PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
ESCUELA DE POSGRADO

Integrative Digital Pathology for Personalized Medicine: Population stratification and early biomarkers findings, consolidating and completing the use of Prostate-Specific Antigen (PSA) and Gleason Score in Prostate Cancer

Tesis para optar el grado académico de Doctora en Ingeniería que presenta:

Laura Elise Marin

Asesora:

Fanny Lys Casado Peña

Lima, 2024

Informe de Similitud

Yo, Fanny Lys Casado Peña, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesora de la tesis titulada: Integrative Digital Pathology for Personalized Medicine: Population stratification andearly biomarkers findings, consolidating and completing the use of Prostate-SpecificAntigen (PSA) and Gleason Score in Prostate Cancer., de la autora Laura Elise Marin, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 9%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 24 de enero del 2025.
- He revisado con detalle dicho reporte y la Tesis, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha:

Lima, 24 de enero del 2025.

Apellidos y nombres de la asesora: Casado Peña, Fanny Lys	
DNI: 40444557	Firma
ORCID: https://orcid.org/0000-0002-8791-626X	

Resumen

El cáncer de próstata, aunque es el segundo cáncer más común en los hombres, no tiene biomarcadores establecidos para predecir el riesgo de recaída y el riesgo de presentar recurrencia bioquímica. Una comprensión más profunda del comportamiento de los tejidos proporcionada por técnicas moleculares puede mejorar la capacidad de pronosticar la probabilidad de recurrencia. Basándose en datos sólidos y correctamente anotados disponibles de grandes cohortes internacionales de pacientes, y en el procesamiento exhaustivo de datos de información fenotípica y genómica, este trabajo propuso y evaluó el papel de los biomarcadores tempranos de recurrencia. Además, se incluyó en el análisis información clínica asociada a datos estructurales y moleculares del tejido para proporcionar una comprensión más profunda del microentorno del cáncer de próstata. Por lo tanto, se entrenaron modelos de aprendizaje profundo para segmentar características morfológicas de imágenes de diapositivas completas, descargadas de repositorios disponibles públicamente. Las características segmentadas estaban asociadas a la proliferación celular, la estructura de la luz y la arquitectura de la región tumoral. A continuación, se predijo el riesgo de presentar recurrencia se predijo entonces mediante algoritmos de aprendizaje automático a partir de las características tisulares mencionadas, y se analizó el papel de la puntuación de Gleason. Al mismo tiempo, se introdujeron en los modelos niveles de expresión genómica pre-procesados para recuperar un subconjunto de genes responsables de la recurrencia. Los resultados indican que, tras la inspección de los biomarcadores, la organización de la matriz extracelular se ha asociado con el riesgo de presentar recurrencia. Además, se establecieron los niveles de PSA como información crítica a la hora de detectar la recurrencia. Los algoritmos de aprendizaje automático entrenados en el genoma y el fenotipo clasificaron a los pacientes con una precisión media del 79 % y el 69,7 %, respectivamente, cuando la recurrencia bioquímica se produjo hasta 22 meses después de su tratamiento final, lo que demuestra que el riesgo de presentar recurrencia bioquímica puede predecirse con éxito cuando se integra la información clínica, fenotípica y genómica.

Abstract

Prostate cancer, although the second most common cancer in men, does not have established biomarkers to predict the risk of relapse and risk of presenting biochemical recurrence. A deeper comprehension of tissue behavior provided by molecular techniques may enhance the ability to forecast the likelihood of recurrence. Based on robust and highly annotated data available from large international cohorts of patients, and extensive data processing of phenotypic and genomic information, this work proposed and evaluated the role of early biomarkers of recurrence. Furthermore, clinical information associated with structural and molecular data from the tissue was included in the analysis to provide a deeper understanding of the prostate cancer microenvironment.

Deep learning models were hence trained to segment morphological features from Whole Slide Images, downloaded from publicly available repositories. Features segmented were associated with cell proliferation, lumen structure, and tumorous region architecture. The risk of presenting recurrence was then predicted via machine learning algorithms from the aforementioned tissue features, and the role of the Gleason score was analyzed. Concurrently, pre-processed genomic levels of expression were fed to models to retrieve a subset of genes responsible for the recurrence.

The results indicate that after inspection of biomarkers, extracellular matrix organization has been associated with the risk of presenting recurrence. In addition, PSA levels were established as critical information when detecting recurrence. Machine learning algorithms trained on the genome and phenotype classified patients with an average precision of 79% and 69.7%, respectively, when the biochemical recurrence occurred up to 22 months after its final treatment, providing evidence that the risk of presenting biochemical recurrence can be successfully predicted when clinical, phenotypic and genomic information are integrated.

Acknowledgments

This work was partially supported by internal funds from Pontificia Universidad Catolica del Peru, which were assigned to support Dr. Daniel Racoceanu's academic involvement.

I would like to thank my thesis advisor, Dr. Fanny Casado, from the Institute for Omic Sciences and Applied Biotechnology at Pontificia Universidad Catolica del Peru, for her unwavering support in completing this study.

Finally, I am also grateful to my entourage, Julie Mezerette, and Julio Gallegos, for their optimism and reassurance during these past years.

Contents

List of Figures

List of Tables

Introduction

Prostate cancer is estimated to be the second most common cancer in men after lung cancer according to worldwide data from 2020 with 1'414,259 incident cases (*Global Cancer Observatory*, n.d.). The American Cancer Society estimated a yearly detection increase of 6% (*National Cancer Institute*, n.d.-a) and 7% in overall mortality attributed to patients diagnosed with the disease. Nowadays, health professionals mainly employ the Gleason classification as a score to catalog prostate cancer by evaluating tissue biopsies, and their respective structures (*American Cancer Society*, n.d.) in order to determine the best therapeutic approach. However, despite completing treatment and being declared in cancer remission, signs of biochemical recurrence, BCR, may appear months or years after the final treatment. Interest in genomic surveillance tools has increased over the years to predict the risk of presenting signs of BCR and forecasting the time of recurrence. The Gleason score alone is ineffective to provide a prediction, as effectuated belatedly when the signs are highly obvious and there is a possible advanced and aggressive tumor or spread metastasis. Whereas the genomic expression could provide crucial information as to the future risk of BCR before the first signs of recurrence, hence increasing the survival of patients, as early treatment might be administrated. Novel molecular classifications of prostate cancer have been proposed to predict the risk of BCR (Long, Johnson, & Osunkoya, 2011)(Verma & Patel, 2011). But the use of genomic information has yet to be proved in comparison to the accuracy of phenotype BCR prediction. The work presented in this

thesis, explores early biomarkers of recurrence, by associating clinical information as well as molecular and structural data from the tissue to provide a deeper understanding of the prostate cancer microenvironment. Specifically, phenotypic features extracted by Deep Learning approaches are contrasted with genomic data of annotated samples of prostate cancer obtained from The Cancer Genome Atlas, TCGA, a publicly available repository of data from cancer patients.

Chapter I

Clinical and Translational Challenges of Prostate Cancer Research

This chapter presents the different technologies currently used in the clinic, their inherent challenges, and promising strategies to develop faster diagnosis, better treatments, and robust surveillance throughout the lifespan in the context of prostate cancer research.

Prostate cancer screening initiates with preliminary tests including Prostate-Specific Antigen, PSA, as the most common one. PSA is a protein released by prostate tissue, present in the blood. A higher level of PSA denotes an abnormality in the prostate organ, which could be prostate cancer or inflammation of the prostate (*National Cancer Institute*, n.d.-b). The diagnosis of prostate hyperplasia is then confirmed by Digital Rectal Examination, DRE, to find any abnormal growth in the prostate by palpation, transrectal ultrasound, or biopsy. Positive prostate hyperplasia is followed by the examination of prostatic tissue by biopsies and classification according to the Gleason score to determine a diagnosis of prostate cancer or benign prostate hyperplasia.

1.1 Gleason score to grade prostate cancer and describe its phenotypes

The Gleason score, developed by Donald F. Gleason in 1960, relies on the cellular architecture of prostate tissue pathological specimens. The Gleason score classifies the tissue from 1-5 according to its aggressiveness:

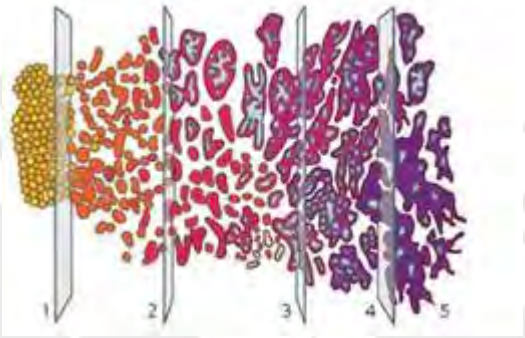


Figure 1.1: Visual description of Gleason's patterns for prostate cancer aggressiveness from 1 to 5. (1) Glands are small and well-formed. Nuclei are organized around the glands. (2) Glands are still well-formed, but larger, with more stroma in the tissue. (3) Recognizable glands, and nuclei start to invade the surrounding stroma. (4) Irregular glands and tumorous nuclei are observed in the stroma. (5) There are no distinguishable glands, nor tumorous nuclei. (*cancer connect*, n.d.)

Prostate cancers have a range of growth patterns. Therefore, the final Gleason score is calculated by adding the score of the two most present patterns. A Gleason score of 7 expresses patterns of either 3+4 or 4+3. However, in the first case, the predominant pattern is 3; while in the second one, the predominant one is 4 (*National Cancer Institute*, n.d.-c).

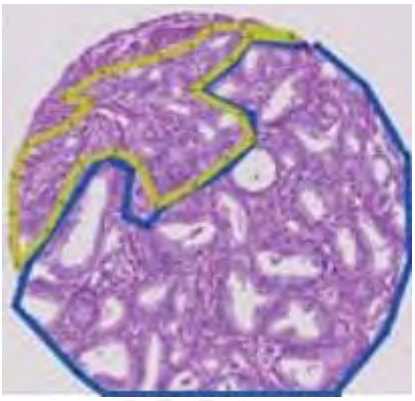
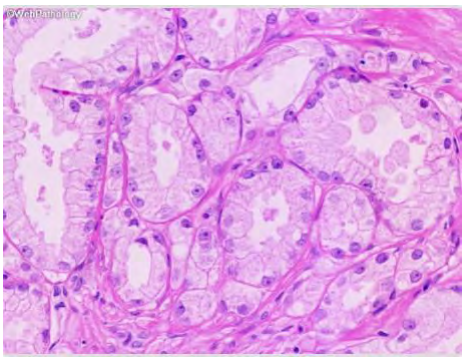
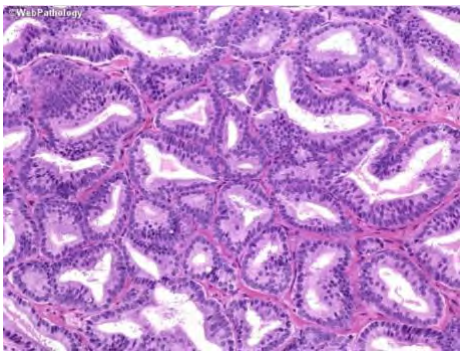


Figure 1.2: Representative pathological slide from tissue biopsy with the annotation of the area of interest to determine the Gleason score. The predominant pattern, enclosed in blue, describes a Gleason score of 3, meanwhile the second one in yellow designates a score of 4. The final Gleason score for the patient will then be 7 (3+4) (Epstein et al., 2005)

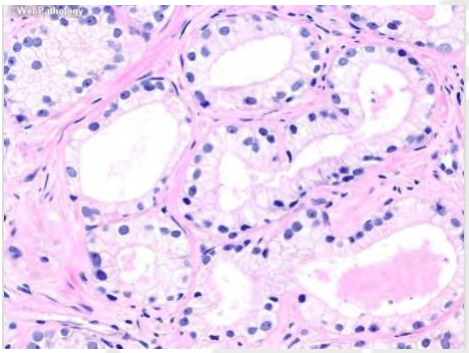
In the following images, multiple combinations of patterns can be appreciated.



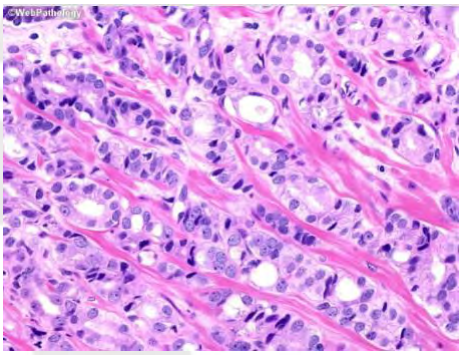
(a) Gleason score 2+2



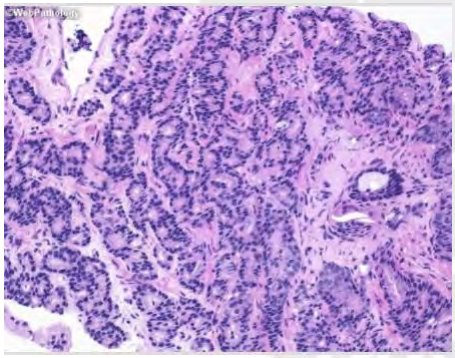
(b) Gleason score 3+3



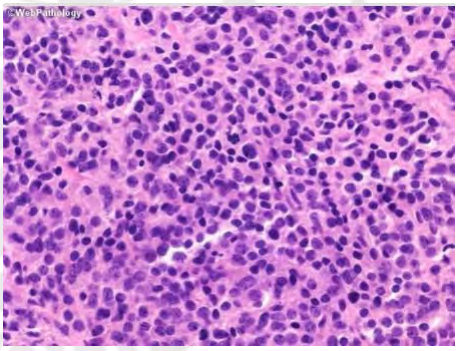
(c) Gleason score 3+3



(d) Gleason score 4+4



(e) Gleason score 4+4



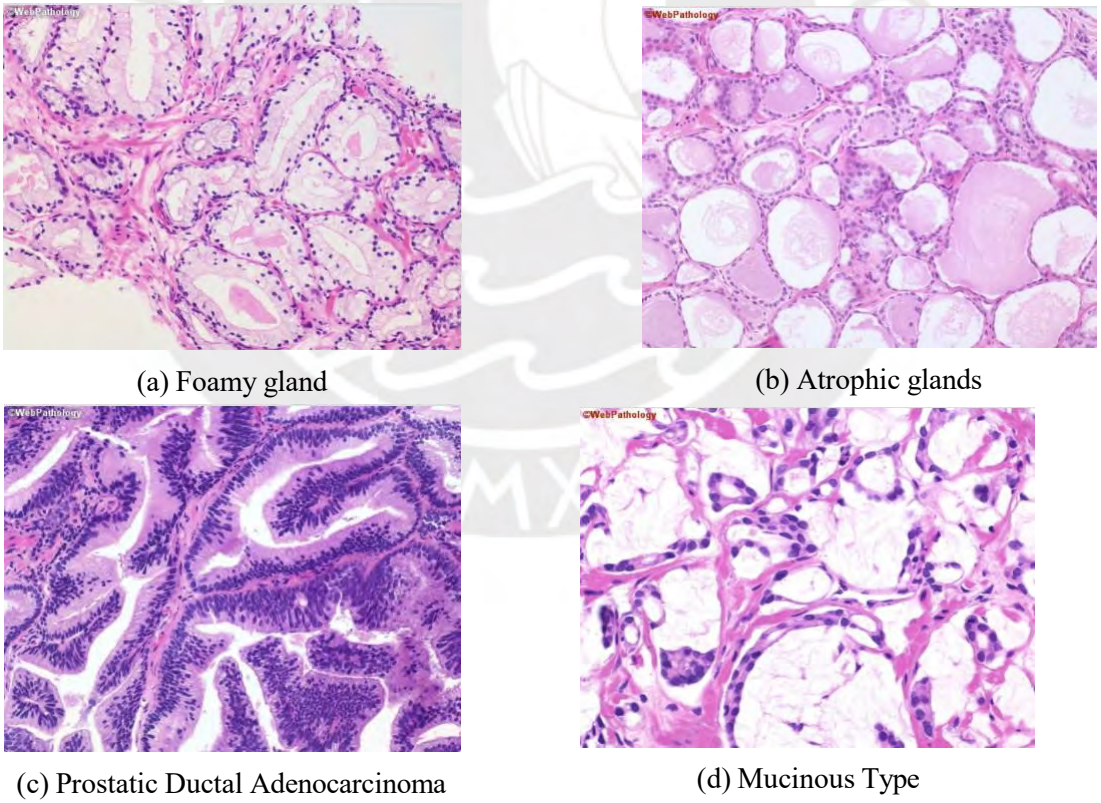
(f) Gleason score 5+5

Figure 1.3: Examples of Gleason grading construction (*web pathology*, n.d.)

In 1.3a, the glands are well organized with prominent nuclei organized around a pale cytoplasm. From the biopsies 1.3b and 1.3c, the nuclei remain organized around the cytoplasm, but a few glands are closer to each other and several fused into one, where no stroma can be found in between the glands. Poorly formed glands can be observed in the image 1.3d, with a

negligible lumen, or occluded one in 1.3e. As stages increase, the organization of the glands decreases until reaching the final stage, the most aggressive one, Gleason 5. No discernible glands with a solid nest of tumor cells are observed, as illustrated in 1.3f.

Prostate cancer phenotypes can provide information of great importance to define the spread and aggressiveness of the tumor by further dissecting the structural characteristics observed. In the previous figure 1.3, different phenotypes are shown from tumors with the same Gleason score. The different types of phenotype are displayed in 1.3b and 1.3c for a tumor with a Gleason score of 6 (3+3). For 1.3b, the cells are arranged in ductal and papillary structures, with abundant cytoplasm and enlarged cells; while in 1.3c, the cytoplasm has a foamy appearance surrounded by small-scaled nuclei. The most frequent phenotype is illustrated below in 1.4 where the foamy type and ductal adenocarcinoma type from 1.3b and 1.3c are magnified.



(a) Foamy gland (b) Atrophic glands

(c) Prostatic Ductal Adenocarcinoma (d) Mucinous Type

Figure 1.4: Most frequent prostate phenotypes (*web pathology*, n.d.)

A more aggressive phenotype is displayed below in figure 1.5, with a Gleason 8 (4+4) scored biopsy.

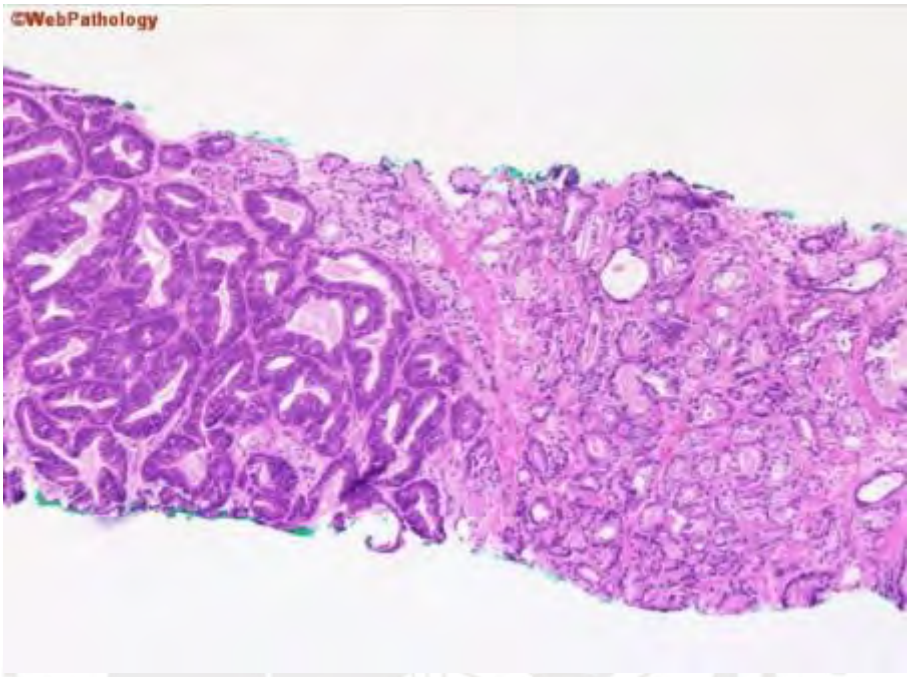


Figure 1.5: Example of biopsy from same Gleason score with different phenotype (*web pathology*, n.d.)

In addition, a signet ring cell type, exhibited in 1.6 illustrates generally a high Gleason where neither glands nor lumen can be outlined.

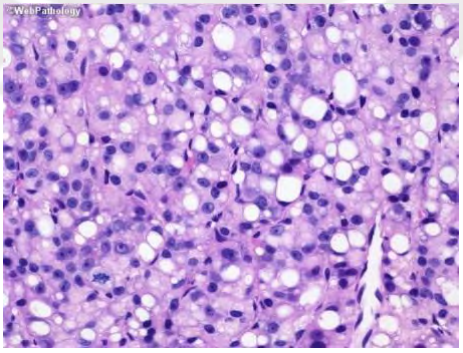


Figure 1.6: Representative figure for a Signet Ring Cell Type structure (*web pathology*, n.d.)

The structure of the glands is consequently decisive in the classification of prostate can-

cer. Communication between glands is mediated by cell adhesion proteins that regulate tissue patterns and establish gland architecture. The extracellular matrix, ECM, is composed of cell adhesion molecules secreted by the cells to provide structural support and tissue organization. As stated above, glands are a layer of cells around a lumen, separated by stroma composed of fibroblasts which secrete mostly collagen type I and III, forming a matrix of tissue embedding the glandular structure. A tumorous ECM tends to be stiffer than healthy tissue, boosting cancer cell proliferation (Nallanthighal, Heiserman, & Cheon, 2019).

At a higher Gleason, many ECM components are underexpressed, proscribing glandular architecture, because collagen is not secreted to maintain tissue structure. Besides collagen, further proteins are part of the ECM, such as Glycosaminoglycans, GAG, known for their roles in cell signaling, Proteoglycans acknowledged to provide hydration to the tissue, Glycoprotein which can act as a molecule receptor, or Elastin which provides elasticity to the ECM. ECM protein expression is critical in cancer progression, as can be witnessed in the glandular architecture.

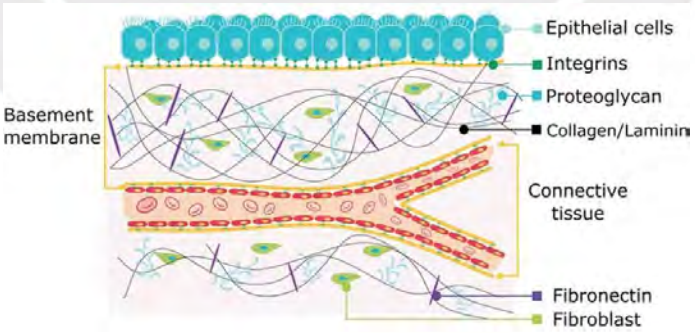


Figure 1.7: Diagram of ECM proteins providing structural support and biochemical reaction sites to tissues (*Merk*, n.d.)

Cells will adhere to ECM and to each other via cell adhesion molecules or adhesion receptors. Nevertheless, adhesion receptors will also contribute to cell signaling. Indeed, adhesion

molecules can communicate with other cells to provide spatial information as a means to assist with cell proliferation, migration, and survival. Adhesion receptors, for example, might transduce the signal of proliferation when tumorous cells are detected as an immune response, or on the contrary cease proliferation. Cell signaling is, as a matter of fact, deregulated when dealing with tumorous cells, hence the presence of clusters of nuclei in a high Gleason score images. No signal to cease cell proliferation will then exist, invading stroma and glands (Moh & Shen, 2019).

1.2 Staging of prostate cancer

To make decisions regarding the course of treatment of patients diagnosed with prostate cancer, it is primordial to refer to the stage of cancer. The stage of the cancer mainly depends on the organization of the malignant cells. In the early stage, prostate cancer does not need immediate treatment, instead, the patient will be kept under what is referred to as active surveillance (Morash et al., 2015). Prostate cancer cells can break away from the tumor, and proliferate through the body, through a process known as metastasis. Some can even reach the bone, in this case, we won't talk about bone cancer but metastatic prostate cancer because it originated in that tissue.

Staging tests are done only if prostate cancer has been diagnosed with a high PSA level and high Gleason score, or if the patient presents pain in the bones. These tests can consist of seminal vesicle biopsy, where fluid from the seminal vesicles is removed and then examined under a microscope by a pathologist. A way to inspect the presence of metastases in the lymph nodes is by removing them surgically in the pelvis, then scrutinizing them. Other external method consists of Magnetic Resonance Imaging, MRI, and Computed Tomography Scan, CTS, to extract detailed pictures of areas inside the body taken from different angles. In some cases, a

small amount of radioactive material can be injected into the bloodstream of the patient and then scanned, hence the material will inform where there are prostate cancer cells, therefore appearing bright on the scan.

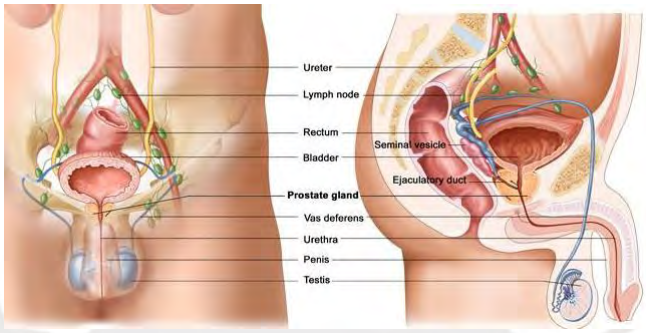


Figure 1.8: Prostate anatomy (*chestnutappeal*, n.d.)

1.2.1 Stage I

Stage I cancers, Gleason score of 6 or less, are exclusively found in the prostate. They usually spread slowly and do not always represent a risk for the patient, however according to the age of the patient, or his medical history, it can represent a danger to his lifespan. In this case, different treatments can be applied, such as radiation therapy or radical prostatectomy (Morash et al., 2015). Treatments available for Stage I prostate cancer are radiation therapy or radical prostatectomy. Both only focus on a specific part of the body, they are local treatments and have to be used according to the size of the tumor, how close the tumor is to sensitive organs, and the location of the tumor.

Radiation therapy consists of using radiation to damage the DNA of cancer cells, slowing down their growth. As the name specifies, external radiation comes from a machine that aims at the cancer. Concurrently, in the case of internal therapy, the radiation is inside the patient's body in the form of seeds, capsules, or ribbons, the closest to the tumor (Gay & Michalski,

2018). Radiation therapy can also be merged with other cancer treatments including surgery and chemotherapy.

Radical prostatectomy consists of removing the part of the prostate gland affected by the cancer cell. It provides the maximum benefits to patients with non-spread cancer. It can be done either internally, accessing the prostate gland from an incision below the belly button, or externally through several small incisions where tools and cameras are inserted, and the radical prostatectomy is performed from outside the body. Radical prostatectomy can be merged with radiation therapy to reduce the size of the cancer before surgery, during the surgery, or after killing any remaining cells (Kesch et al., 2021).

1.2.2 Stage II

Similar to stage I cancer, stage II cancer has not yet grown outside the prostate but it shows a higher Gleason score. If not treated correctly, they are more likely to spread to the surrounding organs, as prostate cancer can advance into the bladder and more frequently into the lymph nodes. Radiation therapy and radical prostatectomy can be used as a treatment, similar to Stage I. In addition, hormone therapy can also be introduced to treat stage II cancer (Morash et al., 2015).

Hormone therapy aims to stop the body from producing testosterone, or by stopping testosterone from reaching the prostate cancer cells. It does not intend to cure the tumor but can control the tumor size and delay some symptoms when complemented with another treatment. It can be used in advance to reduce cancer and slow down its growth, and might release the patients of some symptoms; or after treatment in case the PSA level remains high. Testosterone is a hormone controlling male characteristics such as muscle strength, and growth of the penis

and the testicle. Testosterone is essential for male physiology, but in the case of prostate cancer, it controls the growth of the tumor.

Hormone therapy includes different approaches. Medication to block endogenous production of testosterone, via the luteinizing hormone-releasing hormone, LHRH, or the gonadotropin-releasing hormone, GnRH. They prevent the body's cells from receiving the message to produce testosterone. These medications include leuprolide, goserelin, histrelin and are usually injected under the skin or into a muscle either monthly or every six months, they can also be placed under the skin as an implant, delivering medication over a long period of time. Other medications classified as anti-androgens, taken via oral, block directly the testosterone from reaching the tumor, and so controlling its size. They are often given in association with LHRH since the intake of LHRH first increases the level of testosterone before decreasing it (Gomella, Singh, Lallas, & Trabulsi, 2010). A third approach is the removal of the glands producing testosterone, the testicles; then, testosterone levels will drop quickly, but the procedure is irreversible and may greatly alter the quality of life of the patient.

1.2.3 Stage III

Stage III cancers have grown outside the prostate to surrounding organs such as the bladder or rectum, they also show higher Gleason scores. All the medication cited above can be used and mixed for higher efficiency (Morash et al., 2015). An additional surgical option available for Stage III prostate cancer patients is transurethral resection of the prostate, TURP, which is a surgery aiming to remove part of the prostate through the penis. It does not cure the cancer but is used to relieve the patient of symptoms. As cancer grows, the enlargement of the prostate gland can cause pain and difficulty to urinate, which might indispose the patient. The surgeon

will reach the prostate, using a resectoscope, a metal tube of about 30 cm, bearing wire, light, and a camera guiding it through the urethra to the prostate gland. Once reached, an electric current is used to heat the wires and cut the chosen section of the prostate, fluid will be then pumped into the bladder to flush the removed pieces away (Leslie, Chargui, & Stormont, 2022).

1.2.4 Stage IV

Stage IV cancer can not be cured as they had spread to the lymph node, and can even have reached the bones. Nevertheless, they are treatable to keep cancer under control and improve the quality of life of the patient (Morash et al., 2015).

Chemotherapy in the case of prostate cancer is often used when hormone therapy is no longer controlling the growth of cancer. It is commonly used to slow down cancer growth and increase the quality of life of the patient, but it's very unlikely to cure prostate cancer. Contrary to hormone therapy, chemotherapy employs drugs to stop the growth of cancer cells such as docetaxel, carboplatin, and oxaliplatin.

Delivered either orally or intravenously, they are given in cycles, with a period of time for the patient to recover from the side effects of the drug. Since the drugs prevent the cell from dividing, cancer cells are not the only ones to divide quickly. Indeed, cells such as those in the bone marrow, the intestines, and hair follicles also divide rapidly, hence mediating the known side-effects of chemotherapy including nausea, hair loss, fatigue, or mouth soreness (Nader, Amm, & Aragon-Ching, 2018).

1.3 Time of recurrence

Patients with prostate cancer are treated accordingly to the stage as previously described which takes into consideration the Gleason score, the aggressiveness, the size of the cancer, age, and

race, among other clinically relevant data. Once the patients are cleared of cancer, they still have to be actively screened. Indeed, prostate cancer cells can become evident again months or years after the last treatment, and the latest presence of cancer cells. Therefore, the concept of time of recurrence or disease-free time allows us to establish the time elapsed between remission and relapse. Around 40% of men will experience a recurrence, even though it doesn't always represent a risk to the life of the patient. Time of recurrence depends on the characteristics of the initial tumor; thus, patients with a higher Gleason score with tumors that have spread to the lymph node, or massive cancers are more disposed to present BCR.

In most cases, recurrence is detected using the same tests to diagnose prostate cancer: high levels of PSA, abnormal findings of DRE, and presence of cancer cells in biopsies. However, the cancer cells do not have to be detected in the prostate tissue to call it a recurrence; metastasis in other organs, including the bladder and bones are included together with the recurrence (*Zero Cancer*, n.d.).

When the relapse is confirmed with molecular data, biochemical recurrence, BCR, is considered a high resolution, well defined and accepted clinical endpoint. Komisarof and colleagues (Komisarof, McCall, & Newman, 2017) developed a prediction tool based on the expression pattern of a small set of genes from patients after radical prostatectomy. He proposed changes in the expression of four genes (HBEGF, HOXC13, IGFBP2, and SATB1) which are capable of differentiating non-recurrent tumors from recurrent ones with an accuracy score of 83%. As for Chu J., he investigated the role of eight genes (CHST1, ACOX1, CTBS, GNPNAT1, NAGLU, LPIN3, ASRGL1, HMGCS2) via Cox regression for a five-year BCR (Chu, Li, & Gai, 2018). Stephenson and colleagues (Stephenson, Smith, & Kattan, 2005) established a method to predict the time of recurrence according to the gene expression differences between non-recurrent and recurrent prostate cancer. The method showed increased results, with an accuracy of 89%. Venkatesan and colleagues (Venkatesan, Mudairu-Dawodu, & Duran, 2021) investigated the

use of MRI associated with PSA level and Gleason score to detect locally recurrent prostate cancer in suspected patients. MRI analysis confirmed that patients with low Gleason score and PSA <0.5 ng/mL did not show recurrence, while 88.9% of the patients with Gleason score above 7 and of PSA >1.5 ng/mL exhibited recurrence.

Sharma et al. (Cheaito, Bahmad, Hadadeh, & et al., 2019) and Cheaito et al. (Sharma & Watabe, 2014) both used epithelial-to-mesenchymal transition, EMT, to assess the time of recurrence of patients using immunofluorescence and H&E in paraffin-embedded tissues. EMT can be witnessed with the number of circulating tumor cells, CTC, which corresponds to cells from a primary tumor site that are now circulating in the blood. Cordon-Cardo et al. also extracted cellular elements from H&E, IHC, and immunofluorescence, however, this work gathered information on the intensity of the epithelial cytoplasm, stromal nuclei, and stroma as well as the area of these exact same features relatively to the tissue area (Cordon-Cardo, Kotsianti, Verbel, & et al, 2007) to obtain a classifier capable of predicting the risk of recurrence in patients following radical prostatectomy with an average concordance index of 0.79.

Spatial information remains a challenge in modern medicine to understand the mechanism of different tumors and to create tools able to detect cancer in its earlier form and decrease overall mortality. Investigation on genetic biomarkers has been successful in the past, highlighting a handful of genes as responsible for the time of recurrence, even when they provide no tissue location information. On the other hand, when employing imaging, most of the studies focus on the nuclei, their location from the primary tumor site, in the blood, and the presence of metastases or infiltrating immune cells. In this study, the ability to predict patients' status and time of recurrence were compared when starting with genomic or phenotypic features, respectively. In addition, the influence of the PSA on the prediction was evaluated.

1.4 Prostate cancer datasets publicly available

An strategy to maximize the research opportunities provided by the large amount of data generated when analyzing clinical samples, is to make these datasets publicly available for other researchers focused on evaluating novel approaches to mine data or test novel analytical tools such as the ones proposed in this thesis. The following sections describe the datasets used and the processing necessary to perform the goals of this work based on data coming from genomic expression levels, and Whole Slide Imaging.

1.4.1 Genomic expression levels

The main training dataset consists of genomic expression levels and Whole Slide Imaging, WSI, from biopsies of 500 patients from the fully annotated database The Genome Cancer Atlas, TGCA, accessible from a public web page `https://portal.gdc.cancer.gov`. From the 500 hundred patients, clinical data are accessible, such as the age, race, as well as, time of recurrence, PSA levels, and assigned Gleason scores. Patients from different backgrounds and with diagnoses different from Gleason 4 to 10 and with diverse times of recurrence from a few months to 5 years are included in the set. In this dataset, 401 patients are disease-free, while 99 have experienced recurrence after treatment. Their respective genome was analyzed based on the expression of around 20 000 genes.

To confirm if any of the established models classify patients according to their labels and their time of recurrence correctly, they require to be tested on data different from the ones where the model was developed. For this exact purpose, National Center for Biotechnology Information, NCBI, possesses a public platform, GEO where a myriad of genetic expression datasets are available. Consisting of 106 patients, the GSE54460 dataset was chosen because its annotation is compatible with the training dataset, and it has expression levels of the same 20 000 genes.

Additional datasets considered in the study can be found through the cbio portal, where the TCGA files were previously downloaded. The dataset from the Memorial Sloan Kettering Cancer Center, MSKCC, published in Cancer Cell 2010, including 151 patients, 36 with BCR, and 115 without recurrence, were also downloaded because they also meet the prerequisites for inclusion in this investigation.

1.4.2 Whole Slide Images

The TCGA database contains files from digital scanning of 500 WSI saved with an SVS extension. The average size for an SVS file is over 1 GB, which presented a strong limitation for analysis, adding to additional challenges from the uncommon SVS extension which can not be displayed by most of the operating systems. Various methodologies have been developed to manipulate SVS files. In the following lines, the approach developed by (*IBM Developer*, n.d.) is described. The data were first scaled down by a factor of 32x in order to shrink the dataset. From 85656x71305 pixels, the scaled-down images of 2676x2228 pixels were then transformed into a more common format, such as PNG or JPG by virtue of the open-slide library.

Figure 1.9: Red pen marked whole slide histological image (*IBM Developer*, n.d.)

Since the slides are from public libraries, some could be marked by colored markers that

would need to be removed to secure a reliable image processing. The data is usually passed through a combination of filters to isolate the tissue and discard the background. The final filtered slides, cleared of any marks, were then ready to be divided into tiles. In this interest, the WSI was split into *n* tiles, the same tiles were then resized. Once sliced, each of the tiles was evaluated according to their tissue percentage and assigned a score. Histograms from each tile were retrieved, a high number of purple or pink pixels is linked to a greater score, meanwhile, dark pixels representing the background are associated with lower-scored tiles. Only high-scored slides were kept and saved for further investigation.

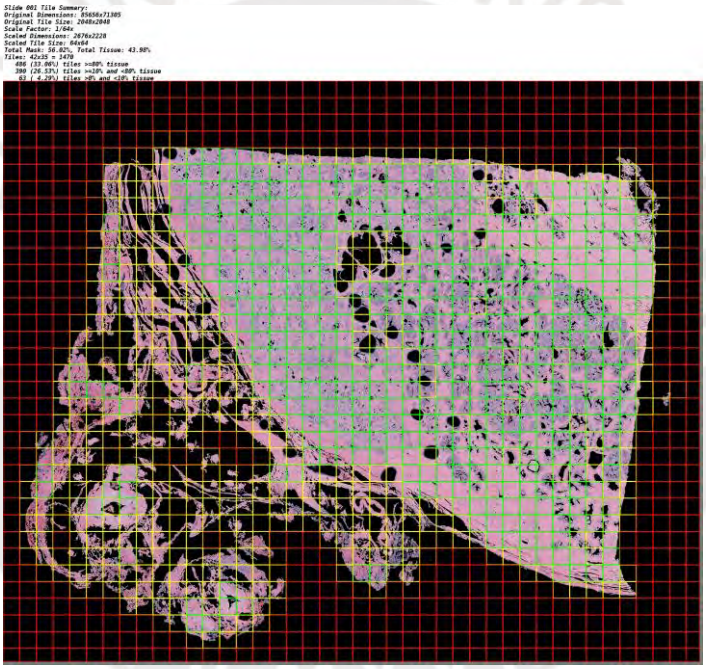
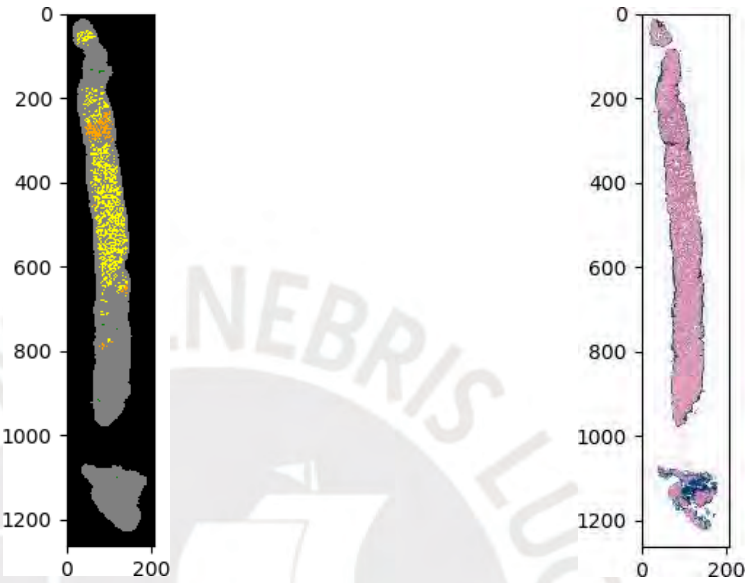


Figure 1.10: Whole slide heat map, tiles with 80% or more of tissue delineated in green, between 10% and 80% of tissues, the tiles are yellow, tiles containing less than 10% color orange is applied, no tissue enhanced in red

Tissue score evaluation

The tissue score was tested with a dataset available on Kaggle, PANDA, consisting of one thousand of WSI and their respective masks. The masks were color-labeled according to their

Gleason score.



(a) Mask from panda set (b) Whole slide image from panda set

Figure 1.11: Whole slide image with its respective mask from panda set

The Gleason score was calculated on the WSI and then superposed onto the mask. This verified if the tiles classified with 80% of tissues correspond to the area labeled with Gleason. The test was done on 40 images, and the optimal number of top tiles was retrieved.

