign relatedness, where ω^{src} and ω^{tgt} denote the source and target coefficients respectively. We follow the transfer scheme proposed in Eqs. (5.11)-(5.12) and illustrated in Figure 5.3.

Data was split using a stratified training-test partition (80-20). Then, in order to validate the model performance on different stages of the data acquisition process, the training set was randomly subsampled in 10 nested subsets with several sizes (10%, 20%, 30%, ..., 100%). Each experiment was repeated 30 times varying the test partition. The

regularization factor (λ) and all the remaining intrinsic hyper-parameters were learned using Stratified K-fold cross-validation (K = 3) over the training set.

For each method, the normalized signed Area Under the gain Curve (sAUC) is measured when compared with training the model using target data only, where the gain is measured in terms of percentage relative gain. Thus, positive gain reflects positive transfer and, analogously, negative gain reflects negative transfer.

We instantiate the proposed sign-transfer method to two linear models: linear regression for the risk prediction task and SVM for the QA task. In each case, we validate the proposed method with fixed sign importance ($\alpha = 1$) - denoted as Sign - and with varying tradeoff between sign agreement and coefficient magnitude ($0 \leq \alpha \leq 1$) - denoted as α-Sign. The proposed regularizers are compared to the state-of-the-art approach, hereafter referred as Diff, where the model is learned using full-observability transfer by regularizing coefficients to be similar to the source-model coefficients [85, 184, 197, 267].

12.2.1 Risk Factors

In this section, we instantiate the proposed partial transfer technique to predict the individual patient's risk when multiple screening strategies are available (i.e., colposcopy using acetic acid - Hinselmann, colposcopy using Lugol iodine - Schiller, cytology, and biopsy). For this purpose, a database with 858 patients including demographic information, habits and historical medical records was collected (see Table 12.1). Several patients decided not to answer some of the questions due to privacy concerns. Hence, the features denoted by bool $\times$ T, T $\in$ {bool, int}, were encoded as two independent values: whether or not the patient answered the question and the reported value. Missing values were filled using the sample mean. Categorical features were encoded using the one-of-K scheme.

Table 12.1: Features acquired in the risk factors dataset.

Feature	**Type**	**Feature**	**Type**
Age	int	IUD (years)	int
# sexual partners	bool $\times$ int	STDs	bool $\times$ bool
Age of 1st sexual intercourse	bool $\times$ int	STDs (how many?)	int
# of pregnancies	bool $\times$ int	Diagnosed STDs	categorical
Smokes?	bool $\times$ bool	STDs (years since first diag.)	int
Smokes? (years & packs)	int $\times$ int	STDs (years last diag.)	int
Hormonal Contraceptives?	bool	Has previous cervical diag.?	bool
Horm. Contr.? (years)	int	Prev. cervical diag. (years)	int
Intrauterine device? (IUD)	bool	Prev. cervical diagnosis	categorical

Table 12.2 shows the results for this task using a regularized linear regression. It was validated that gains achieved by the proposed partial transfer framework were higher than the obtained by the fully observable transfer recently used in the literature. In most cases, the best results were obtained by the α-controlled sign regularization approach.

Table 12.2: sAUC obtained by the TL approaches on the risk prediction task with multiple screening strategies: Hinselmann (H), Schiller (S), Cytology (C) and Biopsy (B). Performance is measured in terms of Rooted Mean Squared Error (RMSE).

Source	**Target**	**Diff**	**Sign**	**α-Sign**	**Source**	**Target**	**Diff**	**Sign**	**α-Sign**
H	S	66.09	66.02	**68.96**	C	H	35.05	34.51	**35.11**
H	C	19.51	24.67	**37.12**	C	S	55.45	53.97	**55.81**
H	B	54.70	52.39	**54.96**	C	B	47.37	47.40	**47.54**
S	H	38.72	36.44	**38.74**	B	H	47.99	47.39	**48.80**
S	C	33.55	34.21	**39.90**	B	S	64.10	61.89	**66.66**
S	B	**45.48**	42.19	45.34	B	C	28.18	34.14	**43.69**

12.2.2 Quality Assessment

Choosing frames with good quality to perform the screening is an important step to improve physician's effectiveness. However, several challenges arise when defining the quality in this context. Thus, quality becomes a subjective concept subject to human preferences. In this work we consider a binary annotation scheme (e.g., good and bad quality) to simplify the presentation of the proposed framework. However, in the future, we will consider ordinal scales (e.g., poor, fair, good, excellent) and pairwise relative preferences (e.g., the image A is better than the image B). The following semantic medical features were considered:

- Image area occupied by each anatomical body part (cervix, external os, and vaginal walls) and occluding objects (speculum and other artifacts).
- The area of each region occluded by artifacts or by SpR.
- The maximum area difference between the four cervix quadrants.
- Fitness goodness of the cervix to a given geometric model: convex hull, bounding box, circle, and ellipse.
- Distance between the image center and the cervix centroid/external os.
- Mean and standard deviation of each RGB and HSV channel in the cervix area and in the entire image.

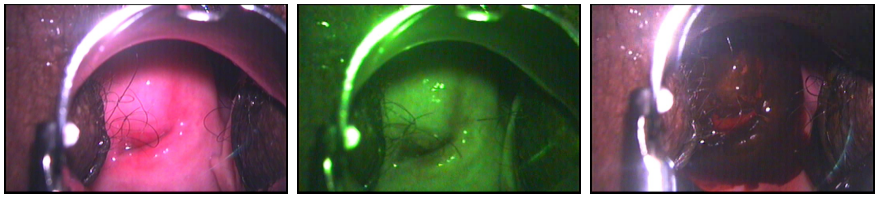


Figure 12.1: Colposcopy modalities. From left to right: Hinselmann, Green light and Schiller.

In a joint collaboration with *Hospital Universitario de Caracas*, a dataset with annotations from 6 experts on about 100 cervigrams per modality (see Figure 12.1) was collected [94]. In the experimental evaluation, each ROI was manually segmented by an expert to simplify the comparison of the TL approaches.

Table 12.3: sAUC obtained by the TL approaches on the quality prediction task with several colposcopic modalities: Hinselmann (H), Green (G) and Schiller (S). Performance is measured in terms of accuracy.

Source	**Target**	**Diff**	**Sign**	α-**Sign**	**Source**	**Target**	**Diff**	**Sign**	α-**Sign**
H	G	53.31	**54.14**	53.83	H	S	**47.82**	46.58	45.73
G	H	64.13	68.05	**68.30**	G	S	47.07	47.98	**48.15**
S	H	**63.73**	62.67	61.02	S	G	47.16	**49.28**	48.54

Table 12.3 shows the results for the binary classification of the subjective image quality using SVM. The target labels are assigned using the mode of the annotations given by the physicians. Contrary to the linear regression case, the version with $\alpha = 1$ obtained better results than the α-Sign approach. This can be explained by the fact that each modality has a few annotated instances per expert (about 100), turning it difficult to correctly estimate the α parameter.

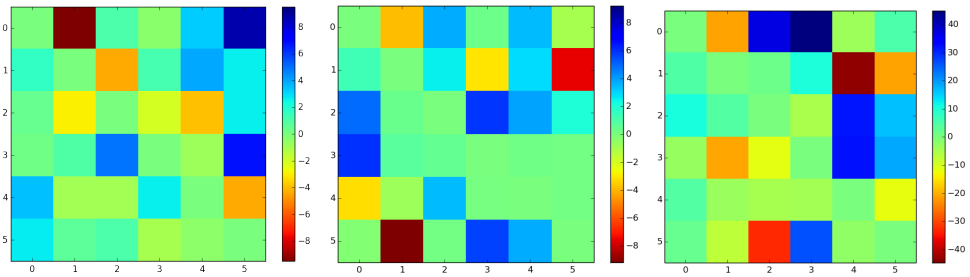


Figure 12.2: Heatmap of the transfer gain obtained by the α-Sign regularizer when compared to the state-of-the-art regularizer. Transfer is done from a given expert's preferences (row) to another expert's preferences (column) between the same modality. The modalities are, from left to right: Hinselmann, Green light and Schiller.

Figure 12.2 shows the gains obtained by the α-Sign version of the regularizer when compared with the state-of-the-art approach on a multi-expert setting. Here, source and target tasks represent different annotators' preferences (i.e., transferring from the i-th expert in the row to the j-th expert in the column). Analogously to previous experiments, the proposed transfer with partial observability obtained the best results in most cases. Schiller was the modality with highest gains. However, it was also the most unstable, being also the one with lowest gains in some cases. Using partial transfer schemes, some experts reflected poor performance as source in some modalities (e.g., expert 2 in Hinselmann) while behaving as good sources in other modalities (e.g., expert 2 in Green). Moreover, since the partial model observability is a weak prior over the model space, the set of models

that achieves an optimal regularization value is infinite, inducing a non-symmetric gain matrix.

12.3 Conclusions

In this work, we validated the performance of the TL framework proposed in Chapter 5 on two cervical cancer screening tasks. First, we validate the impact of transferring the coefficient sign between models aiming to predict the outcome of several cervical cancer screening modalities. Also, we validate the impact of using TL to learn subjective QA models on scenarios with multiple experts and modalities.

In all cases, we validated that applying TL with partial observability induces larger gains than transferring the entire model instantiation. Since diverse applications can found similar properties at a high semantic level, applying a transfer with partial observability of high-level properties of the model structure increases the chances of observing gains when the source and target tasks are not closely related. This work suggests that the analysis of how models encode high-level properties of the domain may improve transfer performance.

Chapter 13

A Deep Learning Approach for the Forensic Evaluation of Sexual Assault

This chapter was published in [93]:

- Kelwin Fernandes, Jaime S. Cardoso, and Birgitte Schmidt Astrup. A deep learning approach for the forensic evaluation of sexual assault. In *Pattern Analysis and Applications*. Springer, 2018

A preliminary version of this work was published in [92]:

- Kelwin Fernandes, Jaime S. Cardoso, and Birgitte Schmidt Astrup. Automated detection and categorization of genital injuries using digital colposcopy. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 251–258. Springer, 2017

Despite the existence of patterns able to discriminate between consensual and non-consensual intercourse, the relevance of genital lesions in the corroboration of a legal rape complaint is currently under debate in many countries. The testimony of the physicians when assessing these lesions has been questioned in court due to several factors (e.g., a lack of comprehensive knowledge of lesions, wide spectrum of background area, among others). Thereby, it is relevant to provide automated tools to support the decision process in an objective manner. In this work, we evaluate the performance of state-of-the-art deep learning architectures for the forensic assessment of sexual assault. We propose a deep architecture and learning strategy to tackle the class imbalance on deep learning using ranking. The proposed methodologies achieved the best results when compared with handcrafted feature engineering and with other deep architectures.

13.1 Introduction

The relevance of genital lesions in the corroboration of a legal rape complaint is currently under debate in many countries [17, 19, 20]. Since genital lesions are frequent in both, consensual and non-consensual intercourse [17, 18], the existence of a pattern of genital injury able to discriminate trauma seen in rape cases and trauma seen following consensual sexual intercourse has been a matter of study in the past [20]. Discriminative patterns were analyzed by several authors [20, 312]. Slaughter et al. [312] suggested that multiple genital lesions at multiple locations are frequent in rape victims, while single lesions in the posterior forchette are predominant in consensual sexual intercourse. Astrup et al. [20] suggested a higher frequency of abrasions, hematomas and multiple lesions in rape cases. Also, Astrup et al. [20] confirmed a higher frequency of lesions in locations other than the 6 o'clock position and the presence of larger and more complex lesions in non-consensual cases.

Despite the existence of such patterns have been validated by several studies, the debate continues. Legal experts suggest the lack of comprehensive knowledge of lesions sustained during consensual sexual intercourse as a key problem [17]. Moreover, the expert responsible for conducting such evaluations as well as the physical analysis of sexual assault victims itself differs around the world [19]. For instance, in the US most examinations are done by specially trained nurses, while in many European countries the examinations are performed by gynecologists [263]. Other countries like Denmark delegate this responsibility to forensic pathologists [19]. Given the wide spectrum of background knowledge of experts and the low inter-evaluator agreement [18], the expert testimony given by these professionals in cases of genital lesions in sexual assaults has been questioned in court in several countries [19].

In previous work, we proposed a preliminary framework for the automated forensic evaluation of sexual assault on digital colposcopies based on image processing and ML techniques [92]. This framework tackled several problems that are relevant from an automation point of view, from the acquisition modality identification and the semantic object segmentation to the lesion categorization and final forensic assessment. While previous attempts to use ML for automated analysis of digital colposcopies have been addressed in the past [94, 95, 158], the proposed project is the first attempt to tackle the forensic evaluation of sexual assault using colposcopic images from a computational perspective. Building an objective data-driven system may increase the reliability of genital trauma findings in legal rape complaints. Also, it would serve as a common basis for the interaction between medical and legal teams. The main contributions of the current chapter are:

- We extend the framework for the forensic evaluation of sexual assault by including the spatial localization of the genital injuries.

- We validate the performance of state-of-the-art deep architectures for segmentation and classification to tackle all the subtasks in the framework.

- We propose learning methodologies to address the problem of data imbalance in segmentation and classification tasks using deep learning, including a deep architecture that extends the ranking methodology proposed in [58] to deep learning.

- We propose a visualization mechanism to understand the image regions with the strongest impact on the final decision.

The rest of the chapter is organized as follows. Section 13.2 summarizes some concepts from the forensic community that are relevant to the comprehensive understanding of the proposed pipeline. Aiming to facilitate the automation of the decision process and to remove external sources of bias, we divided the entire pipeline into several (minor) subtasks which are described in section 13.3.1. In section 13.3.2, we describe the deep learning architectures and strategies that are used in this work for each type of task. Finally, section 13.4 presents the results obtained in the experimental assessment of the proposed methodology and section 13.6 summarizes the findings of this work.

13.2 Basic Concepts and Definitions

In this section, we describe forensic concepts that are fundamental in this work. Specifically, the type of lesions of interest, and the investigative techniques used in their detection. Further details about these concepts can be found in the medical literature [16].

13.2.1 Investigative Techniques

The usual methodologies for the detection of genital injuries cover the naked eye inspection, the colposcope and inspection after application of toluidine blue dye (see Figure 13.1) [16]. In this work, we merely include the two latter which allow automation.

(a) Colposcope with light source (b) Colposcope (c) Toluidine blue dye

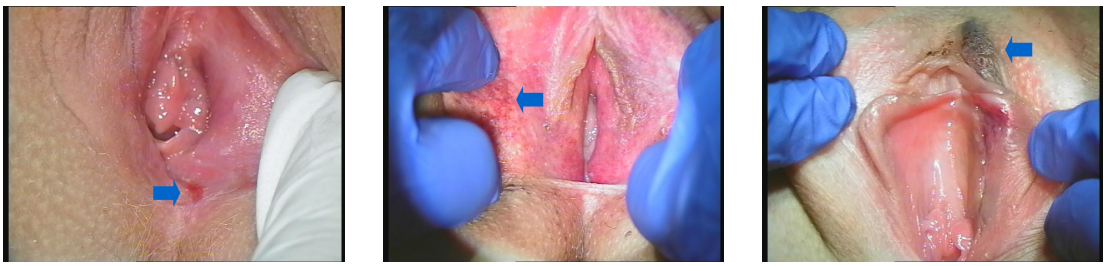
Figure 13.1: Examples of images from several acquisition techniques

Colposcope: The investigator inspects the external genitalia and afterwards the vagina and cervix using digital colposcope. A colposcope is a binocular instrument that magnifies and illuminates the inspected area.

Toluidine Blue Dye: After inspection, a blue dye is applied to the genital mucous membranes and then wiped off. Toluidine blue stains exposed cellular nuclei but not intact mucosa, thus enhancing areas of surface disruption.

