Prostate Histopathology Image Classification and Retrieval using Weakly-Supervised Multimodal Fusion and Representation Learning

Juan Sebastián Lara Ramírez

Universidad Nacional de Colombia
Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2020

Prostate Histopathology Image Classification and Retrieval using Weakly-Supervised Multimodal Fusion and Representation Learning

Juan Sebastián Lara Ramírez

Submitted to the Engineering School of the Universidad Nacional de Colombia, in partial fulfillment of the requirements of the degree of:
Master in Systems and Computer Engineering

Advisor:
Fabio A. González Ph.D.

Research Area:
Machine Learning
Research Group:
MindLab Research Group

Universidad Nacional de Colombia
Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2020

Dedication

To my family Blanca and Patricia

Acknowledgements

I want to express my gratitude to my family, Blanca Ramírez and Patricia Ramŕez. They have always been there for me, and this would not be possible without their unconditional support. I want to thank my advisor Fabio González for his guidance and encouragement. Also, I appreciate the collaboration of all the members from the MindLab research group, with whom I was able to share knowledge and novel scientific ideas. Finally, I want to thank Universidad Nacional de Colombia for funding support, infrastructure, and education.
This work was supported by COLCIENCIAS with the research project "Detección temprana de daño ocular en diabéticos usando un sistema de inteligencia artificial en imágenes de fondo de ojo"number 1101-807-63563 and the Universidad Nacional de Colombia with the 2017-2018 national call for the project Çlasificación de retinopatía diabética y edema macula diabético en imágenes de fondo de ojo mediante redes neuronales convolucionales "number 202010029118.

Abstract

This thesis presents an information fusion strategy for the automatic classification and retrieval of prostate histopathology whole-slide images (WSIs) that incorporates novel machine learning components from deep learning and kernel methods. Its main purpose is to enhance the representation of the WSIs using additional text content extracted from diagnosis reports. This is achieved using the **multimodal latent semantic alignment (M-LSA)** model, which employs a weakly-multimodal-supervised methodology that incorporates text information during the model training to enrich the representation of the WSIs with complementary semantic information. Besides, M-LSA does not require the text data during the prediction phase, which makes it suitable for realistic scenarios where a pathologist may only have the image data. The experimental evaluation demonstrates that the weakly-supervised multimodal enhancement has a significant improvement in the performance during classification and retrieval, further, the proposed model outperforms the state–of–the–art unimodal and multimodal baselines in automatic prostate cancer assessment.

Keywords: Information Fusion, Histopathology Images, Representation Learning, Kernel Methods, Weakly-Supervision, Multimodal Learning.

Resumen

Esta tesis presenta una estrategia de fusión de información para la clasificación y recuperación automática de imágenes de histopatología de próstata incorporando novedosos compenentes de aprendizaje de máquina y aprendizaje profundo. El propósito de la estrategia es mejorar la representación de las imágenes con contenido textual adicional que es extraído de reportes de diagnóstico. Para lograr esto, se propone el modelo **multimodal latent semantic alignment (M-LSA)**, el cual emplea una metodología de supervisión multimodal débil que incorpora información textual durante el entrenamiento para enriquecer la representación de las imágenes con información semántica complementaria. Adicionalmente, M-LSA no requiere la modalidad textual durante la fase de predicción, por lo que el modelo es apropiado para escenarios más realistas donde un patólogo puede tener sólo las imágenes. La evaluación experimental muestra que el enriquecimiento por supervisión débil multimodal presenta una mejora significativa en el despempeño durante clasificación y recuperación, además, el método propuesto supera otros enfoques unimodales y multimodales en el estado del arte del análisis automático de cáncer de próstata.

Palabras clave: Fusión de Información, Imágenes de Histopatología, Aprendizaje de la Representación, Métodos de Kernel, Supervisión Débil, Aprendizaje Multimodal

Esta tesis de maestría se sustentó el 30 de octubre de 2020 a las 8:00am, y fue evaluada por los siguientes jurados:

Ángel Alfonso Cruz Roa (PhD.)
Profesor Facultad de Ingeniería.
Universidad de los Llanos.

Jair Eduardo Rocha González (PhD.)
Profesor Facultad de Medicina.
Universidad Nacional de Colombia.

Contents

Acknowledgements vii

Abstract ix

1 Introduction 2
- 1.1 Problem Statement . . . 3
- 1.2 Objectives . . . 4
  - 1.2.1 General Objective . . . 4
  - 1.2.2 Specific Objectives . . . 4
- 1.3 Main Contributions . . . 4
- 1.4 Thesis Organization . . . 5

2 Related Work 6
- 2.1 Prostate Histopathology Image Classification . . . 6
- 2.2 Prostate Histopathology Image Retrieval . . . 7
- 2.3 Multimodal Medical and Histopathology Data Analysis . . . 9

3 Multimodal Latent Semantic Alignment 11
- 3.1 Data Representation . . . 11
- 3.2 Fusion Strategy . . . 13
- 3.3 Experimental Settings . . . 15
  - 3.3.1 Dataset Description . . . 15
  - 3.3.2 Cancer Detection Performance . . . 15
  - 3.3.3 Image Retrieval Performance . . . 16
  - 3.3.4 Hyperparameter Selection . . . 16
- 3.4 Results and Analysis . . . 16
- 3.5 Conclusion . . . 18

4 Extended Multimodal Latent Semantic Alignment 19
- 4.1 Improved Visual Representation . . . 19
- 4.2 Improved Text Representation . . . 22
- 4.3 Weak Multimodal Supervision . . . 23
- 4.4 Experimental Setup . . . 24
  - 4.4.1 Weak supervision . . . 24

4.4.2 BoVW representation 25
4.4.3 BoN representation 26
4.4.4 Weak multimodal supervision 26
4.5 Results and Analysis 27
4.5.1 Weak Supervision 27
4.5.2 BoVW representation 28
4.5.3 BoN representation 29
4.5.4 Weak multimodal supervision 30
4.6 Conclusion 33

5 Conclusions and future work **34**
5.1 Histopathology Image Classification 34
5.2 Histopathology Image Retrieval 34
5.3 Multimodal Learning 35

References **36**

List of Figures

3-1 Overview of the training and prediction phases of the proposed method for the automatic classification of prostate WSIs. 12
3-2 Conceptual diagram of the multimodal information fusion strategy. 13
3-3 Example data from the TCGA-PRAD, a) whole-slide image; b) surgical pathology report. 15

4-1 Learning procedure of eM-LSA 20
4-2 Training procedure for a CNN model using weak supervision and summarization. 21
4-3 Average TF-IDF representation of the top 50 terms. 26
4-4 Comparison of the evaluation metrics for different codebook sizes for the BoVW representation with and without TF-IDF weighting schema. 29
4-5 Comparison of the evaluation metrics for different N-grams for the BoN representation with and without TF-IDF weighting schema. 30

List of Tables

3-1 Comparison with state-of-the art methods in cancer detection. 17
3-2 Results for the retrieval task, * denotes cases with multimodal queries. . . . 18

4-1 Number of patches per class for each partition. 24
4-2 Performance comparison for the implemented convolutional neural networks. 25
4-3 Performance of different CNN models for high and low Gleason score classification using weak supervison. 28
4-4 Comparison of the classification performance between eM-LSA and eV-LSE. 31
4-5 Comparison with the state-of-the-art methods in the classification task using the improved representations. 32
4-6 Comparison of the retrieval performance between eM-LSA and eV-LSE. . . . 32
4-7 Comparison with the state-of-the-art methods in the retrieval task using the improved representations, the cases with * require multimodal queries. . . . 33

1 Introduction

Prostate cancer (PCa) is the fourth most common cancer worldwide with 1.2 million new cases in 2018 and it has the second-highest incidence of all cancers in men [1]. Currently, the Gleason score (GS) is the standard grading system used to determine the aggressiveness of PCa and determine treatment. Typical scores range from 6 to 10 and cases with higher values are more likely to grow and spread fast [2]. The gold standard for the diagnosis of PCa is the inspection of biopsies or tissue samples. Thanks to the recent improvements in digital microscopy, the diagnosis is increasingly made through the visual inspection of high-resolution scans of a tissue sample or a whole-slide image (WSI) [3]. Digital pathology is focused on the acquisition, management and interpretation of this kind of data, it includes tasks like digital scanning, visualization, computer-assisted diagnosis (CAD), medical image processing, among others [4, 5]. Collections of WSIs and related information like pathology reports can be accessed and stored using picture archiving and communication systems (PACS), this preserves the data in the long term providing a valuable clinical information source. CAD is one the most studied tasks in digital pathology, it generally covers tasks such as the automatic classification or grading of a disease, segmentation of regions of interest, mitosis and necrosis detection, image retrieval, among others [6].

A current problem in the assessment of PCa is the misdiagnosis due to the inter-reader variability, this involves an untreated cancer patient or an unnecessary cost for the treatment of a healthy person. This variability is mainly caused by the high disease heterogeneity in PCa, which could produce several tumor types [7]. Although different expert readers could achieve a substantial agreement, newer readers may misdiagnose and require supervision from a more experienced person. In this regard, a CAD system is a useful tool that addresses the cognitive bias in pathological interpretation, there is evidence that errors made by a computer differ from the pathologists' errors [8]. Therefore, misdiagnoses could be addressed using an automated grading system that supports pathologists providing a second opinion or a retrieval system that finds meaningful cases in medical databases to support pathologists' decisions during uncertain diagnostics, this additional information would allow observing how other experts diagnosed similar cases and the evolution of the patients during and after a specific treatment.

Databases of medical images usually contain additional text data that is often not used by CAD systems [9]. There are diagnostic reports, clinical and related metadata that can be used to improve the performance of current CAD systems. The text modality usually contains semantic content that complements the information in the images. However, a current

challenge is related to the appropriate combination of the image and the text information, especially, considering that these modalities originate from different sources and therefore have different statistical properties [10]. Besides, current multimodal approaches consider additional modalities as inputs for a model, this is a highly-restrictive approach that also requires the text data for new predictions, which is unpractical since it would bias a classification model to the pathologist's decisions or a retrieval model to the uncertain beliefs. On this basis, new methods that exploit unused multimodal data while avoiding cognitive bias are required and must be further studied.

1.1. Problem Statement

The main aim of this thesis is to develop a method that incorporates multimodal information to improve the performance of an image model in the tasks of prostate tissue classification and case-based image retrieval. An appropriate combination of multimodal information remains a challenge since the data of each modality comes from different sources and therefore has different statistical properties [10]. In this regard, recent studies have demonstrated the importance of multimodal learning in the analysis of medical data, especially, multimodal approaches have shown to outperform single-modal image and text approaches [11, 12, 13]. Current related approaches aim to exploit the joint information in multimodal histopathology data, nonetheless, one of their main disadvantages is that they are unfeasible in certain scenarios where a pathologist may only have an image, because, these approaches also require multimodal inputs or queries during the prediction or retrieval of new cases. Considering this, we formulate the following main research question:

- *How to properly incorporate text information to enhance a single-modal model for histopathology whole-slide image automatic analysis?.*

The data from different modalities are often complementary, i.e., all the sources capture information of the same phenomena but there are specific patterns that are only captured in a specific modality. This behavior can be ideally represented using a joint distribution of the sources, which conventional multimodal approaches aim to compute or approximate. However, this distribution cannot be easily estimated when there is information from a single modality only. This makes it harder to design a method that exploits the multimodal information in a weakly-supervised manner, i.e., using low-quality information from other modalities. Nonetheless, we hypothesize that this can be achieved by dealing with two main factors: on the one hand, suitable representation techniques must be used to transform the data into simpler representations that ease the weakly-supervised multimodal fusion; on the other, a proper fusion strategy must allow enhancing the independent representations of each modality instead of enhancing a joint and combined representation.
To address the main research question and the two aforementioned factors, it is necessary to answer the following research questions:

- How to learn a suitable representation of the WSIs that can be improved with text information?
- How to represent the text data to enhance the visual representation?
- How must be the fusion strategy to enhance the independent representation of the images?

1.2. Objectives

Based on these questions, this research has the following goals:

1.2.1. General Objective

- To develop a weakly-supervised multimodal method that incorporates text information to enrich a single-modal model for prostate histopathology image automatic analysis.

1.2.2. Specific Objectives

- To determine an appropriate representation for the WSIs using feature learning.
- To determine a compatible representation for the texts that complement the visual representation.
- To develop an information fusion strategy that allows a weakly-supervised multimodal enhancement of the visual representation.
- To systematically evaluate the effects of the weakly-supervised multimodal fusion in tasks like classification and case-based retrieval on prostate histopathology data.

1.3. Main Contributions

This work presents a novel weakly-multimodal-supervised model for prostate histopathology images. The main contributions of this work are outlined as follows:

- Exploration of different representation learning strategies for prostate histopathology images, achieving state-of-the-art results in the high-and-low classification of prostate histopathology images.
- Exploration of an appropriate representation for the diagnosis reports that eases the enhancement of the visual representation.

- Formulation of a weakly-multimodal-supervised model and a novel training strategy that automatically incorporates semantic information in the visual representation.
- Systematic evaluation of the proposed model on multimodal prostate histopathology data, demonstrating that the weakly-multimodal enhancement outperforms unimodal approaches and is competitive with conventional multimodal strategies.

The following is a list of papers that have been published or accepted in international conferences:

- **Juan S. Lara**, Victor H. Contreras O., Sebastián Ótalora, Henning Müller, Fabio A. González, "Multimodal Latent Semantic Alignment for Automated Prostate Tissue Classification and Retrieval". Manuscript accepted for publication in Medical Image Computing and Computer-Assisted Intervention - MICCAI (2020).
- Victor H. Contreras O., **Juan S. Lara**, Oscar Perdomo, Fabio A. González, "Supervised online matrix factorization for histopathological multimodal retrieval". International Symposium on Medical Information Processing and Analysis (2018).