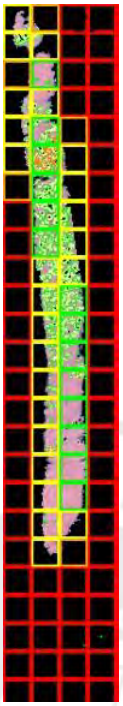


Figure 1.12: Tiles scoring with overlaid whole slide image and mask

The accuracy of the tiles scoring was about 87%, meaning 87% of the area labeled as tumorous was included in the top green tiles by the formula. Nonetheless, some yellow tiles did not contain any area of interest. With the aim to keep only the tiles with tumorous tissue, five tiles were retrieved by WSI with a size of 2048x2048 to maintain most of the glands intact and complete.

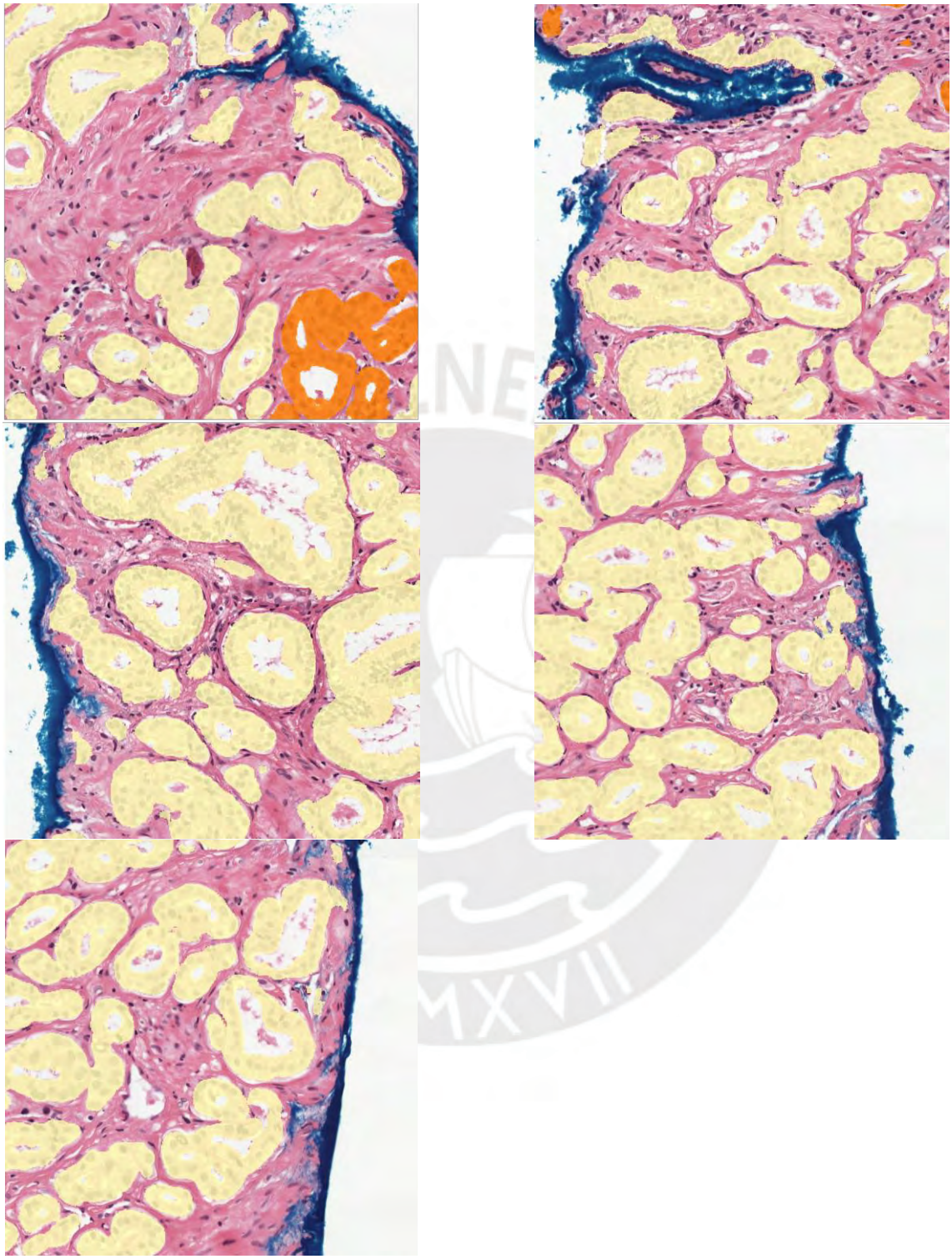


Figure 1.13: Top 5 tiles recovered from tiles scoring, and its overlay masks

Stain normalization

Once the tiles were collected in a PNG extension, they were normalized. In reality, the images were H&E stained, but the staining process or the digital scanning can affect the intensities of the different colors in the image. The unevenness of the stains can affect future segmentation results. Therefore, to avoid any trouble and artifacts, stain normalization is essential.

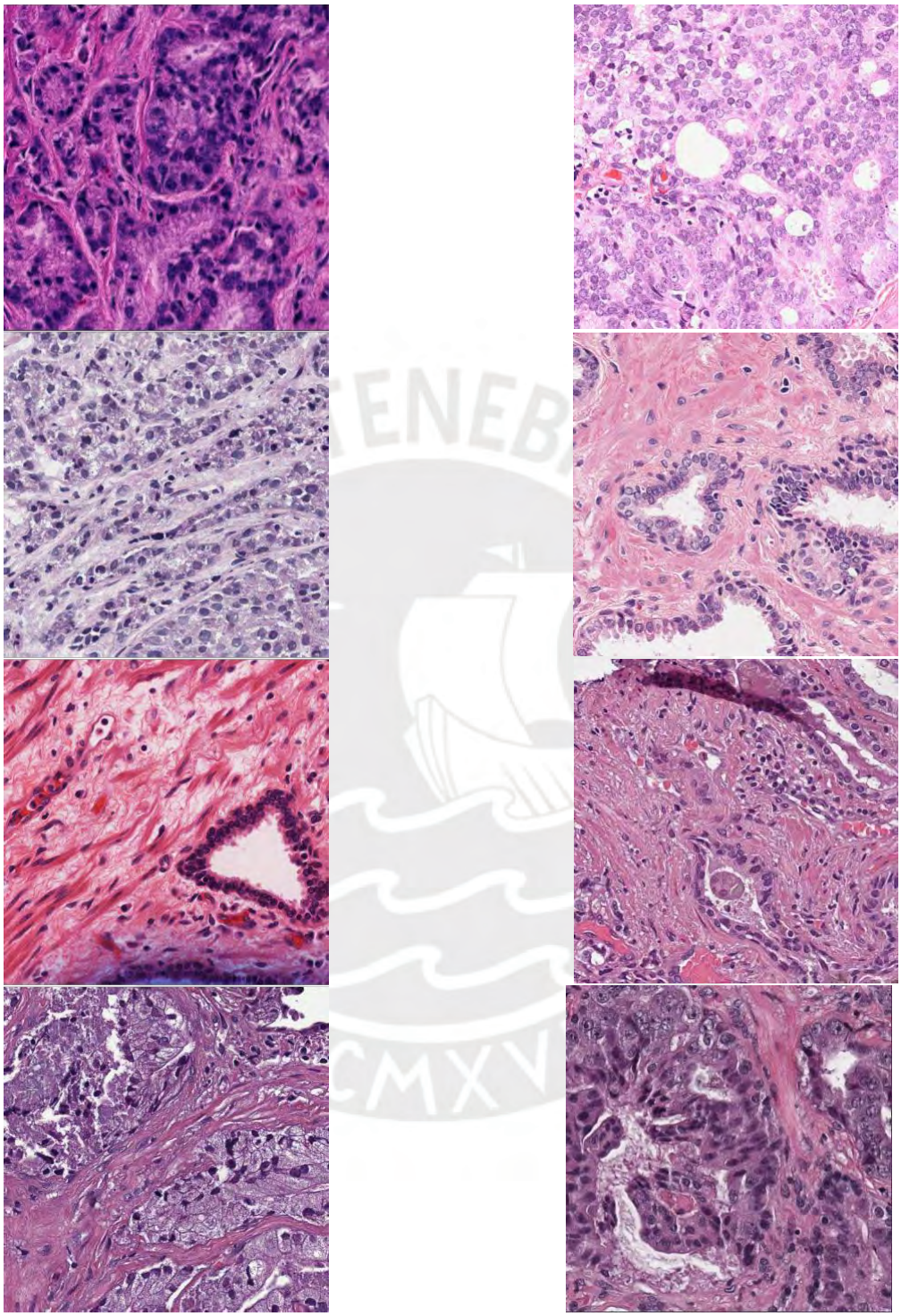


Figure 1.14: Stain disparity in histological images from TCGA dataset

Distinct normalization algorithms can be found in the literature such as Histogram Specification, Macenko Method, Stain Color Descriptor, or Reinhard Method. In most of the approaches, the architecture consists of applying the intensity of the colors from a source image to the target one.

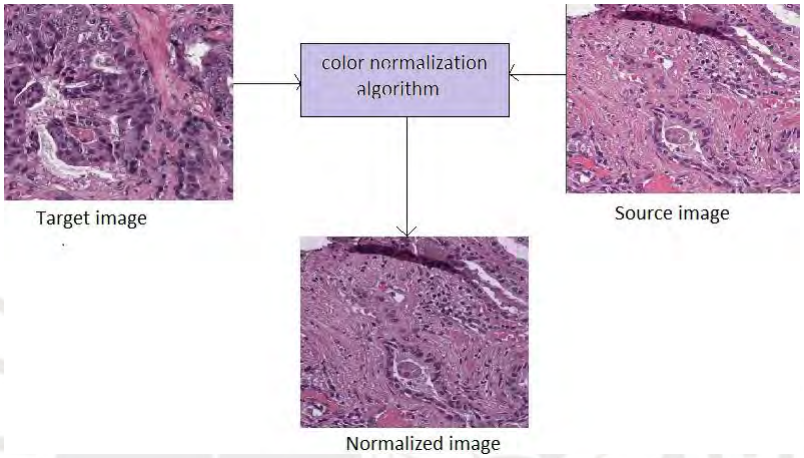


Figure 1.15: Color normalization architecture

From the different methods listed above and according to the requirements, the Reinhard method was chosen. According to Reinhard et al., an algorithm is proposed where the source image originally operating on channels Red, Green, and Blue is converted into the LAB color space. The color information is converted into lightness and color information independent of the device. The mean and standard deviation were extracted and then applied to the target image. The main advantage of the algorithm relies on the preservation of the original contrast from the target image.

In an effort to normalize the histological image, the histomicsTK library was added to our Python environment. The histomicsTK package was used to analyze digital pathology images since it contains various algorithms for color deconvolution, filtering, and color normalization, including Reinhard stain normalization.

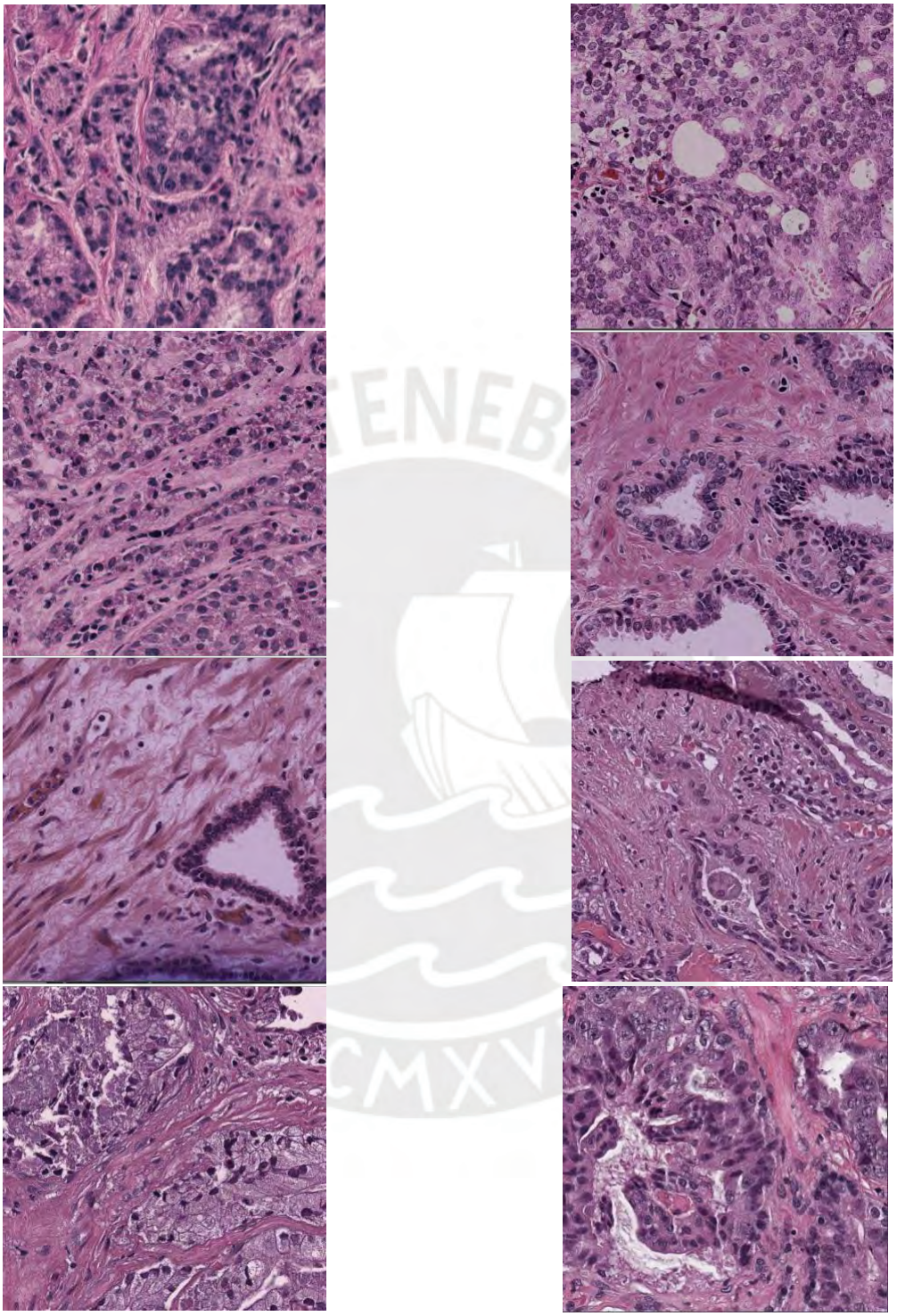


Figure 1.16: Stain disparity after applying Reinhard approach in histological image from TCGA dataset

Chapter II

Prediction of Biochemical Recurrence based on Phenotypic Features

Prostate cancers are classified according to their cellular architecture as shown in figure 1.1, showing the importance to differentiate each of the features. Thus, there is a need to segment the glands, nuclei, and lumina in histological images for any phenotypic variant I, to gather evidence about the state of the cancer. A histological image made up of defined segmented glands, centered around a clear lumen surrounded by cytoplasm and of regular size will be classified as a low Gleason score, benign cancer. Meanwhile, a histological image with diffuse growing tumor cells, without defined glands will be classified as a Gleason grade 4/5 prostate cancer. Therefore, the main point is to segment every phenotypic attribute attached to prostate cancer. Different studies have been focused to detect and classify each element of a histological image of prostate cancer. In general, there are five main approaches, texture-based, nuclei-based, gland-based, or lumen based. In some cases, they will detect the nuclei or lumen via color deconvolution or threshold, and reconstruct the gland from each detected element. However, in most recent studies, the application of artificial intelligence, AI has increased the robustness and accuracy of said segmentation. Naik and colleagues (Naik, Doyle, & Feldman,

2007) proposed potential lumina areas that were detected by a Bayesian classifier, then using a level set curve to detect the lumen. From the lumina, the nuclei were detected. Nguyen and colleagues (Nguyen, Sarkar, & Jain, 2012) detected areas of interest based on the intensity of the nuclei, then the shape was detected by its coordinates. Nevertheless, challenges remain in applying AI to histological images. Building on established strategies (L. He, Long, Antani, & Thoma, 2021) (Al-Kofahi, Lassoued, Lee, & Roysam, 2010), the first approach was designed exclusively via Deep Learning by merging and cascading different neural network techniques. Despite its increased use, Deep Learning still presents interesting challenges when applied to medical imaging. However, it enables to process great amounts of data and images and it has shown better results than most of the classical techniques. According to the results, different strategies were evaluated to achieve a rapid, accurate, and steady protocol to segment glands in WSI. In the following sections, different approaches were examined to evaluate deep learning models and pre-processing steps. The final results would allow us to capture the best prostate cancer phenotypic features in each of the slides to predict biochemical recurrence.

2.1 Segmentation using Convolutional Neural Network approaches

With recent image analysis advancements, deep learning methods have shown broader and superior results. Mask Regional Convolutional Neural Networks (Mask R-CNN) have been successful to segment the different features of the histological slide. Mask R-CNN is an instance segmentation, meaning it can identify different objects at a pixel level, providing, more precise results for small objects, such as nuclei. But before discussing any further the Mask R-CNN methodology, let's first explain what is a Convolutional Neural Network, the backbone of the Mask R-CNN.

2.1.1 Convolutional Neural Network principle

Deep learning methods, and in this case, Convolutional Neural Network (CNN) was designed to recognize patterns, usually generalized from previous knowledge and to adapt it to different environments, similar to the pattern recognition autonomously performed by the human eyes and brain (Indolia & Goswami, 2018).

A CNN consists of two different structures, a convolution layer, and a neural network layer.

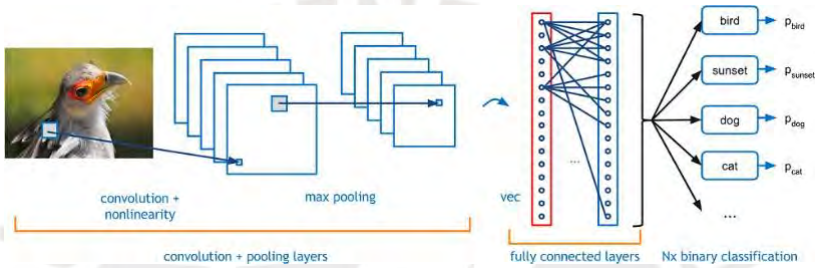


Figure 2.1: Convolutional Neural Network architecture [7]

The convolution layer takes an image as input, made up of three layers, representing the RGB values. The representative array of the image (480x480x3 for example) is passed through a first filter. In order to preserve the spatial relationships between pixels, the filtering is done by sliding a kernel or feature detector over the image in order to generate a feature map. The feature map is composed of the dot products between the kernel and the image represented in the example, in the case of an image whose pixel values range between 0 and 1 2.2.

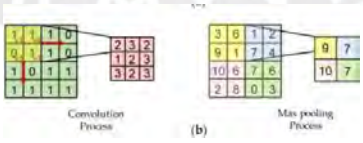


Figure 2.2: Feature map generation and maxpooling layers, (Chu et al., 2020)

Mathematically, the feature value (dot) is represented by:

$$z = wx + b \tag{2.1}$$

with w and b being the weight vectors.

The kernel is therefore used to extract the image feature and to determine the important features including curves, circles, and lines. Different kernels can be applied to the same image, thus creating different feature maps. The accuracy of the result is firmly linked to the number of filters. When this number increases, so does the accuracy. In Figure 2.1 five distinct kernels have been applied to create five feature maps. For instance, to detect a bird such as in Figure 2.1, the essential features are the curves to detect its beak or the shape of its head, the circles as eyes, etc.

Activation function

Most of the real-world data analyzed by the CNN are non-linear. The purpose of the activation function is then to introduce non-linearity to CNN. It eases the process as the model will be able to generalize or adapt to a collection of data. To do so, distinct functions can be introduced including tanh, ReLu, and sigmoid (Nwankpa, Ijomah, Gachagan, & Marshall, 2018). The sigmoid function exists between (0,1) as shown in Figure 2.3. It is therefore mainly used when predicting a probability.

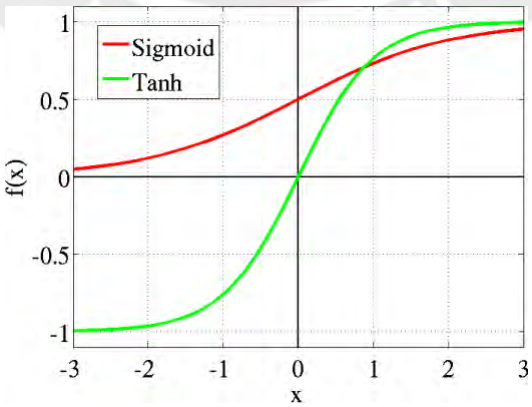


Figure 2.3: Sigmoid and tanh function (Pan et al., 2020)

Unlike the sigmoid function, tanh function ranges between (-1,1). The tanh activation function is mainly employed when the input has to be classified between two classes. Nevertheless, the most popular activation is Rectified Linear Unit (Relu) and ranges from (0, ∞). The ReLu function replaces all negative pixel values in the feature map with zero.

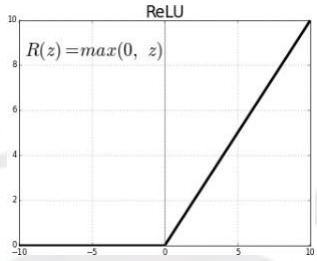


Figure 2.4: ReLu function (Rashid, 2019)

The activation function has to be applied after each convolution. These two layers can be applied in sequence, thus extracting more abstract features.

Max Pooling

A max pooling layer is then applied to the features map in order to reduce their dimensions while retaining the most important information. To do so, a window slides over the feature maps, and extracts the largest elements, as represented in Figure 2.2. The computational cost is hence significantly reduced with the use of CNN (Suarez & Segura, 2018).

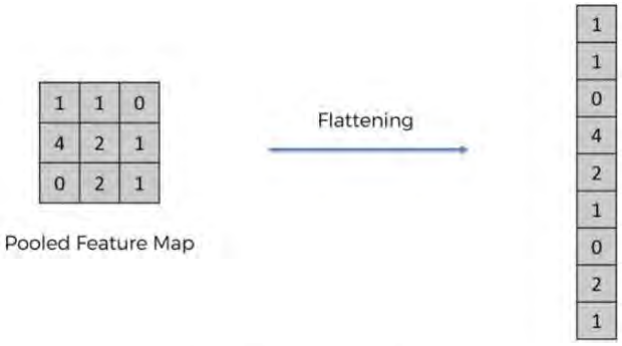


Figure 2.5: Flattening layer to reshape the data
(C. & Subhrasankar, 2019)

Flattening

The last layer before the neural network is to transform the matrix into a single linear vector, that can be fed to neurons 2.5.

Fully connected layer

Each feature is then fed to the input neurons of the neural network. The input neurons are linked to hidden neuron layers, then to the output neurons. The output neurons determine the classification of the image and define the final associated label (Q. Zhang, Wu, & Zhu, 2017). The fully connected layer allows the combination of different features. In the end, in order to recognize a bird, features like beak, feathers, and eyes among others are more than necessary, while components like nose, fur, or mouth won't be of any use.

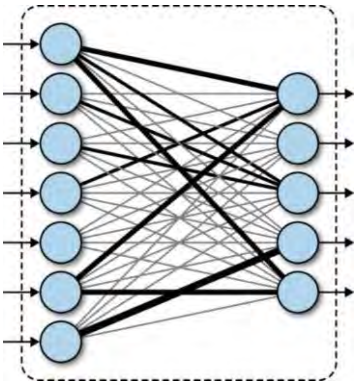


Figure 2.6: Fully connected layer (Tammina, 2019)

Back-propagation

Each of the neurons of layers has randomly initialized weights. Knowing the label of the image, by back-propagation, the weights of the neurons are uploaded at each step until meeting the appropriate label. Back-propagation is used to calculate the gradients of the error of the network by comparing the actual targets to the predicted ones and to handle the gradient descent to update the value of the weights in order to minimize the error. The error is actually represented by the loss function. The main benefit of backpropagation is to reduce the loss function (Li, Cheng, & Shi, 2012).

$$w = wi - n\frac{dL}{dW} \tag{2.2}$$

Where w the update weights,wi represents the initial weights, $\frac{dL}{dW}$ the loss function and n the learning rate.

The learning rate is determined by the user and affects how much each weight is uploaded. A considerable learning weight is represented by an increase between the former weight and the actual one. The model is trained until the loss function approaches zero.

Accuracy and loss

The performance of deep learning models is evaluated on the training and test sets with the accuracy and loss. By definition, accuracy is the number of predictions where the predicted values are equal to the true value (Jung, Bi, & Davuluri, 2019). Loss on the contrary does not represent a percentage but gathers the distance between the true value and the predicted one. In the training process, the loss function is gathered to update the weights 2.1.1. In most models, the aim is to lessen the loss, while increasing the accuracy. Both provide essential information, as low accuracy and low loss mean that the errors are small and abundant, the model is then approaching the optimal weights. Meanwhile, a high loss and high accuracy might indicate fewer but more important errors in some classes.

For the following study, the loss, and accuracy were the only metrics employed for model evaluation. Precision, recall, and F1-score are also an option and are explained in 3.2.4

2.1.2 Mask Regional Convolutional Neural Network principle

The Mask R-CNN approaches is based on CNN. The same method is applied to extract the essential features (edge, corner) by using kernel and activation function (K. He, Gkioxari, & Dollar, 2017).

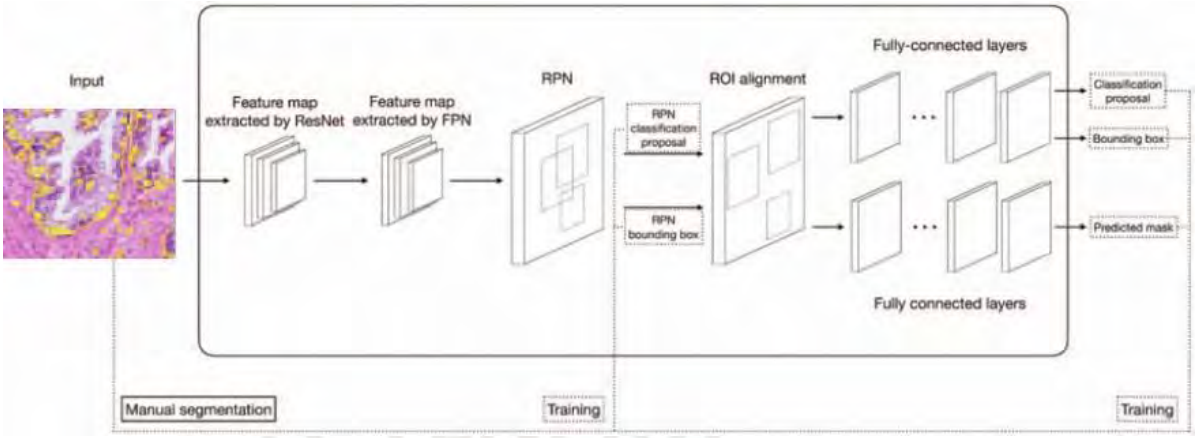


Figure 2.7: Mask R-CNN structure influenced by (Hyun et al., 2021)

Feature Pyramid Network

Feature Pyramid Network, FPN, is added to our Mask R-CNN model as an extension to improve the detection of different scale objects. The bottom-up and top-down pathways composed the FPN. The bottom-up pathway is a usual convolutional network to extract features. As we go through the different layers, the spatial resolution decreases, but the semantic increases. CNN usually operated the final layers for object detection, which generates poor performance for small objects. FPN output feature maps at each last layer of each stage and then fed to the top-down pathway by lateral connection to assist the detector for improved objects location prediction (Lin et al., 2017).

The Top-Down pathway collects the feature maps, up-sampled by a factor of 2 with the nearest neighbor, then undertook a 1x1 2D convolutional layer in order to reduce the channel dimension. Then, both feature maps from the bottom-up and top-down are merged together. A final 3x3 2d convolutional layer generates the final feature map from the merged ones.

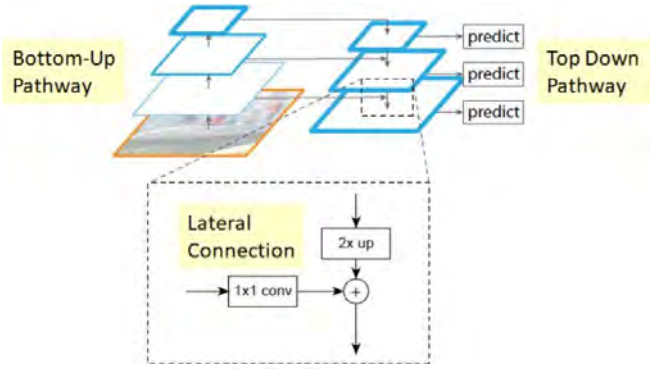


Figure 2.8: Feature Pyramid network (*Review: FPN — Feature Pyramid Network*, n.d.)

2.1.3 Region Proposal Network

Once the feature maps are extracted from the image, a small CNN window scans the feature maps. At each position, the network predicts whether there is an object present or not and also predicts the coordinates of a bounding box that tightly encloses the object. The Regions of Interest (ROI) are then fed to the next layer.

The classification layer of the ROI classifier evaluates each ROI and generates two outputs: the label of the image (planes, person, animal) or discards the ROI as background information. While the regression layer of the classifier refines the bounding-box, surrounding the objects as well. The objects on the image can have different shapes, due to each element scale. However, classifiers can not manage data with variable sizes. Therefore, each bounding box need to go through a max pooling layer as explained on 2.1.1. A convolutional network will then generate the masks of the ROI from the Region Proposal Network. With the combination of the masks, the bounding box will segment the objects instantly with high definition and precision (Shah, Kasukurthi, & Pande, 2019).

2.1.4 Application of Mask R-CNN to segment glands and nuclei from histological images

Mask R-CNN, enables the detection, segmentation, and classification, all at once, of different objects, from the instance segmentation aspect of the model. A Mask R-CNN skeleton from Adulla's work (Abdulla, 2018) available on Github was manipulated. The skeleton Mask R-CNN proposed and trained, can identify on an image or video, humans, bikes, or even cars. A public model has already been trained to detect the nuclei from this same skeleton (*Nuclei Counting and Segmentation*, n.d.). The model will be modified, adapted, and trained to detect the following phenotypic components.

1. Glands
    (a) Foamy Glands
    (b) Atrophic Glands
    (c) Ductal Adenocarcinoma Glands
    (d) Mucinous Glands
2. Cells
    (a) Epithelials
    (b) Fibroblasts
    (c) Tumor cells
3. Artifacts
    (a) Blood vessels

The starting point is to annotate the image according to the elements from the list 2.1.4 soon to be classified. Therefore, with an online annotator, such as VGG `https://www.robots.ox.ac.uk/~vgg/software/via/via.html`, every coordinate and its associated label are saved in a JSON file (polygon: every point, ellipse/circle: center and radius). The plan was to segment simultaneously through the Mask R-CNN the tumor cells, epithelial cells, and glands areas.

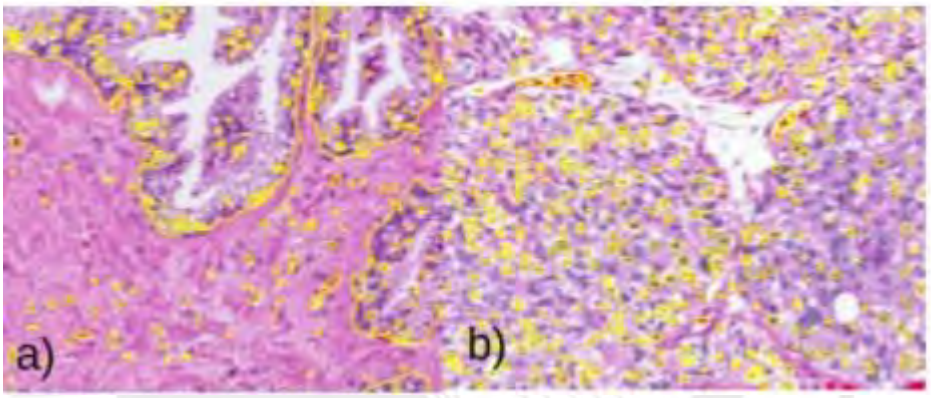


Figure 2.9: Image annotations

One hundred images were analyzed. Opposite to classic CNN, there is no need to load data. Indeed, the model was not trained from scratch, as we started with a weight file, trained with the COCO or imagenet dataset. The COCO and imagenet dataset consists of hundred labels from common objects (peoples, cars, planes). The data is split with 66% of the set corresponding to the training set, while the remaining is used as a testing set.

Images were loaded and masks generated from the JSON annotation were transposed into polygons. The model is backboned with resnet101 (Y. Zhang, Chu, Leng, & Miao, 2020); which is a pre-trained 101 layers deep convolutional network. The bottom-up pathway feature maps are extracted at each level with resnet101, and then fed to the Top Down Pathway following the theory 2.1.3. A 3x3 sliding window then scans the final merged feature map con-

sisting of a 3x3 convolutional layer, along with 2 1x1 convolutional layers. Mask of the ROI was generated by a convolutional network comprising of four convolutional layers of size 256, kernel (3x3), then, a convolutional layer 256, kernel (2x2), and finally a convolutional size 8, corresponding to the 8 classes to segment and classify, with the kernel (1x1). The activation function used in between the model was relu function, while the softmax was used in the output layer. Parameters were tuned and the model was tested until the best results (accuracy and loss) were retrieved. The best model was trained over 1000 epochs, with a learning rate of 0.001.

The Mask R-CNN model was trained for around one month. Nevertheless, as seen in figure2.10, the results are quite deceptive. Only a few glands were segmented and classified despite two months of training. Furthermore, hardly any nuclei have been detected. The same result can be seen in figure 2.11 where the focus is on the nuclei. Only a handful of nuclei have been segmented and classified. Additionally, the nuclei are detected but only with an accuracy of 60%, and none are tumor cells. The Mask R-CNN is thereby classifying tumor cells as healthy ones.

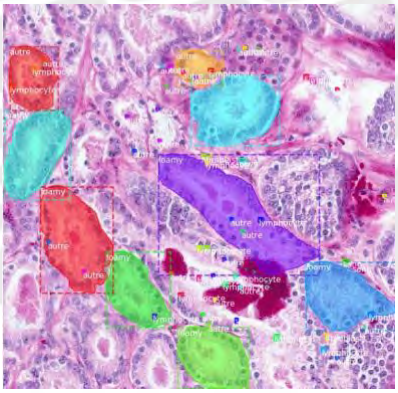


Figure 2.10: Image segmentation and classification according to the Mask R-CNN model

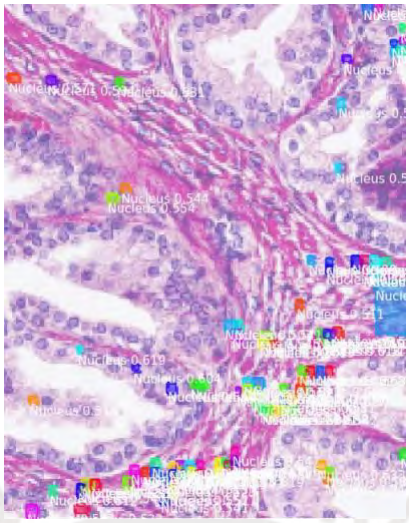


Figure 2.11: Nuclei segmentation and classification according to the Mask R-CNN model

The poor results can be explained by the following clarification. Firstly, it can be linked to the shape of the images. We process images of size 1024x1024x3 in order to have whole glands in the images. The substantial size of the images can actually downgrade our final results. Since nuclei are small-scaled in comparison to the glands, the algorithm won't be able to detect them. Another explanation can be due to the unbalanced data. We have an average of 10 glands per image, against hundreds of nuclei. The results for the under-represented class will be mediocre, as there are fewer observations in the set.

Moreover, the glands were annotated manually and segmented according to the architecture in figure 2.12, due to its unchallenging pattern. A gland is detected if there is the presence of a lumen, cytoplasm, and surrounded by a layer of nuclei. However, with a highly aggressive tumor the shape of the glands is irregular, there is a presence of stroma inside the gland and the disappearance of the lumen. In other words, the model would work on healthy histological images but would show very poor results on Gleason 4 and 5.

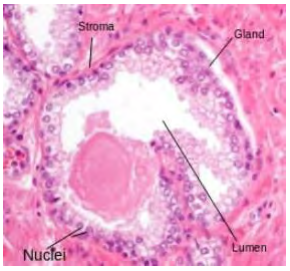


Figure 2.12: Healthy gland architecture

2.1.5 Segmentation of nuclei via Mask R-CNN

The first alternative would be to only focus on nuclei segmentation via Mask R-CNN. In order to do so, the Kaggle 2018 Data Science Bowl was used. In this competition, the web page Kaggle offered a dataset containing numerous segmented nuclei from different conditions adopted as the base of the training dataset. This event was created as a public competition to challenge the nuclei segmentation models.

Additionally to the Kaggle dataset, the nuclei annotated manually previously are added to the set as external data, taking inspiration from the technique used by the 5th winner of the Kaggle competition, Inom Mirzaev (Abdulla, 2018).

Most of the cells are actually transparent and without color, before the Haematoxylin and Eosin (H&E) staining process. A color deconvolution was applied to every slide, to isolate the purple-stained nuclei on the Haematoxylin channel from the pink-stained stroma.

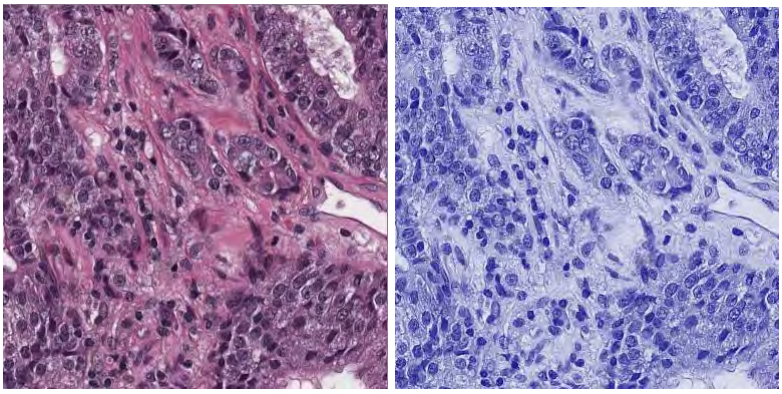


Figure 2.13: RGB initial image and Haematoxylin channel

As seen on Figure 2.13 nuclei are stained purple while the rest is of lighter color, hence increasing the contrast between the region of interest and the background information. The slide, with an initial shape of 2048x2048x3 (Haematoxylin channel) is moreover split into 8 tiles of shape 256x256x3 to increase the overall pace of the training process.

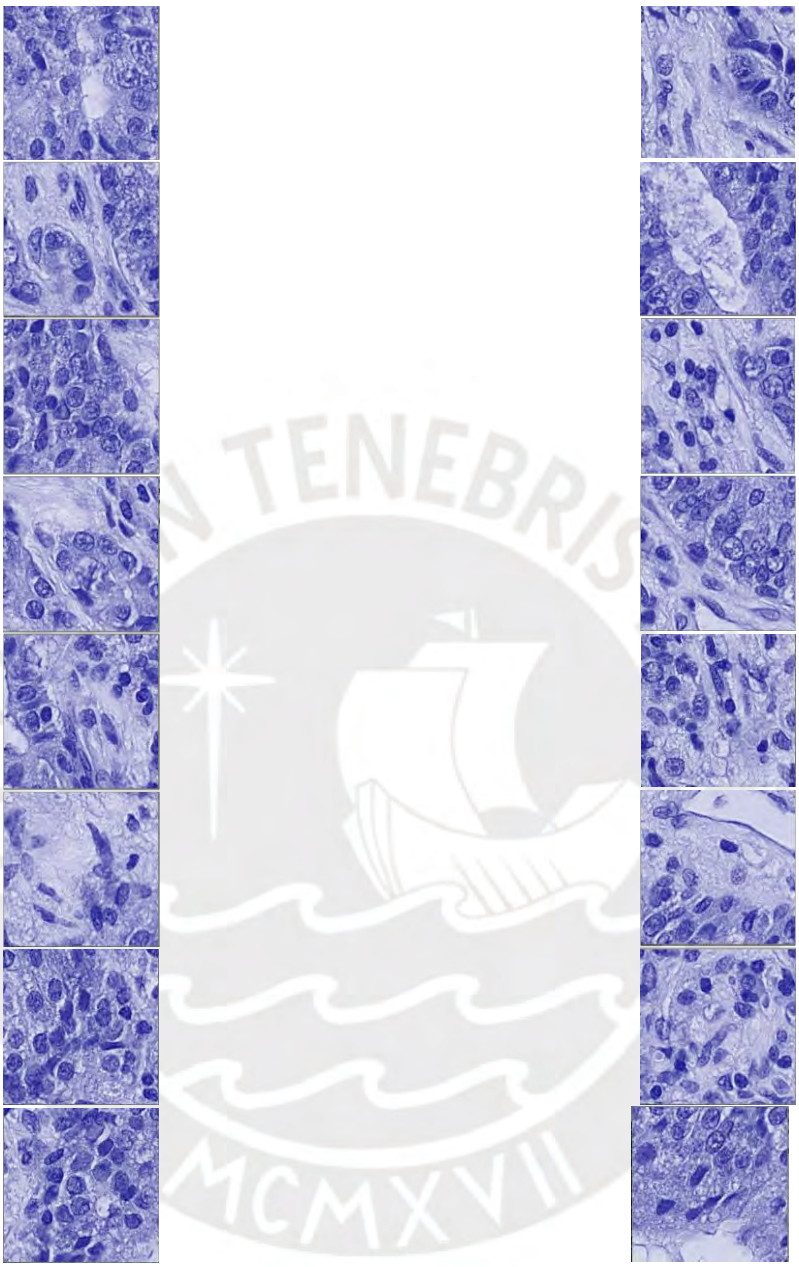


Figure 2.14: Image tilling into 16 sub slides of size 256x256x3

The final set consists of thousands of Haematoxylin stained tiles.

