A DEEP REPRESENTATION MODEL TO CLASSIFY PROSTATE CANCER GLEASON DEGREES

FABIAN ANDRÉS LEÓN PÉREZ

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA

2022

DEEP REPRESENTATION MODEL TO CLASSIFY PROSTATE CANCER GLEASON DEGREES

FABIAN ANDRÉS LEÓN PÉREZ

Research work in partial fulfillment of the requirements for the degree of:
Magíster en Ingeniería de Sistemas e Informática

Advisor:
Fabio Martínez Carrillo
Ph.D in Systems and Computer Engineering

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2022

ACKNOWLEDGEMENTS

The author expresses his acknowledgement:

First of all, to God, for giving me the strength to carry on. to my parents Oliva Pérez Pérez and Henry León Lizcano for their unconditional love and effort that they dedicated to me during all these years, for their advice, their company, and all the values and principles that they taught me throughout my life.

Thanks to my brothers Jhon Henry León Pérez and Diana Marcela Rodríguez Pérez for their company and support.

To Professor Fabio Martínez Carrillo and the Biomedical Image, Vision, and Learning Laboratory research group, for all the shared knowledge, teachings, and patience during this process.

Thanks to Tatiana Carolina Gelvez Barrera and her family, for their company, support, guidance, and patience throughout this work.

Thanks to Virginia Duarte Quintero for all her care, advice, values, principles, and examples that accompanied me throughout my life.

Finally, thanks to the Universidad Industrial de Santander (UIS) and other staff who made this process possible.

CONTENTS

page

LIST OF FIGURES

page

LIST OF TABLES

page

LIST OF APPENDICES

page

ABSTRACT

TITLE: A DEEP REPRESENTATION MODEL TO CLASSIFY PROSTATE CANCER GLEASON DEGREES [*]

AUTHOR: FABIAN ANDRÉS LEÓN PÉREZ [**]

KEYWORDS: PROSTATE CANCER, DEEP LEARNING, EMBEDDED REPRESENTATIONS, GLEASON SCORE.

DESCRIPTION: Histopathological image analysis is the most accurate method to characterize, diagnose, and quantify cancer stages. The Gleason scale is the gold standard, in clinical routine, to stratify the disease aggressiveness, allowing the pathologists to weight image segments according to the architectural and morphological arrangement of the cancer cells. Nonetheless, this task is highly subjective, variable and greatly depends on the expertise and experience of the pathologists, which affects the diagnosis of the disease. For instance, recent studies report a discordance level between 41 pathologists of up to 0.45 in terms of the kappa value over a set of 38 images. Recently, Deep learning models have emerged as an alternative to classify and support cancer stratification tasks, following the Gleason system. However, these models remain limited to learn the complexity of observed histological patterns because of the high variability to represent each cancer degree, with marked overlapping representation among classes. Besides, the representations may be biased to a specific observer, and the training samples may result limited with an inherent class imbalance, available in clinical scenarios. This work introduces a multitask learning approach to represent intra and inter-Gleason relationships from more challenging samples, following two branches: a triplet loss and a conditioned cross-entropy. In such sense, the approach uses a semi-hard triplet distance metric as the main task to tackle Gleason stratification, training with challenging positive and negative patches on a particular Gleason level. Then, an auxiliary task was herein proposed to regularize embedding space, allowing to deal with the high inter and intra appearance of the four grades, considered in Gleason stratification. As a regularizer, the proposed approach uses a cross-entropy rule. The proposed approach was validated on a public dataset with 886 tissue micro-array spots, that in the test subset was delineated independently by two expert uropathologists, according to the Gleason grading system.

---

[*] Research work

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Advisor: Fabio Martínez Carrillo, Ph.D.

The proposed approach achieves a general classification of average accuracy of 66% and 64%, for two experts without statistical difference. Additionally, the proposed approach achieved an average accuracy of 73% in patches where both pathologists were agree, showing the robustness of patterns learning from the approach.

RESUMEN

TÍTULO: UN MODELO DE REPRESENTACIÓN PROFUNDA PARA CLASIFICAR GRADOS DE GLEASON DE CÁNCER DE PRÓSTATA. *

AUTOR: FABIAN ANDRÉS LEÓN PÉREZ **

PALABRAS CLAVE: CÁNCER DE PRÓSTATA, APRENDIZAJE PROFUNDO, REPRESENTACIONES EMBEBIDAS, PUNTUACIÓN DE GLEASON.

DESCRIPCIÓN: El análisis de imágenes histopatológicas es el método más preciso para caracterizar, diagnosticar y cuantificar los estadios del cáncer. La escala de Gleason es el estándar de oro, en la rutina clínica, para estratificar la agresividad de la enfermedad, lo que permite a los patólogos ponderar segmentos de imágenes de acuerdo con la disposición arquitectónica y morfológica de las células cancerosas. No obstante, esta tarea es altamente subjetiva, variable y depende en gran medida de la experticia y experiencia de los patólogos, lo que afecta el diagnóstico de la enfermedad. Por ejemplo, estudios recientes informan un nivel de discordancia entre 41 patólogos de hasta 0,45 en términos del valor kappa en un conjunto de 38 imágenes. Recientemente, los modelos de aprendizaje profundo han surgido como una alternativa para clasificar y apoyar las tareas de estratificación del cáncer, siguiendo el sistema de Gleason. Sin embargo, estos modelos siguen estando limitados para aprender la complejidad de los patrones histológicos observados debido a la alta variabilidad para representar cada grado de cáncer, con una marcada superposición de representación entre clases. Además, las representaciones pueden estar sesgadas a un observador específico, y las muestras de entrenamiento pueden resultar limitadas con un desequilibrio de clase inherente, presente en escenarios clínicos. Este trabajo presenta un enfoque de aprendizaje multitarea para representar las relaciones intra e inter-Gleason de las muestras más desafiantes, siguiendo dos ramas: una pérdida de tripletas y una entropía cruzada condicionada. En tal sentido, el enfoque propuesto usa una métrica de distancia de tripletas semidifícil como tarea principal para abordar la estratificación de Gleason, entrenada con parches positivos y negativos desafiantes en un nivel particular de Gleason. Luego, se propuso aquí una tarea auxiliar para regularizar el espacio embebido, que permitiera lidiar con la alta inter e intra-apariencia de los cuatro grados, considerados en la estratificación de Gleason. Como regularizador, el enfoque propuesto

---

* Trabajo de investigación

** Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Fabio Martínez Carrillo, Ph.D. Codirector:

utiliza una regla de entropía cruzada. El enfoque propuesto se validó en un conjunto de datos públicos con 886 microarreglos de tejido, que en el subconjunto de prueba fue delineado de forma independiente por dos uropatólogos expertos, de acuerdo con el sistema de clasificación de Gleason. El enfoque propuesto logra una clasificación general de exactitud promedio de 66% y 64%, para dos expertos sin diferencia estadística. Además, el enfoque propuesto logró una exactitud promedio del 73% en parches en los que ambos patólogos estaban de acuerdo, lo que demuestra la robustes de los patrones que aprenden del enfoque.

INTRODUCTION

Prostate cancer is the second most common cancer with more than one million new diagnosed cases and more than 350,000 deaths associated with this disease each year. In Colombia, 14,460 new diagnosed cases were reported in 2020 [1], being the second most common cancer in men with an average age of 66 years old [2]. The analysis of histopathological images, from biopsies, is currently the definitive method to characterize and quantify cancer disease [3]. These images are obtained by taking a sample of prostate tissue and staining it with Hematoxylin and Eosin (H&E) for further microscopic analysis. Stained tissue allows experts to visualize a wide range of characteristics of the cytoplasmic, nuclear, and extracellular matrix, which are important biomarkers in the diagnosis of cancer.[4]

The Gleason grading system is the main prognosis standard to quantify and characterize cancer disease from histological images [5]. This system is based on glandular analysis, which characterizes among others, basic geometric structures such as size, color, and shape. In general, this grading system is divided into two general scales: from [1-5] to characterize local regions and from [6-10] to report a global sample analysis. In the first scale ([1-5]) the primary levels present

---

[1] International Agency for Research on Cancer. *The Global Cancer Observatory "GLOBOCAN"*. `https://gco.iarc.fr/today/home`. 2018.

[2] Elena Payá Bosch. "Diseño y desarrollo de un sistema automático de segmentación de glándulas histológicas para identificar el cáncer de próstata en una etapa inicial". In: (2019).

[3] Ana Isabel Ruiz et al. "Actualización sobre cáncer de próstata". In: *Correo Científico Médico* 21.3 (2017).

[4] Andrew H Fischer et al. "Hematoxylin and eosin staining of tissue and cell sections". In: *Cold spring harbor protocols* 2008.5 (2008), pdb–prot4986.

[5] Rebecca Arora et al. "Heterogeneity of Gleason grade in multifocal adenocarcinoma of the prostate". In: *Cancer: Interdisciplinary International Journal of the American Cancer Society* 100.11 (2004), pp. 2362–2366.

well-formed glands, while the last stages do not present glands, instead there are solid nets and necrosis. The second scale ([6-10]) is dedicated output general prognosis score, as the sum of the two most predominant patterns in the sample. For example, A diagnosis of grading 7 is the result of find grades 3 and 4 in the sample (3 + 4 or 4 + 3) [6].

Despite the stratification and coarse quantification from the Gleason system, there is remaining and intrinsic diagnosis variation, mainly associated with pathologist subjectivity to characterize particular visual grade patterns. This variation could be dramatic and inside on clinical decision about following the disease and patient treatment. In fact, several studies have reported this inter pathologist variation, for instance, with a moderate kappa agreement of 0.68, among eight pathologists, in the diagnosis of 81 prostate slides [7]. Also, in [8] 150 slides were sent to three pathologists in two phases, reporting a kappa value of 0.25 and 0.52 respectively, showing an improvement after a training course. Another study among 41 pathologists only reached a value of 0.435, following Gleason standard stratification [9]. This high level of discordance may come from expert biases, such as avoiding extreme ranges, preference for numbers, and confirmation biases among others [10].

---

[6] Jonathan I Epstein et al. "The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma". In: *The American journal of surgical pathology* 40.2 (2016), pp. 244–252.

[7] J Melia et al. "A UK-based investigation of inter-and intra-observer reproducibility of Gleason grading of prostatic biopsies". In: *Histopathology* 48.6 (2006), pp. 644–654.

[8] Alireza Abdollahi et al. "Inter-observer reproducibility before and after web-based education in the Gleason grading of the prostate adenocarcinoma among the Iranian pathologists." In: *Acta Medica Iranica* (2014), pp. 370–374.

[9] William C. Allsbrook et al. "Interobserver reproducibility of Gleason grading of prostatic carcinoma: General pathologist". In: *Human Pathology* 32.1 (2001), pp. 81 –88. DOI: https://doi.org/10.1053/hupa.2001.21135.

[10] Richard J Zarbo, Frederick A Meier, and Stephen S Raab. "Error detection in anatomic pathology". In: *Archives of Pathology and Laboratory Medicine* 129.10 (2005), pp. 1237–1245.

Hence, automatic or semi-automatic methods to classify Gleason score and to stratify and quantify cancer patterns are demanding in routine and clinical analysis. In literature have been reported several approximations, for instance, designing descriptors to code nuclear arrangement, glandular structures, and image texture, which thereafter are mapped to a support vector machine to classify between three and four Gleason grades[11]. In the same line, in [12] was proposed structure tissue features to support the Gleason grading system. These approaches are however limited to associate complex grading patterns with simple and isolated histopathological primitives. Recently, deep learning approaches have been designed to represent tissue microarray images and automatically classify among three, four, and five Gleason grades [13]. Also, Bulten et al. [14] proposed a semi-supervised approach to classify patterns in complete slide images, using convolutional nets to segment and classify gland structures. In [15] was conducted a study to classify local patches using a teacher-student model in a self-supervised task. In spite of advances in deep representations, the variability of intra and inter-Gleason scores remain as an open problem, mainly due to rigid learning with stratified and balanced datasets, which results in unrealistic in the clinical domain.

---

[11] Scott Doyle et al. "Automated grading of prostate cancer using architectural and textural image features". In: *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. 2007, pp. 1284–1287.

[12] Shivang Naik et al. "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology". In: *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. 2008, pp. 284–287.

[13] Eirini Arvaniti et al. "Automated Gleason grading of prostate cancer tissue microarrays via deep learning". In: *Scientific reports* 8.1 (2018), pp. 1–11.

[14] Wouter Bulten et al. "Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study". In: *The Lancet Oncology* 21.2 (2020), pp. 233–241.

[15] Julio Silva-Rodríguez et al. "Self-learning for weakly supervised gleason grading of local patterns". In: *IEEE journal of biomedical and health informatics* 25.8 (2021), pp. 3094–3104.

This work introduces a novel strategy that learns a feature embedding space, from a triplet loss representation, taking batches of positive and negative histological patches, from more challenging Gleason samples. The proposed learning scheme allows to deal with imbalanced set, learning embedding spaces that maximizes differences among disease stages. This approach achieves remarkable results to discriminate among three and four grades, the most challenging levels in clinical routine. Nonetheless, the challenging variability of this problem produces overfitting spaces when the model tries to classify all Gleason grades. To avoid this limitation, an auxiliary task is herein integrated to deal with high inter and intra appearance and structural variations of the Gleason system. Specifically, a cross-entropy is implemented as an auxiliary classification task, helping with inter-class variability of samples, while adding robust representations to the triplet loss main task. This new low-dimensional topological embedding space aims to maintain close the Gleason grade samples belonging to the same class while maximizing the distance among other classes. The achieved deep representation robustly learns semantic relationships, avoiding expert bias, supported by the self-representation capability, learned from intra-class visual patterns. In this case, the model robustness is evaluated in a multi-class classification task across the different Gleason grades showing an average accuracy of 66% and 64%, for two experts without statistical difference.