For reproducibility, we made public the source code with the method implementation and the replication code, also, we shared the preprocessed dataset and the trained models:

- TensorFlow 2.0 implementation in Python of M-LSA and replication code for the experiments: `https://github.com/larajuse/MLSA`.
- Preprocessed TCGA-PRAD dataset: `https://drive.google.com/drive/folders/14pbie6QsN64i0ArpfOnpiyEtFDX8LWqS?usp=sharing`
- Weights and additional metadata of the trained models: `https://drive.google.com/drive/folders/1WPkJ_aCGHnA4SQf_9lmIB_GhBMP92NHH?usp=sharing`

1.4. Thesis Organization

The remaining chapters of this document are organized as follows. Chapter 2 presents a review of the previous related works, it gives an overall background on classification, retrieval, and multimodal analysis of medical and prostate histopathology images. Chapter 3 presents the Multimodal Latent Semantic Alignment (M-LSA), detailing the fusion strategy and demonstrating the effects of the weakly-multimodal-supervision. Chapter 4 presents an improved version of M-LSA that consists of a systematic evaluation of the representation techniques, showing that the proposed model outperforms current state-of-the-art methods. Finally, Chapter 5 summarizes the main findings of this research and provides future research directions on the multimodal data analysis of histopathology images.

2 Related Work

The main purpose of this research is to exploit this kind of multimodal data to improve the performance of current CAD systems, specifically, in the automatic classification and case-based retrieval of prostate tissue images. For this reason in the next sections will be presented background about these tasks, as well as, an overview of the multimodal analysis of medical and histopathology images.

2.1. Prostate Histopathology Image Classification

The automatic grading of histopathology images can be seen as a supervised learning problem, more precisely, a classification model is used to predict the grade of the disease using relevant patterns from the images. In recent years, image processing and machine learning methods have been successfully applied for the automatic grading of different tissue samples, including prostate cancer, breast cancer, colorectal cancer, brain tumors, lung cancer, among others [14]. Several approaches have been developed for the automatic grading of prostate tissues, conventional methods follow a typical image processing pipeline that consists of three main steps:

Preprocessing, it is related to normalization in the images, removal of noise, or any other artifact that hinders the classification. A common strategy is to normalize color and illumination, which reduces the differences between images due to variations during the scanning or the staining [15, 16, 17]. A color deconvolution is usually performed over H&E stained images to separate a WSI into an image for Hematoxylin and an image for Eosin [18, 19]. Finally, several preprocessing strategies aim to extract regions of interest (ROI) that are more related to the GS prediction, for instance, extracting areas around the nuclei, glands, or the lumen [17, 20].

Feature extraction, it is a task that was originally known as feature engineering and aims to determine a set of discriminative features to represent the histopathology images. A typical strategy is to calculate general-purpose image descriptors like multi-wavelet, Gabor, morphological, statistical, and texture features [15]. Other approaches use the ROIs to extract domain-specific descriptors, more precisely, a set of features is computed to capture the properties of the tissue structures [21, 22, 18], for example, distributional and morphological descriptors like the gland area, nuclei area, lumen area, roundness or boundary smoothness, the number of nuclei, among others [21, 22].

Classification, the extracted features are used to train any kind of classification model, for instance: a support vector machine [19, 21, 23], a Bayesian model [24] or boosting [25]. In this regard, two main classification approaches are used for PCa assessment. On the one hand, the WSIs are used for cancer detection, i.e., it is a binary problem in which the goal is to classify between low (6 and 7) and high (8, 9, and 10) GS [20, 16]. On the other hand, the WSIs are used for the automatic grading of PCa, it is a multi-class problem where the Gleason grade must be predicted [26, 27].
In recent years, representation learning has emerged as an important alternative to conventional image processing methods. Convolutional neural networks (CNNs) have become the gold standard in fields like image processing and computer vision, and its application is a current topic of interest in digital pathology [27]. CNNs are being studied for the automatic analysis of prostate tissue images, these kind of models can automatically learn multiple representations from the raw images without the need of feature engineering and with a minimal preprocessing [28]. Currently, training a CNN using the complete WSIs is computationally unfeasible due to the large size of these images, however, there is evidence that the cancer types can be differentiated through cellular-level features that can be detected in a local tissue level [29].
For this reason, a common approach is to train patch-level CNNs to predict local patterns and then aggregate or summarize the information for a global prediction. In this matter, several CNNs architectures have been used for the classification of prostate WSIs, including: LeNet, AlexNet, GoogLeNet [20], OverFeat [28], VGG11 [30] or Inception V3 [26]. These CNNs are usually trained to identify Gleason Patterns (patch-level predictions) and the slice prediction is obtained from a majority vote [20, 16], ensembles [26] or different classifiers that use the predictions or intermediate representations from the CNNs [27, 30]. One of the main advantages of deep-learning-based CAD is that there is evidence that these kinds of systems are comparable to international pathologist experts in cancer detection and grading [26], however, finding appropriate representations and an intelligent combination of the patch-level representations is still a challenge [29].

2.2. Prostate Histopathology Image Retrieval

A medical information retrieval system is a valuable technology when searching for specific medical data, its main purpose is to obtain clinically relevant information from a medical collection. The retrieval of histopathology images can be performed under different paradigms. The most common approach is to use keywords as queries for text-based retrieval. Nonetheless, this approach is highly-subjective and biased since it requires additional semantic abstractions of the images, thus, factors such as expertise or interpretability could affect the search [31]. An alternative is a query-by-example paradigm or content-based image retrieval (CBIR), in which, the query is an example image without additional semantic information and the retrieval is made using low-level feature descriptors. It is unbiased since it does not

depend on subjective queries, however, the performance of CBIR systems is generally lower in comparison with text-based systems [32], this occurs when the image descriptors are not able to capture high-level semantics of images and is known as the semantic gap [11].

A conventional pipeline in CBIR consists on two main components: (1) quantitative descriptors are extracted from the image to obtain a simpler vector representation and then (2) a similarity measure is calculated between the descriptors of each image in the database and the query descriptor, this procedure allows building a rank and inducts an order about the most similar cases in the database [33]. The similarity measure is generally a distance metric and is selected according to the nature of the representation, for instance, the Euclidian distance is used in real and continuous spaces, the Hamming distance is used in binary spaces and the cosine distance is used in probabilistic spaces. Furthermore, the complexity in a CBIR approach is usually covered in the descriptors, i.e., the representation technique must allow embedding the images' information into a simple vector space where linear relationships can be easily found.

Determining a representation technique or an embedding is a difficult task due to the complex nature of the histopathology images. However, dimensionality reduction techniques are preferred because they allow to avoid the curse of dimensionality and allow compact representations that reduce the computational cost when calculating pairwise distances [33]. In this matter, several methods have been proposed for CBIR in PCa assessment, initial approaches used matrix factorization strategies such as principal component analysis or non-negative matrix factorization [33]; however, most of these methods only capture linear correlations which may not be appropriate considering the nature of this data.

Recent studies explored non-linear embedding strategies that build low-dimensional spaces while preserving original adjacencies [34], for example, the out-of-sample extrapolation utilizing semi-supervised manifold learning (OSE-SSL) allows to learn a non-linear representation without a direct computation of singular value decomposition for each query image [33]; the boosted spectrally embedding (BoSE) also uses manifold learning, but, it uses a probabilistic approach that leverages from a boosting strategy and allows to weight the importance of the original features [35, 34]; the explicit shape descriptors (ESDs) learn a morphological similarity that uses fuzzy K-Means for quantitative shape modeling, this allows a representation of the images through morphological descriptors that can be used for retrieval [36].

These methods have demonstrated the importance of non-linear dimensionality reduction in the retrieval of prostate histopathology images, however, they still require an additional representation step that consists of feature engineering. In this regard, there is an interest in the application of deep learning, especially, representation learning has demonstrated to outperform typical computer vision descriptors in CBIR. Some CNN architectures like the deep rank network [37], GoogLeNet [9], or DenseNet [38] have been used for the retrieval of PCa cases and have shown impressive results. Furthermore, representation learning has not been fully explored in the specific case of retrieving prostate histopathology images.

2.3. Multimodal Medical and Histopathology Data Analysis

Databases of medical images usually contain additional text data like diagnostic reports, genomics, clinical and related metadata. This information can be used to improve the performance of current CAD systems using additional content that usually complements the information in the images. However, an appropriate combination of the image and the text information is a challenging task, especially, considering that these modalities originate from different sources and therefore have different statistical properties [10]. Multimodal fusion is an approach that aims to combine the information from different modalities and its application in the medical domain is an active research area.

It is possible to identify three types of information fusion in the medical domain [11, 12]: *early fusion*, in which the fusion is performed at the feature level, i.e., the descriptors of each modality are combined into a single descriptor; *late fusion*, which is used to combine the decisions made from each modality; *hybrid methods*, which integrate strategies from early and late fusion. Related studies in medical data show that an appropriate combination of these two modalities provides better overall performance in comparison with single modal approaches. An exhaustive examination of this behavior has been conducted in recent years [13], especially, the popularization of competitions like ImageCLEF and its medical task ImageCLEFmed have allowed evaluating the effects of different information fusion strategies in the performance of medical image retrieval systems.

Regarding histopathology data, a major concern is that text data is often not used in current CAD systems [9] and the application of multimodal fusion is a current research problem. In this matter, some fusion strategies have been proposed for automatic analysis of histopathology images: Chen et al. [39] proposed the panthomic fusion for the automatic grading and prognosis prediction of glioblastoma multiforme and low-grade glioma cases, it uses a VGG19 network to represent the WSIs, a graph convolutional network based on GraphSAGE to represent the cell spatial distribution and a fully-connected network to represent genomic data, these representations are weighted using a gated-based attention mechanism and a multimodal descriptor is obtained using the Kronecker product. Yan et al. [40] proposed a hybrid deep learning method for breast cancer classification, it uses multiple intermediate representations from a VGG16 network to build a richer representation which is concatenated with hand-crafted features that were extracted from electronic medical reports. Weng et al. [41] proposed a multitask model for metadata prediction (staining method and type of fixation, tissue, procedure, and staining), it incorporates a ResNet50 network for patch-level and slide representations of the WSIs, a BERT network to represent pathology reports, and one-hot encodings for case categories, these representations are fused using vector concatenation and compact bilinear pooling.

One of the main disadvantages of these methods is that they are unfeasible in certain scenarios where a pathologist may only have an image, because, these approaches also require

multimodal inputs during the prediction or retrieval of new cases. In this case, it is desired a fusion strategy that enhances the independent representations of each modality instead of enhancing a joint and combined representation. Some approaches aim to address this issue: Caicedo et al. [3, 42] proposed a non-negative matrix factorization method for the multimodal indexing of multiple organ tissues, it aims to induct a shared latent space for all modalities through an iterative optimization process that independently reconstructs each modality in each step. Cheerla et al. [43] proposed an unsupervised deep multimodal representation for pan-cancer prognosis prediction, it uses a fully-connected encoder to represent clinical data, highway networks to represent the genomic and microRNA data, and a modified SqueezeNet to represent the WSIs, these representations are combined using a similarity-based loss function that is based on siamese networks. These approaches show the feasibility of a new kind of weakly-supervised multimodal fusion strategies that can be used to enrich the representation of the histopathology images.

To our knowledge, there are only two studies that exploit multimodal fusion on prostate histopathology images. First, Jimenez-del-toro et al. [9] proposed a deep learning approach for multimodal case-based retrieval, it uses a GoogLeNet to represent the WSIs and a Doc2Vec embedding to represent diagnosis reports, late fusion is used to combine the information and consists on a convex combination of the rank matrices of the two modalities. Second, Contreras et al. [44] followed this line of work and proposed a hybrid fusion approach that combined the information in two levels, the representations are combined in an early stage using a similarity-based approach, and the decisions of each modality are also combined through a convex combination. Finally, weakly-supervised multimodal fusion has not been used for the automatic classification and retrieval of prostate images, this is a promising application considering the impressive results in similar histopathology applications.

3 Multimodal Latent Semantic Alignment

This chapter presents the Multimodal Latent Semantic Alignment (M-LSA) for cancer detection and similar case retrieval using multimodal histopathology data from prostate whole-slide-images (WSIs) and their diagnostic reports. It simultaneously learns an embedded representation for the WSIs and their associated text content and is trained using a weakly-multimodal-supervised fashion that incorporates text information to enhance and align the representation of the images. M-LSA exploits the complementary information of visual and text modalities while preserving independent representations for each modality, this allows us to compute predictions from data that contains images data only while using a model that contains multimodal information. Part of this work was published in the *International Symposium on Biomedical Imaging (ISBI)* [44] and was accepted for publication in the *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* [45].
An overview of the method is shown in Figure **3-1**. During the training phase, the method incorporates weakly-supervised information from the diagnostic reports to enhance an embedded representation of the WSIs. This enhanced representation is used during the prediction phase to obtain a Gleason score (GS) estimation using the information from the images only. The WSIs are represented using a bag-of-visual words (BoVW) approach and the text content is represented using a bag-of-words (BoW). These representations are embedded and aligned using an information fusion strategy that is described in the following subsections.

3.1. Data Representation

As shown in Fig. **3-1**, the training data $\mathcal{D} = \{(\mathbf{I_1}, \mathbf{d_1}, y_1), (\mathbf{I_2}, \mathbf{d_2}, y_2), \ldots, (\mathbf{I_N}, \mathbf{d_N}, y_N)\}$ is composed of pairs of annotated WSIs $\mathbf{I_i}$ and their diagnostic reports $\mathbf{d_i}$. We represent these multimodal data as a term frequency-inverse document frequency (TF-IDF) matrix for each modality. In this way, the representation of the text content is straightforward. The text preprocessing consists of stop-word removal during the text vocabulary T construction. We use the TF-IDF weighting schema because it benefits the information fusion strategy providing numerical stability while increasing the importance of unique terms and attenuating the common ones.

