UNIVERSITAT POLITÈCNICA DE VALÈNCIA

_ TELECOM ESCUELA TÉCNICA VLC SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

DESIGN AND DEVELOPMENT OF A SYSTEM FOR GRADING CANCER PATTERNS IN PROSTATE HISTOPATHOLOGICAL IMAGES USING DEEP LEARNING ALGORITHMS THAT CONSIDER THE UNCERTAINTY AND INTERPATHOLOGISTS VARIABILITY DURING TRAINING

Author: Marta Martínez Mora

ADVISORS:

Valery Naranjo Ornedo

Julio Silva Rodríguez

Bachelor's thesis submitted in partial fulfillment of the requirements for the Bachelor's degree in Telecommunication Technologies and Services Engineering from the School of Telecommunications Engineering at the Universitat Politècnica de València

Academic year 2021-22

Valencia, March 2022

Abstract

Prostate cancer is the second most prevalent type of cancer worldwide. In 2018, 1.3 million patients were diagnosed and it is estimated that the number of new annual cases will increase by 40.2% in 2030. This pathology is diagnosed from the visual analysis of biopsies by pathologists and the later classification of tissue differentiation according to the Gleason scale. This scale goes from 3 to 5, being inversely proportional to the degree of differentiation. This diagnostic process is a time-consuming task, and suffers from high variability among pathologists. To reduce the workload and increase the level of objectivity, in recent years, diagnostic support systems have been proposed based on deep learning algorithms and, specifically, convolutional neural networks. Despite the promising results obtained by these systems, state-of-the-art techniques are still limited, among other factors, by the bias introduced into the annotation during training.

Therefore, the aim of this thesis is the development of deep learning models that are robust to the variability between pathologists during training, in order to improve their generalization capacity. Following this purpose, loss functions based on Cohen's quadratic statistic will be used, as well as label smoothing. Likewise, this thesis will investigate the possibilities of optimizing convolutional neural network architectures that are used in the state of the art, including attention blocks which include multi-level features into the decision. In order to validate the improvement in generalization capacity, these methods will be trained and tested in different databases, with reference labels given by different pathologists. Finally, the results obtained will be analyzed, these have shown that providing the neural network with information on inter-pathologist variability improves generalization to external databases.

Keywords: Deep Learning, Prostate Cancer, Computer Vision, Gleason Grading

Resumen

El cáncer de próstata es a nivel mundial el segundo tipo de cáncer con mayor prevalencia. En 2018 se diagnosticaron 1.3 millones de pacientes y se estima que el número de casos anuales nuevos aumente en un 40.2% en 2030. Esta patología es diagnosticada a partir del análisis visual de biopsias por medio del patólogo y la clasificación de la diferenciación del tejido según la escala Gleason. Esta escala va de 3 a 5, siendo inversamente proporcional al grado de diferenciación. Este proceso diagnóstico es una tarea que consume grandes cantidades de tiempo, y sufre de una elevada variabilidad entre patólogos. Para reducir la carga de trabajo y aumentar el nivel de objetividad, en los últimos años se han propuesto sistemas de ayuda al diagnóstico basados en algoritmos de deep learning y, en concreto, redes neuronales convoluciones. A pesar de los prometedores resultados obtenidos por estos sistemas, las técnicas punteras aún se ven limitadas, entre otros factores, por el sesgo introducido en la anotación durante el entrenamiento.

Por ello, el objetivo de este TFG es el desarrollo de modelos de deep learning robustos a la variabilidad entre patólogos durante el entrenamiento con tal de mejorar la capacidad de generalización. En esta línea, se pretende utilizar funciones de pérdida basadas en el estadístico cuadrático de Cohen, y en suavizado de etiquetas. Asimismo, se tratará de optimizar las arquitecturas de redes neuronales convolucionales utilizadas en el estado del arte, incluyendo bloques de atención los cuales introducen patrones de diversas escalas en la decisión. Con tal de validar la mejora en la capacidad de generalización, estos métodos serán entrenados y testeados en distintas bases de datos, con etiquetas de referencia dadas por distintos patólogos. Por último se analizarán los resultados obtenidos, los cuales han demostrado que proporcionar a la red neuronal información sobre la varaibilidad interpatólogo mejora la generalización a bases de datos externas.

Palabras clave: Deep Learning, Cáncer de Próstata, Computer Vision, Gradación de Gleason

Resum

El càncer de pròstata és a nivell mundial el segon tipus de càncer amb major prevalença. En 2018 es van diagnosticar 1.3 milions de pacients i s'estima que el nombre de casos anuals nous augmente en un 40.2% en 2030. Aquesta patologia és diagnosticada a partir de l'anàlisi visual de biòpsies per mitjà del patòleg i la classificació de la diferenciació del teixit segons l'escala Gleason. Aquesta escala va de 3 a 5, sent inversament proporcional al grau de diferenciació. Aquest procés diagnòstic és una tasca que consumix grans quantitats de temps, i patix d'una elevada variabilitat entre patòlegs. Per a reduir la càrrega de treball i augmentar el nivell d'objectivitat, en els últims anys s'han proposat sistemes d'ajuda al diagnòstic basats en algoritmes de deep learning i, en concret, xarxes neuronals convolucionals. A pesar dels prometedors resultats obtinguts per aquestos sistemes, les tècniques punteres encara es veuen limitades, entre altres factors, pel biaix introduït en l'anotació durant l'entrenament.

Per això, l'objectiu d'este TFG és el desenvolupament de models de deep learning robustos a la variabilitat entre patòlegs durant l'entrenament amb tal de millorar la capacitat de generalització. En esta línia, es pretén utilitzar funcions de pèrdua basades en l'estadístic quadràtic de Cohen, i en suavitzat d'etiquetes. Així mateix, es tractarà d'optimitzar les arquitectures de xarxes neuronals convolucionals utilitzades en l'estat de l'art, incloent blocs d'atenció els quals incorporen patrons a diverses escales en la decisió. Amb tal de validar la millora en la capacitat de generalització, estos mètodes seran entrenats i provats en distintes bases de dades, amb etiquetes de referència donades per distints patòlegs. Finalment s'analitzaran els resultats obtinguts, els quals han demostrat que proporcionar a la xarxa neuronal informació sobre la varaibilidad interpatòleg millora la generalització a bases de dades externes.

Paraules clau: Deep Learning, Càncer de Pròstata, Computer Vision, Gradació de Gleason

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The first algorithm created to be processed by a machine was developed in 1842 by the mathematician, considered the mother of modern computing, Ada Lovelace. Quoting the British pioneer, ”a machine could act on other things besides numbers, it could compose elaborate musical and scientific pieces of any degree of complexity or extension”. This incredibly visionary conclusion would become a reality decades later thanks to Artificial Intelligence (AI).

The term Artificial Intelligence was first adopted by computer scientist John McCarthy, who defined it as the science of making a machine behave in ways that would be called intelligent if a human were to do so. The revolution in this field came in the 2010s, with the help of neural networks and deep learning algorithms. Advances in these areas have boosted another field already known within artificial intelligence, what we know as Computer Vision. We can describe Computer Vision as the scientific discipline that is responsible for processing and analyzing the content of digital images, emulating human cognitive processes.

Nowadays, we are surrounded by vast amounts of digital content that are subject to being analyzed by machine learning and computer vision algorithms. For this reason, this is one of the fields with loads of applications and a wide variety of use cases. Among them, the automotive industry stands out, with innovative systems such as the Tesla AutoPilot, as well as the manufacturing industry, through the automation of quality control. However we will focus particularly on the one that concerns us in this work, healthcare solutions.

Computer vision is one of the main tools of digital pathology, which facilitates the solution of some of the challenges of current medicine, that involve extremely complex processes that require a large amount of time and resources. Hence, the main objective of digital pathology focuses on streamlining diagnostic processes and facilitating clinical decision-making through the use of technologies such as biopsy analysis via deep learning algorithms. This technique starts with samples that are digitized and shared between remote laboratories, allowing collaboration between pathologists. It lays the foundations for the use of computer-aided diagnosis (CAD) techniques, which use these algorithms to help pathologists interpret multimedia samples such as medical images. Thus, this will be the tool used in this project in order to try to solve the problem that will be described below.

1.1 Motivation

As it will be explained in depth later, the diagnosis of prostate cancer is carried out by analyzing prostate biopsies using the methodology known as Gleason grading. As mentioned above, automatic image analysis algorithms are recently being developed to assist in this task.

Deep learning techniques in computer vision have become very popular in medical diagnosis support systems. However, the use of techniques that provide a priori information such as variability between pathologists, and use of visual patterns at different scales, remains little explored in many applications. In this thesis, we will focus on diagnostic aid systems for prostate cancer grading in histological images.

Interobserver variability can be described as the degree of difference in diagnostic interpretations when a group of patients is treated by two or more physicians[1]. Meanwhile, the diversity in image resolutions and patterns represented tend to vary task-relevant information and impedes CNN training and performance [2]. This information has not been taken into account so far in models that evaluate prostate cancer tissues, and by doing so preventing from improvement their generalization potential.

Thus, the motivation of this thesis lies in the development of a CNN model that introduces and combines these innovative aspects in its architecture and development, in order to be able to analyze patches of prostate samples not only in the training database, but also being robust in external databases.

1.2 State of the art

In order to understand the objective of this thesis, we will make a brief summary of the current state of the art techniques regarding the diagnosis of prostate cancer using deep learning algorithms. For the realization of this paper, several scientific articles have been reviewed that have helped us contextualize the problem, and determine possible development paths for the project.

1.2.1 Deep learning in prostate histopathology

One of the main points of reference in this paper has been *Automated Gleason grading of prostate cancer in tissue microarrays* [3]. In this work, Arvaniti et al. proposes a deep learning system to grade prostate cancer tissue microarrays with Hematoxylin and Eosin staining, the same context in which we begin our experiments. In this article, attention is focused on the annotation of the samples by different pathologists, to then be compared with the results of the trained model, which generates promising results. Thus, in this paper it is clear that there is indeed a disagreement between pathologists, since it is a very complicated task and sometimes a bit subjective. This will lay the groundwork for prostate cancer grading research using computer vision.

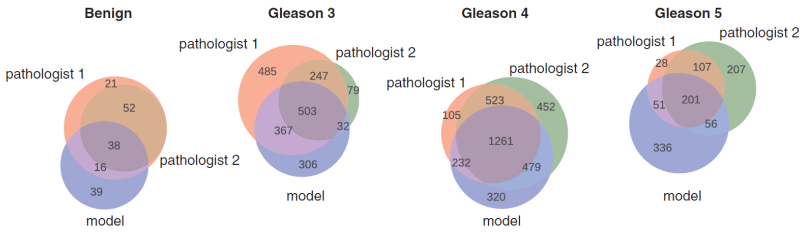


Figure 1.1: Venn diagrams illustrating the overlap in patch-level Gleason annotations produced by the Arvaniti deep learning model and the two pathologists. [3]

The next paper that has served as a guide for our work is *Going deeper through the Gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection*[4], a work carried out by the Institute of Research and Innovation in Bioengineering. In this article, a model designed by the authors themselves is trained from scratch, and focuses its attention on the detection of cribriform patterns in Gleason grade 4 patches, through the retraining of a series of filters in the last layers of the network. The SICAPv2 database is also presented, which will be described later as it is used in the present work. Thus, this paper will also be used as a referent, since in addition to working with the same database, it presents results that even outperform the results of state-of-the-art techniques.

These papers introduce innovative features within state-of-the-art methods for the detection of cancerous patterns, however they do not contemplate the introduction of the variability of agreement between pathologists, eventhough Arvaniti presents it in their paper [3], nor the introduction of features of different scales.

1.2.2 Introducing uncertainty in prostate cancer diagnosis

As explained above, the uncertainty and noise caused by the difference of opinions between pathologists has not yet been studied for the introduction into a deep learning model that analyzes prostate biopsies. However, some articles have been revised that present methods which seek to improve the results of their models by addressing the issue of uncertainty. However, these advances have been seen in more conventional images, such as the interpretation of chest radiographs or eye fundus images. In fact, the paper that has been used as a reference in this thesis, works with the last metioned sort of medical images.

In this thesis, the method proposed in *Cost-Sensitive Regularization for Diabetic Retinopathy Grading from Eye Fundus Images* [5] has been followed as a theoretical basis for our cost sensitive loss function, which introduces the term that takes into account inter-observer disagreement as a regularizing term.