13.2.2 Genital Injuries

The European and Australian categorization of genital injuries is used in this work, which comprises laceration, abrasion and hematoma (see Figure 13.2) [16].



(a) Laceration (b) Abrasion (c) Hematoma

Figure 13.2: Examples of genital injury on digital colposcopy. Injuries are marked with blue arrows.

Laceration: Discontinuity of epidermis and dermis. Caused by a blunt force such as tearing, crushing, or overstretching.

Abrasion: traumatic exposure of lower epidermis or upper dermis. Most often caused by lateral rubbing or sliding against the skin tangentially. The outermost layer of skin is scraped away from the deeper layers.

Contusion/Hematoma/Bruise: Traumatic extravasation of blood in tissues below an intact epidermis. Caused by blunt force.

13.3 Proposed Methodology

In this section, we describe the sub-problems that were identified as relevant for the automation of the forensic evaluation of sexual assault and the proposed deep architectures and learning strategies for learning such problems.

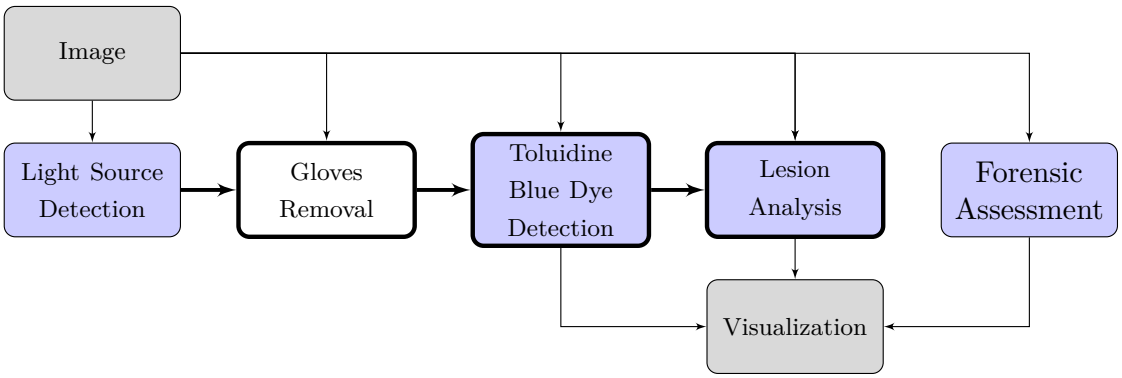


Figure 13.3: Pipeline of the proposed system for the automatic forensic assessment of sexual assault. Thick arrows represent filtering of the Regions of Interest. Blocks highlighted with thick borders were modeled as image segmentation tasks, blocks with blue background were modeled as classification tasks (e.g. absence/presence, type, consensual/rape).

13.3.1 Pipeline

In order to facilitate the final analysis of these images, we subdivided the framework into several subtasks that cover specific parts of the overall forensic assessment. The pipeline can be broadly described by five predictive tasks: light source detection; segmentation of gloves; detection and segmentation of the toluidine blue dye stained regions; detection, classification and segmentation of lesions; and forensic assessment (i.e., discrimination of consensual and non-consensual intercourse). Figure 13.3 illustrates the subtasks involved in the general framework. Blocks highlighted with thick borders were modeled as image segmentation tasks, blocks with blue background were modeled as classification tasks (e.g., absence/presence, type, consensual/rape).

In a preliminary work [92], we proposed supervised learning techniques to handle each one of these subtasks using handcrafted features. Our preliminary approach is mainly based on color quantization, superpixels and image descriptors (e.g., SIFT and SURF). Further details about the methods can be seen in [92]. In this work, we improve those results by considering state-of-the-art deep learning architectures and learning paradigms that tackle the high data diversity and the imbalance class distribution.

13.3.1.1 Light Source Detection

The first step in the proposed pipeline is the detection of the light source. Being able to recognize the use of an external light source pointing to a subset of the image reduces the search space for relevant traits in the image. So, after identifying the presence of such artifact, the ROI is narrowed to the illuminated region. Since the presence of the artificial light source is spatially located in approximately the same round area (see 13.1a), we simplified the task of segmenting the lighted-unlighted areas to a global binary classification task.

13.3.1.2 Gloves Removal and Toluidine Blue Dye Segmentation

Secondly, to remove irrelevant objects that should not impose any kind of bias on the model decision, we segment the gloves in the image to remove them in later stages of the pipeline. Also, we propose to segment the stained regions with toluine blue dye to use such spatial information in later stages. While we are not considering modality-dependent systems due to the low number of images in our dataset, it is relevant to identify the acquisition conditions for future refinements of the proposed system.

13.3.1.3 Lesion Analysis and Forensic Assessment

The final steps of the proposed system to support the expert's decision are the detection and categorization of lesions and the final assessment (i.e., consensual intercourse vs. rape). In this sense, we divided the problem into four main tasks. The lesion analysis task was addressed in three steps: 1) detection of presence of lesions in the image (binary classification task); 2) spatial location of the lesions (segmentation task); and 3) classification of lesions according to the categorization referred in subsection 13.2.2 (multiclass classification).

The final predictive task is the classification of the case as consensual or non-consensual (binary classification).

13.3.2 Deep Architectures and Learning Strategies

In this section we describe the deep architectures used for segmentation (section 13.3.2.1) and classification (section 13.3.2.2) tasks. Also, we propose learning strategies to tackle the problem of data scarcity and imbalance nature of our tasks.

13.3.2.1 Deep Architectures for Segmentation

In previous work [92], we studied the performance of encoder-decoder networks for segmentation. While the results show that these architectures are able to spatially locate the objects of interest, they were not able to discover the true object boundaries in a sharp manner, as done by traditional handcrafted techniques (see Figure 13.4).

Thereby, we considered the state-of-the-art U-net architecture [283] for image segmentation. The U-net architecture is a fully convolutional segmentation network [217] able to learn accurate segmentations with small datasets. This network improves the granularity of the final segmentation by using bypass-merge layers between feature maps at the same level of the encoding-decoding components.

In the original papers, the authors of the U-net architecture use the pixelwise categorical cross-entropy loss function for optimizing the network parameters. However, this loss function does not consider the IR between the objects in the image. Thereby, given that we are interested in segmenting objects that are considerably small such as the genital

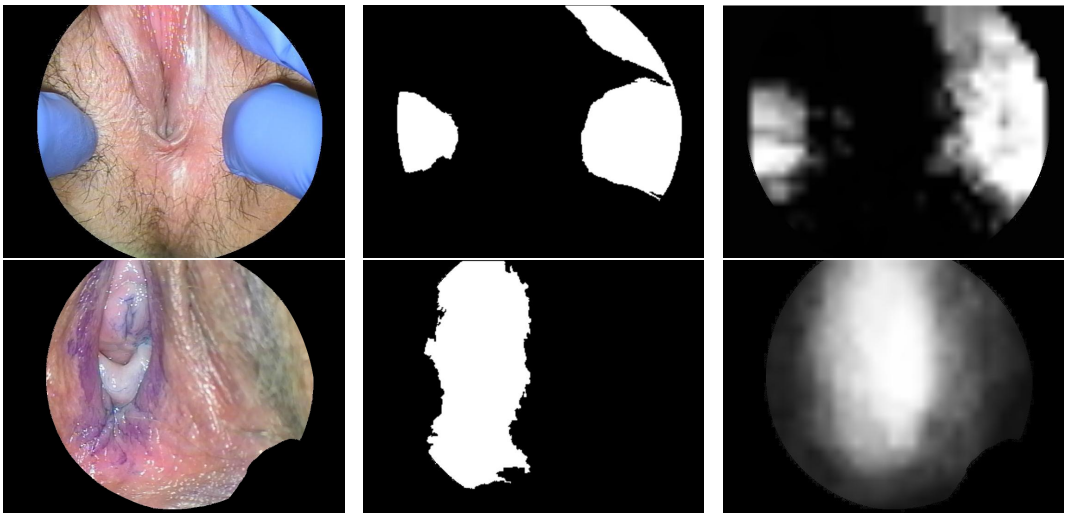


Figure 13.4: Results obtained by the strategies to segment gloves (**top**) and toluidine blue dye (**bottom**). **Left:** original image. **Middle:** Handcrafted features. **Right:** Encoder-Decoder network.

lesions, we introduce an optimization process based on the fuzzy Sørense-Dice coefficient. The hard version of this function is commonly known in classification as the F1 score, and is widely used when the performance of classification techniques is assessed on unbalanced settings [58]. Eq. 13.1 formalizes the fuzzy Dice coefficient between a ground-truth mask M and a probabilistic binary segmentation P, where $*$ is the elementwise multiplication (resembling the fuzzy set intersection).

$$fuzzyDice(M,P) = \frac{2 \parallel M * P \parallel_1}{\parallel M \parallel_1 + \parallel P \parallel_1} \tag{13.1}$$

13.3.2.2 Deep Architectures for Classification

For the classification task, we used the state-of-the-art classification networks ResNet-50 [146] and Inception-v3 [320]. In both cases, we include a final sequence of interleaved dense and dropout layers for learning robust subtask-specific features. In the end, a *softmax layer* is appended to obtain the final classifier. The depth, width, and activation functions are defined using a validation set. Figure 13.5 illustrates the architecture of the proposed network.

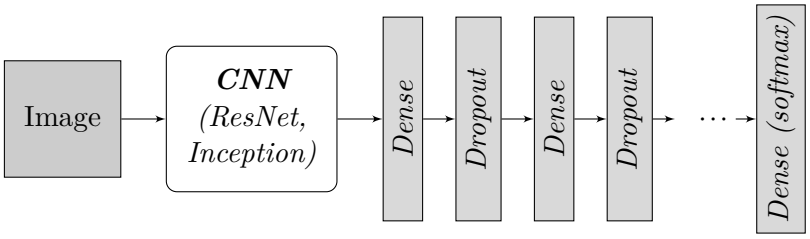


Figure 13.5: Deep Network for classification.

Given that our training data is scarce, we applied the initialization-based transfer learning technique by using networks pretrained on the ImageNet dataset. While in our previous work we constrained the learning process to fine-tuning the final dense layers of the networks, in this work we consider a two-stage process. First, we freeze the original parameters from the pre-trained network while we optimize the additional dense layers. Then, we optimize the entire network.

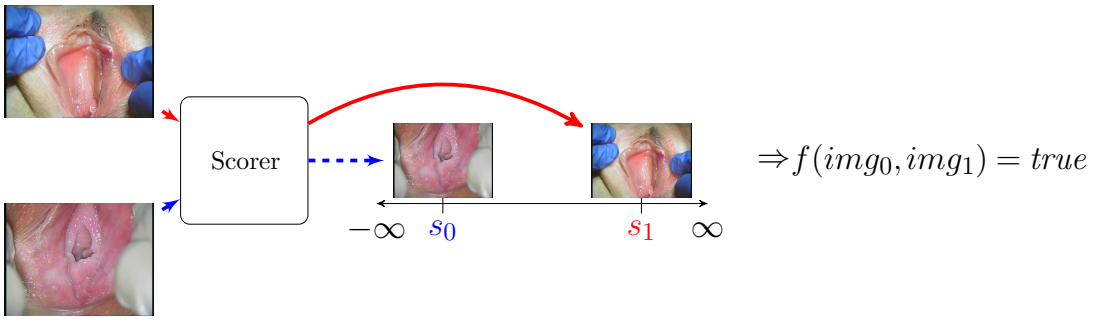
13.3.2.3 Learning in Imbalance Settings

In most medical applications, the class distribution is inherently unbalanced, being the class of interest usually poorly represented while most observations are normal cases. For instance, in most screening programs, the number of patients with the disease is extremely low when compared with the entire population size. The same property holds in our case, where the number of rape victims and the number of patients with genital injuries is relatively low when compared with the total population. In this setting, traditional ML techniques tend to converge to a naive classifier that always predicts the majority class (e.g., healthy, consensual sex, etc.) [58]. A widely used technique to tackle the imbalance problem is using weighted loss functions that tend to balance the contribution of each class in the learning process. These weights are represented by a cost matrix. For instance, in our case, we should define the cost of classifying a rape victim as a consensual case and vice versa. The same applies to missdetecting lesions which can be used as evidence in court. Intuitively, these approaches are strongly avoided by the medical community since it is a highly sensitive topic, being extremely difficult to estimate such costs, involving ethical concerns. An alternative approach would be to train one-class models using observations from one of the classes and to model the problem as an outlier detection one. However, this second approach discards a vast amount of data which may improve discriminability.

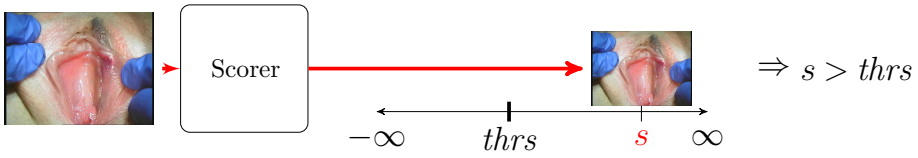
A third approach based on ranking was presented in Chapter 3, outperforming state-of-the-art methods. The idea was illustrated in Figure 3.1 and can be summarized as follows. The binary unbalanced classification task is transformed into a balanced ranking problem by considering all comparisons of pairs of objects from different classes. The ranker is trained to classify which observation in a pair should be ranked first. Since all pairs are compared, the model is balanced and each class contributes with the same impact on the training loss.

The training and inference stages are illustrated in Figure 13.6. During training, both high-dimensional images are independently projected to a one-dimensional space by using a scorer and comparison is done on the linear space. During inference, a single image is used and the right class is given by using a thresholding operator on the space of scores.

When dealing with a linear ranker such as the RankSVM model, the pairwise ranking function is trained on the space of the feature difference between observations $f(a,b) = \omega \cdot (a-b) > 0$, where the scoring function arises from the linear separability of the decision function into $\omega \cdot a > \omega \cdot b$. While the decision function of a deep architecture D on the



(a) Training pairwise scoring rankers.



(b) Inference with pointwise scorer obtained from pairwise scoring rankers.

Figure 13.6: Illustration of training and inference with classifiers obtained with pairwise scoring ranking

difference space $D(a-b)$ is not linearly separable, if we consider a latent feature space $E(\cdot)$ learned by the DNN, we can learn a linear ranker in the form $\omega \cdot (E(a) - E(b)) > 0$. The learning process of the ranking coefficients ω and the embedding function E can be learned in a joint fashion by considering the architecture defined in Figure 13.7, where E is a DNN.

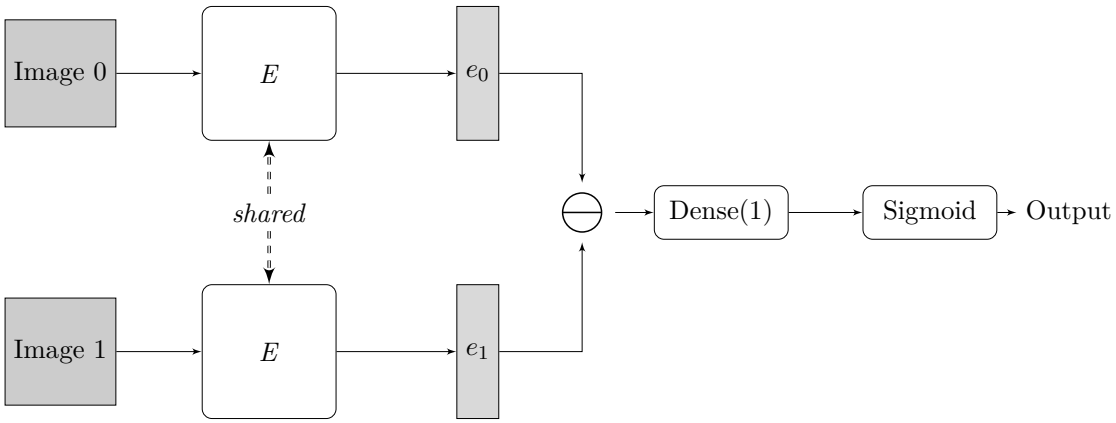


Figure 13.7: Deep Ranking Network.