2.1.6 Mask R-CNN results for segmentation of nuclei

The identical trained Mask R-CNN skeleton from the previous section was applied, however, the model to detect nuclei was different. The first step consisted of increasing the database available, adding external data of thousands of images to the training and test dataset. Therefore, the implementation heavily relied on image augmentation such as image rotation, image cropping, and image scaling. The set is then divided into 90% training and 10% validation cohorts. Tiles from the same image can be found in both the training and testing sets due to the high variance of the pattern.

The Mask R-CNN was trained through 75 epochs :

- 25 epochs with a learning of $1.e^{(-4)}$
- 25 epochs with a learning of $1.e^{(-5)}$
- 25 epochs with a learning of $1.e^{(-6)}$

At the end of the training, we had a loss of 13% on the training dataset with an accuracy of 73%, and a loss of 15% on the testing dataset and 71% accuracy.

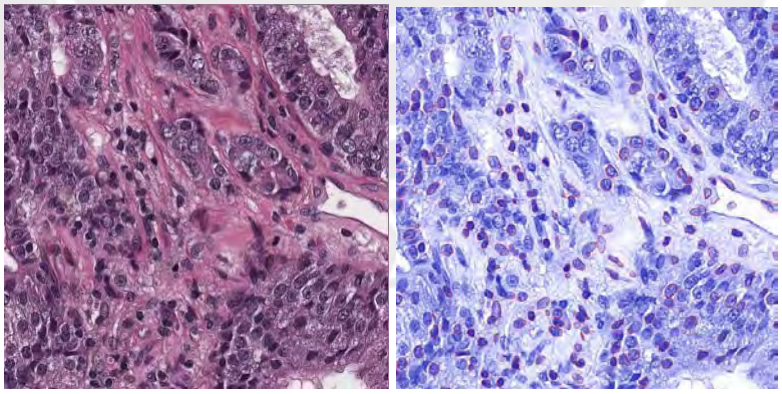


Figure 2.15: Nuclei identification results from the Mask R-CNN implementation

After post-processing the image and joining the tiles, we obtained the image 2.15. Nuclei segmentation and classification did increase, but presented poor results over the edges and

performed unsatisfactorily on tumorous cells. Nevertheless, this approach delivered superior results than the first one but needs more training.

2.1.7 Glands reconstruction from segmented nuclei

With this second approach, only the nuclei were segmented. The glands were then reconstructed through Delaunay's triangulation, as the example shown in figure 2.16. Indeed, glands were organized around a lumen, and delimited by an external layer of nuclei. This solution would then work perfectly for well-arranged glands. Otherwise, with a tumor of a higher score, the poor feasibility of this methodology would deliver crucial hints as to the state of the cancer.

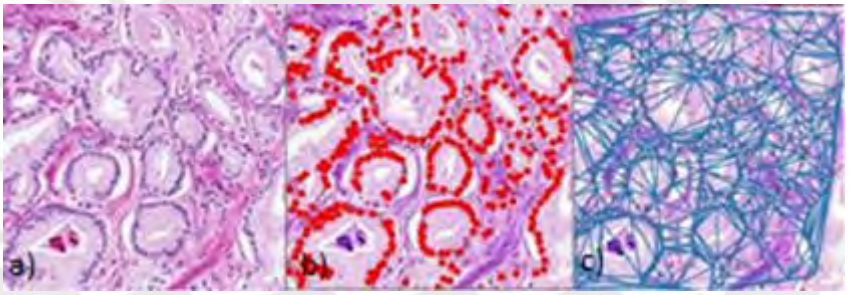


Figure 2.16: a) digital histology image, b) nuclei annotation, c) Delaunay's Triangulation application

By definition, Delaunay's triangulation consists of a triangulation of a set of points, such as no point inside the circumcircle of any triangle. Gland contours are distinctly observed after applying Delaunay's triangulation considering, a higher density of segments can be discerned in the layer of nuclei.

The lack of segmented nuclei as seen in figure 2.11 prohibited us to continue with this methodology. Indeed, Delaunay's triangulation did not allow a correct segmentation of the glands. Although as said above, the poor feasibility of the methodology may actually give

major information, the segmentation won't be conclusive once a hint of disorganization is perceived in the histological images. The training will also have major issues regarding some variants of the phenotype. Foamy and atrophic glands, may be successfully segmented, but ductal adenocarcinoma, mucinous, and signet ring cells may not be.

2.1.8 Segmentation of glands via CNN

A simpler approach than the Mask R-CNN nuclei segmentation is to use a convolutional neural network, CNN, based on the Mask R-CNN. Contrasted to Mask R-CNN, the CNN is less computationally heavy, dwindling the time of training and testing, enabling the experimentation of different models.

In this case, the database is similar to the one employed with the mask R-CNN but with different annotations. Indeed, only the glands were contoured and fed to the CNN 2.1.1. To increase the size of the dataset, data augmentation techniques were practiced. Mainly consisting of manipulating the images, by rotating, shifting, and flipping them to artificially expand the dataset. External datasets were used as well, as a backbone by adding annotated colon cancer images from the Gland Segmentation in Histology Images Challenge Contest held at MICCAI 2015. Out of the set, 66% trained the model, while the remaining 33% served as testing set.

The CNN model represented below demonstrates the best segmentation of glands in prostate cancer images, consisting of five blocks composed in order:

1. Block 1
    (a) 2D convolutional layer, size 64, kernel (3x3) - Conv 1
    (b) LeakyRelU - Conv 1
    (c) Spatial Dropout - Conv 1
    (d) 2D convolutional layer, size 64, kernel (3x3)- Conv 1

(e) LeakyRelU - Conv 1

(f) Spatial Dropout - Conv 1

(g) Average Pooling 2D - Pool 1

2. Block 2

(a) 2D convolutional layer, size 128, kernel (3x3) - Pool 1

(b) LeakyRelU - Conv 2

(c) Spatial Dropout - Conv 2

(d) 2D convolutional layer, size 128, kernel (3x3) - Conv 2

(e) LeakyRelU - Conv 2

(f) Spatial Dropout - Conv 2

(g) Average Pooling 2D - Pool 2

3. Block 3

(a) 2D convolutional layer, size 128, kernel (5x5) - Pool 2

(b) LeakyRelU - Conv 3

(c) Spatial Dropout - Conv 3

(d) 2D convolutional layer, size 128, kernel (5x5) - Conv 3

(e) LeakyRelU - Conv 3

(f) Spatial Dropout - Conv 3

4. Block 4

(a) Concatenate - - Conv 2 & Conv 3

(b) 2D convolutional layer, size 64 , kernel (3x3)- Conv 4

(c) LeakyRelU - Conv 4

(d) Spatial Dropout- Conv 4

(e) 2D convolutional layer, size 64 , kernel (3x3)- Conv 4

(f) LeakyRelU- Conv 4

(g) Spatial Dropout- Conv 4

5. Block 5

(a) Concatenate - - Conv 1 & Conv 5

(b) 2D convolutional layer, size 256 , kernel (7x7)- Conv 5

(c) LeakyRelU - Conv 5

(d) Spatial Dropout- Conv 5

(e) 2D convolutional layer, size 256 , kernel (7x7)- Conv 5

(f) LeakyRelU- Conv 5

(g) Spatial Dropout- Conv 5

6. Output

(a) 2D convolutional layer, size 1 , kernel (1x1)

The dropout layer, explained in section 3.1, benefits the model as it reduces the risk of overfitting. While the relu activation was applied in various blocks, in the final layer, the sigmoid activation was employed to predict the probability of the region of interest being a gland. The model was trained over 200 epochs with a learning rate of 0.0001. Nevertheless, the weights were uploaded with Adam optimizer, using the estimation of first and second moments.

The time of training was reduced, and most of the glands were segmented. Nevertheless, the accuracy of the segmentation diminished as the Gleason score increased. Undeniably, slides with low Gleason, where glands are organized, and the lumen is surrounded by cytoplasm followed by a layer of nuclei were detected fairly well, but as the Gleason increased and so the disorganization of the glands, the model was unable to detect and segment the glands, or cluster of nuclei.

2.2 Segmentation of phenotypic features via U-net

As explained in the previous approaches, the Mask R-CNN and CNN model trained were not the most appropriate options to segment glands and nuclei from high Gleason tumor WSI and specific phenotypes. One alternative is to change the model, by adding layers, and adjusting the learning rate; but as seen, these models can be computationally heavy, and without the proper technology; it may take months to train. Therefore, a U-net approach was tested to segment the glands. As a matter of fact, per its attributes, U-net architecture shrinks the training time and facilitates the segmentation of the required features.

2.2.1 U-net architecture

U-net architecture involves two different parts: The encoder and the decoder, as seen in 2.17. The encoder section is a traditional convolutional and max pooling layer (explained in 2.1.1). At each layer, the number of kernels doubles to learn complex features. The network then learns the object inside the images but is losing its localization. In order to reconstruct the spatial resolution, the feature map is upsampled, or in other words, passed through a transposed convolutional network. Essentially, a transposed convolutional network undoes the process of convolution (Punn & Agarwal, 2020).

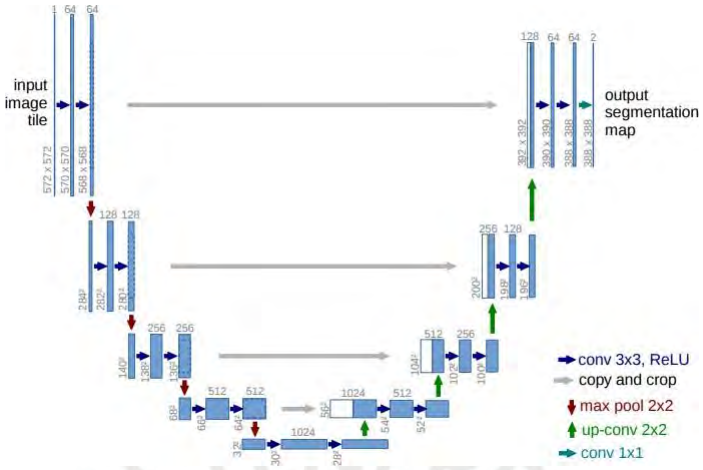


Figure 2.17: U-net architecture (*Humans Image Segmentation with Unet using Tensorflow Keras*, n.d.)

Another issue witnessed in the last approaches stems from the dataset itself. Certainly, most of the annotations considered mostly well-formed glands as in the first approaches the images were manually annotated, thus setting aside clusters of nuclei for lack of experience. In the aforementioned approach, the dataset was composed of the Panda Kaggle dataset, in which tiles were extracted with the scoring tiles. Since five tiles per patient were extracted, the splitting between training, including 90% of patients, and testing was done on the patients' level, rather than on the tiles level.

Since a large and important amount of images were employed, it was necessary to increase the speed of the training process. Instead of feeding them directly to the input layer, they were stocked in a single hdf5 file. Image augmentation namely rotation, flip, and crop was operated. The U-net network skeleton was adapted from (Ronneberger, Fischer, & Brox, 2015). The major model alteration was to adapt the model to our data. The main components of the architecture remained the same. Nevertheless, parameters such as the depth, number of befalling blocks of the U-net, and number of kernels at each layer were adjusted according to our requirements. The original paper (Ronneberger et al., 2015) suggests a number of kernels wf, of

5 and a depth of 6 for images of 512x512 pixels. The number of kernels increases analogously with the number of layers following $2 ** (wf + layer_number)$. Nonetheless, the slides fed to the model were considerably larger. Small-scaled phenotype elements may hence be overlooked by the initial architecture. To ensure high-speed training, while avoiding the likelihood of over-fitting, wf is altered to 6, while depth remained at 6. Implemented with an SGD optimizer, the model with a learning rate of 0.1, a weight decay of 0.0005, and a momentum of 0.9 proffered the highest accuracy while maintaining low loss on the test samples.

The U-net model trained for less than one week, outperforming the CNN or Mask R-CNN approaches in terms of training velocity. Moreover, the algorithm detected glands of different shapes, from well-formed to irregular ones, along with clusters of nuclei as seen on the left of the image 2.18. After training, the model performed efficiently on the remaining 10% test set, with an accuracy of 85% and a loss below 10%.

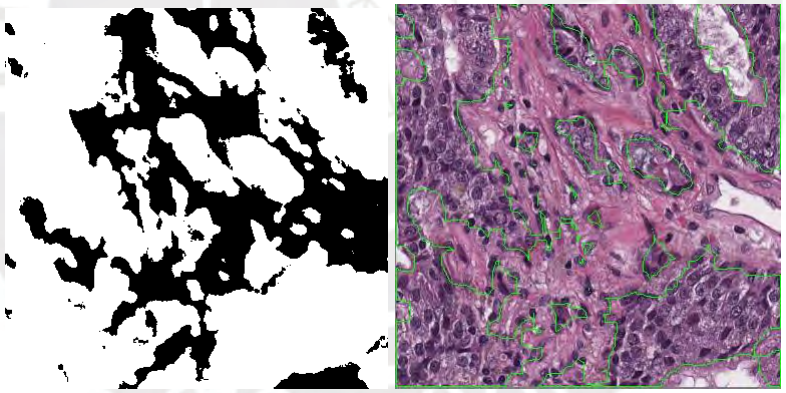
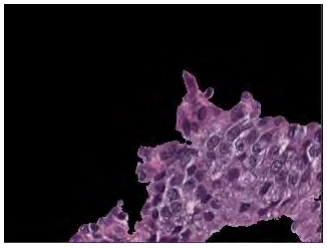
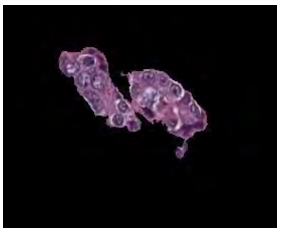


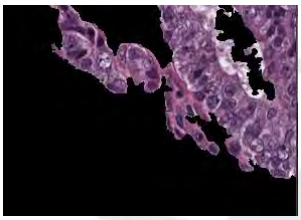
Figure 2.18: Gland segmentation from U-net model



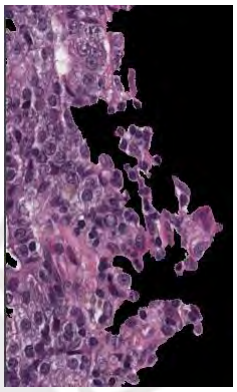
(a) Gland number 1



(b) Gland number 2



(c) Gland number 3



(d) Gland number 4

Figure 2.19: Glands segmentation obtained via U-Net approach

From the accuracy and swiftness of the results, the U-net model was the most reliable to segment the glands in the chosen dataset. Despite the high accuracy and low loss, an essential characteristic must be underlined. As a matter of fact, in 2.19a, 2.19c, and at the edge of 2.19d, the lumina were not recognized as part of the glands, and remained as background. This information is crucial when splitting and reconstructing the glands in 2.3.

Some masks of the glands were not completed and were left open. To resolve the issue, and to ensure that every gland gets segmented correctly, the U-net results were dilated by a kernel 5×5. The 5x5 kernel window slides over every pixel of the tiles, and the output pixel was equal to the maximum value of the neighborhood 25 pixels. In 2.20a, the red square represents the kernel centered on pixel 250. The output pixel took the maximum value out of the 25 neighborhood pixels, being 250, and so its neighboring pixels of value 42 and 26.

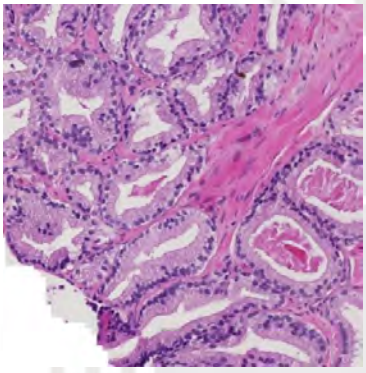
15	22	56	255	24	58	69	52
45	12	102	200	0	15	240	23
89	100	120	152	140	23	75	85
95	41	23	100	12	5	10	255
200	140	152	250	42	26	2	8
20	56	44	85	74	101	114	44
10	14	87	120	152	0	48	87
105	89	47	56	22	10	15	0

(a) Input matrix

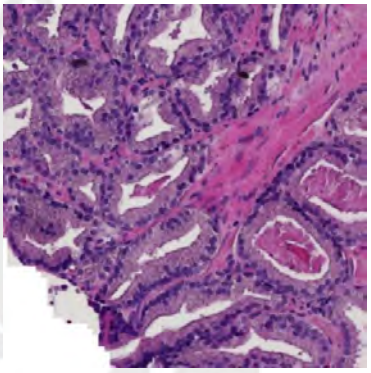
			250				

(b) Output matrix for centered pixel

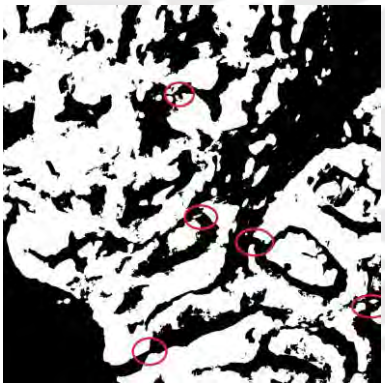
Figure 2.20: Morphological dilatation



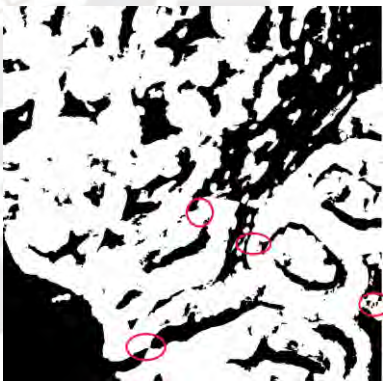
(a) Dilated by kernel 5x5 RGB tiles



(b) Original RGB tiles



(c) U-net results for RGB non dilated tiles



(d) U-net results for RGB 5x5 dilated tiles

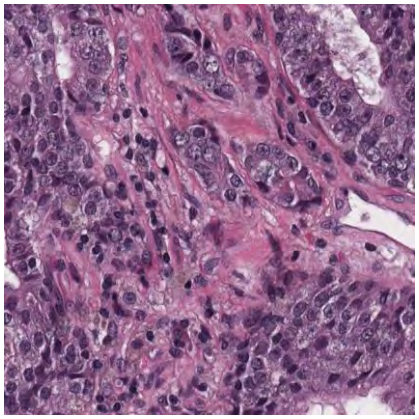
Figure 2.21: Differentiation in U-net results using dilatation

Dilatation, although subtle, did fix the issue, as there are fewer holes in the masks. Nevertheless, spatial accuracy was lost when the lumen automatically decreased. Further morphological image processing was needed to be applied to properly segment the lumen.

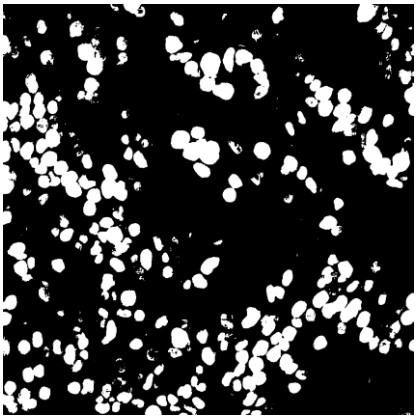
2.2.2 U-net nuclei results on histological annotations

Given the speed of the U-net approach, it is undoubtedly the leading proposal out of these approaches. To collect the phenotypic features with the highest precision, and segment to the closest nuclei, distinctive image pre-processing was tested to segment the nuclei with the same model.

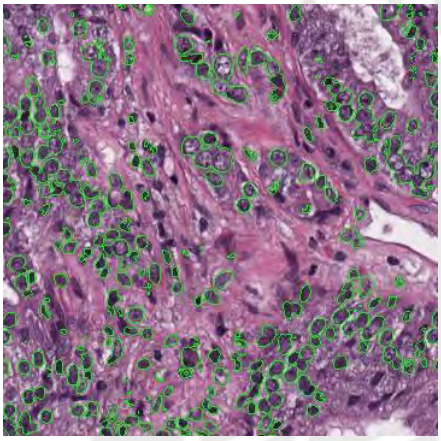
The first test consisted of simply applying the same exact U-net model to a substantially smaller set consisting of 60 H&E stained images and respective segmented nuclei, with the aim of segmenting nuclei. The dataset from Section 2.1.5 where cells from a distinctive type of cancer were used. Equivalently to the method in 2.2.1, the images were stocked in a hdf5 file, however, the Kaggle dataset has no information about patients' backgrounds, hence each image was considered as a unique patient separating them into 90% training set, and 10% testing set. After training, an accuracy of 83% and binary cross entropy loss of 8% was obtained on the testing set.



(a) Haematoxylin and Eosin staining



(b) nuclei mask from the U-net model

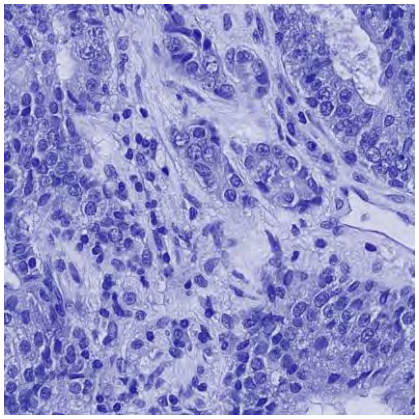


(c) U-net segmentation output

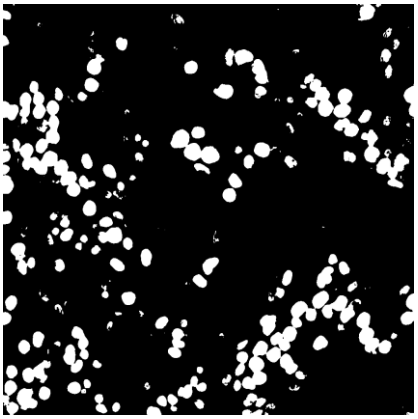
Figure 2.22: Nuclei segmentation obtained via U-Net approach with an RGB image

As observed, the first test was quite successful with a low loss, and high accuracy. Most of the nuclei were identified, however, some remained labeled as background, especially around the edges.

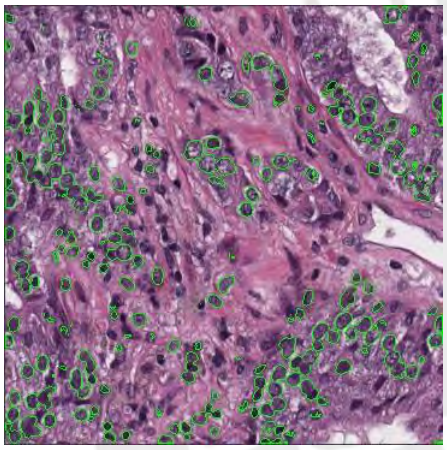
From the characteristic of the TCGA images and H&E staining, the nuclei are colored blue, while the glands in pink. This singular attribute might be helpful to palliate the issue identified in the first test. As a matter of fact, since it is possible to separate the different stained elements, by applying the U-net model only on the hematoxylin stain, where the nuclei were highlighted, the accuracy might improve. Nevertheless, with a loss of 19% and an accuracy of 78%, this hypothesis was proven to be erroneous.



(a) Haematoxylin staining



(b) nuclei mask from the U-net mode



(c) U-net segmentation output

Figure 2.23: Nuclei segmentation obtained via U-Net approach from a Haematoxylin image

As witnessed above, the accuracy was actually lower, and the loss was higher. The results can be clearly seen in 2.23c, fewer nuclei were identified and segmented. Nevertheless, by adding 2.23c and 2.22c, the total number of nuclei identified increased.

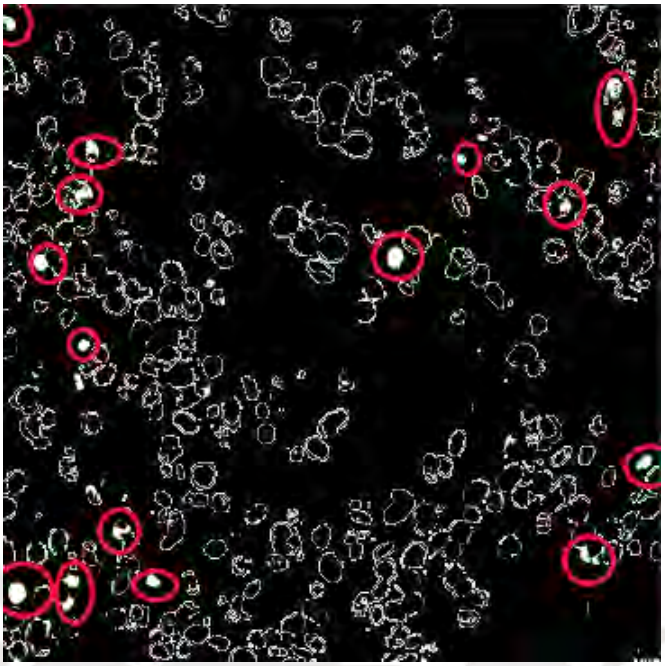


Figure 2.24: Mask resulting from subtracting U-net output on RGB and Haematoxylin images

From 2.24, the attention can be brought to the red circles. As a matter of fact, applying the U-net model to Haematoxylin stained images, allowed spotting a couple of nuclei overlooked when applied to RGB images. The results might seem moderate, but our glands separation and reconstruction approach relies heavily on a righteous nuclei segmentation.

The U-net offers a great alternative to segment elements of different scales from small datasets, proving its robustness and ability to be generalized. Also, from its accessible implementation and efficient training, U-net displays great results to segment histological images.

2.2.3 Improving the segmentation of the nuclei

One of the most significant challenges, when working on histological images, is the proximity of its elements given that some nuclei and glands are too close to each other for the algorithm to recognize them individually. This issue is even greater in the case of prostate cancer. Indeed, as

mentioned in 1.1, the higher the Gleason Score is, the less organized the glands are. This matter can become quite problematic while retrieving the phenotypic features. To address this, the focus was made on splitting the nuclei that have been recognized as one, then on reconstructing the glands from the lumen to separate any merged glands.

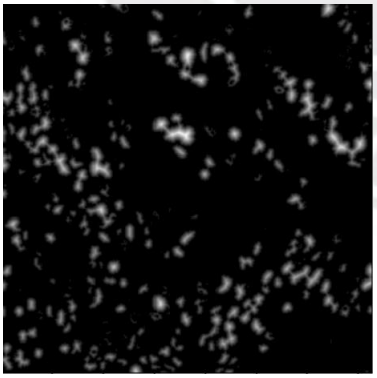
The first step consists on splitting the nuclei that have been merged as observed in 2.25. According to the shape of the object to separate, various procedures can be directed, however the best strategy to segment round objects such as nuclei is by distance Transform Watershed (Gamarra, Eduardo, Banerjeelante, Hurtado, & San-Juan-Vergara4, 2019).

(a) Mask from u-net merging 4 nuclei into 1

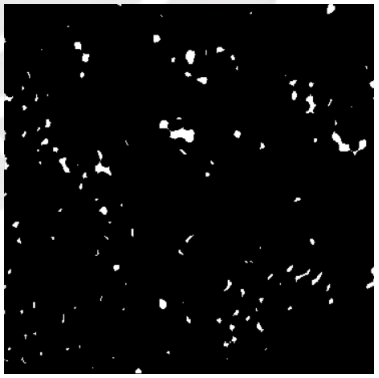
(b) Contour of the merged nuclei

Figure 2.25: Overlapping nuclei

The distance transform of the binary image's complement resulting from the U-net segmentation is computed from the distance of each pixel to the closest non-zero value-pixel, creating "watershed ridge lines" and "catchment basins" (G.Wang, Mang, H.Cai, & et al, 2016).



(a) Distance transform



(b) Distance transform threshold

Figure 2.26: Watershed to mask

Only the regions with the highest intensity from distance transform were kept. The back-

ground was now distinguished from each separated nucleus; labels were then associated with each region and watershed applied to divide the nuclei.

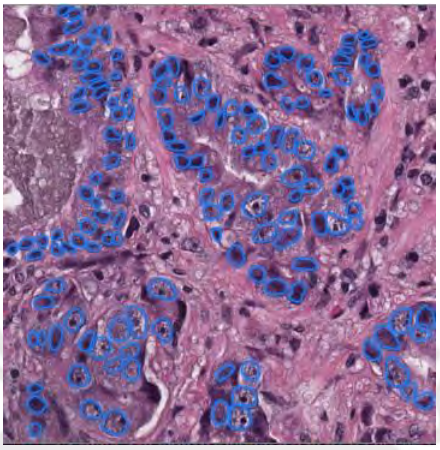


Figure 2.27: Ground truth for nuclei contours

The intensity value is extracted from ground truth contoured by a professional 2.27.

(a) Distance transform threshold

(b) Watershed mask

Figure 2.28: Watershed result to separate merged nuclei

From the illustrated results in 2.28a and 2.28b, nuclei area was lost in the process. None withstanding, the number of nuclei is more precisely extracted and so is the density of nuclei. As nuclei do not fluctuate tremendously in size, the density of nuclei can be calculated from the number of instances in the slide.

Some elements in the image were detected and classified as nuclei, such as in image 2.29. To correct any wrongdoing of the algorithm, the distance between the center of the lumen (to be explained in the following subsection 2.2.4) and each nucleus of the glands was calculated.

After computing the median, each of the elements detected as nuclei, which distance from the lumen is too far away from the median, were removed.

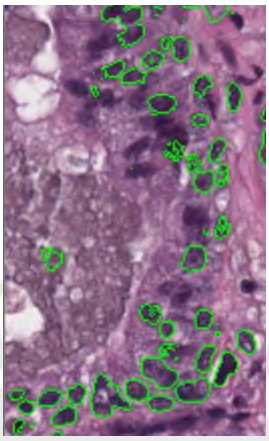


Figure 2.29: Erroneous segmented nuclei

2.2.4 Identification of the lumen

From the glands' segmentation results, all glands were worked on separately, as shown in figure 2.19. Lumen is an important feature to define the stage of cancer. A clear, well-defined lumen, surrounded by cytoplasm and a layer of nuclei, is a sign of healthy tissues. This can clearly be seen in the gland number 2, as there is no lumen in the glands, with irregular shape and presence of tumorous nuclei. The U-net model was trained to detect the lumen inside the glands, by considering them part of the background and creating a new contour inside the glands, as seen in 2.18. Instinctively, we would want to work only on the segmented background to isolate lumen by color condition, nonetheless the lumen wouldn't be matched with its righteous gland's contour.

Figure 2.30: Lumen segmented from background

By using the hierarchy tools from Open-cv library, it was possible to extract the lumen, keeping the glands who have a child contour.

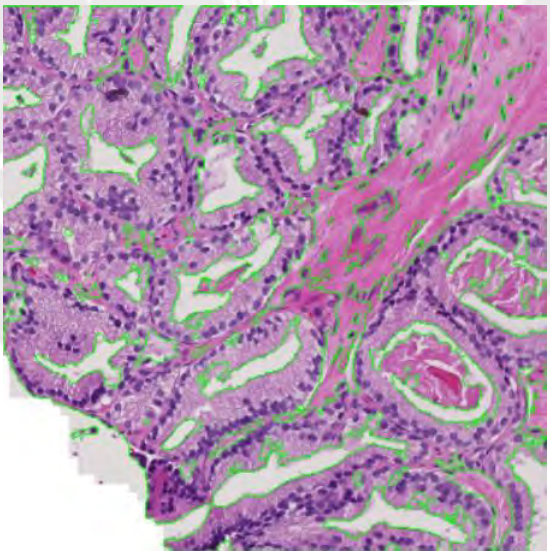


Figure 2.31: Lumen inside contours

2.2.5 Hierarchy of the contours

Open Computer vision is a library available in Python and C++ dedicated to image processing. Open CV proposes a wide panel of tools, from morphological transformation to threshold among others. In this case, the focus was on the contour tools. Contours are continuous points with the same color or intensity that define a curve. Different options are available to retrieve the contours, for this matter, all contours were extracted from the U-net output, external and internal as seen in 2.18. When talking about a gland with a presence of a lumen, the presence of internal contours also called a child contour is required, while the outer is known as the parents. The hierarchy of the glands is expressed according to the following forms and displayed in 2.32.

1. If the contour has not internal/child contour
    (a) [−1, −1, −1, −1]
2. If the contour has internal/child contour
    (a) [−1, −1, −3, −1]
    (b) element in the 3rd position is different from -1
3. If the contour has a external/parent contour
    (a) [8, −1, −1, −1]
    (b) element in the 1st position is different from -1

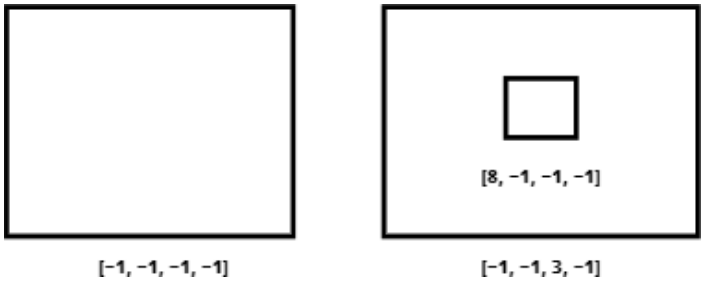


Figure 2.32: Hierarchy of the contours and its expressed form

Looping through the contours, only the one with parent contours is kept to ensure the recollection of a lumen. The instinctive thinking would be to loop only through the contours with child contours. The main issue there relies upon the case where the contour involves more than one gland. Indeed, some glands were too close to each other, fused together, or just non-well organized to be individually identified. Those same contours comprehend more than one lumen. The hierarchy can only give us the presence or absence of child contour and in any case the number of it. To anticipate the issue, the approach focused only on contour with parent contours. These contours went through some size and color selection to ensure the recollection of lumina while discarding artifacts. As a matter of fact, with close glands, stroma in between glands were considered as child contour as in Figure 2.33.

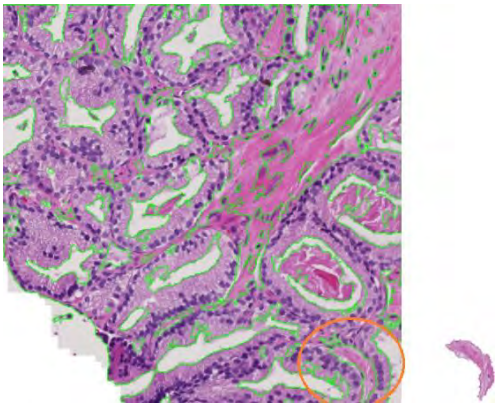


Figure 2.33: Artifact considered as lumen by the contour hierarchy

Identification of edge lumina

The previous methodology was performed for any lumen that is not placed on the edges. When the lumen was placed on the edge, its contour was not considered a child contour, such as in figure 2.34. To solve this issue, the glands and lumen were considered as one, by enclosing all their points into the same plane.

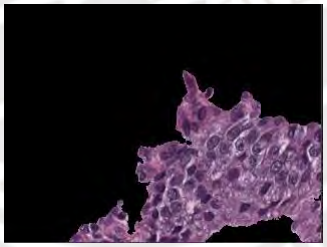


Figure 2.34: Segmented gland with lumen on the edge

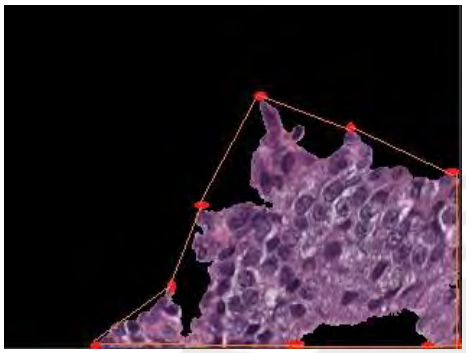
Each point of the gland's contour was connected to each other with segments. The Convex Hull is defined as the simplest closed area with minimum perimeter containing the set of points by:

$$\sum_{j=1}^{N} \lambda_j p_j : \lambda_j \geq 0 \qquad (2.3)$$

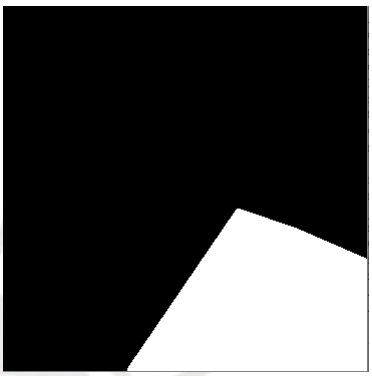
for all j and

$$\sum_{j=1}^{N} \lambda_j = 1 \tag{2.4}$$

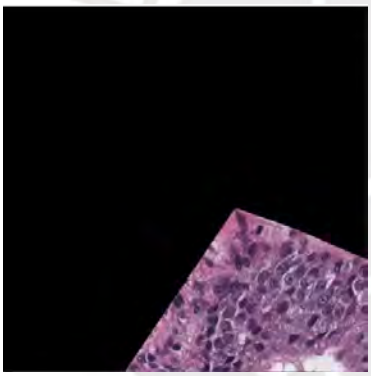
where λ_j is a parameter to be defined and p a set of points.



(a) Simplest closed area with the minimum parameter for the set of points corresponding to the contour of the glands



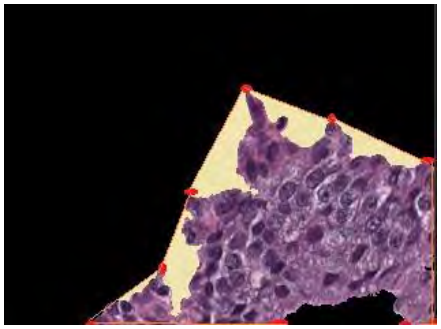
(b) Convex final mask envelope



(c) Convex final glands RGB envelope

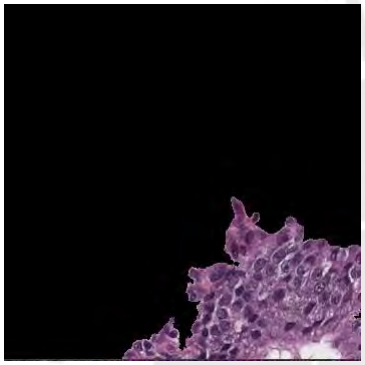
Figure 2.35: Convex closure of the edges mask form external points

When creating the Convex Hull, part of the stroma was incorporated in the envelope, as seen in 2.36a. With the intent to retain exclusively the lumen and discard stroma, the convex hull 2.35c and actual glands were subtracted 2.34, and add a size and color conditions were applied to only retain the lumen.



(a) Stroma incorporated in convex hull envelope

(b) Subtraction results consisting of stroma and lumen



(c) Final glands result with lumen

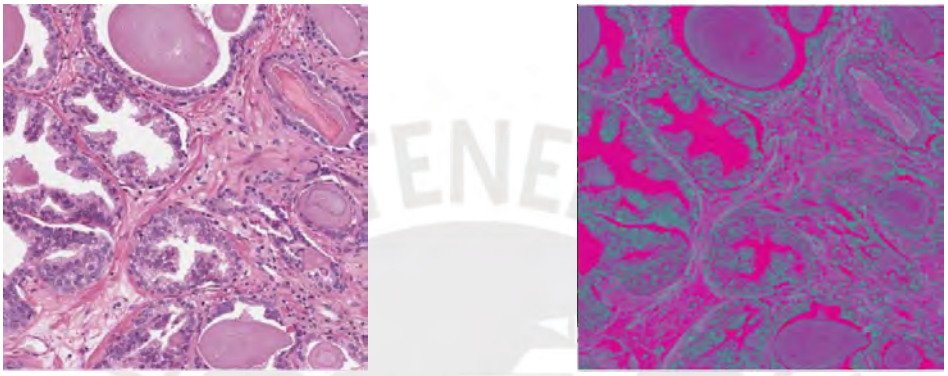
Figure 2.36: Detection edge lumen from edge masks by subtracting convex envelops to ROI and applying color condition

2.3 Separation of merged glands

At this point, the nucleus, glands, and lumen were properly detected and segmented. However, we were still facing one issue. In gland number 3 in figure 2.19 and image 2.21, the lumen can be clearly seen as the center of the glands. However, because of the proximity of other irregular glands, the algorithm considered them both part of the same gland. To resolve the issue, the glands were "reconstructed" via various methodologies.