1.3 Structure of the thesis

This section aims to briefly outline the general structure of the thesis:

- In chapter 2, the general and specific objectives of the work will be established, delving into the novelties introduced by this work with respect to the state of the art described above.
- During chapter 3, the entire related theoretical framework will be explained for a better understanding of the problem proposed, and therefore of the solutions given. This section has been divided into two, the necessary biological background, since due to the nature of the degree studied, this field has had to be investigated, as well as the basic knowledge of neural networks.
- Throughout chapter 4, the material necessary to carry out this work has been described, from the databases used, to the hardware and software used.
- Chapter 5 digs into the methodology followed for the execution of the various experiments carried out. These experiments pursue the investigation of the topics introduced in 1.1, and the methodology that explains why these experiments have been carried out is explained during this section.
- During chapter 6, the metrics used for the evaluation of the results are defined, as well as the experiments carried out for each research sub-topic and the results obtained through them.
- Finally, in chapter 7, the work will be finished describing the conclusions drawn after the analysis of the results and metrics obtained.

Chapter 2

Objectives

The ultimate goal of this thesis is the development of a prostate cancer diagnostic system that uses deep learning algorithms that take into account the uncertainty and variability of opinions among pathologists, as well as the integration of patterns at different scales. All this is done to improve diagnostic generalization, through a better understanding of the components that help a deep learning algorithm to determine the accurate classification of samples.

However, to achieve this global objective, sub-objectives have been proposed and achieved in order to guide us through the process of carrying out the project.

- Familiarization with the basic neural networks used for the diagnosis of prostate cancer.
- Understanding of the process that explains prostate cancer proliferation and the method used to grade it, Gleason grading.
- Description of the database used, since understanding how it is formed and how many samples there are for each case led to changes in the basic training model.
- Training of a foundational model based on a VGG19 architecture, to later add incremental improvements.
- Study and implement methods to mitigate the imbalance of classes present in the dataset, also investigating how to extract the model at the best time, as well as perform an optimization of the apparent hyperparameters to be used.
- Development of custom loss functions, converting the algebraic description of the methods in the papers studied into code that can be implemented in our network.
- Implementation of non-traditional layers included in the modified architecture, such as the attention layer with the gated mechanism, as well as an extra activation layer, in order to integrate multi-level features.
- Carrying out different ablation experiments to finish setting not only the hyperparameters, but also the variable parameters present in the improvements introduced.
- Evaluation of the results obtained after performing different experiments, through the analysis of figures of merits and the graphical representation, illustrated by the confusion matrices.

- Identify the problems and limitations observed, propose possible improvements and future lines of research in which the results of this project could be improved.

Chapter 3

Theoretical framework

3.1 Biological background

3.1.1 Prostate cancer

Cancer is nowadays the second cause of death worldwide, provoking one in every six deaths and specially affecting countries with intermediate and low incomes [6]. The World Health Organization (WHO) describes cancer as a general term used for designating a wide range of illnesses that can affect any part of the human anatomy. It can be explained as the rapid and uncontrollable reproduction of cells, that grow outside their normal cycle and start multiplying at a higher rate possibly forming tumors, which are an excess of tissue caused by these abnormal created cells that will not die or be repaired. Tumors can be benign or cancerous, these last ones also called malignant, can extend into contiguous tissues of the body as well as into other organs, developing what is known as metastasis, which is the cause of most of cancer deaths.[7]

Cancer classifies as one of the main challenges for current medicine since 19.3 million cases have been registered in 2020, and the number is expected to grow to 27.5 million new cancer cases by 2040. Within the bast variety of cancer, more than 200 types, the four most common in terms of new cases were: breast (2.26 million cases), lung (2.21 million), colon and rectum (1.93 million) and prostate (1.41 million)[8].

This project will be focused on prostate cancer, which remains the most frequent diagnosed cancer in Spanish men in 2021 with 35.764 new cases, and expected to grow as the time passes [9].

The prostate is a gland that belongs to the male reproductive system, located between the bladder and the penis and surrounding a section of the urethra, its main function is to secrete a fluid that protects sperm and creates semen when mixed with it in the urethra [10]. There is no clear and identifiable cause of prostate cancer, however there are several factors that are known to increase the risk of suffering from this type of cancer. These include ethnic group (men of African and Caribbean descent are more prone to have it), family history, the type of diet followed by the patient or age [11].

Prostate cancer is a condition that represents a risk for all men, however the older a man is, the greater the probability of being diagnosed with it. The 90% of new cases correspond to men older

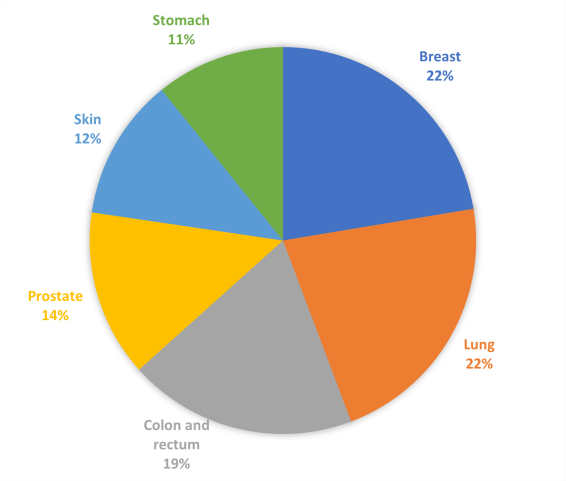


Figure 3.1: Most common cancers in 2020

than 65 years old, and the average age of diagnose is around 75. Generally it is a slow developing cancer, that causes no apparent symptoms unless its stage is advanced, that is why early detecting and treating are without a doubt a vital aspect of this illness [12].

With the purpose of granting patients quality of life by ensuring an early diagnosis there are different types of testing that can be performed following a routine and beginning about age 45.[13]

- **Prostate Specific Antigen (PSA) Test:** a blood test that measures the level of the PSA protein, produced by the prostate. If it is higher than normal it can be one of the first signs of prostate cancer, however it can also indicate the existence of other conditions, that is why alternative tests are recommended.
- **Digital Rectal Exam (DRE):** since the prostate is located immediately in front of the rectum it can be evaluated through inserting a gloved lubricated finger inside of it by a doctor, who will determine if it presents any irregularities in size, shape or texture.
- **PCA-3 test:** this analysis examines the concentration of prostate cancer gene 3 (PCA3) using a urine sample. It is usually performed after a DRE, and confirms the further need of doing a prostate biopsy, since it is a specific biomarker for prostate cancer, unlike PSA.
- **Prostate biopsy guided by transrectal ultrasound (TRUS):** this test is performed if there are previous reasons to suspect there could be a tumor, such as the PSA test mentioned before. It is carried out by inserting a small ultrasound probe through the rectum, which will radiate sound waves in order to make up an echo pattern that will form an image. If tumorous tissues are detected by the specialist, a needle will take small samples of it in order to be studied.

3.1.2 Histopathology Images

In most of prostate cancer cases, the only way to form an accurate diagnosis in order to properly plan a treatment if necessary, is to perform a prostate biopsy. By doing this simple procedure as explained before, pathologists are able to evaluate the severity of the cancer by grading it and

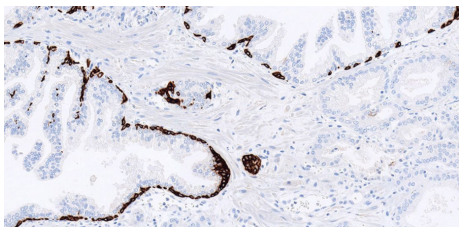
propose a fitting medical therapy. Although it generally is the decision making test, it is a bit of an invasive procedure, so it has some side effects like bleeding, pain and possible infection. This is the main reason why biopsies are the final test and are performed after the other procedures described have been tried [13].

Once the samples have been recollected, it is time for expert pathologists to interpret them. In order to understand how do these specialists do it, it is necessary to explain the aim of histology and the processes from when the samples are taken until a diagnosis is made.

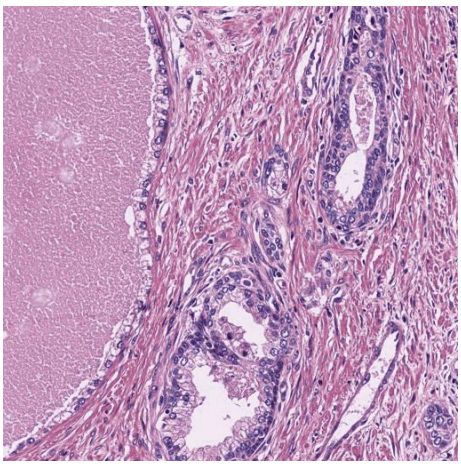
Histology, also known as microscopic anatomy, is the branch of biology that studies the structure, composition and characteristics of animal and plant tissues. Following that logic, the aim of histopatholgy is to examine samples of tissue taken from the living body, to study whether there has been any changes related to a disease or disorder [14]. For the samples to be studied under a microscope, the tissues taken have to undergo through several techniques.

The first step towards preparing a sample is tissue fixation. With the purpose that it preserves its cell and molecule structures for further handling, samples are submerged in a formalin solution immediately after they are taken. After its fixation, paraffin is used as an embedding compound so the tissue has the right consistency to allow thin slices to be cut. Afterwards, a microtome is used to obtain almost transparent slices in such manner that the light can pass through them. These thin sections are colorless, therefore the samples is not yet ready for examination under a microscope until the last step: they have to be stained with a dye in order to increase the contrast between the different cellular components[15].

Despite the fact that there are various staining methods, the most common staining protocol used for prostate tissue samples is hematoxylin and eosin dying. Ribosomes, chromatin (genetic material) within the nucleus, and other components are stained with hematoxylin, which gives them a dark blue-purple tint. Eosin generates an orange-pink-red tinge in the cytoplasm, cell wall, collagen, connective tissue, and other structures that surround and support the cell[16]. An example of how well every part can be studied in contrast of other staining methods is shown in 3.2



(a)



(b)

Figure 3.2: (a) Basal Cell Cocktail 34ßE12+p63 staining (b) H&E staining Arvaniti database, non cancerous sample

3.1.3 Gleason Grading

Cancer staging and grading is one of the most important steps in a medical diagnose, it describes how large a tumor is and whether it has spread. Tumor grade is determined by the abnormality of tumor cells and tissue studied under a microscope. Tumors that are poorly differentiated, meaning they contain abnormal-looking cells and may lack typical tissue structures, usually develop and spread faster than well differentiated ones, those that appear with more normal structures. Following these differences, pathologists designate a numerical grade following different systems depending on the type of tumor and where it is located. This information will be vital in order to properly assign a treatment.

The Gleason grading system was developed by the pathologist Donald F Gleason in the 1960s and 1970s together with the Veterans Administration Cooperative Urological Research Group (VACURG), and it is the most used grading method for prostatic carcinoma in the world. It is based on the interpretation of H&E stained samples, the pathologist studies the characteristics of tissue structure and histologic glandular growth patterns, which are categorized into nine possible basic patterns. Those nine patterns are distributed into the five grades that describe this system, their morphologies can be seen in Figure 3.3, this grades are used to establish a numerical grade from 2 to 10, adding the primary and secondary patterns found in the sample. The primary pattern is the most common one, identifiable by simple visual inspection, and the secondary pattern is the more predominant after the first. If only one grade can be spotted, the final score will be the grade in question doubled. [17]

The resulting number, also called Gleason score will determine the Gleason grade. The approach for assigning this Gleason grade has changed through the years and different International Society of Urological Pathology (ISUP) revisions of the method in 2005 and 2014, resulting in the classification shown in table 3.1. This manner of allocating grades was validated in a multi-institutional study and described in a Johns Hopkins Hospital investigation project, and even though it still has some deficiencies that can be improved, it makes Gleason Grading one of the most potent prognostic predictors for prostatic adenocarcinomas[18].

Table 3.1: Gleason Score Grade Grouping

Grade Group	**GG1**	**GG2**	**GG3**	**GG4**	**GG5**
Gleason Score	≥ 6	3+4=7	4+3=7	8	9 and 10

3.2 Artificial Neural Networks

In the first half of the 20^{th} century the British mathematician and logician Alan Turing suggested to consider the question ”Can machines think?” in [19], this was one of the first works about what would develop in the future as Artificial Intelligence (AI). It explored the most important concepts of AI and laid the groundwork of this field, which is now widely used and described as the aptitude of a digital computer to execute tasks that normally require human intelligence such as the ability to reason, visual perception, pattern recognition or the capability to learn form past experience [20].