In our case, we used the same pre-trained deep architectures that were aforementioned described by ignoring the final *softmax* layer. Namely, our scorer is a deep CNN (e.g., ResNet, Inception) that projects images to a real-valued score. At training, images from both classes are sampled and the tuples with observations from different classes are built. So, for each pair of images i_0 and i_1 with labels $l_0 = false$, $l_1 = true$, the training pairs $\langle (i_0, i_1),\ true \rangle$ and $\langle (i_1, i_0),\ false \rangle$ are fed to the ranker. It is important to note that each

batch is balanced since the ranker receives both comparisons (i_0, i_1) and (i_1, i_0).

Training models with a class distribution that differs from the one observed during inference may require additional post-processing [125]. Thus, the decision threshold that maps scores to classes should be fine-tuned on the final evaluation setting. As in Chapter 3, we choose the thresholds that maximize the F1-score on an independent validation set. The F1-score is widely applied on settings with highly unbalanced distribution to assess the performance of classifiers, being more robust than other metrics such as accuracy.

13.3.2.4 Regularization

While handcrafted features can induce invariance to certain properties of interest such as illumination, rotation, translation, and scale, DNN require a large amount of data to infer such properties implicitly. Furthermore, given the high flexibility of the decision function induced by DNN, they have the capability of overfitting to the training data, which is even more evident in small datasets. Thereby, besides traditional L_2 regularization of the coefficients traditionally used in ML, we considered other regularization mechanisms that have been widely used in the DL community.

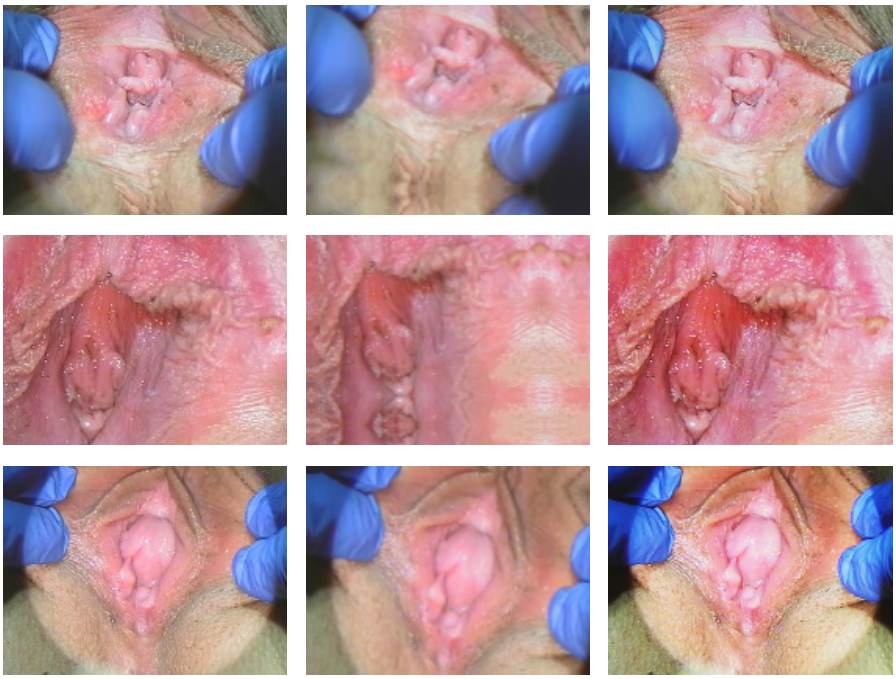
The first technique that we use in all our networks is early stopping. This technique consists in stopping the optimization technique without exhausting the allowed resources (i.e., reaching the global optimum or using a maximum number of iterations). This procedure has been considered a regularization technique that allows to solve the bias-variance trade-off [284, 353]. In our case, we stop the learning process after N iterations without improving the model performance on the validation loss.

Also, we used dropout, a technique that can be understood as randomly turning off hidden units in the network. This mechanism leads to learning redundant latent representation that improves the robustness of the network. Furthermore, dropout can be understood as a regularization mechanism by implicitly combining an ensemble of structurally different networks [319].

Finally, we augmented the training set by applying random transformations to the training data at each iteration. The set of transformations that we use cover:

- Horizontal flips of the images.
- Affine transformations: random rotations, translations, shearing and scaling.
- Stretching of the color histogram.

Border regions were filled by reflecting the information from the original images. Figure 13.8 illustrates the effect of these transformations on sample images. These transformations (exception for color transformations) were also applied to the ground-truth masks in the segmentation tasks to ensure consistency between the randomly transformed images and the ground-truth.



(a) Input images (b) Affine transformation (c) Contrast stretching

Figure 13.8: Data augmentation

At each iteration, all images are passed through the transformations with random perturbations (e.g., rotation degree, scaling rate, etc.). This prevents the network from memorizing information from specific pixels in the image by learning on a virtually infinite training set. The dataset is augmented proportionally to the number of training iterations.

13.4 Experiments

In the experimental assessment of the proposed methods a dataset with 394 images collected by the *Southern Denmark Sexual Assault Referral Centre* was used (78 images from non-consensual intercourse and 316 from consensual intercourse). For further details about the acquisition process refer to the source [18]. To keep the validation protocol consistent with preliminary work [92], we use the same random training, validation and test partitions from [92]. The partitions follow a standard 60-20-20 distribution. The splits were done randomly in a stratified fashion by keeping the distribution of images with and without artificial light, with each color of gloves and from consensual and non-consensual cases was kept constant. Since the frequency of hematoma in our dataset was very limited, abrasion and hematoma cases were combined as a single class in the categorization of lesions. The performance of each classification strategy was measured using accuracy (Acc), F1-score (F1) and macro-averaged area under the ROC curve (AUC). Table 13.1 shows the class distribution for each classification subtask. For the segmentation tasks,

the fuzzy Dice coefficient (see Eq. (13.1)) was used to validate the performance of all the proposed techniques.

13.4.1 Hyper-parameter fine-tuning

Table 13.1: Class distribution per task.

Sub-task	Classes	Class distribution
Light Source	without-with	84.81 - 15.19
Toluidine blue dye	without-with	44.30 - 55.70
Lesion detection (binary)	without-with	79.75 - 20.25
Lesion categorization	none-laceration-others	74.68 - 16.46 - 8.86
Consensual/Rape	consensual-rape	79.75 - 20.25

Table 13.2: Hyper-parameter configuration for the DNN.

(a) Segmentation Tasks

Parameter	Range
Image resolution	$128 \times 128, 256 \times 256$
Depth	$2, 3, 4, 5$
Convolution size	$3 \times 3, 5 \times 5$

(b) Classification Tasks

Parameter	Range
Network	Inception v3, ResNet 50
Dense Activations	ReLU, logistic
Dense width	64, 128, 256
Dense depth	0, 1, 2

The best configuration for each network was chosen using a grid search strategy on the parameter space defined in the Tables 13.2a and 13.2b. For the convolutional layers, we used Rectifier Linear Units [123]. We used a maximum number of 500 and 100 iterations for the segmentation and classification networks respectively. The batch size was set to 16. We used the AdaDelta optimization strategy for fitting all the architectures [359].

For the U-net strategy, we used two consecutive convolutional layers for each pooling layer as suggested by the authors [283]. In this sense, the depth parameter in Table 13.2a refers to the number of blocks with two convolutional layers and one pooling layer in the encoding part of the network, being symmetric for the decoding part.

The early stopping parameter was empirically set to 20% of the maximum number of iterations (i.e., 100 and 20 for segmentation and classification tasks respectively) and the best model was chosen using the validation loss.

13.4.2 Results

Tables 13.3 and 13.4 show the results for the segmentation and classification subtasks respectively on the test set. We compare the proposed deep learning strategies with the hand-crafted methodology proposed in [92]. The hand-crafted pipeline for classification

relies on features from color and texture information of superpixels in the image and Bag-of-Words of SIFT and SURF descriptors. The methodology for segmentation uses segmentation-by-classification of superpixels with features extracted from texture, color, and shape. The underlying models used for both kind of tasks cover ensembles (e.g., RF, AdaBoost, Gradient Boosting), SVM and LR. Further details about the hand-crafted methodology are presented in [92].

Table 13.3: Summary of the results for the segmentation subtasks. **Trad** denotes the methodologies proposed in [92] using traditional CV and ML techniques. **Deep** denotes the architectures using U-net. Performance is measured in terms of fuzzy Dice coefficient.

Sub-task	**Trad**	**Deep**
Gloves	84.99	**87.63**
Toluidine blue dye	84.30	**88.03**
Lesion	14.62	**79.75**

Table 13.4: Summary of the results for the classification subtasks. **Trad** denotes the methodologies proposed in [92] using traditional CV and ML techniques. **Class** and **Rank** refers to the base networks trained as classifiers and rankers respectively.

Sub-task	**Accuracy**			**F1-score**			**ROC AUC**		
	Trad	**Deep**		**Trad**	**Deep**		**Trad**	**Deep**	
		Class	**Rank**		**Class**	**Rank**		**Class**	**Rank**
Light Source	**100.0**	94.9	**100.0**	**100.0**	93.3	**100.0**	**100.0**	**100.0**	**100.0**
Toluidine blue dye	93.7	**100.0**	98.7	96.6	**100.0**	98.9	99.2	**100.0**	**100.0**
Lesion detection (binary)	68.4	**78.5**	77.2	25.8	41.0	**52.9**	56.1	76.0	**78.1**
Lesion categorization	72.2	**76.0**	–	27.7	**51.9**	–	61.7	**74.9**	–
Consensual/Rape	84.8	**87.3**	84.8	55.2	72.7	**77.4**	74.6	93.5	**95.3**

The results are overall satisfactory, being able to provide positive predictive results for all the proposed subtasks.

While in previous work [92], poor performance on the segmentation subtasks using deep networks was observed, the combination of the Dice coefficient-based loss function and the U-net architecture allowed to surpass the performance of traditional techniques in all cases. As can be observed in Figure 13.9, even with the U-net architecture, the cross-entropy loss tends to produce soft segmentations that rarely assign high probabilities to the injured tissues. For the segmentation of gloves, the best network had the simplest parametrization, with depth 2 and kernels with size 3×3. On the other hand, for the segmentation of blue stained regions and lesions, the best-performing networks had kernels with size 5×5. In terms of depth, the segmentation of blue stained regions favored parametrizations with high depth (5) while the segmentation of lesions performed the best with low-depth architectures (2). The intuition behind this idea is the small size of lesions, which may require a low loss of resolution, intrinsic of very deep networks. Moreover, the ground-truth masks of the blue stained areas are very coarse and are formed by a single large

(a) Input images

(b) Cross-entropy loss

(c) f-Dice

Figure 13.9: Comparison of the U-net segmentations with cross-entropy loss and fuzzy Dice coefficient

.

object in most cases, which may be obtained with deep networks. In contrast, lesions masks have high levels of details and several small lesions may be observed in a single image.

As observed in related areas, deep learning strategies performed better than traditional pipelines in most classification tasks in terms of accuracy, F1, and AUC. As was validated in the literature [58], using the proposed classification as ranking strategy induced classifiers with high ROC AUC and F1. However, the final classification performance is worse than with traditional strategies. Thereby, further research should be conducted on the threshold decision process. Despite the limited size of the database, models with high expressiveness such as the DeepRank strategy achieved a high test/validation performance ratio of 0.91 and 0.89 in terms of F1-score and ROC AUC respectively on the detection of lesions, which was the task with the worst performance. On the other tasks, the test/validation ratio is close to one. Thereby, models do not seem to be overfitting to the training set.

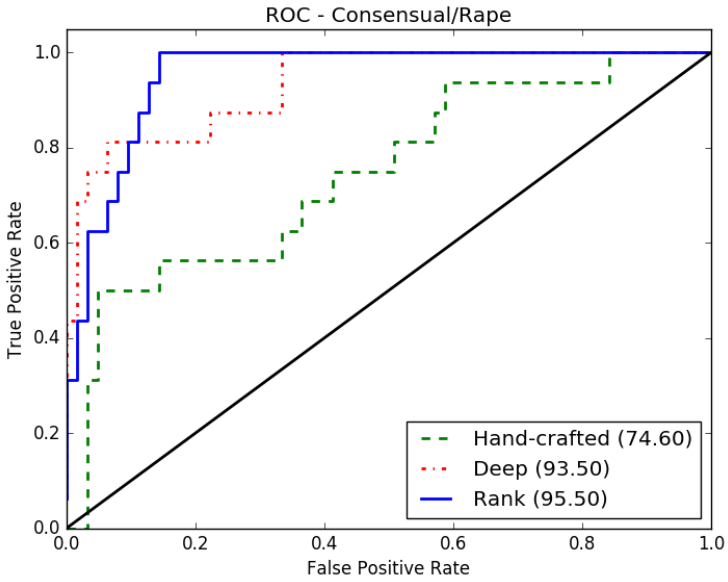


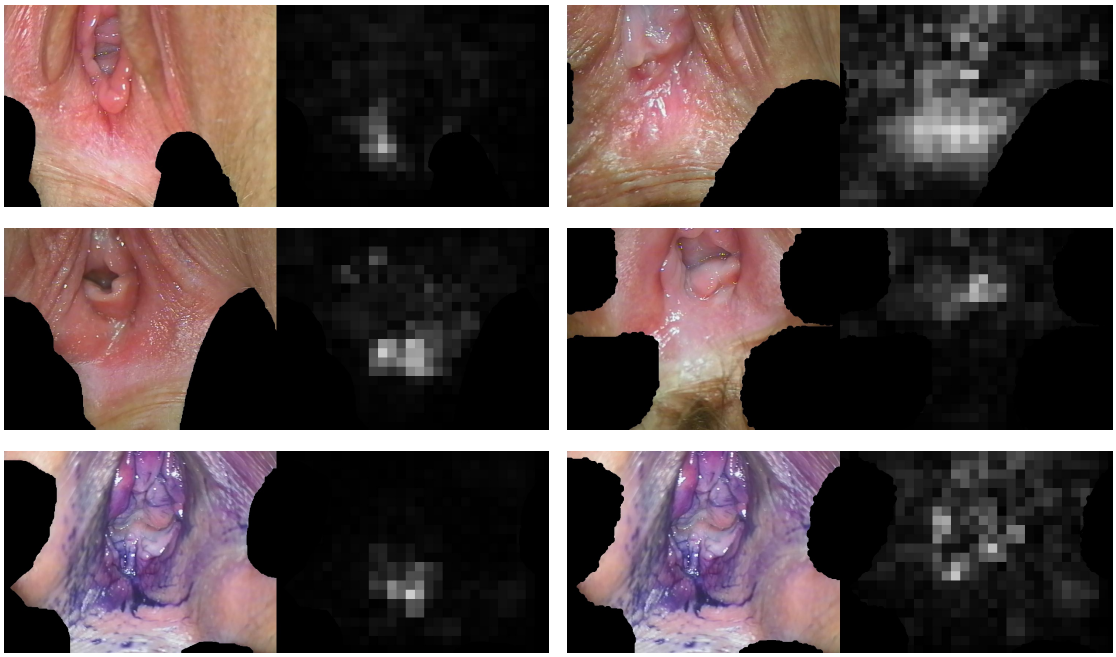
Figure 13.10: Receiver operating characteristic curve (ROC curve) of the models for forensic assessment.