2.3.1 Pre-processing of the masks

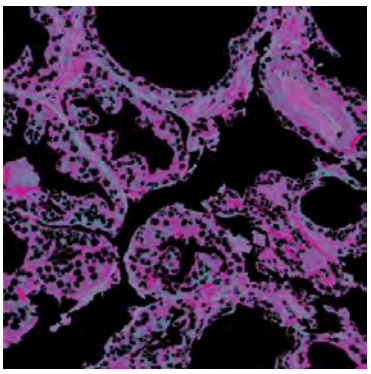
Segmented masks can contain nuclei, and cytoplasm but also in some cases stroma. To clean up the mask, the image was first converted in HSV, and the hue saturation value and color space were modified to solely operate with color intensity, disregarding color components.



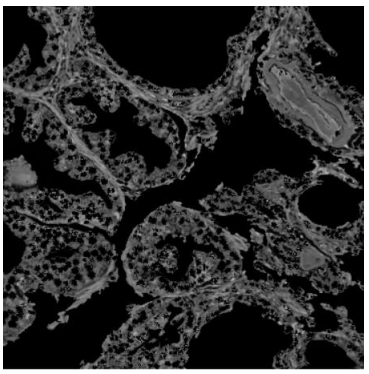
(a) Image in RGB color space (b) Image in HSV color space

Figure 2.37: RGB color space conversion to HSV

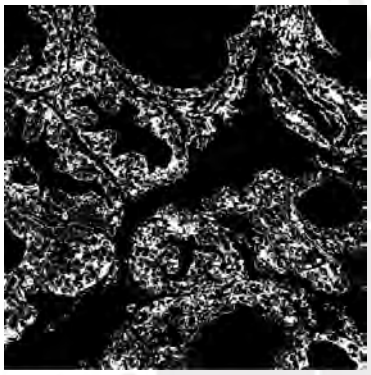
To focus on eliminating the stroma in the gland's region; the nuclei, and lumina were subtracted, and pinkish color pixels in the second channel of the image were deleted as seen in Figure 2.38b, given that stroma is enhanced with pixels of bright intensity. Pixels above 0.35, the color intensity closest to the pinkish color representing the stroma, were discarded, to generate the final mask, then applied to the HSV and RGB color space image.



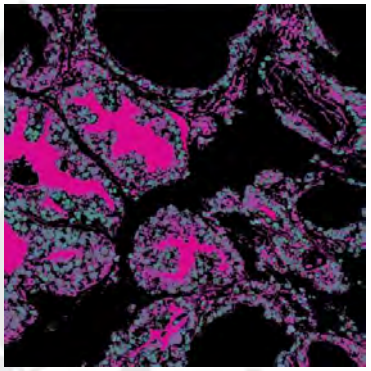
(a) Image in HSV color space without segmented nuclei and lumen



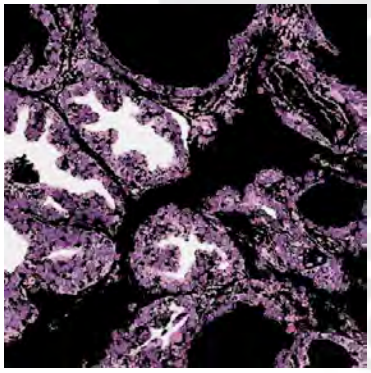
(b) Image in HSV color space from 2nd channel



(c) Mask with filtered stroma



(d) Final HSV image with filtered stroma



(e) Final RGB image with filtered stroma

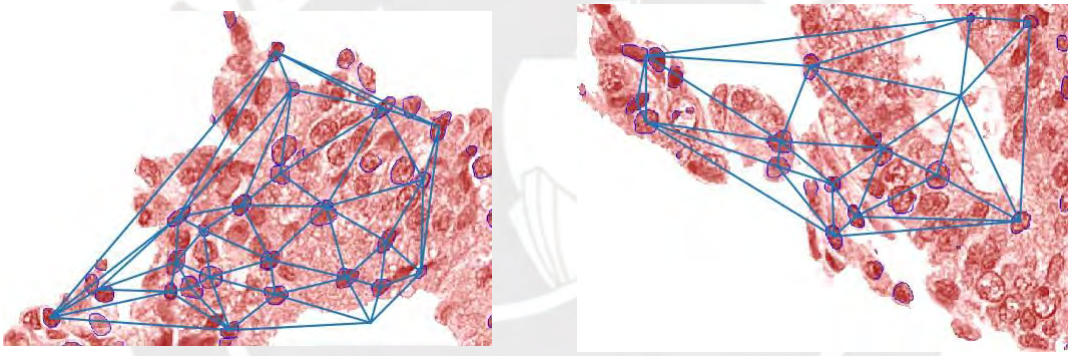
Figure 2.38: Pre-processing methodology to remove stroma from ROI with color constraint on HSV color space

The pre-processed image is then used in the following steps. Nevertheless, for demonstration purposes, the following results are displayed on non-pre-processed images.

2.3.2 Delaunay's triangulation to reconstruct the glands

Our hypothesis relies on the evidence that the glands are organized around a lumen followed by a layer of nuclei. The center of the perceived lumen was linked to the center of the nuclei detected on the image, via Delaunay's triangulation. If the nucleus had direct contact with the center of the lumen, then it was kept and assigned as part of the nuclei layer.

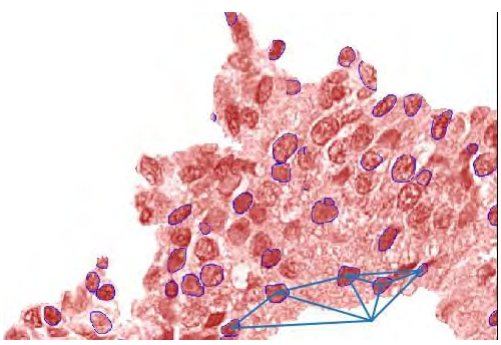
Delaunay's triangulation connects each center of the elements via a vector as shown below and stocks information as the indices of each triangle's point, the coordinates of each cell, and the indices of neighboring triangles (Cazals & Giesen, 2004).



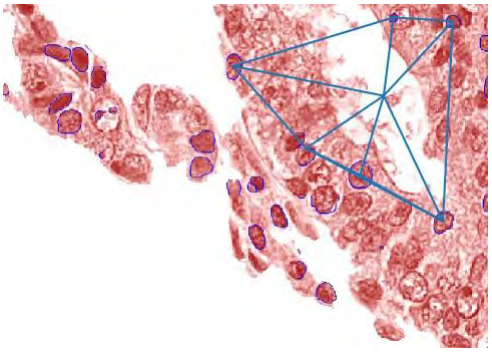
(a) Gland number 1 and its surrounded nuclei (b) Gland number 3 and its surrounded nuclei

Figure 2.39: Delaunay's triangulation between the lumen center and the surrounding nuclei

As said in the section above, in order to reconstruct the glands, only the nuclei directly connected to the center of the lumen were kept, and considered as the external layer of the gland. As a matter of fact, in prostate cancer phenotypes unlike other organs, a single layer of nuclei surrounds the lumen. Therefore, only the direct segments from the center of the lumen to the center of the closest nuclei are needed. When there is more than one layer of nuclei, iteration of this methodology could be applied, if necessary.



(a) Gland number 1



(b) Gland number 3

Figure 2.40: Delaunay's triangulation between the lumen center and its connected nuclei

The aforementioned approach would perform well for circular glands, while elliptical ones might not be reconstructed properly, such as in 2.41.

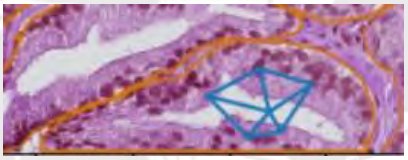


Figure 2.41: Delaunay's triangulation between the lumen center and the surrounded nuclei for non-circular gland

Rather than selecting the center of the lumen as the starting point of the Delaunay, several points served as points of departure of the Delaunay, evidenced by the corner of the lumen.

Delaunay's triangulation from Harris corner to reconstruct the glands

As stated above, employing the center of the lumen as the Delaunay starting point, might not work properly on non-circular glands. As an alternative, the corners of the lumen were evaluated to reconstruct the glands. Harris corner slides a window over the pixel, to verify a sudden shift in intensity in all directions, considering that a corner is the junction of two vectors (Vino & Sappa, 2013).

Figure 2.42: Corner detection in the lumen by Harris corner detection

All corners detected by the algorithm were linked to the nearest and directly connected nuclei by Delaunay's triangulation.

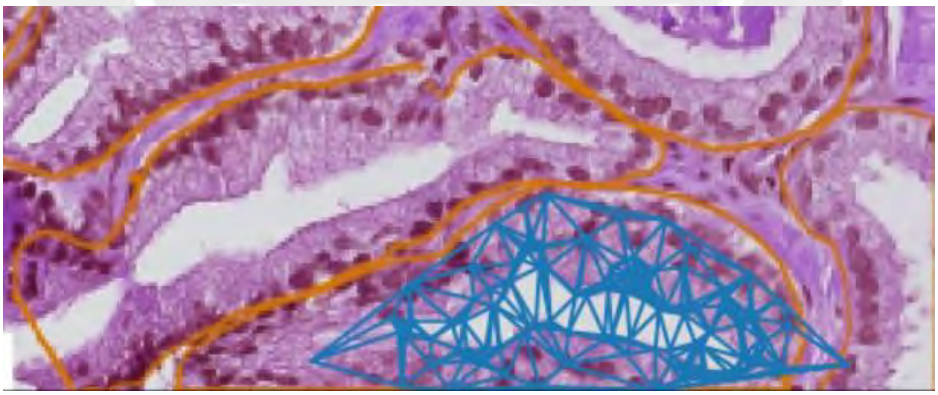


Figure 2.43: Final Delaunay's triangulation from the corner of the lumen to the directly connected nuclei

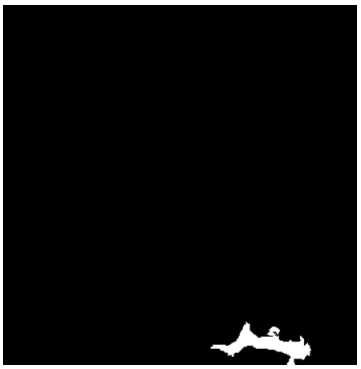
Application of convex hull to the Delaunay's triangulation

With Delaunay's triangulation, the reconstructed contour of the gland would stop at the center of the nuclei directly connected to the center of the lumen. Nevertheless, to consider the nuclei as a whole, the convex hull is created taking as coordinates, the contour of the lumen, and the center of the nuclei selected by Delaunay's triangulation.

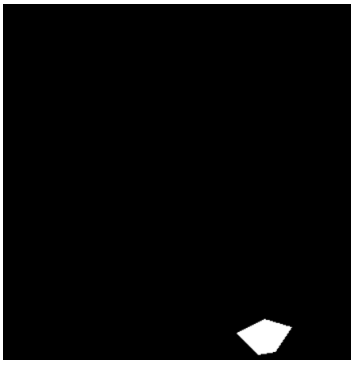
(a) Convex mask envelope from Delaunay's triangulation

(b) Convex envelope from Delaunay's triangulation

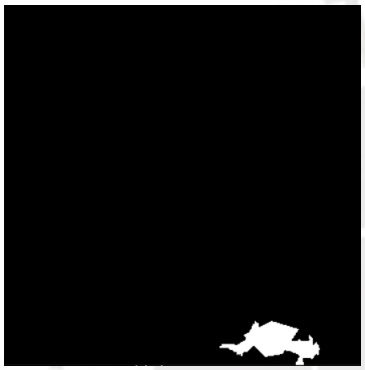
Figure 2.44: Result from applying convex to the Delaunay's triangulation's results



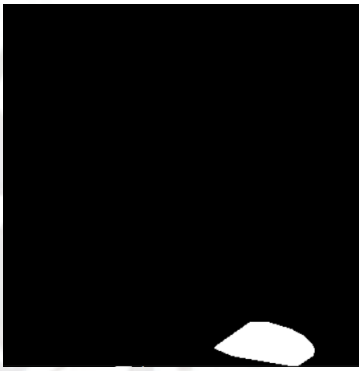
(a) Mask of the lumen for the second gland



(b) Convex envelope from Delaunay's triangulation for the second gland



(c) Lumen and convex hull envelop added for the second gland

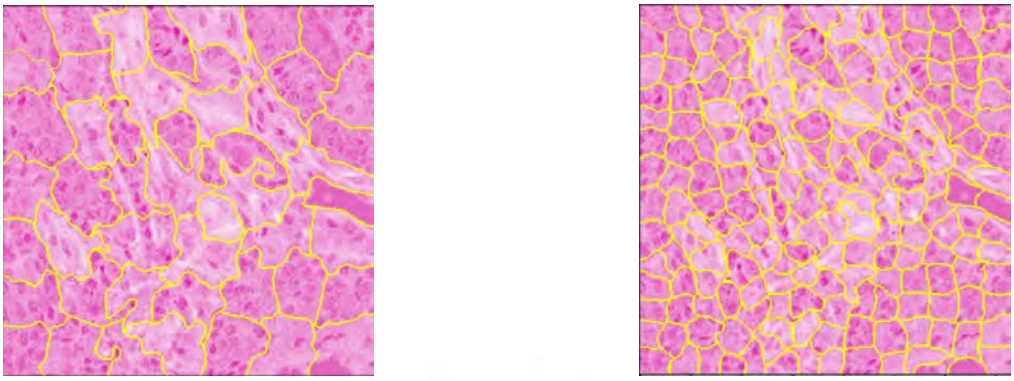


(d) Final convex envelope from the added masks for the second gland

Figure 2.45: Result from applying convex hull to the results of Delaunay's triangulation for the second gland

2.3.3 Simple Linear Iterative Clustering

The application of convex hull retrieves major information as part of the glands in comparison to Delaunay's triangulation. Nevertheless, it enhances the geometrical shape of the contour. With the aim to increase the accuracy of the gland reconstruction, while maintaining a more natural shape, the Eosin-stained slides were segmented according to their pixel values via Simple Linear Iterative Clustering, SLIC (Chen, Zhang, & Zhang, 2017). By applying the SLIC on Eosin stained slide, glands were highlighted, hence improving pixel clustering.



(a) SLIC on Eosin stained with 50 segments

(b) SLIC on Eosin stained with 200 segments

(c) Superposition of convex hull envelop with SLIC on Eosin stained with 50 segments

(d) Superposition of convex hull envelop with SLIC on Eosin stained with 200 segments

(e) Final gland segmentation

(f) Final gland segmentation

Figure 2.46: SLIC segmentation

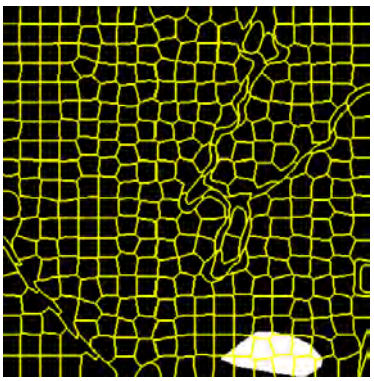
SLIC clusters pixels with similar intensity and proximity in the image. SLIC algorithm takes as input an image in LAB color space, in addition to the fitting number of approximately equally-sized multi-segments, also called superpixels K. The algorithm starts with K cluster

centers, which will move to the lowest gradient position in a 3x3 neighborhood computed as:

$$G(x, y) = I(x + 1, y) - I(x - 1, y)^2 + I(x, y + 1) - I(x, y - 1)^2 \quad (2.5)$$

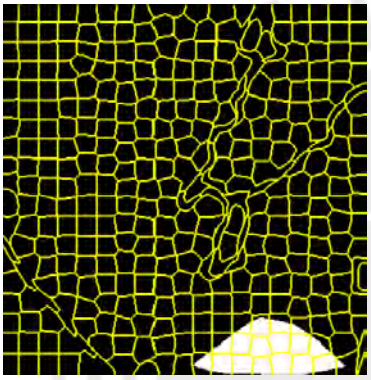
with $I(x, y)$ corresponding to the pixel in LAB color space at position (x,y). The pixel is associated with the closest cluster center, then new centers are calculated. This exact process is iterated until convergence (Chen et al., 2017).

The convex hull envelope from the previous section was superposed to the SLIC segmented slide. Each cluster that enclosed part of the convex hull shape was hence considered as part of the glands. According to the SLIC algorithm, images were separated into N segments. A lower SLIC might add part of the stroma in the gland contour, while a higher SLIC might exclude important information.



(a) Superposition convex hull and SLIC from approach 1 with center of lumen

(b) Final gland contour with approach 1



(c) Superposition convex hull and SLIC from approach 2 with corner of lumen

(d) Final gland contour with approach 1

Figure 2.47: Final gland segmentation superposing convex hull envelope and SLIC

The optimal N label was calculated according to the contour emitted by the oncologist on a set of slides and approaches compared to ground truth.

2.3.4 Geodesic Active Contour

Unlike the previous idea of engaging the closest nuclei to the lumen to reconstruct the gland, this time, the intensity of the pixel surrounding the lumen by Active Contour was evaluated.

Active Contour, by contrast to SLIC, is a supervised method since it requires a point of departure to construct a closed contour for a region of interest. By definition, active contours

are active models operating under external and internal forces to determine their curvature. Geodesic active contour models rely on Curve shortening flow, where the curve will expand by moving its point perpendicularly (Caselles, Kimmel, & Sapiro, 1997).

Figure 2.48: Curve shortening flow from convex hull

New points were included in the contour by level set function, employing the value of the gradients defined by the change in the color or intensity in an image. The model was then started from the convex hull coordinates, or lumen, and evolved until reaching a pixel with a value below the specified threshold. Glands were delimited by their layer of nuclei, which were considered as the threshold to stop the active contour. Biopsy images are complex, and have significant intensity variation, which could have an impact on the detected final contour by the geodesic model. Therefore, to increase edge contours, while smoothing internal pixels, the RGB contour masks were dilated and eroded.

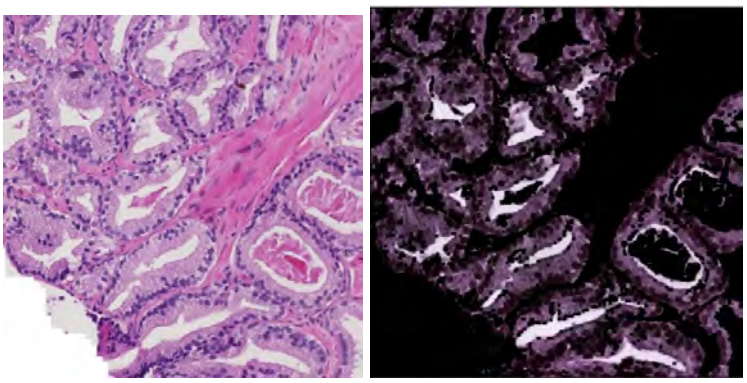
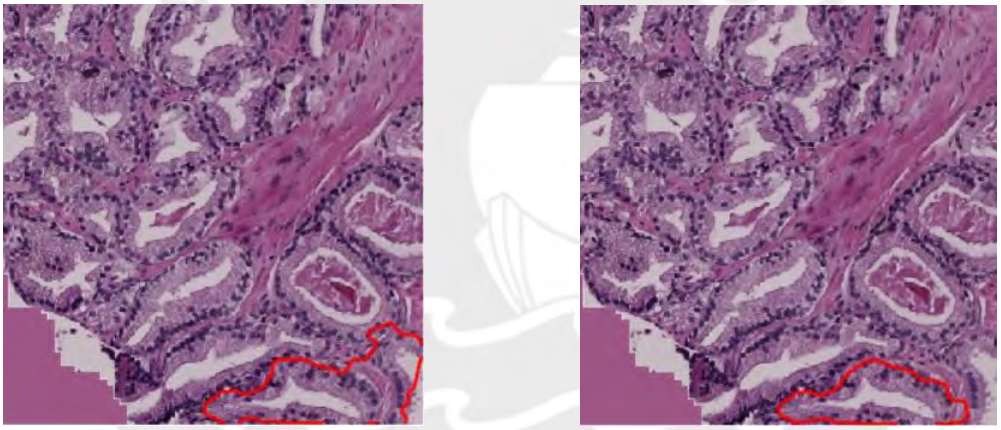


Figure 2.49: Dilatation and erosion

Then, different thresholds were tested until a suitable fit was found.



(a) Active contour with a threshold of 0.15 (b) Active contour with a threshold of 0.25

Figure 2.50: Effect of threshold variable on active contour segmentation

Notwithstanding, the contours were calculated on the dilated and eroded representations of the image, hence the final contour on the non-modified RGB slide might not include properly the layer of nuclei. The edge of the active contour was then superposed to the pre-processed SLIC segmentation. The average color intensity of the SLIC region was retrieved, Figure 2.51, and segments with an average pixel intensity below 0.20, corresponding to purple, the color of stained nuclei, and crossed through by the active contour were kept and added to the final ROI.

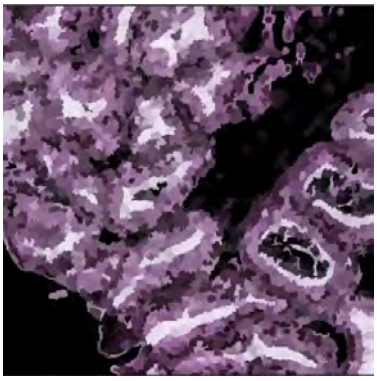


Figure 2.51: Visualization of SLIC segmentation according to segments average

2.3.5 Lumina false positive

When the cancer is advanced, glands are disorganized and should not be considered as one, but our methodology considered that each lumen is attached to a gland. As seen in 2.52, lumina are highlighted and an active contour departed from it to delimit the non-existent glands as in 2.53.

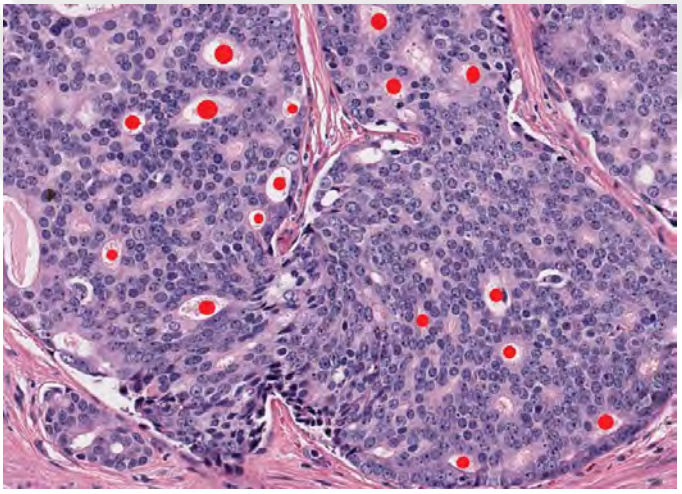


Figure 2.52: Region of interest classified as lumen

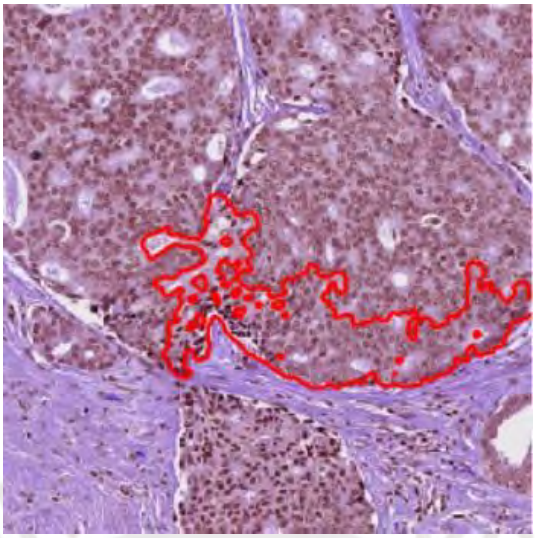
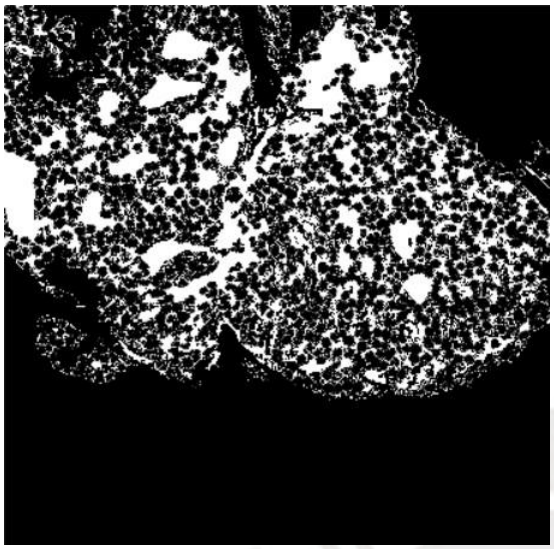


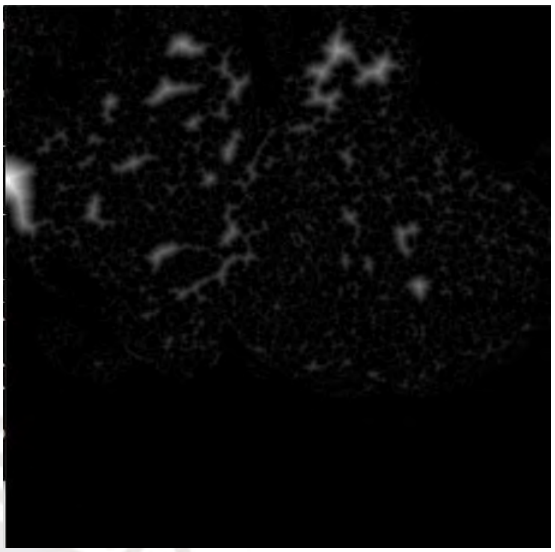
Figure 2.53: Active contour from lumen

The main reason comes from the fact that from the segmentation of the lumen to the reconstruction of the lumen, only pixel intensity was employed to classify the elements and not the texture. The main indicator of the gland is the presence of cytoplasm, when present in between the lumen and the layer of nuclei the contour should be detected, otherwise the lumen should be discarded.

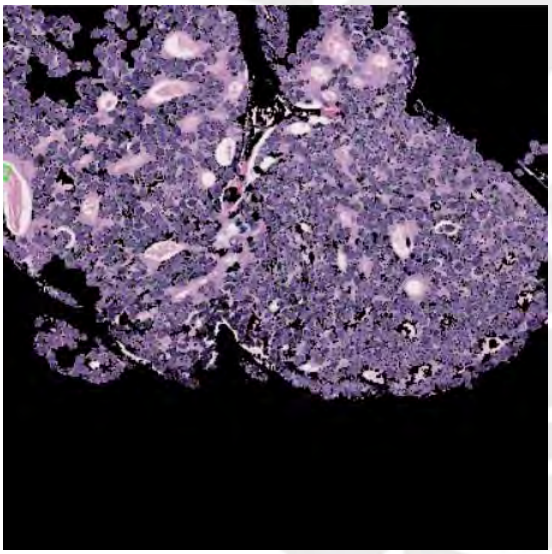
The grey-level co-occurrence matrices could be used to detect the homogeneity and dissimilarity of the tissue around the lumen. Nevertheless, in some cases, the cytoplasm layer is too thin to extract this specific information. Instead, the segmented nuclei were subtracted to the segmented gland, and then the Euclidean distance was calculated as done previously 2.2.3. The bright pixels corresponded to the central regions of the glands when nuclei were organized as an external layer, while darker pixels coincided with the nuclear regions. From the normalized distance transform, pixels with a value above 0.8 were kept and labeled as lumen. In the case that a layer of nuclei does not surround the lumen, the result from the distance transform should not coincide with any lumen identified previously.



(a) Contour mask without nuclei



(b) Distance transform



(c) Area labelled as lumen from distance transform in green

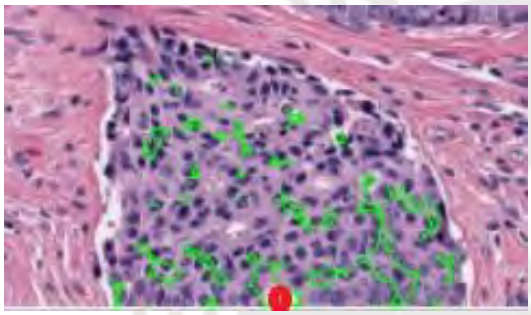
Figure 2.54: Lumen classified as lumen from distance transform

Only the area concurring with the segmented lumen was kept and passed through an active contour. In 2.54c, from the distance transform, only 1 region of interest was kept, but did not coincide with the lumen detected in 2.52 so it was discarded in the active contour.

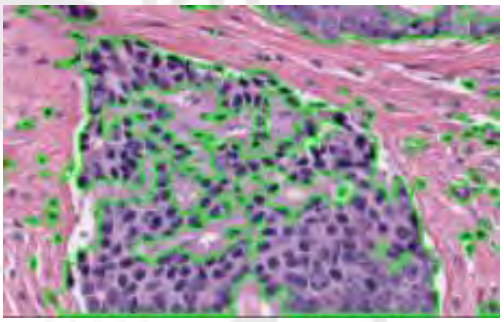
For this specific contour, no glands were identified, even though the lumen was recognized,

which provided decisive data as to the cancer stage of the patients.

A different condition was included to retain or discard the lumina, which consisted of comparing the value of the area of the segmented lumen, and the area of the lumen labeled by distance transform. When the distance transform labeled lumen had an area superior to the area of the classified lumen, then it was rejected. In 2.55a, one area coincides with the lumen in 2.55b, however, it was rejected as the total area was superior because part of the tissue was erroneously labeled as lumen.



(a) Area contour labeled as lumen from distance transform



(b) Area contour classified as lumen

Figure 2.55: Lumen area requirements

The common area was therefore retained and tested through Delaunay's triangulation and active contour.

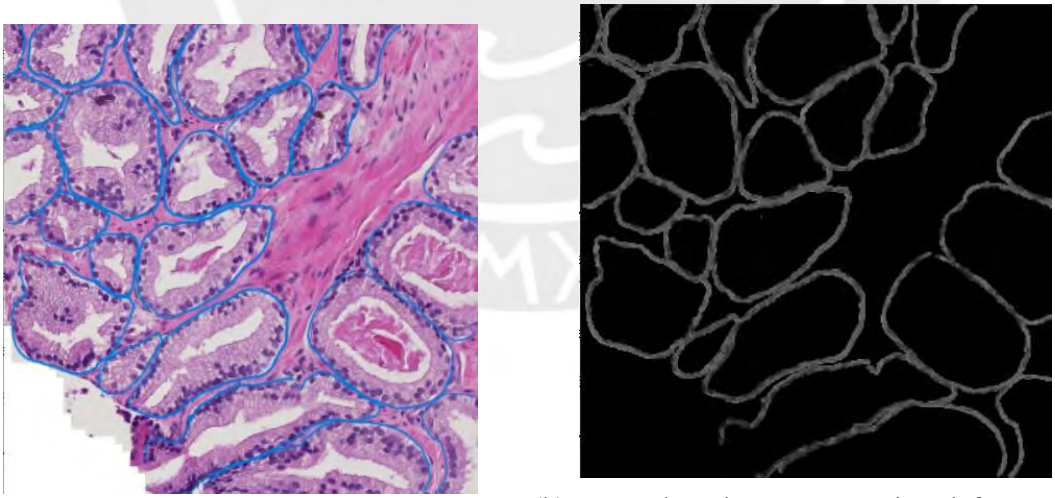
2.4 Evaluation of the reconstruction of glands from ground truth contour

With the aim to evaluate the proposed methodology, the segmentation was compared to contours defined by an experienced oncologist.

The ideal number of segments in the SLIC algorithm, and the point of departure of Delaunay's triangulation were also tested by checking false positive (FP), true positive (TP), false negative (FN), and true negative (TN) from the predicted segmentation to the ground truth contour. True positive corresponds to the area segmented by our approach as glandular, which was also considered glandular by the ground truth. Equivalently, the true negative was defined as the area labeled as background by the approach proposed in this thesis and the ground truth. Contrariwise, false negative was equivalent to the area not segmented by the previously described approach but defined as part of the gland by the ground truth, and false positive was the area predicted as glands, only by the proposed approach. Numerically, the intersection over union defined by the following expression

$$IoU = \frac{TP}{TP + FP + FN} \tag{2.6}$$

expresses the accuracy of the segmentation. A IoU of 1, displays perfect segmentation between ground truth and prediction.

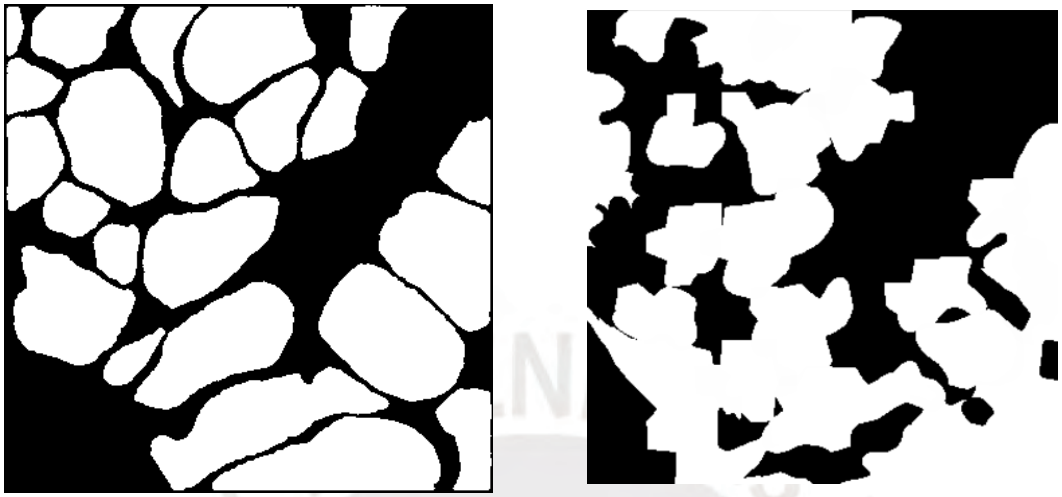


(a) Ground truth contour in RGB image

(b) Ground truth contour retrieved from HSV color space

Figure 2.56: Ground truth contour retrieved from a oncologist

The ground truth mask from 2.57a was eroded for greater visualization of adjacent glands.



(a) Ground truth mask (b) Example of predicted mask

Figure 2.57: Ground truth mask from oncologist's contour compared with mask from aforementioned methodology

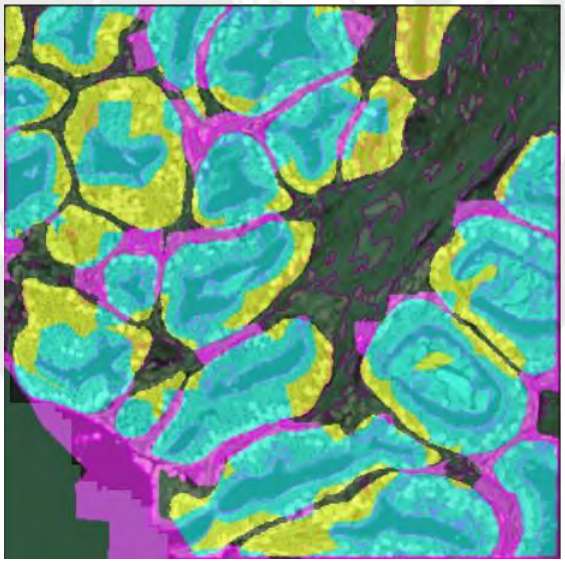
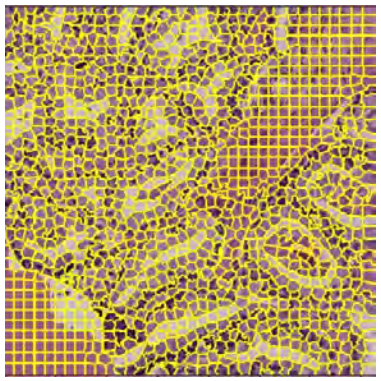


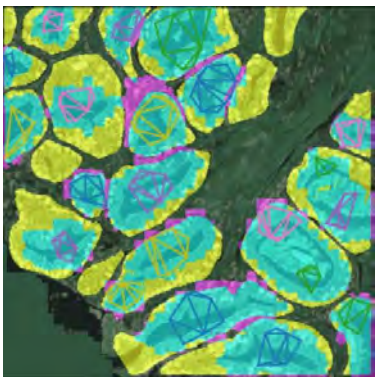
Figure 2.58: Example of confusion matrix visualization, where magenta corresponds to FP, yellow to FN, cyan to TP and black to TN

2.4.1 Evaluation of the reconstruction of glands from the center of the lumen and SLIC application

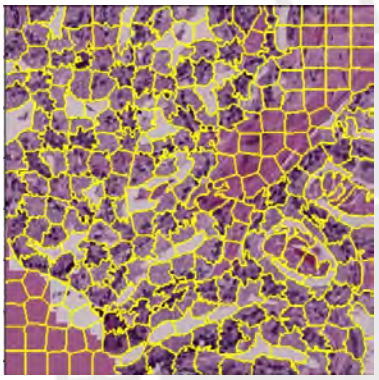
The first approach was to use the center of the lumen and add to the envelope the nuclei directly connected to it. The images were segmented into 1400 sub-segments, with a superpixel algorithm. As displayed in 2.59b, most of the glands were not correctly segmented, as too narrow, and not enough nuclei were taken into account, with an IoU of 0.40. A decrease in the number of segments in the superpixel algorithm could resolve the issue, as the wider segments comprise more gland elements, exposed in 2.59c. Glands were indeed fuller, but FP increases, showing no improvement in the IoU metric, while other glands remained improperly segmented. The major problem persisted by the lack of nuclei, clustering the glands a few centimeters from the center.



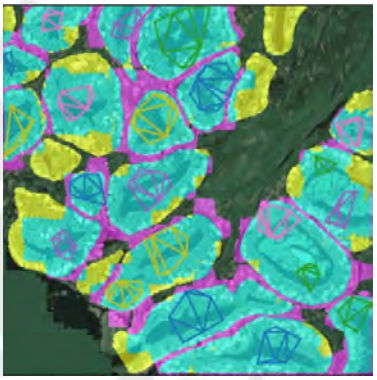
(a) SLIC segmentation with 1400 segments



(b) Confusion matrix visualization with lumen center and 1400 segments



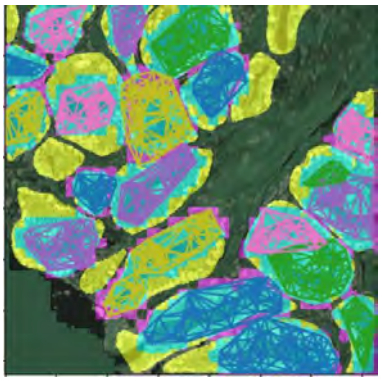
(c) SLIC segmentation with 300 segments



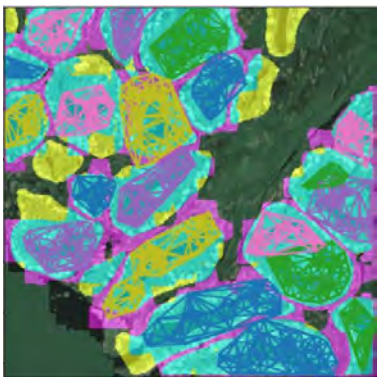
(d) Confusion matrix visualization with lumen center and 300 segments

Figure 2.59: Confusion matrix visualization for the approach with the center of lumen , where magenta corresponds to FP, yellow to FN, cyan to TP and black to TN

The second approach considered the corners of the lumen and added all nuclei directly connected to them. When two approaches were compared while employing 1400 segments, TP expand, indicating, more glands were correctly segmented with an IoU of 0.6. However, when the lumen was not proportionate to the glands, most of the glands were inappropriately labeled.



(a) Confusion matrix visualization with lumen corners and 1400 segments



(b) Confusion matrix visualization with lumen corners and 300 segments

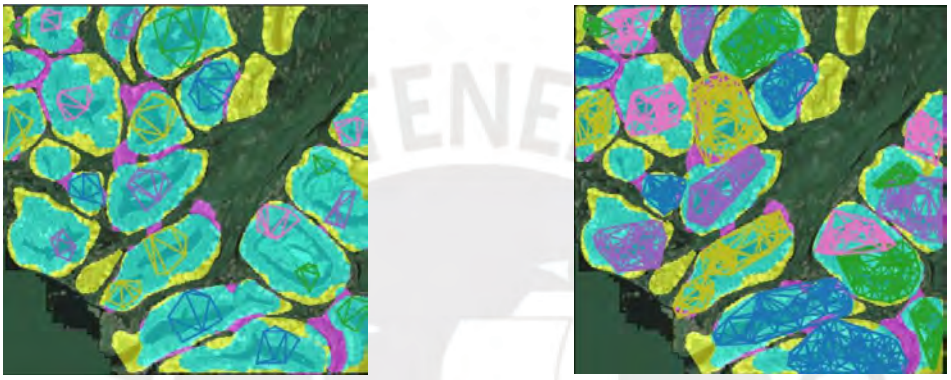
Figure 2.60: Confusion matrix visualization for the approach with the corner of lumen, where magenta corresponds to FP, yellow to FN, cyan to TP and black to TN

Indisputably, when the lumen was rather slim compared with the rest of the glands, the segmentation was not completed, as some nuclei were found in the cytoplasm in between the lumen and the layer of nuclei. The approach described in this thesis considered these nuclei as part of the external layer of nuclei and paused the segmentation in those regions, as in 2.61. Decreasing the number of segments can palliate the issue, but may increase the FP in other cases.