Research in this field has permitted many useful and intricate creations that are able to help in

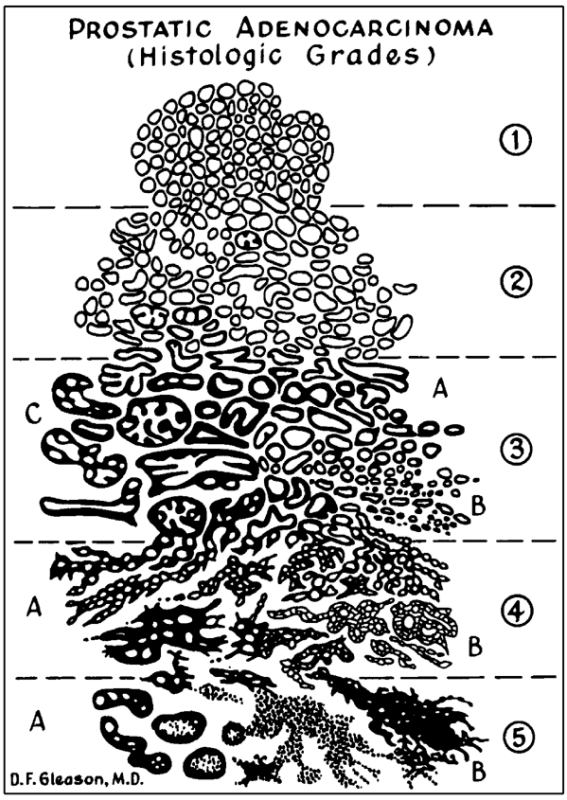


Figure 3.3: Gleason Grades: standard drawing

different areas, such as the medical field. AI helps one of the main challenges of modern medicine, dealing with enormous amounts of data and using them in order to help clinicians formulating diagnosis, or assisting in the decision making of fitting treatments [21]. One of the main technologies used are Artificial Neural Networks (ANNs), this AI technique was developed following the nature of human brain neural network. This inspiration is clearly reflected on its description and in 3.2. ANNs replicate the way brains perform certain tasks by modeling networks of highly interconnected processing units called neurons[22]. These processing elements and its connections store experiential knowledge that is acquired through a learning process, and is made available for future use [23].

Table 3.2: Similarities between biological neural networks and artificial neural networks

Biological Neural Networks	Artificial Neural Networks
Stimulus	Input
Receptors	Input Layer
Neural Net	Processing Layers
Neuron	Processing Element
Effectors	Output Layer
Response	Output and an entry

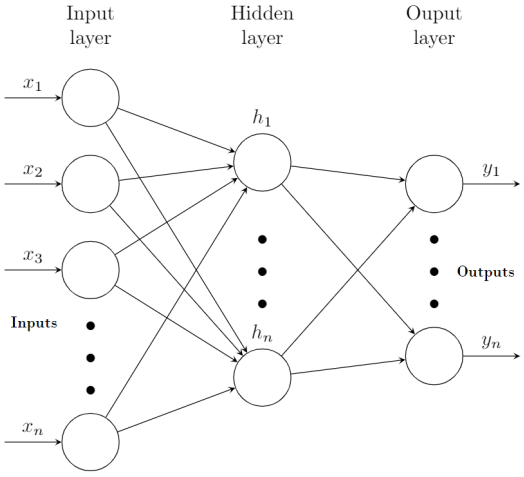


Figure 3.4: Multilayered Artificial Neural Network

3.2.1 Multilayer perceptron

In this section the Multilayer Perceptron (MLP) will be explained. MLP has been the most extensively used neural network architecture since it was described, and it started with Rosenblatt's perceptron machine in 1958. It was first defined as hardware aimed to perceive and recognize images, hence its name, instead of what is currently identified as, an algorithm. [24]

The single-layer perceptron designed by Rosenblatt was a linear classifier that relied on the neuron as its computational element. This fundamental unit can be explained by identifying its three different parts: The synapses or connecting links that will carry the input weights w_{kj} which will be multiplied by the input signal x_j, in order to arrive at the neuron k. This will be linearly combined in a summing junction and then an external bias b_k will be included in order to increase classification accuracy. The last part of the neuron is called activation function $\varphi(\cdot)$, which is applied to the weighted sum obtained before, and it simply is a predefined threshold that will determine the output of the neuron depending on the range $u_k + b_k$ is found. This can be easily understood by studying the structure of its diagram 3.5. [24]

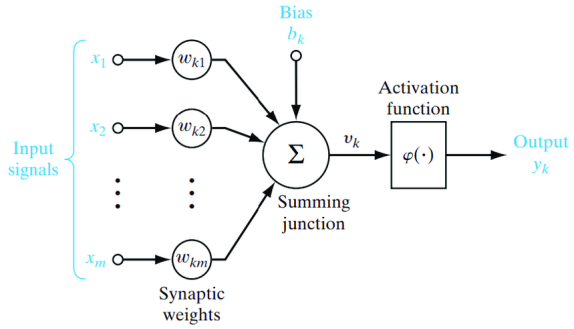


Figure 3.5: McCulloch-Pitts model of a neuron

The shallow perceptron did not include multiple layers, this meant a limitation in terms of the problems it was able to solve, such us non linear problems. In this context, the MLP was created, it is defined as a deep artificial network with a fully connected structure. MLPs are made up of an

input layer that receives the vector of data, and an output layer that makes judgments and classifies the input data. In between these layers one or more hidden layers are stacked together. This layers are truly the computational core of MLPs and have the capability of approximating any continuous function. [23]

3.2.2 Learning algorithms

To conclude, this section is going to define the most important part of any ANN, its learning process. The capacity to learn is a vital characteristic of intelligence, in ANN it can be viewed as the step in which the networks parameters, such as connection weights, are updated iterative so the performance of the model is improved.

Supervised, unsupervised, and hybrid learning are the three primary learning paradigms. This paper will focus on supervised learning, since the databases used contain the ground truths associated with each image. For this type of learning, the network is given the true answer (ground truth) and, accordingly, weights are adjusted to help the network provide responses that are as near to those known accurate answers as allowed. [23]

In order to fully understand how ANNs work, two fundamental mechanisms are explained:

- **Forward propagation:** typical of feed-forward networks, it describes the direction of the input data an how it is processed. Input is fed only in the forward direction through the network, starting at the first layer, and being processed through each hidden layer finally creating an output.
- **Back-propagation:** it is the process of adjusting and fine-tuning the synaptic weights depending on the error obtained in previous steps or epochs.

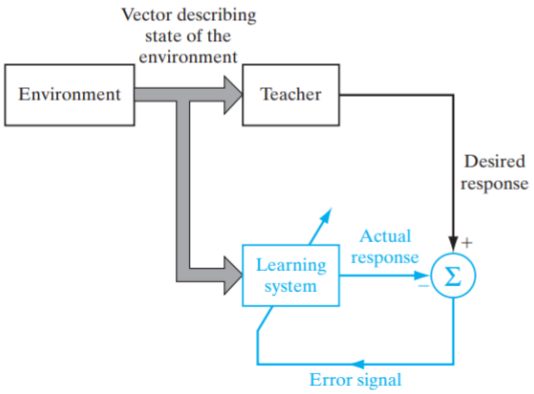


Figure 3.6: Diagram of supervised learning

Optimization functions are used to find the weights that will enhance this process of feed-forward and then back-propagation. In order to achieve this some techniques locate the absolute minimums of the error signal, and therefore help the model to get results faster. The most used function is the gradient descent method, namely the stochastic gradient descent, which tries to reach the minimum of the loss surface by descending it in predefined steps. The idea behind this

algorithm is to start at a random point on the loss function and travel down its slope until the lowest point is reached, that is, when it converges [23]. The steps in which the algorithms moves down the function is determined by the gradient of it and by the learning rate, a parameter that can be changed depending on the convergence tendency observed. A large learning rate may cause to jump across the desired point, since it takes broad steps, that is why low learning rates are used. However this may cause a slow approach to convergence, for this reason programmers use creative coding methods such as learning rates schedulers.

3.2.3 Data usage in training, validation and test processes

Deep learning bases its potential in learning through examples, extracting the common patterns and fundamental characteristics from samples of data. That is why the quantity of data available and how it is treated to make the most of it is one of the most important things in ANN modeling.

As it has been stated earlier, this paper focuses on supervised learning, namely in prostate biopsies classifying. Since generally there is a limited amount of data of this sort available, it is important to follow a strategy in dividing the samples accessible in each part of the network training.

There are different ways of partitioning the data, and the common idea behind it is to avoid overly optimistic estimates in order to guarantee that the model will not fail when predicting new data. In this document we will explain the most common partition method, double partition. The core concept in this technique is to segregate all of the data into three datasets, training, validation and test.

- **Training dataset:** this partition of data is used to fit the model. By training the algorithm with this data it learns from its patterns and we obtain the actual weight values used to predict.
- **Validation dataset:** this sample of data provides information about how good does the model fit still in the development process. It is the first unbiased evaluation of the accuracy achieved by the algorithm, we use this results to fine-tune the model parameters.
- **Test dataset:** it provides the final unbiased evaluation of a model. The test set is used once a model is completely trained, and it is normally chosen to represent as well as possible the diverse classes that the algorithm would face.

The remaining aspect to define is the dataset split ratio, it usually depends on the total number of samples available and on the particular model we are training. That is why there is no exact rule on the percentage of the data that should go on each partition, however there are some recommendations. For datasets with a standard number of samples, the recurrent ratio is 60:20:20, that is 60% of the samples will be part of the training set, 20% to the validation set and the remaining to the test set.

3.3 Convolutional Neural Networks

Computer vision (CV) is an interdisciplinary field of AI which includes methods to acquire, process and study digital images and videos with the purpose that computers gain meaningful information

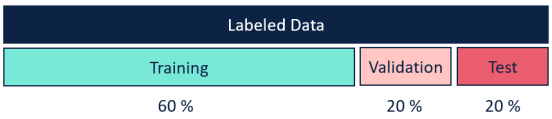


Figure 3.7: Dataset split ratio

from them. It tries to mimic the human vision system, especially its pattern recognition or object detection abilities[25], but in much less time, which stands an incredible advantage to certain fields of science such as the medical field. Deep learning (DL) has allowed to overcome even the highest challenging problems in this area with the help of one of its most potent an popular tools, Convolutional Neural Networks (CNNs).

The core of CNNs and their potential resides in the use of kernels, matrices of values that are multiplied with the input data in order to detect its features. Various kernels stacked and linked together are known as filters, and the main idea behind them is to detect whether a feature is present on the given input or not. To better understand how CNNs work it is important to describe the behavior of their different layers[26].

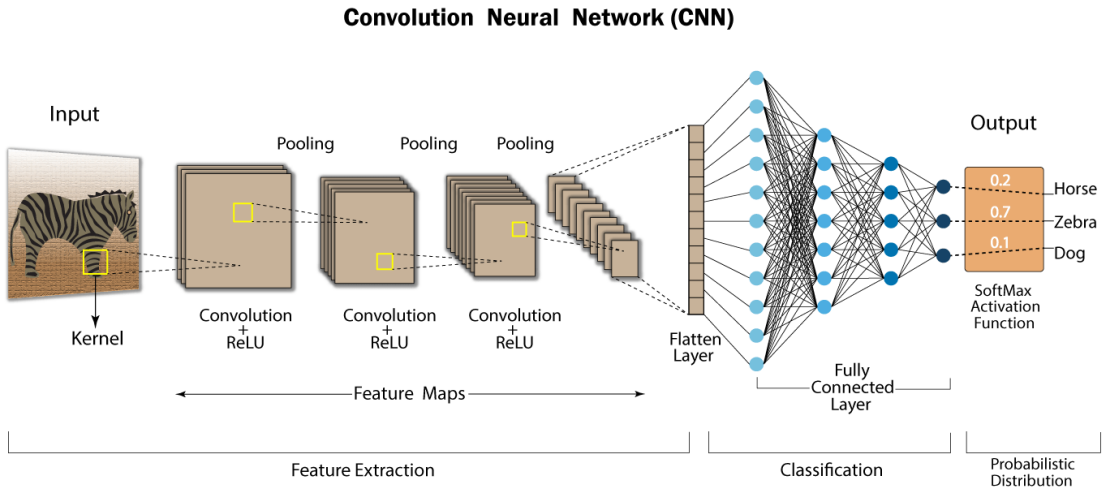


Figure 3.8: Example of a Convolutional Neural Network for image classfication

- **Convolutional layers:** in this step, filters formed by small kernels are applied to the original image, as well as intermediate feature maps, by means of the convolution operation. This operation is applied by a method called sliding window, where patches of the input matrix sized as the kernels convolve. When the windows slides through the birders of the matrix more adjacent values are needed, this is called zero-padding.

