Multimodal Latent Semantic Alignment for Automated Prostate Tissue Classification and Retrieval

Juan S. Lara[1], Victor H. Contreras O.[1], Sebastián Otálora[2], Henning Müller[2], and Fabio A. González[1]

[1] Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, Bogotá, Colombia
[2] Institute of Information Systems, HES-SO (University of Applied Sciences and Arts Western Switzerland), Sierre, Switzerland

Abstract. This paper presents an information fusion method for the automatic classification and retrieval of prostate histopathology whole-slide images (WSIs). The approach employs a weakly-supervised machine learning model that combines a bag-of-features representation, kernel methods, and deep learning. The primary purpose of the method is to incorporate text information during the model training to enrich the representation of the images. It automatically learns an alignment of the visual and textual space since each modality has different statistical properties. This alignment enriches the visual representation with complementary semantic information extracted from the text modality. The method was evaluated in both classification and retrieval tasks over a dataset of 235 prostate WSIs with their pathology report from the TCGA-PRAD dataset. The results show that the multimodal-enhanced model outperforms unimodal models in both classification and retrieval. It outperforms state–of–the–art baselines by an improvement in WSI cancer detection of 4.74% achieving 77.01% in accuracy, and an improvement of 19.35% for the task of retrieving similar cases, obtaining 64.50% in mean average precision.

Keywords: Multimodal Fusion · Histopathology Images · Prostate Cancer

1 Introduction

Prostate cancer (PCa) is the fourth most common cancer worldwide with 1.2 million new cases in 2018 and it has the second-highest incidence of all cancers in men [17]. Currently, the Gleason score (GS) is the standard grading system used to determine the aggressiveness of PCa and determine treatment. Typical scores range from 6 to 10 and cases with higher values are more likely to grow and spread fast [12]. The gold standard for the diagnosis of PCa is the inspection of biopsies or tissue samples. Thanks to the recent improvements in digital microscopy, the diagnosis is increasingly made through the visual inspection of high-resolution scans of a tissue sample or a Whole-Slide Image (WSI) [2]. Digital pathology is focused on the management of this kind of data. Collections of WSIs and related information like pathology reports can be accessed and stored using Picture Archiving and Communication Systems (PACS). This preserves the data in the long term providing a valuable clinical information source. Computer-Assisted Diagnosis (CAD) is one of the most studied tasks in digital pathology. It generally covers tasks such as the automatic classification or grading of a disease, segmentation of regions of interest, mitosis and necrosis detection, image retrieval, among others [9].

Databases of medical images usually contain additional text data that is often not used by CAD systems [8]. There are diagnostic reports, clinical and related metadata that can be used to improve the performance of current CAD systems. Text usually contains semantic content that complements the information in the images. However, a current challenge is related to the appropriate combination of the image and the text information, especially, considering that these modalities originate from different sources and therefore have different statistical properties [1]. Multimodal fusion is an approach that aims to combine information from different modalities or information sources. Its application in the medical domain is an active research area [3,10] and it has not been fully explored for the analysis of multimodal prostate histopathology data. Related studies show that an appropriate combination of the image and text modalities provides better overall performance in comparison with single modal approaches. For instance, Jimenez-del-Toro et al. [8] proposed a multimodal retrieval method that combines deep learning with rank fusion, their results show that multimodal queries outperform both image and text in the retrieval of prostate tissue cases. In the same manner, Contreras et al. [6] proposed a method that combines the data in an early representation level besides the rank fusion, showing that there is joint information that can be exploited at different fusion levels.

One of the main disadvantages of these methods is that they are unfeasible in certain scenarios where a pathologist may only have an image, because, these approaches also require multimodal inputs during the prediction or retrieval of new cases. This motivates a fusion strategy that enhances the independent representations of each modality instead of enhancing a joint and combined representation. In this regard, some strategies have been proposed for the analysis of histopathology data: Caicedo et al. [2] proposed a non-negative matrix factorization method for the multimodal indexing of multiple organ tissues, it aims to induct a shared latent space for all modalities through an iterative optimization process that independently reconstructs a modality in each step. Cheerla et al. [5] proposed an unsupervised deep multimodal representation for pan-cancer prognosis prediction, it combines information from clinical data, genomics, microRNA, and WSIs, using deep representation learning and a loss function based on siamese networks. These approaches show the feasibility of a new kind of weakly-supervised multimodal fusion that can be used to enrich the representation of the histopathology images and can be explored in the analysis of PCa.