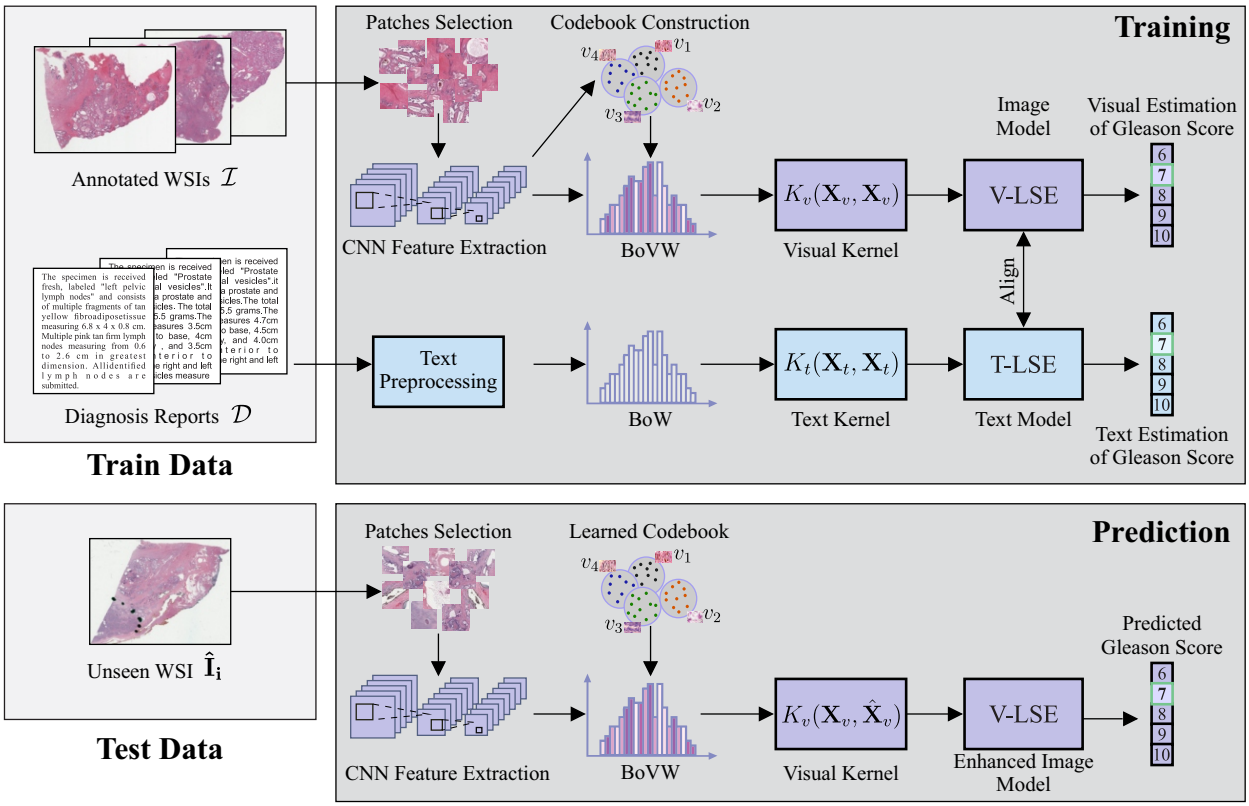

Figure 3-1: Overview of the training and prediction phases of the proposed method for the automatic classification of prostate WSIs.

For the images, a codebook or visual vocabulary $\mathcal{V}$ of size $|\mathcal{V}|$ is constructed to represent a WSI as a bag-of-visual-words (BoVW). The BoVW contains a distribution $P(\mathcal{V} = v_i|\mathcal{I} = \mathbf{I_j})$ of certain visual word v_i in an image $\mathbf{I_j}$. To compute this, 2000 non-overlapping patches p_{ij} of size 256×256 are selected from each WSI using the blue ratio (view Eq. 3-1) as filtering criteria (for obtaining most severe cancer areas) [46] as it is done in [20], where I_R, I_G, I_B are the red, green and blue components respectively.

$$BR(I_R, I_G, I_B) = \frac{100 \cdot I_B}{1 + I_R + I_G} \cdot \frac{256}{1 + I_R + I_G + I_B} \tag{3-1}$$

Then, a feature representation of the patches is computed using the GoogLeNet CNN architecture that was pre-trained for the binary classification of the GS, we use this network because it has demonstrated to outperform other architectures in the automatic diagnostic of prostate tissues [20]. Each patch is described with the feature vector that outputs the last average pooling layer of GoogleNet, which is commonly used for feature extraction. The codebook is constructed using K-means over the CNN descriptors. More precisely, a visual word is a cluster in the CNN representation space and a visual document is constructed by

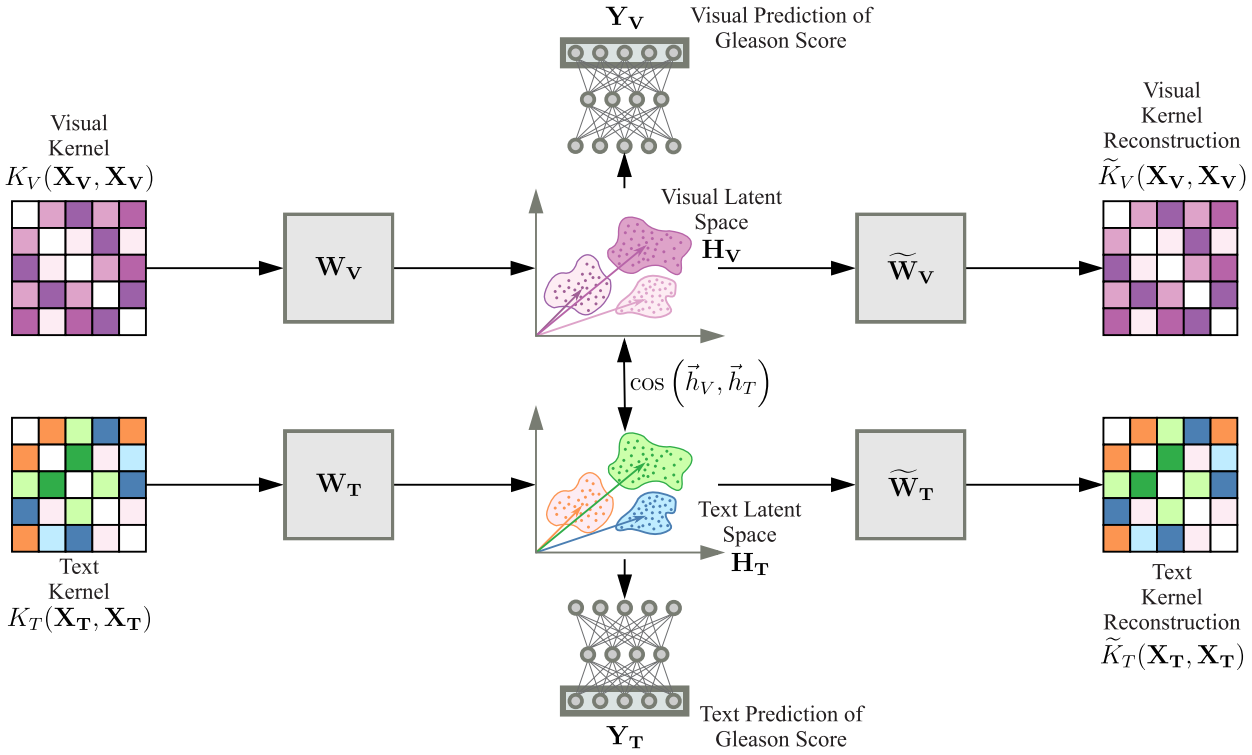

Figure 3-2: Conceptual diagram of the multimodal information fusion strategy.

assigning each patch descriptor to their closest centroid. As shown in Fig. **3-1**, this procedure allows computing the BoVW by counting the number of patches in each cluster. Besides, TF-IDF is also used to weight the distribution of visual words.

3.2. Fusion Strategy

An overall description of the information fusion strategy is shown in Fig. **3-2**. This strategy uses a reformulation of kernel matrix factorization that allows solving the problem through gradient-based optimization techniques as originally proposed in [47, 48]. The main idea of the strategy is to take advantage of the reformulation and include certain constraints that incorporate supervised and weakly-supervised semantic information. This can be seen as an extension of the original embedding learning problem, in which the goal is not only to find a low-dimensional latent space but to learn an aligned latent space for each modality that also contains information about the GS.

The information fusion strategy requires to compute two kernel matrices: on the one hand, a matrix $K_V(\mathbf{X_V}, \hat{\mathbf{X}}_\mathbf{V})$ is calculated applying a visual kernel function K_V on the TF-IDF representations $\mathbf{X_V} \in \mathbb{R}^{N \times |\mathcal{V}|}$ of the training WSIs (N is the number of observations and $|\mathcal{V}|$ is the visual codebook size) and a matrix $\hat{\mathbf{X}}_\mathbf{V}$ that can be the visual training or the

test TF-IDF representations. On the other hand, a matrix $K_T(\mathbf{X_T}, \hat{\mathbf{X}}_\mathbf{T})$ is calculated using the equivalent text matrices and functions. The main purpose of the kernel functions is to capture the complex nature of each modality to obtain a simpler representation, i.e., the data is transformed into a feature space where linear relations are more likely to be found. In this case, the feature spaces of each modality $\mathcal{X}_i \ \forall \ i \ \in \{V, T\}$ are mapped into a latent space $\mathbf{H_i} \in \mathbb{R}^{N \times L}$ of dimension L, through a linear transformation that uses a weight matrix $\mathbf{W_i} \in \mathbb{R}^{N \times L}$ and the kernel matrix $K_i(\mathbf{X_i}, \hat{\mathbf{X}}_\mathbf{i}) \in \mathbb{R}^{N \times N}$, as shown in Eq. 3-2.

$$\mathbf{H_i} = K_i(\mathbf{X_i}, \hat{\mathbf{X}}_\mathbf{i})\mathbf{W_i} \tag{3-2}$$

The latent representations are used to obtain a reconstruction of each kernel $\widetilde{K_i}(\mathbf{X_i}, \hat{\mathbf{X}}_\mathbf{i})$, this is achieved through a linear projection inducted by a weight matrix $\widetilde{\mathbf{W}_\mathbf{i}}$:

$$\widetilde{K_i}(\mathbf{X_i}, \hat{\mathbf{X}}_\mathbf{i}) = \mathbf{H}_i\widetilde{\mathbf{W}_\mathbf{i}} \tag{3-3}$$

The GS predictions are obtained using a deep learning architecture that transforms the latent representations $\mathbf{H_i}$ into predictions $\widetilde{\mathbf{Y}_\mathbf{i}}$. This is achieved using several non-linear transformations $f(\cdot)$ and a set of weights $\mathbf{W}_{ANN_i}$ as shown in Eq. 3-4.

$$\widetilde{\mathbf{Y}_\mathbf{i}} = f(\mathbf{H_i}, \mathbf{W}_{ANN_i}) \tag{3-4}$$

The complete loss function is presented in Eq. 3-5, it combines the three following errors: (1) the reconstruction of each kernel $\widetilde{K_i}(\mathbf{X_i}, \hat{\mathbf{X}}_\mathbf{i})$ allows us to estimate a reconstruction error $\overset{1}{\mathcal{J}_i}$, which is the mean squared error between the input and the output of each i modality and was derived using the kernel trick in [47]. This error is the basis of matrix factorization and is a non-supervised way to learn latent factors; (2) The GS predictions are used to calculate the *categorical cross-entropy* $\overset{2}{\mathcal{J}_i}$, which is used as an estimate of how different the predictions to one-hot encodings are of the GS ground truth $\mathbf{Y_i}$; (3) the cosine similarity $\cos(\vec{h_1}, \vec{h_2})$ is computed between latent vectors of each modality $\vec{h_1} \in \mathbf{H_1}$ and $\vec{h_2} \in \mathbf{H_2}$, it measures the degree of alignment between the visual and text latent spaces, and allows us to calculate an alignment error $\overset{3}{\mathcal{J}}$. The alignment term promotes the learning of close latent spaces, this allows the mutual enrichment of the visual and textual latent representations.

$$\begin{aligned}
&\overset{1}{\mathcal{J}_i} = \frac{1}{2}\sum_{\vec{x_j} \in \mathbf{X_i}} (1 - 2K(\vec{x_j}, \mathbf{X_i})\widetilde{K_i}(\vec{x_j}, \mathbf{X_i})^T + \widetilde{K_i}(\vec{x_j}, \mathbf{X_i})K(\mathbf{X_i}, \mathbf{X_i})\widetilde{K_i}(\vec{x_j}, \mathbf{X_i})^T) \\
&\overset{2}{\mathcal{J}_i} = -\sum_{\vec{y_j} \in \mathbf{Y_i}, \vec{\tilde{y}_j} \in \widetilde{\mathbf{Y_i}}} \left\langle \vec{y_j}, \log \vec{\tilde{y}_j} \right\rangle \qquad \overset{3}{\mathcal{J}} = \frac{1}{2}\sum_{\vec{h_1} \in \mathbf{H_1}, \vec{h_2} \in \mathbf{H_2}} (\cos(\vec{h_1}, \vec{h_2}) - 1)^2 \\
&\mathcal{J} = \alpha_1\overset{1}{\mathcal{J}_V} + \alpha_2\overset{1}{\mathcal{J}_T} + \beta_1\overset{2}{\mathcal{J}_V} + \beta_2\overset{2}{\mathcal{J}_T} + \gamma\overset{3}{\mathcal{J}}
\end{aligned} \tag{3-5}$$

3.3. Experimental Settings

To validate the advantages of weak multimodal supervision and M-LSA, we evaluate the performance of the proposed model for cancer detection and image retrieval. In this section, we describe the used dataset, the classification and retrieval pipelines, and the hyperparameter selection strategy.

3.3.1. Dataset Description

We use the TCGA-PRAD dataset, it is comprised of images and diagnostic reports from prostate cancer tissue with Gleason scores between 6 and 10 as shown in Fig. **3-3**. The data is available via the cancer genome atlas (TCGA), which is a publicly available large collection of digital pathology and other data that contains a set of 500 cases of prostate adenocarcinoma (PRAD). We use a subset with 235 cases as suggested in our baseline [20, 9]. The dataset was divided into the same baseline partitions for cross-validation: 141 cases for training, 48 for validation, and 46 for testing.