Figure 13.10 shows the ROC curves for the final task of the proposed pipeline (consensual/rape). This task is the most sensitive since it is the most difficult to corroborate by the human experts. As can be seen in the graph, the proposed deep learning strategies dominate the performance of hand-crafted methodologies. Since both deep strategies have the same underlying predictive model, their overall performance is similar in terms of area. However, the ranking-based model is able to achieve optimal performance in terms of true positives with a low number of false positives. This kind of curves is preferred in this context since it can be corrected by further evidence. This preference is observed in most medical applications, where screening strategies with high sensitivity are preferred since it would motivate further examination [59].

The low performance of the lesion detection and categorization tasks when compared to the consensual/non-consensual problem may be the consequence of two reasons. On the one hand, the models may be learning external patterns during the acquisition process of rape victims, in which case further validation is required to ensure that both groups were handled indistinguishably. On the other hand, there might be potential patterns that haven't been quantified by the medical community. The latter would be more interesting, motivating further research on understanding the decisions of the networks.

13.5 Deep Visualization

As was mentioned in Section 13.1, despite the existence of patterns able to discriminate between consensual and non-consensual intercourse, the relevance of the specialist judgment



(a) Lesions. (b) Forensic evaluation.

Figure 13.11: Visualization of the most relevant regions for the binary classification of lesions and forensic evaluation.

is often questioned in court. This would be exacerbated by the presence of computer-based medical decision support systems despite being strongly data-driven and free of any type of cultural bias. This matter is very relevant when dealing with black box predictive systems such as DNN which decisions are difficult to interpret.

In this sense, DSS should aim to provide an explanatory decision instead of the single predictive outcome (e.g., lesions vs. non-lesion, consensual vs. rape). Several strategies have been proposed to analyze the decision process of DNN, from low-level visualization of the convolutional layers [83, 198, 372] to high-level analysis of the sensitivity to certain occlusions [360].

In this work, we use a technique based on the occlusion sensitivity proposed in [360]. The idea is to occlude with a rectangle portions on the image and to study the impact on the network outcome. While in [360], the occlusion is done with a gray rectangle, we used the average color of the image to resemble the background regions that were occluded in the preliminary steps of the pipeline (e.g., gloves, light effect, etc.). Then, the probability differences are accumulated through a sliding window with varying resolution. The final result is an attention map that reflects the regions with the most relevant activations in the decision process. We used occluding blocks of size 50×50 and 100×100 with stride equals half of the block side (the original image resolution is 768×576). The total number of occlusions is about 800. Final probabilities are normalized to the range $0 - 1$ to improve visualization. While this idea does not reflect *why* did the network make that decision;

it illustrates **what** regions are relevant in the decision. This may improve the decision process of the human expert by indicating relevant points in the image that suggest a given decision.

Figure 13.11 shows the visualization results for some sample images on the classification tasks of lesion detection and forensic evaluation. As can be seen in the images, the network focuses on regions near the vagina and its surrounding areas. The high relevance of the posterior forchette and surrounding regions is an interesting property, which has been suggested as a strong decision factor by the medical community [20, 312].

13.6 Conclusions

Despite the existence of patterns able to discriminate between consensual and non-consensual intercourse have been proved, the relevance of genital lesions in the corroboration of a legal rape complaint is currently under debate in many countries. Being the lack of comprehensive knowledge of lesions a driving factor in the acceptance of this type of evidence in courts, it is fundamental to provide objective methods to support the expert's decision.

In this work, we proposed a framework that covers the preliminary steps in the automated detection of genital injuries and the forensic assessment on digital colposcopies using CV and ML. We proposed a method to visualize the areas of interest in the automated decision process to improve the support offered to the specialists. The proposed system aims to objectify the forensic evaluation of sexual assault by using data-driven ML models. Therefore, we covered aspects from the detection of genital injuries, to the forensic assessment, to the explanation of the decision using visualization techniques. From a technical point of view, we compared the performance of traditional pipelines with handcrafted features and deep learning approaches in several subtasks obtaining the best results with deep learning strategies in all tasks. Also, we proposed extensions for the learning process of segmentation and classification networks to handle imbalance settings, either in terms of the object size in segmentation tasks or class distribution in classification tasks.

As future work, we intend to explore alternative mechanisms for the visualization and explanation of the network's predictions. Namely, enhancing the input image in order to maximize the network confidence by means of gradient backpropagation as typically done in the construction of adversarial examples [126, 258, 259]. Also, we plan to explore ideas based on metric learning to provide similar historical examples to the human evaluator when inspecting a new patient. In this work, we used initialization-based TL from networks pre-trained on the ImageNet dataset. While such initialization has proved to be useful in several applications, it is not clear whether such features are relevant for the initialization of this kind of images. In this sense, we plan to work on transferring knowledge from networks pre-trained on close domains such as nudity and pornography detection. These tasks have gained some traction which databases that are openly available nowadays [11, 21, 220].

Last, we will consider multitask learning settings, where a common feature representation is jointly learned for all tasks.

Part III

Conclusions

Chapter 14

Conclusions

Granting universal access to cervical cancer screening is a major challenge nowadays, concentrating the attention of key players in the area, from companies such as Intel and MobileODT [237] to global organizations such as the WHO. In a speech entitled "Grand challenges for the next decade in global health policy and programmes", the former Director-General of the WHO, Margaret Chan, highlighted the need for integration of cervical cancer control into existing human immunodeficiency virus (HIV) services [42]. On February 2018, the current Director-General of the WHO Tedros Adhanom said[1]:

> *"One woman dies of cervical cancer every two minutes. Each one is a tragedy. We're working to scale up HPV vaccination, screening and treatment for cervical cancer to save many more lives."*

The fast growth and expansion of world population require the involvement of automated methodologies in order to ensure the massive scalability of such health-care problems. The present thesis described an effort to develop ML and CV methods to assist the digital colposcopy procedure with two target applications: cervical cancer and forensic assessment of sexual assault. As relevant as having an end-to-end automated system for the automation of these tasks is ensuring the acceptance of such system by the medical community. Therefore, in this thesis, we addressed both fundamental and applied topics in the area. First, we extrapolate typical requirements raised by our application to other areas, proposing general contributions that could be easily instantiated to other applications. The first part of this thesis (see Chapters 2-8) summarizes this line of research. Then, we instantiated several of our ideas and delved strategies to solve application-specific tasks in the second part (see Chapters 9-13).

The outcomes of this thesis have been published in several international conferences and specialized journals. This work benefited from collaborations with international research teams (i.e., Venezuela, Spain and Canada) and local colleagues. Recent agreements with active drivers in the area will enable us to instantiate several of the proposed ideas and to formulate new contributions in the field.

[1] https://twitter.com/drtedros/status/965690574853505024

14.1 Fundamental Contributions

- We proposed ranking as an alternative and general approach to model a vast amount of learning tasks, typically modeled as classification or regression tasks. We argue that ranking models are a closer representation of human preference models, being a step forward on the generation of human-like machine decisions. First, we presented an interpretable ranking model (Chapter 2, [97]). Second, we validate the gains of using ranking as the underlying learning paradigm when tackling traditional tasks such as binary (and ordinal) classification with imbalanced class distribution (Chapter 3, [58, 60–62, 264]). We achieved competitive performance, surpassing state-of-the-art techniques used to address unbalanced datasets. Improving the performance of ML models with this condition is a key issue given the difficulty in collecting large amounts of data from patients with a specific disease. As the final contribution of this research line, we present a novel learning paradigm that focuses on automating the "easy" cases and delegating "difficult" samples to a human examinator (Chapter 4, [59]).

- We addressed transfer learning as a way to learn robust models on scenarios where data is scarce and difficult to acquire, such as the ones studied in this thesis. We proposed a general transfer learning technique (Chapter 5, [89]) and instantiated it to the main major learning tasks: regression, classification, ranking and recommender systems. Then, we applied the proposed framework to the cervical cancer screening tasks of predicting the individual patient's risk and the quality assessment of digital colposcopies (Chapter 12, [95]).

- The study of classifiers that properly handle directional –angular or periodic– data was studied in Chapter 6. Angular data is often used to reference the cervix and vagina by gynecologists and forensic pathologists. Thus, we proposed a generalization to Logistic Regression (Chapter 6, [88]) to handle mixed (i.e., directional and linear) data.

- We drew a link between shallow and deep methodologies in Chapter 7 by extending a traditional descriptor in CV –Local Binary Patterns– to incorporate deep ideas. Achieving a configurable balance between traditional and deep techniques leads to a unified way of introducing expert and data-driven knowledge.

- We proposed a new paradigm for semantic image segmentation (Chapter 8), where the outcome is achieved by iteratively estimating the quality of a segmentation mask and by applying local refinements based on the quality estimation. This approach attempts to unify traditional and deep segmentation techniques. On the one hand, traditional methods such as region growing lack of high expressiveness inducing segmentations based on low-level information such as intensity level but have the

capability of building and improving partial solutions. On the other hand, deep methodologies reach high abstraction levels but are unable to recover from misdetections. Thus, the proposed contribution attempts to reduce the gap between both strategies. Also, we believe it is closer to the way we perceive concepts as humans, by refining and improving our perception of the truth.

14.2 Applied Contributions

Besides our contributions on general ML and CV topics, we applied our ideas to application-specific tasks. In the area of cervical cancer screening, we performed a literature review, categorizing the main areas of research, limitations and open challenges. We acquired, annotated and released a database that will serve to validate the performance of automated techniques by other members of the community. We worked on several of the main tasks, including:

- Temporal segmentation of the videos according to the colposcopy protocol into its four main modalities and removal of transition intervals (Chapter 10).
- Inference of the risk to develop cervical cancer at a patient level to optimize healthcare access (Chapter 12).
- Quality Assessment of digital colposcopies to select the best frames to perform the diagnosis (Chapter 12). The proposed methodology can learn robust models adapted to the specific preferences of the human evaluator.
- Segmentation of the main anatomical parts of the cervix and surrounding objects observed in a typical colposcopy (Chapter 11). The proposed methodology addressed the segmentation of all the objects of interest in a joint fashion, avoiding suboptimal cascade methodologies that cannot recover from errors committed in the preliminary stages.

Finally, our applied contribution on the automation of the forensic assessment of rape and sexual assault victims, we proposed an end-to-end methodology that covers basic tasks such as the modality recognition and removal of external objects to the identification and characterization of genital injuries and the final forensic assessment. We studied a visualization methodology as a preliminary attempt to provide an explanation about the model's decision.

14.3 Final Remarks and Future Work

Although several contributions were exposed in this thesis, achieving state-of-the-art results in several of the aforementioned tasks, the development of a CAD system for the

decision support of digital colposcopies is far from being solved. Through this thesis, we highlighted open problems and future lines of work on each chapter, being the list closer to being illustrative than to being exhaustive. Digital colposcopy and its two main applications, cervical cancer screening and forensic analysis, is an inexhaustible source of inspiration for both, fundamental and applied contributions in ML and CV. The high complexity of the data, the diverse acquisition settings and medical patterns that can be identified turns this area an ideal niche for exploring novel techniques in these areas. Here, we would like to highlight some final open problems that permeate all the tasks surrounding the development of CAD systems and, with special interest, for digital colposcopies.

In addition to being growing in volume, acquisition modalities for medical diagnosis have been growing in the last years. In this thesis, we discussed some underlying modalities in the digital colposcopy. However, other sources of data can be found in medical records, histological images, etc. With equal importance, the diversity of predictive tasks that can be considered in the usual development of these systems difficults the development of end-to-end automation. Therefore, the development of general pipelines able to cope with the specificities of each modality and task while keeping the learning process robust and reproducible is a must. While current trends in deep learning simplified some parts of the pipeline, the added complexity and lack of interpretability of the resulting models may raise difficulties on their acceptance.

In this thesis, we sketched some ideas on how to reach that goal, techniques such as ranking present promising results as a general approach to model several types of tasks with an unified learning scheme; transfer learning of high-level properties may enable us to selectively transfer knowledge across modalities and even across predictive tasks; the integration of shallow and deep strategies may restore the control on how much domain knowledge are we willing to embed in our tasks. In collaboration with Fraunhofer Portugal, an FCT project entitled "CLARE: Computer-Aided Cervical Cancer Screening" was submitted and accepted for funding. The project will strengthen the contributions presented in this work, focusing on the joint automated analysis of cytology and digital colposcopy for the early detection of cervical cancer.

In the future, we would like to observe end-to-end learning pipelines, capable of addressing diverse tasks, from segmentation to classification and taking advantage of multimodal, poorly annotated (e.g., weakly-supervised and semi-supervised) databases.

Humans are capable of working in teams and discussing ideas because they can argue and debate instead of dictating and answer. The holy grail of ML models in medicine is the construction of interpretable and self-explanatory models, where the physician is aware of the *why* and not only of the *what*. Relevant decision support on medicine should go beyond a binary label. In order to impact on the discovery of findings, CAD systems should be able to illustrate the human expert with explanations. Concrete lines (but limited) lines of work on this line include the explanation by example, the analysis of the impact that influenced the decision and the analysis of the impact of a decision in the next stages of

the diagnosis (i.e., treatment, surgical options, and potential side effects).

References

[1] The digital mammography dream challenge. `https://mobileodt.com/products/eva-colp/`, 2017. [Online; accessed 22-March-2018].

[2] Héctor-Gabriel Acosta-Mesa, Nicandro Cruz-Ramírez, Karina Gutiérrez-Fragoso, Rocío-Erandi Barrientos-Martínez, and Rodolfo Hernández-Jiménez. Assessing the possibility of identifying precancerous cervical lesions using aceto-white temporal patterns. *Decision Support Systems, Advances in*, pages 107–116, 2010.

[3] Héctor-Gabriel Acosta-Mesa, Nicandro Cruz-Ramírez, and Rodolfo Hernández-Jiménez. Aceto-white temporal pattern classification using k-nn to identify precancerous cervical lesion in colposcopic images. *Computers in biology and medicine*, 39(9):778–784, 2009.

[4] Héctor-Gabriel Acosta-Mesa, Nicandro Cruz-Ramírez, Rodolfo Hernández-Jiménez, and Daniel-Alejandro García-López. Modeling aceto-white temporal patterns to segment colposcopic images. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 548–555. Springer, 2007.

[5] Héctor-Gabriel Acosta-Mesa, Fernando Rechy-Ramírez, Efrén Mezura-Montes, Nicandro Cruz-Ramírez, and Rodolfo Hernández Jiménez. Application of time series discretization using evolutionary programming for classification of precancerous cervical lesions. *Journal of biomedical informatics*, 49:73–83, 2014.

[6] Héctor-Gabriel Acosta-Mesa, B Zitova, HV Rios-Figueroa, Nicandro Cruz-Ramirez, A Marin-Hernandez, Rodolfo Hernandez-Jimenez, Bertha E Cocotle-Ronzon, and Efrain Hernandez-Galicia. Cervical cancer detection using colposcopic images: a temporal approach. In *Sixth Mexican International Conference on Computer Science (ENC'05)*, pages 158–164. IEEE, 2005.

[7] Ramesh Agarwal and Mahesh V Joshi. Pnrule: A new framework for learning classifier models in data mining (a case-study in network intrusion detection). In *Proceedings of the 2001 SIAM International Conference on Data Mining*, pages 1–17. SIAM, 2001.

[8] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In *European conference on computer vision*, pages 469–481. Springer, 2004.

[9] Amir Alush, Hayit Greenspan, and Jacob Goldberger. Lesion detection and segmentation in uterine cervix images using an arc-level MRF. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages 474–477. IEEE, 2009.