Besides the fusion strategy, a multimodal system requires an appropriate representation of the images. In this matter, deep learning has become the state of the art in many applications of computer vision, digital pathology, and the automatic analysis of prostate tissue images. Specifically, models like the Convolutional Neural Networks (CNNs) can learn high-level representations from the raw images without requiring handcrafted feature extraction and with minimal preprocessing. Since WSIs are large images, the CNNs are usually trained to identify patch-level patterns and the information is summarized for the global prediction through a majority vote, bag-of-features, or ensembles [7,11,14]. There is evidence that CNN-based CAD systems are comparable to international pathology experts in prostate cancer detection and grading [14]. Nonetheless, the integration of semantic information in deep learning models for the automatic analysis of prostate images remains a challenge.

This work addresses the problem of cancer detection and similar case retrieval using multimodal histopathology data from prostate WSIs and their diagnostic reports. To this end, we present Multimodal Latent Semantic Alignment (M-LSA), a model to simultaneously learn an embedded representation for the WSIs and their associated text content, it is trained using a weakly-supervised approach that uses text information to enhance the representation of the images. The method exploits the complementary information of visual and text modalities, which leads to better classification and retrieval performance as shown by the experimental evaluation. The remainder of this document is organized as follows: Section 2 presents M-LSA, showing details about the information fusion strategy and the representation techniques; Section 3 describes the experimental

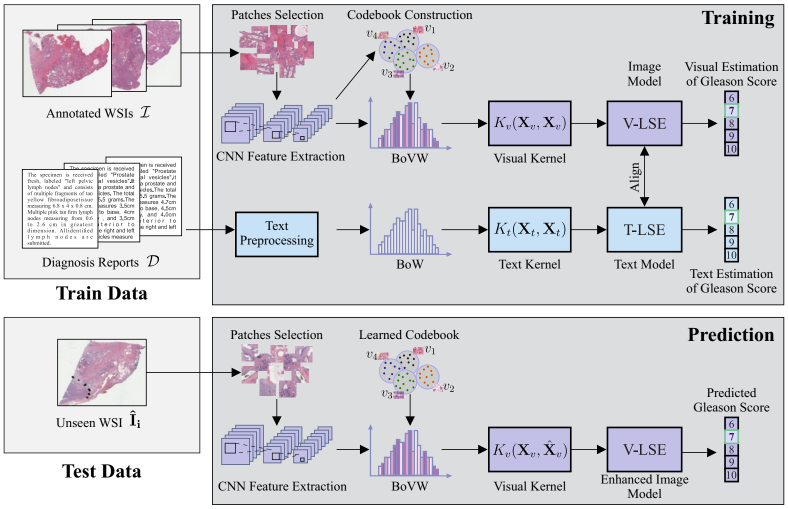

Fig. 1. Overview of the training and prediction phases of the proposed method for the automatic classification of prostate WSIs.

evaluation used to assess the effects of the weak semantic supervision in classification and retrieval tasks; Section 4 presents the experimental results and the analysis; Section 5 shows the conclusions.

2 Overview of Multimodal Latent Semantic Alignment

An overview of the method is shown in Figure 1. During the training phase, the method incorporates weakly-supervised information from the diagnostic reports to enhance an embedded representation of the WSIs. This enhanced representation is used during the prediction phase to obtain a GS estimation using only the information from the images. The WSIs are represented using a Bag-of-Visual Words (BoVW) approach and the text content is represented using a Bag-of-Words (BoW). These representations are embedded and aligned using an information fusion strategy that is described in the following subsections.

2.1 Data Representation

As shown in Fig. 1, the training data are composed of pairs of annotated WSIs and their diagnostic reports. We represent these multimodal data as a term frequency-inverse document frequency (TF-IDF) matrix for each modality. In this way, the representation of the text content is straightforward. The text preprocessing consists of stop-word removal during the text vocabulary T construction. We use the TF-IDF weighting schema because it benefits the information fusion strategy providing numerical stability while increasing the importance of unique terms and attenuating the common ones.