1. FUNDAMENTALS AND PREVIOUS WORK

1.1. Prostate cancer

The prostate is a small gland located under the bladder and in front of the rectum, responsible of produces the seminal fluid that nourishes and transports sperm. Prostate cancer starts as the abnormal growth of cells out of control that can be spread across the body colonizing distant organs [16]. Around the world prostate cancer caused 375,304 deaths (14.8% of all the deaths of cancer in men), including 3,846 deaths in Colombia (6.9% of all the deaths of cancer in men)[1]. This disease is rarely diagnosed in people younger than 50 years old (<0·1% of the diagnoses), being the most common diagnosis in people with more than 65 years (85% of the diagnoses) [17]. Today, there is not a clear risk and factors of prostate cancer, but several studies has established the most common causes:

- **Age:** It's the most common malignancy diagnosis in old people [18], reporting that almost over 30% of 50 years old men that died for other causes were diagnosed with prostate cancer after autopsy[19].
- **Race** For reasons that are not determined, prostate cancer risk varies among different racial groups. For example, African people have more risk to be diagnosed with cancer

---

[16] Ginevra Doglioni, Sweta Parik, and Sarah-Maria Fendt. "Interactions in the (pre) metastatic niche support metastasis formation". In: *Frontiers in oncology* 9 (2019), p. 219.

[17] Henrik Grönberg. "Prostate cancer epidemiology". In: *The Lancet* 361.9360 (2003), pp. 859–864.

[18] Freddie Bray et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 68.6 (2018), pp. 394–424.

[19] Peter T Scardino, Robert Weaver, and A Hudson M'Liss. "Early detection of prostate cancer". In: *Human pathology* 23.3 (1992), pp. 211–222.

and also, present a more advanced and aggressive disease [20].

- **Family history:** About 20% of prostate cancer diagnoses come from patients with a family history of the disease. However, the development of the disease can not be only associated with genetics, but also can be associated with shared environmental patterns and lifestyles

- **genetics:** Other studies report that the risk of prostate cancer for genetics it's about 5%, which can increase when high-penetrance genetic "risk" alleles are inherited [21].

- **Obesity:** Most of the time, obese people present different levels of metabolic and sex hormones, that are involved in prostate development[22]. Also, some studies addressed a more aggressive progression of the disease and worse outcomes in obese people[23].

The early detection and the correct diagnosis of prostate cancer are highly correlated with patients' survival [24]. Actually, as a clinical control is normal that men with more than 50 years old have any early and fast prostate cancer testing such as prostate-specific antigen (PSA) or digital rectal exam (DRE). However, this cancer screenings report a high risk of over-diagnosis and over-treatment [25]. Exist several methods and procedures to correct diagnosis of cancer

---

[20] Ina Wu and Charles S Modlin. "Disparities in prostate cancer in African American men: what primary care physicians can do". In: *Cleve Clin J Med* 79.5 (2012), pp. 313–20.

[21] Gayathri Sridhar et al. "Association between family history of cancers and risk of prostate cancer". In: *Journal of Men's Health* 7.1 (2010), pp. 45–54.

[22] Russell Bailey McBride. *Obesity and aggressive prostate cancer bias and biomarkers*. Columbia University, 2012.

[23] Peter Greenwald. "Clinical trials in cancer prevention: current results and perspectives for the future". In: *The Journal of nutrition* 134.12 (2004), 3507S–3512S.

[24] Takashi Fukagai et al. "Discrepancies between Gleason scores of needle biopsy and radical prostatectomy specimens". In: *Pathology international* 51.5 (2001), pp. 364–370.

[25] Michael J Barry et al. "Screening for prostate cancer–the controversy that refuses to die". In: *New England Journal of Medicine* 360.13 (2009), p. 1351.

aggressiveness, such as ultrasound or magnetic resonance imaging but this analysis remains limited in terms of specificity, characterization, and prognosis of the disease [26].

Currently, biopsies are the main method to quantify the aggressiveness of the disease, through histological analysis of microscopic images. This procedure is performed using a thin needle that is inserted into the prostate to collect tissue to generate microscopical observation analysis. The digitization of such microscopical slides results in histopathology images that allow among others to observe and analyze suspicious prostate tissue. These samples are namely stained with a combination of Hematoxylin and Eosin (H&E) to enhance the observation of glands, nuclei, cytoplasm, and connective tissue. Pathologists then are dedicated to localizing and characterizing prominent findings, associated with cancer disease, and related to structural disorders of gland segments. Figure 1 is illustrated two typical histological samples, identifying the main glandular structures of interest relate to disease findings. These findings are stratified and grouped into the Gleason standard protocol to correlate observations with cancer severity, as described in the next subsection.

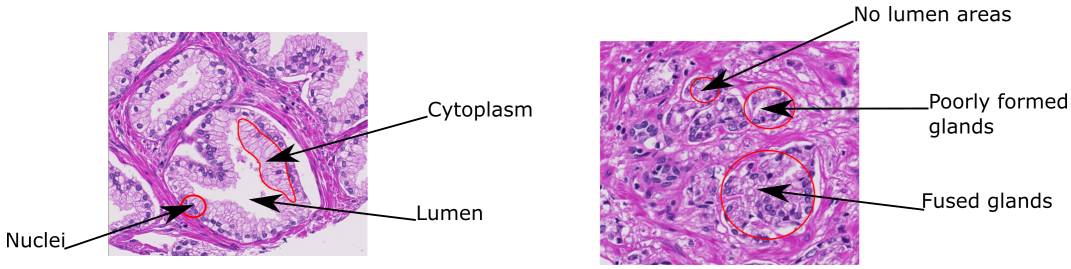

Figure 1. In the left, benign tissue with well defined glands, formed by lumen, cytoplasm and nuclei. In the right a malignant tissue, with glands without lumen areas, fused and malformed

[26] Enrique Bley and Andrés Silva. "Diagnóstico precoz del cáncer de próstata". In: *Revista médica clínica Las Condes* 22.4 (2011), pp. 453–458.

1.2. Gleason grading system

The Gleason grading system is the dominant method world around to measure and report the prostate cancer aggressiveness w.r.t tumor areas observations from histopathological images. The grading is carried out at low magnification (X10-40) allowing to observe mainly glandular structure patterns and their evolution caused by the disease [27]. According to structural observed patterns, the Gleason system coarsely identifies five primary observational classes, which correlate with levels of cancer disease, annotated from one to five. Each analyzed image can have one or several Gleason primary levels. In such way, each of the observed regions is labeled with one level, obtaining a regional quantification of each slide [28][29][30]. A set of primary grade examples are illustrated in Figure 2. To summarize a final diagnosis, a second Gleason scale is defined by summing up the two most predominant primary regions, into a scale from [6-10]. For instance, for the first stage of the disease, the Gleason score establishes a level six, obtained from predominant primary patterns between $(1-3)$. In contrast, an advanced stage could be evidenced as a level nine or then, obtained from different combinations of primary levels, such as: $\{(4+5),(5+4),(5+5)\}$. This system has been recurrently updated by the International Society of Urological Pathology [31]. A more detailed description of primary observational scales is thereafter:

---

[27] Peter A Humphrey. "Gleason grading and prognostic factors in carcinoma of the prostate". In: *Modern pathology* 17.3 (2004), pp. 292–306.

[28] Jonathan I Epstein. "An update of the Gleason grading system". In: *The Journal of urology* 183.2 (2010), pp. 433–440.

[29] Oleksandr N Kryvenko and Jonathan I Epstein. "Prostate cancer grading: a decade after the 2005 modified Gleason grading system". In: *Archives of pathology & laboratory medicine* 140.10 (2016), pp. 1140–1152.

[30] CF Kweldam, GJ van Leenders, and T van der Kwast. "Grading of prostate cancer: a work in progress". In: *Histopathology* 74.1 (2019), pp. 146–160.

[31] Jonathan I Epstein et al. "A contemporary prostate cancer grading system: a validated alternative to the Gleason score". In: *European urology* 69.3 (2016), pp. 428–435.

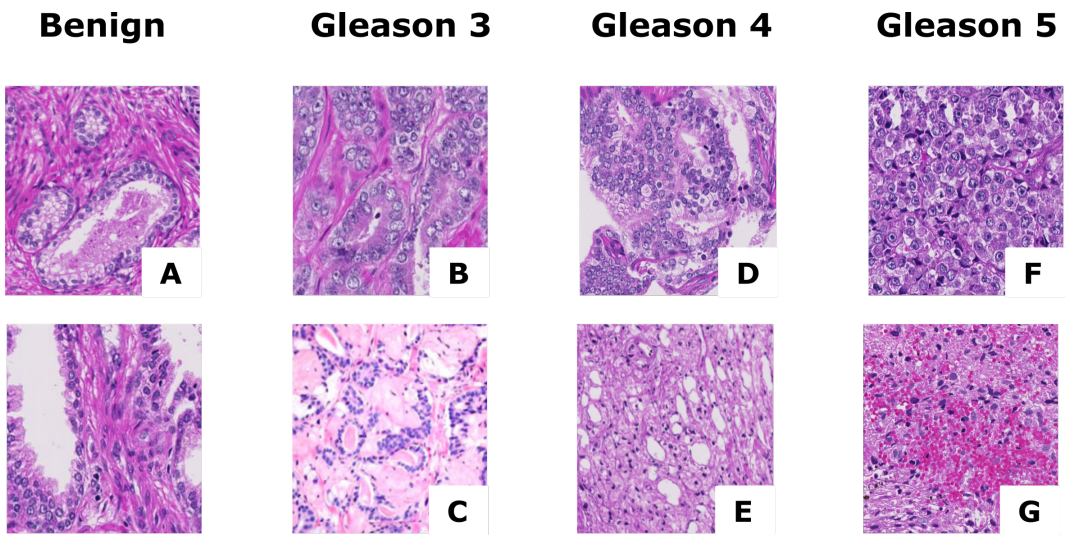

Figure 2. A) well-formed and uniform glands, B) discrete glans with stroma separation, C) collagenous micronodular, D) fused and poorly formed glands, E) cribriform pattern, F) solid nests, G) cell necrosis

- **Gleason patterns 1 and 2** have rare growth patterns of medium-sized acini. In these stages is observed stroma separation with an average distance of less than one gland diameter. In this primary patterns are described also size variations of the neoplastic glands and a slight irregularity in the periphery of the nodule. However, these patterns are nowadays not reported by recommendation in the last Gleason update. This decision is mainly supported by the poor reproducibility even among experts, poor correlation with the radical prostatectomies, and such patterns are not always observed in tissue samples[32].

- **Gleason pattern 3** is a clearly infiltrative neoplasm, *i.e.,* malignant cells start to surround and invade non-cancerous tissue. In this pattern is observed well-differentiated glands but with irregular shape and size, showing stroma separation. The observed intervening in the stroma and in some benign glands is associated with neoplastic glands density. The carcinoma glands are also separated by benign glands comprising less than 50% of the tumor area. Finally, Desmoplasmatic stroma and collagenous micronodular

[32] Jonathan I Epstein. *Gleason score 2–4 adenocarcinoma of the prostate on needle biopsy: a diagnosis that should not be made.* 2000.

also can be shown in this pattern with entrapped acinar carcinoma (See en Figure 2.2.B, 2.2.C)

- **Gleason pattern 4** shows poor formation of glands with very few lumen areas and elongated nests. Also, in this pattern are observed fused glands, displaying multiple merged glandular strands. This grade is described also by dilated cancer glands with variable sizes and nests with lumen formation. Finally, this grade are taking into account the expansible areas of carcinoma cells without intervening stroma or vasculature (see Figure 2.2.D, 2.2.E).

- **Gleason pattern 5** is the least differentiated pattern of prostate carcinoma. This level is mainly characterized by solid and larger nests, sheets, and comedo necrosis in single glands and cords without lumen formation. This pattern doesn't show glandular features as lumen formation or gland contours, instead, individual cells are shown and neoplasms without glandular differentiation (see in Figure 2.2.F, 2.2.G).

Despite of reported advantages of stratification and coarsely quantification of this grading scale, and its universal use in clinical routine, there exists also evidence of large intra and inter-observer variability, leading to misdiagnosis and remarkable differences depending on the expertise of the observed. For instance, in an intra and inter-observer investigation [7], were sent 81 prostate slides to 11 experts interested in urological histopathology. The experts report a kappa value of 0.54 showing high variation in intermediate grades 2-4 with a kappa value of 0.33. In [9] were sent 38 slides to a set of 41 general pathologists. The ground truth of the 38 cases was determined by consensus for 10 urological pathologists. In the end, a kappa of 0.435 was reported for the 41 experts. In [8] three pathologists classify a total of 150 tissue samples. In the first step, the pathologist only reaches a kappa value of 0.25 (fair agreement). Then all pathologists attend a free web-based course, reporting a kappa of 0.52 (moderate agreement) showing a remarkable improvement.

1.3. Computational support for Gleason pattern classification

Computational tools have emerged as an ideal alternative to support the quantification from Gleason system, pretending to reduce subjectivity, inter pathologist variation, and also augmenting sensibility on visual pattern analysis. These systems also have the possibility to better support the transition definition among disease levels. Different strategies have been proposed in the literature, starting with the handcrafted or engineering approaches, mainly dedicated to the model gland and cell features to create descriptors for the classification of Gleason patterns. For instance, in [33] was proposed an automatic system for detecting, segmenting, and extracting gland features, in Lab color space, and using glandular component features, classified with a KNN classifier. Also, in [34] were coded features, as k-means centroids to represent tissue and cell properties. These features were used to train a SVM to classify benign or malign tissue. Nguyen et al. [35] use the glandular structure including lumen areas, and epithelial cells, among other features, to classify benign tissue and Gleason pattern three and four using SVM. Despite important advances in modeling histological features, these approaches are limited to some restricted conditions such as color, light, and shapes that not always are satisfied. Furthermore, gland structures are poorly formed in high levels of the disease being difficult to extract gland features in these scores.