[10] Amir Alush, Hayit Greenspan, and Jacob Goldberger. Automated and interactive lesion detection and segmentation in uterine cervix images. *IEEE transactions on medical imaging*, 29(2):488–501, 2010.

[11] Ana Paula Brand ao Lopes, Sandra Eliza Fontes de Avila, Anderson Nunes Alves Peixoto, Rodrigo Silva Oliveira, and Arnaldo de Albuquerque Araújo. A bag-of-features approach based on hue-sift descriptor for nude detection. In *Proceedings of the XVII European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland, 2009.

[12] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[13] Andreas Argyriou, Massimiliano Pontil, Yiming Ying, and Charles A Micchelli. A spectral regularization framework for multi-task structure learning. In *Advances in neural information processing systems*, pages 25–32, 2007.

[14] Yusuf Artan and Xiaolei Huang. Combining multiple 2ν-svm classifiers for tissue segmentation. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 488–491. IEEE, 2008.

[15] Juan D García Arteaga and Jan Kybic. Automatic landmark detection for cervical image registration validation. In *Medical Imaging*, pages 65142S–65142S. International Society for Optics and Photonics, 2007.

[16] Birgitte Schidt Astrup and Annemette Wildfang Lykkebo. Post-coital genital injury in healthy women: A review. *Clinical Anatomy*, 28(3):331–338, 2015.

[17] Birgitte Schmidt Astrup, Jens Lauritsen, Pernille Ravn, and Jørgen Lange Thomsen. Genital lesions after consensual sexual intercourse: They are frequent and they last for several days. In *19th World meeting of the International Association of Forensic Sciences*, 2011.

[18] Birgitte Schmidt Astrup, Jens Lauritsen, Jørgen Lange Thomsen, and Pernille Ravn. Colposcopic photography of genital injury following sexual intercourse in adults. *Forensic science, medicine, and pathology*, 9(1):24–30, 2013.

[19] Birgitte Schmidt Astrup, Pernille Ravn, Jens Lauritsen, and Jørgen Lange Thomsen. Nature, frequency and duration of genital lesions after consensual sexual intercourse—implications for legal proceedings. *Forensic science international*, 219(1):50–56, 2012.

[20] Birgitte Schmidt Astrup, Pernille Ravn, Jørgen Lange Thomsen, and Jens Lauritsen. Patterned genital injury in cases of rape–a case–control study. *Journal of forensic and legal medicine*, 20(5):525–529, 2013.

[21] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo De A AraúJo. Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding*, 117(5):453–465, 2013.

[22] Yusuf Aytar and Andrew Zisserman. Tabula rasa: Model transfer for object category detection. In *2011 International Conference on Computer Vision*, pages 2252–2259. IEEE, 2011.

[23] Francis R Bach. Considering Cost Asymmetry in Learning Classifiers. *Jmlr*, 7:1713–1741, 2006.

[24] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[25] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. In *Journal of Machine Learning Research*, pages 1345–1382, 2005.

[26] Oren Barkan, Jonathan Weill, Lior Wolf, and Hagai Aronowitz. Fast high dimensional vector multiplication face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1960–1967, 2013.

[27] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[28] Shai Ben-David and Ruth Urner. Domain adaptation as learning with auxiliary information. In *New Directions in Transfer and Multi-Task-Workshop@ NIPS*, 2013.

[29] Ewert Bengtsson and Patrik Malm. Screening for cervical cancer using automated analysis of pap-smears. *Computational and mathematical methods in medicine*, 2014, 2014.

[30] Kristin P Bennett and Olvi L Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization methods and software*, 1(1):23–34, 1992.

[31] Silvia Bessa, Inês Domingues, Jaime S. Cardoso, Pedro Passarinho, Pedro Cardoso, Vitor Rodrigues, and Fernando Lage. Normal breast identification in screening mammography: A study on 18 000 images. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 325–330. IEEE, 2014.

[32] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[33] Edwin V Bonilla, Kian M Chai, and Christopher Williams. Multi-task gaussian process prediction. In *Advances in neural information processing systems*, pages 153–160, 2007.

[34] Richard Booth, Yann Chevaleyre, Jérôme Lang, Jérôme Mengin, and Chattrakul Sombattheera. Learning conditionally lexicographic preference relations. In *ECAI*, pages 269–274, 2010.

[35] Craig Boutilier, Ronen I Brafman, Carmel Domshlak, Holger H Hoos, and David Poole. CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *J. Artif. Intell. Res.(JAIR)*, 21:135–191, 2004.

[36] Michael Bräuning and Eyke Hüllermeier. Learning conditional lexicographic preference trees. *Workshop on Preference learning: problems and applications in AI, ECAI*, pages 11–15, 2012.

[37] Phil Brodatz. *Textures: a photographic album for artists and designers*. Dover Pubns, 1966.

[38] Attila Budai, Rüdiger Bock, Andreas Maier, Joachim Hornegger, and Georg Michelson. Robust vessel segmentation in fundus images. *International journal of biomedical imaging*, 2013, 2013.

[39] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pages 89–96, New York, New York, USA, 2005. ACM Press.

[40] Jaime S. Cardoso and Maria J Cardoso. Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. *Artificial intelligence in medicine*, 40(2):115–126, 2007.

[41] Jaime S. Cardoso and Joaquim F. Costa. Learning to classify ordinal data: The data replication method. *Journal of Machine Learning Research*, 8(Jul):1393–1429, 2007.

[42] Margaret Chan. Grand challenges for the next decade in global health policy and programmes. `http://www.who.int/dg/speeches/2017/address-university-washington/en/`, 2017. [Online; accessed 21-March-2018].

[43] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[44] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[45] Sheng Chen, Haibo He, and Edwardo a Garcia. RAMOBoost: Ranked Minority Oversampling in Boosting. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 21(10):1624–42, 2010.

[46] Weiwei Cheng, Michaël Rademaker, Bernard De Baets, and Eyke Hüllermeier. Predicting partial orders: ranking with abstention. In *Machine Learning and Knowledge Discovery in Databases*, pages 215–230. Springer, 2010.

[47] Weiwei Cheng, Michaël Rademaker, Bernard De Baets, and Eyke Hüllermeier. Predicting partial orders: Ranking with abstention. In *Machine Learning and Knowledge Discovery in Databases*, pages 215–230. Springer, 2010.

[48] Youngmin Cho and Lawrence K Saul. Kernel methods for deep learning. In *Advances in neural information processing systems*, pages 342–350, 2009.

[49] Jae Young Choi, Konstantinos N Plataniotis, and Yong Man Ro. Using colour local binary pattern features for face recognition. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 4541–4544. IEEE, 2010.

[50] Wei Chu and S Sathiya Keerthi. New approaches to support vector ordinal regression. In *Proceedings of the 22nd international conference on Machine learning*, pages 145–152. ACM, 2005.

[51] Isabelle Claude, Renaud Winzenrieth, Philippe Pouletaut, and J-C Boulanger. Contour features for colposcopic image classification by artificial neural networks. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 1, pages 771–774. IEEE, 2002.

[52] Corinna Cortes and Mehryar Mohri. AUC Optimization vs. Error Rate Minimization. *Advances in Neural Information Processing Systems*, pages 313–320, 2003.

[53] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[54] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.

[55] Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. In *EUROSIS*, 2008.

[56] David Cossock and Tong Zhang. Subset ranking using regression. In *Learning theory*, pages 605–619. Springer, 2006.

[57] Koby Crammer, Yoram Singer, et al. Pranking with ranking. In *NIPS*, volume 14, pages 641–647, 2001.

[58] Ricardo Cruz, Kelwin Fernandes, Jaime S. Cardoso, and Joaquim F. Pinto Costa. Tackling class imbalance with ranking. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 2182–2187. IEEE, 2016.

[59] Ricardo Cruz, Kelwin Fernandes, Joaquim F. Pinto Costa, and Jaime S. Cardoso. Constraining Type II Error: Building Intentionally Biased Classifiers. In *International Work-Conference on Artificial Neural Networks*, pages 549–560. Springer, 2017.

[60] Ricardo Cruz, Kelwin Fernandes, Joaquim F. Pinto Costa, María Pérez Ortiz, and Jaime S. Cardoso. Combining ranking with traditional methods for ordinal class imbalance. In *International Work-Conference on Artificial Neural Networks*, pages 538–548. Springer, 2017.

[61] Ricardo Cruz, Kelwin Fernandes, Joaquim F. Pinto Costa, María Pérez Ortiz, and Jaime S. Cardoso. Ordinal class imbalance with ranking. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 3–12. Springer, 2017.

[62] Ricardo Cruz, Kelwin Fernandes, Joaquim F. Pinto Costa, María Pérez-Ortiz, and Jaime S. Cardoso. Binary ranking for ordinal class imbalance. In *Pattern Analysis and Applications*. Springer, 2018.

[63] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 193–200. ACM, 2007.

[64] Abhishek Das, Avijit Kar, and Debasis Bhattacharyya. Elimination of specular reflection and identification of ROI: The first step in automated detection of cervical cancer using digital colposcopy. In *Imaging Systems and Techniques (IST), 2011 IEEE International Conference on*, pages 237–241. IEEE, 2011.

[65] Abhishek Das, Avijit Kar, and Debasis Bhattacharyya. Preprocessing for automating early detection of cervical cancer. In *Information Visualisation (IV), 2011 15th International Conference on*, pages 597–600. IEEE, 2011.

[66] Abhishek Das, Avijit Kar, and Debasis Bhattacharyya. Implication of technology on society in asia: Automated detection of cervical cancer. In *Technology and Society in Asia (T&SA), 2012 IEEE Conference on*, pages 1–4. IEEE, 2012.

[67] Abhishek Das, Avijit Kar, and Debasis Bhattacharyya. Detection of abnormal regions of precancerous lesions in digitised uterine cervix images. In *Electrical Engineering Congress (iEECON), 2014 International*, pages 1–4. IEEE, 2014.

[68] Abhishek Das, Avijit Kar, and Debasis Bhattacharyya. Early detection of cervical cancer using novel segmentation algorithms. *Invertis Journal of Science & Technology*, 7(2):91–95, 2014.

[69] Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006.

[70] S Hashem Davarpanah, Fatimah Khalid, Lili Nurliyana Abdullah, and Maryam Golchin. A texture descriptor: Background local binary pattern (bglbp). *Multimedia Tools and Applications*, 75(11):6549–6568, 2016.

[71] Jesse Davis and Pedro Domingos. Deep transfer via second-order markov logic. In *Proceedings of the 26th annual International Conference on Machine Learning*, pages 217–224. ACM, 2009.

[72] Misha Denil and Thomas Trappenberg. Overlap versus imbalance. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6085 LNAI:220–231, 2010.

[73] Dieter Devlaminck, Willem Waegeman, Bruno Bauwens, Bart Wyns, Patrick Santens, and Georges Otte. From circular ordinal regression to multilabel classification. In *Proceedings of the 2010 workshop on preference learning, European conference on machine learning*, 2010.

[74] Michael J DeWeert, Jody Oyama, Elisabeth McLaughlin, Ellen Jacobson, Johan Hakansson, Gary S Bignami, Ulf P Gustafsson, Paul Troy, Violeta Poskiene, Kristina Kriukelyte, et al. Analysis of spatial variability in hyperspectral imagery of the uterine cervix in vivo. In *Biomedical Optics 2003*, pages 67–76. International Society for Optics and Photonics, 2003.

[75] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[76] Pedro Domingos. MetaCost: A General Method for Making Classifiers Cost-Sensitive. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 55:155–164, 1999.

[77] Mark Dredze, Alex Kulesza, and Koby Crammer. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning*, 79(1-2):123–149, 2010.

[78] John S Dryzek and Christian List. Social choice theory and deliberative democracy: a reconciliation. *British journal of political science*, 33(1):1–28, 2003.

[79] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[80] Othmane El Meslouhi, Mustapha Kardouchi, Hakim Allali, and Toufiq Gadi. Semi-automatic cervical cancer segmentation using active contours without edges. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2009 Fifth International Conference on*, pages 54–58. IEEE, 2009.

[81] Daniel Ellsberg. Classic and current notions of" measurable utility". *The Economic Journal*, 64(255):528–556, 1954.

[82] M Elter, R Schulz-Wendtland, and T Wittenberg. The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process. *Medical Physics*, 34(11):4164–4172, 2007.

[83] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660, 2010.

[84] A Evgeniou and Massimiliano Pontil. Multi-task feature learning. *Advances in Neural Information Processing Systems*, 19:41, 2007.

[85] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117. ACM, 2004.

[86] Jianping Fan, David KY Yau, Ahmed K Elmagarmid, and Walid G Aref. Automatic image segmentation by integrating color-edge extraction and seeded region growing. *IEEE transactions on image processing*, 10(10):1454–1466, 2001.

[87] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[88] Kelwin Fernandes and Jaime S. Cardoso. Discriminative directional classifiers. *Neurocomputing*, 207:141–149, 2016.

[89] Kelwin Fernandes and Jaime S. Cardoso. Hypothesis transfer learning based on structural model similarity. *Neural Computing and Applications*, pages 1–14, 2018.

[90] Kelwin Fernandes and Jaime S. Cardoso. Ordinal image segmentation using deep neural networks. In *Neural Networks (IJCNN), 2018 International Joint Conference on*. IEEE, 2018.

[91] Kelwin Fernandes, Jaime S. Cardoso, and Birgitte Astrup. A deep learning approach for the forensic evaluation of sexual assault. *Pattern Analysis and Applications*, 2018.

[92] Kelwin Fernandes, Jaime S. Cardoso, and Birgitte Schmidt Astrup. Automated detection and categorization of genital injuries using digital colposcopy. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 251–258. Springer, 2017.

[93] Kelwin Fernandes, Jaime S. Cardoso, and Birgitte Schmidt Astrup. A deep learning approach for the forensic evaluation of sexual assault. In *Pattern Analysis and Applications*. Springer, 2018.

[94] Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. Temporal segmentation of digital colposcopies. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 262–271. Springer, 2015.

[95] Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. Transfer learning with partial observability applied to cervical cancer screening. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 243–250. Springer, 2017.

[96] Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. Automated methods for the decision support of cervical cancer screening using digital colposcopies. *IEEE Access*, 2018.

[97] Kelwin Fernandes, Jaime S. Cardoso, and Hector Palacios. Learning and ensembling lexicographic preference trees with multiple kernels. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 2140–2147. IEEE, 2016.

[98] Kelwin Fernandes, Jaime S. Cardoso, and Hector Palacios. Learning and ensembling lexicographic preference trees with multiple kernels. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, 2016.

[99] Kelwin Fernandes, Davide Chicco, Jaime S. Cardoso, and Jessica Fernandes. Supervised deep learning embeddings for the prediction of cervical cancer diagnosis. *PeerJ Computer Science*, 2018.

[100] Kelwin Fernandes, Ricardo Cruz, and Jaime S. Cardoso. Deep image segmentation by quality inference. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, 2018.

[101] Kelwin Fernandes, Ricardo Cruz, and Jaime S. Cardoso. Image segmentation by quality inference. In *Neural Networks (IJCNN), 2018 International Joint Conference on*. IEEE, 2018.

[102] Kelwin Fernandez and Carolina Chang. Teeth/palate and interdental segmentation using artificial neural networks. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 175–185. Springer, 2012.

[103] Daron G Ferris, J Thomas Cox, Louis Burke, Mark S Litaker, Diane M Harper, Michael J Campion, Mitchell D Greenberg, Lisa McShane, and Lap Ming Wun. Colposcopy quality control: establishing colposcopy criterion standards for the National Cancer Institute ALTS trial using cervigrams. *Journal of lower genital tract disease*, 2(4):195–203, 1998.