For the images, a codebook or visual vocabulary V is constructed to represent a WSI as a Bag-of-Visual-Words (BoVW). The BoVW contains a distribution $P(V = v_i|\mathcal{I} = I_j)$ of certain visual words v_i in an image I_j. To compute this, 2000 non-overlapping patches of size 256×256 are selected from each WSI using the blue ratio as filtering criteria (for obtaining most severe cancer areas) [4] as it is done in [7]. Then, a feature representation of the patches is computed using the GoogLeNet CNN architecture that was pre-trained for the binary classification of the GS, we use this network because it has demonstrated to outperform other architectures in the automatic diagnostic of prostate tissues [7]. Each patch is described with the feature vector that outputs the last average pooling layer of GoogLeNet, which is commonly used for feature extraction. The codebook is constructed using K-means over the CNN descriptors. More precisely, a visual word is a cluster in the CNN representation space and a visual document is constructed by assigning each patch descriptor to their closest centroid. As shown in Fig. 1, this procedure allows computing the BoVW by counting the number of patches in each cluster. Besides, TF-IDF is also used to weight the distribution of visual words.

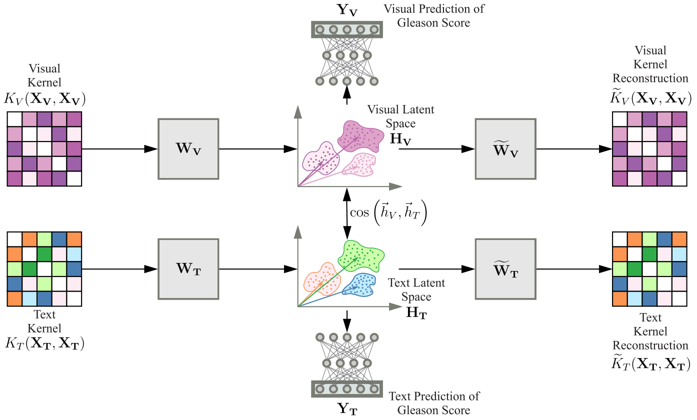
$\mathbf{Y_V}$
Visual Prediction of Gleason Score
Visual Kernel $K_V(\mathbf{X_V}, \mathbf{X_V})$
$\mathbf{W_V}$
Visual Latent Space $\mathbf{H_V}$
$\widetilde{\mathbf{W}}_\mathbf{V}$
Visual Kernel Reconstruction $\widetilde{K}_V(\mathbf{X_V}, \mathbf{X_V})$
$\cos\left(\vec{h}_V, \vec{h}_T\right)$
$\mathbf{W_T}$
Text Latent Space $\mathbf{H_T}$
$\widetilde{\mathbf{W}}_\mathbf{T}$
Text Kernel $K_T(\mathbf{X_T}, \mathbf{X_T})$
Text Kernel Reconstruction $\widetilde{K}_T(\mathbf{X_T}, \mathbf{X_T})$
$\mathbf{Y_T}$
Text Prediction of Gleason Score

Fig. 2. Conceptual diagram of the multimodal information fusion strategy.

2.2 Information Fusion

An overall description of the information fusion strategy is shown in Fig. 2. This strategy uses a reformulation of kernel matrix factorization that allows solving the problem through gradient-based optimization techniques as originally proposed in [15,16]. The main idea of the strategy is to take advantage of the reformulation and include certain constraints that incorporate supervised and weakly-supervised semantic information. This can be seen as an extension of the original embedding learning problem, in which the goal is not only to find a low-dimensional latent space but to learn an aligned latent space for each modality that also contains information about the GS.

The information fusion strategy requires to compute two kernel matrices: on the one hand, a matrix $K_V(\mathbf{X_V}, \hat{\mathbf{X}}_\mathbf{V})$ is calculated applying a visual kernel function K_V on the TF-IDF representations $X_v \in \mathbb{R}^{N \times |V|}$ of the training WSIs (N is the number of observations and $|V|$ is the visual codebook size) and a matrix $\hat{\mathbf{X}}_\mathbf{V}$ that can be the visual training or the test TF-IDF representations. On the other hand, a matrix $K_T(\mathbf{X_T}, \hat{\mathbf{X}}_\mathbf{T})$ is calculated using the equivalent text matrices and functions. The main purpose of the kernel functions is to capture the complex nature of each modality to obtain a simpler representation, i.e., the data is transformed into a feature space where linear relations are more likely to be found. In this case, the feature spaces of each modality, $X_i \in \{V, T\}$, are mapped into a latent space $\mathbf{H_i}$ of dimension L, through a linear transformation $K_i(\mathbf{X_i}, \hat{\mathbf{X}}_\mathbf{i})\mathbf{W_i}$ that uses a weight matrix $\mathbf{W_i}$ that must be learned.