  Behind each convolutional block there is an activation function, normally presented as a layer, the most common activation functions are the signoid function and Rectified Linear Unit (ReLU) function.

- **Pooling layers:** these layers perform an operation called downsampling to reduce the width and height dimensions of the network. It does not affect the depth of the network, however it

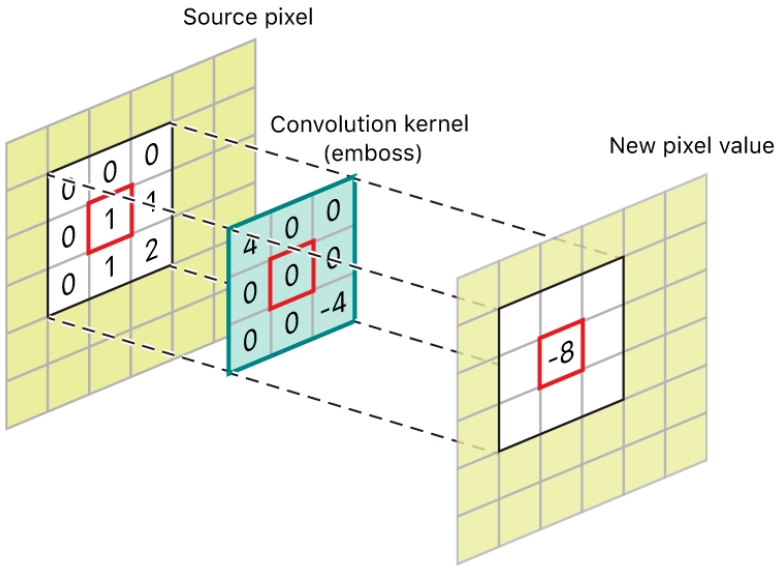
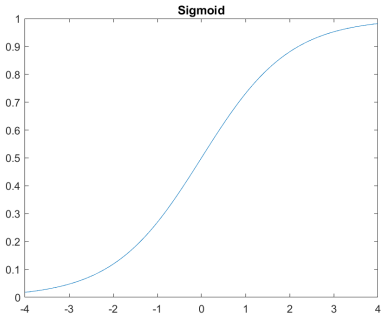
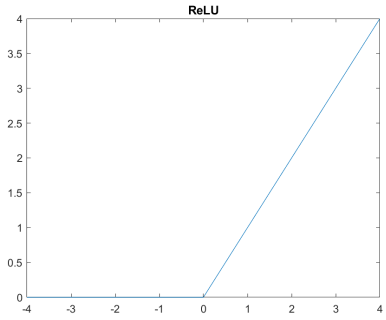


Figure 3.9: Convolution Process



(a) Sigmoid Activation Function



(b) ReLU Activation Function

Figure 3.10: Representation of most used activation functions

does translate in a loss of information. This loss is diminished by the fact that it decreases the computational cost of the algorithm and increases the training speed. There are two different strategies in order to decrease the size of the layers: average pooling and max pooling. The first one takes the average value of the section covered by the kernel, and the second one takes the maximum value in the filter section. Max pooling is widely used since it also acts as a noise reduction filter.

- **Fully-connected layers:** placed before the classification output, they convert the two dimensional final feature maps into a vector with the aggregated information. This flattened layer is the input of what essentially is a MLP with fully-connected nodes, that will finally determine the classification of the input after passing through one last activation function. The most common activation functions at this point are softmax and sigmoid, depending on the number of classes.

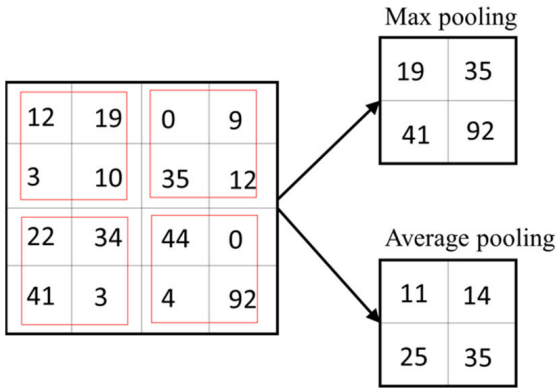


Figure 3.11: Types of pooling

3.3.1 CNN Architectures

Over the past few decades there has been numerous breakthroughs in the development of CNN architectures. This is largely due to the variety of contests and challenges proposed by the AI community. Initiatives such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) have contributed to great advances in the field of deep learning, achieving through various competitions the error rates reduction of the most famous models, and thus increasing their precision. The 4 most relevant CNNs in the recent history of deep learning will be shown below [27].

- **LeNet-5:** This neural network was proposed by the computer scientist Yann LeCun in 1989. The prototype of this architecture was born in order to help the United States Postal Service identify handwritten postal codes. It became one of the pioneering works for the development of what we know now as deep learning.

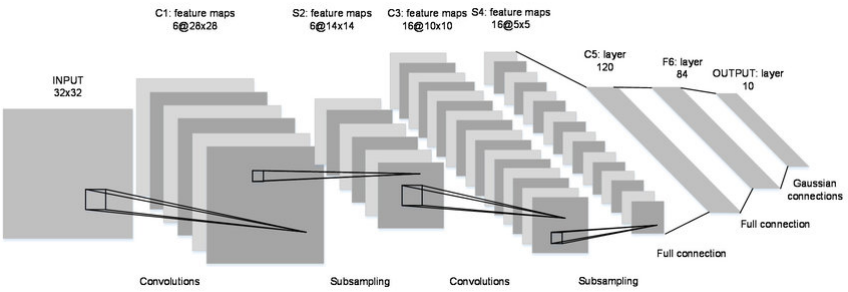


Figure 3.12: Structure of LeNet architecture

- **AlexNet:** Name of the convolutional network formed by 5 convolutional layers developed by Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. This model won the ImageNet challenge in 2012, and it was the first architecture to use elements such as max-pooling layers or the ReLU activation function [28].

- **VGG-16:** Created by the Visual Geometry Group of the Oxford University in 2014, it made the term ”deep” highly common. In their article ”Very Deep Convolutional Networks for Large-Scale Image Recognition” [29] it was demonstrated that the most intuitive way to im-

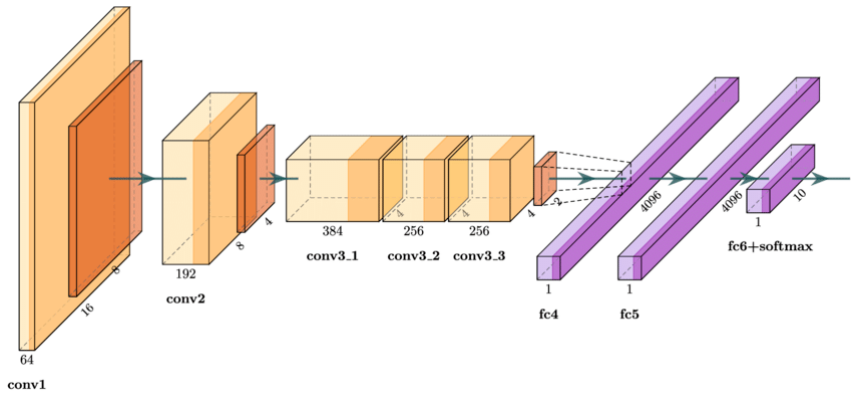


Figure 3.13: Structure of AlexNet architecture

prove results obtained by previous networks was to add more layers. Therefore they developed this architecture with 16 layers, 13 convolutional and 3 fully-connected, using smaller filters but stacked together, which helped them achieve 92.7% accuracy.

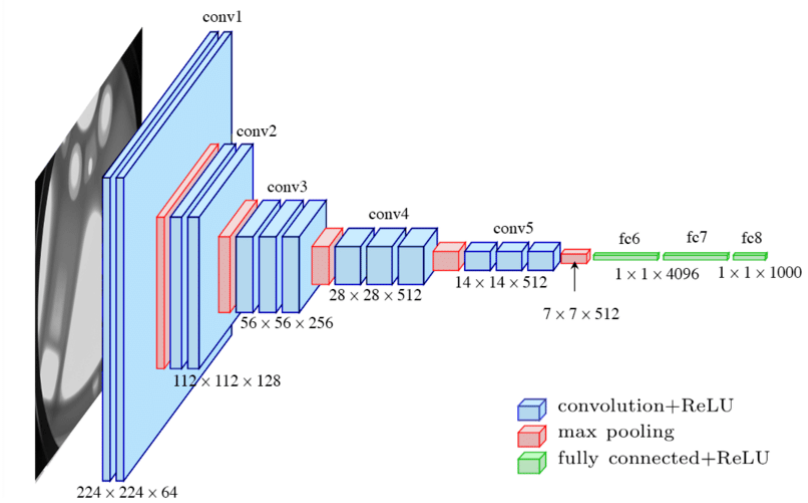
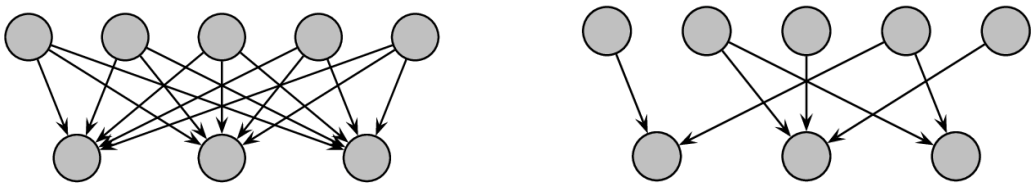


Figure 3.14: Structure of VGG-16 architecture

- **InceptionNet or GoogleNet:** This architecture won the ImageNet challenge in 2014. It was proposed in the paper "Going deeper with convolutions" [30], that studies the way to increase the size of the networks but in an efficient way, since an excessive enlargement makes the model easily wander into overfitting. Another consequence is reflected in an increase of the computational cost beyond the existing resources. To overcome these problems, Inception v1 neural network is proposed, which bases its potential on connecting the layers sparsely instead of fully-connected as shown in the image 3.15.
- **ResNet:** This sort of architecture works with the concept of residual blocks to boost its performance. These blocks respond to the fact that in very deep networks the gradient ends up adopting very small values that keeps the weights from updating. This, together with the decrease in training precision, is attempted to be solved by this technique, which includes



(a) Densely connected architecture **(b) Sparsely connected architecture**

Figure 3.15: Connection differences between dense and sparsely connected architectures

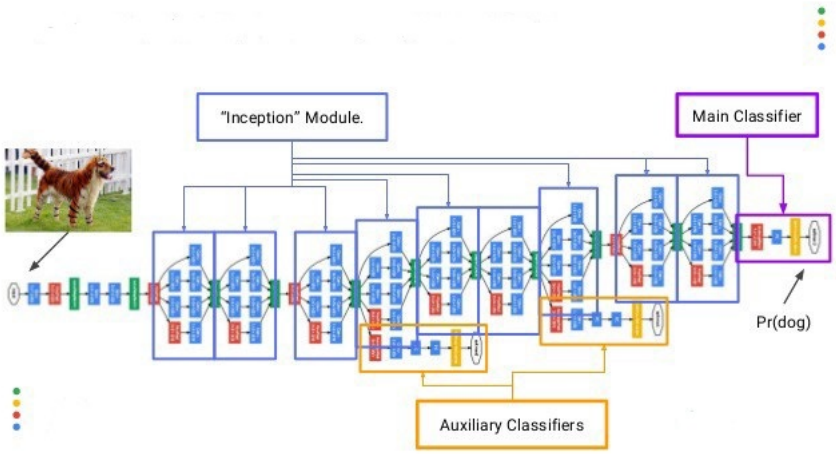


Figure 3.16: Structure of Inception v1 architecture

generic information from the first layers in the subsequent ones. By doing this, deeper layers have their own information, very specific, but also they have the most general data so the result is not completely wrong.