Recently, deep learning approaches have shown remarkable results to represent complex visual features from a hierarchical and non-linear representation. In prostate cancer, different

---

[33] Kien Nguyen, Anil K Jain, and Ronald L Allen. "Automated gland segmentation and classification for gleason grading of prostate tissue images". In: *2010 20th International Conference on Pattern Recognition*. IEEE. 2010, pp. 1497–1500.

[34] Subrata Bhattacharjee et al. "Analysis of Biopsy Tissue Images based on Color Moment Technique and Morphology of Cell Nuclei". In: ().

[35] Shivang Naik et al. "Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information". In: *MIAAB workshop*. Citeseer. 2007, pp. 1–8.

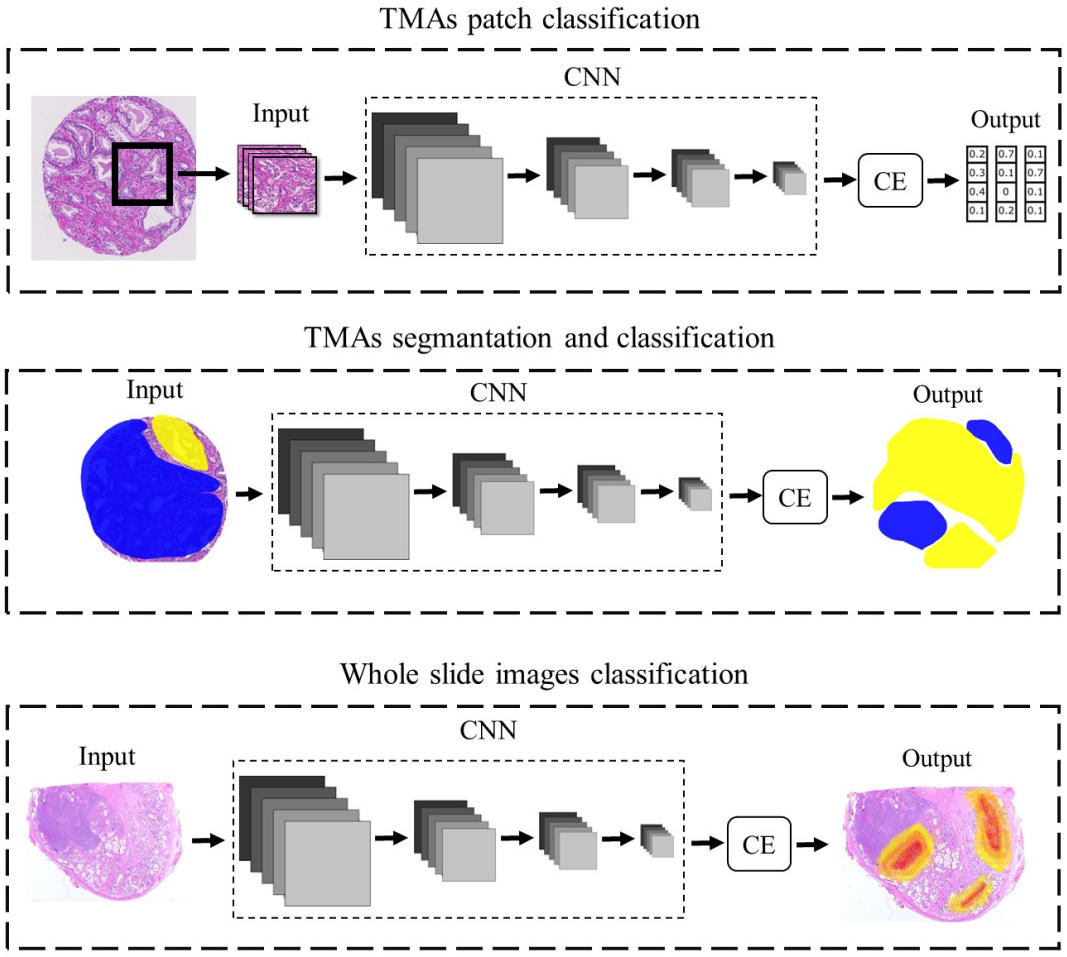

Figure 3. In general, the reported strategies in prostate cancer classification may group as: TMA patch classification, TMA segmentation, and Whole slide image classification. Typical models and tasks are limited to rigid models and training schemes. These strategies don't include a mechanism to deal with unbalanced data sets and they may be biased by a specific observer during training.

strategies and tasks have been proposed, which may be grouped in TMA patch classification, TMA segmentation, and Wholes slide image classification. In Figure 3 is illustrated the general pipeline of such strategies. These models and strategies normally follow rigid training and optimization schemes such as cross-entropy (CE) that don't deal directly with imbalanced data and human biased present in the annotation process. For example Eirini, et al. [13] propose a mobilnetV3 architecture to classify Gleason patches from tissue micro-arrays. The model was

trained using annotations from one pathologist and evaluated with two expert references. This approach however uses a classical training scheme that results sensible to unbalanced data and high variability in annotation classes. In [36] was proposed two integrated phases to procure a robust deep learning system able to obtain Gleason maps regarding whole slide images. First, a searching architecture is used to adjust the parameter of an Xception network. Then the trained Xception is used to predict patches with a unique grade. Then, the predicted patches are assembled in heat maps of the slide. From such maps are extracted features related to the tumor percentage and the relative percentages of each Gleason pattern. These cumulative steps result in a sophisticated framework that may propagate errors from patch prediction to measures over slide map. David, et al. [37] also proposed an Xception to predict patches into an assistance biopsies analysis tool. The patches are used to reconstruct the biopsy, allowing to measure the percentage of Gleason tumors. The tool validation evidence the necessity to assist routine procedures on biopsy analysis. In [38] was used whole slide images with a staining normalization to train 3 independent convolutional architectures. Two compact nets, the Lenet and Alexnet were dedicated to learn local features, while the third architecture was selected the GoogleNet to learn more global and complex features such as gland structures and cell arrangement. Cai, et al. [39] proposes to use unstained prostate tissues to avoid tissue fixation and staining problems, using multiphoton microscopy (MPM) based on second-harmonic generation

---

[36] Kunal Nagpal et al. "Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens". In: *JAMA oncology* 6.9 (2020), pp. 1372–1380.

[37] David F Steiner et al. "Evaluation of the Use of Combined Artificial Intelligence and Pathologist Assessment to Review and Grade Prostate Biopsies". In: *JAMA Network Open* 3.11 (2020), e2023267–e2023267.

[38] Oscar Jiménez del Toro et al. "Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score". In: *Medical Imaging 2017: Digital Pathology*. Vol. 10140. International Society for Optics and Photonics. 2017, 101400O.

[39] Jianyong Cai et al. "Automated Gleason grading of prostate cancers via deep learning in label-free multiphoton microscopic images". In: *Multiphoton Microscopy in the Biomedical Sciences XX*. vol. 11244. International Society for Optics and Photonics. 2020, 112442A.

(SHG) to get an optical high-resolution image. Then, an Inception-V3 was used to classify the Gleason patterns. This procedure nevertheless has a complex experimental setup and there is no evidence to be used in clinical routine. Another example is proposed by Ali, et al. [40] that uses a DeepLabV3 with MobileNetV2 as a backbone to classify Gleason scores. In [15] [41] is proposed Weakly Supervised models to classify Gleason grades from whole slide images. These models are implemented under the multiple instance learning (MIL) paradigm, grouping training instances as bags, considering a positive bag if at least one instance is positive. However, the models are optimized and evaluated to a specific observer, losing the inter observer perspective. Bultren, et al. [14] propose a semi-automated labeling method to classify Gleason scores in an entire prostate biopsy. For doing so, the authors proposed a workflow that includes several processing steps and deep learning nets dedicated to different tasks, such as the extraction of background, delineation of tumor areas, and finally the classification of Gleason patterns. In spite of advances from deep representations, the variability intra and inter Gleason scores is still an open and challenging problem. Among others, these approaches require a significant amount of data to deal with variability in a particular domain, fact that restricts the performance of histopathological problems. Also, in realistic scenarios, in this domain, there is significant unbalanced data, which results in a limitation for typical deep training strategies. The next table shows a brief summary of the previous works.

---

[40] Ali Asghar Khani et al. "Towards Automatic Prostate Gleason Grading Via Deep Convolutional Neural Networks". In: *2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*. IEEE. 2019, pp. 1–6.

[41] Yechan Mun et al. "Yet another automated Gleason grading system (YAAGGS) by weakly supervised deep learning". In: *NPJ Digital Medicine* 4.1 (2021), pp. 1–9.

Autor	Description	Gleason task
Julio, et al. 2021	Weakly Supervised approach using teacher-student model	Gleason classification
Mun, et al. 2021	Weakly Supervised approach using multiple instance learning	Gleason classification
Cai, et al. 2020	InceptionV3 model to classify unstained tissue using multiphoton microscopy	Tissue classification
Kunal, et al. 2020	Deep learning system based on Xception, to classify and extract Gleason tumor percentages to train a SVM	Groups classification
David, et al. 2020	AI-based assistance tool using Xception architecture for the classification of prostate biopsies	Groups classification
Bultren, et al. 2018	semi-automated labeling method to classify Gleason scores in an entire prostate biopsy using three different concatenate neuronal networks.	Segmentation and classification
Eirini, et al. 2018	MobilNetV3 architecture to classify Gleason patches from tissues microarrays, evaluated using two pathologists.	Patch classification
Nathan, et al. 2018	Gland segmentation architectures comparison to classify Gleason patterns in whole slide images.	Glands classification
Oscar, et al. 2017	Three independent convolutional networks used to classify whole slide images with staining normalization	Binary tissue classification
Subrata, et al. 2015	K-means and Watershed algorithms used to color-based segmentation and separation of overlapping cells, classifying features with SVM	Benign and malign gland classification

Table 1. Most recent studies include the use of weakly and unsupervised methods for the classification of prostate cancer. these methods aim to reduce the inherent error include in the pathologist annotations, helping to select the most representative samples.

2. RESEARCH PROBLEM

Histopathological images are the main method to characterize and quantify cancer disease. Regarding prostate cancer, the Gleason scoring system allows to measure and classify the aggressiveness of the disease from glandular and cellular structure observations. In such standard, a set of visual patterns, associated with pathology evolution, are defined to support pathologist disease characterization, over particular microscopic tissue samples. Nonetheless, this kind of pattern differentiation is really complex and challenging even for an expert pathologist, being highly expert dependent, which introduces an inherent high variation in the diagnosis of the disease. In fact, in the literature, there exist multiple works that quantitatively evidence such variability in such characterization. For instance, reporting kappa agreement values until 0.52 among 3 expert pathologists in a total of 150 samples, or a study that reports a dramatic agreement of 0.435 in a total of 38 slides, among 41 pathologists.

Hence, computational approaches emerge as an alternative to support observational pattern quantification and to help with diagnosis reports from the Gleason scale. The problem is then to define models and build visual representations that deal with the high inter and intra variability for each of the scale levels. In seminal works, the engineering methods were proposed to follow precise cell modeling, aiming to emulate and recover pathologist patterns. However, these approaches are limited to representing some specific histological primitives, resulting in rigid methodologies to support prostate cancer levels. More recently, the deep learning representations have shown some promising results in the stratification of cancer disease. The main advantage of such patterns is the learning of very deep and hierarchical representations that cover a wide feature spectrum that altogether represents particular cancer instances.

Despite reported advances, there exist a significant limitation to support classification because of the high variability of classes. Also, such deep architectures to successfully achieve a particular classification task require a huge amount of training annotated data, which should be stratified among different classes. These requirements are not easily achieved in a histopathological

domain, where the annotation task is tedious, highly variable among pathologists, and during the clinical routine is obtained unbalanced datasets. To illustrate the complexity of Gleason patterns classification, Figure 4 illustrated an embedding space that results from training an inception network on a set of 641 samples with three Gleason grades. As observed there exist a total overlapping among classes, which suggests the study of new representation alternatives to achieve a better class representation, overall in close Gleason degrees.

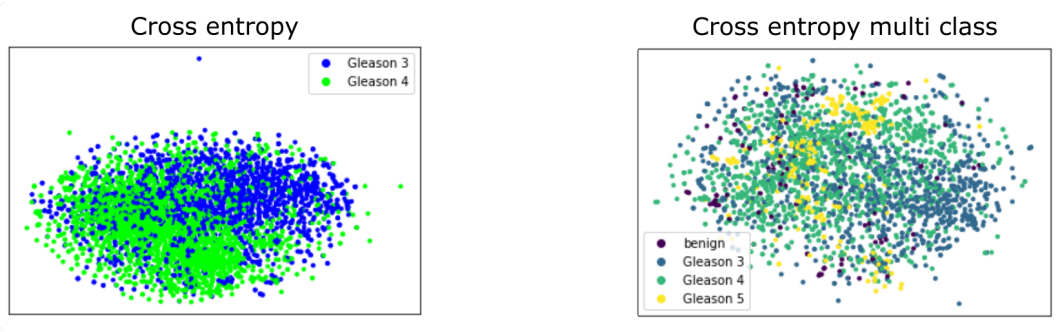

Figure 4. Embedding Gleason representation from a typical classification network. High semantic overlapping and diffused class boundaries are observed, leading classification variability and weakens Gleason patterns representations.