Figure 2.61: Nuclei considered as external nuclei layer

2.4.2 Evaluation of the reconstruction of glands with active contour

As a means to increase TP while maintaining low FP, and as an alternative to Delaunay's triangulation, an approach employing expanding active contour was applied. Different starting points were evaluated, consisting of the convex hull comprising the lumen center and connected nuclei or the corner of the lumen and attached nuclei.

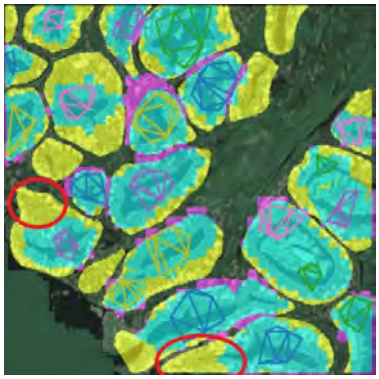


(a) Confusion matrix visualization with Morphological Geodesic Active Contours from center of the lumen

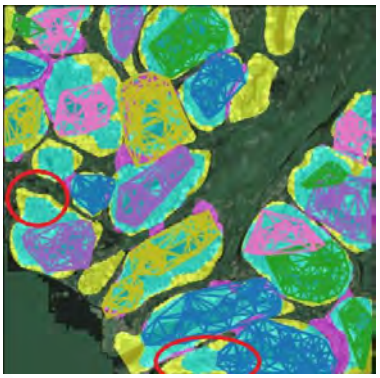
(b) Confusion matrix visualization with Morphological Geodesic Active Contours from corner of the lumen

Figure 2.62: Confusion matrix visualization with Morphological Geodesic Active Contours results segmentation

Active contour enhanced the segmentation, especially in cases of non-circular glands, as presented in 2.63a and 2.63b, while maintaining low FP with an IoU increasing to 0.71.



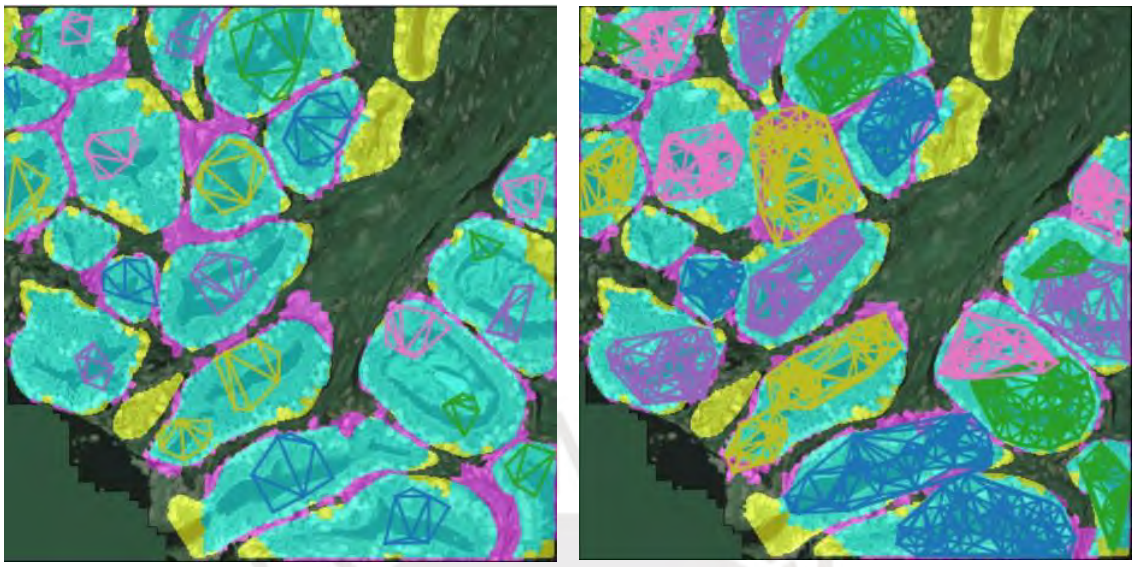
(a) Confusion matrix visualization with Delaunay and SLIC



(b) Confusion matrix visualization Result with Active Contour

Figure 2.63: Confusion matrix visualization for non-circular glands, difference of segmentation between active contour and SLIC results circled in red

Nevertheless, most of the glands were not thoroughly sectioned, a lower threshold would enlarge them but would increase the FP and include stroma in the glandular area. Alternatively, the masks provided by the active contour were superposed with the SLIC algorithm to increase TP.



(a) Confusion matrix visualization with Morphological Geodesic Active Contours from the center of the lumen superposed with SLIC 1400 segments

(b) Confusion matrix visualization with Morphological Geodesic Active Contours from the corner of the lumen superposed with SLIC 1400 segments

Figure 2.64: Final confusion matrix visualization

The final approach provides a segmentation closer to the contours obtained by the oncologist with a precision of 81% over 46 patients, 230 tiles, and an IoU of 0.87. Notwithstanding, the segmentation is not flawless as with higher proximity of phenotypical elements, glands without the presence of lumen 2.65g were not considered as glands. Categorically, the U-net architecture can detect clusters of nuclei as well as glands. In spite of this matter, when glands were too close to each other, and considered under one contour by the U-net, the methodology proposed in this work was not capable of reconstructing the gland from the non-existent lumen. A similar result occurred when the edges of the glands were not clearly defined 2.65e and 2.65a. The U-net was then unable to detect correctly the border of the lumen, overestimating or underestimating its area.

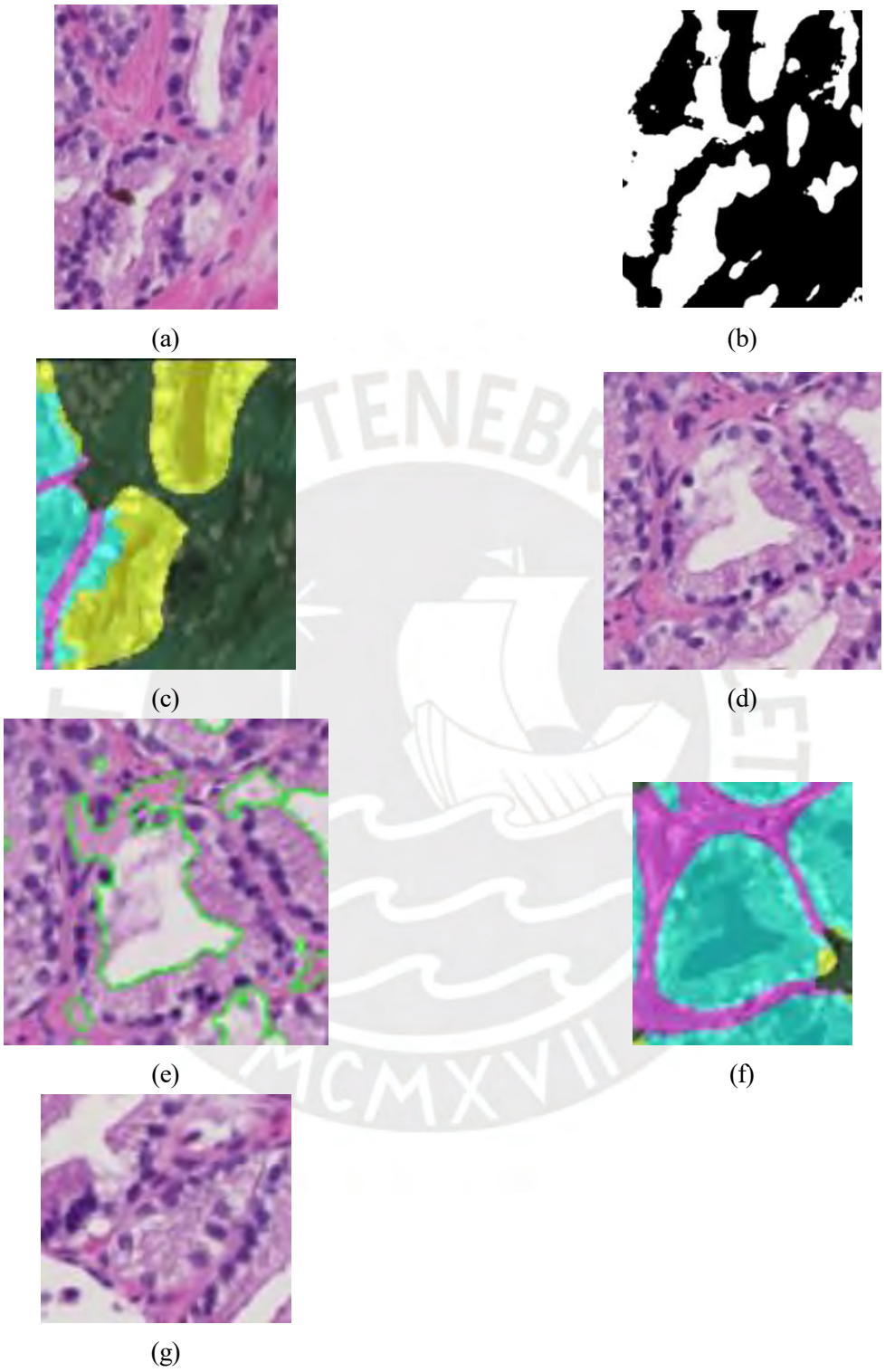


Figure 2.65: Final segmentation results

2.4.3 Spatial arrangement and cell adhesion of prostate tissues

In summary, this chapter showed evidence that the glands, the nuclei, and the lumina were successfully identified on tiles from selected datasets of WSI.

Gland organization plays a major part in prostate cancer detection and grading, as explained previously. High Gleason scored patients, present disorganized glands, as deregulated cell adhesion molecules prohibit the formation and maintenance of healthy glands. Cancer and healthy cells invade the glands, as they are not holding themselves into layers, losing their structure; hence the importance of collecting evidence of the cell adhesion deficiency, by investigating first the morphology of the glands, calculated by the variance from the distance between the center of the lumen and membrane of the glands. If the variance is low, the gland is of regular shape; on the other hand, high variance results when the shape is irregular. Other features consist of the number of glands with lumen, the number of clusters of nuclei, the ratio area of the gland/cluster to the size of the detected contour by U-net, the ratio of lumen to glands, ratio area of the reconstructed glands to the original contours, glands fused or individuals, the distance between glands, as well as the number of detected lumina discarded. When more than one lumen was observed, the convex hull strategy allowed us to calculate the density of nuclei in between the lumina. In 2.52, every single lumen was discarded, in between these lumina, the density of nuclei is high, as few stroma/vessels were present, indicating high disorganization and low cell adhesion. The density of nuclei is also computed for each contour, as the proliferation of nuclei is a significant indicator of tumor aggressiveness.

The structures of the glands are controlled by cell adhesion molecules and their extracellular matrix, ECM. ECM is composed of structural proteins such as collagen and cells with structural functions such as fibroblasts, which proffer a support system to organize the cell. Cell adhesion controls the tissue structure as it stops the proliferation of cells when other cells are detected, hence the presence of collagen peripheral to the glands. When glands are attached

or fused together, cell adhesion is damaged and the cells within the tissue are unable to detect surrounding components. When there is no stroma, nor fibroblasts in between the glands, the secretion of collagen is hindered (Frantz, Stewart, & Weaver, 2010).

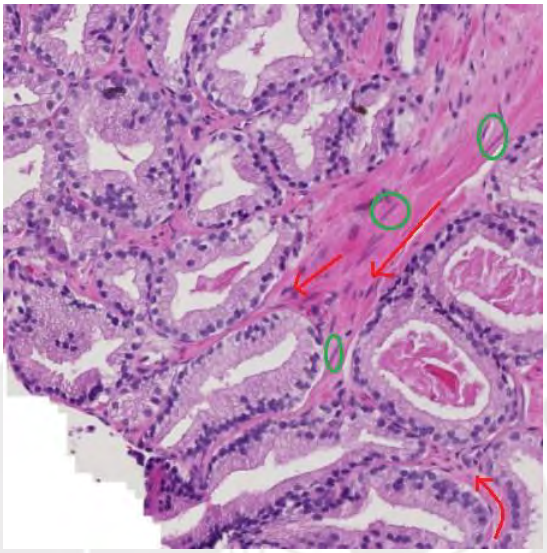


Figure 2.66: Extracellular matrix components, arrows indicate the collagen fibers and green circles are the fibroblasts

The presence of a lumen displays an organized structure and thus can provide information on the cell adhesion properties within the tissue, same as the morphology of the glands. The area of reconstructed glands was calculated and compared to the area of the original contour. Indeed, since the U-net model was trained to segment glands, a ratio too significant can prove the presence of a cluster of nuclei, and as well exhibit poor tissue structure because the cells peripheral to the membrane of the glands do not secrete the proteins necessary for the ECM. The minimum distance in between the glands provided information as to the condition of the ECM. In healthy tissue, it should have stroma in between the glands, and the presence of fibers of collagen.

For each variable, the mean, variance, minimum and maximum values were retrieved from each tile (five tiles per patient). In case one of the variables was equal to 0, the variance, mean, minimum and maximum were assigned 0. In the event that the variable was encountered only once, the variance was assigned 0, but the mean, minimum and maximum were equal. These variables were retrieved by tiles since prostate cancer is known to be a highly heterogeneous disease, and multi-cancer foci were expected.

Chapter III

Prediction of Biochemical Recurrence based on Genomic Information

While overwhelming by its numbers and complexity, genomic data can provide key information regarding the patient's present and future health. The genomic information collected is quantitative and, unlike phenotypic features, does not need to be segmented. Therefore, our efforts in this chapter focus on the evaluation of classification strategies to improve BCR prediction.

The entire TCGA dataset contained over 20 000 genes presenting the challenge of needing extensive computing power. In addition, finding a final gene signature responsible for the time of recurrence might be complicated by the noise from the 20 000 genes. In order to decrease the number of genes, the variance, the variability from the mean of each gene in the 499 patients was calculated. A low variance indicates similar expression in every patient and shows no change according to the patient's status. Therefore, during this pre-processing step, only genes with a variance superior to 1 were kept, with a final dataset of 7 800 genes. Advice from Joseph Pinto

The genomic expression also needed normalization, to model proportional change, a log2 was applied. Indeed, log2 helps in calculating the fold-change, and thus measuring which genes

are down-regulated, or up-regulated between samples.

3.1 Classification of the status of patients using fully connected neural networks

The first approach incorporates the genetic expression with the time of recurrence via a fully connected neural network, FCNN, as explained in 2.1.1. To represent spatially the FCNN in 2.6, each gene represents one input neuron of the FCNN.

The first idea was to develop an approach only employing deep learning, from the phenotype, to the genetic, using and testing various models, such as CNN, M-RCNN, U-net, hence have the FCNN linking the genomic expression to the BCR. Deep learning is relatively new in the biomedical field but shows more accurate results than general approaches. FCNN can be employed for such investigation since the volume of data is numerous, however, the risk of overfitting remains elevated. More classic approaches might miss important components, compromising the discovery of early biomarkers, while the high number of neurons and layers of FCNN can assist in discovering even the more subtle correlations.

Underfitting and Overfitting The structure of the model, as well as the shape of the dataset, will have a direct incidence on the training and testing scores. The architecture of the model depends mainly on the complexity of the dataset. If the dataset fed to the model has easy-to-recognize trends and patterns, the neural network might need fewer layers. However, a balance should be stricken with its complexity, by changing the number of neurons and hidden layers. Indeed, a simple model wouldn't enable the learning of the patterns in the training dataset, providing low accuracy and high loss score, classifying incorrectly most of the labels. To deal with underfitting, one can either increase the complexity of the neural network or increase the train-

ing time, as it may be possible that the neural network has just not found the optimal weights yet to permit correct classification. On the contrary, a highly complicated model could lead the neural network to overfit by giving a high accuracy score on the training dataset, but a low score once provided with a new dataset. Due to the complexity of the model, it may learn explicitly the pattern from the training dataset, and will not be able to generalize it. Another cause of overfitting might come from the lack of data, with a small dataset, in the training process, the model can get used to the data, thus predicting correctly its outcome without having properly learned any trends. The main solution relies on a more diverse dataset by adding instances.

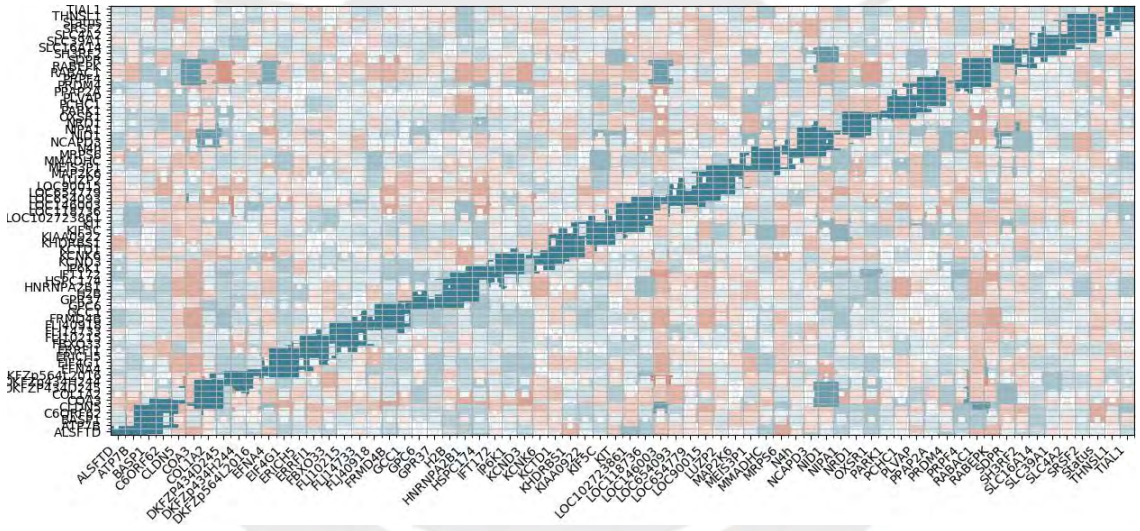


Figure 3.1: Heatmap generated from public dataset representing the correlation between each gene and the status of the patients

The data used comprised 500 patients, and their genetic expression, by calculating their correlation to the status, it is obvious that the low correlation between them will demand a model of higher complexity. Another issue can surge from an uneven dataset. Indeed, out of 500 patients, only 99 patients present BCR, and from these 99 only 73 patients have complete clinical data. One way to address this issue would be to oversample the minority class, by

adding copies of data from recurred patients with white noise.

One of the simplest methods to check if the model is either overfitting or underfitting is to investigate the accuracy and loss results for a testing set. An overfitted model will provide excellent results in the training process, with an accuracy close to 90% while maintaining low loss, but will be incapable of deducing correctly a case never seen, thus poor accuracy and high loss in the testing process. On the other hand, an under-fitted model won't be performing well in the training process, with an average accuracy of 45%.

K-fold Cross-Validation To train the model, the dataset needs to be separated into two groups, the training batch, and the testing batch. Thus, the training of the model is effectuated with the training dataset and then tested on new cases. A random split in the cohort is one course of action. However, the model can get used to the training dataset and thus provoke overfitting. K-fold cross-validation consists of separating the data set into k-folders, one fold is used as the testing data set while the rest is the training one, shifting at each iteration, preventing the model to familiarize with the training dataset.

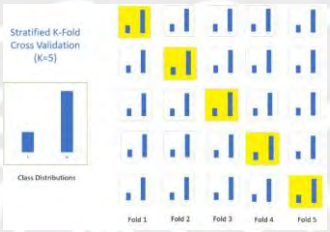


Figure 3.2: Stratified K-Fold Cross Validation (*Improve Your Model Performance using Cross Validation*, n.d.)

Addition of drop out Adding drop out to the model can be a solution to better address the issue of overfitting. As a matter of fact, during the training phase, dropout will randomly ignore N number of neurons. By doing so, the weights are learned by a fraction, in each training iteration, instead of altogether (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov,

2014).

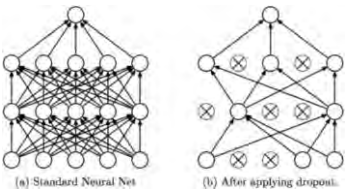


Figure 3.3: Process of drop out from (Zaidan et al., 2019)

3.1.1 FCNN model

The 500 patients and their respective, 7800 genes were fed to the model. When using a FCNN, the main two arguments to adjust, are the number of layers, and the number of neurons in the layers. Starting with a rather modest model composed of three dense layers with 100 neurons each and the ReLu activation function, which according to literature (Banerjee, Mukherjee, & Pasiliao, 2019), is the default activation, providing overall better results with easier training. As for the dataset, it was split into 10 folds. After training the model for a while, no correlation was found from the inadequate complexity of the model, as the accuracy did not surpass 45%. A decision was then made to increase the number of layers and neurons. Since more than 7 000 genes were fed to the model, each layer had around thousands of neurons. These two variables were increased until reaching adequate accuracy and loss. Tests were effectuated, interchanging the number of layers and neurons until reaching acceptable results. It was achieved using eight dense layers.

1. number of neurons in the 1st layer:1900
2. number of neurons in the 2nd layer:1700
3. number of neurons in the 3rd layer:1000
4. number of neurons in the 4th layer:1000

5. number of neurons in the 5th layer:700
6. number of neurons in the 6th layer:800
7. number of neurons in the 7th layer:500
8. number of neurons in the 8th layer:100

Nevertheless, given the complexity of the model, and the faint correlation between the genes it was overfitting. To resolve the issue, the data set was evened out by appending patients with BCR as white noise to the pre-existent ones. Nevertheless, the issue persisted. Another strategy was to introduce a dropout after each layer, but overall, the model operated poorly in testing patients. The loss and accuracy curves can be seen in 3.4 after adding the dropout layers. The overfitting can definitely be observed in the training phase, as the loss rapidly decreases, while the accuracy is close to perfection. Other similar datasets from prostate cancer patients, could also be added to the original dataset, to overcome overfitting.

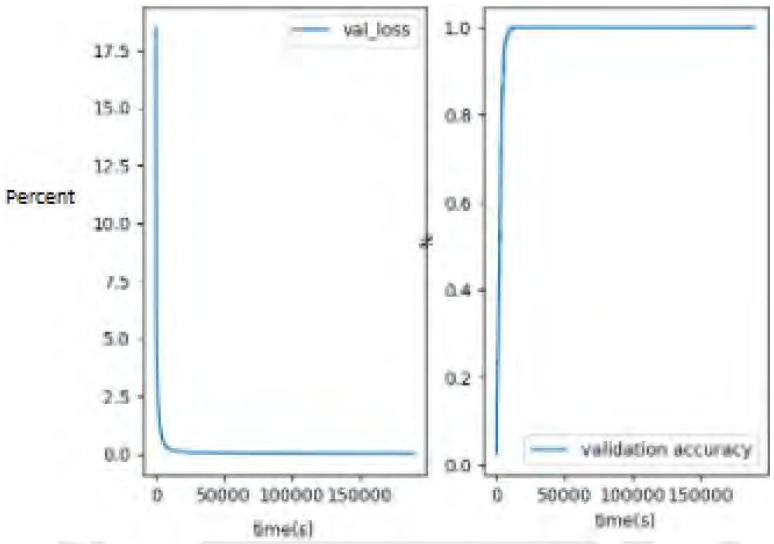


Figure 3.4: Loss and accuracy curves in the training phases with 10 folds cross validation

Based on these results, we concluded that FCNN and deep learning approaches are not the most suitable method to retrieve the genes and their importance to predict patient status. Often mentioned as a black box, another difficulty presents in this methodology is to retrieve and analyze the role of each of the genes in the prediction. Indeed, various layers and neurons are used, thus the complexity to discern which genes played a dominant role in the final classification might not be evident. Training an FCNN can also be time-consuming since thousands of iterations are made during the training process.

Although, the weights associated with each predictive gene could be retrieved with Local Interpretable Model-agnostic Explanation, LIME, and the training process reiterated. LIME algorithms can be applied to any machine learning model, and are mainly used to expand the insight into the underlying mechanics of our models. By perturbing the input of the model, the predictions will change. LIME will inspect exactly how the prediction adjusted from the original one to understand its mechanics (de Sousa I, Vellasco, & da Silva, 2019).

The LIME algorithm was applied to the FCNN results, to get a glimpse of the genes with greater importance in the classification of the patients. In Figure 3.5, a perturbed set was fed to the trained model. Group 0 coincides with patients with BCR and group 1 without BCR. The perturbed patients were classified as with BCR since the DKK1 gene expression was below -1.23, and HFSX1 above 0.53. The numbers close to the bar represent the weights associated with the genes. These two genes were consequently identified as responsible for the recurrence of the cancer. The LIME algorithm also displayed why the patients were not classified as without risk of presenting BCR.

Figure 3.5: Result from the LIME algorithms explaining the result of the prediction from the model

In summary, the use of FCNN is not a suitable strategy to interpret the action of each gene to define patient status. Nevertheless, LIME may provide hints as to the role of some genes in the prediction. By using deep learning, the prediction of the time of recurrence might improve on new datasets, but in the end, the weights associated with each input neuron will be unknown.

Corresponding results could give primordial information as to the role of the genes in the prediction. However, as explained above, the model was overfitted, due to the complex corre-

lation between genes and the status of patients. Moreover, the model had a large computational time, with quite a prolonged training process (around 2 weeks with a CPU). Therefore, this approach was discarded.

3.2 Classification of the status of patients using common classifiers

The following section focuses on three classifiers: logistic regression, decision tree, and Support Vector Machine, SVM. These classifiers are robust techniques used in the classification of data and are employed according to the shape of the data. Incontestably, the objective of each classifier is to notice the decision boundary which separates the classes, while logistic regression can only produce a linear decision boundary, decision trees generate a geometrical shape, and SVM a circular one.

Logistic regression model Different types of logistic regression can be introduced. In the first step of our approach, we focus on binary logistic regression in order to classify the patient as recurrent or non-recurrent. Logistic regression, similar to the linear regression represented by $Z = wx + b$, uses an equation as representation (Sperandei, 2014). Each input value is combined by weights (w and b in the equation) to predict the output y. In the case of logistic regression, the model is represented by:

$$y = \frac{e^{(b0 + b1 * x)}}{1 + e^{(b0 + b1 * x)}} \tag{3.1}$$

with y the output value, x the input one, b0 and b1 the associated weights. The main goal is to determine the b weights in order to properly predict the output value. The cost function is defined by the sigmoid function explained in section 2.1.1 and illustrated in figure 2.3:

$$h(x) = sigmoid(y) - log(h(x))ify = 1 - log(1 - h(x))ify = 0 \quad (3.2)$$

The logistic regression can also serve in classification questions, by replacing nominal values with numeric ones.

Decision Tree model A decision tree model is created by learning decision rules from the data features (Song & Lu, 2015). A decision tree is therefore composed of: - Root node: first split, which decides in how many set the data should be divided

- Splitting: Dividing a node into sub-nodes
- Decision Node: if sub-node is split into further sub-nodes or not
- Leaf: Terminal node predicting outcome.

The process of splitting the subset is repeated until the subset at a leaf has the same value as the target value. To determine the most accurate node, the cost of each split is calculated. In other words, at each split the predictive label is compared to the actual one with the following cost function:

$$Costfunction = sum(y - prediction)^2 \quad (3.3)$$

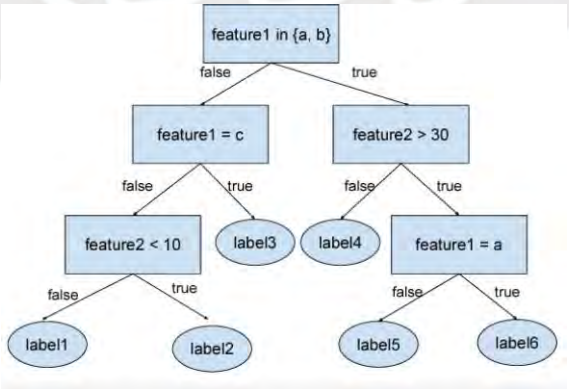


Figure 3.6: Decision Tree architecture,(Tike & Tavarageri, 2017)

The attribute selection measurement can be tested through iteration, with different options available including Entropy, Gain Ratio, or Information Gain. By default, the model decides on the attributes based on the Gini index. A shift in the attributes might lead to higher accuracy in the prediction. Another criterion that can affect the classification is the increase in the maximum depth of the tree. This argument can be tricky, as a significant number of nodes can lead to overfitting, but a low one might underfit.

Support Vector Machine model Support Vector Machine or SVM model proposes an approach with less computation power than logistic regression or decision tree approaches. The main point of the SVM is to find a hyperplane in a dimension that will distinctly classify the data. The hyperplane to be found needs to have the maximum margin, meaning the maximum distance between the different classes of data (S. Huang et al., 2018).

The loss function maximizes the margin by the following equation:

$$c(x, y, f(x)) = (1 - y * f(x)) \tag{3.4}$$

When the prediction label and the target one are of the same sign, the cost is equal to 0. If the sign differs, then the loss function is calculated, and the orientation of the hyperplane changes until both signs coincide.

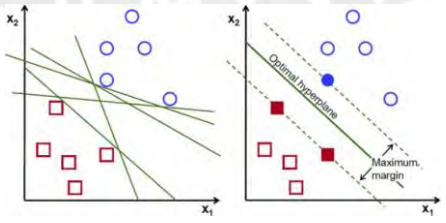


Figure 3.7: Support Vector Machine theory

Regularization is one of the most important hyperparameters to tune in while working with

machine learning, such as SVM and logistic regression. The dataset comprised of more than 7 000 genes results in noise, which can lead to overfitting. To avoid such a nuisance, L1 regularization also called LASSO or L2, called Ridge regression can be employed when training the model. While training, weights are updated at each iteration to decrease the loss functions. With complex datasets, those weights can increase and will not be able to generalize when provided with unseen data. The regularization will shrink the weights towards zero; thus reducing the variance of the model, and avoiding overfitting. On the other hand, when computing regularization to a model, the risk of underfitting surfaces increases because the weights will be so close to zero that a precise update of the weights can not be achieved. L1 and L2 regularization differ in how they are applying the shrinkage. Whilst L1 adds the shrinkage to the sum of the absolute value of coefficients, L2 adds the penalty to the square of the magnitude of the coefficients. Other parameters can be useful, such as the maximum of iteration and the weights associated with the classes.

3.2.1 Dataset processing

As stated above, the genetic expression was pre-processed by logarithm 2, however, different tests were performed to improve the final result, and to increase the standardization of the model.

Normalization Normalization consists of re-scaling the value of the gene's expression into a range between [0,1], where X_{min} and X_{max} are the minimum and maximum values of the attributes.

$$z = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{3.5}$$

Standardization Standardization consists of re-scaling the gene's expression to have a mean of 0 and a standard deviation of 1 by the following equation, with σ the standard deviation from

the mean and μ the mean.

$$z = \frac{(x - \mu)}{\sigma} \tag{3.6}$$

Discretization The genetic expression is represented by quantitative data. Nonetheless, the genetic expression of all patients is clearly different and distant from each other. To decrease the degree of freedom of the classifier, the genetic expression was transformed into a nominal value. As a matter of fact, the genetic expression of each of the patients can be divided into k intervals, by equal-width or equal frequency intervals by applying unsupervised binning discretization defined by

$$width = (maximum_{value} - minimum_{value})/k \tag{3.7}$$

Equal frequency discretization assigns the variables in a way that each class will have the same n bins or number of instances. Concurrently, equal-width divides the attributes into n groups of equal size (Jung, Bi, & Davuluri, 2015). Discretization will boost the correlation between the attributes and the target variable, creating a more intelligible model, and improving the speed and accuracy of the classifier.

3.2.2 Establishing the dataset

Different datasets were fed to the models. The most obvious strategy would be to assign directly the 500 patients to the classifier. However, the possibility for the model to train properly, and have appropriate predictions would be limited due to the aforementioned imbalanced data where 80% of the patients do not present recurrence. Therefore, different datasets were used, where patients were chosen according to their status and time of recurrence.

Based on previous studies (Alkharabsheh et al., 2022) (Q. Wei & Dunbrack, 2013), it was evident that the dataset must be balanced and not extensive for any machine learning classifier to operate properly. Thus, From the 500 patients, the first approach was to separate them into

three different datasets.

- First dataset consists of patients with prompt BCR.
- Second dataset consists of patients with average BCR.
- Third dataset consists of patients with prompt and belated BCR.

Table 3.1: Prompt recurrence dataset

Patient's status	Number of patients	Range of time(months)
Disease free	81	0.7 - 12.98
Recurrent	36	1.6 - 14.22
Total	84	

Table 3.2: Average recurrence dataset

Patient's status	Number of patients	Range of time(months)
Disease free	295	14.7 - 51.1
Recurrent	71	12.9 - 44.1
Total	364	

Table 3.3: Prompt and belated recurrence data set

Patient's status	Number of patients	Range of time (months)
Disease free	57	0.7 - 165.1
Recurrent	61	1.7 - 44.1
Total	118	

The training dataset resides within the TCGA set. Once the model is trained, it needs to be tested on new, never seen patients by the classifiers known as test datasets. If these new patients are classified correctly, then the gene signature will be approved for further studies. GSE54460 from NCBI and MSKCC from cbio portal were used as test dataset.

Table 3.4: GSE54460 test dataset

Patient's status	Number of patients	Range of time(months)
Disease free	55	
Recurrent	51	2.6 - 82.29
Total	106	

Table 3.5: MSKCC test dataset

Patient's status	Number of patients	Range of time(months)
Recurrent	104	
Disease Free	36	1.38 - 115
Total	235	

3.2.3 Selection of attributes

After discarding the genes with a variance throughout the patients below 1, the total number of genes was reduced from 20 000 to around 7 000. However, a high number of genes might hinder the training process. Instead, as in the previous (Espichan & Villanueva, 2018) work, the aim was to extract a feature gene template. For that purpose, computational efficiency was improved by removing the irrelevant features that can be mistaken as noise in the models (Raschka, 2018). On the other hand, the selected genes for each feature selection were compared with the objective to provide information about the relevance of genes in the prediction of recurrence. Indeed, some genes were present in each of the feature selection methods, exposing their importance in the prediction of the time of recurrence. In the next lines, different strategies of attribute selection used for this work are explained and the results of their application will be discussed in the following section.

Sequential Forward Selection, SFS The dataset begins without any genes. The genes are then added only if the classifier performance improves until a desired feature gene template is

reached.

Sequential Backward Selection, SBS Selection starts with the full list of genes in the dataset, then the genes are removed if they don't increase the performance of the classifier. The selection stops once a desired feature gene template is reached.

Sequential Forward and Backward Floating Selection The floating sequential selections are variants of the SFS and SBS. They have an additional option of inclusion and exclusion to remove features that were once included. If a newly added gene provides improved results with the removal of a gene previously included, then it is excluded.

Correlation-based Feature Selection A batch of genes can also be selected by Correlation-based Feature Selection, CFS. The CFS algorithm ranks the features according to the correlation between the attributes and the target. With this intent, the correlation score, and merit score is calculated by

$$MeritScore = \sqrt{\frac{lt_c}{l + l(l-1)t_f}}$$

defined by the evaluation of a subset S, including l features, t_f the correlation score between two features, and t_c the correlation value between features and patient's status: either free of BCR or with recurrence. The merit score extends from 0 to 1, with 1 corresponding to genes with higher correlation with the label, and lowest inter-correlation. It is up to the user to choose the merit score to take into account, however, usually sets at 0.5. According to our numerous datasets, only the genes, which scored above 0.60 are kept to train the model while the rest are removed from the dataset.

Classifier Attributes Evaluation Each attribute is evaluated with a classifier chosen by the user. Logistic regression is adopted, since, as displayed in the next section, it grants a greater

classification. The data is fitted to the logistic regression and the weights associated with each attribute retrieved. The 50 genes with the greatest weights are kept for classification.

Gain Ratio The selector evaluates the significance of each gene based on the gain ratio with the following formula, H representing the entropy.

$$GainRatio = (H(Class) - H(class|Attribute))/H(Attribute)$$

The gain Ratio varies between 0 and 1 for each attribute. Those who contribute the most to the classes are selected according to the limit determined by the users, while the rest is discarded.

3.2.4 Model evaluation

Each classifier needs to be evaluated. A plethora of evaluators exist, such as simply splitting the data set into training and test set. Usually, the training set consists of 66% of the total set, while the remaining is utilized to test the model. The three additional main common methods are presented in the following section, and taken from (Cheng, Garrick, & Fernando, 2017).

Leave One Out Cross Validation Leave one out cross validation (LOOCV) isolates 1 patient as the test set and the rest as the training set. For example, if we feed the full set of 500 patients, 499 are used for the training set, and the model is evaluated on the remaining 1 patient. In other words, 500 models will be created, where all patients are, at one point, drawn as the test data.

Leave P out Leave P out provides indices to split the train and test into p observation. Those are operated as the testing set, while the remaining trains the model. The choice of the p number belongs to the user.

Stratified K-Fold Cross Validation Similar to the K-Fold cross validation explained in chapter 3.1, the evaluation of the model is done on 1 of the N folds at each iteration. Stratified K-Fold assures that each of the folds contains balanced data from the different labels.

ROC curve, Precision, Recall, and F1 score Every model is evaluated with the precision variable and ROC curves. Precision measures how close the model prediction is to reality. The higher the precision is, the closer the data point is to the regression line.

$$Precision = \frac{number of \quad True positive}{Number of True Positives + Number of False Positives}$$

The recall can be interpreted as the number of positive correctly predicted amidst the total number of positive samples.

$$Recall = \frac{number of \quad True positive}{Number of True Positives + Number of False Negatives}$$

The F1-score, a combination of both precision and recall, is pertinent for an unbalanced dataset (W. Wegier, 2020).

$$F1score = 2 * \frac{Precision * Recall}{Precision + Reall}$$

The ROC curve, the receiver operating characteristic, is a graphic representation of the performance of the model by categorizing the true positives and the negatives. The Area Under the Curve is also calculated, as a significant score implies, enhanced classification, nevertheless, an Area Under the Curve of 1, can also indicate overfitting.

3.3 Strategies to predict the status of patients using common classifiers

3.3.1 Use the TCGA full dataset

The first idea was to actually enroll the full dataset, with its 500 patients and 7800 genes, while comparing the efficiency of each classifier and the redundancy of the genes. With all 7800 genes kept in the training process, and each classifier evaluated through a 10-fold cross-validation, according to the confusion matrix and the precision displayed in Table 3.6, the logistic regression and SVM classifiers were able to learn to classify patients without BCR, however, no genes were present in the final logistic regression and SVM models. It was clear by the poor results, that a selection of attributes had to be completed in advance to reduce the noise.

Next as shown in Table 3.6, models were built with sklearn library in Python, and the best results are displayed for the different tests tuning the optimization parameters until reaching adequate loss and accuracy. The decision tree displays superior results by increasing the depth to 500 nodes, and Gain Ratio selection attributes, with no sign of overfitting nor underfitting, with a ROC Area of 0.52. SVM model did not overfit or underfit either but displayed improved results with linear kernel. The logistic regression model may have overfit since no instances were correctly classified in new cases when the patients show signs of BCR. L2 regularization was by default applied, and tests were done with L1 regularization, but results remained similar, due possibly to the unbalanced dataset, as explained below. Nonetheless, the results were not as expected.

Table 3.6: 1st experiment table of results with 7000 genes and 500 patients with a 10 Kfold validation

	SVM		**Logistic regression**		Decision Tree	
	DF	Recurred	DF	Recurred	DF	Recurred
Disease Free	35	5	40	0	33	7
Recurred	7	2	9	0	7	3
Precision	0.82	0.22	0.81	0	0.83	0.26
ROC area	0.52		0.5		0.55	

For this purpose, in the second experiment, for all three classifiers, the 7800 genes were tested through SFS abs SBS, as well as sequential forward and Backward floating selection and their respective final selection was evaluated. When each gene was assessed and the classifier correctly trained and evaluated through LOOCV and Leave P-out, with a P value of 100 samples corresponding to 20% of the total sample, 12 final genes signature from the three classifiers and four distinct attributes selections were highlighted, and compared. The function of the repetitive genes was scrutinized with their respective proteins.

The process took a fair extent of time, as the 7 800 genes underwent selection 12 times. None withstanding, after two months of training, a result emerged only for the logistic classifier. Nonetheless, the final results were quite deceptive. Not surprisingly with an accuracy of 63% and a loss of 45%, manifesting low learning, as one out of two patients was classified incorrectly. The final evaluation of the classifiers was not concluded, and different approaches need to be further inspected to increase the overall velocity. The genes identified by sequential selection were not satisfactory. A possible interpretation for the modest results comes from the unbalanced dataset where out of the 500 patients, only 73 showed BCR. To train correctly any classifier, the dataset must be comprised of the same amount of instances. Concerning the increase in speed, distinctive attributes selection was tested, however, the selection was done before the training process, and not during, as done for sequential selection. For example, the

correlation of the genes with the patients' status was used.