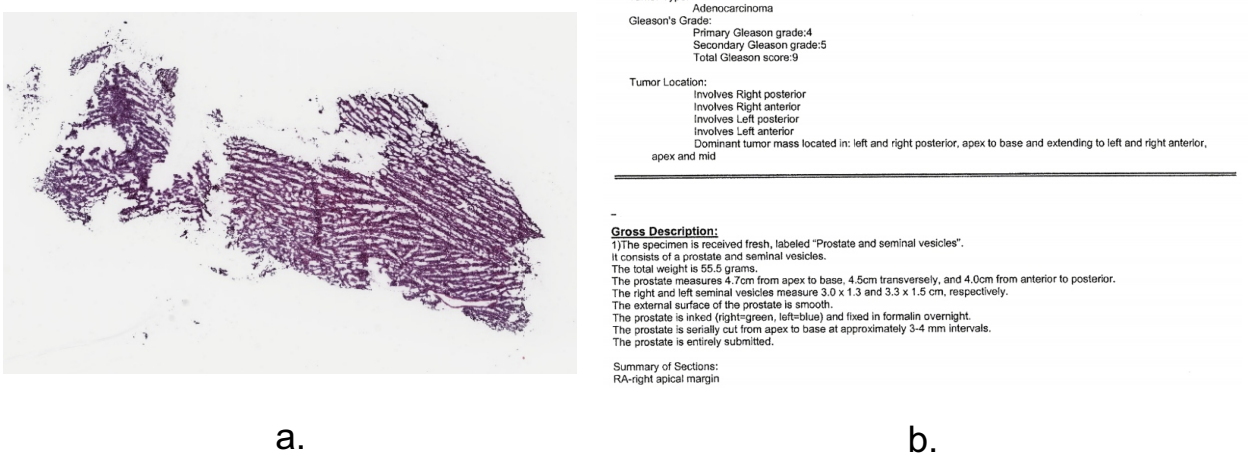

Figure 3-3: Example data from the TCGA-PRAD, a) whole-slide image; b) surgical pathology report.

3.3.2. Cancer Detection Performance

The proposed method is evaluated on the automatic classification of low (GS 6 and 7) and high (GS 8, 9, and 10) grades, as this stratification changes the treatment decision. We aim to evaluate the effects of the semantic enhancement, for this reason, two versions of the proposed model are trained: (1) A visual latent semantic embedding *V-LSE*, which is a version of the proposed model that does not include the alignment, i.e., it is a model that is

only trained using the WSIs. (2) *M-LSA*, which is a V-LSE model that is enhanced using the semantic information of the reports during training and is evaluated as shown in Fig. **3-1**. In this case, we evaluate the performance in terms of classification accuracy, which is the metric used in similar studies [20, 27].

3.3.3. Image Retrieval Performance

In this case, the models are trained to classify the five different categories of GS and the softmax outputs are used as an indexer. A single experiment consists of a simulated query, i.e., an example image is taken from the test set and the softmax outputs are calculated. Finally, these outputs are compared to the training set and using the cosine similarity with all the test cases a ranking is constructed. Similar to our baseline studies [9, 44], a case is relevant to the query if they share the same GS. The performance is evaluated in terms of mean average precision (MAP), GM-MAP, and precision at the top 10 (P@10) and 30 (P@30) retrieved results.

3.3.4. Hyperparameter Selection

The validation set is used to determine an appropriate combination of hyperparameters. We use a random search due to a large number of combinations. The model's weights are estimated through the Adam optimization algorithm ($lr = 10^{-3}, \beta_1 = 10^{-1}, \beta_1 = 10^{-2}$) using the training set and a combination of hyperparameters is selected using the validation loss as criteria. The loss parameters are configured as follows: $\beta_1 = \beta_2 = 5$, $\alpha_1 = \alpha_2 = \gamma = 1$; the visual codebook size $|V|$ is explored in a range between 100 and 1000; the latent dimension L is explored between 10 and 100; the activation functions of the ANNs are explored between ReLU, sigmoid and linear; the ANNs have two hidden layers and the number of units in each layer is explored in a range between 16 and 256; a dropout probability of 0.2 is added to the ANN weights for regularization; finally, some common kernels for histogram-based representations such as the linear, cosine, χ^2 and RBF are evaluated. The last two kernels have an additional hyper-parameter γ that must be determined, the range of the visual χ^2 kernel is $\gamma_V \in [10^{-3}, 10]$, the range for the text χ^2 kernel is $\gamma_T \in [10^{-4}, 10^{-1}]$, the range for the visual RBF kernel is $\gamma_V \in [10^{-2}, 100]$ and the range for the text RBF kernel is $\gamma_T \in [10^{-3}, 10]$. There are a total of 16 possible kernel combinations, for each one 100 random combinations of hyperparameters are used. The generated parameters for the visual modality are also used to train the V-LSE.

3.4. Results and Analysis

Table **3-1** presents the results for cancer detection. The proposed method is compared with similar studies that use comparable evaluation strategies on the same dataset. In the first

baseline study [20] a GoogLeNet is used to represent the patches and to summarize the information through a majority vote, V-LSE achieves an equivalent performance. This behavior is reasonable considering that we are using the same CNN for the representation and a unimodal model should achieve similar performance. The second baseline study [16] presents a modified AlexNet architecture and summarizes using a majority vote. The authors specify that they included more training data. Thus, an important advantage of M-LSA is that it achieves a similar performance including text content instead of more training data. This means it can obtain better performance when limited training data are available. Also, weak supervision allows us to find a better visual latent representation through the automatic incorporation of text content. There is no need to assign additional local labels to model a visual vocabulary as it is usually done in similar approaches.

Table 3-1: Comparison with state-of-the art methods in cancer detection.

Method	Accuracy
GoogLeNet [20]	73.52
Modified AlexNet [16]	76.90
V-LSE	74.02
M-LSA	**77.01**

The weak supervision allows us to find a more appropriate feature space that may not be found using the image content only, the results show that M-LSA outperforms V-LSE in cancer detection, achieving the best performance using an RBF kernel for both modalities, whereas V-LSE achieves it using a linear kernel. Likewise, compared to the linear alignment case of M-LSA, the RBF kernel achieves an accuracy improvement of 2.25 %, which shows the advantage of a non-linear alignment, especially, the importance of the kernel functions lies in their capacity to transform the representations to a feature space in which it is more likely to align the embeddings from different modalities. This is important considering that the representations learned in a deep neural network may not share linear relations with other modalities, thus, the kernel methods are valuable to model the complex nature of multimodal data.

The retrieval results are shown in Table **3-2**, presenting a comparison with the state-of-the-art retrieval methods that have been used to search PCa cases on the same dataset. It can be noticed that the semantic enhanced M-LSA model outperforms other image retrieval approaches. It is important to highlight that M-LSA only uses an image as a query, whereas other multimodal retrieval approaches require a multimodal query during the testing phase, which may not be suitable in realistic environments with new and uncertain cases where pathologists may not have a diagnosis report.

The proposed methodology represents an important opportunity for clinical translation, a CAD system can include M-LSA to provide a second opinion or retrieve similar cases. Contrary to other multimodal approaches, it does not require the text information during

Table 3-2: Results for the retrieval task, * denotes cases with multimodal queries.

Method	MAP	GM-MAP	P@10	P@30
Image Retrieval [9]	0.5113	0.3921	0.4500	0.4600
Text Retrieval [9]	0.4092	0.3561	0.4913	0.3775
Multimodal Retrieval* [9]	0.5404	0.4196	0.5217	0.4884
KLSE* [44]	0.6263	**0.4843**	0.5667	**0.6326**
Visual TF-IDF	0.4390	0.3486	0.3717	0.3667
Text TF-IDF	0.3574	0.3143	0.3848	0.3377
V-LSE	0.5881	0.3966	0.5000	0.4949
M-LSA	**0.6450**	0.4187	**0.5752**	0.5500

the prediction phase, which is important during uncertain diagnostics where the findings may not be clear and the annotations may be erroneous or cognitively biased.

3.5. Conclusion

We present a novel information fusion strategy for improving image representations using weak semantic supervision from diagnostic reports. The method uses the text information of diagnostic reports attached to histopathology cases as a source of weak supervision during training. During prediction it only uses visual information, same as unimodal visual methods, however, the experimental results showed that the use of multimodal information during training greatly improves the performance when compared to unimodal approaches. The proposed methodology shows that it is possible to exploit the multimodal information in medical databases that currently is not being fully exploited, considering a realistic environment in which pathologists may only have a WSI as the input query.

4 Extended Multimodal Latent Semantic Alignment

This chapter presents the extended Multimodal Latent Semantic Alignment (eM-LSA), which is a model that extends the original M-LSA model [45] in three main aspects as depicted in Fig. **4-1**: (1) *improved image representation*, in which different state-of-the-art convolutional neural networks (CNN) are compared for the weak-supervised automatic patch-level classification of high and low Gleason scores (GS). A detailed evaluation of the bag-of-visual-words (BoVW) summarization is conducted to determine the effects of different codebook sizes; (2) *improved text representation*, in which the original bag-of-words (BoW) for the diagnosis reports are extended to include n-grams representations. A detailed evaluation of a bag-of-ngrams (BoN) is performed to determine the effects of the number of grams; (3) *weak multimodal supervision*, in which a general alignment function is incorporated for better modeling of the multimodal relationships between the different latent spaces of each modality, in this case, eM-LSA model is contrasted with an extended visual latent semantic embedding (eV-LSE) to evaluate the effects of the weak multimodal supervision for the improved image and text representations.

4.1. Improved Visual Representation

CNNs have demonstrated to achieve performances that are comparable to international pathology experts in prostate cancer diagnosis [26]. Likewise, an important advantage is that these kinds of models leverage from general-purpose computing on graphics processing units (GPGPU), which makes them suitable for large datasets. This is especially important when dealing with histopathology image data, considering that the whole-slide-images (WSIs) are high-resolution images which size is about 80000×60000 pixels and the total size for each image is about 15GB. Currently, there are two main approaches to train a CNN with prostate histopathology images: on the one hand, if patch-level annotations are available, a CNN is trained to detect local findings or the Gleason patterns. On the other hand, if only the global labels or the Gleason score is available, the CNN is trained as a local feature extractor or predictor, and the decision is made after a summarization step.
Although it is desirable to have specialized annotations for each image patch, most of the medical databases for histopathology data only contain the global labels, for this reason, the second approach has gained special attention and is promising for digital pathology [49].

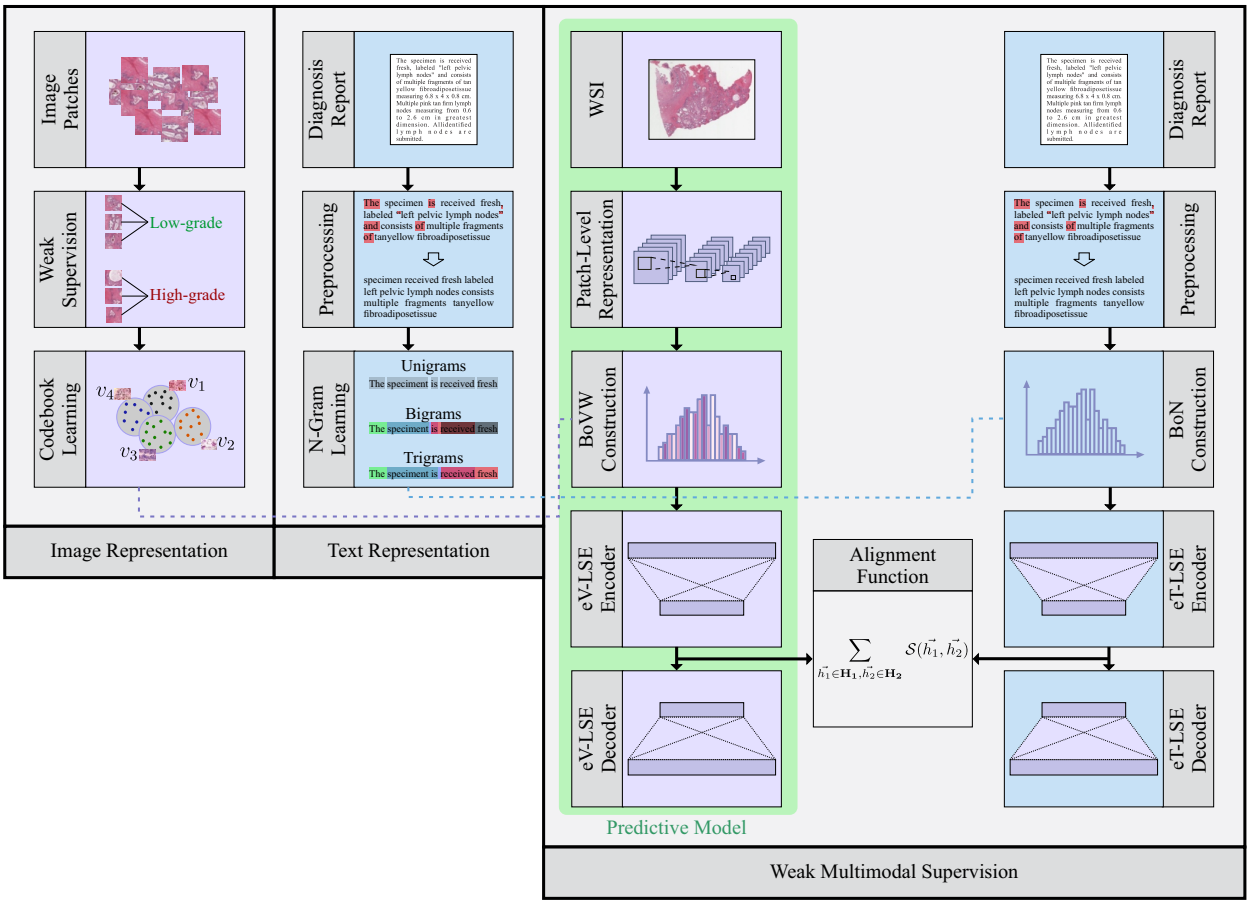

Figure 4-1: Learning procedure of eM-LSA

Furthermore, the challenge is to determine a feature representation for the image patches and an appropriate strategy to combine the local information. In this matter, M-LSA can be used to enhance a model during the summarization step, outperforming other similar approaches that share the same representations. Moreover, although a systematic evaluation of deep learning models for prostate cancer detection was performed in [20], the architectures that were evaluated are out-of-date and better results may be achieved using recent deep learning architectures that have demonstrated impressive results in similar tasks.
For the systematic evaluation of the CNN models and the learned representations, we propose the procedure that is depicted in Fig. **4-2**, it is composed of two phases:

- **Weak supervision**: this phase consists of the patch-level training of the CNN model. The overall idea is to learn a feature vector representation for each image patch that is highly correlated with the Gleason patterns. Moreover, we consider the case in which there are not local annotations for the images, thus, there is not a specific label y_{ij} for a given patch p_{ij} in the image $\mathbf{I}_i$, instead, there is a global label y_i for the whole

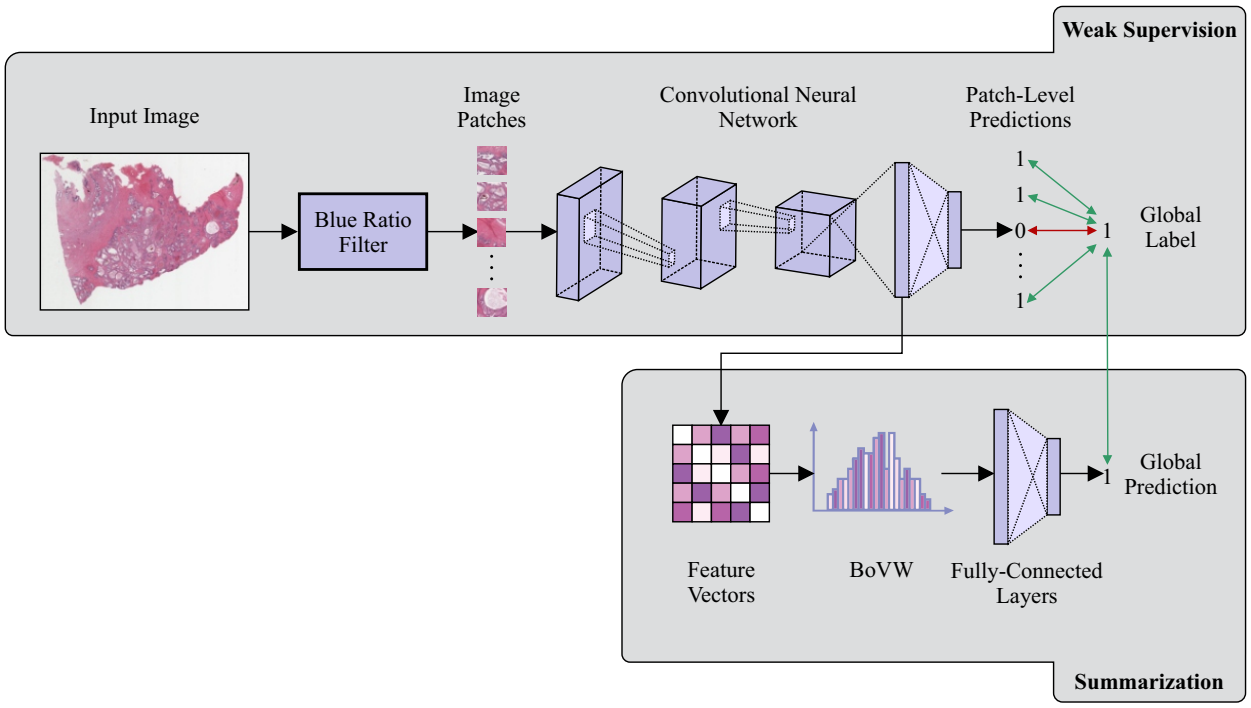

Figure 4-2: Training procedure for a CNN model using weak supervision and summarization.

image. This is a challenging problem considering the nature of the WSIs, because, not all the regions in an image are important for the classification and different patches can be associated with different Gleason patterns. The original M-LSA model used a pre-trained network for feature extraction, moreover, the weak supervision is incorporated in the complete eM-LSA pipeline, since different CNN representations can be more suitable for the BoVW summarization.

The CNNs are hence trained following a weakly-supervised approach, in which each patch is used to predict the global label y_i. This is a problem in which the global label may be imprecise and noisy for a given patch, furthermore, we hypothesize that a large number of patches must be related to the associated Gleason score, and a deep neural network must be able to capture these properties in the long term. The complete training process consists of three main steps: (1) *Preprocessing*, to mitigate the effects of the large size of the image and to filter redundant and non-informative image regions, a set of N_i non-overlapping image patches are extracted from an image $\mathbf{I_i}$ using the blue ratio as criteria. (2) *Estimation*, each image patch p_{ij} is feed to a convolutional neural network that obtains an estimation $\tilde{y}_{ij}$ of the Global label. (3) *Optimization*, the patch-level predictions are compared with the global label y_i, and a loss function is used to determine how similar are the predictions and the ground truth,

in this case, we consider the problem of high and low Gleason score classification, thus, the optimization consists on minimizing the binary cross-entropy for the N images in the training set:

$$\mathcal{J}_{weak} = -\sum_{i=1}^{N}\sum_{j=1}^{N_i} y_i \log\left(\tilde{y}_{ij}\right) + (1-y_i)\log\left(1-\tilde{y}_{ij}\right) \tag{4-1}$$

- **Summarization**: In this phase, the information from all the image patches is aggregated into a single descriptor, we extract a feature vector $\vec{z}_{ij} \in \mathbf{Z_i}$ of each patch p_{ij} using an intermediate representation in the CNN. Then, the feature vectors $\mathbf{Z_i}$ are vector quantized into a codebook of size $|V|$ and a BoVW $\vec{x}_i \in \mathbf{X}$ representation is constructed using the counts of the hard assignments as it was described in the chapter 3. The BoVW for all the training images $\mathbf{X_V} \in \mathbb{R}^{N\times|V|}$ are weighted using TF-IDF with sub-linear scaling and a multilayer perceptron (MLP) is used to obtain a global prediction $\tilde{y}_i$ of the Gleason score. This prediction is compared with a ground truth y_i and the MLP weights are estimated from the optimization of the binary cross-entropy:

$$\mathcal{J}_{summary} = -\sum_{i=1}^{N} y_i \log\left(\tilde{y}_i\right) + (1-y_i)\log\left(1-\tilde{y}_i\right) \tag{4-2}$$

4.2. Improved Text Representation

Although a specific representation for the data modalities is not required, it is reasonable to consider a text representation that is compatible with the image representation. In this case, a reasonable text representation is a BoW since the images are represented as a BoVW, Choosing a different representation would require a more complex kernel function for the alignment to make sense. A BoW is a probabilistic representation that has been already studied for text classification and retrieval [45], further, in this case we explore a generalization that measures the distribution $P(\mathcal{S} = s_l|\mathcal{D} = d_i)$ of term sequences. This is known as a Bag-of-Ngrams (BoN) and is constructed using counts of sequences $s_l = [t_1, t_2, \ldots, t_{N_n}]$ of size N_n. This representation is weighted using the TF-IDF schema and allows us to build a matrix $\mathbf{X}_t \in \mathbb{R}^{N\times|T|}$ ($|T|$ is the total number of possible N_n-grams) to represent the documents.
In contrast with the BoW representation, an important advantage of the BoN representation is that it partially takes into account the sequence information of the terms in the text [9]. For instance, a conventional BoW would not be able to capture the sequence information in the following example: "*Gleason Score: 8, Primary pattern: 4, Secondary pattern: 4, score: 4+4*", because it does not preserve the location relations between the numbers and the words, it would generate counts of the tokens ["4","8","4 + 4","*Gleason*","*pattern*","*Primary*",

"*Secondary*","*Score*"], which are independent and do not allow to discriminate between the number "4.ªssociated to the Gleason pattern and the number "8.ªssociated to the Gleason score. Whereas, a BoN is more appropriated since it would preserve the relative location of the words in the 3-grams tokens ["*Gleason*","*Score* : ","8"], ["*Primary*","*pattern* : ","4"], ["*Secondary*","*pattern* : ","4"], which allow to associate the numbers with the categories.

4.3. Weak Multimodal Supervision

The overall idea of weak multimodal supervision is to incorporate semantic information into an image model during the learning phase. This behavior was originally explored in the fusion strategy of M-LSA, which constrained the intermediate or latent representations of each modality to be aligned with the other. This approach outperformed fully unimodal or multimodal strategies and showed the advantages of weak multimodal supervision. However, M-LSA forces the latent spaces of each modality to be aligned according to the cosine similarity, which may not appropriate for all the cases. For this reason, we propose a more general alignment for the weak multimodal supervision in eM-LSA as it is shown in Fig. **4-1**. The complete weakly-multimodal-supervised model is composed of two models (one for each modality) that share information through an alignment function. eM-LSA contains an extended visual latent semantic embedding (eV-LSE) model for the images and an extended text latent semantic embedding (eT-LSE) for the texts.

Both eV-LSE and eT-LSE use the improved representations that are learned in the image and text representation phases. This allows us to compute a pair of BoVW-BoN for each image-text pair. More precisely, if we consider a training set $\mathcal{D} = \{(\mathbf{I}_1, \mathbf{d}_1, y_1), (\mathbf{I}_2, \mathbf{d}_2, y_2), \ldots, (\mathbf{I}_N, \mathbf{d}_N, y_N)\}$ that contains triplets of WSIs $\mathbf{I}_i$, diagnosis reports $\mathbf{d}_i$ and labels y_i; eM-LSA computes an intermediate $\mathbf{X}_v \in \mathbb{R}^{N\times|V|}$ BoVW representation for the images and a $\mathbf{X}_T \in \mathbb{R}^{N\times|T|}$ BoN for the texts. The remaining components for both eV-LSE and eT-LSE are common and can be described in three steps:

First, a kernel matrix $K_i(\mathbf{X_i}, \hat{\mathbf{X}}_\mathbf{i})$ is computed between the intermediate representation $\mathbf{X_i}$ and the matrix $\hat{\mathbf{X}}_\mathbf{i}$ that can be associated with the training, validation or test intermediate representations for the modality $i \in \{V, T\}$. This matrix is linearly projected to a latent space $\mathbf{H_i} \in \mathbb{R}^{N\times L}$ using an encoder network (view Eq. 3-2). Second, a decoder network transforms the latent representation $\mathbf{H_i}$ into a reconstruction of the kernel $\hat{K}_i(\mathbf{X_i}, \hat{\mathbf{X}}_\mathbf{i})$ and incorporates supervised supervision from y_i using several fully connected layers as it was described in the section 3.2. Third, a differentiable dissimilarity or alignment function $\mathcal{S}$ is computed for each pair of $\mathbf{H_V}$ and $\mathbf{H_T}$, which differs from the original M-LSA formulation that considered $\mathcal{S}$ as the cosine similarity. Furthermore, it is an additional hyperparameter that must be explored in eM-LSA.

The resultant weakly-multimodal-supervised loss function for the optimization of eM-LSA

is:

$$\overset{1}{\mathcal{J}}_i = \frac{1}{2}\sum_{\vec{x_j}\in\mathbf{X_i}} (1 - 2K(\vec{x_j}, \mathbf{X_i})\widetilde{K_i}(\vec{x_j}, \mathbf{X_i})^T + \widetilde{K_i}(\vec{x_j}, \mathbf{X_i})K(\mathbf{X_i}, \mathbf{X_i})\widetilde{K_i}(\vec{x_j}, \mathbf{X_i})^T)$$

$$\overset{2}{\mathcal{J}}_i = -\sum_{\vec{y_j}\in\mathbf{Y_i},\vec{\tilde{y}_j}\in\widetilde{\mathbf{Y}_\mathbf{i}}} \left\langle \vec{y_j}, \log \vec{\tilde{y}_j} \right\rangle \qquad \overset{3}{\mathcal{J}} = \sum_{\vec{h_1}\in\mathbf{H_1},\vec{h_2}\in\mathbf{H_2}} \mathcal{S}(\vec{h_1}, \vec{h_2}) \tag{4-3}$$

$$\mathcal{J} = \alpha_1\overset{1}{\mathcal{J}}_V + \alpha_2\overset{1}{\mathcal{J}}_T + \beta_1\overset{2}{\mathcal{J}}_V + \beta_2\overset{2}{\mathcal{J}}_T + \gamma\overset{3}{\mathcal{J}}$$

4.4. Experimental Setup

We propose four main experiments to evaluate the representations for each modality and the effects of the weakly-multimodal-supervision. The experiments are performed on the TCGA-PRAD dataset using the same partitions that were used in [9, 20, 44], i.e., training: 141, validation: 48, test: 46.

4.4.1. Weak supervision

For the patch-level training, the WSIs are divided into patches, an average of $N_i = 2000$ patches are extracted for each image. The number of patches for each class is described in Table **4-1**. There are a total of 268730 training patches, 91540 validation patches, and 87740 test patches.

Table 4-1: Number of patches per class for each partition.

Partition	Low-grade	High-grade
Training	124186	144544
Validation	40760	50780
Test	39942	47798

We explore the five different CNN architectures that are presented in Table **4-2**. Since a complexity analysis can not be easily performed over these deep learning models, a comparison of the floating operations per second (FLOPS) and their number of parameters is presented instead. We choose these five models considering the impressive performances that they have shown in similar tasks and that these models have demonstrated to be appropriate for general purpose transfer learning.