Research Question

How much contribute different training schemes over deep representation to classify Gleason patterns?

3. OBJECTIVES

General Objective

To propose a deep strategy to classify visual features following a Gleason scale stratification.

Specific Objectives

- To select a public histopathological dataset annotated by pathologists according to the Gleason scale.
- To define a deep training scheme to deal with intra and inter-Gleason level variability.
- To train a backbone deep convolutional net from the defined training scheme to support histological image classification.
- To project the resultant embeddings on a low-dimensional space as a strategy to separate samples among different Gleason classes.
- To evaluate the proposed strategy in terms of the capability to classify visual regions according to the Gleason stratification.

4. PROPOSED APPROACH

Learning semantic distances may be a key issue in histopathological Gleason classification to deal with pattern variability without constrained stratification and artificial balanced of data. For instance, the resulting embedding space from different deep training strategies is observed in Figure 4. As observed, the typical representation is overlapped from a classical classification framework with no separation of classes. Specifically, the reported classification results are the contribution of boundary and outlier samples that result in differences among other classes but especially of the same class. This research introduces a deep representation dedicated to built an embedding representation that integrates an auxiliary task to deal with the high inter and intra appearance of the Gleason system. Firstly, the learning of a distance metric, from a triplet loss scheme, allows fixing a representation from sample batches of positive and negative histological patches. Then an auxiliary task, that follows a cross-entropy rule, works as a regularizer to help with the inter-class variability of samples while adding robust representations to the main task. *The complete content of this section has been published as research articles in Biomedical Physics & Engineering Express*[42] *and also presented in the 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)* [43].

4.1. Learning distance metric

New distance metrics and training associations have been extensible and explored in recent years as alternatives to deal with rigid learning schemes. The main advantage of these new schemes is

---

[42] Fabian Leon and Fabio Martínez Carrillo. "A multitask deep representation for Gleason score classification to support grade annotations". In: *Biomedical Physics & Engineering Express* (2022).

[43] Fabian León and Fabio Martínez. "Learning a Triplet Embedding Distance to Represent Gleason Patterns". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 3229–3232.

to learn optimized embedded representations, generalizing examples within and between classes. For instance, Hadsell et al. [44] propose a contrastive loss, that uses tuples of positive and negative samples, minimizing the euclidean distance between a pair of examples with the same class and maximizing distance between pairs with different classes. This difference is relaxed by an α factor. Formally, being x_i and x_j two samples with their respective classes y_i and y_j, the paired loss can be formulated as:

$$Loss\left(x_i, x_j\right) = \{y_i = y_j\} \left\|x_i - x_j\right\|_2^2 + \{y_i \neq y_j\} \max\left(0, \alpha - \left\|x_i - x_j\right\|_2\right)^2$$

In contrast to cross-entropy, this entropy loss approach learns variability between the same class (entropy) and between different classes (cross-entropy), allowing to optimize the architecture without imposing any constrain in the model representation. In this line, Sumit, et al. [45] propose a siamese network that reinforces learning with shared weights from two integrated networks. This approach was firstly used in face verification, a clear domain where is impossible to obtain a standard and well-balanced dataset. In this learning scheme, the complementing nets give high-level embedding vectors, that allow building a distance that search reduces the distance of sample person face and maximizes in other tuples configurations. The siamese representation results in the learning of a low-dimensional feature space that is able to recognize sub-set face instances with respect to the possible rest of the samples. The contrastive loss implemented in such work only uses batch pairwise relationships to optimize the embedded space, being necessary for several training epochs to learn all data relations. Such fact could limit the approaching of abundant false-positive samples, which are natural in this kind of domain.

---

[44] Raia Hadsell, Sumit Chopra, and Yann LeCun. "Dimensionality reduction by learning an invariant mapping". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2. IEEE. 2006, pp. 1735–1742.

[45] Sumit Chopra, Raia Hadsell, and Yann LeCun. "Learning a similarity metric discriminatively, with application to face verification". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. IEEE. 2005, pp. 539–546.

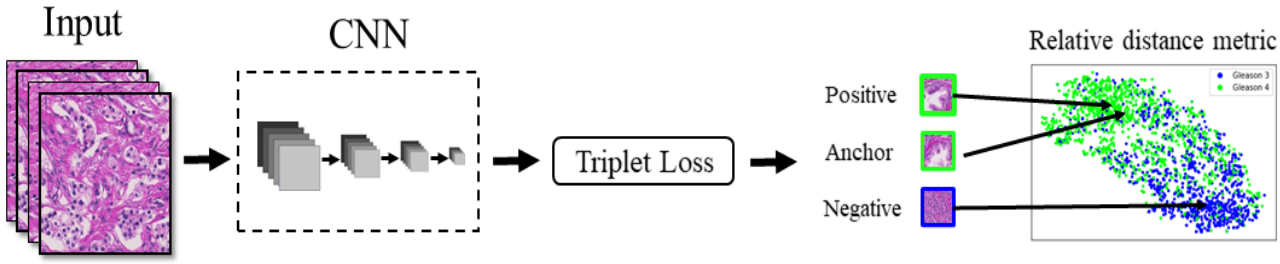

Figure 5. In the figure is illustrated an example of the triplet loss minimization rule that includes three examples tuples to learn semantic relationships between and within Gleason grades, minimizing the separation between similar pairs and maximizing space between dissimilar pairs.

To cope with such limitation, in [46] was proposed a triplet loss strategy, that includes three example tuples on batch training configuration. In this case, the distance learning to be trained to try directly with an unbalanced hypothesis, trying to put close similar class images and being distant from other classes with an α factor. Formally, the triplet is defined by an anchor representation sample fx_i, a positive sample representation fx_i^+, with same class, and a negative sample fx_i^-, *i.e.,* a different Gleason grade class w.r.t the anchor. Hence, a distance D is formulated to minimize separation between similar pairs and maximize space regarding the anchor and the negative class, defined as:

$$D(fx_i, fx_i^+) + \alpha < D(fx_i, fx_i^-)$$

Such relationship from deep convolutional representation f allows to learn semantic embedding spaces that trend to deal with a better Gleason grade separation. Properly the triplet loss can be defined as a hinge loss as follow:

$$TL(x_i, x_i^+, x_i^-) = max\{0, \alpha + D(fx_i, fx_i^+) - D(fx_i, fx_i^-)\}$$

---

[46] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.

The main advantage of this learning scheme, regarding typical cross-entropy, is the self-representation grade capability, which allows to directly model the strong intra-variability reported on the Gleason score. The triplet loss can approximate KL divergence by learning cross-entropy w.r.t to negative samples, but also entropy regarding the same Gleason sample. This fact allows to create of robust embedding that among others reduces sensibility to noise samples, avoids adversarial examples, and enhances boundary margins among classes, a fundamental issue in cancer degree characterization. An example of a triplet loss scheme is shown in figure 5.

4.2. An auxiliary task to deal with Gleason representation

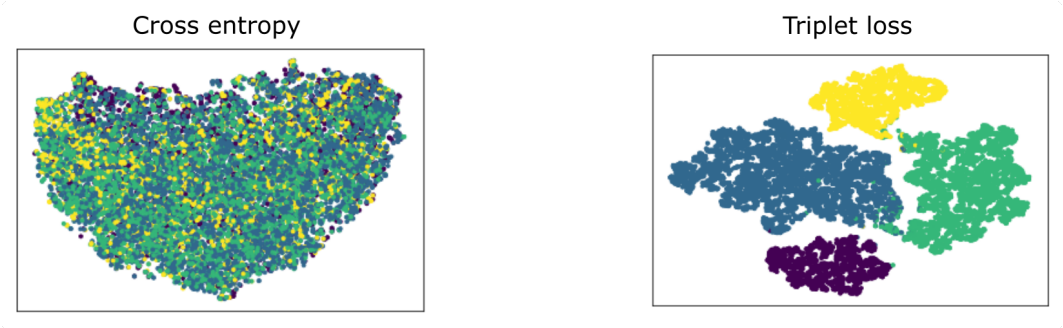

Figure 6. Embeddings representations from the last convolutional layer. Classical cross-entropy approach (left) shows high overlapping and diffused boundaries. Triplet loss (right) shows strong class separation, overfitting Gleason grades including human biases.

Despite the inter and intra class differentiation of distance metrics, the resultant deep embedding representations are in general limited to well-defined and high confidence data. This fact in a histopathological context is rarely appreciated, resulting in labeled data with weak confidence and biases by the experience of the expert. As shown in Figure 6 left, the Cross-entropy rule for multi-Gleason classification, is still showing high overlapping difficulting the topological discrimination of Gleason grades. The triplet embedding representation as shown in Figure 6 right, results in a model with well-defined boundaries with at least appreciate overlapping between classes. However, this triplet loss scheme may overfit because the intrinsic visual variability of each Gleason Grade converges in a collapse mode due to expert bias annotations.

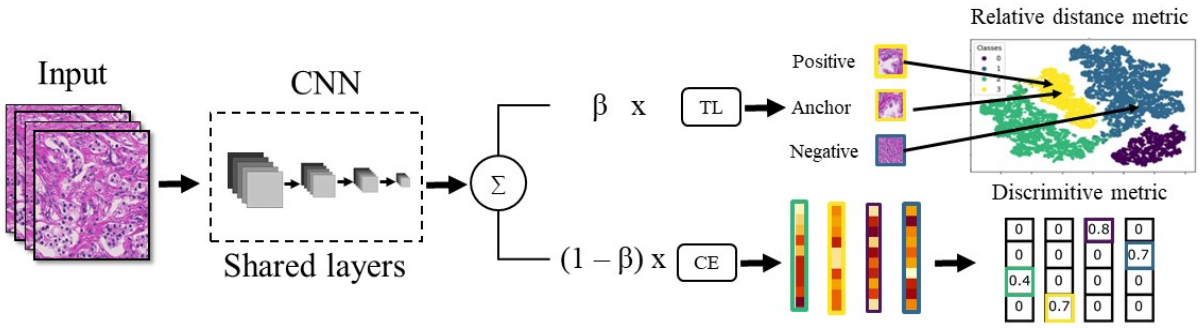

Figure 7. The Proposed multitask approach. In this proposed strategy is implemented a triplet loss as the main task to fully characterize Gleason similarities and a cross-entropy task that forces to maintain grade clusters into the representation and generalize representation. This complementary task has a regularization role and therefore a low impact during objective training.

Here, it was implemented a new strategy based on a multitask learning scheme. This strategy includes a triplet loss as the main task to fully characterize intra and inter-Gleason class similarities and a cross-entropy rule. From such scheme, it is possible to avoid expert bias annotations taking advantage of intra and inter-Gleason class similarities. This representation is complemented with an auxiliary task, implemented from a cross-entropy that forces to maintain grade clusters into the representation and generalize representation. This complementary task has a regularization role and therefore a low impact during objective training. [47]. An example of this scheme is shown in Figure 7. Formally, the CE is defined as:

$$CE(y_j, \hat{y}_j) = -\sum_{j=1}^{N} y_j log(\hat{y}_j) \tag{1}$$

Where y_j denote the experts' annotations and $\hat{y}_j$ is the model probabilities. Then, a linear combination of both losses is herein considered to complement triplet loss representation, as:

$$\zeta = CE(y_j, \hat{y}_j) + TL(x_i, x_i^+, x_i^-) \tag{2}$$

---

[47] Lukas Liebel and Marco Körner. "Auxiliary tasks in multi-task learning". In: *arXiv preprint arXiv:1805.06334* (2018).

The contribution of each learning scheme is weighted by β, to better exploit representations. This scheme achieves better performance, where the triplet loss learns semantic concepts between Gleason grades, and the CE applies a soft parameter restriction to the model. The proposed approach used the softmax branch to obtain the image probabilities, avoiding the use of an extra classifier.

4.2.1. Negative mining To select a suitable configuration to negative tuples is a crucial step to ensure the fast convergence of the neural network. Particularly, the negative mining can be defined as a triplet examples in three categories:

- **Easy triplets** Triplets where the loss is 0 it means that:

$$D(fx_i, fx_i^+) + \alpha < D(fx_i, fx_i^-)$$

- **Hard triplets** Triplets where the distance between the positive and the anchor is high than the distance between the negative and the anchor.

$$D(fx_i, fx_i^+) > D(fx_i, fx_i^-)$$

- **Semi-hard negatives** Triplets where the distance between the negative and the anchor is farther than the anchor and positive distance but is within the α factor, and still contribute positive loss.

$$D(fx_i, fx_i^+) < D(fx_i, fx_i^-) < D(fx_i, fx_i^+) + \alpha$$

In general, easy triplets have loss equal to 0 and don't contributes nothing to the train and should be discarded. Hard triplets leads to noisy gradients that cannot push effectively two samples being difficult to train in the early and middle epochs. Instead, Semi-hard negatives still contribute to positive loss, push examples efficiently especially for early and middle epochs,

reducing the variance in the gradients, using the *alpha* factor to select triplets [48].

4.3. Convolutional backbone coding

As a backbone for the triplet loss configuration was adopted a CNN representation that fully characterizes cancer histological patterns, stratified according to Gleason degrees. Specifically, in this scheme, it was adopted an InceptionV3 architecture [49]. This architecture factorizes symmetric and asymmetric block convolutions, reducing the number of connections/parameters without decreasing the network efficiency. The main issue to train such large net architectures is to have a sufficient amount of data. To overcome this limitation and to take advantage of previously learned representation, this scheme was initially used as transfer learning. Two different transfer learning approaches were here validated. Firstly, the selected backbone net was initialized from ImageNet and then adjusted with a triplet loss. A second transfer learning option was validated in two steps: First, the model was trained with a classical cross-entropy loss with a transfer learning from ImageNet. Then, the previous weights were used to adjust the embedding space from a coarse histopathological space using the triplet loss.
For the multitask scheme any deep convolutional backbone can be integrated to operate into this space. Particularly, in this scheme taking advantage of the additional data, was adopted the Xception architecture that includes convolutional layers but fully exploits inception layers. In brief, such layers approach a factorization principle to reduce complexity in nets and gain deeper representation. Also, the network built kernels that learn neural relationships through filters, allowing combine features to recover a robust sample representation [50]. This architecture

---

[48] Chao-Yuan Wu et al. "Sampling matters in deep embedding learning". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2840–2848.