Since working with the full dataset provided unsatisfactory results, the patients were separated into three distinctive groups, prompt BCR, average BCR, and prompt and belated BCR. Next, we tested the hypothesis that genes with contrasting times of recurrence might be responsible for the patient's status. So, by isolating each group, a specific gene signature might be identified and BCR predicted.

3.3.2 Patients with early BCR and instances selected with CFS and Gain Ratio

Our next experiment consisted of conserving only the patients with early/prompt BCR or that have been tested for BCR in the following 12 months after the last check-up. Indeed, to increase the life expectancy of patients, the focus needs to be executed on patients with a high risk of presenting early BCR to activate surveillance. In this case, the data were intuitively normalized between [0,1] and then the attributes selected by Gain Ratio and CFS. The following tables summarize the evaluation of each classifier, the number of patients classified correctly, precision, and the ROC area. Each classifier was evaluated through a 10-fold cross-validation to limit the risk of overfitting. Results are explained independently in the ensuing sections. Only the TCGA dataset was used as a training and testing set.

Table 3.7: Number of genes selected by CFS and GainRatio selection attributes

CFS	**GainRatio**
56	119

Table 3.8: Results for patients with early BCR and instances selected with CFS

	Logistic regression		**SVM**		Decision Tree	
	DF	Recurred	DF	Recurred	DF	Recurred
Disease Free	39	9	41	7	38	10
Recurred	15	21	16	20	15	21
Precision	0.72	0.700	0.72	0.74	0.72	0.68
ROC area	0.78		0.705		0.688	

Table 3.9: Results for patients with early BCR and instances selected using CFS with reduced attributes

	Logistic regression		**SVM**		Decision Tree	
	DF	Recurred	DF	Recurred	DF	Recurred
Disease Free	40	8	42	6	39	9
Recurred	10	26	15	22	14	22
Precision	0.80	0.76	0.52	0.75	0.79	0.70
ROC area	0.905		0.91		0.711	

Table 3.10: Results for patients with early BCR and Gain Ratio attributes selector

	Logistic regression		**SVM**		Decision Tree	
	DF	Recurred	DF	Recurred	DF	Recurred
Disease Free	36	12	40	8	31	17
Recurred	19	17	17	19	17	19
Precision	0.655	0.700	0.702	0.704	0.646	0.528
ROC area	0.643		0.681		0.587	

Table 3.11: Results for patients with early BCR and instances selected with Gain Ratio reduced attributes

	Logistic regression		**SVM**		Decision Tree	
	DF	Recurred	DF	Recurred	DF	Recurred
Disease Free	39	9	45	3	39	9
Recurred	14	22	20	16	16	20
Precision	0.736	0.710	0.69	0.84	0.709	0.69
ROC area	0.667		0.691		0.684	

The results below were achieved by keeping the default parameters, except an increase of the maximum depth for the Decision Tree, and linear kernel for SVM.

Classifier model with Logistic regression and CFS attributes selector Applying the logistic classifier to the 56 genes selected by the CFS, the outcome for the classification of the patients according to their status was the following equation. As observed, by CFS attribute selection, 56 genes were identified with high correlation with the label, however, only 14 genes were displayed in the equation. By reducing the number of attributes to 56 from 14 genes, the number of patients classified correctly, the precision and ROC curves increased between 0.10% to 0.20%.

The probability for the patient to present BCR is displayed in the signature below. The genes expression of the patients and their associated weights are summed and the final score passes through the following formula.

$$Probability of BCR = e^{finalscore}/(e^{finalscore} + e^{finalscore})$$

Class Disease Free with 56 attributes:

$$\begin{aligned} 3.38 + [HSPA1A] * 1.52 + [HSPA1B] * 2.02 + \\ [NFIX] * -1.32 + [C5orf10] * -1.85 + \\ [DBNL] * -1.61 + [OLFML3] * 1.12 + \\ [NIPA2] * -1.54 + [ITGBL1] * -1.45 + \\ [NTM] * -1.44 + [LOC95070] * -1 + [CRLF \\ 1] * 1.15 + [TARSL2] * -1.77 + \\ [NAA40] * -2.69 \end{aligned} \tag{3.8}$$

The equation is also displayed in 3.8.

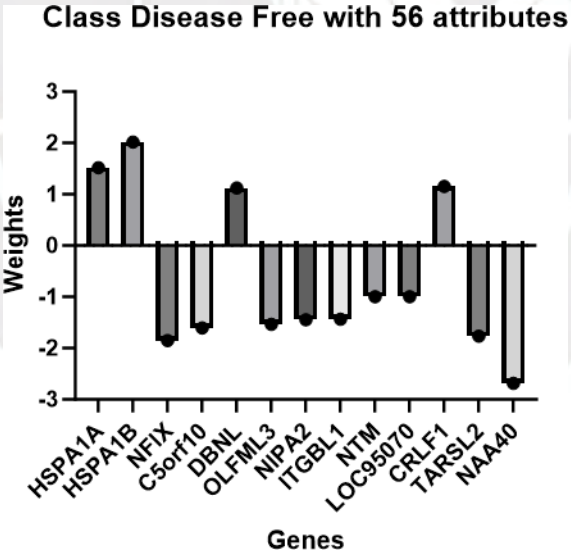
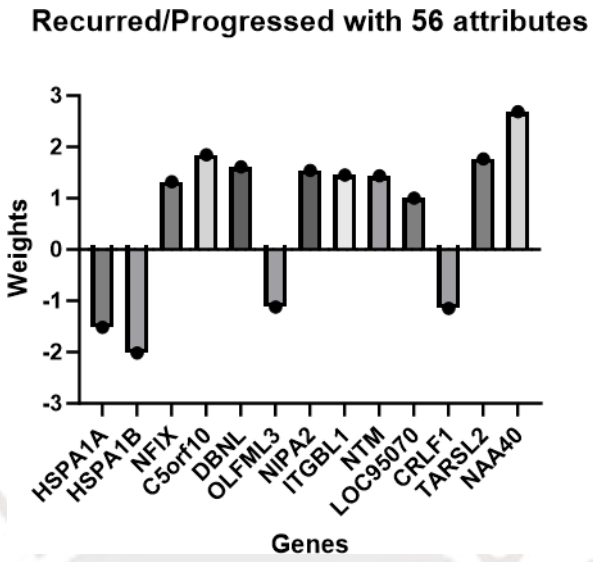
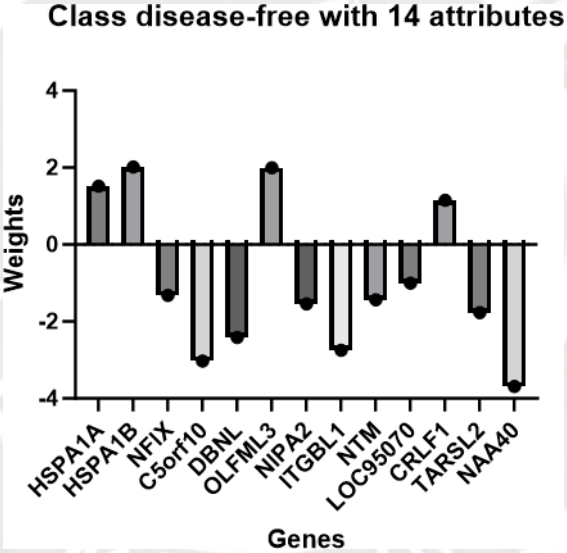


Figure 3.8: Weights associated with the genes to predict the non-recurrence of prostate cancer

From now on, the genes signature will be only presented as a bar graphic, but the full equations are

available in the annex.





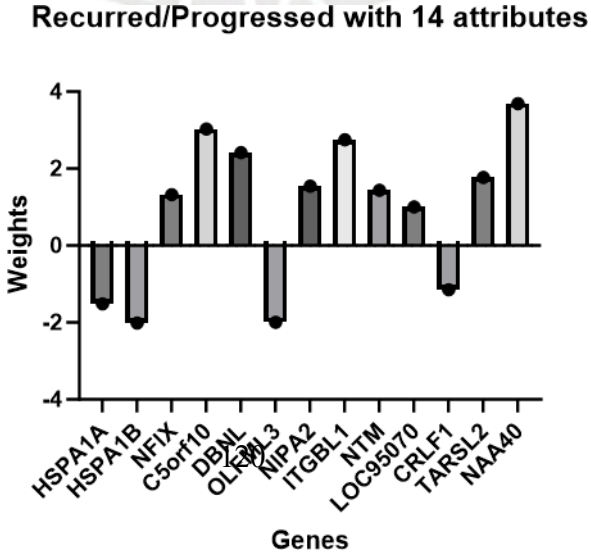
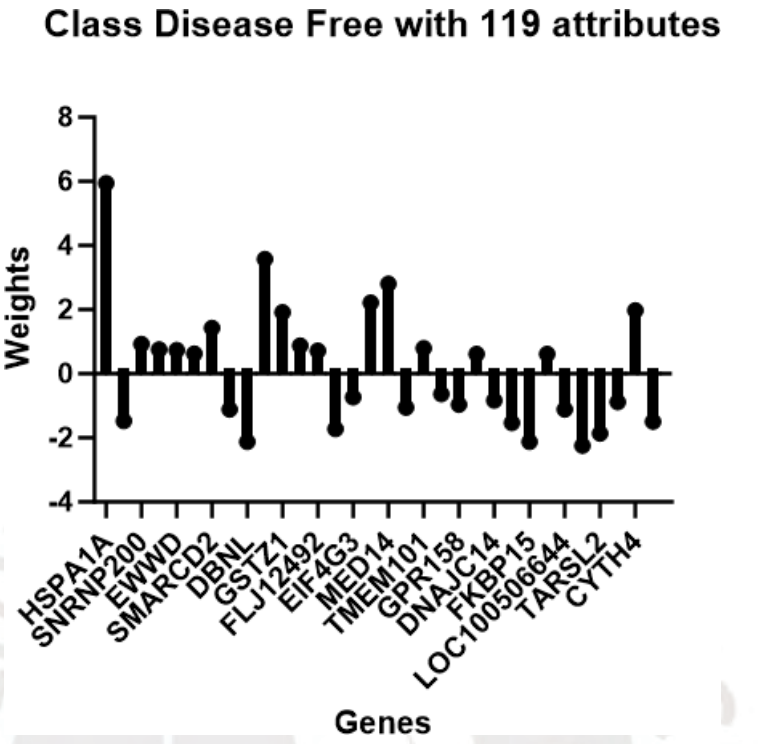


Figure 3.9: Graphic representation of the different experimentation with a variation of number of genes

By reducing the number of attributes improvement as seen in Figure 3.9 in the classifier was witnessed, however, the results for recurred patients were not precise enough, and thus set aside.

Classifier model with Logistic regression and Gain Ratio attributes selector In the case of Gain Ratio, the attributes were presented according to their gain, meaning the genes with more entropy were first on the list. In total, with 119 genes identified, more noise was added, explaining the inferior results in comparison with the CFS selection attributes. Instead of using the 119 genes, an alternative approach was to choose the top first genes according to the Gain Ratio selection, to reduce the noise, and thus as expected improve the results. This was started with the first 30 genes, as there might be enough genes to correctly allocate the patients to its status and keep the noise at a reduced level. By doing so, the precision increased as well as the number of patients correctly classified, as shown in 3.11. But the recurrent patients remained mostly wrongly distributed.



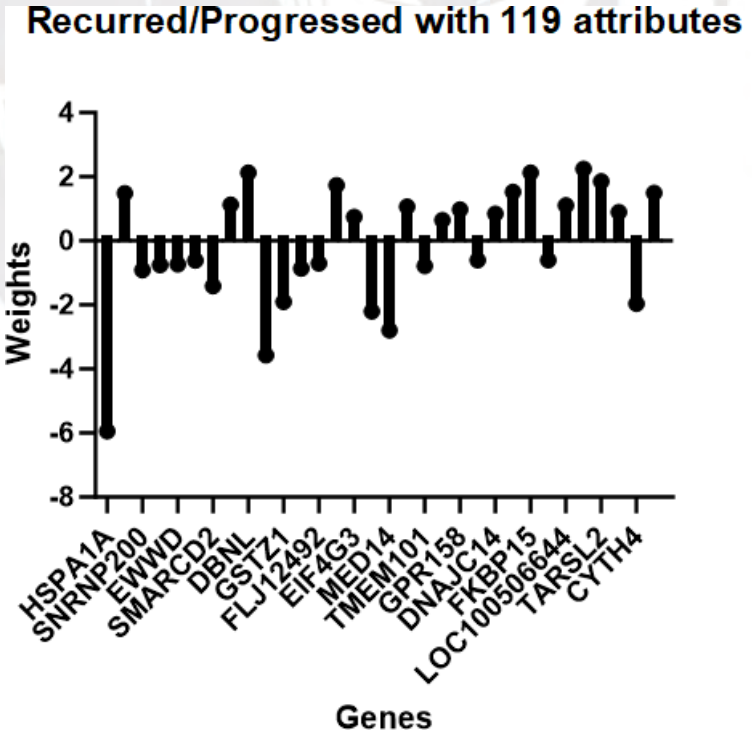
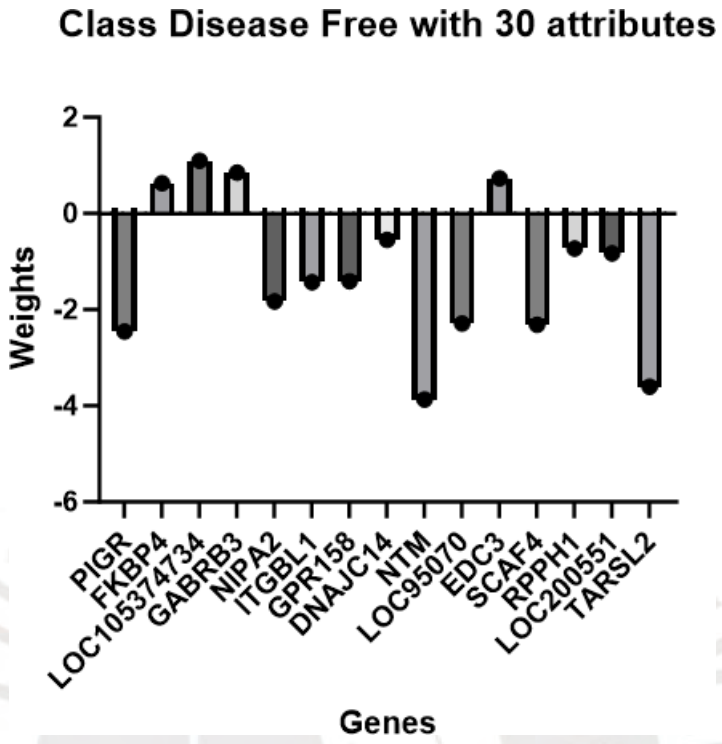


Figure 3.10: Graphic representation of the genes and associated weights with 119 genes



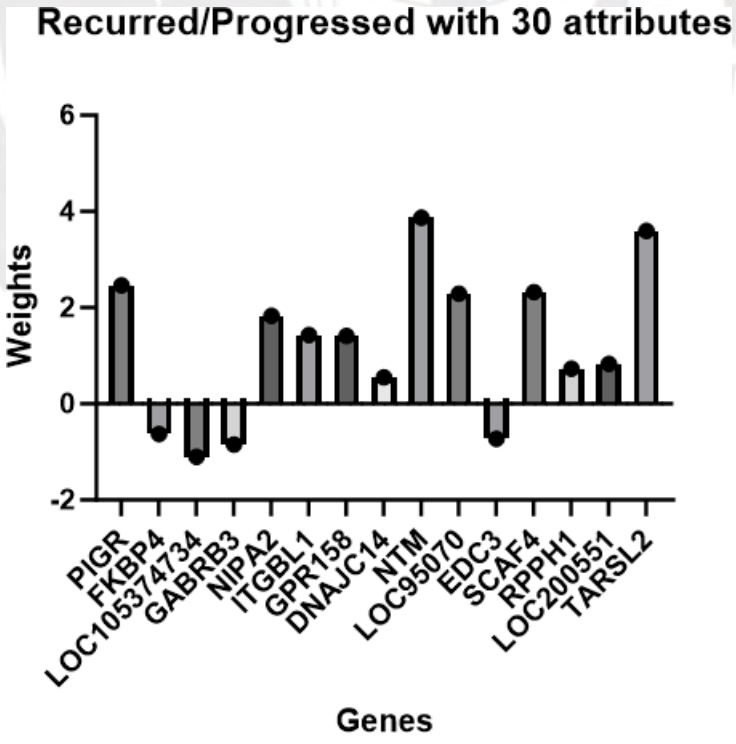


Figure 3.11: Graphic representation of the genes and associated weights with 30 genes

The change of attribute selectors as exhibited in Figure 3.10 and 3.11, did not improve the classification, as most of the patients with recurrence were miss-classified. Further machine learning models were investigated preserving the same dataset and attribute selectors.

Classifier model with SVM and CFS attributes selector Most of the attributes were engaged in the SVM classifiers, nonetheless nearly all of them have insignificant weights associated, meaning that their expressions don't have a critical influence in the final classification, as displayed in Table 3.12, only 8 genes out of 56 genes have weights superior to 1.

Table 3.12: Weights assigned to each attribute after training

Weights	Genes	Weights	Genes	Weights	Genes
-1.03	HSPA1A	0.78	PIGR	-0.69	KIAA1324
-1.02	HSPA1B	0.88	NFIX	-0.46	FKBP4
-0.35	LOC105374734	-0.29	TNRC18	-0.54	EWWD
1.11	C5orf10	0.14	ADIPOR1	0.50	ATP5CL1
1.17	DBNL	-0.75	OLFML3	0.01	PSMD7
-0.02	UBTF	-0.433	CHTF8	0.2611	FLJ12492
-0.3318	FADS1	0.7646	PPFIA2	0.9903	NIPA2
0.5594	MINK1	0.4934	ASAP2	-0.3559	ARHGEF2
-0.3019	CUL1	1.029	NTM	0.0015	TMEM99
1.2708	LOC95070	-0.1338	ROGDI	1.2708	LOC95070
0.0354	BTBD9	-0.5776	EDC3	0.3346	ATG2A
- 0.0386	LOC100130886	0.6082	SCAF4	0.4427	RNF40
-0.2406	PPM1M	-0.4174 9	UTS2B	-0.5968	JARID2
-0.9236	CRLF1	0.1965	RPH3AL	1.3071	NAA40
0.1482	UBE2R2	0.2281	RPPH1	0.3202	FAHD1
-0.5972	MYCBP	0.5309	PPP2R5B	-1.0107	TARSL2
0.4063	FANCC	-0.3262	JIP2	0.7628	ITGBL1
0.3765	INTS8	-0.3401	PPP1R26	-0.181	LOC102724376
-0.5666	GPR158	-0.0926	DNAJC14		

Classifier model with SVM and GainRatio attributes selector As it was performed for CFS attributes selection, most of the genes were handled in the classification, but only a few of them have associated weights above 1.

Table 3.13: Weights assigned to each attribute after training

Weights	Genes	Weights	Genes	Weights	Genes
0.6833	HSPA1A	-0.2336	PIGR	0.2479	KIAA1324
0.7842	GLTSCR2	-0.2522	NFIX	-0.1081	FKBP4
-0.1081	LOC105374734	-0.542	IFI4	0.002	SYTL1
-0.5253	GPT2	-0.5253	SNRNP200	-0.2983	RCN1
-0.0683	ALDH3A2	-0.2569	LARP1	-0.191	BAZ2A
-0.98	SMARCD2	0.0306	CREBBP	0.4928	C17ORF62
1.0727	DBNL	-0.2959	OLFML3	-0.3519	GSTZ1
-0.7173	TNRC18	-0.1168	UQCR11	-0.4147	EWWD
0.2388	ATP5O	0.165	DFNA44	0.0674	USP9Y
-0.3555	SLC31A1	0.1817	VPS4A	-0.5745	GABRB3
0.0624	KIDINS220	-0.3568	STO	-0.1262	FLJ12492
0.3676	LOC196752	0.6557	SUGP2	0.4873	PPFIA2
0.7101	NIPA2	0.4945	EIF4G3	-0.7702	ABCC5
-0.6127	MED14	-0.316	PEFLIN	-0.0729	C17ORF49
0.107	MINK1	0.1452	RTS	0.2318	PCNXL3
-0.1498	KHSRP	-0.302	RPRD2	0.0664	NARFL
-0.0245	DKFZp547L134	-0.0887	WHSC1	0.1359	FN3KRP1
0.2198	FRYL	0.3277	UBE2R2	0.1806	SLC37A3
-0.2138	FLJ10821	0.1302	PRO2309	-0.0867	JIP2
0.0498	ODF2	0.5145	MGC12217	-0.6105	TMEM101
0.942	SMG7	0.0137	PPFIA1	-0.6323	JAGN1
0.365	IGSF3	0.1951	LOC102724376	0.4219	GPR158
-0.0183	KIAA2018	-0.1068	KDM3A	-0.1618	NIT2
-0.2272	CRAMP1L	0.3604	DNAJC14	0.603	SMEK1
-0.2812	DYS	0.4884	SH3D1B	0.0149	CDK14
0.1199	LOC145448	0.3437	ASAP2	0.519	CUL1
0.8385	HDAC6	-0.0131	KIAA0884	0.2272	DENND4B
-0.6012	ITSN1	0.1841	PCDHGB2	-0.0357	CIAPIN1
-0.0493	TMEM99	0.3727	BTBD9	-0.3305	EDC3
0.4733	ATG2A	0.0059	FKBP15	0.2257	FAM193A
0.2957	LOC100130886	0.0665	SCAF4	-0.2039	BROX
0.9651	SUSD4	-0.6635	TECPR2	-0.7866	TMEM41A
-0.2143	SAP130	-0.3092	PPM1M	0.1535	RPPH1
0.1014	SUCNR1	-0.7806	CRY1	0.1484	SDHAF3
0.0422	RNF135	0.3281	LOC200551	0.1289	USP37
0.3469	LOC100506644	-0.0225	TBC1D7	-0.534	OTUD3
-0.4073	RPH3AL	0.5602	TARSL2	0.5895	FANCC
0.5793	FAIM	-0.3705	MSI2	-0.1818	WTIP
-0.0511	MYLK3	1.1974	PI4KAP1	0.5079	USP42
-0.2057	CYTH4				

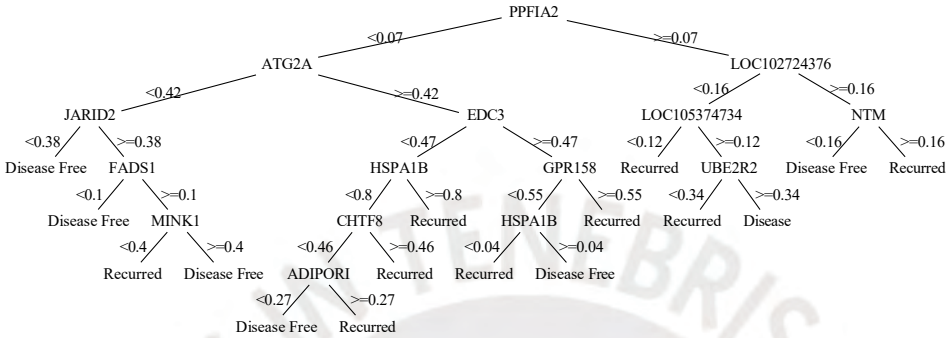
Table 3.14: Weights assigned to each attributes after training with reduced genes selection

Weights	Genes	Weights	Genes	Weights	Genes
1.3711	PIGR	-0.5454	FKBP4	-0.7745	LOC105374734
0.0607	RCN1	-0.6385	ALDH3A2	0.3761	MRPL12
-0.4754	TNRC18	0.6502	USP9Y	-0.2645	ATP5CL1
-0.9198	GABRB3	0.3886	KIAA0284	0.9039	NIPA21
0.28284	MINK1	0.2918	RNF40	0.066	FN3KRP
-0.4682	UBE2R2	0.275	JIP2	0.4199	ITGBL1
0.7967	GPR158	0.0642	DNAJC14	0.6192	ASAP2
1.4054	NTM	0.2366	TMEM99	1.5095	LOC95070
-0.433	EDC3	0.1321	POFUT2	0.5205	MGC11316
1.0715	SCAF4	-0.1338	ROGDI	1.2708	LOC95070
0.0354	BTBD9	-0.5776	EDC3	0.3346	ATG2A
- 0.0386	LOC100130886	0.6082	SCAF4	0.4479	RPPH1
-0.0765	LOC200551	0.706	TARSL2		

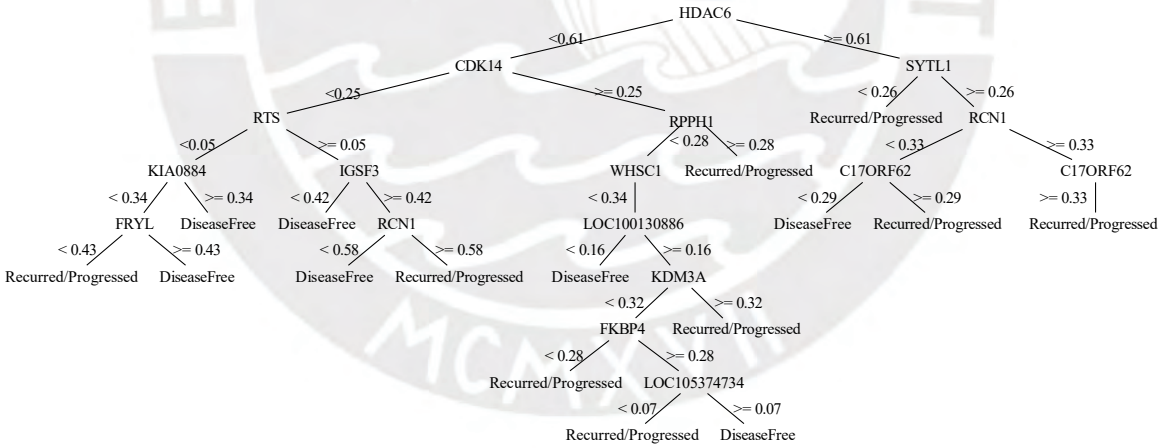
Since fewer genes were engaged in the classifier, the precision increased, but the weights associated with each gene remained low, and thus this strategy was overruled.

Classifier model with Decision Tree and CFS attributes selector The Decision Tree had the potential to offer overall better results, thanks to a complex tree where the classes can be allocated with

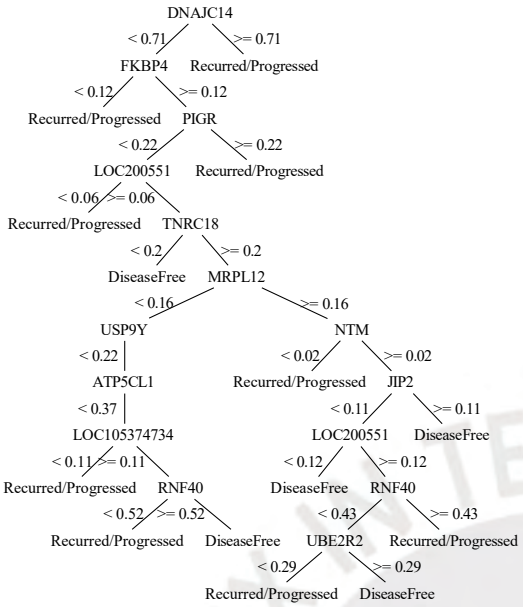
a myriad of alternatives. However, the results remained mediocre and were not further investigated. The classification did improve, but the experiment was discarded, as only a few recurred patients were accurately allocated to their respective labels.



Classifier model with Decision Tree and Gain Ratio attributes selector The complexity of the Decision tree and the quality of its results do not enable a complete gene analysis. By using the 119 genes identified by Gain Ratio, noise is added, and results are inadequate.



While using 30 genes, the Tree provides direct information as to the genes responsible for the patient's status, such as when the expression of DNAJC14 is higher or equal to 0.71. Nonetheless, improvements need to be achieved for these results to be conserved.



After testing different classifiers and attribute selectors, and according to the various results, the dataset consisting of patients with early BCR was momentarily put aside to evaluate the other proposed datasets.

3.3.3 Patients with average BCR and instances selected with CFS and Gain Ratio

According to the previous results, classifiers provided better results if the attributes selected coincided with the one presented in the final signature. Therefore, only the results with fewer attributes were presented. In the case of Gain Ratio attributes selection, for the next attempts only the top 30 genes were selected. The results for the three classifiers and two attribute methods are presented in the ensuing tables 3.16 and 3.17. Again, the classifiers were evaluated through a 10 k-fold cross-validation from the TCGA set and normalized genes.

Table 3.15: Number of genes selected after CFS and Gain Ratio attribute selectors

CFS	GainRatio
68	30

Table 3.16: Results for average BCR, and CFS reduced attributes

	Logistic regression		**SVM**		Decision Tree	
	DF	Recurred	DF	Recurred	DF	Recurred
Disease Free	262	33	286	9	272	23
Recurred	49	21	64	7	55	16
Precision	0.72	0.41	0.817	0.438	0.842	0.400
ROC area	0.61		0.534		0.599	

Table 3.17: Results for average BCR, and Gain Ratio reduced attributes

	Logistic regression		**SVM**		Decision Tree	
	DF	Recurred	DF	Recurred	DF	Recurred
Disease Free	288	7	295	0	258	37
Recurred	67	4	71	0	44	27
Precision	0.81	0.364	0.806	0	0.854	0.422
ROC area	0.562		0.50		0.627	

No further explanations are established, as to the deceptive results. The models were able to classify correctly most of the instances without recurrence, but they misclassified the patients with recurrence. The best model could not classify correctly even half of the patients with BCR. The dataset was hence, discarded.

3.3.4 Patients with early and late recurrence, and instances selected with CFS and Gain Ratio

The third dataset investigated was composed of patients with early and late recurrence. From the past two experiments, the early recurrence dataset gave superior results, but could only be applied to a certain part of the population. In order to increase the spread of the prediction, patients with belated BCR were computed. Overall, the genes selected by CFS delivered preferable results and were solely manipulated

in the analysis shown below. The results by Gain Ratio are discussed later in the subsection. Only the leading classifiers were kept and further investigated, concurrently the decision tree was discarded. All classifiers were evaluated through 10 k-fold cross-validations on the TCGA dataset.

The classifiers improved their classification with the new dataset and proposed satisfactory results. In order to confirm the gene's signature, the GSE54460 was supplied as a test dataset. By operating on a new never-seen dataset, models were approved or discarded. Notwithstanding, the results for the dataset were deceptive, as only half of the new patients were correctly classified by the trained classifiers. The model, not trained for patients with a time of recurrence between 15 months and 50 months, could not behave on the aforementioned attributes in the test dataset. These new patients were not allocated properly. Standardization of the training and testing datasets might be able to improve moderately the score. The classifiers trained with the genes selected by Gain Ratio were not tested with GSE54460, as the results were already unsatisfactory.

Table 3.18: Number of genes selected with CFS and Gain Ratio attribute selectors

CFS	**GainRatio**
44	30

Table 3.19: Results for early and late recurrence using CFS reduced attributes

	Logistic regression		**SVM**		Decision Tree	
	DF	Recurred	DF	Recurred	DF	Recurred
Disease Free	52	9	57	4	40	21
Recurred	8	49	5	52	23	24
Precision	0.867	0.845	0.919	0.929	0.635	0.618
ROC area	0.934		0.923		0.626	

Table 3.20: Results for early and late recurrence for the GSE54460 test dataset

	Logistic regression		**SVM**		Decision Tree	
	DF	Recurred	DF	Recurred	DF	Recurred
Disease Free	30	21	28	23	29	22
Recurred	16	29	18	27	20	25

Table 3.21: Results for early and late recurrence using Gain Ratio reduced attributes

	Logistic regression		**SVM**		Decision Tree	
	DF	Recurred	DF	Recurred	DF	Recurred
Disease Free	45	16	48	13	45	16
Recurred	21	36	21	36	25	32
Precision	0.682	0.692	0.693	0.735	0.643	0.667
ROC area	0.735		0.709		0.659	

3.3.5 Evened dataset and class partition

The previous experiments did not meet our expectations, thus other directions were taken. Firstly, the dataset is now composed of all the recurred TCGA patients, covering most of BCR time, as obtaining a gene signature to predict patients with BCR proffered the most difficulty. The same number of patients without BCR are selected. However, only BCR-free patients with similar disease-free time are picked. To decrease the degree of freedom and thus by definition increase the classification, the new dataset was standardized since the attributes followed a Gaussian distribution 3.3.7 and discretized with equal width. The number of intervals was changed through the experiments, starting with two intervals, and incremented until reaching satisfactory results. By narrowing the intervals, further investigation was concluded, but precision was lost. The genes were still being selected through CFS attributes evaluator from the standardized and discretized TCGA dataset. However, the dataset comprised three datasets. The set was consequently split into training involving the TCGA data and testing including the GSE and

MSKCC cases.

Cohorts, reflecting real-world circumstances, encompass ordinarily more patients without BCR. Nevertheless, the unbalanced data set can create a bias towards the sample presented in greater numbers (Tasci, Zhuge, Camphausen, & Krauze, 2022). Most prostate cancer relapsing prediction either disregards the mater (Oh et al., 2017) (Zhao, Z.Tao, & Li, 2022) or applies balancing techniques namely oversampling, undersampling or hybrid (Rajendran, Jayabalan, & Thiruchelvam, 2020), (Beinecke & Heider, 2021), (H. Chan, Chattopadhyay, Chuang, & Lu, 2021). In spite of that, oversampling infers duplicating minority classes, and discarding variant expression, while undersampling removes an instance of the majority class, increasing the risk of eliminating useful information. Integrating multi-omic data from GSE and MSKCC, the intent is to increase the accuracy, robustness, and greater statistical power by including more specimens in the study. The objective is to focus on the cause of the prediction rather than the final prediction score.

Only logistic regression was handled, as it provided better classification. The results illustrated below represent the models with superior classification.

Table 3.22: Number of patients in training and test set for evened dataset

Number of patients in training set	**Number of patients in test dataset**
143	241

Table 3.23: Status of the patients in the training and testing set

Number of patients in training set with BCR	**Number of patients in training set without BCR**
71	73
Number of patients in test set with BCR	**Number of patients in test set without BCR**
90	170

Table 3.24: Results without discretization

	Logistic regression	
	DF	Recurred
Disease Free	121	13
Recurred	54	53
Precision	0.691	0.803
ROC area	0.839	
Correctly Classified Instances	174	
incorrectly Classified Instances	67	

Table 3.25: Results from standardized dataset without discretization

	Logistic regression	
	DF	Recurred
Disease Free	113	24
Recurred	38	70
Precision	0.748	0.745
ROC area	0.782	
Correctly Classified Instances	183	
incorrectly Classified Instances	62	

Table 3.26: Results from standardized dataset using two-interval discretization

	Logistic regression	
	DF	Recurred
Disease Free	117	17
Recurred	39	68
Precision	0.750	0.800
ROC area	0.851	
Correctly Classified Instances	185	
incorrectly Classified Instances	56	

Table 3.27: Results from standardized dataset using three-interval discretization

	Logistic regression	
	DF	Recurred
Disease Free	120	14
Recurred	41	66
Precision	0.745	0.825
ROC area	0.844	
Correctly Classified Instances	186	
incorrectly Classified Instances	55	

Table 3.28: Results from standardized dataset using five-interval discretization

	Logistic regression	
	DF	Recurred
Disease Free	112	22
Recurred	56	51
Precision	0.667	0.698
ROC area	0.728	
Correctly Classified Instances	163	
incorrectly Classified Instances	78	

Table 3.29: Results from standardized dataset using ten-interval discretization

	Logistic regression	
	DF	Recurred
Disease Free	109	26
Recurred	40	66
Precision	0.722	0.702
ROC area	0.779	
Correctly Classified Instances	175	
incorrectly Classified Instances	66	

The ROC area values were reasonably elevated, indicating a proper classification. Nevertheless, by thoughtfully looking at the results, the high precision mostly comes from patients without BCR. In some instances, 1/3 of patients with BCR were misdiagnosed, resulting altogether in deceptive classification due to the unbalanced test data, as 143 patients present BCR, while 253 do not.

3.3.6 Readjustment of the dataset

As explained in the last section, the models were able to classify correctly most of the patients without BCR, but lack accuracy regarding cases with recurrence. To resolve the issue, the new test dataset included the entire MSKCC set, along with the patients with BCR from the GSE54460, excluding the 55 patients without BCR. Consequently, the dataset was evened to verify how well the model can classify the 51 patients with BCR from the GSE54460 dataset. In order to select patients without BCR from the dataset, we retained those who had a comparable time to recurrence. We conjectured that by analyzing patients with distinct diagnoses but similar time to BCR, we could identify a more precise set of signature genes for forecasting prostate cancer recurrence. An alternative solution to an unbalanced dataset is to shift the metric score to g-mean, balanced accuracy, and F1-score

Table 3.30: Number of patients in the training and test sets after dataset readjustment

Number of patients in training set	**Number of patients in test set**
143	194

Table 3.31: Status of patients in each set

Number of patients in training set with BCR	**Number of patients in training set without BCR**
71	73
Number of patients in test set with BCR	**Number of patients in test set without BCR**
90	104

Table 3.32: Results from the standardized dataset using two-interval discretization

	Logistic regression	
	DF	Recurred
Disease Free	76	33
Recurred	25	62
Precision	0.752	0.653
ROC area	0.793	
Correctly Classified Instances	138	
incorrectly Classified Instances	58	

Table 3.33: Results from the standardized dataset using three-interval discretization

	Logistic regression	
	DF	Recurred
Disease Free	82	27
Recurred	29	58
Precision	0.739	0.682
ROC area	0.778	
Correctly Classified Instances	140	
incorrectly Classified Instances	56	

Table 3.34: Results from the standardized dataset using five-interval discretization

	Logistic regression	
	DF	Recurred
Disease Free	76	33
Recurred	19	68
Precision	0.800	0.673
ROC area	0.824	
Correctly Classified Instances	144	
incorrectly Classified Instances	52	

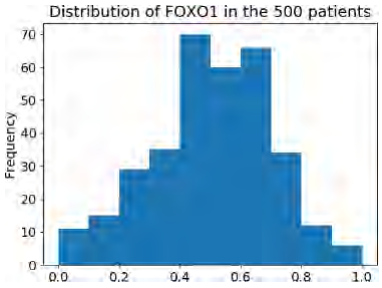
Table 3.35: Results from the standardized dataset using ten-interval discretization

	Logistic regression	
	DF	Recurred
Disease Free	82	27
Recurred	24	63
Precision	0.774	0.700
ROC area	0.864	
Correctly Classified Instances	145	
incorrectly Classified Instances	51	

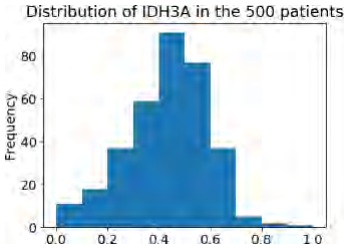
This approach provided improved classification. By discretizing the values of the genes into ten intervals, the best precision for both status and the highest number of instances correctly classified was reached. However, with $\frac{1}{3}$ of the cases misclassified, the score needed to be improved.

3.3.7 Standardization versus normalization

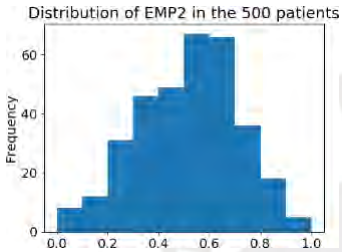
In the previous section, the dataset was standardized or normalized intuitively. However, no strict rules exist for their application. From literature (Ali, Faraj, & Koya, 2014), normalization operates better if the data distribution does not follow a Gaussian distribution, while standardization is more helpful when following a Gaussian distribution. Some random genes distribution along the 500 patients are displayed below, showing indeed Gaussian distribution.



(a) FOXO13 distribution



(b) IDH3A distribution



(c) EMP2 distribution

Figure 3.12: Genes distribution

Standardization is also mainly used when attributes differ in units and scale. Machine learning generally works better with standardized values, as weight may update faster. However, there are no ground rules to which one should be applied to the data. Tests need to be done with raw, normalized, and standardized attributes and the performance of the model with each data set had to be compared.