To train the CNN models we follow a transfer learning approach, in which the models are initialized with the ImageNet pre-trained weights and fine-tuned using the histopathology images. The last fully-connected layers of each model are replaced (as it is usually done for fine-tuning) by a fully-connected layer (with 1024 units, ReLU activation function, and a

Table 4-2: Performance comparison for the implemented convolutional neural networks.

Model	FLOPS	Number of parameters	Feature vectors size
DenseNet-201	4G	20M	1920
Inception V3	6G	24M	2048
Xception	8G	23M	2048
ResNet-152 V2	11G	60M	2048
Inception-ResNet V2	12G	56M	1536

dropout rate of 0.2 for regularization) and an output layer (two units for binary classification and a softmax activation function). The global labels are transformed into one-hot encodings and the model's weights are estimated using the Adam optimization algorithm with a learning rate of 10^{-7}, and a batch size of 32 for 20 epochs. Sample-wise data augmentation is used to enhance the training patches, including the following random image transformations: brightness adjustment between 0 and 0.3, contrast adjustment between 0.7 and 1, hue adjustment between 0 and 0.1, saturation adjustment between 0.7 and 1, random rotations between 0 and 360 degrees, and random vertical and horizontal image flips. The CNNs are trained using the training partition and the best hyperparameters are selected using the lowest validation loss as criteria. All the models are trained on an NVIDIA RTX 2080 TI GPU using the TensorFlow 2.0 framework. The performance of the models is evaluated in terms of the classification accuracy, precision, recall, and F1 score.

4.4.2. BoVW representation

To evaluate the effects of the summarization step, a feature vector representation for each image patch is extracted from the CNN with the lowest validation loss. Then a K-Means model is used to build a codebook from the feature vector representations of all the training patches, the codebook size is explored in the range $|V| \in [100, 2000]$. The trained K-Means model assigns each patch to a specific code-vector, and a BoVW is computed as counts of the assignments. To assess the impact of the TF-IDF weighting schema, we evaluate two representations that correspond to the cases with and without TF-IDF. To evaluate the performance of each representation, an MLP model (two hidden layers of 64 units, ReLU activation function, a dropout rate of 0.2, and an output softmax layer of size two) is trained to classify high and low Gleason scores. The MLP models are optimized using the Adam optimization algorithm with a learning rate of 10^{-4}, a batch size of 32 for 100 epochs. A number of 10 trials are used for each codebook size to evaluate the average and the standard deviation for the accuracy, precision, recall, and F1 score. The models are fitted using the training BoVW and the best hyperparameters were selected using the lowest validation loss as selection criteria.

4.4.3. BoN representation

The main objective of these experiments is to assess the BoN representation, in this regard, we explore sequences in the range $N_n \in [1, 2, \ldots, 7]$. Similar to the previous experiments, the TF-IDF weighting schema is assessed and the same MLP architecture used in the experiments for the visual representation is trained to classify between high and low Gleason scores.The preprocessing consists of stop-words, uppercase, and special character removal. The resultant texts have a vocabulary of 4959 different words and the most important terms according to the TF-IDF weighting schema are shown in Fig. **4-3**. This shows that the pathology reports are structured data with a small vocabulary that contains domain-specific terms.

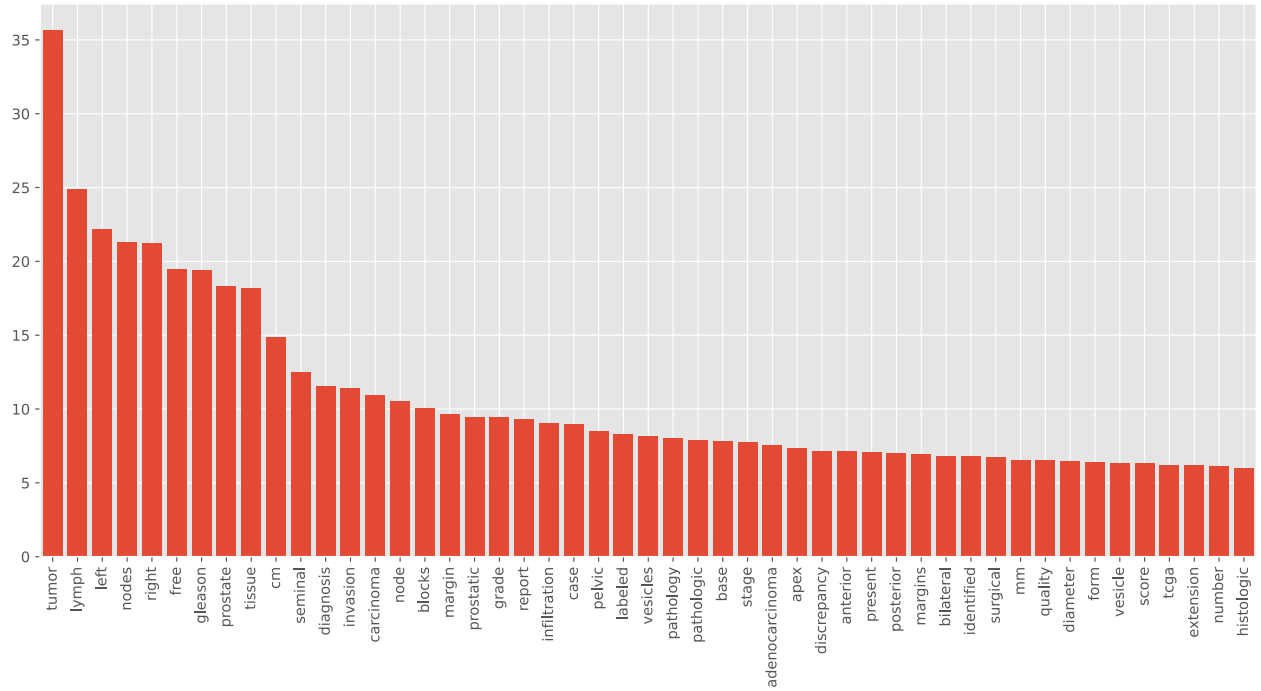

Figure 4-3: Average TF-IDF representation of the top 50 terms.

Similarly, the model is trained 10 times with different initial random weights, and the average and standard deviation of the accuracy, precision, recall, F1 score, and vocabulary size are computed.

4.4.4. Weak multimodal supervision

Similar to the original experiments that were performed in M-LSA, an appropriate combination of hyperparameters is chosen using the validation loss as selection criteria. A random search is conducted due to a large number of hyperparameter combinations. The model's weights are estimated through the Adam optimization algorithm ($lr = 10^{-4}, \beta_1 = 10^{-1}, \beta_1 =$

10^{-2}) using the training set. The loss parameters are configured as follows: $\beta_1 = \beta_2 = \gamma = 1$, $\alpha_1 = \alpha_2 = 0{,}1$; the visual codebook size K_v is selected from the experiments on the visual representation; the latent dimension L is explored between 600 and 1000; the activation functions of the ANNs are explored between ReLU, sigmoid and linear; the ANNs have two hidden layers and the number of units in each layer is explored in a range between 32 and 128; a dropout probability of 0.2 is added to the ANN weights for regularization; some common kernels for histogram-based representations such as the linear, cosine, χ^2 and RBF are evaluated. The last two kernels have an additional hyper-parameter γ that must be determined, the range of the visual χ^2 kernel is $\gamma_V \in [10^{-3}, 10]$, the range for the text χ^2 kernel is $\gamma_T \in [10^{-4}, 10^{-1}]$, the range for the visual RBF kernel is $\gamma_V \in [10^{-2}, 100]$ and the range for the text RBF kernel is $\gamma_T \in [10^{-3}, 10]$. Some common distance metrics such as the euclidean distance, cosine similarity, and Manhattan distance are explored for the alignment.
There are a total of 16 possible kernel combinations, 100 random combinations of hyper-parameters are used for each one. We train a unimodal model (eV-LSE) and a multimodal enhanced model (eM-LSA) to evaluate the effects of the weak multimodal supervision. The generated parameters for the visual modality are also used to determine an appropriate combination for the unimodal version. A random search is conducted using binary labels for classification and as a multiclass problem for retrieval, the predicted labels are used as an indexer and the Gleason score is used as the relevance criteria. The models with the best hyperparameters are trained for 10 trials with different initial weights to report the average and standard deviation of the performance metrics. The classification performance is evaluated in terms of accuracy, precision, recall, and F1 score. The retrieval performance is evaluated in terms of the mean average precision (MAP), geometrical mean average precision (GM-MAP), and precision at top 10 (P@10) and 30 (P@30) retrieved results.

4.5. Results and Analysis

The results are structured according to the four proposed experiments. Likewise, to ensure a fair comparison, the methods are only contrasted with studies that use similar or comparable data partitions for cross-validation. The models are trained only once for the weakly supervision experiments since the training takes a long time due to a large number of image patches (view Table **4-1**). Moreover, training a neural network with large data (even infinite due to the data augmentation) ensures a robust estimation. For the other experiments, the training is performed over different trials, allowing a direct assessment of the model's bias and variance.

4.5.1. Weak Supervision

Table **4-3** presents the results for prostate cancer detection. The implemented CNN architectures are compared with a baseline study [20] that presented a systematic comparison

of different deep learning models in the same task. Although this study held robust experimentation, it was published in 2017 and the evaluated models are mostly out-of-date. In this matter, the main evaluation metric in this work is the classification accuracy, which is appropriate and interpretable in balanced datasets (view Table **4-1**). Based on this, Xception is the network with the best results, showing an overall increment of 1.34 % in the accuracy on the test set, which represents an approximate reduction of 535 miss-classified cases in comparison with the state-of-the-art results. According to Table **4-2**, the Xception network is in the middle in terms of computational cost. The models that were evaluated in [20] have lower FLOPS and a lower number of parameters (for instance, the GoogleNet has 0.7M parameters and 1.5G FLOPS) but lower classification performances, which makes them suitable for low-resource applications, showing a trade-off between computational cost and classification performance.

Table 4-3: Performance of different CNN models for high and low Gleason score classification using weak supervison.

Model	Accuracy	Precision	Recall	F1
LeNet [20]	0.6947	-	-	-
AlexNet [20]	0.7116	-	-	-
GoogLeNet [20]	0.7352	-	-	-
ResNet152 V2	0.6993	0.6715	**0.8770**	0.7606
Inception-ResNet V2	0.7128	0.7013	0.8236	0.7576
Inception V3	0.7146	0.6922	0.8572	0.7659
DenseNet201	0.7346	**0.7495**	0.7703	0.7597
Xception	**0.7486**	0.7445	0.8198	**0.7803**

The additional evaluation metrics have an interesting interpretation considering the clinical connotation of the Gleason Score. There is evidence of the correlation between this grading system and certain clinical factors such as the metastasis risk, local and regional recurrence, patient outcome, prognostic factor, among others [20, 50, 51, 7]. Choosing the best model is not straightforward, it depends on the scope of the clinical application. For instance, the DenseNet201 model achieves high precision, meaning that it is a good model for the prediction of the positive class (high Gleason scores); whereas, the ResNet152 V2 achieves a high recall, showing that it is a good predictor for the negative class (low Gleason scores). The model with the best overall performance is the Xception network, it presents the best accuracy (valid for balanced data) and the best F1 score (considers the precision and recall).

4.5.2. BoVW representation

Fig. **4-4** presents the results of the visual representation experiments. It is shown a comparison of the classification performance for the BoVW representation for different codebook

sizes and the cases with and without TF-IDF. These representations were computed using the Xception as a feature extractor, we selected this network because it presented the lowest validation loss. The best results were achieved using a codebook size of 1700 with TF-IDF weighting, this representation yields the following performance metrics: accuracy 0.8529±0.0165, precision 0.8230±0.0140, recall 0.9293±0.0224, F1 0.8728±0.0147. The TF-IDF schema not only provides better results, but it also reduces the variability of the models as it is shown in Fig. **4-4**. These results show the advantages of using the BoVW summarization strategy with the improved representations of the Xception, especially, this improved image representation allows performances that considerably outperform the previous unimodal and multimodal results (view Table **3-1**).