[49] C *et al.* Szegedy. "Rethinking the inception architecture for computer vision". In: *CVPR*. 2016, pp. 2818–2826.

[50] François Chollet. "Xception: Deep learning with depthwise separable convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.

in the base has a total of 71 layers with a remarkable set of 23 million of training parameters. An additional condition of this net, which results interestingly to represent Gleason observations is the integration of skip connections, to obtain residual features that deal with vanishing gradient problems when there exists a limited set of training data and relatively large deep representations. With such advantages, the Xception result ideal to represent Gleason patch observations, capturing complex visual patterns, but also, being efficient face to relatively few and unbalanced training samples.

4.4. Experimental setup

4.4.1. Data The evaluation of the proposed scheme was carried out over a public dataset provided by HARVARD Data-verse [51]. This dataset contains a total of 886 H&E tissue images at 40x resolution (0.23 microns per pixel) divided into 5 TMAs. For validation of the schemes was followed the training, testing, and evaluation test suggested by the authors of the dataset. Particularly, the training has observations captured from the TMAs named as 83, 111, and 79. The validation samples were recovered from the TMA named 81 which shows a major balance among Gleason classes. For the test, the authors suggest the TMA 199 which has the major number of samples and also was annotated by an extra pathologist to quantify intra-observer agreement and the train and evaluation set was annotated by one uropathologist. To train the models a path extraction was performed following the same scheme as Eirini, et al [13]. allowing to code and to capture Gleason patterns from the images efficiently. In total 20024 patches were extracted from the 886 tissue images. 17785 patches were used for training divided into: 2076 patches for benign tissue, 7226 for grade three, 5207 for grade four, and 4541 for grade five. For test, from pathologist one, was extracted 4237 patches coding: 127 benign patches, 1602 for grade three 2121 for grade four, and 387 for grade five.

---

[51] Eirini Arvaniti et al. *Replication Data for: Automated Gleason grading of prostate cancer tissue microarrays via deep learning.* Version V1. 2018. DOI: 10.7910/DVN/OCYCMP.

Data augmentation. A patch extraction from

Data augmentation. process was herein carried out due to the lack of samples in some Gleason grades and also to avoid limitations like overfitting. The variability of samples was increased using horizontal-vertical flips and shifts, as well as patch image rotations. These variations correspond to geometrical transformation that naturally could be present during the computation of sample and they are admissible as a local variation of cell representation. Other samples generated include random zooms and image reflections.

4.4.2. Model parameters For the evaluation of triplet loss, the net was trained, following two steps: 1) only top layers are trained with 5 epochs, and 2) the model is fine-tuned using 20 epochs. An Adam optimizer was used with a learning rate of 0.001. Regarding triplet loss configuration, after a hyperparameter tuning, an empirical α factor of 10 was fixed, forcing an optimal and stable Gleason pattern differentiation without decreasing the network efficiency. Feature embedding vectors were fixed with a dimension of 512, following a semi-hard triplet negative mining. Regarding multitask approach, the backbone model also was initialized with ImageNet weights allowing to take advantage of common patterns of the initial layers using 5 epochs. Then a fine-tuning was carried out over the whole net using 15 epochs to fit specific histopathological patterns. For such fine tunning was implemented an ADAM optimizer with a learning rate of 0.0001. For the loss parameter β it was set as 0.7, with the triplet loss as the main contribution to the loss, and the cross-entropy as a regularization term. The α factor in this case for multi-class classification was fixed as 5 allowing proper separation between Gleason grades with optimal convergence. Also, was implemented semi-hard negative mining with an embedding vector of size 512 that act as descriptors of the input sample and recover all complex patterns learned from a set of samples.

4.4.3. Statistical analysis The proposed strategy was validated regarding the capability to classify Gleason grades. This work considered the following metrics:

- **The Accuracy** represents the global ratio of images well classified, measuring the observed agreement between observers and models. This metric describes how the model performs across all the classes. Is calculated as the ratio of correct Gleason predictions (True Positives (TP) and True Negative (TN)) to the total number of samples, which also include False Positives ((FP)) and False Negative (FN). Then, the accuracy is defined as: $TP + TN/TP + FP + TN + FN$. This metric nevertheless may be biased because of the imbalanced nature of the dataset.

- **The Precision** quantifies the number of positive class predictions that actually belong to the positive class. This metric is calculated as the ratio of positive Gleason samples correctly classified (TP) to the total number of Gleason samples classified as positive (TP and FP) as: $TP/TP + FP$

- **The Recall** measures the model's ability to predict all the positive Gleason regarding the FN ratio, and it is defined as : $TP/TP + FN$

- **F1-score** is an harmonic average that combine the precision and the recall as: $2(Precision \cdot recall)/Precision + recall$

- **The Kappa value** is a typical metric to establish the inter-rater agreement and uncertainty by observers. This statistic accounts for the possibility that models and experts agree on at least some observations due to chance, describe as: $p_0 - p_e/1 - p_e$, where p_0 is the relative observed agreement among raters and p_e is the hypothetical probability of chance agreement. The kappa value is defined from 0 to 1 where 0 corresponds to a poor agreement and 1 is a perfect agreement.

4.4.4. Baseline validation Baseline approaches based on distance metrics attempts to optimize an embedding space, reducing the distance between samples that belong to the same class and maximizing distance between dissimilar class. However, these methods don't implement a discriminative classification. From resultant embedding vectors is then necessary to

consider a discriminative machine learning algorithm to separate among classes. To properly validate the embedding space of baseline approaches such as triplet loss and siames loss, here it was implemented a classifier, as an extra step. In this case, was implemented one of the most common classifier K-nearest neighbors. This classifier allows to get image predictions based on the nearest labeled samples.

5. EVALUATION AND RESULTS

A first experiment was conducted to build embedding spaces, using different learning representation schemes. Taking into account that the main Gleason challenge discrimination between intermediate classes, the experiments were carried out first to differentiate between grades two and three. The set of resultant embedding vectors are projected into a two-dimensional space, following a t-SNE projection. As shown in Figure 8, the embedding space that results from the triplet loss approach achieves a significant separation of grades, which results promising to understand cancer severity patterns. In contrast, the embedding space from classical cross-entropy has a significant overlapping of classes.

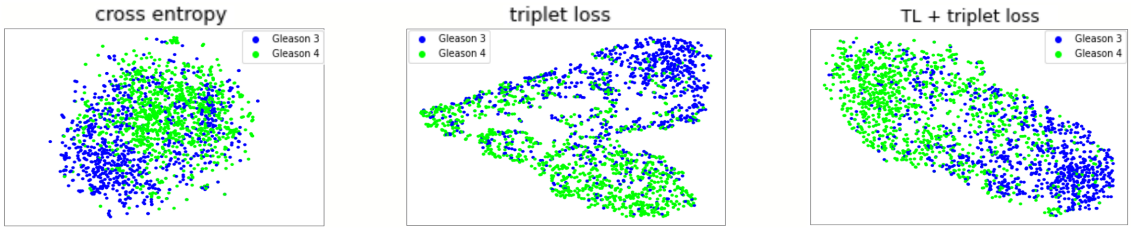

Figure 8. Embeddings from last inceptionV3 layers for the pathologist 1. Triplet embedding with transfer learning from classical schemes shows better Gleason differentiation, optimizing the space learning semantic concepts to contrast fine-grained Gleason patterns.

Table 2 summarizes the Gleason grade classification obtained for different learning schemes. In this bi-modal classification task, the two pathologists obtained an average accuracy of 71%, which evidenced the challenge of visual pattern quantification. Each scheme used the InceptionV3 and Xpception nets. In general, the schemes that implement triplet loss achieve better scores, being the recovered samples closer to the centroid of the embedding class. The best configuration was achieved by the InceptionV3 (73%), using the classical transfer learning with a further triplet loss and for Xception (70%) adjusting directly from triplet loss. The achieved results show the robustness of the proposed learning scheme, being more confident the classified samples, *i.e.,* the points in embedding space are closer to their respective centroid.

Scheme	Accuracy	
	InceptionV3	Xception
Cross-entropy loss	71% ± 1.89%	67% ± 1.76%
Triple loss	0.71% ± 2.14%	0.69% ± 1.47%
Transfer learning + Triple loss	**0.73% ± 1.65%**	**0.70% ± 2.35%**

Table 2. Evaluation results between inceptionV3 and Xception architectures for the different setup experiments.

Figure 9 shows a more detailed analysis, obtained from the confusion matrices between the two pathologist references. It should be noted that there exists a remarkable difference between pathologists to classify Gleason grade three, obtaining an average accuracy (agreement) of 47%. The proposed approach, in contrast, achieves balanced results between grades and among pathologists. This fact stands out the robustness of representation and clearly evidences the significant support that could provide in clinical scenarios and in challenging decisions. Despite of remarkable results of this strategy, the embedding space that includes more Gleason grades results in more challenging, showing remarkable restriction. In fact, the triplet-loss may collapse and overfit training representation, resulting in a less generalizable scheme.

Figure 9. Confusion matrix of triplet loss with transfer learning for Gleason score 3 and 4 . This approach show remarkable Gleason differentiation, especially for Gleason score 3, with more robustness and stable classification.

To overcome such limitations and extend the analysis to all Gleason classes, this work was robustness embedding representation by integrating a triplet loss analysis, with an auxiliary

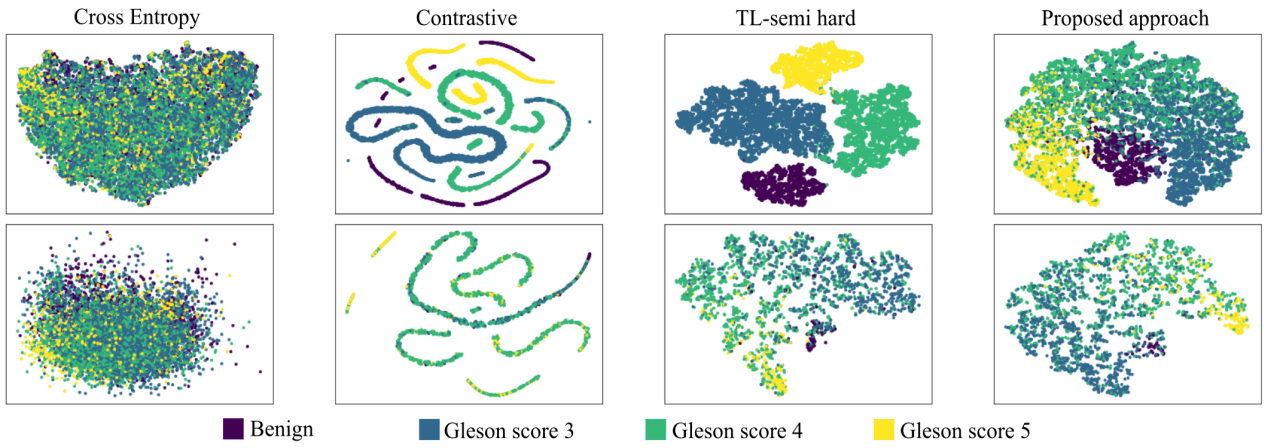

Figure 10. This plot illustrates the resultant embedding representation, achieved from two embedding vectors, obtained from the last convolutional layer of respective representations. To obtain a low dimensional representation it was implemented the t-SNE technique. The first row illustrates resultant representations, during training, for a set of samples. For train, the CE shows high overlapping performance, with diffused boundaries, while contrastive loss and TL-semi hard show bias and overfitting performance. Contrary, the proposed approach shows a more suitable Gleason representation, avoiding overfitting expert biases. The test projection results are more challenging in all representations but with different levels of overlapping among classes. Nonetheless, the proposed approach and the TL show better data distribution, where dominant centroids of the classes are preserved.

cross-entropy task. Hence, a multitask representation was proposed by integrating to the triplet loss representation, regularized by the cross-entropy. Such integration is weighted by a β parameter that defines the linear contribution to the final classification decision of Gleason score classes. The evaluation was carried out by comparing the proposed approach in the classification of all Gleason grades with other baseline approaches such as the cross-entropy (CE), the contrastive loss, and the triplet loss with semi-hard negative mining (TL). For such experiment, all compared training schemes were validated with a convolutional Xception net. These experiments were in turn compared with the scheme proposed by Eirini, *et al*[13]. The approaches were validated following the next classification metrics: F1 score, precision, recall, accuracy, and also the agreement kappa value with respect to the two expert pathologist.