Table 3.36: Results from the raw dataset without discretization

	Logistic regression	
	DF	Recurred
Disease Free	85	18
Recurred	33	61
Precision	0.720	0.772
ROC area	0.809	
Correctly Classified Instances	146	
Incorrectly Classified Instances	50	

Table 3.37: Results from the standardized dataset without discretization

	Logistic regression	
	DF	Recurred
Disease Free	85	24
Recurred	31	56
Precision	0.780	0.644
ROC area	0.795	
Correctly Classified Instances	141	
incorrectly Classified Instances	55	

Table 3.38: Results from normalized updated dataset without discretization

	Logistic regression	
	DF	Recurred
Disease Free	83	19
Recurred	26	68
Precision	0.761	0.782
ROC area	0.839	
Correctly Classified Instances	151	
incorrectly Classified Instances	45	

Standardization didn't improve extensively the classification in comparison to the raw attributes, only five more patients were correctly classified than when raw data were being fed to the model. Even without discretization, normalized gene expression provided superior results than standardized and raw ones. An additional aspect is the number of genes selected in the logistic regression result. Indeed, an approximation of 40 genes was present in the logistic regression signature with raw and standardized data, but only 20 with the normalized one. With a reduced signature, an enhanced detection of the genes responsible for the BCR was obtained.

Discretization: equal width - equal frequency Subsequently, numerous discretization approaches exist, either splitting the expression of the genes into intervals containing the same amount of attributes, or the same width. The normalized 173 patients were divided into N numbers of intervals, with both equal width and equal frequency discretization. The precision, accuracy, and additionally the number of genes in the final models were evaluated to decide which procedure to be retained for the following experiments.

Table 3.39: Comparison of the method to estimate the status of patients with the fewest number of genes and higher accuracy without discretization

	Logistic regression	
	Equal width	Equal frequency
Precision	75	75
Accuracy	75	77
Number of genes	19	31

Table 3.40: Comparison of the method to estimate the status of patients with the fewest number of genes and higher accuracy with five intervals

	Logistic regression	
	Equal width	Equal frequency
Precision	81	77
Accuracy	81	77
Number of genes	10	38

Table 3.41: Comparison of the method to estimate the status of patients with the fewest number of genes and higher accuracy with ten intervals

	Logistic regression	
	Equal width	Equal frequency
Precision	77	71
Accuracy	75	71
Number of genes	7	16

Altogether, equal width results exceeded equal frequency from accuracy to the number of genes highlighted. In the next section, discretization was executed exclusively with equal width.

Discretization: Number of bins and final signature One question remaining was how to apply the discretization. First, data were discretized into N random intervals, to inspect a possible increased accuracy. Nonetheless, it might be a possibility that each gene performs better with different number of intervals. Different strategies were investigated to determine the number of bins for discretization (Cebeci & Yildiz, 2017).

Table 3.42: Table of methods to estimate the number of bins for discretization

Name of the rules	Formula	Number of bins
Square root	$n^{\frac{1}{2}}$	12
Cencov	$n^{\frac{1}{3}}$	5
Rice	$2n^{\frac{1}{3}}$	10
Terrell-Scott	$(2n)^{3^{\frac{1}{}}}$	6
Sturge	$1+\log_2 n$	8
Brooks-Carruthers	$5\log_{10} n$	11
Freedman-Diaconis	$\frac{R}{IQRn^{\frac{-1}{3}}}$	
Scott	$\frac{R}{3.5\sigma n^{\frac{-1}{3}}}$	

The first six rules produced the same number of bins for each attribute, whether normalized or standardized, considering they solely included the variable n that represented the number of elements

per attribute. Meanwhile, the last two rules take into account the difference between the maximum and minimum value or the range of the dataset, R, and IQR which describes the middle 50% of values from the lowest to the highest and σ the standard deviation of each attribute. Nonetheless, if working with normalized data, the range of the data will always be equal to 1, while working with standardized, the standard deviation will be close to 0. The NumPy library histogrambinedges calculates the edges of the bin according to the rules explained above. Freedman and Scott had slower computation time, since the bin intervals were calculated for each of the 7800 genes and supplied overall inferior results.

Table 3.43: Results from normalized and discretized data

	Square root	Cencov
Correctly Classified Instances	154	159
Incorrectly Classified Instances	42	37
	Terrell-Scott	Sturge
Correctly Classified Instances	141	141
Incorrectly Classified Instances	55	55
	Freedman-Diaconis	Scott
Correctly Classified Instances	138	138
Incorrectly Classified Instances	58	58
	Rice	Brooks-Carruthers
Correctly Classified Instances	148	144
Incorrectly Classified Instances	48	52

Cencov clearly granted the best results, with a five-bin discretization, where the precision of both statuses is over 0.800. The following signature, containing 10 genes, can determine if the patients present a risk of BCR in the near future. Out of the 55 patients with BCR from the GSE54460 set, 15 patients were incorrectly classified, confirming the veracity of the model.

Table 3.44: Results from normalized data using five-bin discretization

	Logistic regression	
	DF	Recurred
Disease Free	84	18
Recurred	19	75
Precision	0.816	0.806
ROC area	0.857	
Correctly Classified Instances	159	
incorrectly Classified Instances	37	

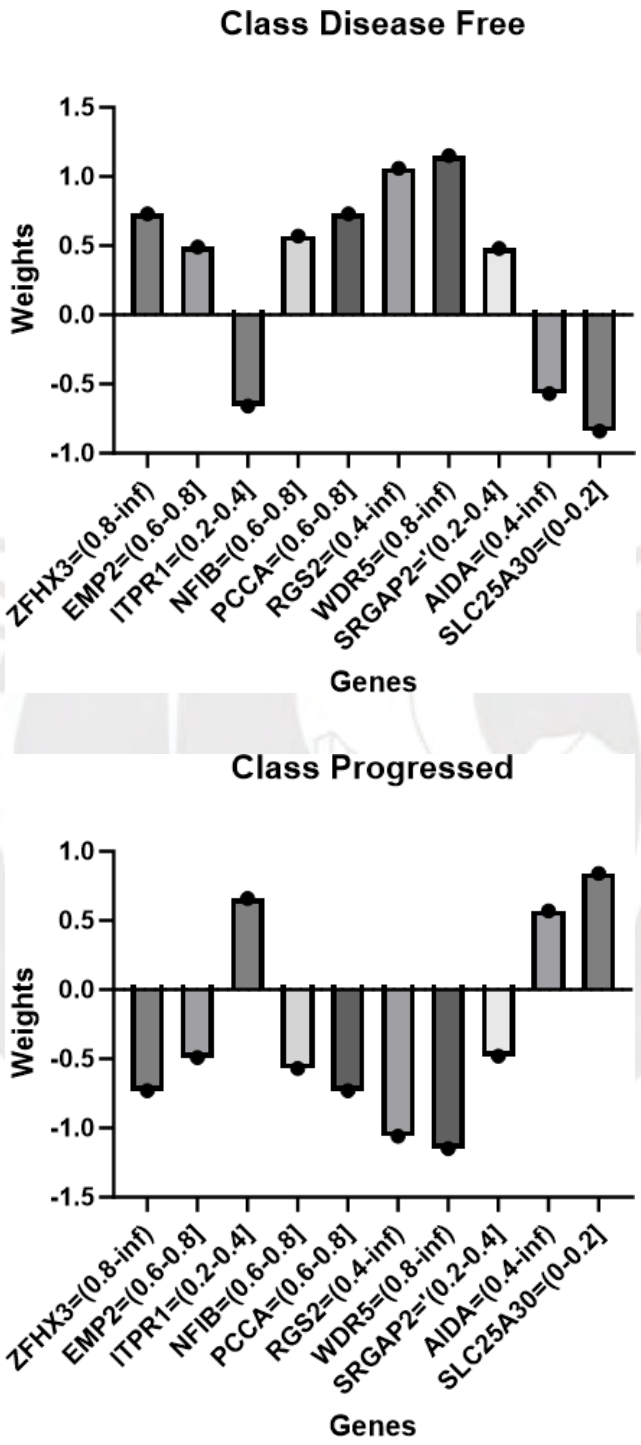


Figure 3.13: Graphic representation of the genes and associated weights after discretization

The risk of presenting BCR can be predicted with an accuracy of 81% through the ten genes signature including ZFHX3, EMP2, ITPR1, NFIB, PCCA, RGS2, WDR5, SRGAP2, AIDA, SLC25A30, displayed in Figure3.13. Previous studies achieved similar results. Nevertheless, the aforementioned genes differ from the precedent investigation, advancing the role of tumor extracellular matrix (ECM) and cell proliferation in recurrence. As previously explained, the ECM is the sole warrant of the organization of the glands, regulating cell proliferation and migration. Alterations in its composition have been associated with tumor cell migration and the presence of metastasis (Winkler, Abisoye-Ogunniyan, & Metcalf, 2017) and (Stewart, Cooper, & Sikes, 2004).

These results are consistent with previous pre-clinical research in prostate cancer. It has been demonstrated in mice that deletion of the ZFHX3 gene boosts cell proliferation, and it disrupts tissue organization and gland formation during development (X. Sun et al., 2015). According to (Hu, Zhang, & Chen, 2019), increased levels of ZFHX3 are associated with better survival, hence explaining the weights correlated with the expression of ZFHX3 in the patients' signatures. The expression above 0.8, $[ZFHX3 = (0.8 - 1)]$ was correlated with a positive weight of 0.73 to classify patients as low risk of BCR. Differing, and within the same interval, $[WDR5 = (0.8 - 1)]$, was conventionally higher in prostate cancer than in healthy tissue. Its interaction with androgen signaling infers its role to accelerate prostate cancer cell proliferation (Kim et al., 2014).

The gene NFIB has been elated to prostate hyperplasia (Grabowska et al., 2015). Although a relationship between prostate enlargement and BCR has not been previously studied, hyperplasia can be acknowledged as an aftermath of a troubled ECM. Also, ITPR1 does not have a previously reported link to prostate cancer, but it proved to directly affect the process of apoptosis in colorectal cancer and ovarian cancer (Duca et al., 2021). Its depletion can deepen the loss of apoptotic control, consequently preserving cancer cells in the organs and disrupting their structural characteristics.

The other genes in the signature that participate in the classification of patients have not been previously mentioned in prostate cancer research. Therefore, our approach shows novelty and may support further research into the role of these novel genes and their pathways might show promise for better screening, diagnosis, or treatment.

3.4 Predicting the time of recurrence for each patient

3.4.1 TCGA full dataset

We first tried, intuitively, to predict the time of recurrence using only the raw data. Solely patients with BCR from TCGA, GSE, and MSKCC were included in the set to predict the exact time of recurrence by logistic regression, and 10-k-fold validation using normalized only data.

Next, genes were discretized into intervals to predict the BCR still using logistic regression. In both cases, attributes were chosen by CFS, after pre-processing. Results for the test set supplied to the newly trained regressors can be found in 3.45 and 3.46. Discretization was made randomly in this first attempt which served as a trial with logistic regression.

Table 3.45: Results from testing set without using discretization

	Linear regression
Mean absolute error	19.31
Relative absolute error	98.35 %

Table 3.46: Results from testing set using two-interval discretization

	Logistic regression
Mean absolute error	15.02
Relative absolute error	76.94 %

When regression was applied, contrasting model evaluators needed to be defined, such as the absolute square error and mean absolute error. Generally, the absolute error represents the amount of error between the predicted status and the actual one with the formula $Absoluteerror = |x_p - x_r|$. The average of all absolute errors is illustrated by the mean absolute error, where all absolute errors are calculated, added, and then divided by the number of errors. The 76% relative absolute error compares the

mean error from the model to the errors generated by a naive model. According to these results, the model did not fit our requirements, as the errors were extremely large.

3.4.2 Classifying the patients into two groups by logistic classification and CFS

To simplify the model, the patients were classified into groups according to their time of recurrence. The dataset was processed in the same manner as in the previous section: Normalized, discretized, and then passed through an attribute selector. Patients were clustered in proportionate groups, while the formulas employed previously were applied to determine the most advantageous number of intervals for the gene attributes. The training set includes 71 patients with BCR from the TCGA set, while the test encompasses 72 subjects.

Table 3.47: Table of methods to estimate number of bins for discretization

Name of the rules	Formula	Number of bins
Square root	$n^{\frac{1}{2}}$	9
Cencov	$n^{\frac{1}{3}}$	4
Rice	$2n^{\frac{1}{3}}$	8
Terrell-Scott	$(2n)^{\frac{1}{3}}$	5
Sturge	$1+\log_2 n$	7
Brooks-Carruthers	$5\log_{10} n$	9
Freedman-Diaconis	$\frac{R}{IQRn^{\frac{-1}{3}}}$	
Scott	$\frac{R}{3.5\sigma n^{\frac{-1}{3}}}$	

Table 3.48: Results from raw dataset without using discretization

	Logistic regression	
	Square root	Cencov
Correctly Classified Instances	52	58
Incorrectly Classified Instances	31	25
	Terrell-Scott	Sturge
Correctly Classified Instances	65	60
Incorrectly Classified Instances	18	23
	Freedman-Diaconis	Scott
Correctly Classified Instances	37	36
Incorrectly Classified Instances	53	54
	Rice	Brooks-Carruthers
Correctly Classified Instances	58	52
Incorrectly Classified Instances	25	31

Table 3.49: Results from normalized data using five-interval discretization

	Logistic regression	
	(0-22]	>22
(-0-22]	30	9
>22	9	35
Precision	0.769	0.795
ROC area	0.822	

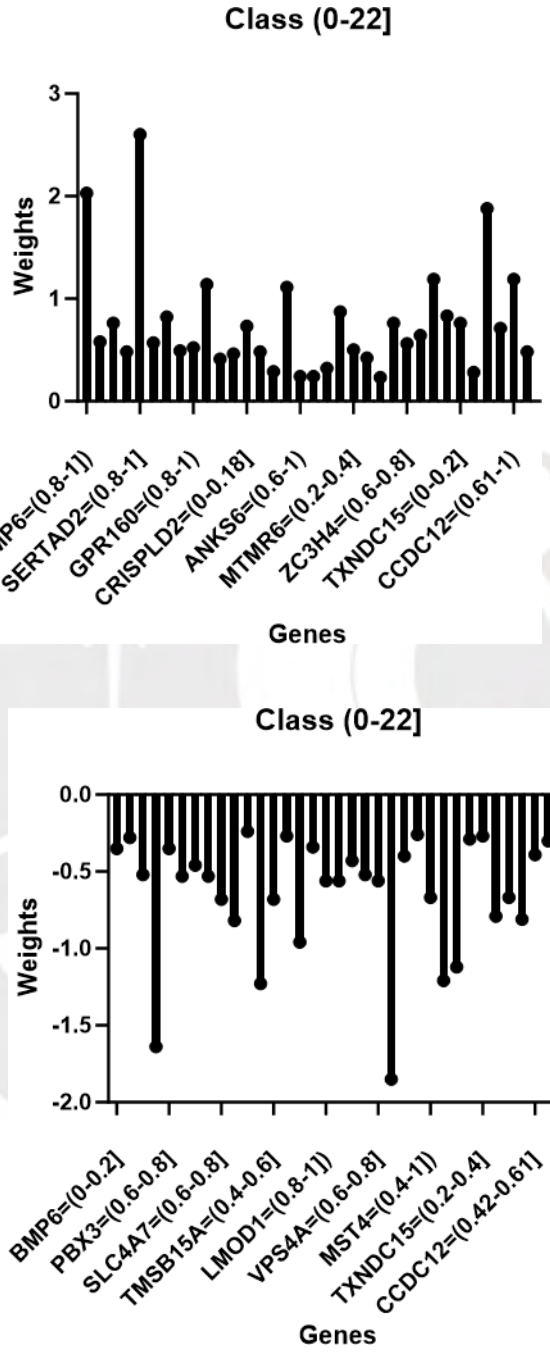


Figure 3.14: Graphic representation of the genes and associated weights to predict the risk of presenting BCR in the first 22 months

To predict the risk of presenting recurrence after 22 months, the sign of each weight must be switched. The full equations are available in the annex

By discretizing the genes into five intervals, the risk of BCR in a two-year span can be predicted as displayed in 3.14. Nevertheless, the objective was to further narrow down the groups for the time of recurrence. The gene signature is a lot more complex than the one employed to classify the status of patients, as some genes are present more than once. According to their expression and interval, different weights were assigned to them, as evidence of their importance in the classification. For instance, FAM50A is present 4 times, suggesting that in spite of its expression, the gene has an implication in the prediction. Genes SLC25A25, ZHX2, SRGAP3, and BMP6 emerge three times in the signature, also conveying their influence.

3.4.3 Classifying the patients into three groups by logistic classification and CFS

Patients were now divided into three equitable groups, while the normalized and discretized genes were selected through CFS. The following tables illustrate the optimum results.

Table 3.50: Results with discretization

	Logistic regression	
	Square root	Cencov
Correctly Classified Instances	57	44
Incorrectly Classified Instances	26	39
	Terrell-Scott	Sturge
Correctly Classified Instances	48	55
Incorrectly Classified Instances	41	28
	Freedman-Diaconis	Scott
Correctly Classified Instances	138	136
Incorrectly Classified Instances	58	60
	Rice	Brooks-Carruthers
Correctly Classified Instances	53	57
Incorrectly Classified Instances	30	26

The discretization tested above proffers unsatisfactory results since only half of the instances were classified correctly. As the time of recurrence intervals was reduced, the difficulty to predict BCR increased. Undeniably, fewer patients were part of the groups, thus interfering with the classification. On the assumption, the model was properly trained for classification, it can not be asserted as suitable, since it will only be tested through a handful of patients.

Since the classification for patients with a risk of BCR 22 months after their last treatments were concluded, the focus was on cases where the BCR is lower than 22 months. Nonetheless, from the training and test sets only 73 patients fit the requirements, 24 within the TCGA data set. Said new set, will be too modest to certify the veracity of the model.

While previous studies (Gongwei et al., 2021) (Wu et al., 2020) focus on classifying patients between low and high risk of BCR with a variation between three and five years, this work applies discretization to predict the actual time of recurrence within a two-year time period. Thus, the present approach shows promise to promote surveillance efforts of patients with more risks of presenting BCR within two years,

which can translate into improving their life expectancy and providing less burdensome treatments.

Comparatively, to the former 10 genes' signature, 42 genes participate in the prediction of the BCR. Nonetheless, a handful of genes were mentioned multiple times at different intervals, manifesting their influence in the prognostics. The majority of the repetitive genes support the working hypothesis of a role for the ECM in the BCR. For instance, BPM6, appearing twice in the final signature, has been related to the promotion of the migration and invasion of prostate cancer cells (Darby, Cross, Brown, Hamdy, & Robson, 2008). Also in other cancers, SRGAP3 contributes as a tumor suppression function in breast cancer and promotes anchorage-independent cell growth when expressed at low levels (Kazanietz & Caloca, 2017).

Although most of the genes point out the roles of the ECM in the BCR, further investigations need to focus on the remaining genes, provided the scarcity of information related to their role in tumorous tissue. With a deeper comprehension of the role of the genes, novel biomarkers of phenotypes could be proposed, and earlier signs of BCR standardized along with new indicators of the aggressiveness of cancer.

Chapter IV

Performance comparison of classifiers trained on genomic and phenotypic data

From the previous chapters, we have implemented a tool predicting the risk of recurrence with an average accuracy of 81.1%, while forecasting the risk of BCR within the first 2 years after final treatment with an accuracy of 78%.

With the goal to improve patient stratification, the phenotypic features from 2.4.3 were linked to the risk of recurrence via logistic regression analogously with the method from 3.3.7 and compared to genomic results. Since GSE and MSKCC datasets did not include whole slide information, the models generated in this chapter without exception were evaluated with 10-k-folds including only the TCGA patients.

4.1 The influence of the Gleason score on the risk of BCR

In the first experiment, only the Gleason score was investigated, as it is automatically handled in prostate cancer detection and classification. Turned into nominal values, the Gleason scores were inspected through logistic regression.

Table 4.1: Results when using Gleason score to predict the status of patients

	Logistic regression	
	With BCR	Without BCR
With BCR	59	14
Without BCR	29	41
Precision	0.808	0.586
ROC area	0.653	

Table 4.2: Results when using Gleason score to predict the time to biochemical recurrence based on number of months

	Logistic regression	
	0 - 24	> 24
0 - 24	11	62
>24	4	66
Precision	0.156	0.941
ROC area	0.510	

From this experiment, it is concluded that Gleason alone is not sufficient to predict the risk of BCR, which can be attributed mostly to the heterogeneous nature of prostate cancer.

4.2 The influence of the phenotypic features on the risk of BCR

The Gleason score is based on the heterogeneous nature of prostate cancer as it relies on leading and sub-dominant patterns. Nevertheless, Gleason exclusively resorts to the tumor architecture, ignoring significant information such as the nuclei proliferation or inter-glands spatial location that was retrieved by the approaches used in this work. Therefore, phenotypic features from 2.4.3 were added to the Gleason score.

Table 4.3: Phenotypic features retrieved with the proposed methodology

Name of phenotypic features	Methodology
Number of glands without lumen	Total number by tiles
Number of lumen discarded	Total number by tiles
Number of Unet contour with more than one lumen (fused glands & high Gleason)	Total number by tiles
Number of Unet contour with one lumen (organized tissue & low Gleason)	Total number by tiles
Morphology of the glands	Variance of the distance between the center of the lumen and membrane of the reconstructed glands
Unet contour area	mean/variance/minimum/maximum
Ratio between Unet region area to total image area	mean/variance/minimum/maximum
Density of nuclei in cluster of nuclei	mean/variance/minimum/maximum
Density of stroma in cluster of nuclei	mean/variance/minimum/maximum
Ratio lumen area to Unet contour area	mean/variance/minimum/maximum
Ratio number lumen discarded to total lumen in contour	mean/variance/minimum/maximum
Ratio reconstructed area to Unet contour	mean/variance/minimum/maximum

The number of instances was increased by five since these values were extracted from each of the five tiles for each patient considering multiple distinct cancer foci are displayed.

Table 4.4: Results from Gleason score, discretized phenotypic features, using three-intervals equal width to predict status of patients

	Logistic regression	
	With BCR	Without BCR
With BCR	292	73
Without BCR	104	246
Precision	0.737	0.771
ROC area 0.818		

Adding the phenotype features did increase the accuracy of the prediction as shown with the follow-

ing signature.

$$
\begin{aligned}
&ClassRecurrent : 0.21 + [Numberoflumendiscarded = (20.33 - 40.67]] * 0.37 + \\
&[numberofglandfused = (2.33 - 4.67]] * 0.45 + \\
&[Numberofclusterofnuclei = (19.33 - 1)] * -0.77 + \\
&[Numberofglandsnonmerged = (9107.71 - 18215.43]] * -0.18 \\
&+ [Maximumratiocontourtowholetiles = (0.01 - 0.02]] * 1.5 \\
&+ [Maximumdensityincontour = (0.21 - 1)] * -1.17 + \\
&[Meanarealumentoglandscontour = (0 - 42.66]] * -0.26 + \\
&[Maxareaglandsreconstructedtocontourarea = (558942.93 - 1117885.86]] * 0.51 + \\
&[Morphologyglandsmeans = (0 - 705.48]] * 0.62 + \\
&[Morphologyglandsvariance = (0 - 1106.61]] * -0.35 + \\
&[Morphologyglandsminimum = (2523.73 - 1)] * -0.73 + \\
&[GLEASON_SCORE = 6] * -0.91 + [GLEASON_SCORE \\
&= 8] * 0.37 + \\
&[GLEASON_SCORE = 9] * 1.16 + \\
&[GLEASON_SCORE = 10] * 1.33
\end{aligned}
\tag{4.1}
$$

Most of the selected phenotypic features by the logistic method are displayed in patients with high Gleason scores, demonstrating that patients with aggressive tumors have more risk of showing BCR. If we glance at the weights associated with the Gleason score in the recurrent class, greater weights were attached to higher Gleason scores, where Gleason 6 carries negative weights. Therefore, patients

are more likely to not present a recurrence in the future, indicating there is an influence of the Gleason scores in the recurrence, consolidating the working hypothesis suggesting that patients with more aggressive tumors are more likely to present BCR.

By considering the attributes associated with positive weights, glands with a more circular shape were more likely to present fewer risks of BCR, which concurs with the Maximum ratio contour of the whole tiles variable. According to the aforementioned variable, modest U-net segmented contours coincide with low-risk patients. For the U-net to segment small-scaled contours, glands, or clusters must be detached from each other by stroma, comparable with low Gleason score patients.

Negative weighted attributes indicate that tiles presenting BCR enclose a broad number of clusters of nuclei, contours with a high density of nuclei, and glands with small lumina with non-circular morphology. Predominantly, the proliferation of nuclei, assignable to low cell adhesion, is responsible for the risk of recurrence.

Table 4.5: Results from Gleason score and discretized phenotypic features using five intervals with equal width to the time of recurrence in months

	Logistic regression	
	0 - 24	> 24
0 - 24	107	80
> 24	61	128
Precision	0.637	0.615
ROC area	0.681	

Overall the accuracy did increase but remained mediocre to predict the time of recurrence. In the signature only two attributes were considered, the Gleason score when it's equal to 6, and the area between the discarded lumina. However, the precision is too low to confirm the aforementioned signature.

4.3 The influence of the PSA on the risk of BCR

PSA plays a major role in prostate cancer detection, as it is the first step once patients express discomfort. According to its level, a biopsy is then requested by their oncologists. It is a requirement so every patient has that information available. Nonetheless, PSA alone is not an indicator of cancer, as the protein can be released by prostate injuries, inflammations, or enlargements, can be caused by urinary tract infections, or simply as patients age (*Prostate-Specific Antigen (PSA) Test*, n.d.). Depending on the guidelines used, the limit level of PSA is set around 4 ng/ml, but others prefer to start below, at 3 or 2.5 ng/ml, as about 15% of men with a PSA level below 4 will present cancer phenotypes in their respective biopsies. Above 4 ng/ml medical professionals will likely order further screening such as DRE or biopsy. Men with a PSA level between 4 and 10 ng/ml have a 25% more chance of enduring prostate cancer, while at a level above 10 ng/ml, the chance rises to 50% (*Prostate-specific antigen (PSA) blood test*, n.d.).

Previously mentioned levels are applicable for the first detection of prostate cancer. Nonetheless, they might not be the most pertinent with recurrence. As a matter of fact, only 9 patients out of 154 have a PSA level above 4 ng/ml. Conforming to the European Association of Urology, two consecutive PSA tests with levels above 0.2 ng/ml are an indication of recurrence (Venclovas, Jievaltas, & Milonas., 2019).

Combining PSA levels with the gene expression from the confirmed 10 genes might assure the usage of genomics levels of expression to forecast the risk of recurrence. To prove the hypothesis as well as to evaluate the role of the PSA in the risk of presenting BCR, PSA levels of TCGA patients were combined with the phenotypic features summarized as Gleason scores, and genomic expression respectively to inquire about a change in the accuracy of BCR prediction via a 10-fold logistic model. For genomic data, the gene signature from 3.3.7 incorporates intervals. For all subjects, when the gene expression befalls in the interval, a value of 1 is assigned, if not, 0 is assigned to simplify the model.

Table 4.6: Results when adding the PSA levels to both phenotypic and genomic data

	Phenotypic data and PSA level		Genomic data and PSA level	
	With BCR	Without BCR	With BCR	Without BCR
With BCR	296	69	65	8
Without BCR	80	270	9	62
Precision	0.730	0.750	0.871	0.863
ROC area	0.78		0.83	

The following signature states that PSA levels below 0.04 ng/ml did not influence the prediction of recurrence, however above 0.04 the weights associated with PSA increase respectively with its value, confirming the statement from European Association of Urology 4.1.

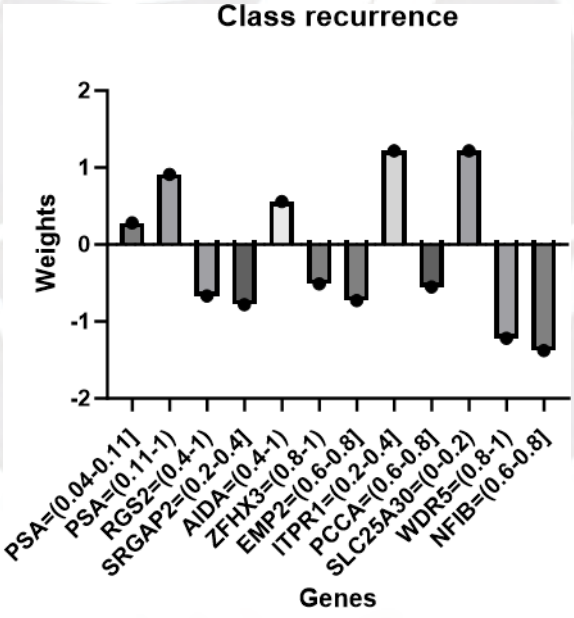


Figure 4.1: Graphic representation of the genes and PSA value associated weights to predict the risk of presenting BCR

Table 4.7: Results obtained when adding the PSA levels to both phenotypic and genomic data to predict time of recurrence in months

	Phenotype and PSA		Genomic and PSA	
	0 - 24	> 24	0 - 24	> 24
0 - 24	106	81	33	2
> 24	62	127	5	30
Precision	0.697	0.631	0.711	0.733
ROC area	0.68			

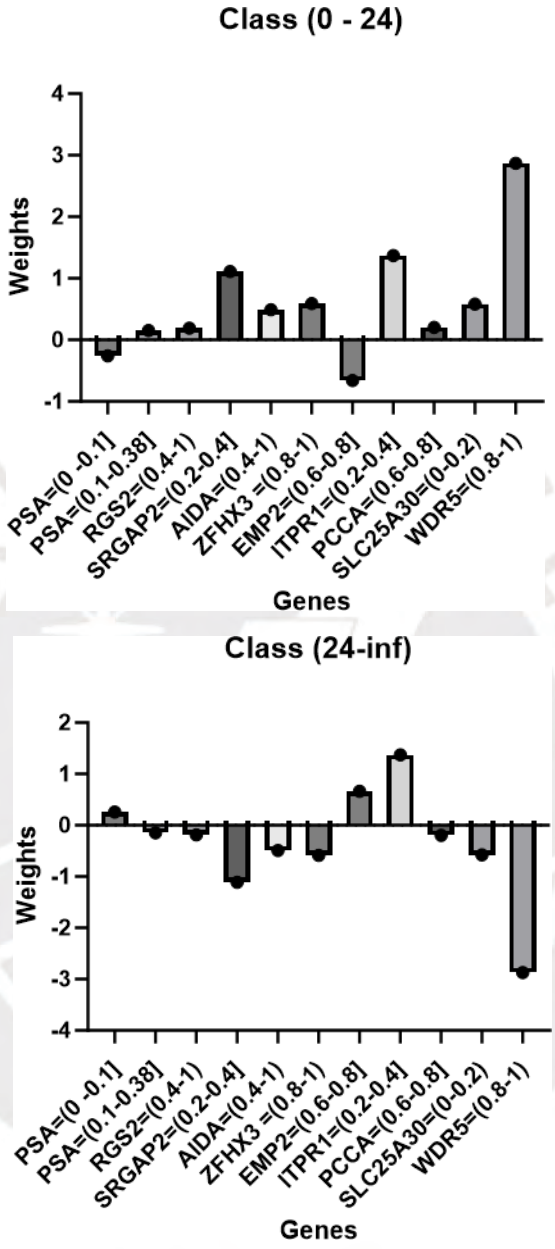


Figure 4.2: Graphic representation of the genes and PSA value with associated weights to predict the risk of presenting BCR before of after 24 months

From the table above, it is evident that PSA levels are powerful data, as the prediction of the risk of BCR increased significantly while the prediction of the time of recurrence remains equal. The addition of the PSA levels to the genetic information allows the preservation of the same gene template to predict

the risk of presenting BCR, and its actual range.

If we inspect the gene signature from 3.4.2 to predict the time of recurrence, a template of 43 genes was presented. Notwithstanding, satisfactory results were achieved with the 10 genes formerly detected, suggesting their extensive roles in the recurrence. The 10 isolated genes were qualified to predict early BCR with an accuracy of 71%. A PSA level below 0.38 ng/ml was for the most part correlated with a risk of BCR within a two-year time-lapse.

In both gene signatures 4.2, EMP2, ITPR1, AIDA, and SLC25A30 are established as evident predictors of recurrence. Genes SLC25A30 and ITPR1 are under-expressed in both signatures, while on the contrary AIDA is over-expressed. In contrast, high expression of EMP2 manifests a low risk of recurrence.

Conclusions and Recommendations

Comparatively, discretized genomic information was proven to be more efficient than the Gleason score alone to boost the accuracy of the predictions. Gleason's score relies on tissue architecture but overlooks cell proliferation and luminal knowledge, considering those key elements, the prediction of BCR was improved but remains deficient to stratify patients.

PSA levels corroborate as a significant and compelling tool in BCR prognostics. A slight variation of 0.05 ng/ml can alter the exactness of the final classification with both genomic and phenotypic attributes. Ultimately, by combining genomic information with PSA levels, the designed model was capable of predicting the status of 87% of the cohort correctly. Whilst, no standardization rule has been extracted in this research as to the optimal number of intervals for discretization, five-bin equal-width discretization provides the best results in both BCR status classification and BCR time prediction.

In the present study, genomic biomarkers associated with the risk for the patients to present BCR were uncovered. Those 10 genes, though not usually associated with prostate cancer, provide significant information about the role of cell adhesion proteins in the recurrence, reinforcing the gland's structure as the focal point of prognosis and diagnosis. A 10 genes signature ground on tumor-adjacent normal tissue by Rui (Rui et al., 2021) was explicitly linked to cell-to-cell signaling, an element of the ECM (Brownlee, 2002).

The main objective of this study was to evaluate how a template of genes expresses itself phenotypically. Spatial gene expression is an ongoing challenge in the medical field. Because the phenotype could be forecasted from gene expression, patients could be treated with adapted strategies.

Whilst technological advancements in the medical field keep on expanding, spatial gene expression

remains an arduous task considering that genes can not be individually pinpointed but understood as a set with fluctuating correlations. The initial strategy was to link a set of phenotypic features, to the genes via multi-output classifiers. The results, an accuracy oscillating between 48% and 52%, were deceiving as no classifier actually succeed in detecting the correlation between the attributes.

A fully connected neural network, with considerable depth, might have been more fitted for such a complex assignment for future reference. As an alternative solution, a deep learning model can also be trained, with the WSI as an input and the prediction of the 10 genes expression as an output. Precise prediction of the gene expression is demanding, and might not provide the required results. A possible course of action consists of classifying the WSI according to the discretized gene expression, which coincides with the final signature discovered during this research. Once trained, the features extracted by the deep learning model should be analyzed to isolate phenotypic biomarkers, and gene expression spatially distinguished. This approach presented different challenges, mainly due to the difficulty to work directly with WSI, given their large sizes and with high morphological heterogeneity (Dimitriou, Arandjelovic´, & Caie, 2019). The approach employed in this investigation, restricted the phenotype to 5 tiles, overlooking information and phenotypes outside those selected tiles. WSI, on the other hand, includes the entire tissue, introducing more phenotypic elements to the model. Along with the dataset, the complexity, and architecture of the deep learning model will be challenging, as the correlation between the WSI and gene expression might not be discernible by a simple model.

Maneuvering with WSI from biopsy did not simplify the approach, as spatial gene expression has first been introduced via in silico investigation. Gene expression values are extracted from the whole slide, hence pinpointing where the activity is occurring exactly can be arduous. Inadequate phenotypic feature predictions can be attributed to the fact that recurrence can happen outside the prostate glands. Further, studies need to be concluded taking into account the presence of metastases in the bladder, lymph node, presented in the clinical data, bone, or other surrounding organs. Genomic information could also be inquired to inspect the role of their expression on the location of metastases (Patel, Patel, & Siddiqui, 2015),(Mahdy, Patil, & Parajuli, 2019), and (Craig, Woulfe, Sinclair, & Malone, 2015).

The proposed phenotype protocol can also assist oncologists in automatically indicating the region

of interest in the WSI and pointing down tumorous architecture as well as segmenting the glands, nuclei, and clusters of nuclei.

The application of computer-aided diagnosis, CAD, tools using multi-parametric MRI or Quantitative Ultrasound (C. Wei et al., 2019) have shown progress in the detection and classification of cancer (Giannini, Mazzetti, Cappello, & et al., 2021), (Hambrock, Vos, van de Kaa, Barentsz, & Huisman, 2013). In the majority of prostate glands, CAD relies on MRI imaging, as less invasive, but biopsies CAD may be an option for more complex cases. Oncologists rely heavily on biopsies to detect and identify the aggressiveness of prostate cancer, for their reliability, effectiveness, and cost-efficient prospects. Despite, the accuracy of the lessened BCR prediction with phenotype variables, it could provide an alternative to health facilities that may not have the financial or technological resources to comply with genomic studies. The Unet embedded solution is a great alternative to warrant correct histopathological segmentation for distinct cancer types, as it is easily trained with scatter datasets. In low-resource sanitary systems of Latin America, LA, the lack and unequal distribution of highly-qualified healthcare providers (Goss, Lee, & et al., 2013) (Sussman et al., 2022) lengthens the diagnosis waiting time and directly diminishes the quality of the medical care (Strasser-Weippl, Chavarri-Guerra, & et al., 2015). Prediction of the risk of prostate BCR in LA is persistently bounded by the inadequate training of caregivers and excessive cost and wait lists for cancer treatment (Oliveira, de Lorena Sobrinho, & da Cruz Gouveia Mendes, 2022).

The proposed solution, combining advanced artificial intelligence (AI) methods and histological images, improves personalized medicine and boosts clinical expertise (Guo & Li, 2018). Its straightforward and practical application can distribute uniformly access to the technologies required for cancer diagnosis and prostate BCR prediction(Reis et al., 2020).

The acquired knowledge from the trained Unet architecture model to segment glands, clusters of nuclei, and nuclei could also be applied to related challenges, and generalized to other types of cancer by transfer learning. Rare histological variants of prostate adenocarcinoma, without a large available dataset for training, could also benefit from transfer learning, as the Unet model is already built and trained for similar tasks. Hence the corresponding weights could be employed as the starting point to

learn and segment rare histological patterns (Lee, Miller, & Epstein, 2010) (Kumar & Mukherjee, 2010) (Osunkoya, 2018).

As it is right now, the methodology proposed here segments the features but does not attribute a Gleason score to the patients. An additional model can be trained with the same dataset and extracted features to predict the Gleason score from WSI, hence proposing a complete tool, from segmentation to classification to oncologist. The proposed model would highlight in the WSI the tumorous regions and associated Gleason score, providing a quantitative estimation of the disease enabling efficient and reliable clinical decisions. The methodology enables the comprehension of the role of morphological features in the recurrence.

Over the past years, various approaches have been proposed to segment prostate glands, from texture analysis to the use of deep learning, demonstrating the relevance and need of this work. The main strength of the methodology discussed in this thesis resides in the reconstruction of the glands. When using deep learning, most studies use the contour generated by the model as final segmentation (Salvi et al., 2021a), which provides factual results with low Gleason patients. With highly heterogeneous and substantial variation in appearance, segmentation of prostate glands always remains a challenge. Ren (Ren, Sadimin, Foran, & Qi, 2017), proposed an approach with 83% of accuracy, superposing CNN segmented glandular region with superpixel regions. Meanwhile, Massimo (Salvi et al., 2021b) combined CNN and active contours for a precision of 91%. Xu (Xu et al., 2016) suggested a three-channel CNN, assigned to segment foreground from background pixels, perceived gland boundaries, and detected individual glands, respectively. Nevertheless, clear and well-defined contours, additionally with single lumina are required for a successful implementation of these approaches. To resolve these issues, Yali (Yali et al., 2022) advanced a pyramid semantic parsing network, predominantly focused on a binary classification between non-tumorous and tumorous regions, which would perform poorly on contiguous glands or clusters of nuclei. The developed U-net model can segment glands without discernible lumen, often neglected (L. Chan, Hosseini, Rowsell, Plataniotis, & Damaskinos, 2019) (Jia, Huang, Chang, & Y.Xu, 2017) by other studies. New efforts in computer vision have shown impressive results in histopathology image segmentation. Implementation of Deeplabv3 models exhibited high accuracy in

Gastric Cancer Segmentation (M. Sun et al., 2019) but was unfit on invariant stain pattern. Superior segmentation performances might be attained with Vision Transformers (ViT) algorithms (Ikromjanov et al., 2022) fused with ANN (Y. Zhang, Liu, & Hu, 2021). Nevertheless, considerate annotated datasets are required and the substantial scarce interpretability of ViT (K. He et al., 2021) hinders their application for this study. From its simple implementation and efficient training, the U-net exhibits great potential for transfer learning on diverse image sources. Moreover, patients with a Gleason of 7 are usually problematic as in between category (Sim et al., 2008), heavily dependent on the predominant pattern. Our models overcome these hurdles by engaging specific morphological attributes.

The gland reconstruction can help professionals to have access to data related to the lumen, or the proximity of the glands. Future work needs to be directed at ECM characteristics, that are responsible for the tissue architecture, targeting the stroma surrounding the glands and its composition. In addition, other commonly disregarded features of the prostate gland such as the presence of collagen fibers, fibroblasts, and blood vessels associated with cell adhesion quality, might prove critical to fully understanding cancer growth patterns and risk of recurrence. Studying and collecting collagen changes in the tissue (J. Huang et al., 2021), might facilitate the prediction of the time of recurrence.