3.3.2 Categorical Cross Entropy

In previous sections, the importance of learning algorithms such as gradient descent has been explained. However, these algorithms must be accompanied by a loss function, since they need a way to quantify how well the model is working in order to optimize it. To evaluate this behavior, loss functions measure how far the value estimated by the model is from the real value. Different loss functions perform the necessary calculations in different ways. The decision of which cost function to use is one of the most important when describing the model, as it has a significant impact on how the model learns and therefore on the final results in terms of accuracy and other metrics.

The choice depends on the type of problem being treated, in this particular case we will use the categorical cross entropy as a cost function, since it is a very common loss function in classification problems with more than two classes. This function is given by the following formula:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * log(p_{ij}) \tag{3.1}$$

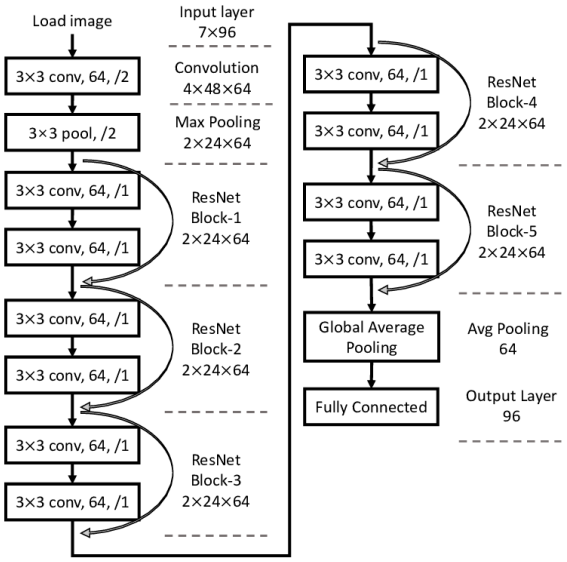


Figure 3.17: Structure of Resnet architecture

Entropy, in the computer science area, comes from a term in information theory, and explains the amount of information, or uncertainty, that a message carries according to its probability distribution. Therefore, the cross entropy measures the difference of that amount of information between two random variables. If the probability difference between the predicted and the real one grows, so will the loss of cross entropy. For this to work properly when implemented in an algorithm, the labels must be encoded in one hot encoding and the classes must be mutually exclusive.

Chapter 4

Materials

4.1 Datasets

We can differentiate the datasets used in this work depending on the purpose of the process they have been used for. In the case of the training and perfecting of the structure, parameters and functions used within the DL model, the dataset provided by the SICAP project was used. The SICAP program is a Spanish project funded by the Ministry of Economy and Industry in which several research groups from Spanish universities, such as CVBLab from Universidad Politécnica de Valencia, cooperate with Hospital Clínico Universitario de Valencia in order to develop a software for computer aided diagnosis. This is intended to help with prostate cancer automatic diagnosis via image processing with deep learning algorithms, and the possibility of integrating said system in routine clinical practice. The SICAPv2 database contains 155 prostate whole slide images, local level annotated by a team of professional and experienced pathologists using the Gleason grading system. We refer as whole slide imaging to the digitization of ordinary glass slides, in this case the digital imaging of prostate biopsies[31]. However the sets used will not be the WSIs explained, but 512^2 pixel patches extracted and downsampled at 10 $\times$ resolution, extracted from the original slides. The optimum size for binary classification cancer vs. non cancerous was obtained in [32].

On the other hand, when evaluating the model trained two external databases were used. This is one of the principal focuses of the work done, since when trying to create a tool that can be used in regular clinical practice, generalization is one of the main traits we are searching for. The testing of our model on external source databases helps cover this issues and it has done with the two of the most used public databases existing. The first database, shared by Arvaniti et al. is composed of 886 Tissue Micro Arrays (TMAs), and is also of good use since the author proposes some inter-pathologist confusion matrices [3] that will be used in the experiments; and the second one published by Gertych has 625 patches of prostate whole slide images; the number of patches that belong to each class of the different datasets are given in table 4.1.[33]

In order to test the DL model with the external datsets there has to be a preprocessing stage, where images are treated and prepared due to the difference in color caused by manual cell sectioning and hematoxylin and eosin (H&E)stain concentration. With the purpose of standardizing the color variation a channel-wise histogram was used followed by the metohod described in [34].

Table 4.1: Distribution of the annotated patches Gleason grades

Database	NC	GG3	GG4	GG5
SICAP	4417	1635	3622	665
SICAP (TEST)	644	393	853	232
ARVANITI	115	274	210	104
GERTYCH	32	95	216	70

4.2 Hardware

One of the most important factors when studying machine learning algorithms is the computational budget that one has available. This is due to the fact that implementing techniques that bring us closer to high accuracy in computer vision problems entail a great computational effort. Therefore, the hardware available is one of the vital factors for the development of a project. Thus, the hardware elements that have been used during this work will be specified below.

The algorithm programming has been carried out with a personal computer, as well as the displaying of the results. This computer is a Microsoft Surface Laptop 3, which counts with the following characteristics:

Operative System	Windows 10 64-bit
Processor	Intel®Core™i5-1035G7
Frequency	1.20 GHz
GPU	Intel®Iris®Plus Graphics

Table 4.2: Laptop specifications

The algorithm has been deployed and executed through MobaXterm into the CVBlab computing servers, which count with fitting hardware resources to train the model; specifically they use NVIDIA Titan V Graphical Processing Unit.

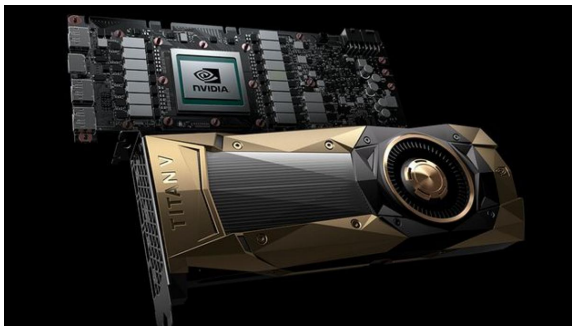


Figure 4.2: NVIDIA Titan V GPU

4.3 Software

The environment used for developing the algorithms and perfecting the models was PyCharm 2020.3.5, a Python Integrated Development Environment designed by JetBrains, a software de-

velopment company from the Czech Republic. PyCharm provides the tools for an intuitive programming using Python 3.9 , an object-oriented high-level programming language. It offers great versatility since it is an open-source language with countless libraries and modules developed by a great community, hence a perfect tool for a remarkable variety of fields, from web development to artificial intelligence; for this reason it is widely used in deep learning projects, and it will be used on this thesis. Two of the most notable libraries used are Tensorflow an Keras, being the first one an end-to-end open source platform for machine learning model building that makes the most of intuitive high-level APIs (Application Programming Interface) like Keras, and grants the possibility of taking advantage of the whole potential GPUs have to offer, by carrying out tensor operations in an efficient and productive way, creating highly scalable programs.

As for the formatting and drafting of this document the LaTeX typsetting system for was used with the aid of LaTeX editor, Overleaf. This software system is widely used for scientific and technical document preparation, since it helps researchers to focus on the content instead of the presentation.

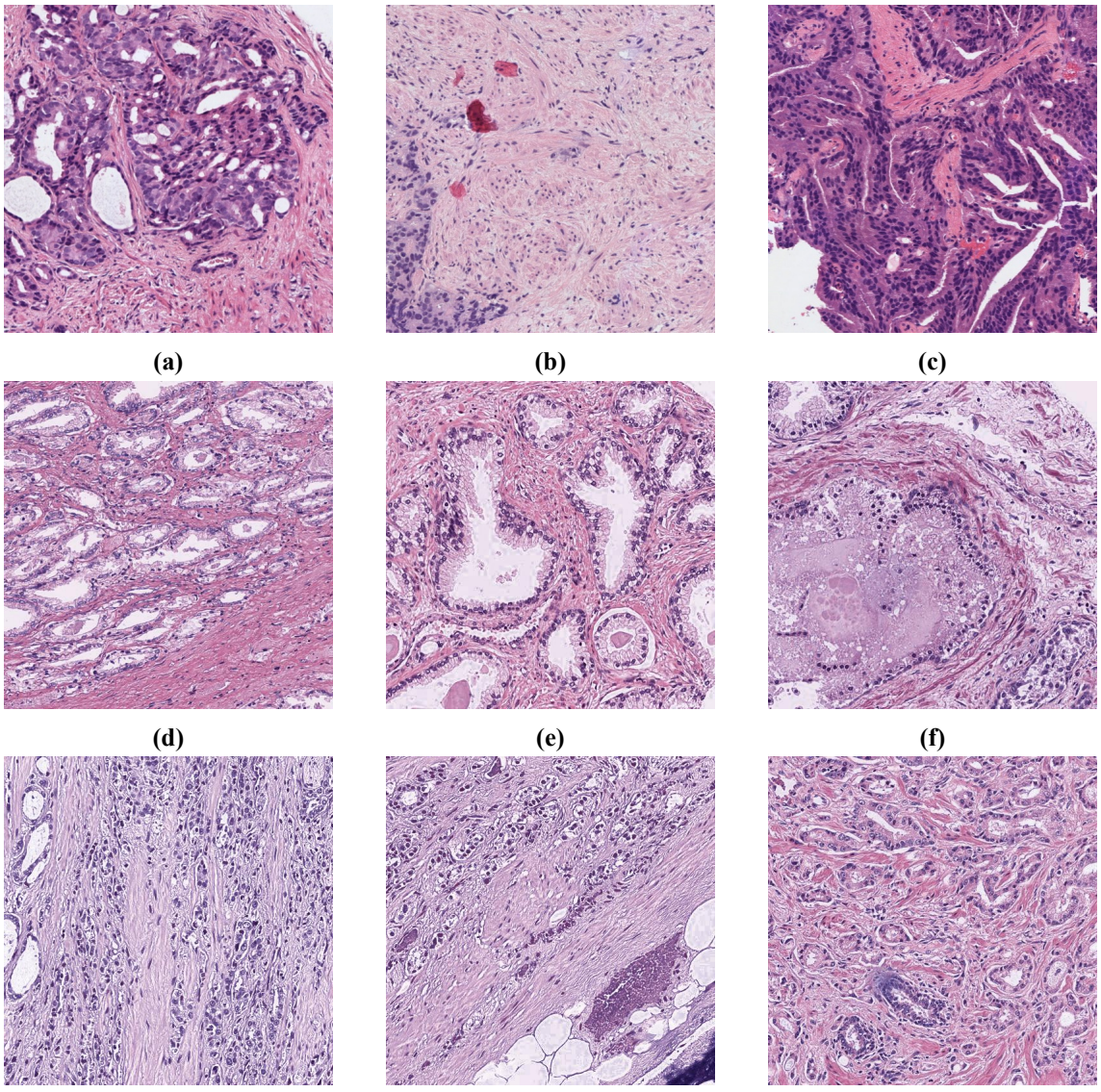


Figure 4.1: Sample patches: (a) and (c) correspond to Gleason grade 4 structures and (b) contains grade 5 patterns, all of them pertain to SICAP database. Samples from the second row were extracted from the Arvaniti dataset, where (d) presents grade 3 shapes, (e) has grade 4 morphology and (f) correlates with a non cancerous tissue. As for the last row, all of the samples (g), (h), (i) are classiefied as Gleason grade 4.

Chapter 5

Methodology

5.1 Incorporating Uncertain on Automatic Gleason grading

The Gleason grading system was updated in 2005 by the International Society of Urological Pathologists (ISUP) due to the constant evolving of prostate cancer testing methods. Regardless of this persevering growth in prostate cancer diagnosis, Gleason grading remains the main classifying method, and despite the 2005 improvement in the sorting procedure, it is still a grading system based on a subjective microscopic analysis. This implies that, inevitably, there will be disagreements between different pathologist, since the outcome of their decision will be influenced by the type of facility they work on, as well as their training and expertise.

This section will describe some tools and procedures used in deep learning to examine, quantify and incorporate interobserver agreement into a deep learning algorithm, helping with imbalanced databases along with taking into consideration the noise that those agreements or disagreements present, in order to develop an algorithm that is going to be more similar to human pathologist behavior.