4.5.3. BoN representation

Fig. **4-5** presents the results of the text representation experiments. It is shown a comparison of the classification performance of the BoN representation for different N-gram sizes and the

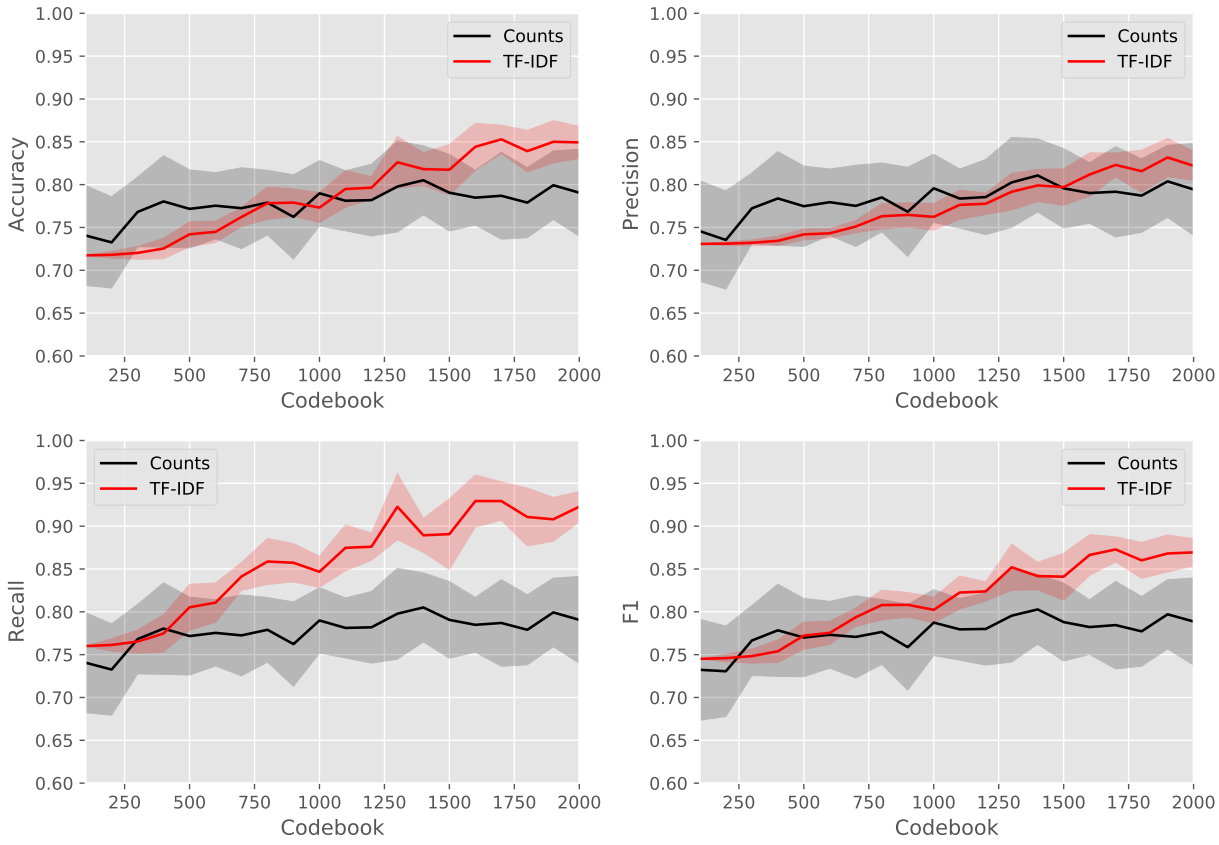

Figure 4-4: Comparison of the evaluation metrics for different codebook sizes for the BoVW representation with and without TF-IDF weighting schema.

cases with and without TF-IDF. Considering the accuracy metric, the best representation is a BoN of bigrams and TF-IDF, it yields the following performance metrics: accuracy 0.8862±0.0157, precision 0.8368±0.0131, recall 0.9827±0.0304, F1 0.9036±0.0145. Similar to the visual representation results, the TF-IDF ensures better performances and reduces the variability. Besides, there is a clear difference between the performance of the visual and text modalities, specifically, the text representation provides a relative average improvement of 3.33 %. This behavior is well-known and common in multimodal datasets.

4.5.4. Weak multimodal supervision

The results of the weak multimodal supervision are structured in two parts. First, the proposed model is evaluated for high and low Gleason score classification. And second, it is evaluated for case-based prostate image retrieval.

- **Classification**: Table **4-4** presents the performance evaluation of eV-LSE and eM-LSA on the test set. For the multimodal case, we present the kernel combinations with

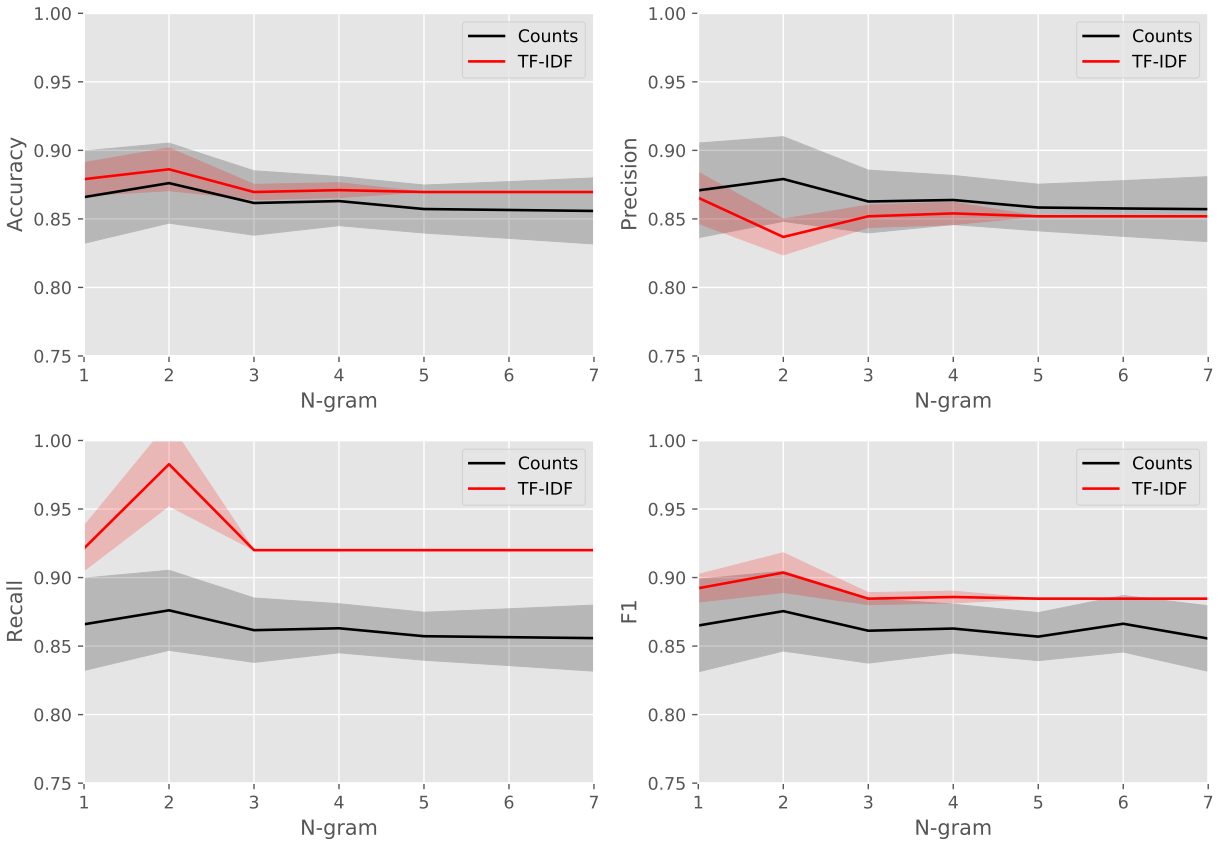

Figure 4-5: Comparison of the evaluation metrics for different N-grams for the BoN representation with and without TF-IDF weighting schema.

the lowest validation loss for each possible visual kernel. These results show that the weak multimodal supervision provides an overall average improvement in comparison with the unimodal case. Likewise, eM-LSA outperforms eV-LSE in all the cases. The image model that was trained included the weak multimodal supervision achieves an accuracy of 0.8848±0.0258, whereas the unimodal image model achieves an accuracy of 0.8717±0.0258. The best results correspond to a non-linear kernel function, showing that, although eM-LSA operates over deep representations, a linear assumption is not enough for the spaces of each representation. Also, the best results for eM-LSA were achieved using different alignment functions for the different kernel combinations (cosine for the best model of each possible kernel), showing that there is a different similarity function for each inducted kernel space.

Table 4-4: Comparison of the classification performance between eM-LSA and eV-LSE.

	K_V	K_T	Accuracy	Precision	Recall	F1
eV-LSE	Linear	-	0.8652±0.0163	0.8289±0.0153	**0.9480±0.0256**	0.8843±0.0144
	Cosine	-	**0.8717±0.0227**	0.8404±0.0250	0.9440±0.0196	**0.8890±0.0189**
	χ^2	-	0.8696±0.0000	**0.8519±0.0000**	0.9200±0.0000	0.8846±0.0000
	RBF	-	0.8717±0.0181	0.8448±0.0155	0.9360±0.0196	0.8880±0.0158
eM-LSA	Linear	Linear	0.8783±0.0242	0.8604±0.0316	0.9280±0.0240	0.8925±0.0204
	Cosine	**Cosine**	**0.8848±0.0258**	**0.8639±0.0259**	0.9360±0.0265	**0.8983±0.0223**
	χ^2	Cosine	0.8761±0.0100	0.8486±0.0107	**0.9400±0.0200**	0.8918±0.0092
	RBF	Linear	0.8804±0.0146	0.8571±0.0108	0.9360±0.0196	0.8948±0.0133

Table **4-5** presents a comparison of different state-of-the-art classification strategies and the proposed methods. One of the main improvements appears from the new image representations using the Xception network, for instance, a complete unimodal strategy like Xception + BoVW has a relative average improvement of 8.28 % over the multimodal enhanced GoogleNet + M-LSA representation. Similarly, including supervised information in the summarization strategy results in an overall improvement in comparison with unsupervised strategies like the BoVW. The multimodal enhanced model Xception + eM-LSA achieves the best performance in this task, this is especially important since it is a model that does not require the text information during the prediction phase and its performance is equivalent to the unimodal text models presented in the section 4.5.3. Finally, the presented methodology proposes a new paradigm in information fusion that has not been completely explored in histopathology, it allows to enhance image models using semantic information during the training and does not require the text data for the prediction of new cases.

- **Retrieval**: Table **4-6** presents the retrieval performance of eV-LSE and eM-LSA using the test set as query. Similar to the classification case, we select the kernel combinations

Table 4-5: Comparison with the state-of-the-art methods in the classification task using the improved representations.

Strategy	Accuracy
GoogleNet + Majority vote [20]	0.7352
AlexNet + Majority vote [16]	0.7690
GoogleNet + BoVW [45]	0.7370
GoogleNet + V-LSE [45]	0.7402
GoogleNet + M-LSA [45]	**0.7701**
Xception + BoVW	0.8529 ± 0.0165
Xception + eV-LSE	0.8717 ± 0.0227
Xception + eM-LSA	**0.8848 ± 0.0258**

with the lowest validation loss for each possible visual kernel. The multimodal enhanced models presented outperformed unimodal models in terms of the four evaluated metrics. eM-LSA achieves a maximum MAP of 0.7549±0.0071, whereas the eV-LSE model achieves a MAP of 0.7317±0.0058. In this task, the best results are also achieved with non-linear kernel functions, which shows the advantages of combining kernel methods with deep learning. Likewise, the best eM-LSA model uses a Euclidean alignment with a visual RBF kernel and a text Linear kernel, showing that in some cases the cosine similarity is not the best option.

Table **4-7** presents a comparison of different state-of-the-art retrieval strategies and the proposed methods. These results show the importance of a good image representation strategy, especially, there is a clear difference in the original V-LSE and M-LSA results that used the GoogleNet network and the latest results that use the Xception. Besides, the presented methodology presents a novel alternative for conventional multimodal retrieval in the medical domain, allowing to exploit unused text data from medical databases while allowing unimodal image queries.

Table 4-6: Comparison of the retrieval performance between eM-LSA and eV-LSE.

	K_V	K_T	MAP	GM-MAP	P@10	P@30
eV-LSE	Linear	-	0.7266±0.0145	0.5179±0.0160	0.6652±0.0174	0.6494±0.0163
	Cosine	-	0.7256±0.0159	0.5172±0.0156	0.6522±0.0257	**0.6502±0.0194**
	χ^2	-	**0.7317±0.0058**	**0.5241±0.0060**	**0.6717±0.0065**	0.6497±0.0068
	RBF	-	0.7015±0.0155	0.4905±0.0168	0.6369±0.0196	0.6226±0.0159
eM-LSA	Linear	χ^2	0.7358±0.0163	0.5279±0.0164	0.6631±0.0243	0.6633±0.0205
	Cosine	Cosine	0.7468±0.0129	0.5389±0.0132	0.6805±0.0170	0.6764±0.0153
	χ^2	Cosine	0.7356±0.0076	0.5281±0.0087	0.6761±0.0065	0.6544±0.0096
	RBF	**Linear**	**0.7549±0.0071**	**0.5461±0.0088**	**0.7022±0.0099**	**0.6851±0.0076**

Table 4-7: Comparison with the state-of-the-art methods in the retrieval task using the improved representations, the cases with * require multimodal queries.

Method	MAP	GM-MAP	P@10	P@30
GoogleNet [9]	0.5113	0.3921	0.4500	0.4600
Doc2Vec [9]	0.4092	0.3561	0.4913	0.3775
Rank fusion* [9]	0.5404	0.4196	0.5217	0.4884
KLSE* [44]	0.6263	**0.4843**	0.5667	**0.6326**
BoVW [45]	0.4390	0.3486	0.3717	0.3667
BoW [45]	0.3574	0.3143	0.3848	0.3377
V-LSE [45]	0.5881	0.3966	0.5000	0.4949
M-LSA [45]	**0.6450**	0.4187	**0.5752**	0.5500
eV-LSE	0.7317±0.0058	0.5241±0.0060	0.6717±0.0065	0.6497±0.0068
eM-LSA	**0.7549±0.0071**	**0.5461±0.0088**	**0.7022±0.0099**	**0.6851±0.0076**

4.6. Conclusion

In this study we performed a systematic evaluation of the different components in M-LSA, showing the advantages of the multimodal enhancement. We evaluated the state-of-the-art CNN architectures for weak supervision, which allowed the construction of improved patch representations without requiring local annotations. In a like manner, a rigorous evaluation of histogram summarization strategies for prostate histopathology images was performed, showing the effects of different codebook sizes and the TF-IDF weighting schema. The results are promising for digital pathology, especially, for the automatic diagnosis of prostate cancer and case-based image retrieval. We propose a new methodology that has not been explored in the automatic assessment of prostate cancer, which encourages the development of new multimodal strategies more appropriate for a clinical scenario.

5 Conclusions and future work

5.1. Histopathology Image Classification

Several strategies for the automatic classification of histopathology images were proposed in this thesis. Prostate cancer was the scope of the experimentation, however, the proposed methods are general enough and its adaptation for any other multimodal medical application is straightforward and requires minor modifications. We aim to address two ongoing challenges in the automatic assessment of histopathology images: on the one hand, classification and representation image models are obtained from data with low-quality labels using weak supervision, this differs from conventional approaches in which the models are trained using local annotations only. The achieved results pose important implications, especially, the neural networks achieve acceptable performances when they are taught to learn local patterns from global labels, this shows that the current deep learning models can automatically represent the Gleason Patterns without the need of prior or specific knowledge. On the other hand, a detailed exploration of the BoVW summarization strategy allowed us to achieve the best state-of-the-art results in high and low Gleason score classification, showing the advantages of the combination of classical histogram representation strategies with novel deep learning representations.