[104] Daron G Ferris, Raymond A Lawhead, Eileen D Dickman, Nina Holtzapple, Jill A Miller, Stephanie Grogan, Shabbir Bambot, Anant Agrawal, and Mark L Faupel. Multimodal hyperspectral imaging for the noninvasive diagnosis of cervical neoplasia. *Journal of Lower Genital Tract Disease*, 5(2):65–72, 2001.

[105] Adelaide Figueiredo. Discriminant analysis for the von mises-fisher distribution. *Communications in Statistics-Simulation and Computation*, 38(9):1991–2003, 2009.

[106] Adelaide Figueiredo and Paulo Gomes. Discriminant analysis based on the watson distribution defined on the hypersphere. *Statistics*, 40(5):435–445, 2006.

[107] Peter C Fishburn. Utility theory for decision making. Technical report, Research analysis corp McLean VA, 1970.

[108] NI Fisher and AJ Lee. Regression models for an angular response. *Biometrics*, pages 665–677, 1992.

[109] Peter Flach and Edson Takashi Matsubara. A simple lexicographic ranker and probability estimator. In *ECML*, pages 575–582. Springer, 2007.

[110] Aldrin Barreto Flores, Leopoldo Altamirano Robles, Rosa Maria Morales Tepalt, and Juan D Cisneros Aragon. Identifying precursory cancer lesions using temporal texture analysis. In *The 2nd Canadian Conference on Computer and Robot Vision (CRV'05)*, pages 34–39. IEEE, 2005.

[111] Centers for Disease Control, Prevention (CDC, et al. Cervical cancer screening among women aged 18-30 years-united states, 2000-2010. *MMWR. Morbidity and mortality weekly report*, 61(51-52):1038, 2013.

[112] Eibe Frank and Mark Hall. A simple approach to ordinal classification. *Machine Learning: ECML 2001*, pages 145–156, 2001.

[113] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *The Journal of machine learning research*, 4:933–969, 2003.

[114] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[115] Johannes Fürnkranz and Eyke Hüllermeier. *Preference learning*. Springer, 2010.

[116] Javier Galbally, Sébastien Marcel, and Julian Fierrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE transactions on image processing*, 23(2):710–724, 2014.

[117] Fei Gao, Dacheng Tao, Xinbo Gao, and Xuelong Li. Learning to rank for blind image quality assessment. *IEEE transactions on neural networks and learning systems*, 26(10):2275–2290, 2015.

[118] Juan D Garcıa-Arteaga, Jan Kybic, Jia Gu, and Wenjing Li. Geometric and information constraints for automatic landmark selection in colposcopy sequences, 2007.

[119] Juan D García-Arteaga, Jan Kybic, and Wenjing Li. Automatic colposcopy video tissue classification using higher order entropy-based image registration. *Computers in biology and medicine*, 41(10):960–970, 2011.

[120] Juan David Garcıa-Arteaga. *Multichannel Image Information Similarity Measures: Applications to Colposcopy Image Registration.* PhD thesis, Faculty of Electrical Engineering, Czech Technical University, 2012.

[121] Juan David García-Arteaga, Jan Kybic, and Wenjing Li. Elastic image registration for movement compensation in digital colposcopy. *Buiosignal: Analysis of Biomedical Signals and Images, Brno, Czech Republic June*, pages 236–238, 2006.

[122] Jochen Garcke and Thomas Vanck. Importance weighted inductive transfer learning for regression. In *Machine Learning and Knowledge Discovery in Databases*, pages 466–481. Springer, 2014.

[123] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

[124] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[125] Víctor González-Castro, Rocío Alaiz-Rodríguez, Laura Fernández-Robles, Roberto Guzmán-Martínez, and Enrique Alegre. Estimating class proportions in boar semen analysis using the hellinger distance. *Trends in Applied Intelligent Systems*, pages 284–293, 2010.

[126] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[127] Shiri Gordon and Hayit Greenspan. Segmentation of non-convex regions within uterine cervix images. In *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 312–315. IEEE, 2007.

[128] Shiri Gordon and Hayit Greenspan. An agglomerative segmentation framework for non-convex regions within uterine cervix images. *Image and Vision Computing*, 28(12):1682–1701, 2010.

[129] Shiri Gordon, Gali Zimmerman, and Hayit Greenspan. Image segmentation of uterine cervix images for indexing in pacs. In *Computer-Based Medical Systems, 2004. CBMS 2004. Proceedings. 17th IEEE Symposium on*, page 298. IEEE, 2004.

[130] Shiri Gordon, Gali Zimmerman, Rodney Long, Sameer Antani, Jose Jeronimo, and Hayit Greenspan. Content analysis of uterine cervix images: initial steps towards content based indexing and retrieval of cervigrams. In *Medical Imaging 2006: Image Processing*, volume 6144, page 61444U. International Society for Optics and Photonics, 2006.

[131] Hayit Greenspan, Shiri Gordon, Gali Zimmerman, Shelly Lotenberg, Jose Jeronimo, Sameer Antani, and Rodney Long. Automatic detection of anatomical landmarks in uterine cervix images. *IEEE Transactions on Medical Imaging*, 28(3):454–468, 2009.

[132] Jia Gu and Wenjing Li. Automatic image quality assessment for uterine cervical imagery. In *Medical Imaging 2006: Image Perception, Observer Performance, and Technology Assessment*, volume 6146, page 61461B. International Society for Optics and Photonics, 2006.

[133] Peng Gu, Won-Mean Lee, Marilyn A Roubidoux, Jie Yuan, Xueding Wang, and Paul L Carson. Automated 3d ultrasound image segmentation to aid breast cancer image interpretation. *Ultrasonics*, 65:51–58, 2016.

[134] Zhenhua Guo, Lei Zhang, and David Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663, 2010.

[135] Zhenhua Guo, Lei Zhang, and David Zhang. Rotation invariant texture classification using lbp variance (lbpv) with global matching. *Pattern recognition*, 43(3):706–719, 2010.

[136] Ulf P Gustafsson, Elisabeth McLaughlin, Ellen Jacobsen, Johan Hakansson, Paul Troy, Michael J DeWeert, Katarina Svanberg, Sara Palsson, Marcelo Soto Thompson, Sune Svanberg, et al. In-vivo fluorescence and reflectance imaging of human cervical tissue. In *Medical Imaging 2003*, pages 521–530. International Society for Optics and Photonics, 2003.

[137] Ulf P Gustafsson, Elisabeth McLaughlin, Ellen Jacobson, Paul Troy, Michael J DeWeert, P Sara, Marcelo Soto Thompson, Sune Svanberg, Aurelija Vatkuviene, Katarina Svanberg, et al. Fluorescence and reflectance monitoring of human cervical tissue in vivo: a case study. In *Biomedical Optics 2003*, pages 100–110. International Society for Optics and Photonics, 2003.

[138] Pedro Antonio Gutiérrez, María Pérez-Ortiz, Javier Sanchez-Monedero, Francisco Fernández-Navarro, and Cesar Hervas-Martinez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2016.

[139] K Gutiérrez-Fragoso, HG Acosta-Mesa, N Cruz-Ramírez, and R Hernández-Jiménez. Automatic classification of acetowhite temporal patterns to identify precursor lesions of cervical cancer. In *Journal of Physics: Conference Series*. IOP Publishing, 2013.

[140] Karina Gutiérrez-Fragoso, Héctor Gabriel Acosta-Mesa, Nicandro Cruz-Ramírez, and Rodolfo Hernández-Jiménez. Optimization of classification strategies of acetowhite temporal patterns towards improving diagnostic performance of colposcopy. *Computational and mathematical methods in medicine*, 2017, 2017.

[141] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016.

[142] F Maxwell Harper and Joseph A Konstan. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2015.

[143] Eric Hayman, Barbara Caputo, Mario Fritz, and Jan-Olof Eklundh. On the significance of real-world conditions for material classification. In *European conference on computer vision*, pages 253–266. Springer, 2004.

[144] Haibo He and Edwardo A. Garcia. Learning from Imbalanced Data Sets. *IEEE Transactions on knowledge and data engineering*, 21(9):1263—-1284, 2010.

[145] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[146] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[147] World Health Organization. Reproductive Health, World Health Organization. Chronic Diseases, and Health Promotion. *Comprehensive cervical cancer control: a guide to essential practice*. World Health Organization, 2006.

[148] Marko Heikkilä, Matti Pietikäinen, and Cordelia Schmid. Description of interest regions with center-symmetric local binary patterns. In *Computer vision, graphics and image processing*, pages 58–69. Springer, 2006.

[149] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. *Advances in neural information processing systems*, pages 115–132, 1999.

[150] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support vector learning for ordinal regression. In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, volume 1, pages 97–102. IET, 1999.

[151] Rolando Herrero, Mark H Schiffman, Concepción Bratti, Allan Hildesheim, Ileana Balmaceda, Mark E Sherman, Mitchell Greenberg, Fernando Cárdenas, Víctor Gómez, Kay Helgesen, et al. Design and methods of a population-based natural history study of cervical neoplasia in a rural province of costa rica: the guanacaste project. *Revista Panamericana de Salud Pública*, 1(6):411–425, 1997.

[152] Jason Hill, Enrique Corona, Jingqi Ao, Sunanda Mitra, and Brian Nutter. Information theoretic clustering for medical image segmentation. In *Advanced Computational Approaches to Biomedical Engineering*, pages 47–70. Springer, 2014.

[153] Tin Kam Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, 1998.

[154] Shengguo Hu, Yanfeng Liang, Lintao Ma, and Ying He. MSMOTE: Improving classification performance when training data is imbalanced. *2nd International Workshop on Computer Science and Engineering, WCSE 2009*, 2:13–17, 2009.

[155] Di Huang, Caifeng Shan, Mohsen Ardabilian, Yunhong Wang, and Liming Chen. Local binary patterns and its application to facial image analysis: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6):765–781, 2011.

[156] Sheng Huang, Mingchen Gao, Dan Yang, Xiaolei Huang, Ahmed Elgammal, and Xiaohong Zhang. Unbalanced graph-based transduction on superpixels for automatic cervigram image segmentation. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 1556–1559. IEEE, 2015.

[157] Xiaolei Huang and Gavriil Tsechpenakis. Medical image segmentation. *Information Discovery on Electronic Health Records*, 10:251–289, 2009.

[158] Xiaolei Huang, Wei Wang, Zhiyun Xue, Sameer Antani, L Rodney Long, and Jose Jeronimo. Tissue classification using cluster features for lesion detection in digital cervigrams. In *Medical Imaging*, pages 69141Z–69141Z. International Society for Optics and Photonics, 2008.

[159] Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16):1897–1916, 2008.

[160] Daniel P. Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.

[161] Qiang Ji, John Engel, and Eric Craine. Texture analysis for classification of cervix lesions. *IEEE Transactions on medical imaging*, 19(11):1144–1149, 2000.

[162] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[163] Jing Jiang. Multi-task transfer learning for weakly-supervised relation extraction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1012–1020. Association for Computational Linguistics, 2009.

[164] Lei Jiang, Jian Zhang, and Gabrielle Allen. Transferred correlation learning: An incremental scheme for neural network ensembles. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–8. IEEE, 2010.

[165] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[166] Kaggle. Cervical Cancer Screening. `https://www.kaggle.com/c/cervical-cancer-screening`, 2015. [Online; accessed 7-March-2018].

[167] Kaggle. Intel & MobileODT Cervical Cancer Screening. `https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening`, 2017. [Online; accessed 5-March-2018].

[168] Kaggle. Intel & MobileODT Cervical Cancer Screening Competition, 1st Place Winner's Interview: Team "Towards Empirically Stable Training". `https://tinyurl.com/KaggleMobileFirst`, 2017. [Online; accessed 5-March-2018].

[169] Chetak Kandaswamy, Luís M Silva, and Jaime S. Cardoso. Source-target-source classification using stacked denoising autoencoders. In *Pattern Recognition and Image Analysis*, pages 39–47. Springer, 2015.

[170] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014.

[171] Ashish Kapoor and Rosalind W Picard. Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 677–682. ACM, 2005.

[172] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

[173] Shogo Kato, Kunio Shimizu, and Grace S Shieh. A circular-circular regression model. *Statistica Sinica*, 18(2):633, 2008.

[174] Robert P Kauffman, Stephen J Griffin, Jon D Lund, and Paul E Tullar. Current recommendations for cervical cancer screening: do they render the annual pelvic examination obsolete? *Medical Principles and Practice*, 22(4):313–322, 2013.

[175] Navdeep Kaur, Nikson Panigrahi, and Ajay Mittal. Automated cervical cancer screening using transfer learning. In *International Conference on Recent Advances in Engineering Science and Management 2017*, 2017.

[176] Shehroz S Khan and Michael G Madden. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3):345–374, 2014.

[177] Edward Kim and Xiaolei Huang. A data driven approach to cervigram image analysis and classification. In *Color Medical Image analysis*, pages 1–13. Springer, 2013.

[178] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[179] M Kirby and Rick Miranda. Circular nodes in neural networks. *Neural Computation*, 8(2):390–402, 1996.

[180] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[181] Joseph B Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964.

[182] Miroslav Kubat, Robert Holte, and Stan Matwin. Learning when negative examples abound. In *Journal of Chemical Information and Modeling*, volume 53, pages 146–153. Springer, 1997.

[183] V Kudva, K Prasad, and S Guruvare. Detection of specular reflection and segmentation of cervix region in uterine cervix images for cervical cancer screening. *IRBM*, 38(5):281–291, 2017.

[184] Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *ICML (3)*, pages 942–950, 2013.

[185] Ilja Kuzborskij and Francesco Orabona. Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, 106(2):171–195, 2017.

[186] Gustaf Kylberg. The kylberg texture dataset v. 1.0. External report (Blue series) 35, Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, Uppsala, Sweden, September 2011.

[187] Christopher T Lam, Marlee S Krieger, Jennifer E Gallagher, Betsy Asma, Lisa C Muasher, John W Schmitt, and Nimmi Ramanujam. Design of a novel low cost point of care tampon (pocket) colposcope for use in resource limited settings. *PloS one*, 10(9):e0135869, 2015.

[188] Christopher T Lam, Jenna Mueller, Betsy Asma, Mercy Asiedu, Marlee S Krieger, Rhea Chitalia, Denali Dahl, Peyton Taylor, John W Schmitt, and Nimmi Ramanujam. An integrated strategy for improving contrast, durability, and portability of a pocket colposcope for cervical cancer screening and diagnosis. *PloS one*, 13(2):e0192530, 2018.

[189] Jérôme Lang, Jérôme Mengin, and Lirong Xia. Aggregating conditionally lexicographic preferences on multi-issue domains. In *Principles and Practice of Constraint Programming*, pages 973–987. Springer, 2012.

[190] Jérôme Lang, Leendert Van Der Torre, and Emil Weydert. Utilitarian desires. *Autonomous agents and Multi-agent systems*, 5(3):329–363, 2002.

[191] Jérôme Lang and Lirong Xia. Voting in combinatorial domains. *Handbook of Computational Social Choice*, 2014.

[192] Holger Lange. Automatic detection of multi-level acetowhite regions in RGB color images of the uterine cervix. In *Medical Imaging*, pages 1004–1017. International Society for Optics and Photonics, 2005.

[193] Holger Lange. Automatic glare removal in reflectance imagery of the uterine cervix. In *Medical Imaging 2005: Image Processing*, volume 5747, pages 2183–2193. International Society for Optics and Photonics, 2005.

[194] Holger Lange, Ross Baker, Johan Hakansson, and Ulf P Gustafsson. Reflectance and fluorescence hyperspectral elastic image registration. In *Proceedings of SPIE*, volume 5370, pages 335–345, 2004.