5.1.1 Weighted Kappa loss function

As it has been explained in 3, loss functions are one of the crucial attributes in machine learning and deep learning algorithms, since they determine how well is the programmed algorithm fitting to the data at one's disposal. They evaluate the algorithm's performance based on how far the predicted value is from the actual value. However, the method used to calculate this difference depends on the context of the question one is trying to solve and can diverge a lot.

Weighted Kappa loss was introduced in *Weighted kappa loss function for multi-class classification of ordinal data in deep learning* [35], where Quadratic Weighted Kappa (QWK) is proposed as a loss function for ordinal classification problems. Cohen's Kappa is a statistic measure that quantifies the level of agreement between raters on the classification of elements into different classes. It is commonly used in machine learning model testing, since it measures the quality of an algorithm's performance by quantifying the inter-rater reliability. It calculates the agreement between two independent raters but removing the effect of randomness, since the two raters could agree accidentally. This last part is adjusted by computing the agreement by chance and subtracting it from the actual observed agreement.

However, this metric does not take into account the level of difficulty that the classification of medical imaging presents. The grading of this type of images such as biopsies can be challenging, and determining the severity of cancer that a sample of tissue has, can have different outcomes depending on the pathologist that analyses it. In order to help minimize this problem, weighted kappa is defined. It measures how reliable a model and its interpretations are by considering a table of predefined weights, that represent the disagreement between observers and punish the wrong classification of distant classes.

Quadratic Weighted Kappa can be described as follows:

$$\kappa = 1 - \frac{\sum_{i,j} \mathbf{W}_{ij} \mathbf{O}_{ij}}{\sum_{i,j} \mathbf{W}_{ij} \mathbf{E}_{ij}} \tag{5.1}$$

Where the three main matrices, W, O and E are the weight matrix, the matrix of observed scores, or confusion matrix, and the outer product between the actual number of times an outcome has happened and the predicted frequency of those outcomes respectively. The weights in W matrix designate the penalty of misclassifying class i as class j and the other way around,and is calculated following:

$$\mathbf{W}_{ij} = \frac{(i-j)^2}{(c-1)^2} \tag{5.2}$$

Where c is the number of classes. As it can be inferred form the expression, the cost of misclassifying farther classes will have more impact than faulting in the classification of closer classes, since classifying a healthy sample as a grade 5 adenocarcinoma would be an extremely severe error, but categorizing a grade 4 ill tissue with grade 3 one could be easier to compensate.

5.1.2 Cost Sensitive learning

As it has been explained before, annotating prostate samples can be a noisy procedure with interobserver disagreement. Cost sensitive classifiers can be another method, like quadratic kappa loss, to help with class imbalanced scenarios and take into account this difference in opinion. In this paper we will implement cost sensitive loss as a regularizer term within a loss function.The foundation from which we can rest our final function on is a common base loss function, like cross-entropy, so then we can add the cost sensitive term multiplied by a coefficient. The new loss function will adopt the following structure:

$$\mathcal{L}^{CS}(\hat{y}, y) = \mathcal{L}^{BASE}(\hat{y}, y) + \lambda \langle \hat{y}, M(y, \cdot) \rangle \tag{5.3}$$

Where M is the average matrix of $M^{(2)}$, which represents the part of the cost that increments at the same time that distance does, and $I - M^*_{opht}$ the component that represents the information in relation to inter-pathologist variability when evaluating prostate tissue microarrrays (TMAs), in the form of normalized confusion matrices.

$$M = (M^{(2)} + I - M^*_{path})/2, \qquad M^{(2)}_{ij} = \|i - j\|_2^2 \tag{5.4}$$

The aim of taking into account the disagreement within ratings is simple, to punish incorrect predictions harder when the label assigned by the annotator is expected to be trustworthy than when it is unreliable, and the algorithm can be more permissive. In doing so, unlike when using the weighted kappa loss function, we diverge from standard learning algorithms, which presume that all mistakes will have identical costs to penalize various sorts of faults while training, and instead we will encode those penalties as a criterion for the loss function.

5.2 Attention-Based Pooling for multiscale feature learning

Aside from attempting to enhance the accuracy of a model by experimenting with new or improved function losses, there is still a lot of room for improvement in a deep learning algorithm, from hyperparameters tuning to choosing the right optimization. For that reason, researches are constantly improving and discovering new methods to boost the performance of machine learning algorithms.

This section will describe one of the techniques used in the experiments, which focuses on how patterns are perceived in the processing of prostate tissue images, and what value does the algorithm give to those patterns in order to classify a sample, a method known as Attention Mechanism.

5.2.1 Attention Layers

Machine learning has helped us develop innovative solutions for problems that have existed for along time now, however researcher are still working on comprehending methods that are already working or trying to unfold new ones. One way of trying to discover what can be done in order to improve machine learning algorithms is to compare it to human behavior, something that has been really helpful in deep learning, and image processing particularly. This is somehow what attention mechanisms do, they recreate the process that a human brain does when processing data through vision, identifying different patterns, and prioritizing the ones that we deem more important, by focusing on them and bringing attention to them, and therefore giving the rest of sequences less importance presenting them in a lower resolution.

Attention mechanisms are a fundamental part of modern and innovative architectural designs, they have been proven to be crucial in the improvement of most CNN-based procedures. Therefore, in this thesis we will use attention layers that will determine which pieces of data are more relevant, using weights that will be trained during the process and will attract the DL model to those crucial parts.

Within this context, we can define two types of different attention mechanisms: soft and hard attention. The latter model focuses only on a section of the image, choosing the one that is most

likely to be relevant when classifying the image, while in soft attention the entire image is taken into account but the model will consider more important some parts of the image. This requires a higher cost in terms of computation and memory, however it involves a simple training process, while in hard attention the objective is not differentiable and therefore requires a difficult training process.

However, there are different ways to implement these types of attention mechanisms, depending on the stage of development in which it is implemented: "a posteriori" network analysis, or integrated attention mechanisms [36].The last-mentioned one does not carry out a subsequent analysis for the interpretation of the predictions but rather integrates this mechanism in the same training with parameters that will change and learn.

In this paper we will focus on the so-called gated attention mechanism, which works with intermediate features maps in a way that is more efficient, concentrating on the structures that it considers most relevant and thus achieving a suppression of non-relevant areas, which will accomplish a more accurate model with a not much higher computational cost. [37] This is achieved through the establishment of gates in the information flow, these gates will administer what data will be propagated to the rest of the layers [38], and can be defined as follows:

$$a_k = \frac{\exp\left\{\mathbf{w}^\top \left(\tanh\left(\mathbf{V}\mathbf{h}_k^\top\right) \odot \operatorname{sigm}\left(\mathbf{U}\mathbf{h}_k^\top\right)\right)\right\}}{\sum_{j=1}^{K} \exp\left\{\mathbf{w}^\top \left(\tanh\left(\mathbf{V}\mathbf{h}_j^\top\right) \odot \operatorname{sigm}\left(\mathbf{U}\mathbf{h}_j^\top\right)\right)\right\}} \tag{5.5}$$

5.2.2 Coupling multi-scale features

This section will describe an additional technique focused on helping with the issue of visual saliency, that is, the process used by cognitive systems to guide attention to certain regions or characteristics when classifying images. As explained in Section 3, convolutional networks (CNN) are neural networks that consist of a certain number of layers of different types. The reasoning behind pattern recognition through the combination of multi-scale features is to extract deep features from images at different stages to improve accuracy. [39]

The interest in using this technique in the case at hand, arises from the fact that there are notable differences in the complexity of the patterns present in samples of grade 5 and grade 3. Since grade 3 patterns are visually very complex, our hypothesis is that taking into account lower level features, through attention mechanisms in intermediate layers, the algorithm will improve when interpreting less complex patterns, such as those present in patches of degree 5, while maintaining good performance among the rest of grades.

In state-of-the-art experiments carried out using this technique, intermediate feature maps are extracted and used as inputs for different convolutional network models. This maps are representations of the images, and are present after convolutional layers in different scales and sizes depending on their location in the architecture. Once instanced and processed by the networks, every output is merged for final classification [40]. In this paper, the representations mentioned above will be taken as input for the attention layers described in the previous section. All this will be better explained in the next section.

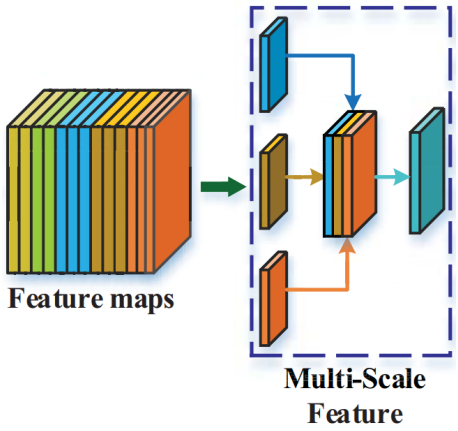


Figure 5.1: Merging multi-scale features

Chapter 6

Experiments and Results

This chapter will explain the results obtained during the various experiments and the strategy followed to reach them. In order to achieve this, we will first emphasize the main purpose of the project, since the method that has been followed for the succession of experiments and the interpretation of the results that these have generated, are better understood in the contextualization of the project's objectives.

Thus, the purpose of this project is to train a CNN for cancer grading in patches extracted from prostate biopsies. As one might expect, we started from a basic model to gradually increase the complexity to achieve a robust model even in external databases. These advances were introduced with the aim of solving the problems and limitations that appeared during the training phase, or the aspects that were identified as subject to improvement, and they will be explained following this progressive logic.

As explained in the previous section 4, this project has been carried out with the SICAPv2 database, one of the largest public databases containing pixel-level annotated prostate biopsies. After the preprocessing, explained in that same section, there is a considerable enlargement of the collection, as described in 6.1 table. Here we can observe that in 763 Gleason Grade 4 (GG4) patches, the existence of cribiform patterns was specified during annotation. The latter will not be relevant in our project but it is worth mentioning as it is part of our database. The four fold cross-validation method was used to partition the dataset, however in this project we only used the first partition as the main dataframe, see Table 6.2.

In the first section, the metrics used to evaluate the performance of our model are explained, while in the second section we will begin to describe the experiments that we carried out in order to include interobserver variability. Then in the third section the attention and feture learning

Table 6.1: SICAPv2 database description. Amount of whole slide images and their respective biopsy-level primary label (first row) and number of patches of each one of the Gleason categories (second row) [4]

	Non cancerous	Grade 3	Grade 4 (cribiform)	Grade 5	Total
#WSIs	37	60	69 (36)	16	182
#Patches	4417	1636	3622 (763)	665	10340

Table 6.2: Number of patches for each grade

	Patches			
Group	Non Cancerous	GG3	GG4 (Cribriform)	GG5
Training	3773	1242	2769	433
Test	644	$6-393$	$9-853(145)$	$2-232$

mechanisms will be explained.

6.1 Specification of used metrics

In order to properly evaluate the performance of a machine learning algorithm, there are different types of evaluation metrics at one's disposal. Choosing the right metric is essential, since it will determine the interpretation of the results a model is giving. For instance, a model may apparently have good accuracy, but behave poorly with other statistic measures. The metrics used in this project are classification accuracy, *F1-score* and Cohen's quadratic *Kappa* score. Lastly confusion matrices are used to visualize the complete performance of the model.

- **Confusion Matrix:** We will begin by explaining the structure of a confusion matrix, since it has some key concepts in order to understand the definition of the metrics proposed. A confusion matrix is a table-like representation method that exposes ground-truth labels versus model predictions [41].

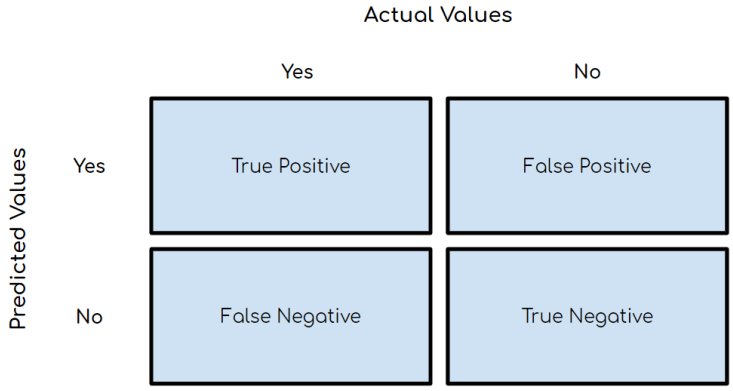


Figure 6.1: Understanding confusion matrices

The binary case will be explained since the multi-class case is an extension of this, however our confusion matrices will be 4 x 4 since there are 4 labels. When contrasting ground-truths or the actual values, with the values that the model has predicted, 4 cases can happen:

  - True Positive (TP), the predicted value and the original coincide and the statement is positive.