Firstly, the topological distribution of each training scheme is projected in a low-dimensional space using a t-SNE method. As shown in Figure 10 the points learned from a cross-entropy report high overlapping, without any possibility to separate among Gleason patterns. Alterna-

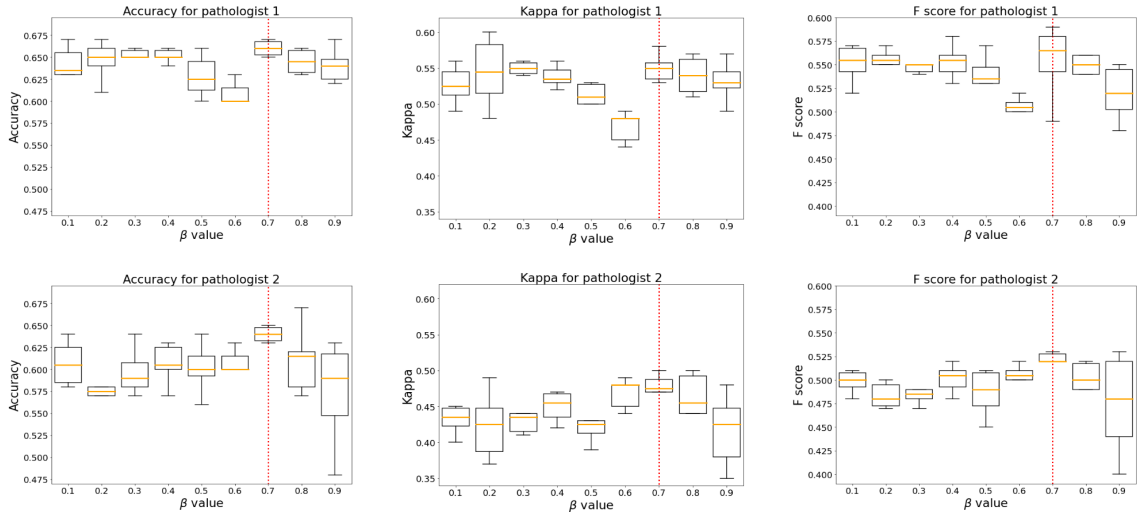

Figure 11. Validation of proposed approach, measuring the contribution of auxiliary task through different values of β parameter. Each row shows the performance of the proposed approach regarding each pathologist, independently. The results are presented in boxplots that summarize performance with respect to the accuracy, the kappa agreement, and the F-score. In average, the $\beta = 0.7$ has a better performance, which induces a major contribution to the triplet loss, but with an important role of the CE (30%) that avoids overfitting from expert annotations.

tively, the contrastive and the TL report an interesting training separation but biased topology with respect to the training pathologist. In consequence, for the test there exist a marked overlapping among degrees with a high distance among learned and evaluated distributions. The proposed approach achieves a better generalization of training, being resultant topology more adaptable for discrimination among Gleason classes. The bias reduction may be attributed to an auxiliary cross-entropy task that avoids overfitting of the representation.

In a second evaluation for multitask approach, it was performed an ablation study to analyze how the contribution of both tasks determines the generalization and score of the final deep representation. In such case, the β parameter is a linear operator (between $[0-1]$), that varies for different values, measuring at each experiment the capability of the proposed approach to correctly classify patches. Figure 11 summarize each of the experiments with respect to each pathologist (rows) and regarding the different metrics (columns). The results are presented as box plots for each β parameter. In general, there are no significant differences among mean scores

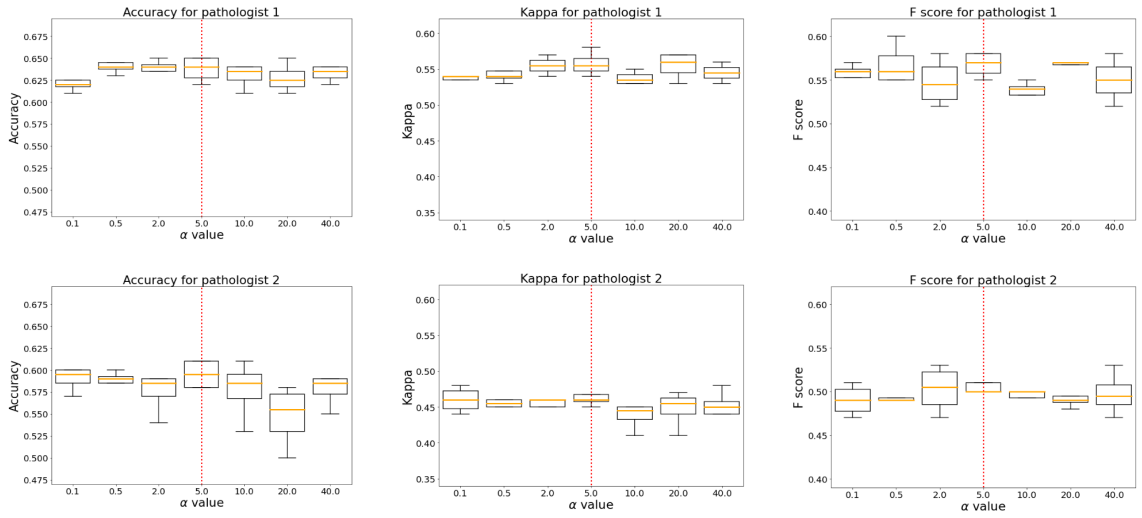

Figure 12. Fig 4 illustrates the validation of the proposed approach measuring the contribution of the α factor that margin positive and negative pairs of samples. Each row shows the performance of the proposed approach regarding each pathologist, independently. The results, summarized in Box plots show the performance with respect to the accuracy, the kappa agreement, and the F-score. The value of $\alpha = 5$ shows in average for both pathologists a much better performance, learning a proper separation between samples without increasing the model loss and the optimal convergence.

achieved for the different configurations. Nonetheless, the variability of prediction has important changes with respect to the contribution of each learning task. For instance, for $\beta = 0.1$, the cross-entropy auxiliary task has a major contribution to the multitask proposed model, showing more confident results (less variance). Contrary, for a major contribution of triplet loss representation ($\beta = 0.9$), there exists high variability among projected prediction samples. The best configuration is achieved with a $\beta = 0.7$, from which the triplet loss contributes to learning semantic concepts between the different Gleason grades. while the cross-entropy regularizes the model avoiding overfit annotation of the pathologist 1

In the same line, a similar study was implemented to better understand the α factor that measures how much distance exists between positive and negatives samples in the embedding space of the triplet loss. In this case, the α factor varies from different values from 0.1 to 40. As shown in Figure 12, it seems, that especially for extreme values, some of the metrics decrease the model performance. This fact can be attributed due to computational precision problems in the

Loss	Pathologist 1					Pathologist 2				
	Accuracy	**Kappa**	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Kappa**	**Precision**	**Recall**	**F1-score**
CE	0.57±0.012	0.49±0.013	0.52±0.004	0.52±0.01	0.5±0.011	0.49±0.033	0.37±0.032	0.46±0.012	0.51±0.017	0.44±0.015
Contrastive	0.60±0.01	0.5±0.012	0.58±0.011	0.53±0.1	0.53 ±0.01	0.56±0.01	0.43±0.01	0.53±0.11	0.52±0.013	0.49±0.014
TL	0.62±0.012	0.54±0.014	0.6±0.049	0.54±0.011	0.54±0.012	0.58±0.015	0.45±0.014	0.52±0.017	0.54±0.022	0.5±0.017
Eirini, et al	0.63	0.55	0.59	**0.58**	0.57	0.61	**0.49**	0.50	**0.53**	0.51
Proposed approach	**0.66±0.024**	**0.55±0.017**	**0.65±0.015**	0.53±0.015	**0.58±0.018**	**0.64±0.013**	0.47±0.028	**0.57±0.021**	0.51±0.028	**0.52±0.02**

Table 3. Classification metrics along the different training schemes using a softmax layer for Gleason classification.

multiple multiplications of very low values or very high values and the not optimal convergence of the model loss. In this case for this work, we select a middle value of $\alpha = 5.0$, allowing to the model learn a proper separation between the different Gleason grades without increasing and forcing the model loss.

Thirdly, the proposed approach was validated with respect to the cross-entropy branch that gives an automatic classification. For baseline approaches a softmax layer was adapted to validate an end-to-end training. Table 3 summarizes the performance achieved with respect to each pathologist. As observed, the proposed approach outperforms classical learning strategies, with a remarkable difference with respect to the typical CE rule (at least a difference of 9% in accuracy and precision). Also, there is reported a gain of 4% in F1-score for close contrastive and TL approaches. Interestingly, we conducted a t-student statistical test among the output results of the proposed approach for each pathologist. In such case, we obtain non-statistical differences ($p > 0.05$) between the results achieved by the proposed approach and the annotations carried out for each pathologist. Such fact may suggest a notable generalization of the proposed approach to represent Gleason patterns, avoiding expert bias introduced during training. In contrast, for the same t-student statistical test, the baseline strategies show significant differences ($p < 0.05$) between both pathologists, which suggests an inductive bias to the annotations of the first pathologist. The results reported by Eirini, et al [13] were also included in the comparison, which reports a relative best kappa agreement and recall, with respect to pathologists 2. This fact may be associated with the multiclass nature of the problem and the imbalance of the dataset. In such case, the metrics may reflect a better classification for benign tissue and Gleason grade five, which has the minor samples in the dataset.

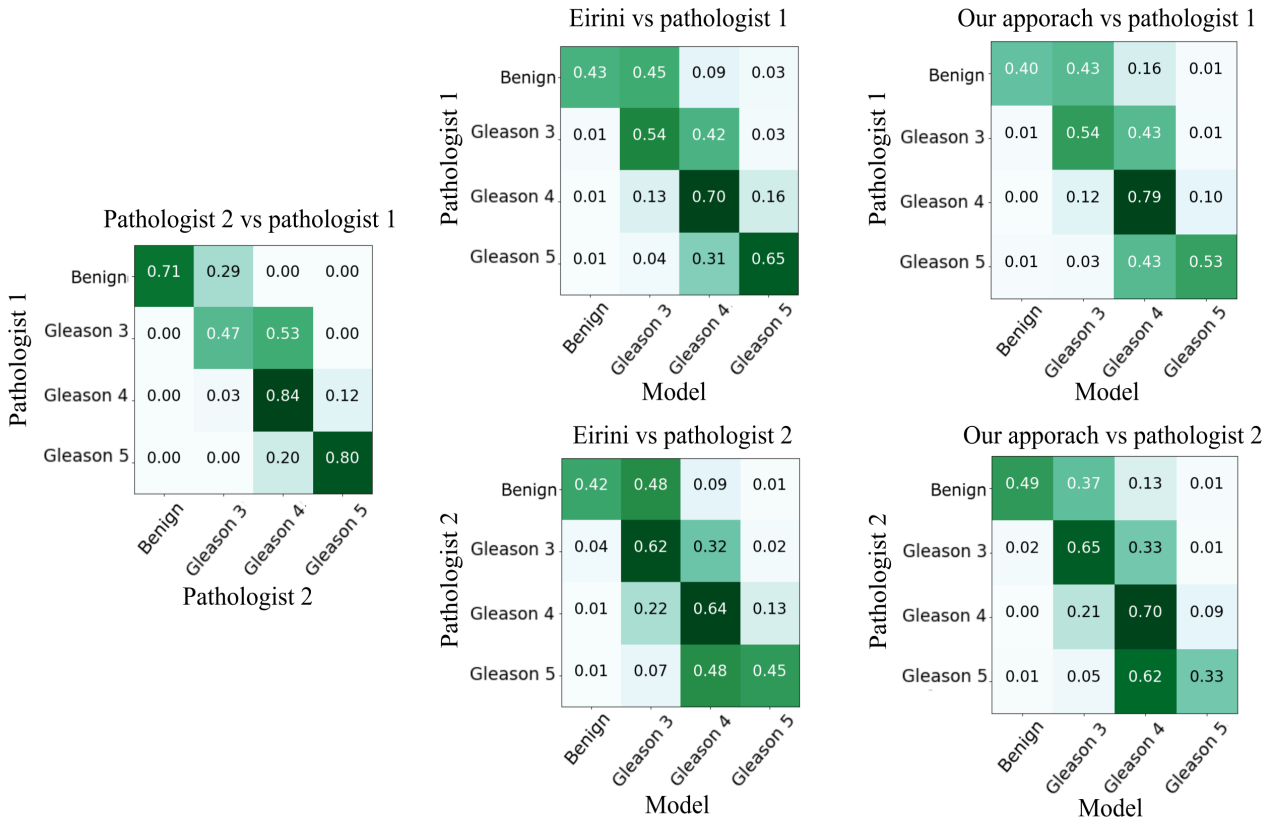

Figure 13. Confusion matrix for the two experts pathologist, Eirini, et al., and the proposed approach. As can be seen, the proposed approach achieves better Gleason differentiation, especially for patches where pathologists have a high degree of disagreement.