[195] Holger Lange and Daron G Ferris. Computer-aided-diagnosis (CAD) for colposcopy. In *Medical Imaging*, pages 71–84. International Society for Optics and Photonics, 2005.

[196] Pat Langley and Stephanie Sage. Induction of selective bayesian classifiers. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 399–406. Morgan Kaufmann Publishers Inc., 1994.

[197] Changki Lee and Myung-Gil Jang. A prior model of structural SVMs for domain adaptation. *ETRI Journal*, 33(5):712–719, 2011.

[198] Honglak Lee, Peter Pham, Yan Largman, and Andrew Y Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems*, pages 1096–1104, 2009.

[199] Hang Li. *Learning to Rank for Information Retrieval and Natural Language Processing*. Morgan & Claypool Publishers, 2011.

[200] Ping Li, Qiang Wu, and Christopher J Burges. Mcrank: Learning to rank using multiple classification and gradient boosting. In *Advances in neural information processing systems*, pages 897–904, 2007.

[201] Wenjing Li, Daron G Ferris, and Rich W Lieberman. Computerized image analysis for acetic acid induced intraepithelial lesions. In *Medical Imaging*, pages 69143A–69143A. International Society for Optics and Photonics, 2008.

[202] Wenjing Li, Jia Gu, Daron Ferris, and Allen Poirson. Automated image analysis of uterine cervical images. In *Medical Imaging*, pages 65142P–65142P. International Society for Optics and Photonics, 2007.

[203] Wenjing Li and Allen Poirson. Detection and characterization of abnormal vascular patterns in automated cervical image analysis. In *International Symposium on Visual Computing*, pages 627–636. Springer, 2006.

[204] Wenjing Li, Marcelo Soto-Thompson, and Ulf Gustafsson. A new image calibration system in digital colposcopy. *Optics express*, 14(26):12887–12901, 2006.

[205] Xiaodong Li, Weijie Mao, and Wei Jiang. Extreme learning machine based transfer learning for data classification. *Neurocomputing*, 174:203–210, 2016.

[206] Mingpei Liang, Gaopin Zheng, Xinyu Huang, Gaolin Milledge, and Alade Tokuta. Identification of abnormal cervical regions from colposcopy image sequences. In *International Conference on Computer Graphics, Visualization and Computer Vision*. Václav Skala-UNION Agency, 2013.

[207] M Lichman. UCI Machine Learning Repository, 2013.

[208] Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. Facing the cold start problem in recommender systems. *Expert Systems with Applications*, 41(4):2065–2073, 2014.

[209] Max A Little, Patrick E McSharry, Stephen J Roberts, Declan AE Costello, Irene M Moroz, et al. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*, 6(1):23, 2007.

[210] Jun Liu, Ling Li, and Lei Wang. Acetowhite region segmentation in uterine cervix images using a registered ratio image. *Computers in biology and medicine*, 93:47–55, 2018.

[211] Li Liu, Lingjun Zhao, Yunli Long, Gangyao Kuang, and Paul Fieguth. Extended local binary patterns for texture classification. *Image and Vision Computing*, 30(2):86–99, 2012.

[212] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[213] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory Undersampling for Class Imbalance Learning. *IEEE Transactions on Systems, Man and Cybernetics*, 39(2):539–550, 2009.

[214] Xudong Liu and Miroslaw Truszczynski. Aggregating conditionally lexicographic preferences using answer set programming solvers. In *Algorithmic Decision Theory*, pages 244–258. Springer, 2013.

[215] Xudong Liu and Miroslaw Truszczynski. Learning partial lexicographic preference trees over combinatorial domains. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[216] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1377–1385, 2015.

[217] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[218] Mingsheng Long, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. *arXiv preprint arXiv:1605.06636*, 2016.

[219] Philip M Long and Rocco A Servedio. Discriminative learning can succeed where generative learning fails. In *Learning Theory*, pages 319–334. Springer, 2006.

[220] Ana Paula B Lopes, Sandra EF de Avila, Anderson NA Peixoto, Rodrigo S Oliveira, Marcelo de M Coelho, and Arnaldo de A Araújo. Nude detection in video using bag-of-visual-features. In *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on*, pages 224–231. IEEE, 2009.

[221] Pedro L López-Cruz, Concha Bielza, and Pedro Larrañaga. The von Mises Naive Bayes classifier for angular data. In *Advances in Artificial Intelligence*, pages 145–154. Springer, 2011.

[222] Pedro L López-Cruz, Concha Bielza, and Pedro Larrañaga. Directional naive Bayes classifiers. *Pattern Analysis and Applications*, 18(2):225–246, 2013.

[223] Shelly Lotenberg, Shiri Gordon, and Hayit Greenspan. Shape priors for segmentation of the cervix region within uterine cervix images. *Journal of digital imaging*, 22(3):286–296, 2009.

[224] O. L. Mangasarian, W. N. Street, and W. H. Wolberg. Breast Cancer Diagnosis and Prognosis Via Linear Programming. *Operations Research*, 43(4):570–577, 1995.

[225] Kanti V Mardia, John T Kent, Zhengzheng Zhang, Charles C Taylor, and Thomas Hamelryck. Mixtures of concentrated multivariate sine distributions with applications to bioinformatics. *Journal of Applied Statistics*, 39(11):2475–2492, 2012.

[226] Vanesa Margariti, Michalis Zervakis, and Costas Balas. Wavelet and physical parametric analysis of the acetowhitening optical effect: comparative evaluation of performances in non-invasive diagnosis of cervical neoplasia. In *Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine*, pages 1–4. IEEE, 2010.

[227] Aldo Marquez-Grajales, Hector-Gabriel Acosta-Mesa, Efrén Mezura-Montes, and Rodolfo Hernández-Jiménez. Cervical image segmentation using active contours and evolutionary programming over temporary acetowhite patterns. In *Evolutionary Computation (CEC), 2016 IEEE Congress on*, pages 3863–3870. IEEE, 2016.

[228] D Pretty Mary, Vinolia Anandan, and KG Srinivasagan. An effective diagnosis of cervical cancer neoplasia by extracting the diagnostic features using crf. In *Computing, Electronics and Electrical Technologies (ICCEET), 2012 International Conference on*, pages 563–570. IEEE, 2012.

[229] Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2015.

[230] Peter McCullagh, John A Nelder, and P McCullagh. *Generalized linear models*, volume 2. Chapman and Hall London, 1989.

[231] Grit Mehlhorn, Christian Münzenmayer, Michaela Benz, Andreas Kage, Matthias W Beckmann, and Thomas Wittenberg. Computer-assisted diagnosis in colposcopy: results of a preliminary experiment? *Acta cytologica*, 56(5):554–559, 2012.

[232] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 5437–5440. IEEE, 2013.

[233] OE Meslouhi, Hakim Allali, Toufiq Gadi, and Mustapha Kardouchi. Colposcopic image registration using opponentSIFT descriptor. *Mediterranean Telecommunication Journal*, 1(2):74–79, 2011.

[234] Efrén Mezura-Montes, Héctor-Gabriel Acosta-Mesa, Darío-del-Sinaí Ramírez-Garcés, Nicandro Cruz-Ramírez, and Rodolfo Hernández-Jiménez. An image registration method for colposcopic images. *Computational and mathematical methods in medicine*, 2013, 2013.

[235] Nuno Miranda, Cristina Portugal, Paulo Jorge Nogueira, Carla Sofia Farinha, Ana Lisette Oliveira, Maria Isabel Alves, and José Martins. Portugal doenças oncológicas em números, 2015. *Portugal Doenças Oncológicas em números, 2015*, pages 7–65, 2016.

[236] Andriy Mnih and Ruslan R Salakhutdinov. Probabilistic matrix factorization. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1257–1264. Curran Associates, Inc., 2008.

[237] MobileODT. Eva colposcope. `https://mobileodt.com/products/eva-colp/`, 2018. [Online; accessed 21-February-2018].

[238] Jennifer A Mooney, Peter J Helms, and Ian T Jolliffe. Fitting mixtures of von mises distributions: a case study involving sudden infant death syndrome. *Computational Statistics & Data Analysis*, 41(3):505–513, 2003.

[239] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S. Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.

[240] Suleiman Mustafa, Steve Adeshina, Mohammed Dauda, and Wole Soboyejo. Classification of cervical cancer tissues using a novel low cost methodology for effective screening in rural settings. In *Electronics, Computer and Computation (ICECCO), 2014 11th International Conference on*, pages 1–4. IEEE, 2014.

[241] Loris Nanni, Carlo Fantozzi, and Nicola Lazzarini. Coupling different methods for overcoming the class imbalance problem. *Neurocomputing*, 158:48–61, 2015.

[242] Loris Nanni, Alessandra Lumini, and Sheryl Brahnam. Local binary patterns variants as texture descriptors for medical image analysis. *Artificial intelligence in medicine*, 49(2):117–125, 2010.

[243] Natalia Neverova, Christian Wolf, Graham W Taylor, and Florian Nebout. Multi-scale deep learning for gesture detection and localization. In *Workshop at the European Conference on Computer Vision*, pages 474–490. Springer, 2014.

[244] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841, 2002.

[245] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. In *NIPS*, volume 14-2, pages 849–856, 2001.

[246] M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[247] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[248] Mohammad Norouzi, Mani Ranjbar, and Greg Mori. Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2735–2742. IEEE, 2009.

[249] Haydemar Núñez, Luis Gonzalez-Abril, and Cecilio Angulo. A post-processing strategy for SVM learning from unbalanced data. *ESANN 2011 proceedings, 19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 195–200, 2010.

[250] International Federation of Gynecology and Obstetrics. Guía global para la prevención y control del cáncer cervicouterino. Technical report, International Federation of Gynecology and Obstetrics, 2009.

[251] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Gray scale and rotation invariant texture classification with local binary patterns. In *European Conference on Computer Vision*, pages 404–420. Springer, 2000.

[252] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[253] Hélder P Oliveira. Prediction of breast deformities: A step forward for planning aesthetic results after breast surgery. In *Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20-23, 2017, Proceedings*, volume 10255, page 267. Springer, 2017.

[254] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

[255] Rivka Oxman. The thinking eye: visual re-cognition in design emergence. *Design Studies*, 23(2):135–164, 2002.

[256] V Pallavi and K Payal. Automated analysis of cervix images to grade the severity of cancer. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3439–3442. IEEE, 2011.

[257] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

[258] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519. ACM, 2017.

[259] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[260] Sun Y Park, Dustin Sargent, Richard Lieberman, and Ulf Gustafsson. Domain-specific image analysis for cervical neoplasia detection based on conditional random fields. *Medical Imaging, IEEE Transactions on*, 30(3):867–878, 2011.

[261] Sun Young Park, Michele Follen, Andrea Milbourne, Anais Malpica, Nick MacKinnon, Mia K Markey, Rebecca Richards-Kortum, Calum MacAulay, and Helen Rhodes. Automated image analysis of digital colposcopy for the detection of cervical neoplasia. *Journal of biomedical optics*, 13(1):014029–014029, 2008.

[262] Sun Young Park, Dusty Sargent, Rolf Wolters, and Richard W Lieberman. Semantic image analysis for cervical neoplasia detection. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, pages 160–165. IEEE, 2010.

[263] Jason Payne-James, Anthony Busuttil, and William Smock. *Forensic medicine: clinical and pathological aspects*. Cambridge University Press, 2003.

[264] María Pérez-Ortiz, Kelwin Fernandes, Ricardo Cruz, Jaime S. Cardoso, Javier Briceño, and César Hervás-Martínez. Fine-to-coarse ranking in ordinal and imbalanced domains: An application to liver transplantation. In *International Work-Conference on Artificial Neural Networks*, pages 525–537. Springer, 2017.

[265] María Pérez-Ortiz, Pedro Antonio Gutiérrez, and César Hervás-Martínez. Projection-based ensemble learning for ordinal regression. *IEEE transactions on cybernetics*, 44(5):681–694, 2014.

[266] María Pérez-Ortiz, Aurora Sáez, Javier Sánchez-Monedero, Pedro Antonio Gutiérrez, and César Hervás-Martínez. Tackling the ordinal and imbalance nature of a melanoma image classification problem. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 2156–2163. IEEE, 2016.

[267] Michaël Perrot and Amaury Habrard. A theoretical analysis of metric hypothesis transfer learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1708–1717, 2015.

[268] Dzung L Pham, Chenyang Xu, and Jerry L Prince. Current methods in medical image segmentation. *Annual review of biomedical engineering*, 2(1):315–337, 2000.

[269] BD Focal Point. BD SurePath. `http://www.bd.com/tripath/labs/fpscreening.asp`, 2018. [Online; accessed 21-February-2018].

[270] Daniel Povey, Stephen M Chu, and Balakrishnan Varadarajan. Universal background model based speech recognition. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4561–4564. IEEE, 2008.

[271] PS Rama Praba and H Ranganathan. Comparing different classifiers for automatic lesion detection in cervix based on colour histogram. *Journal of Computer Applications (JCA)*, 6(1), 2013.

[272] PS Rama Praba and H Ranganathan. Wavelet transform based automatic lesion detection in cervix images using active contour. *Journal of Computer Science*, 9(1):30, 2013.

[273] VG Prabitha, S Suchetha, JL Jayanthi, P Rema, KV Baiju, Nita Sukumar, Anita Mathews, Paul Sebastian, and N Subhash. Multi-spectral diffuse reflectance imaging for detection of cervical lesions: a pilot study. *International Journal of Engineering Science and Innovative Technology*, 3, 2014.

[274] R.C. Prati, G.E. Batista, and M.C. Monard. Class imbalances versus class overlapping: an analysis of a learning system behavior. *MICAI 2004: Advances in Artificial Intelligence*, pages 312–321, 2004.

[275] Viara Van Raad and Andrew P Bradley. Active contour model based segmentation of colposcopy images of cervix uteri using gaussian pyramids. In *6th International Symposium on Digital Signal Processing for Communication Systems (DSPCS'02)*, 2002.

[276] Julien Rabin, Julie Delon, and Yann Gousseau. Circular Earth Mover's Distance for the comparison of local features. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[277] Trout Rader. The existence of a utility function to represent preferences. *The Review of Economic Studies*, pages 229–232, 1963.

[278] PS Ramapraba, MP Chitra, and Prem Kumar. Effective lesion detection of colposcopic images using active contour method. *Biomedical Research*, 2017.

[279] PS RamaPraba and H Ranganathan. Automatic lesion detection in colposcopy cervix images based on statistical features. In *Global Trends in Information Systems and Software Applications*, pages 424–430. Springer, 2012.

[280] PS Ramapraba and H Ranganathan. Performance evolution of various wavelets in cervical lesion detection. *Indian Journal of Computer Science and Engineering*, 4(6), 2013.

[281] Jianfeng Ren, Xudong Jiang, and Junsong Yuan. Noise-resistant local binary pattern with an embedded error-correction mechanism. *IEEE Transactions on Image Processing*, 22(10):4049–4060, 2013.

[282] Robert W Robinson. Counting unlabeled acyclic digraphs. In *Combinatorial mathematics V*, pages 28–43. Springer, 1977.

[283] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[284] Lorenzo Rosasco, Andrea Tacchetti, and Silvia Villa. Regularization by early stopping for online learning algorithms. *stat*, 1050:30, 2014.

[285] Farnaz Rouhbakhsh, Fardad Farokhi, and Kaveh Kangarloo. Effective feature selection for pre-cancerous cervix lesions using artificial neural networks. *International Journal of Smart Electrical Engineering*, 1(3), 2012.

[286] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[287] Ulrich Rückert and Stefan Kramer. Kernel-based inductive transfer. In *Machine Learning and Knowledge Discovery in Databases*, pages 220–233. Springer, 2008.