  - True Negative (TN), the prediction is correct but the statement negative.
  - False Positive (FP) or Type 1 Error, incorrectly classifies the negative class as positive.
  - False Negative (FN) or Type 2 Error, incorrectly classifies the positive class as negative.

  To summarize, the elements present in the diagonal represent the accurate prediction for each class, whilst the off-diagonal ones denote the incorrect predictions.

- **Classification Accuracy:** This is one of the most basic measurements in classification models and represents the ratio between the number of correct predictions and the total number of predictions made. In terms of the concepts explained above, it would be calculated as follows:

$$Acc = \frac{FP + FN}{TP + TN + FP + FN} \tag{6.1}$$

- **Cohen's Kappa:** This statistical coefficient has been explained in the methodology section 5, since the loss function that uses this factor is one of the main research points of this paper. Even though loss functions not only explain the behavior of the model but are used to train it, while metrics monitor and quantify the performance in training and testing, Cohen's quadratic kappa definition remains the same as the one given 5.1.1.

- **F1-score:** This metric is defined by two other concepts, precision and recall. Precision gives information about the relationship between the correct positive predictions among all the positive predictions. Otherwise, recall represents the fraction of positive samples that were correctly predicted. Following the analogy exposed while explaining confusion matrices:

$$Precision = \frac{TP}{TP + FP} \tag{6.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{6.3}$$

  Thus, F1-score combines this to measures into a single one. This is achieved by computing the harmonic mean of precision and recall, which is commonly used as an average of ratios [41].

$$F_1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{6.4}$$

6.2 Incorporating Uncertain on Automatic Gleason grading

One of the innovative aspects that have been investigated during the development of this paper has been the incorporation of uncertainty and inter-pathologist variability, and how the fact that this exists can help us when training the algorithm. In order to integrate this method to reduce the disagreement noise, two experiments were completed.

After various considerations and different ablation experiments in both analyses, the logic of starting from a basic loss function, such as Categorical Cross Entropy, and then integrating the

Table 6.3: Hyperparameters specification

Input Shape	Batch Size	Optimizer	Learning Rate	Epochs
224, 224, 3	15	Stochastic Gradient Descent	0.0001	100

Table 6.4: Figures of merit for the best three tries

Method	ACC	F1S					k2
		NC	GG3	GG4	GG5	Avg	
				SICAP			
cce	0.75	0.88	0.64	0.74	0.52	0.74	0.80
cce + kappa ($\alpha = 10$)	0.69	0.81	0.52	0.70	0.49	0.69	0.72
cce + cost sensitive ($\alpha = 10$)	0.78	0.87	0.66	0.80	0.59	0.77	**0.83**
				GERTYCH			
cce	0.62	0.62	0.66	0.67	0.16	0.58	0.52
cce + kappa ($\alpha = 10$)	0.66	0.70	0.66	0.70	0.41	0.64	0.61
cce + cost sensitive ($\alpha = 10$)	0.72	0.79	0.73	0.77	0.37	0.69	**0.66**
				ARVANITI			
cce	0.57	0.42	0.71	0.59	0.04	0.53	0.48
cce + kappa ($\alpha = 10$)	0.62	0.66	0.70	0.60	0.26	0.60	**0.64**
cce + cost sensitive ($\alpha = 10$)	0.61	0.58	0.71	0.62	0.21	0.59	0.57
				AVERAGE			
cce	0.65	0.64	0.67	0.67	0.24	0.62	0.60
cce + kappa ($\alpha = 10$)	0.66	0.72	0.63	0.67	0.39	0.64	0.66
cce + cost sensitive ($\alpha = 10$)	0.70	0.75	0.70	0.73	0.39	0.68	**0.69**

differential term by means of a coefficient, has been followed. Thus, the same hyperparameters have been used for both experiments, defined in Table 6.3, however the regularizer term and the coefficient that multiplies it have changed. This structure was explained in section 5 and can be seen in 5.1.2 equation.

Finally, it should be pointed out that the values used to represent inter-observer agreement, as explained in section 5, are described by the following agreement matrix 6.5.

$$M^*_{path} = \begin{bmatrix} 0.71 & 0.29 & 0.00 & 0.00 \\ 0.00 & 0.47 & 0.53 & 0.00 \\ 0.00 & 0.03 & 0.84 & 0.12 \\ 0.00 & 0.00 & 0.20 & 0.80 \end{bmatrix} \tag{6.5}$$

Out of the different experiments that have been carried out, the results obtained using the best hyperparameters for each setting will be presented in Table 6.4. Note that although in most cases it is correlated, we do not look only at the accuracy of the model, but also at the k2 and f1-score of the most problematic classes, such as GG4 due to the complexity of its patterns or the GG5 as a result of the imbalance in the existing samples.

In pursuance of forming our conclusions, we will explain the apparent meaning of the results

presented in table 6.4. In the first place, it should be noted that the model trained only with CCE loss function is presented in order to establish a comparison for the following methods. Thus, we can observe that the loss function with the kappa term is of little use in the training data set, however, it stands out in external databases, being even the most strong one when tested in Arvaniti images. On the other hand, the loss function that includes the cost sensitive term remains robust both in the training database and in external databases and stays, on average, the best of all the experiments.

In order to illustrate and properly visualize the meaning of the metrics represented, we will display a series of confusion matrices. Figures 6.2 and **??** clearly reflect the impact of improved models, since those that have greater Cohen's kappa have more elements that tend to be on the diagonal.

6.3 Incorporating Attention-Based Pooling for multiscale feature learning

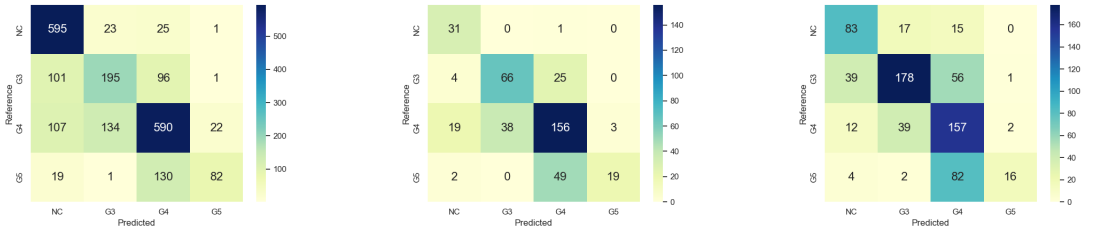
Once loss functions have been improved, we plan to incorporate another differential element. In this experiment, the structure of the neural network has been altered so that it takes the outputs of intermediate layers and passes them through an attention mechanism. Along these lines, experiments have been carried out in initial and more advanced blocks. Firstly, the outputs are extracted from the last layers of each block and passed through the gated attention layer and then they are embedded back to the network in order to pass through the fully connected layers and classify the images.

Again the same hyperparameters 6.3 have been used between experiments, which at he same time coincide with previous experiments. In addition, the same database has been used for training, and the same external databases for testing, to ensure robustness.

As we can infer from the results seen in Table 6.5, concatenating our attention layer to the last layer of the fourth convolutional block, makes the model not only more accurate in the SICAP database, but also stronger in external databases. This is due to a better perfomance when classifying GG5 patches, which supports our hypothesis that low-level pattern input helps to classify images with seemingly simpler patterns. For this reason, despite the fact that, globally, a very drastic improvement is not noticeable, it helps to enhance performance, as it can also be seen in figures 6.4 and 6.5. So much so, that if the previous experiment table is observed 6.4, it can be noted that on average, training attention weights and applying them on the last convolutional block, provides the best reults among every experiment tried.

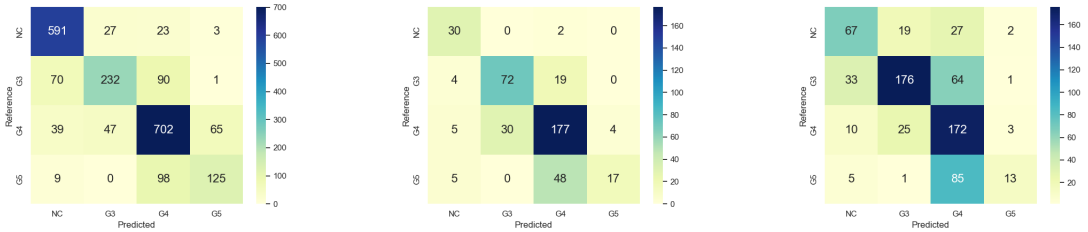
Table 6.5: Metrics for the Attention experiments

Method	ACC	F1S				k2
		NC	GG3	GG4	GG5	Avg	
				SICAP			
Cost Sensitive Loss	0.78	0.87	0.66	0.80	0.59	0.77	**0.83**
Gated Attention 2nd block	0.72	0.86	0.56	0.75	0.20	0.69	0.78
Gated Attentiom 4th block	0.77	0.87	0.66	0.78	**0.52**	0.76	0.80
				GERTYCH			
Cost Sensitive Loss	0.72	0.79	0.73	0.77	0.37	0.69	**0.66**
Gated Attention 2nd block	0.62	0.79	0.67	0.65	0.08	0.57	0.59
Gated Attention 4th block	0.68	0.66	0.70	0.72	**0.42**	0.66	**0**.61
				ARVANITI			
Cost Sensitive Loss	0.61	0.58	0.71	0.62	0.21	0.59	0.57
Gated Attention 2nd block	0.62	0.57	0.75	0.61	0.06	0.58	0.59
Gated Attention 4th block	0.65	0.64	0.73	0.64	**0.35**	0.63	**0.66**
				AVERAGE			
Cost Sensitive Loss	0.70	0.75	0.70	0.73	0.39	0.68	0.69
Gated Attention 2nd block	0.65	0.74	0.66	0.67	0.11	0.61	0.65
Gated Attention 4th block	0.70	0.72	0.70	0.71	**0.43**	0.68	**0.69**



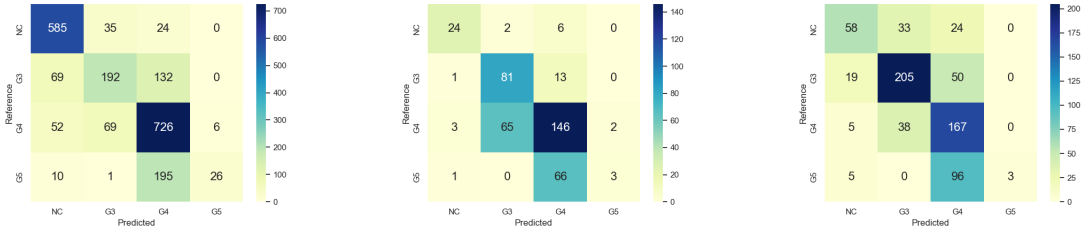
(a) SICAP **(b) GERTYCH** **(c) ARVANITI**

Figure 6.2: Confusion matrices for the Kappa Loss experiments ($\alpha = 10$) in the three different databases



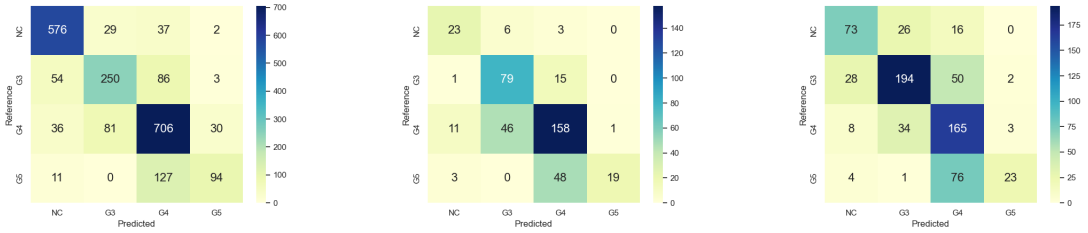
(a) SICAP **(b) GERTYCH** **(c) ARVANITI**

Figure 6.3: Confusion matrices for Cost Sensitive experiments ($\alpha = 10$) in the three different databases



(a) SICAP **(b) GERTYCH** **(c) ARVANITI**

Figure 6.4: Confusion matrices for the second block attention experiments in the three different databases



(a) SICAP **(b) GERTYCH** **(c) ARVANITI**

Figure 6.5: Confusion matrices for the fourth block attention experiments in the three different databases

Chapter 7

Conclusions and future work

As explained in various sections, the main objective of this work was to incorporate uncertainty and variability among pathologists in a CNN model that analyzes patches of prostate biopsies. Along these lines and according to both the general aim, the individual objectives and the results explained in the previous section 6, we can draw several conclusions.