The experiments were performed in the TCGA-PRAD dataset, which is a large patch-level dataset but a small sample-level dataset. For this reason, the kernel methods behaved so well, since its computing can be easily performed after the summarization of the WSIs. Moreover, this approach can not be directly applied in large sample-level datasets, considering the memory complexity for the computation of the kernel methods. In this regard, several approaches can be explored, for instance, the original reformulation of the kernel matrix factorization algorithm was proposed upon active learning using the budget approach. M-LSA uses a complete budget but it can be modified for large samples. Conversely, there are other recent promising approaches to scale kernel methods that can be used in M-LSA, including the random Fourier features, Nyström method, among others.

5.2. Histopathology Image Retrieval

A medical image retrieval system is an important tool that allows the accurate and appropriate use of the information stored in medical databases. It allows an automated and unbiased

search that semantically compares the information between different images depending on the representation complexity. In this matter, although neural networks have shown impressive representation capabilities in similar tasks, the application of deep learning for medical information retrieval has been barely explored and represents a promising research direction. Deep representation learning must be especially considered since there is evidence that neural networks can learn high-level concepts from low-level features, therefore, using deep intermediate representations is an appropriate way to describe complex clinical-pathological relations that can not be easily captured with other representation techniques.
In this study, M-LSA was used to compute enhanced image descriptors that were used in a conventional case-based image retrieval procedure. Moreover, the complete process can be reformulated to include additional learning components, for instance, metric learning can be used to determine a deep similarity function for a better comparison of the query and target descriptors, or, reinforcement learning can be used to train an end-to-end model that learns to retrieve similar images while enhancing the image representations.

5.3. Multimodal Learning

In this thesis, a novel information fusion strategy that incorporates semantic information into the image models during the training was proposed. It is a new approach that opens a new paradigm in prostate histopathology and permits exploiting unused text information in medical databases while preserving the properties of unimodal models to predict or retrieve new cases. In this matter, the main aim of current multimodal methods is to merely improve the evaluation performance by considering other modalities as additional model's inputs, this is highly restrictive since it requires the availability of all the modalities for a single prediction and may not be suitable in the practice. Moreover, the proposed method considers a more realistic scenario in which the predictions require the only image modality and exploits available multimodal data during the training.
Future work is aimed to formulate an end-to-end model that combines the representations for each modality with the fusion strategy. This would require reformulating the BoVW and the clustering strategy as differentiable components that can be integrated into deep learning models, the representations could be fine-tuned and the model would achieve higher performances. Likewise, the proposed alignment function can be replaced with a metric learning strategy, which would automatically adapt for a given kernel combination instead of being an additional hyperparameter. Finally, we used classical representation strategies for the text modality that are compatible and similar to the visual representation, however, recent advances in natural language processing show impressive performances using novel text embeddings that are based on transformers and recurrent neural networks, these embeddings can be incorporated to improve the proposed method.

References

[1] WCRF, "Worldwide cancer data. Global cancer statistics for the most common cancers," 2018.

[2] PCEC, "Gleason Score, Prostate Cancer Grading & Prognostic Scoring," 2020.

[3] J. C. Caicedo, J. A. Vanegas, F. Páez, and F. A. González, "Histology image search using multimodal fusion," *Journal of Biomedical Informatics*, vol. 51, pp. 114–128, 2014.

[4] P. W. Hamilton, P. Bankhead, Y. Wang, R. Hutchinson, D. Kieran, D. G. McArt, J. James, and M. Salto-Tellez, "Digital pathology and image analysis in tissue biomarker research," *Methods*, vol. 70, no. 1, pp. 59–73, 2014.

[5] S. Al-Janabi, A. Huisman, and P. J. Van Diest, "Digital pathology: current status and future perspectives," *Histopathology*, vol. 61, no. 1, pp. 1–9, 2012.

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

[7] K. Trpkov, "Contemporary Gleason Grading System," in *Genitourinary Pathology: Practical Advances*, pp. 13–32, Springer-Verlag New York, 2015.

[8] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep Learning for Identifying Metastatic Breast Cancer," in *Internacional Conference on Learning Representations*, pp. 1–6, 2016.

[9] O. Jiménez del Toro and et al., "Deep Multimodal Case - Based Retrieval for Large Histopathology Datasets," in *MICCAI 2017 workshop on Patch-based image analysis*, vol. 2, pp. 149–157, 2017.

[10] J. Arevalo and et al., "Gated multimodal networks," *Neural Computing and Applications*, vol. 1, 2020.

[11] Y. Cao, S. Steffey, H. Jianbiao, D. Xiao, C. Tao, P. Chen, and H. Müller, "Medical Image Retrieval: A Multimodal Approach," *Cancer Informatics*, vol. 13, pp. 125–136, 2014.

[12] A. Mourão, F. Martins, and J. a. Magalhães, "Multimodal medical information retrieval with unsupervised rank fusion," *Computerized Medical Imaging and Graphics*, vol. 39, pp. 35–45, 2015.

[13] A. G. Seco de Herrera, R. Schaer, D. Markonis, and H. Müller, "Comparing fusion techniques for the ImageCLEF 2013 medical case retrieval task," *Computerized Medical Imaging and Graphics*, vol. 39, no. Medical visual information analysis and retrieval, pp. 46–54, 2015.

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

[15] C. Mosquera-Lopez, S. Agaian, A. Velez-Hoyos, and I. Thompson, "Computer-Aided Prostate Cancer Diagnosis from Digitized Histopathology: A Review on Texture-Based Systems," *IEEE Reviews in Biomedical Engineering*, vol. 8, pp. 98–113, 2015.

[16] J. Ren, I. Hacihaliloglu, E. A. Singer, D. J. Foran, and X. Qi, "Unsupervised Domain Adaptation for Classification of Histopathology Whole-Slide Images," *Frontiers in Bioengineering and Biotechnology*, vol. 7, 2019.

[17] O. Eminaga, M. Abbas, C. Kunder, A. M. Loening, J. Shen, D. Brooks, C. P. Langlotz, and D. L. Rubin, "Plexus Convolutional Neural Network (PlexusNet): A novel neural network architecture for histologic image analysis," *arXiv:1908.09067 [q-bio.QM]*, 2019.

[18] H. Xu, S. Park, and T. H. Hwang, "Computerized Classification of Prostate Cancer Gleason Scores from Whole Slide Images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. PP, no. X, pp. 1–1, 2019.

[19] E. Esteban, A. Colomer, V. Naranjo, C. D. Vera, and U. D. Valencia, "Granulometry-Based Descriptor for Pathological Tissue Discrimination in Histopathological Images," in *IEEE International Conference on Image Processing (ICIP)*, pp. 1413–1417, IEEE, 2018.

[20] O. Jiménez del Toro, M. Atzori, S. Otálora, M. Andersson, K. Eurén, M. Hedlund, P. Rönnquist, and H. Müller, "Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score," in *SPIE Medical Imaging*, pp. 1–9, 2017.

[21] S. Bhattacharjee, H. G. Park, C. H. Kim, D. Prakash, N. Madusanka, J. H. So, N. H. Cho, and H. K. Choi, "Quantitative analysis of benign and malignant tumors in histopathology: Predicting prostate cancer grading using SVM," *Applied Sciences (Switzerland)*, vol. 9, no. 15, 2019.

[22] D. Karimi, G. Nir, L. Fazli, P. C. Black, L. Goldenberg, and S. E. Salcudean, "Deep Learning-Based Gleason Grading of Prostate Cancer from Histopathology Images - Role of Multiscale Decision Aggregation and Data Augmentation," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–14, 2019.

[23] D. Wang, D. J. Foran, J. Ren, H. Zhong, I. Y. Kim, and X. Qi, "Exploring automatic prostate histopathology image gleason grading via local structure modeling," in *International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2015-Novem, pp. 2649–2652, 2015.

[24] P. W. Huang and C. H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *IEEE Transactions on Medical Imaging*, vol. 28, no. 7, pp. 1037–1050, 2009.

[25] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1205–1218, 2012.

[26] P. e. a. Ström, "Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study.," *The Lancet. Oncology*, vol. 2045, no. 19, pp. 1–11, 2020.

[27] K. e. a. Nagpal, "Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer," *npj Digital Medicine*, vol. 2, no. 1, 2019.

[28] H. Källén, J. Molin, A. Heyden, C. Lundström, and Å. Kalle, "Towards grading gleason score using generically trained deep convolutional neural networks," in *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pp. 1163–1167, 2016.

[29] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, 2016.

[30] J. Li, W. Li, A. Gertych, B. S. Knudsen, W. Speier, and C. W. Arnold, "An attention-based multi-resolution model for prostate whole slide image classification and localization," *arXiv:1905.13208 [cs.CV]*, 2019.

[31] G. Saposnik, D. Redelmeier, C. C. Ruff, and P. N. Tobler, "Cognitive biases associated with medical decisions: a systematic review," *BMC Medical Informatics and Decision Making*, vol. 16, no. 1, pp. 1–14, 2016.

[32] H. Müller and T. M. Deserno, "Content-Based Medical Image Retrieval," in *Biomedical Image Processing*, ch. 19, Springer, 2011.

[33] R. Sparks and A. Madabhushi, "Out-of-Sample Extrapolation utilizing Semi-Supervised Manifold Learning (OSE-SSL): Content Based Image Retrieval for Histopathology Images.," *Scientific reports*, vol. 6, p. 27306, jun 2016.

[34] A. Sridhar, S. Doyle, and A. Madabhushi, "Content-based image retrieval of digitized histopathology in boosted spectrally embedded spaces," *Journal of Pathology Informatics*, vol. 6, no. 1, p. 41, 2015.

[35] A. Sridhar, S. Doyle, and A. Madabhshi, "Bossted spectral embedding (BoSE): applications to content-based image retrieval of histopathology," in *IEEE International Symposium on Biomedical Imaging*, pp. 1897–1900, 2011.

[36] R. Sparks and A. Madabhushi, "Content-based image retrieval utilizing explicit shape descriptors: applications to breast MRI and prostate histopathology," *Medical Imaging 2011: Image Processing*, vol. 7962, p. 79621I, 2011.

[37] N. Hegde, J. D. Hipp, Y. Liu, M. Emmert-Buck, E. Reif, D. Smilkov, M. Terry, C. J. Cai, M. B. Amin, C. H. Mermel, P. Q. Nelson, L. H. Peng, G. S. Corrado, and M. C. Stumpe, "Similar image search for histopathology: SMILY," *npj Digital Medicine*, vol. 2, no. 1, pp. 1–9, 2019.

[38] R. Schaer, S. Otálora, O. Jimenez-del Toro, M. Atzori, and H. Müller, "Deep Learning - Based Retrieval System for Gigapixel Histopathology Cases and the Open Access Literature Roger," *Journal of Pathology Informatics*, vol. 9, no. 1, 2019.

[39] R. J. Chen, M. Y. Lu, J. Wang, D. F. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood, "Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *arXiv preprint arXiv:1912.08937*, 2019.

[40] R. Yan, F. Ren, X. Rao, B. Shi, T. Xiang, L. Zhang, Y. Liu, J. Liang, C. Zheng, and F. Zhang, "Integration of multimodal data for breast cancer classification using a hybrid deep learning method," in *International Conference on Intelligent Computing*, pp. 460–469, Springer, 2019.

[41] W.-H. Weng, Y. Cai, A. Lin, F. Tan, and P.-H. C. Chen, "Multimodal multitask representation learning for pathology biobank metadata prediction," *arXiv preprint arXiv:1909.07846*, 2019.

[42] J. A. Vanegas, J. C. Caicedo, F. A. González, and E. Romero, "Histology image indexing using a non-negative semantic embedding," in *MICCAI International Workshop on Medical Content-Based Retrieval for Clinical Decision Support*, pp. 80–91, Springer, 2011.

[43] A. Cheerla and O. Gevaert, "Deep learning with multimodal representation for pancancer prognosis prediction," *Bioinformatics*, vol. 35, no. 14, pp. i446–i454, 2019.

[44] V. H. Contreras, J. S. Lara, O. J. Perdomo, and F. A. González, "Supervised online matrix factorization for histopathological multimodal retrieval," in *14th International Symposium on Medical Information Processing and Analysis*, vol. 10975, p. 109750Y, International Society for Optics and Photonics, 2018.

[45] J. S. Lara, V. H. Contreras O., S. Otálora, H. Müller, and F. A. González, "Multimodal latent semantic alignment for automated prostate tissue classification and retrieval," in *International Conferene on Medical Image Computing and Computer Assisted Intervention*, MICCAI, 2020.

[46] H. Chang, L. Loss, and B. Parvin, "Nuclear segmentation in H&E sections via multireference graph cut (MRGC)," in *International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4, 2012.

[47] J. A. Vanegas, *Large-scale Non-linear Multimodal Semantic Embedding*. Doctoral thesis, Universidad Nacional de Colombia, 2017.

[48] J. A. Vanegas, H. J. Escalante, and F. A. Gonzalez, "Semi-supervised Online Kernel Semantic Embedding for Multi-label Annotation," *Lecture Notes in Computer Science*, pp. 693–701, 2018.

[49] S. Otálora, M. Atzori, A. Khan, O. Jimenez-del Toro, V. Andrearczyk, and H. Müller, "A systematic comparison of deep learning strategies for weakly supervised gleason grading," in *Medical Imaging 2020: Digital Pathology*, vol. 11320, p. 113200L, International Society for Optics and Photonics, 2020.

[50] P. A. Humphrey, "Gleason grading and prognostic factors in carcinoma of the prostate," *Modern pathology*, vol. 17, no. 3, pp. 292–306, 2004.

[51] G. Sauter, S. Steurer, T. S. Clauditz, T. Krech, C. Wittmer, F. Lutz, M. Lennartz, T. Janssen, N. Hakimi, R. Simon, *et al.*, "Clinical utility of quantitative gleason grading in prostate biopsies and prostatectomy specimens," *European urology*, vol. 69, no. 4, pp. 592–598, 2016.