The complete loss function is presented in Eq. 1, it combines the three following errors: (1) each space is linearly projected using a weight matrix $\widetilde{\mathbf{W}_\mathbf{i}}$ to obtain a reconstruction of each kernel $\widetilde{K_i}(\mathbf{X_i}, \hat{\mathbf{X}}_\mathbf{i})$. This allows us to estimate a reconstruction error $\overset{1}{\mathcal{J}}_i$, which is the mean squared error between the input and the output of each i modality and was derived using the kernel trick in [15]. This error is the basis of matrix factorization and is a non-supervised way to learn latent factors; (2) an artificial neural network (ANN) with a softmax activation output is used to obtain the predictions $\widetilde{\mathbf{Y_i}}$ of the GS. These predictions are used to calculate the *categorical cross-entropy* $\overset{2}{\mathcal{J}}_i$, which is used as an estimate of how different the predictions to one-hot encodings are of the GS ground truth $\mathbf{Y_i}$; (3) the cosine similarity $cos(\boldsymbol{h_1}, \boldsymbol{h_2})$ is computed between latent vectors of each modality $\boldsymbol{h_1} \in \mathbf{H_1}$ and $\boldsymbol{h_2} \in \mathbf{H_2}$. It measures the degree of alignment between the visual and text latent spaces, and allows us to calculate an alignment error $\overset{3}{\mathcal{J}}$. The alignment term promotes the learning of close latent spaces, this allows the mutual enrichment of the visual and textual latent representations.

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

3 Experimental Settings

TCGA-PRAD Dataset: The dataset is comprised of images and diagnostic reports from prostate cancer tissue with Gleason scores between 6 and 10. The data are available via The Cancer Genome Atlas (TCGA), which is a publicly available large collection of digital pathology and other data that contains a set of 500 cases of prostate adenocarcinoma (PRAD). We use a subset with 235 cases as suggested in our baseline [7,8]. The dataset was divided into the same baseline partitions for cross-validation: 141 cases for training, 48 for validation, and 46 for testing.

Cancer Detection Performance: The proposed method is evaluated on the automatic classification of low (GS 6 and 7) and high (GS 8, 9, and 10) grades, as this stratification changes the treatment decision. We aim to evaluate the effects of the semantic enhancement. For this reason, two versions of the proposed model are trained: (1) A visual latent semantic embedding *V-LSE*, which is a version of the proposed model that does not include the alignment, i.e., it is a model that is only trained using the WSIs. (2) *M-LSA*, which is a V-LSE model that is enhanced using the semantic information of the reports during training and is evaluated as shown in Fig. 1. In this case, we evaluate the performance in terms of classification accuracy, which is the metric used in similar studies [7,11].

Image Retrieval Performance: In this case, the models are trained to classify the five different categories of GS and the softmax outputs are used as an indexer. A single experiment consists of a simulated query, i.e., an example image is taken from the test set and the softmax outputs are calculated. Finally, these outputs are compared to the training set, and using the cosine similarity with all the test cases a ranking is constructed. Similar to our baseline studies [8,6], a case is relevant to the query if they share the same GS. The performance is evaluated in terms of Mean Average Precision (MAP), GM-MAP, and precision at top 10 (P@10) and 30 (P@30) retrieved results.

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

4 Results and Analysis

Table 1 presents the results for cancer detection. The proposed method is compared with similar studies that use comparable evaluation strategies on the same dataset. In the first baseline study [7] a GoogLeNet is used to represent the patches and to summarize the information through a majority vote, V-LSE achieves an equivalent performance. This behavior is reasonable considering that we are using the same CNN for the representation and a unimodal model should achieve similar performance. The second baseline study [13] presents a modified AlexNet architecture and summarizes using a majority vote. The authors specify that they included more training data. Thus, an important advantage of M-LSA is that it achieves a similar performance including text content instead of more training data. This means it can obtain better performance when limited training data are available. Also, weak supervision allows us to find a better visual latent representation through the automatic incorporation of text content. There is no need to assign additional local labels to model a visual vocabulary as it is usually done in similar approaches.

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

Method	Accuracy
GoogLeNet [7]	73.52
Modified AlexNet [13]	76.90
V-LSE	74.02
M-LSA	**77.01**

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

The retrieval results are shown in Table 4, presenting a comparison with the state-of-the-art retrieval methods that have been used to search PCa cases on the same dataset. It can be noticed that the semantic enhanced M-LSA model outperforms other image retrieval approaches. It is important to highlight that M-LSA only uses an image as a query, whereas other multimodal retrieval approaches require a multimodal query during the testing phase, which may not be suitable in realistic environments with new and uncertain cases where pathologists may not have a diagnosis report.