In the first place, the task of annotating samples and carrying out a diagnostic process is very time consuming and laborious in all cases, but especially in prostate cancer. That said, developing a DL model that is capable of providing competent results both in the training databases and in external databases can be very helpful, and not only relieves pathologists workload but is also crucial when determining a patient's treatment.

Going back to the principal point of this paper, it has been possible to demonstrate by means of the experiments carried out in this project, that taking into account inter-observer variability improves the results in external bases, and therefore improves generalization. As we have seen, taking into account this difference of opinion in the training database, does not imply a real improvement in testing, since having trained on that data set, the algorithm already infers the diagnostic patterns. However, when testing in Arvaniti and Gertych databases, very significant improvements have been observed, boosting the main figures of merit up to 15 percent, an outcome that in external databases is extremely encouraging.

Kappa Loss function does no take into account the specific variability of prostate cancer problem since it tries to remove the random component of two pathologists having the same opinion. Therefore, as expected, the results are much less strong than those provided by the Cost Sensitive Loss, which does take this fact into account by introducing the matrix that describes the interobserver agreement of biopsy grading using the Gleason scale. As it has been observed, very good results are obtained both in SICAP and in external databases, also considerably improving its score in the most problematic grades in the diagnosis stage.

Through the preliminary introduction of attention layers, it has been shown that the recognition of low-level patterns has improved. This confirms our initial hypothesis, since adding intermediate features has enhanced the classification of grade 5 patches, which were characterized by patterns that, despite being low-level, they were difficult for the algorithm to classify. By introducing this layer and concatenating the output probabilities with the dense layer, attention is drawn to these features.

Finally, in this section we will try to outline future lines of research. In this paper, a neural network has been improved incrementally through various experiments. Along these lines, future work routes could be centered around the conclusions drawn in this paper. That is to say, by delving into deep learning models that take into account low-level patterns and how they can be incorporated to obtain a more complete result, but above all investigating algorithms that include the variability between pathologists, since more than promising results have been achieved.

Bibliography

[1] Hamid R. Tizhoosh et al. “Searching Images for Consensus: Can AI Remove Observer Variability in Pathology?” In: *American Journal of Pathology* 191 (10 Oct. 2021), pp. 1702–1708. ISSN: 15252191. DOI: 10.1016/J.AJPATH.2021.01.015. URL: https://doi.org/10.1016/j.ajpath.2021.01.015.

[2] Nanne van Noord and Eric Postma. “Learning scale-variant and scale-invariant features for deep image classification”. In: *Pattern Recognition* 61 (Feb. 2016), pp. 583–592. ISSN: 00313203. DOI: 10.1016/j.patcog.2016.06.005. URL: https://arxiv.org/abs/1602.01255v2.

[3] Eirini Arvaniti et al. “Author Correction: Automated Gleason grading of prostate cancer tissue microarrays via deep learning (Scientific Reports, (2018), 8, 1, (12054), 10.1038/s41598-018-30535-1)”. In: *Scientific Reports* 9 (1 Dec. 2019). DOI: 10.1038/S41598-019-43989-8.

[4] Julio Silva-Rodríguez et al. “Going deeper through the Gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection”. In: *Computer Methods and Programs in Biomedicine* 195 (Oct. 2020), p. 105637. ISSN: 18727565. DOI: 10.1016/J.CMPB.2020.105637. URL: www.elsevier.com/locate/cmpb.

[5] Adrian Galdran et al. “Cost-Sensitive Regularization for Diabetic Retinopathy Grading from Eye Fundus Images”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12265 LNCS (2020), pp. 665–674. DOI: 10.1007/978-3-030-59722-1_64.

[6] International Agency for Research on Cancer. “Prostate Source: Globocan 2020 Number of new cases in 2020, both sexes, all ages”. In: (2020). URL: https://gco.iarc.fr/today.

[7] Seyed Hossein Hassanpour and Mohammadamin Dehghani. “Review of cancer from perspective of molecular”. In: *Journal of Cancer Research and Practice* 4 (4 Dec. 2017), pp. 127–129. ISSN: 23113006. DOI: 10.1016/J.JCRPR.2017.07.001. URL: http://dx.doi.org/10.1016/j.jcrpr.2017.07.001.

[8] Hyuna Sung et al. “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries”. In: *CA: a cancer journal for clinicians* 71 (3 May 2021), pp. 209–249. ISSN: 1542-4863. DOI: 10.3322/CAAC.21660. URL: https://pubmed.ncbi.nlm.nih.gov/33538338/.

[9] Sociedad Española de Oncología Médica. *El cáncer en cifras - SEOM: Sociedad Española de Oncología Médica*. 2021. ISBN: 9788409380299.

[10] *Cáncer de Próstata - SEOM: Sociedad Española de Oncología Médica © 2019*. URL: https://seom.org/info-sobre-el-cancer/prostata.

[11] B. Alcalá et al. *Libro Blanco de la Carga Socioeconómica del Cáncer de Próstata en España*. Fundación Weber. ISBN: 9788494770371.

[12] Henrik Grönberg. "Prostate cancer epidemiology". In: *The Lancet* 361 (9360 Mar. 2003), pp. 859–864. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(03)12713-4.

[13] Joan Morote, Xavier Maldonado, and Rafael Morales-Bárrera. "Cáncer de próstata". In: *Medicina Clínica* 146.3 (2016), pp. 121–127. ISSN: 0025-7753. DOI: https://doi.org/10.1016/j.medcli.2014.12.021. URL: https://www.sciencedirect.com/science/article/pii/S002577531500041X.

[14] *histology | physiology | Britannica*. URL: https://www.britannica.com/science/histology.

[15] *GENERALIDADES DE LAS TÉCNICAS UTILIZADAS EN HISTOLOGÍA*. Editorial Médica Panamericana, 2007.

[16] John D Bancroft and Christopher Layton. "The hematoxylins and eosin". In: *Bancroft's theory and practice of histological techniques* (2012), pp. 173–186.

[17] Peter. A. Humphrey. "Gleason grading and prognostic factors in carcinoma of the prostate." In: *Modern pathology : an official journal of the United States and Canadian Academy of Pathology* 17(3), 292–306 (Mar. 2004). DOI: 10.1038/modpathol.3800054.

[18] Jennifer Gordetsky and Jonathan Epstein. "Grading of prostatic adenocarcinoma: current state and prognostic implications". In: *Diagnostic Pathology* 11 (1 Mar. 2016). DOI: 10.1186/S13000-016-0478-2. URL: /pmc/articles/PMC4784293/%20/pmc/articles/PMC4784293/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4784293/.

[19] A M Turing. "COMPUTING MACHINERY AND INTELLIGENCE". In: *Computing Machinery and Intelligence. Mind* 49 (1950), pp. 433–460.

[20] Gheorghe Tecuci. "Artificial intelligence". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4 (2 Dec. 2011), pp. 168–180. ISSN: 19395108. DOI: 10.1002/WICS.200. URL: https://dl.acm.org/doi/abs/10.1002/wics.200.

[21] A. N. Ramesh et al. "Artificial intelligence in medicine." In: *Annals of the Royal College of Surgeons of England* 86 (5 Sept. 2004), p. 334. DOI: 10.1308/147870804290. URL: /pmc/articles/PMC1964229/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1964229/.

[22] "Definition of artificial neural networks with comparison to other networks". In: *Procedia Computer Science* 3 (Jan. 2011), pp. 426–433. ISSN: 1877-0509. DOI: 10.1016/J.PROCS.2010.12.071.

[23] Simon Haykin. *Neural Networks and Learning Machines Third Edition*. Prentice Hall, 2009. ISBN: 9780131471399.

[24] "Artificial neural networks: A tutorial". In: *Computer* 29 (3 Mar. 1996), pp. 31–44. DOI: 10.1109/2.485891.

[25] Deborah Walters. "Computer Vision". In: GBR: John Wiley and Sons Ltd., 2003, pp. 431–435. ISBN: 0470864125.

[26] Athanasios Voulodimos et al. "Deep Learning for Computer Vision: A Brief Review". In: *Computational Intelligence and Neuroscience* 2018 (2018). ISSN: 16875273. DOI: 10.1155/2018/7068349.

[27] Laith Alzubaidi et al. “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions”. In: *Journal of Big Data 2021 8:1* 8 (1 Mar. 2021), pp. 1–74. ISSN: 2196-1115. DOI: 10.1186/S40537-021-00444-8. URL: https://link.springer.com/articles/10.1186/s40537-021-00444-8.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60 (6 June 2017), pp. 84–90. ISSN: 15577317. DOI: 10.1145/3065386.

[29] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (Sept. 2014). URL: https://arxiv.org/abs/1409.1556v6.

[30] Christian Szegedy et al. “Going Deeper with Convolutions”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 07-12-June-2015 (Sept. 2014), pp. 1–9. ISSN: 10636919. DOI: 10.1109/CVPR.2015.7298594. URL: https://arxiv.org/abs/1409.4842v1.

[31] Navid Farahani, Anil V Parwani, and Liron Pantanowitz. “Whole slide imaging in pathology: advantages, limitations, and emerging perspectives”. In: *Pathology and Laboratory Medicine International* 7 (June 2015), pp. 23–33. DOI: 10.2147/PLMI.S59826. URL: https://www.dovepress.com/whole-slide-imaging-in-pathology-advantages-limitations-and-emerging-p-peer-reviewed-fulltext-article-PLMI.

[32] Ángel E. Esteban et al. “A new optical density granulometry-based descriptor for the classification of prostate histological images using shallow and deep Gaussian processes”. In: *Computer Methods and Programs in Biomedicine* 178 (Sept. 2019), pp. 303–317. DOI: 10.1016/J.CMPB.2019.07.003. URL: https://doi.org/10.1016/j.cmpb.2019.07.003.

[33] Julio Silva-Rodriguez et al. “Self-Learning for Weakly Supervised Gleason Grading of Local Patterns”. In: *IEEE Journal of Biomedical and Health Informatics* 25 (8 Aug. 2021), pp. 3094–3104. DOI: 10.1109/JBHI.2021.3061457.

[34] Abhishek Vahadane et al. “Structure-preserved color normalization for histological images”. In: *Proceedings - International Symposium on Biomedical Imaging* 2015-July (July 2015), pp. 1012–1015. DOI: 10.1109/ISBI.2015.7164042.

[35] Jordi de la Torre, Domenec Puig, and Aida Valls. “Weighted kappa loss function for multi-class classification of ordinal data in deep learning”. In: *Pattern Recognition Letters* 105 (2018), pp. 144–154. ISSN: 01678655. DOI: 10.1016/j.patrec.2017.05.018. URL: https://doi.org/10.1016/j.patrec.2017.05.018.

[36] Saumya Jetley et al. *LEARN TO PAY ATTENTION*. Tech. rep. arXiv: 1804.02391v2.

[37] Jo Schlemper et al. “Attention gated networks: Learning to leverage salient regions in medical images | Elsevier Enhanced Reader”. In: *Elsevier B.V* (2019).

[38] Yann N Dauphin et al. *Language Modeling with Gated Convolutional Networks*. 2017.

[39] Pierre Buyssens and Abderrahim Elmoataz. *Multiscale Convolutional Neural Networks for Vision–Based Classification of Cells*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 342–352. ISBN: 978-3-642-37444-9. DOI: 10.1007/978-3-642-37444-9_27.

[40] Guanbin Li and Yizhou Yu. “Visual saliency based on multiscale deep features”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 07-12-June-2015 (Oct. 2015), pp. 5455–5463. ISSN: 10636919. DOI: 10.1109/CVPR.2015.7299184.

[41] Nathalie Japkowicz. “Why Question Machine Learning Evaluation Methods? (An illustrative review of the shortcomings of current methods)”. In: (2006).