To better analyze such classification score at each evaluated grade, Figure 13 (top row) shows the confusion matrices, computed for the proposed approach and the work proposed by Eirini, et al, regarding pathologist 1. As expected, the baseline approach has a better performance in grade five but reports remarkable limitations to classify grade four. It should be also noted that both models mistake some patch classifications of grade four or five by the benign tissue. Particularly in such mistakes (about 1%), the methods fail to classify one patch sample. This fact may be associated with the isolated analysis of cropped patches, but doesn't represent a significant risk in final diagnosis support. In fact, we hypothesize that experts achieve more coherent annotations without miss-classifications between extreme grades because of the multi-scale analysis of microarray-spot samples. This analysis allows to observe the general pattern of complete

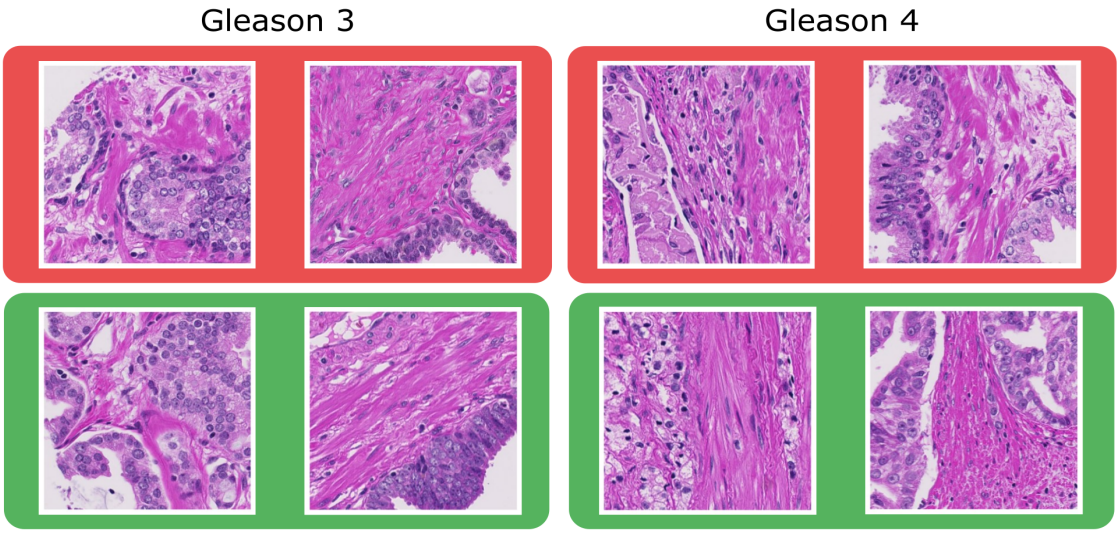

Figure 14. Some examples of patches labeled as Gleason three and four. At the top row are illustrated benign images misclassified as the degree of three or four. The bottom, rows are showed some correct samples, correctly classified. As observed, the miss-classification of benign tissue obeys to the strong similarity of visual patterns.

images but also zoom in into specific regions, observing particular structural configurations, but with global reference. Contrary, both models are designed to exploit local information coded from patches that lost global reference of the total microarray spot.

Particularly for histopathological analysis, the intermediate grades are more challenging, and computational tools may better contribute to the Gleason classification task. The same performance is observed with respect to pathologist 2, as illustrated in the bottom row of Figure 13. In the same line, the proposed approach achieves a better average performance in challenging intermediate grades but has several limitations with respect to benign tissue and the five grade. Figure 14 is illustrated the challenging problem of visual patterns among close Gleason grades. In fact, some patches labeled as benign tissue have a very close visual appearance with respect to the three and four degrees, which difficult for the generalization of the model from such training data. In such case, the proposed approach achieves a robust characterization of textural patterns, including much of the variability reported in observable patterns.

Also, it was measured the generated latent space with respect to the capability to discriminate

Loss	Pathologist 1					Pathologist 2				
	Accuracy	Kappa	Precision	Recall	F1-score	Accuracy	Kappa	Precision	Recall	F1-score
CE	0.16±0.012	0.01±0.005	0.2±0.015	0.26±0.012	0.12±0.013	0.18±0.014	0.01±0.007	0.19±0.013	0.19±0.011	0.12±0.015
Contrastive	0.6±0.021	0.49±0.013	0.61±0.021	0.51±0.008	0.5±0.17	0.55±0.018	0.4±0.018	0.53±0.012	0.49±0.022	0.49±0.19
TL	**0.63±0.006**	0.51±0.013	**0.62±0.018**	0.52±0.011	0.55±0.016	**0.56±0.008**	**0.42±0.024**	**0.58±0.011**	0.51±0.021	**0.51±0.017**
Proposed approach	0.62±0.014	**0.53±0.011**	0.61±0.008	**0.53±0.012**	**0.56±0.021**	0.54±0.014	0.4±0.023	0.54±0.016	**0.52±0.021**	0.48±0.017

Table 4. Classification metrics along the different training schemes using KNN classifier for Gleason classification.

among grades. This space is built from the projection of input patches in corresponding embedding vectors. For doing so, we measure distances among latent vectors and follow a KNN (K-nearest neighborhood) classifier. Table 4 summarizes the results for the different approaches with respect to the two pathologists. As can be observed, the representations that are focused on positive and negative tuples, during training, exhibit a better performance regarding the classical cross-entropy training. This fact is justified by the high reported intra-class variability of this problem. In fact, the cross-entropy scheme has a full overlapping of the representation. More specifically, the proposed approach achieves a gain in the representation with respect to the other approaches dedicated to model the topology of embedding space. For instance, the kappa in the proposed approach has a gain of 2% with respect to the TL, and contrastive loss, using as reference the pathologist 1.

Finally, we conducted an experiment taking into account only the subset of patches where both pathologists have an agreement. In such case, the data subset has a total of 2932 patches with a full agreement for both pathologists. With this experiment, we expect to measure the capability of nets to model visual patterns more than be biased by annotations of a particular pathologist. The whole baseline was then used to predict over this subset using the softmax classification, as well as, measuring closeness in topological embedding space. In the case of the Eirini, et al work, the architecture was run over this subset. Table 5 shows the achieved results for the whole baseline and for the proposed approach. In such case, the proposed approach reaches a remarkable difference in all the metrics with respect to the other schemes with a difference of at least 4% and 7% in accuracy and precision. With respect to Eirini, et al., the proposed approach has better performance with respect to all the metrics with a difference of 5% and 2% in terms of accuracy and kappa.

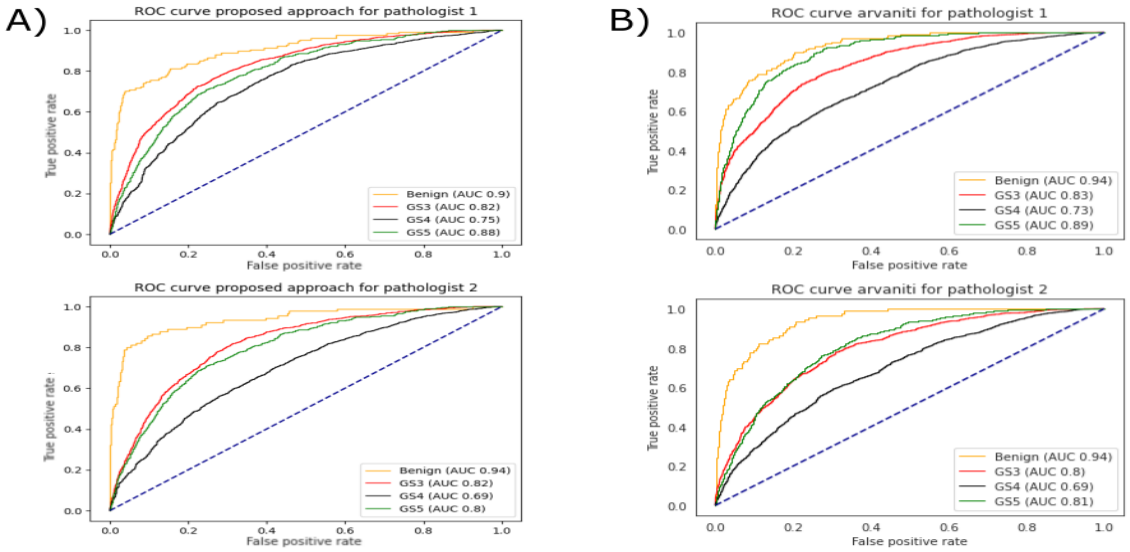

Figure 15. The ROC curves describe the performance of the proposed approach and the baseline, regarding the capability to produce output predictions for each Gleason grade. These curves allow to observe the invariance level of the results face several probability thresholds in output values. Column (a) shows the results achieved by the proposed approach. Column (b) shows the results of the baseline approach, proposed by Arvantini *et al.*

Loss	Softmax					KNN				
	Accuracy	Kappa	Precision	Recall	F1-score	Accuracy	Kappa	Precision	Recall	F1-score
CE	0.58±0.026	0.51±0.026	0.53±0.004	0.57±0.011	0.53±0.012	0.16±0.032	0.01±0.001	0.17±0.021	0.26±0.019	0.1±0.18
Contrastive	0.66±0.0	0.56±0.023	0.6±0.014	0.6±0.008	0.59±0.005	0.65±0.005	0.54±0.018	0.59±0.009	0.6±0.011	0.58±0.008
TL	0.69±0.022	0.61±0.019	0.60±0.018	0.6±0.029	0.59±0.02	**0.7±0.018**	0.6±0.016	**0.62±0.008**	0.6±0.015	**0.6±0.021**
Eirini, et al.	0.68	0.61	0.59	0.61	0.59					
Proposed approach	**0.73±0.008**	**0.63±0.12**	**0.67±0.017**	**0.61±0.017**	**0.62±0.011**	0.69±0.018	**0.61±0.012**	0.61±0.018	**0.61±0.019**	0.59±0.018

Table 5. Classification metrics along the different training schemes using a softmax and KNN classifier for patches were both pathologist are agree.

Complementary, we analyze the output prediction of each Gleason grade using different threshold values in the output prediction of the model, allowing to establish the robustness of results and the discrimination capability of the resultant representation. The results are summarized from ROC-AUC curves (the receiver operating characteristic- area under the curve), as illustrated in Figure 15. This analysis was also extended to the output predictions from the baseline approach, proposed for Eirini, *et al.* As observed in this figure, the best performance of both approaches is achieved for Bening tissue, regarding both pathologists. In such case, the performance can be attributed to the high confidence of the pathologist to diagnose benign tissue,

which result common in clinical routine. These results suggest that the proposed approach may be implemented in clinical scenarios to easily detect and separate control tissue on the samples, allowing pathologists to focus on more challenging tissue. In the same direction, the proposed approach achieved a remarkable detection for Gleason three, which represents one of the principal sources of errors and miss-classification on standard routine. Contrary, the approach proposed by Eirini, *et al* shows a better performance of Gleason five, which represents at advanced evolution of the cancer disease. As expected, for both models were recovered better results regarding the pathologist 1, fact associated that delineation patterns are also used in the training dataset.

6. DISCUSSION

This research work introduced a novel deep embedding representation that built a topological space and minimize intra-class-variance while maximizing distance among Gleason classes. The triplet loss embeddings characterize the most challenging visual histopathological samples (Gleason 3 and Gleason 4), allowing to support the local classification of patch samples. The approach reaches an average accuracy of 73%, with a gain of 2% compared with the annotations of two expert uropatholigst. However, this representation can be biased by human errors during training, showing a kind of overfitting in a multi-class Gleason classification and limited only to binary classifications. Then, the proposed approach is strengthened from a multitask approximation, integrating a triplet loss learning with a cross-entropy auxiliary task to avoid bias in certain sub-class clusters and expert annotations. The proposed approach outperforms the state-of-the-artwork, achieving an average F1-score of 0.58 and 0.52, with respect to two independent uropathologist annotations. In fact, in an exhaustive baseline comparison with alternative and classical CE learning, the proposed approach achieves a gain of at least 9% in terms of accuracy, which corresponds to a total of 381 more samples, predicted effectively. Besides, the achieved topological space achieves a proper generalization of histopathological visual patterns, without overfitting of training expert labels, showing a non-statistical differences ($p < 0.05$) of classifications with respect to both experts.

The main challenge of histopathological analysis is the quantification and association of visual patterns with cancer disease levels, even associated with the Gleason standard. In fact, in the literature has widely reported a moderate agreement among experts [52] K Mulay et al. "Gleason scoring of prostatic carcinoma: impact of a web-based tutorial on inter-and intra-observer variability". In: *Indian Journal of Pathology and Microbiology* 51.1 (2008), p. 22, which results

[52] Alireza Abdollahi et al. "Inter/intra-observer reproducibility of Gleason scoring in prostate adenocarcinoma in Iranian pathologists". In: *Urology journal* 9.2 (2012), pp. 486–490.

in low concordance among expert observations, noise labels, that include diverse and variable cellular structures, and high variability on visual labels with a low agreement that difficult the development of learning models that follow a semantic minimization via the discrimination among class probabilities. Recently, deep convolutional approaches have emerged as an alternative to support pathologists in the classification of Gleason patterns. These models include supervised and semi-supervised methods in a semi-automatic classification task using glandular structures [14] [53]. However, in the advanced stages of the disease is only appreciated solid and larger nests and comedo necrosis, being difficult to characterize such glandular structures [54]. Contrary to these approaches, the proposed methodology follows a learning of patch regions that result flexible to annotation protocols. Actually, the proposed strategy codes cellular observations into a cropped region, which thereafter is embedded in a latent vector to summarize complex visual relationships. These visual descriptors are herein optimized from positive and negative samples relative to a particular patch annotation, which allows to model visual variability and construct a topological space. From such geometrical construction, it is possible to better discriminate visual patterns.

Related to automatic classification, some approaches have also addressed the deep representation of tissue patches [13]. These approaches however use classical learning schemes that follow cross-entropy rules and minimize only positive samples with respective labels. In such sense, this learning has a poor generalization, trend to overfit representation with respect to the specific annotations, and lost generalization to cover the wide variability in histopathological observations. In such a way, the triplet loss learning deal with such limitations by incorporating into the learning the cross-entropy and also the entropy among samples of the same class. In our last

---

[53] Nathan Ing et al. "Semantic segmentation for prostate cancer grading by convolutional neural networks". In: *Medical Imaging 2018: Digital Pathology*. Vol. 10581. International Society for Optics and Photonics. 2018, 105811B.

[54] Jonathan I Epstein. "An update of the Gleason grading system". In: *The Journal of urology* 183.2 (2010), pp. 433–440.

proposal, this triplet loss learning is regularized from an auxiliary task that achieves a better net generalization. In fact, the proposed approach achieved a gain of 3% with respect to both pathologists included in the study and compared with the state-of-the-art baseline proposed by Eirini, et al [13]. These facts may suggest that our model learns more robust and reliable patches to be extended to other pathologists. However, despite the lower kappa and recall values for pathologist two, the proposed approach shows better Gleason patches differentiation, especially in patches where both pathologists have a high disagreement.