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

Method	MAP	GM-MAP	P@10	P@30
Image Retrieval [8]	0.5113	0.3921	0.4500	0.4600
Text Retrieval [8]	0.4092	0.3561	0.4913	0.3775
Multimodal Retrieval* [8]	0.5404	0.4196	0.5217	0.4884
KLSE* [6]	0.6263	**0.4843**	0.5667	**0.6326**
Visual TF-IDF	0.4390	0.3486	0.3717	0.3667
Text TF-IDF	0.3574	0.3143	0.3848	0.3377
V-LSE	0.5881	0.3966	0.5000	0.4949
M-LSA	**0.6450**	0.4187	**0.5752**	0.5500

The proposed methodology represents an important opportunity for clinical translation, a CAD system can include M-LSA to provide a second opinion or retrieve similar cases. Contrary to other multimodal approaches, it does not require the text information during the prediction phase, which is important during uncertain diagnostics where the findings may not be clear and the annotations may be erroneous or cognitively biased.

5 Concluding Remarks

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

Acknowledgments: This work was supported by COLCIENCIAS with the research project "Detección Temprana de daño ocular en Diabéticos usando un sistema de Inteligencia Artificial en Imágenes de Fondo de Ojo" number 1101-807-63563 and the Universidad Nacional de Colombia with the 2017-2018 national call for the project "Clasificación de retinopatía diabética y edema macular diabético en imágenes de fondo de ojo mediante redes neuronales convolucionales" number 202010029118. Sebastián Otálora has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825292 (ExaMode, `htttp://www.examode.eu/`) and from Minciencias through the call 756 for PhD studies.

References

1. Arevalo, J., et al.: Gated multimodal networks. Neural Computing and Applications **1** (2020)
2. Caicedo, J.C., Vanegas, J.A., Páez, F., González, F.A.: Histology image search using multimodal fusion. Journal of Biomedical Informatics **51**, 114–128 (2014)
3. Cao, Y., Steffey, S., Jianbiao, H., Xiao, D., Tao, C., Chen, P., Müller, H.: Medical Image Retrieval: A Multimodal Approach. Cancer Informatics **13**, 125–136 (2014)
4. Chang, H., Loss, L., Parvin, B.: Nuclear segmentation in H&E sections via multi-reference graph cut (MRGC). In: International Symposium on Biomedical Imaging (ISBI). pp. 1–4 (2012)
5. Cheerla, A., Gevaert, O.: Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics **35**(14), i446–i454 (2019)
6. Contreras, V., et al.: Supervised online matrix factorization for histopathological multimodal retrieval. In: International Symposium on Medical Information Processing and Analysis. vol. 10975, pp. 1–8 (2018)
7. Jiménez del Toro, O., et al.: Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score. In: SPIE Medical Imaging. pp. 1–9 (2017)

8. Jiménez del Toro, O., et al.: Deep Multimodal Case – Based Retrieval for Large Histopathology Datasets. In: MICCAI 2017 workshop on Patch-based image analysis. vol. 2, pp. 149–157 (2017)
9. Komura, D., Ishikawa, S.: Machine Learning Methods for Histopathological Image Analysis. Computational and Structural Biotechnology Journal **16**, 34–42 (2018)
10. Mourão, A., Martins, F., Magalhães, J.: Multimodal medical information retrieval with unsupervised rank fusion. Computerized Medical Imaging and Graphics **39**, 35–45 (2015)
11. Nagpal, K., et al.: Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. npj Digital Medicine **2**(1) (2019)
12. PCEC: Gleason Score, Prostate Cancer Grading & Prognostic Scoring (2020), `https://www.prostateconditions.org/about-prostate-conditions/prostate-cancer/newly-diagnosed/gleason-score`
13. Ren, J., Hacihaliloglu, I., Singer, E.A., Foran, D.J., Qi, X.: Unsupervised Domain Adaptation for Classification of Histopathology Whole-Slide Images. Frontiers in Bioengineering and Biotechnology **7** (2019)
14. Ström, P., et al.: Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. The Lancet. Oncology **2045**(19), 1–11 (2020)
15. Vanegas, J.A.: Large-scale Non-linear Multimodal Semantic Embedding. Doctoral thesis, Universidad Nacional de Colombia (2017)
16. Vanegas, J.A., Escalante, H.J., Gonzalez, F.A.: Semi-supervised Online Kernel Semantic Embedding for Multi-label Annotation. Lecture Notes in Computer Science pp. 693–701 (2018)
17. WCRF: Worldwide cancer data. Global cancer statistics for the most common cancers (2018), `https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data`