[288] César San Martin and Sang-Woon Kim, editors. *Virus Texture Analysis Using Local Binary Patterns and Radial Density Profiles*, volume 7042 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2011.

[289] Masakazu Sato, Koji Horie, Aki Hara, Yuichiro Miyamoto, Kazuko Kurihara, Kensuke Tomio, and Harushige Yokota. Application of deep learning to the classification of images from colposcopy. *Oncology letters*, 15(3):3518–3523, 2018.

[290] William W Cohen Robert E Schapire and Yoram Singer. Learning to order things. In *Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference*, volume 10, page 451. MIT Press, 1998.

[291] Philippe Schmid-Saugeon, Jonathan D Pitts, B Kaufman-Howard, Alex Zelenchuk, and Diane M Harper. Time-resolved imaging of cervical acetowhitening. *Draft paper*, 2004.

[292] Michael Schmitt and Laura Martignon. On the complexity of learning lexicographic strategies. *The Journal of Machine Learning Research*, 7:55–83, 2006.

[293] Bernhard Schölkopf, Kah-Kay Sung, Chris JC Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio, and Vladimir Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *Signal Processing, IEEE Transactions on*, 45(11):2758–2765, 1997.

[294] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[295] David Sculley. Combined regression and ranking. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 979–988. ACM, 2010.

[296] John W Sellors and Rengaswamy Sankaranarayanan. *Colposcopy and treatment of cervical intraepithelial neoplasia: a beginner's manual*. Diamond Pocket Books (P) Ltd., 2003.

[297] Ashis SenGupta and Supratik Roy. A simple classification rule for directional data. In *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*, pages 81–90. Springer, 2005.

[298] Ashis Sengupta and Fidelis I Ugwuowo. A classification method for directional data with application to the human skull. *Communications in Statistics—Theory and Methods*, 40(3):457–466, 2011.

[299] Ana F Sequeira, Joao C Monteiro, Ana Rebelo, and Hélder P Oliveira. Mobbio: a multimodal database captured with a portable handheld device. In *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, volume 3, pages 133–139. IEEE, 2014.

[300] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical programming*, 127(1):3–30, 2011.

[301] Caifeng Shan. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters*, 33(4):431–437, 2012.

[302] Linda Shapiro and George C Stockman. Computer vision. 2001. *ed: Prentice Hall*, 2001.

[303] Lavanya Sharan, Ruth Rosenholtz, and Edward Adelson. Material perception: What can you see in a brief glance? *Journal of Vision*, 9(8):784–784, 2009.

[304] Neeraj Sharma and Lalit M Aggarwal. Automated medical image segmentation techniques. *Journal of medical physics/Association of Medical Physicists of India*, 35(1):3, 2010.

[305] Yue Shi, Martha Larson, and Alan Hanjalic. List-wise learning to rank with matrix factorization for collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 269–272. ACM, 2010.

[306] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural

networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[307] Caroline Silva, Thierry Bouwmans, and Carl Frélicot. An extended center-symmetric local binary pattern for background modeling and subtraction in videos. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP 2015*, 2015.

[308] Daniel L Silver, Ryan Poirier, and Duane Currie. Inductive transfer with context-sensitive neural networks. *Machine Learning*, 73(3):313–336, 2008.

[309] Patrice Y Simard, David Steinkraus, John C Platt, et al. Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR*, volume 3, pages 958–962, 2003.

[310] Priscyla W Simões, Narjara B Izumi, Ramon S Casagrande, Ramon Venson, Carlos D Veronezi, Gustavo P Moretti, Edroaldo L da Rocha, Cristian Cechinel, Luciane B Ceretta, Eros Comunello, et al. Classification of images acquired with colposcopy using artificial neural networks. *Cancer informatics*, 13:119, 2014.

[311] Vikrant Bhar Singh, Nalini Gupta, Raje Nijhawan, Radhika Srinivasan, Vanita Suri, Arvind Rajwanshi, et al. Liquid-based cytology versus conventional cytology for evaluation of cervical pap smears: experience from the first 1000 split samples. *Indian Journal of Pathology and Microbiology*, 58(1):17, 2015.

[312] Laura Slaughter, Carl RV Brown, Sharon Crowley, and Roxy Peck. Patterns of genital injury in female sexual assault victims. *American journal of obstetrics and gynecology*, 176(3):609–616, 1997.

[313] Dezhao Song, Edward Kim, Xiaolei Huang, Joseph Patruno, Héctor Muñoz-Avila, Jeff Heflin, L Rodney Long, and Sameer Antani. Multimodal entity coreference for cervical dysplasia diagnosis. *IEEE transactions on medical imaging*, 34(1):229–245, 2015.

[314] Suvrit Sra. A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of i s (x). *Computational Statistics*, 27(1):177–190, 2012.

[315] Yeshwanth Srinivasan, Enrique Corona, Brian Nutter, Sunanda Mitra, and Sonal Bhattacharya. A unified model-based image analysis framework for automated detection of precancerous lesions in digitized uterine cervix images. *IEEE Journal of Selected Topics in Signal Processing*, 3(1):101–111, 2009.

[316] Yeshwanth Srinivasan, Dana Hernes, Bhakti Tulpule, Shuyu Yang, Jiangling Guo, Sunanda Mitra, Sriraja Yagneswaran, Brian Nutter, Jose Jeronimo, Benny Phillips, et al. A probabilistic approach to segmentation and classification of neoplasia in uterine cervix images using color and geometric features. In *Medical Imaging*, pages 995–1003. International Society for Optics and Photonics, 2005.

[317] Yeshwanth Srinivasan, Brian Nutter, Sunanda Mitra, Benny Phillips, and Eric Sinzinger. Classification of cervix lesions using filter bank-based texture mode. In *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, pages 832–840. IEEE, 2006.

[318] Yeshwanth Srinivasan, Shuyu Yang, Brian Nutter, Sunanda Mitra, Benny Phillips, and Rodney Long. Challenges in automated detection of cervical intraepithelial neoplasia. In *Medical Imaging*, pages 65140F–65140F. International Society for Optics and Photonics, 2007.

[319] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

[320] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[321] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[322] Thinprep. `https://healthdxs.com/en/thinprep/`, 2018. [Online; accessed 21-February-2018].

[323] Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. Learning categories from few examples with multi model knowledge transfer. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):928–941, 2014.

[324] M Traversi, M Falagario, and C Guaragnella. CADdy–Colposcopy Learning Machine for Computer Aided Diagnosis. In *Consumer Electronics. Berlin (ICCE-Berlin), 2013. ICCEBerlin 2013. IEEE Third International Conference on*, pages 1–4. IEEE, 2013.

[325] Jirí Trefnỳ and Jirí Matas. Extended set of local binary patterns for rapid object detection. In *Computer Vision Winter Workshop*, pages 1–7, 2010.

[326] Athanasios Tsanas, Max A Little, Patrick E McSharry, and Lorraine O Ramig. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *Biomedical Engineering, IEEE Transactions on*, 57(4):884–893, 2010.

[327] Bhakti Tulpule, Shuyu Yang, Yeshwanth Srinivasan, Sunanda Mitra, and Brian Nutter. Segmentation and classification of cervix lesions by pattern and texture analysis. In *The 14th IEEE International Conference on Fuzzy Systems, 2005. FUZZ'05.*, pages 173–176. IEEE, 2005.

[328] Viara Van Raad. Image analysis and segmentation of anatomical features of cervix uteri in color space. In *Signal Processing Conference, 2005 13th European*, pages 1–4. IEEE, 2005.

[329] Viara Van Raad. A new vision approach for local spectrum features in cervical images via 2D method of geometric restriction in frequency domain. In *Computer Vision for Biomedical Image Applications*, pages 125–134. Springer, 2005.

[330] Viara Van Raad, Zhiyun Xue, and Holger Lange. Lesion margin analysis for automated classification of cervical cancer lesions. In *Medical Imaging*, pages 614454–614454. International Society for Optics and Photonics, 2006.

[331] Maria João M Vasconcelos, Luís Rosado, and Márcia Ferreira. Principal axes-based asymmetry assessment methodology for skin lesion image analysis. In *International symposium on visual computing*, pages 21–31. Springer, 2014.

[332] Ching-Wei Wang, Cheng-Ta Huang, Jia-Hong Lee, Chung-Hsing Li, Sheng-Wei Chang, Ming-Jhih Siao, Tat-Ming Lai, Bulat Ibragimov, Tomaž Vrtovec, Olaf Ronneberger, et al. A benchmark for comparison of dental radiography analysis algorithms. *Medical image analysis*, 31:63–76, 2016.

[333] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.

[334] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, and Tie-Yan Liu. A theoretical analysis of NDCG ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013)*. Citeseer, 2013.

[335] Gang Wu and Ey Chang. Class-boundary alignment for imbalanced dataset learning. *The Twentieth International Conference on Machine Learning (ICML)*, pages 49–56, 2003.

[336] Meng Xi, Liang Chen, Desanka Polajnar, and Weiyang Tong. Local binary pattern network: a deep learning approach for face recognition. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 3224–3228. IEEE, 2016.

[337] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *Advances in Neural Information Processing Systems*, pages 341–349, 2012.

[338] Jun Xiong, Lei Wang, and Jia Gu. Image segmentation of the acetowhite region in cervix images based on chromaticity. In *2009 9th International Conference on Information Technology and Applications in Biomedicine*, pages 1–4. IEEE, 2009.

[339] Haiyong Xu, Kevin Nichols, and Frederic P Schoenberg. Kernel regression of directional data with application to wind and wildfire data in los angeles county, california. *Forest Science*, 57(4):343–352, 2011.

[340] Jun Xu and Hang Li. Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 391–398. ACM, 2007.

[341] Tao Xu, Xiaolei Huang, Edward Kim, L Rodney Long, and Sameer Antani. Multi-test cervical cancer diagnosis with missing data estimation. In *SPIE Medical Imaging*, pages 94140X–94140X. International Society for Optics and Photonics, 2015.

[342] Tao Xu, Edward Kim, and Xiaolei Huang. Adjustable adaboost classifier and pyramid features for image-based cervical cancer diagnosis. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 281–285. IEEE, 2015.

[343] Tao Xu, Cheng Xin, L Rodney Long, Sameer Antani, Zhiyun Xue, Edward Kim, and Xiaolei Huang. A new image data set and benchmark for cervical dysplasia classification evaluation. In *International Workshop on Machine Learning in Medical Imaging*, pages 26–35. Springer, 2015.

[344] Tao Xu, Han Zhang, Xiaolei Huang, Shaoting Zhang, and Dimitris N Metaxas. Multimodal deep learning for cervical dysplasia diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 115–123. Springer, 2016.

[345] Tao Xu, Han Zhang, Cheng Xin, Edward Kim, L Rodney Long, Zhiyun Xue, Sameer Antani, and Xiaolei Huang. Multi-feature based benchmark for cervical dysplasia classification evaluation. *Pattern recognition*, 63:468–475, 2017.

[346] Gengjian Xue, Li Song, Jun Sun, and Meng Wu. Hybrid center-symmetric local pattern for dynamic background subtraction. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1–6. IEEE, 2011.

[347] Gengjian Xue, Jun Sun, and Li Song. Dynamic background subtraction based on spatial extended center-symmetric local binary pattern. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1050–1054. IEEE, 2010.

[348] Zhiyun Xue, Sameer Antani, L Rodney Long, Jose Jeronimo, and George R Thoma. Comparative performance analysis of cervix roi extraction and specular reflection removal algorithms for uterine cervix image analysis. In *Medical Imaging*, pages 65124I–65124I. International Society for Optics and Photonics, 2007.

[349] Zhiyun Xue, Sameer Antani, L Rodney Long, and George R Thoma. An online segmentation tool for cervicographic image analysis. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 425–429. ACM, 2010.

[350] Zhiyun Xue, L Rodney Long, Sameer Antani, and George R Thoma. Automatic extraction of mosaic patterns in uterine cervix images. In *Computer-Based Medical Systems (CBMS), 2010 IEEE 23rd International Symposium on*, pages 273–278. IEEE, 2010.

[351] Jianwei Yang, Shizheng Wang, Zhen Lei, Yanyun Zhao, and Stan Z Li. Spatio-temporal lbp based moving object segmentation in compressed domain. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 252–257. IEEE, 2012.

[352] Liu Yang, Steve Hanneke, and Jaime Carbonell. A theory of transfer learning with applications to active learning. *Machine learning*, 90(2):161–189, 2013.

[353] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[354] Lahav Yeffet and Lior Wolf. Local trinary patterns for human action recognition. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 492–497. IEEE, 2009.

[355] I-Cheng Yeh, King-Jang Yang, and Tao-Ming Ting. Knowledge discovery on RFM model using bernoulli sequence. *Expert Systems with Applications*, 36(3):5866–5871, 2009.

[356] Haiyan Yin, Hua Yang, Hang Su, and Chongyang Zhang. Dynamic background subtraction based on appearance and motion pattern. In *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013.

[357] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.

[358] Yang Yu, Junzhou Huang, Shaoting Zhang, Christophe Restif, Xiaolei Huang, and Dimitris Metaxas. Group sparsity based classification for cervigram segmentation. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1425–1429. IEEE, 2011.

[359] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[360] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[361] Richard S Zemel, Christopher KI Williams, and Michael C Mozer. Lending direction to neural networks. *Neural Networks*, 8(4):503–512, 1995.

[362] Baochang Zhang, Yongsheng Gao, Sanqiang Zhao, and Jianzhuang Liu. Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor. *IEEE transactions on image processing*, 19(2):533–544, 2010.

[363] Jian Zhang, Zoubin Ghahramani, and Yiming Yang. Flexible latent variable models for multi-task learning. *Machine Learning*, 73(3):221–242, 2008.

[364] Shaoting Zhang, Junzhou Huang, Wei Wang, Xiaolei Huang, and Dimitris Metaxas. Cervigram image segmentation based on reconstructive sparse representations. In *SPIE Medical Imaging*, pages 762313–762313. International Society for Optics and Photonics, 2010.

[365] Shengping Zhang, Hongxun Yao, and Shaohui Liu. Dynamic background modeling and subtraction using spatio-temporal local binary patterns. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 1556–1559. IEEE, 2008.

[366] Zi-qian Zhang, Tie-xiang Wen, and Jia Gu. Cervical cancerous region detection using spectral matting technique: A pilot study on color space. *Bulletin of Advanced Technology Research*, 2011.

[367] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6), 2007.

[368] Yang Zhao, Wei Jia, Rong-Xiang Hu, and Hai Min. Completed robust local binary pattern for texture classification. *Neurocomputing*, 106:68–76, 2013.

[369] Yu-qian Zhao, Irene J Chang, Fang-hui Zhao, Shang-ying Hu, Jennifer S Smith, Xun Zhang, Shu-min Li, Ping Bai, Wen-hua Zhang, and You-lin Qiao. Distribution of cervical intraepithelial neoplasia on the cervix in chinese women: pooled analysis of 19 population based screening studies. *BMC cancer*, 15(1):485, 2015.

[370] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.

[371] Zhaohui Zheng, Keke Chen, Gordon Sun, and Hongyuan Zha. A regression framework for learning ranking functions using relative relevance judgments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 287–294. ACM, 2007.

[372] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

[373] Chao Zhu, Charles-Edmond Bichot, and Liming Chen. Multi-scale color local binary patterns for visual object classes recognition. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3065–3068. IEEE, 2010.

[374] Gali Zimmerman, Shiri Gordon, and Hayit Greenspan. Automatic landmark detection in uterine cervix images for indexing in a content-retrieval system. In *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006.*, pages 1348–1351. IEEE, 2006.