7. CONCLUSIONS AND FUTURE WORK

This work introduced a multitask learning approach to deal with histopathology patches classification, following the Gleason system. The proposed approach takes advantage from a triplet loss strategy to introduce a distance metric that allows to define a geometrical space that closes samples with the same degree while trying to maximize the distance of different classes. This distance metric may converge in overfitting representation and therefore the proposed approach considered to regularize the learning from a second classification task, which considers a typical cross-entropy. This regularization deals with high inter and intra appearance and structural variations of the Gleason system. This learning scheme was run over a backbone defined as an Xception architecture that combines residual blocs and a factorization principle, learning more complex and deeper patterns of the Gleason system. The model was evaluated from a public histological dataset that include annotations of two expert uropathologists. The proposed strategy outperforms baseline approaches over such data, and statistical test evidence non-statistical differences among pathologists and the computational strategy. The model showed potential capabilities to support annotations with overlapped degrees annotations, but also in educational scenarios to train pathologists with few experience. For example, benign tissue and Gleason score five are manageable grades to diagnosis where pathologists have high agree with an accuracy of 71% and 80% respectively. However, Gleason score 3 and 4 are challenging grades where the pathologist has significant disagreements, while the proposed approach reaches high concordance with an accuracy of 73% and 74% in a binary classification task and 54% and 79% in a multi-classification task. Despite remarkable advances in histopathological modeling, the Gleason grade characterization remains as an open problem due to the intrinsic visual similarity among grades, and the strong dependencies on expert observations. Future work includes the development of a study to measure intra and inter-expert variability, complemented with information about the formation and experience level of pathologists. From samples delineated from these pathologies, novel strategies may emerge to select and weight positive and negative sam-

ples with respect to beliefs of each expert. Also, new deep representations and training schemes will be proposed to avoid bias annotations and to offer new self-supervised representations.

BIBLIOGRAPHY

Abdollahi, Alireza et al. "Inter-observer reproducibility before and after web-based education in the Gleason grading of the prostate adenocarcinoma among the Iranian pathologists." In: *Acta Medica Iranica* (2014), pp. 370–374 (cit. on pp. 14, 22).

Abdollahi, Alireza et al. "Inter/intra-observer reproducibility of Gleason scoring in prostate adenocarcinoma in Iranian pathologists". In: *Urology journal* 9.2 (2012), pp. 486–490 (cit. on p. 53).

Allsbrook, William C. et al. "Interobserver reproducibility of Gleason grading of prostatic carcinoma: General pathologist". In: *Human Pathology* 32.1 (2001), pp. 81 –88. DOI: https://doi.org/10.1053/hupa.2001.21135 (cit. on pp. 14, 22).

Arora, Rebecca et al. "Heterogeneity of Gleason grade in multifocal adenocarcinoma of the prostate". In: *Cancer: Interdisciplinary International Journal of the American Cancer Society* 100.11 (2004), pp. 2362–2366 (cit. on p. 13).

Arvaniti, Eirini et al. "Automated Gleason grading of prostate cancer tissue microarrays via deep learning". In: *Scientific reports* 8.1 (2018), pp. 1–11 (cit. on pp. 15, 24, 38, 44, 47, 54, 55).

Arvaniti, Eirini et al. *Replication Data for: Automated Gleason grading of prostate cancer tissue microarrays via deep learning.* Version V1. 2018. DOI: 10.7910/DVN/OCYCMP (cit. on p. 38).

Barry, Michael J et al. "Screening for prostate cancer–the controversy that refuses to die". In: *New England Journal of Medicine* 360.13 (2009), p. 1351 (cit. on p. 18).

Bhattacharjee, Subrata et al. “Analysis of Biopsy Tissue Images based on Color Moment Technique and Morphology of Cell Nuclei”. In: () (cit. on p. 23).

Bley, Enrique and Andrés Silva. “Diagnóstico precoz del cáncer de próstata”. In: *Revista médica clínica Las Condes* 22.4 (2011), pp. 453–458 (cit. on p. 19).

Bray, Freddie et al. “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. In: *CA: a cancer journal for clinicians* 68.6 (2018), pp. 394–424 (cit. on p. 17).

Bulten, Wouter et al. “Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study”. In: *The Lancet Oncology* 21.2 (2020), pp. 233–241 (cit. on pp. 15, 26, 54).

Cai, Jianyong et al. “Automated Gleason grading of prostate cancers via deep learning in label-free multiphoton microscopic images”. In: *Multiphoton Microscopy in the Biomedical Sciences XX*. Vol. 11244. International Society for Optics and Photonics. 2020, 112442A (cit. on p. 25).

Cancer, International Agency for Research on. *The Global Cancer Observatory "GLOBOCAN"*. `https://gco.iarc.fr/today/home`. 2018 (cit. on pp. 13, 17).

Chollet, François. “Xception: Deep learning with depthwise separable convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258 (cit. on p. 37).

Chopra, Sumit, Raia Hadsell, and Yann LeCun. “Learning a similarity metric discriminatively, with application to face verification”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. IEEE. 2005, pp. 539–546 (cit. on p. 32).

Doglioni, Ginevra, Sweta Parik, and Sarah-Maria Fendt. “Interactions in the (pre) metastatic niche support metastasis formation”. In: *Frontiers in oncology* 9 (2019), p. 219 (cit. on p. 17).

Doyle, Scott et al. “Automated grading of prostate cancer using architectural and textural image features”. In: *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. 2007, pp. 1284–1287 (cit. on p. 15).

Epstein, Jonathan I. “An update of the Gleason grading system”. In: *The Journal of urology* 183.2 (2010), pp. 433–440 (cit. on p. 20).

— “An update of the Gleason grading system”. In: *The Journal of urology* 183.2 (2010), pp. 433–440 (cit. on p. 54).

— *Gleason score 2–4 adenocarcinoma of the prostate on needle biopsy: a diagnosis that should not be made*. 2000 (cit. on p. 21).

Epstein, Jonathan I et al. “A contemporary prostate cancer grading system: a validated alternative to the Gleason score”. In: *European urology* 69.3 (2016), pp. 428–435 (cit. on p. 20).

Epstein, Jonathan I et al. “The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma”. In: *The American journal of surgical pathology* 40.2 (2016), pp. 244–252 (cit. on p. 14).

Fischer, Andrew H et al. “Hematoxylin and eosin staining of tissue and cell sections”. In: *Cold spring harbor protocols* 2008.5 (2008), pdb–prot4986 (cit. on p. 13).

Fukagai, Takashi et al. “Discrepancies between Gleason scores of needle biopsy and radical prostatectomy specimens”. In: *Pathology international* 51.5 (2001), pp. 364–370 (cit. on p. 18).

Greenwald, Peter. "Clinical trials in cancer prevention: current results and perspectives for the future". In: *The Journal of nutrition* 134.12 (2004), 3507S–3512S (cit. on p. 18).

Grönberg, Henrik. "Prostate cancer epidemiology". In: *The Lancet* 361.9360 (2003), pp. 859–864 (cit. on p. 17).

Hadsell, Raia, Sumit Chopra, and Yann LeCun. "Dimensionality reduction by learning an invariant mapping". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2. IEEE. 2006, pp. 1735–1742 (cit. on p. 32).

Humphrey, Peter A. "Gleason grading and prognostic factors in carcinoma of the prostate". In: *Modern pathology* 17.3 (2004), pp. 292–306 (cit. on p. 20).

Ing, Nathan et al. "Semantic segmentation for prostate cancer grading by convolutional neural networks". In: *Medical Imaging 2018: Digital Pathology*. Vol. 10581. International Society for Optics and Photonics. 2018, 105811B (cit. on p. 54).

Khani, Ali Asghar et al. "Towards Automatic Prostate Gleason Grading Via Deep Convolutional Neural Networks". In: *2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*. IEEE. 2019, pp. 1–6 (cit. on p. 26).

Kryvenko, Oleksandr N and Jonathan I Epstein. "Prostate cancer grading: a decade after the 2005 modified Gleason grading system". In: *Archives of pathology & laboratory medicine* 140.10 (2016), pp. 1140–1152 (cit. on p. 20).

Kweldam, CF, GJ van Leenders, and T van der Kwast. "Grading of prostate cancer: a work in progress". In: *Histopathology* 74.1 (2019), pp. 146–160 (cit. on p. 20).

Leon, Fabian and Fabio Martínez Carrillo. "A multitask deep representation for Gleason score classification to support grade annotations". In: *Biomedical Physics & Engineering Express* (2022) (cit. on p. 31).

León, Fabian and Fabio Martínez. "Learning a Triplet Embedding Distance to Represent Gleason Patterns". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 3229–3232 (cit. on p. 31).

Liebel, Lukas and Marco Körner. "Auxiliary tasks in multi-task learning". In: *arXiv preprint arXiv:1805.06334* (2018) (cit. on p. 35).

McBride, Russell Bailey. *Obesity and aggressive prostate cancer bias and biomarkers*. Columbia University, 2012 (cit. on p. 18).

Melia, J et al. "A UK-based investigation of inter-and intra-observer reproducibility of Gleason grading of prostatic biopsies". In: *Histopathology* 48.6 (2006), pp. 644–654 (cit. on pp. 14, 22).

Mulay, K et al. "Gleason scoring of prostatic carcinoma: impact of a web-based tutorial on inter-and intra-observer variability". In: *Indian Journal of Pathology and Microbiology* 51.1 (2008), p. 22 (cit. on p. 53).

Mun, Yechan et al. "Yet another automated Gleason grading system (YAAGGS) by weakly supervised deep learning". In: *NPJ Digital Medicine* 4.1 (2021), pp. 1–9 (cit. on p. 26).

Nagpal, Kunal et al. "Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens". In: *JAMA oncology* 6.9 (2020), pp. 1372–1380 (cit. on p. 25).

Naik, Shivang et al. “Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology”. In: *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. 2008, pp. 284–287 (cit. on p. 15).

Naik, Shivang et al. “Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information”. In: *MIAAB workshop*. Citeseer. 2007, pp. 1–8 (cit. on p. 23).

Nguyen, Kien, Anil K Jain, and Ronald L Allen. “Automated gland segmentation and classification for gleason grading of prostate tissue images”. In: *2010 20th International Conference on Pattern Recognition*. IEEE. 2010, pp. 1497–1500 (cit. on p. 23).

Payá Bosch, Elena. “Diseño y desarrollo de un sistema automático de segmentación de glándulas histológicas para identificar el cáncer de próstata en una etapa inicial”. In: (2019) (cit. on p. 13).

Ruiz, Ana Isabel et al. “Actualización sobre cáncer de próstata”. In: *Correo Científico Médico* 21.3 (2017) (cit. on p. 13).

Scardino, Peter T, Robert Weaver, and A Hudson M’Liss. “Early detection of prostate cancer”. In: *Human pathology* 23.3 (1992), pp. 211–222 (cit. on p. 17).

Schroff, Florian, Dmitry Kalenichenko, and James Philbin. “Facenet: A unified embedding for face recognition and clustering”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823 (cit. on p. 33).

Silva-Rodríguez, Julio et al. “Self-learning for weakly supervised gleason grading of local patterns”. In: *IEEE journal of biomedical and health informatics* 25.8 (2021), pp. 3094–3104 (cit. on pp. 15, 26).

Sridhar, Gayathri et al. “Association between family history of cancers and risk of prostate cancer”. In: *Journal of Men’s Health* 7.1 (2010), pp. 45–54 (cit. on p. 18).

Steiner, David F et al. “Evaluation of the Use of Combined Artificial Intelligence and Pathologist Assessment to Review and Grade Prostate Biopsies”. In: *JAMA Network Open* 3.11 (2020), e2023267–e2023267 (cit. on p. 25).

Szegedy, C *et al.* “Rethinking the inception architecture for computer vision”. In: *CVPR*. 2016, pp. 2818–2826 (cit. on p. 37).

Toro, Oscar Jiménez del et al. “Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score”. In: *Medical Imaging 2017: Digital Pathology*. Vol. 10140. International Society for Optics and Photonics. 2017, 101400O (cit. on p. 25).

Wu, Chao-Yuan et al. “Sampling matters in deep embedding learning”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2840–2848 (cit. on p. 37).

Wu, Ina and Charles S Modlin. “Disparities in prostate cancer in African American men: what primary care physicians can do”. In: *Cleve Clin J Med* 79.5 (2012), pp. 313–20 (cit. on p. 18).

Zarbo, Richard J, Frederick A Meier, and Stephen S Raab. “Error detection in anatomic pathology”. In: *Archives of Pathology and Laboratory Medicine* 129.10 (2005), pp. 1237–1245 (cit. on p. 14).

APPENDICES

Anexo A. Academic Products

Journals

- Leon, F., and Carrillo, F. M. (2022). A multitask deep representation for Gleason score classification to support grade annotations. Biomedical Physics Engineering Express.
  Status: Published.

Conference papers

- León, F., and Martínez, F. (2021, November). Learning a Triplet Embedding Distance to Represent Gleason Patterns. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC) (pp. 3229-3232). IEEE.
  Status: Published.

Collaborations

- Plazas, M., Ramos-Pollán, R., León, F., and Martínez, F. Towards reduction of expert bias on Gleason score classification via a semi-supervised deep learning strategy. In Proc. of SPIE Vol (Vol. 12032, pp. 120322O-1).
  Status: Published.

- Gómez, A., León-Pérez, F., Plazas-Wadynski, M., and Martínez-Carrillo, F. (2021). Segmentación multinivel de patrones de Gleason usando representaciones convolucionales en imágenes histopatológicas. TecnoLógicas, 24(52), 176-196.
  Status: Published.