This study, by integrating a variety of specimens in the cohort, with patients graded from Gleason 5 to 10 adds considerably to the knowledge. Methodology, affordable and suitable for a plethora of health facilities, to point out tumorous region from WSI has yet to be presented in the literature. An Unet model is advanced, efficient to segment single and cluster of nuclei, and glands within the same model and hyper-parameters, easy to use and to compile without relying on a substantial data set. The reconstruction of the glands allows professionals to inspect patients in between Gleason and highly heterogeneous biopsies. Moreover, the 10 genes signature advanced, reinforces the newly-uncovered role of the ECM in the recurrence, while providing an in-depth analysis of the preprocessing step for omic cohorts.

References

Abdulla, W. (2018). Mask r-cnn for object detection and instance segmentation on keras and tensorflow.

Ali, P., Faraj, R., & Koya, E. (2014). Data normalization and standardization: a technical report. *Mach Learn Tech Rep*. doi: 10.13140/RG.2.2.28948.04489

Alkharabsheh, K., Alawadi, S., Kebande, V., Crespo, Y., Fernández-Delgado, M., & Taboada, J. (2022). A comparison of machine learning algorithms on design smell detection using balanced and imbalanced dataset: A study of god class. *Information and Software Technology*. doi: 10.1016/j.infsof.2021.106736

Al-Kofahi, Y., Lassoued, W., Lee, W., & Roysam, B. (2010). Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Trans Med Imaging*. doi: 10.1109/TBME.2009.2035102

American cancer society. (n.d.). Retrieved from `https://www.prostateconditions.org/~{}about-prostate-conditions/prostate-cancer/newly-diagnosed/gleason-score`

Banerjee, C., Mukherjee, T., & Pasiliao, E. J. (2019). An empirical study on generalizations of the relu activation function. *Proceedings of the 2019 ACM Southeast Conference*. doi: 10.1145/3299815.3314450

Beinecke, J., & Heider, D. (2021). Gaussian noise up-sampling is better suited than smote and

adasyn for clinical decision making. *BioData Mining*. doi: 10.1186/s13040-021-00283-6

Brownlee, C. (2002). Role of the extracellular matrix in cell-cell signalling: paracrine paradigms. *Curr Opin Plant Biol*. doi: 10.1016/s1369-5266(02)00286-8

C., N., & Subhrasankar, C. (2019). A novel approach to age classification from hand dorsal images using computer vision. *ICCM*.

cancer connect. (n.d.). Retrieved from `https://news.cancerconnect.com/prostate-cancer/prostate-cancer-what-you-need-to-know-about-the-gleason-score`

Caselles, V., Kimmel, R., & Sapiro, G. (1997). Geodesic active contours. *International Journal of Computer Vision*. doi: 10.1109/ICCV.1995.466871

Cazals, F., & Giesen, J. (2004). Delaunay triangulation based surface reconstruction: Ideas and algorithms. *INRIA*. doi: 10.1007/978-3-540-33259-6_6

Cebeci, Z., & Yildiz, F. (2017). Unsupervised discretization of continuous variables in a chicken egg quality traits dataset. *Turkish Journal of Agriculture - Food Science and Technology*. doi: 10.24925/turjaf.v5i4.315-320.1056

Chan, H., Chattopadhyay, A., Chuang, E., & Lu, T. (2021). Development of a gene-based prediction model for recurrence of colorectal cancer using an ensemble learning algorithm. *Frontiers in Oncology*. doi: 10.3389/fonc.2021.631056

Chan, L., Hosseini, M., Rowsell, C., Plataniotis, K., & Damaskinos, S. (2019). Histosegnet: Semantic segmentation of histological tissue type in whole slide images. *IEEE International Conference on Computer Vision*. doi: 10.1109/ICCV.2019.01076

Cheaito, K., Bahmad, H., Hadadeh, O., & et al. (2019). Emt markers in locally-advanced prostate cancer: Predicting recurrence. *Front Oncol*. doi: 10.3389/fonc.2019.00131

Chen, X., Zhang, F., & Zhang, R. (2017). Medical image segmentation based on slic superpix-

els model. *Computer Science*.

Cheng, H., Garrick, D., & Fernando, R. (2017). Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. *Sci Biotechnol*. doi: 10.1186/s40104-017-0164-6

chestnutappeal. (n.d.). Retrieved from https://chestnutappeal.org.uk/prostate-cancer-is-the-most-common-cancer-in-men-yet-so-few-know-where-it-is/

Chu, J., Cai, J., Song, H., Zhang, Y., & Wei, L. (2020). A novel bilinear feature and multi-layer fused convolutional neural network for tactile shape recognition. *Sensors*.

Chu, J., Li, N., & Gai, W. (2018). Identification of genes that predict the biochemical recurrence of prostate cancer. *Oncol Lett*. doi: 10.3892/ol.2018.9106

Cordon-Cardo, C., Kotsianti, A., Verbel, D., & et al. (2007). Improved prediction of prostate cancer recurrence through systems pathology. *J Clin Invest*. doi: 10.1172/JCI31399.

Craig, J., Woulfe, J., Sinclair, J., & Malone, S. (2015). Isolated brain metastases as first site of recurrence in prostate cancer: case report and review of the literature. *Curr Oncol*. doi: 10.3747/co.22.2542

Darby, S., Cross, S., Brown, N., Hamdy, F., & Robson, C. (2008). Bmp-6 over-expression in prostate cancer is associated with increased id-1 protein and a more invasive phenotype. *The Journal of Pathology*. doi: 10.1002/path.2292

de Sousa I, P., Vellasco, M. B. R., & da Silva, E. C. (2019). Local interpretable model-agnostic explanations for classification of lymph node metastases. *Sensors (Basel)*. doi: 10.3390/s19132969

Dimitriou, N., Arandjelović, O., & Caie, P. (2019). Deep learning for whole slide image analysis: An overview. *Frontiers Medicine*. doi: 10.3389/fmed.2019.00264

Duca, R., Massillo, C., Dalton, G., Farré, P., Graña, K., Gardner, K., & Siervi, A. D. (2021).

Mir-19b-3p and mir-101-3p as potential biomarkers for prostate cancer diagnosis and prognosis. *American journal of cancer research*.

Epstein, J., Allsbrook, W., Amin, M., & LL.Egevad. (2005). The 2005 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. *The American Journal of Surgical Pathology*. doi: 10.1097/01.pas.0000173646.99337.b1

Espichan, A., & Villanueva, E. (2018). A novel ensemble method for high-dimensional genomic data classification. *IEEE International Conference on Bioinformatics and Biomedicine*. doi: 10.1109/BIBM.2018.8621386

Frantz, C., Stewart, K., & Weaver, V. (2010). The extracellular matrix at a glance. *Journal of Cell Science*. doi: 10.1242/jcs.023820

Gamarra, M., Eduardo, Z., Banerjeelante, H., Hurtado, L., & San-Juan-Vergara4, H. (2019). Split and merge watershed: a two-step method for cell segmentation in fluorescence microscopy images. *Biomed Signal Process Control*. doi: 10.1016/j.bspc.2019.101575

Gay, H., & Michalski, J. (2018). Radiation therapy for prostate cancer. *Mo Med*.

Giannini, V., Mazzetti, S., Cappello, G., & et al. (2021). Computer-aided diagnosis improves the detection of clinically significant prostate cancer on multiparametric-mri: A multi-observer performance study involving inexperienced readers. *Diagnostics (Basel)*. doi: 10.3390/diagnostics11060973

Global cancer observatory. (n.d.). Retrieved from `https://gco.iarc.fr/`

Gomella, L., Singh, J., Lallas, C., & Trabulsi, E. (2010). Hormone therapy in the management of prostate cancer: evidence-based approaches. *Therapeutic Advances in Urology*. doi: 10.1177/1756287210375270

Gongwei, L., Wei, O., Yucong, Z., Guoliang, S., Jiahua, G., Zhiquan, H., & Heng, L. (2021). Identification of a dna repair gene signature and establishment of a prognostic nomogram

predicting biochemical-recurrence-free survival of prostate cancer. *Frontiers in Molecular Biosciences*. doi: 10.3389/fmolb.2021.608369

Goss, P., Lee, B., & et al., T. B.-C. (2013). Planning cancer control in latin america and the caribbean. *Lancet Oncol*. doi: 10.1016/S1470-2045(13)70048-2

Grabowska, M., Kelly, S., Reese, A., Cates, J., Case, T., Zhang, J., . . . Matusik, R. (2015). Nfib regulates transcriptional networks that control the development of prostatic hyperplasia. *Endocrinology*. doi: 10.1210/en.2015-1312

Guo, J., & Li, B. (2018). The application of medical artificial intelligence technology in rural areas of developing countries. *Health Equity*. doi: 10.1089/heq.2018.0037

G.Wang, Mang, S., H.Cai, & et al. (2016). Integrated watershed management: evolution, development and emerging trends. *J. For. Res.*. doi: 10.1007/s11676-016-0293-3

Hambrock, T., Vos, P., van de Kaa, C. H., Barentsz, J., & Huisman, H. (2013). Prostate cancer: computer-aided diagnosis with multiparametric 3-t mr imaging–effect on observer performance. radiology. *Radiology*. doi: 10.1148/radiol.12111634

He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., . . . Shen, D. (2021). Transformers in medical image analysis: A review. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. doi: 10.1016/j.imed.2022.07.002

He, K., Gkioxari, G., & Dollar, P. (2017). Mask r-cnn. *Journal of Chongqing University of Posts and Telecommunications, Natural Science Edition*. doi: 10.48550/arXiv.1703.06870

He, L., Long, L., Antani, S., & Thoma, G. (2021). A cervical histopathology dataset for computer aided diagnosis of precancerous lesions. *IEEE Trans Med Imaging*. doi: 10 .1109/TMI.2021.3059699

Hu, Q., Zhang, B., & Chen, R. (2019). Zfhx3 is indispensable for erβ to inhibit cell proliferation via myc downregulation in prostate cancer cells. *Oncogenesis*. doi: 10.1038/s41389-019 -0138-y

Huang, J., Zhang, L., Wan, D., Zhou, L., Zheng, S., Lin, S., & Qiao, Y. (2021). Extracellular matrix and its therapeutic potential for cancer treatment. *Signal Transduct Target Ther*.

Huang, S., Cai, N., Pacheco, P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (svm) learning in cancer genomics. *Cancer genomics proteomics*. doi: 10.21873/cgp.20063

Humans image segmentation with unet using tensorflow keras. (n.d.). Retrieved from https://medium.com/analytics-vidhya/humans-image-segmentation-with-unet-using-tensorflow-keras-fd6cb43b06e5

Hyun, L., Bo, K., Cheng-Kun, Y., Chao-Yuan, Y., & Jongmin, L. (2021). Measurement of laryngeal elevation by automated segmentation using mask r-cnn. *Medicine*. doi: 10.1097/MD.0000000000028112

Ibm developer. (n.d.). Retrieved from https://developer.ibm.com/~{}technologies/data-science/articles/an-automatic-method-to-identify-tissues-from-big-whole-slide-images-pt

Ikromjanov, K., Bhattacharjee, S., Hwang, Y., Sumon, R., Kim, H., & Choi, H. (2022). Whole slide image analysis and detection of prostate cancer using vision transformers. *International Conference on Artificial Intelligence in Information and Communication*. doi: 10.1109/ICAIIC54071.2022.9722635

Improve your model performance using cross validation. (n.d.). Retrieved from https://www.analyticsvidhya.com/blog/2018/05/improve-model-performance-cross-validation-in-python-r/

Indolia, S., & Goswami, A. (2018). Conceptual understanding of convolutional neural network-a deep learning approach. *Procedia Computer science*. doi: 10.1016/j.procs.2018.05.069

Jia, Z., Huang, X., Chang, E., & Y.Xu. (2017). Constrained deep weak supervision for histopathology image segmentation. *IEEE Trans Med Imaging*. doi: 10 .48550/ arXiv.1701.00794

Jung, S., Bi, Y., & Davuluri, R. (2015). Evaluation of data discretization methods to derive platform independent isoform expression signatures for multi-class tumor subtyping. *BMC Genomics*. doi: 10.1186/1471-2164-16-S11-S3

Jung, S., Bi, Y., & Davuluri, R. (2019). Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines. *bioRxiv*. doi: 10.1101/743138

Kazanietz, M., & Caloca, M. (2017). The rac gtpase in cancer: From old concepts to new paradigms. *Cancer Res*. doi: 10.1158/0008-5472.CAN-17-1456

Kesch, C., Heidegger, I., Kasivisvanathan, V., Kretschmer, A., Marra, G., Preisser, F., . . . Gandaglia, G. (2021). Radical prostatectomy: Sequelae in the course of time. *Frontiers in Surgery*. doi: 10.3389/fsurg.2021.684088

Kim, J., T.Banerjee, Vinckevicius, A., Luo, Q., Parker, J., Baker, M., . . . Chakravarti, D. (2014). A role for wdr5 in integrating threonine 11 phosphorylation to lysine 4 methylation on histone h3 during androgen signaling and in prostate cancer. *Molecular Cell*. doi: 10.1016/j.molcel.2014.03.043

Komisarof, J., McCall, M., & Newman, L. (2017). A four gene signature of recurrent prostate cancer. *Oncotarget*. doi: 10.18632/oncotarget.13837

Kumar, A., & Mukherjee, S. (2010). Metastatic ductal carcinoma of the prostate: a rare variant responding to a common treatment. *Can Urol Assoc J*.

Lee, T., Miller, J., & Epstein, J. (2010). Rare histological patterns of prostatic ductal adenocarcinoma. *Pathology*.

Leslie, S., Chargui, S., & Stormont, G. (2022). Transurethral resection of the prostate. *StatPearls*.

Li, J., Cheng, J., & Shi, J. (2012). Brief introduction of back propagation neural network algorithm and its improvement. *Advances in Intelligent and Soft Computing*. doi: 10 .1007/978-3-642-30223-7_87

Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. *arXiv*.

Long, Q., Johnson, B., & Osunkoya, A. (2011). Protein-coding and microrna biomarkers of recurrence of prostate cancer following radical prostatectomy. *Am J Pathol*. doi: 10.1016/j.ajpath.2011.03.008

Mahdy, A., Patil, R., & Parajuli, S. (2019). Biochemical recurrence in prostate cancer and temporal association to bone metastasis. *Am J Case Rep*. doi: 10.12659/AJCR.918569

Merk. (n.d.). Retrieved from `https :// www .sigmaaldrich .com / FR / fr/technical -documents/technical -article/cell -culture -and -cell -culture -analysis / 3d -cell -culture / ecm -gel -product -protocols`

Moh, M., & Shen, S. (2019). The roles of cell adhesion molecules in tumor suppression and cell migration. *Cell Adh Migr*. doi: 10.4161/cam.3.4.9246

Morash, C., Tey, R., Agbassi, C., Klotz, L., McGowan, T., Srigley, J., & Evans, A. (2015). Active surveillance for the management of localized prostate cancer: Guideline recommendations. *Canadian Urological Association Journal*. doi: 10.5489/cuaj.2806

Nader, R., Amm, J. E., & Aragon-Ching, J. (2018). Role of chemotherapy in prostate cancer. *Asian J Androl*. doi: 10.4103/aja.aja_40_17

Naik, S., Doyle, S., & Feldman, M. (2007). Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information.

Nallanthighal, S., Heiserman, J., & Cheon, D. (2019). The role of the extracellular matrix in

cancer stemness. *frontiers*. doi: 10.3389/fcell.2019.00086

National cancer institute. (n.d.-a). Retrieved from https://seer.cancer.gov/~{}statfacts/html/prost.html

National cancer institute. (n.d.-b). Retrieved from https://seer.cancer.gov/~{}types/prostate/psa-fact-sheetl

National cancer institute. (n.d.-c). Retrieved from https://www.prostateconditions.org/~{}about-prostate-conditions/prostate-cancer/newly-diagnosed/gleason-score

Nguyen, K., Sarkar, A., & Jain, A. (2012). Structure and context in prostatic gland segmentation and classification. *Structure and Context in Prostatic Gland Segmentation and Classification*. doi: 10.1007/978-3-642-33415-3_15

Nuclei counting and segmentation. (n.d.). Retrieved from https://github.com/matterport/Mask_RCNN/tree/master/samples/nucleus

Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. *ArXiv*. doi: 10.48550/arXiv.1811.03378

Oh, J., Park, S., Lee, S., Hong, S., Lee, S., Kim, T., . . . Byun, S. (2017). Genetic risk score to predict biochemical recurrence after radical prostatectomy in prostate cancer: prospective cohort study. *Oncotarget*. doi: 10.18632/oncotarget.18275

Oliveira, F., de Lorena Sobrinho, J., & da Cruz Gouveia Mendes, A. (2022). Profile of judicialization in access to antineoplastic drugs and their costs: a cross-sectional, descriptive study based on a set of all lawsuits filed between 2016 and 2018 in a state in the northeast region of brazil. *BMC Public Health 22*. doi: 10.1186/s12889-022-14199-1

Osunkoya, A. (2018). Mucinous and secondary tumors of the prostate. *Mod Pathol 31*.

Pan, D., Zeng, A., Jia, L., Huang, Y., & Song, X. (2020). Detection of alzheimer's disease

using magnetic resonance imaging: A novel approach combining convolutional neural networks and ensemble learning. *Front Neurosci*. doi: 10.3389/fnins.2020.00259

Patel, P., Patel, J., & Siddiqui, S. (2015). Recurrence of prostate cancer with cutaneous metastasis after radical prostatectomy. *Case Reports in Urology*. doi: 10.1155/2015/825175

Prostate-specific antigen (psa) blood test. (n.d.). Retrieved from `https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging`

Prostate-specific antigen (psa) test. (n.d.). Retrieved from `https://www.cancer.gov/types/prostate/psa-fact-sheet`

Punn, N. S., & Agarwal, S. (2020). Inception u-net architecture for semantic segmentation to identify nuclei in microscopy cell images. *ACM Transactions on Multimedia Computing, Communications, and Applications*. doi: 10.1145/3376922

Rajendran, K., Jayabalan, M., & Thiruchelvam, V. (2020). Predicting breast cancer via supervised machine learning methods on class imbalanced data. *International Journal of Advanced Computer Science and Applications*. doi: 10.14569/IJACSA.2020.0110808

Raschka, S. (2018). Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*. doi: 10.21105/joss.00638

Rashid, A. (2019). Applicaciones de deep-learning a la predicción de radación solar. doi: 10.19053/01211129.v29.n54.2020.11751

Reis, R., Alías-Melgar, A., Martínez-Cornelio, A., Neciosup, S., Sade, J., Santos, M., & Villoldo, G. (2020). Prostate cancer in latin america: Challenges and recommendations. cancer control. *Cancer Control*. doi: 10.1177/1073274820915720

Ren, J., Sadimin, E., Foran, D., & Qi, X. (2017). Computer aided analysis of prostate histopathology images to support a refined gleason grading system. *Medical Imaging 2017: Image Processing*. doi: 10.1117/12.2253887

Review: Fpn — feature pyramid network. (n.d.). Retrieved from https://towardsdatascience.com/review-fpn-feature-pyramid-network-object-detection-262fc7482610

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Computer Vision and Pattern Recognition*. doi: 10.48550/arXiv.1505.04597

Rui, Z., Yuanfa, F., Jianheng, Y., Zhaodong, H., Yuxiang, L., Qingbiao, C., . . . Weide, Z. (2021). Prediction of biochemical recurrence-free survival of prostate cancer patients leveraging multiple gene expression profiles in tumor microenvironment. *Frontiers in Oncology*. doi: 10.3389/fonc.2021.632571

Salvi, M., Bosco, M., Molinaro, L., Gambella, A., Papotti, M., Acharya, U., & Molinari, F. (2021a). A hybrid deep learning approach for gland segmentation in prostate histopathological images. *Artificial Intelligence in Medicine*. doi: 10.1016/j.artmed.2021.102076

Salvi, M., Bosco, M., Molinaro, L., Gambella, A., Papotti, M., Acharya, U., & Molinari, F. (2021b). A hybrid deep learning approach for gland segmentation in prostate histopathological images. *Artificial Intelligence in Medicine*. doi: 10.1016/j.artmed.2021.102076

Shah, S., Kasukurthi, N., & Pande, H. (2019). Dynamic region proposal networks for semantic segmentation in automated glaucoma screening. *ISBI 2019*. doi: 10.1109/ISBI.2019.8759171

Sharma, S., & Watabe, K. (2014). Biomarkers and mechanisms associated with recurrent prostate cancer. *Front Biosci*. doi: 10.2741/4211

Sim, H., Telesca, D., Culp, S., Ellis, W., Lange, P., True, L., & Lin, D. (2008). Tertiary gleason pattern 5 in gleason 7 prostate cancer predicts pathological stage and biochemical recurrence. *J Urol*.

Song, Y., & Lu, Y. (2015). Mlxtend: Providing machine learning and data science utilities

and extensions to python's scientific computing stack. *Shangai Arch Psychiatry*. doi: 10.21105/joss.00638

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochem Med*. doi: 10.11613/BM.2014.003

Srivastava, N., Hinton, E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*.

Stephenson, A., Smith, A., & Kattan, M. (2005). Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *cancer*. doi: 10.1002/cncr.21157

Stewart, D., Cooper, C., & Sikes, R. (2004). Changes in extracellular matrix (ecm) and ecm-associated proteins in the metastatic progression of prostate cancer. *Reproductive biology and endocrinology*. doi: 10.1186/1477-7827-2-2

Strasser-Weippl, K., Chavarri-Guerra, Y., & et al., C. V.-G. (2015). Progress and remaining challenges for cancer control in latin america and the caribbean. *Lancet Oncol*. doi: 10.1016/S1470-2045(15)00218-1

Suarez, V., & Segura, I. (2018). Evaluation of pooling operations in convolutional architectures for drug-drug interaction extraction. *BMC Bioinformatics*. doi: 10.1186/s12859-018-2195-1

Sun, M., Zhang, G., Dang, H., Xingqun, Q., Zhou, X., & Chang, Q. (2019). Accurate gastric cancer segmentation in digital pathology images using deformable convolution and multi-scale embedding networks. *IEEE Trans Med Imaging*. doi: 10.1109/ACCESS.2019.2918800

Sun, X., Xing, C., Fu, X., Li, J., Zhang, B., Frierson, H., & Dong, J. (2015). Additive effect of zfhx3/atbf1 and pten deletion on mouse prostatic tumorigenesis. *Journal of Genetics and Genomics*. doi: 10.1016/j.jgg.2015.06.004

Sussman, L., Garcia-Robledo, J., Ordóñez-Reyes, C., Forero, Y., Mosquera, A., Ruíz-Patiño, A., . . . Cardona, A. (2022). Integration of artificial intelligence and precision oncology in latin america. *Front Med Technol*. doi: 10.3389/fmedt.2022.1007822

Tammina, S. (2019). Transfer learning using vgg16 with deep convolutional neural network for classifying images. *International Journal of Scientific and Research Publications*.

Tasci, E., Zhuge, Y., Camphausen, K., & Krauze, A. (2022). Bias and class imbalance in oncologic data-towards inclusive and transferrable ai in large scale oncology data sets. *Cancers (Basel)*. doi: 10.3390/cancers14122897

Tike, A., & Tavarageri, S. (2017). A medical price prediction system using hierarchical decision trees. *2017 IEEE International Conference on Big Data (Big Data)*.

Venclovas, Z., Jievaltas, M., & Milonas., D. (2019). Significance of time until psa recurrence after radical prostatectomy without neo- or adjuvant treatment to clinical progression and cancer-related death in high-risk prostate cancer patients. *Front. Oncol.*. doi: 10.3389/fonc.2019.01286

Venkatesan, A., Mudairu-Dawodu, E., & Duran, C. (2021). Detecting recurrent prostate cancer using multiparametric mri, influence of psa and gleason grade. *Cancer Imaging*. doi: 10.1186/s40644-020-00373-4

Verma, M., & Patel, P. (2011). Biomarkers in prostate cancer epidemiology. *Cancers (Basel)*. doi: 10.3390/cancers3043773

Vino, G., & Sappa, A. (2013). Revisiting harris corner detector algorithm: A gradual thresholding approach. *ICIAR*. doi: 10.1007/978-3-642-39094-4_40

web pathology. (n.d.). Retrieved from `https://www.webpathology.com`

Wei, C., Zhang, Y., Malik, H., Zhang, X., Alqahtani, S., Upreti, D., . . . Nabi, G. (2019). Prediction of postprostatectomy biochemical recurrence using quantitative ultrasound shear wave elastography imaging. *Frontiers Oncology*. doi: 10.3389/fonc.2019.00572

Wei, Q., & Dunbrack, R. (2013). The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS One*. doi: 10.1371/journal.pone.0067863

Winkler, J., Abisoye-Ogunniyan, A., & Metcalf, K. (2017). Concepts of extracellular matrix remodelling in tumour progression and metastasis. *Cancer research*. doi: 10 .1038/ s41467-020-18794-x

Wu, X., Lv, D., Lei, M., Cai, C., Zhao, Z., Eftekhar, M., . . . Liu, Y. (2020). A 10-gene signature as a predictor of biochemical recurrence after radical prostatectomy in patients with prostate cancer and a gleason score ≥ 7. *Oncology letters*. doi: 10.3892/ol.2020 .11830

W. Wegier, P. K. (2020). Application of imbalanced data classification quality metrics as weighting methods of the ensemble data stream classification algorithms. *Entropy (Basel)*. doi: 10.3390/e22080849

Xu, Y., Li, Y., Wang, Y., Liu, M., Fan, Y., Lai, M., & Chang, E. (2016). Gland instance segmentation using deep multichannel neural networks. *IEEE transactions on bio-medical engineering*. doi: 10.48550/arXiv.1611.06661

Yali, Q., H.Yujin, Peiyao, K., Hai, X., Xiaoliu, Z., Jiuwen, C., . . . Baiying, L. (2022). Automatic prostate gleason grading using pyramid semantic parsing network in digital histopathology. *Frontiers in Oncology*. doi: 10.3389/fonc.2022.772403

Zaidan, M., Dada, L., Alghamdi, M., Al-Jeelani, H., Lihavainen, H., Hyvärinen, A., & Hussein, T. (2019). Mutual information input selector and probabilistic machine learning utilisation for air pollution proxies. *Applied Sciences*. doi: 10.3390/app9204475

Zero cancer. (n.d.). Retrieved from `https :// zerocancer .org / ~{}learn / survivors/recurrence`

Zhang, Q., Wu, Y. N., & Zhu, S. (2017). Interpretable convolutional neural networks. *Journal of Chongqing University of Posts and Telecommunications, Natural Science Edition*. doi:

10.1109/CVPR.2018.00920

Zhang, Y., Chu, J., Leng, L., & Miao, J. (2020). Mask-refined r-cnn: A network for refining object details in instance segmentation. *Sensors*. doi: 10.3390/s20041010

Zhang, Y., Liu, H., & Hu, Q. (2021). Transfuse: Fusing transformers and cnns for medical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. doi: 10.48550/arXiv.2102.08005

Zhao, Y., Z.Tao, & Li, L. (2022). Predicting biochemical-recurrence-free survival using a three-metabolic-gene risk score model in prostate cancer patients. *BMC Cancer*. doi: 10.1186/s12885-022-09331-8

Annex

Class Recurred/Progressed with 56 attributes

$$
\begin{aligned}
&-3.38 + [HSPA1A] * -1.52 + [HSPA1B] * -2.02 + [NFIX] \\
&* 1.32 + [C5orf10] * 1.85 + \\
&[DBNL] * 1.61 + [OLFML3] * -1.12 + \\
&[NIPA2] * 1.54 + [ITGBL1] * 1.45 + \\
&[NTM] * 1.44 + [LOC95070] * 1 + \\
&[CRLF1] * -1.15 + [TARSL2] * 1.77 + \\
&[NAA40] * 2.69
\end{aligned}
\tag{0.2}
$$

Class disease-free with 14 attributes:

$$4.27 + [HSPA1A] * 1.52 + [HSPA1B] * 2.02 + [NFIX] * -1.32 + [C5orf10] * -3.03 + [DBNL] * -2.42 + [OLFML3] * 2 + [NIPA2] * -1.54 + [ITGBL1] * -2.75 + [NTM] * -1.44 + [LOC95070] * -1 + [CRLF1] * 1.15 + [TARSL2] * -1.77 + [NAA40] * -3.69 \quad (0.3)$$

Recurred/Progressed with 14 attributes:

$$-4.27 + [HSPA1A] * -1.52 + [HSPA1B] * -2.02 + [NFIX] * 1.32 + [C5orf10] * 3.03 + [DBNL] * 2.42 + [OLFML3] * -2 + [NIPA2] * 1.54 + [ITGBL1] * 2.75 + [NTM] * 1.44 + [LOC95070] * 1 + [CRLF1] * -1.15 + [TARSL2] * 1.77 + [NAA40] * 3.69 \quad (0.4)$$

Class Disease Free with 119 attributes:

$$1.07 + [HSPA1A] * 5.94 + [PIGR] * -1.48 + [SNRNP\,200] * 0.92 + [TNRC18] * 0.76 + [EWWD] * 0.73 + [SLC31A1] * 0.62 + [SMARCD2] * 1.42 + [C17ORF\,62] * -1.12 + [DBNL] * -2.13 + [OLFML3] * 3.57 + [GSTZ1] * 1.91 + [GABRB3] * 0.87 + [FLJ\,12492] * 0.71 + [PPFIA2] * -1.73 + [EIF\,4G3] * -0.74 + [ABCC5] * 2.21 + [MED14] * 2.8 + [SLC37A3] * -1.06 + [TMEM\,101] * 0.79 + [SMG7] * -0.64 + [GPR158] * -0.97 + [NIT\,2] * 0.61 + [DNAJC14] * -0.84 + [HDAC6] * -1.53 + [FKBP\,15] * -2.12 + [CRY\,1] * 0.61 + [LOC100506644] * -1.11 + [PI4KAP\,1] * -2.24 + [TARSL2] * -1.86 + [USP\,42] * -0.89 + [CYTH4] * 1.97 + [FAIM] * -1.49 \quad (0.5)$$

Recurred/Progressed with 119 attributes:

$$
\begin{aligned}
&-1.07 + [HSPA1A] * -5.94 + [PIGR] * 1.48 + \\
&[SNRNP200] * -0.92 + [TNRC18] * -0.76 + \\
&[EWWD] * -0.73 + [SLC31A1] * -0.62 + \\
&[SMARCD2] * -1.42 + [C17ORF62] * 1.12 + \\
&[DBNL] * 2.13 + [OLFML3] * -3.57 + [GSTZ1] \\
&* -1.91 + [GABRB3] * -0.87 + [FLJ12492] * \\
&-0.71 + [PPFIA2] * 1.73 + [EIF4G3] * 0.74 + \\
&[ABCC5] * -2.21 + [MED14] * -2.8 + \\
&[SLC37A3] * 1.06 + [TMEM101] * -0.79 + \\
&[SMG7] * 0.64 + [GPR158] * 0.97 + [NIT2] * \\
&-0.61 + [DNAJC14] * 0.84 + \\
&[HDAC6] * 1.53 + [FKBP15] * 2.12 + [CRY \\
&1] * -0.61 + [LOC100506644] * 1.11 + \\
&[PI4KAP1] * 2.24 + [TARSL2] * 1.86 + \\
&[USP42] * 0.89 + [CYTH4] * -1.97 + \\
&[FAIM] * 1.49
\end{aligned}
\tag{0.6}
$$

Class Disease Free with 30 attributes:

$$
\begin{aligned}
&3.36 + [PIGR] * -2.46 + [FKBP4] * 0.63 + \\
&[LOC105374734] * 1.1 + [GABRB3] * 0.85 + \\
&[NIPA2] * -1.83 + [ITGBL1] \\
&* -1.43 + [GPR158] * -1.41 + [DNAJC14] \\
&* -0.55 + [NTM] * -3.87 + [LOC95070] * \\
&-2.29 + [EDC3] * 0.73 + [SCAF4] * -2.32 \\
&+ [RPPH1] * -0.73 + \\
&[LOC200551] * -0.83 + [TARSL2] * -3.6
\end{aligned}
\tag{0.7}
$$

Class Disease Free with 30 attributes:

$$
\begin{aligned}
&-3.36 + [PIGR] * 2.46 + [FKBP4] * -0.63 + \\
&[LOC105374734] * -1.1 + [GABRB3] * -0.85 + \\
&[NIPA2] * 1.83 + [ITGBL1] * 1.43 + \\
&[GPR158] * 1.41 + [DNAJC14] * 0.55 + \\
&[NTM] * 3.87 + [LOC95070] * 2.29 + \\
&[EDC3] * -0.73 + [SCAF4] * 2.32 + \\
&[RPPH1] * 0.73 + [LOC200551] * 0.83 + \\
&[TARSL2] * 3.6
\end{aligned}
\tag{0.8}
$$

Class DiseaseFree :

$$
\begin{aligned}
&-1.5 + [ZFHX3 = '(0.8 - inf)'] * 0.73 + [EMP2 = '(0.6 - 0.8]'] * 0.49 + \\
&[ITPR1 = '(0.2 - 0.4]'] * -0.66 + [NFIB = '(0.6 - 0.8]'] * 0.57 + [PCCA = ' \\
&(0.6 - 0.8]'] * 0.73 + [RGS2 = '(0.4 - inf)'] * 1.06 + [WDR5 = '(0.8 - inf)'] \\
&* 1.15 + [SRGAP2 = '(0.2 - 0.4]'] * 0.48 + \\
&[AIDA = '(0.4 - inf)'] * -0.57 + [SLC25A30 = '(-inf - 0.2]'] * -0.84
\end{aligned}
\tag{0.9}
$$

Class Progressed :

$$1.5 + [ZFHX3 = '(0.8 - inf)'] * -0.73 + [EMP2 = '(0.6 - 0.8]'] * -0.49 + [ITPR1 = '(0.2 - 0.4]'] * 0.66 + [NFIB = '(0.6 - 0.8]'] * -0.57 + [PCCA = '(0.6 - 0.8]'] * -0.73 + [RGS2 = '(0.4 - inf)'] * -1.06 + [WDR5 = '(0.8 - inf)'] * -1.15 + [SRGAP2 = '(0.2 - 0.4]'] * -0.48 + [AIDA = '(0.4 - inf)'] * 0.57 + [SLC25A30 = '(-inf - 0.2]'] * 0.84 \quad (0.10)$$

Class (0-22] :

$$
\begin{aligned}
&— 0.16 + [BMP6 = (0 - 0.2]] * -0.35 + [BMP6 = (0.4 - 0.6]] * -0.28 + [BMP6 \\
&= (0.8 - 1])] * 2.03 + [EPHB3 = (0.6 - 0.8]] * 0.58 + \\
&[KDSR = (0.2 - 0.4]] * 0.32 + [GM2A = (0.2 - 0.4]] * -0.52 + [IFI16 = \\
&(0.2 - 0.4]] * -1.64 + [PBX3 = (0.6 - 0.8]] * -0.35 + [PSME2 = (0.4 - 1])] \\
&* 0.87 + [PROM1 = (0.6 - 0.8]] * -0.53 + [MTMR6 = (0.2 - 0.4]] * 0.5 + \\
&[FAM50A = (0 - 0.2]] * 0.76 + [FAM50A = (0.4 - 0.6]] * -0.46 + [FAM \\
&50A = (0.6 - 0.8]] * 0.48 + [FAM50A = (0.8 - 1])] * -0.53 + [SLC4A7 = \\
&(0.6 - 0.8]] * -0.68 + [SERTAD2 = (0.6 - 0.8]] * -0.82 + [SERTAD2 = \\
&(0.8 - 1])] * 2.6 + [SRGAP3 = (0 - 0.2]] * -0.24 + [SRGAP3 = (0.4 - 0.6]] \\
&* 0.42 + [SRGAP3 = (0.8 - 1])] * 0.57 + [FERMT2 = (0.74 - 1])] * -1.23 + \\
&[TMSB15A = (0 - 0.2]] * 0.23 + [TMSB15A = (0.4 - 0.6]] * -0.68 + \\
&[PIM2 = (0.2 - 0.4]] * -0.27 + [ZHX2 = (0.2 - 0.4]] * -0.96 + \\
&[ZHX2 = (0.4 - 0.6]] * 0.76 + [ZHX2 = (0.8 - 1])] * 0.82 + [SEPHS2 \\
&= (0.2 - 0.4]] * 0.56 + [ZC3H4 = (0.4 - 0.6]] * -0.34 + [ZC3H4 = (0.6 - \\
&0.8]] * 0.49 + [LMOD1 = (0.8 - 1])] * -0.56 + [DCAF13 = (0 - 0.2]] * \\
&0.52 + [DCAF13 = (0.2 - 0.4]] * -0.56 + \\
&[GPR160 = (0.6 - 0.8]] * -0.43 + [GPR160 = (0.8 - 1])] * 1.14 +
\end{aligned}
\tag{0.11}
$$

$$[VPS4A = (0 - 0.2]] * -0.52 + [VPS4A = (0.6 - 0.8]] * -0.56 + [PYCR2 = (0 - 0.2]] * -1.85 + [C8orf55 = (0.2 - 0.4]] * -0.4 + [FAM49B = (0 - 0.2]] * 0.52 + [FAM49B = (0.2 - 0.4]] * -0.26 + [TPPP\ 3 = (0.2 - 0.4]] * 0.64 + [MST4 = (0 - 0.2]] * 0.41 + [MST4 = (0.4 - 1])] * -0.67 + [UHRF1BP1 = (0.6 - 0.8]] * 1.19 + [WSB2 = (0 - 0.2]] * -1.21 + [C16orf62 = (0.2 - 0.4]] * 0.83 + [C16orf62 = (0.4 - 0.6]] * -1.12 + [C3orf14 = (0 - 0.2]] * -0.29 + [C3orf14 = (0.2 - 0.4]] * 0.46 + [TXNDC15 = (0 - 0.2]] * 0.76 + [TXNDC15 = (0.2 - 0.4]] * -0.27 + [CRISPLD2 = (0 - 0.18]] * 0.73 + [CRISPLD2 = (0.54 - 0.72]] * 0.48 + [C9orf69 = (0 - 0.22]] * -0.79 + [C9orf69 = (0.61 - 1])] * 0.29 + [SLC25A25 = (0.22 - 0.42]] * 0.28 + [SLC25A25 = (0.61 - 0.81]] * 1.88 + [SLC25A25 = (0.81 - 1])] * -0.67 + [AGAP4 = (0.6 - 1])] * 0.71 + [SLC38A10 = (0.6 - 0.8]] * 1.11 + [SLC38A10 = (0.8 - 1])] * -0.81 + [CCDC12 = (0.42 - 0.61]] * -0.39 + [CCDC12 = (0.61 - 1])] * 1.19 + [ANKS6 = (0.6 - 1])] * 0.24 + [STAC3 = (0 - 0.2]] * -0.3 + [SLC25A30 = (0 - 0.18]] * 0.48 + [SLC25A30 = (0.18 - 0.35]] * -0.63 + [IAH1 = (0.2 - 0.4]] * 0.24 \quad (0.12)$$

Class >22 :

$$0.16 + [BMP6 = (0 - 0.2]] * 0.35 + [BMP6 = (0.4 - 0.6]] * 0.28 + [BMP6 = (0.8 - 1])] * -2.03 + [EPHB3 = (0.6 - 0.8]] * -0.58 + [KDSR = (0.2 - 0.4]] * -0.32 + [GM2A = (0.2 - 0.4]] * 0.52 + [IFI16 = (0.2 - 0.4]] * 1.64 + [PBX3 = (0.6 - 0.8]] * 0.35 + [PSME2 = (0.4 - 1])] * -0.87 + [PROM1 = (0.6 - 0.8]] * 0.53 + [MTMR6 = (0.2 - 0.4]] * -0.5 + [FAM50A = (0 - 0.2]] * -0.76 + [FAM50A = (0.4 - 0.6]] * 0.46 + [FAM50A = (0.6 - 0.8]] * -0.48 + [FAM50A = (0.8 - 1])] * 0.53 + [SLC4A7 = (0.6 - 0.8]] * 0.68 + [SERTAD2 = (0.6 - 0.8]] * 0.82 + [SERTAD2 = (0.8 - 1])] * -2.6 + [SRGAP3 = (0 - 0.2]] * 0.24 + [SRGAP3 = (0.4 - 0.6]] * -0.42 + [SRGAP3 = (0.8 - 1])] * -0.57 + [FERMT2 = (0.74 - 1])] * 1.23 + [TMSB15A = (0 - 0.2]] * -0.23 + [TMSB15A = (0.4 - 0.6]] * 0.68 + [PIM2 = (0.2 - 0.4]] * 0.27 + [ZHX2 = (0.2 - 0.4]] * 0.96 + [ZHX2 = (0.4 - 0.6]] * -0.76 + [ZHX2 = (0.8 - 1])] * -0.82 + [SEPHS2 = (0.2 - 0.4]] * -0.56 + [ZC3H4 = (0.4 - 0.6]] * 0.34 + \quad (0.13)$$

$$[ZC3H4 = (0.6 - 0.8]] * -0.49 + [LMOD1 = (0.8 - 1])] * 0.56 + [DCAF 13 = (0 - 0.2]] * -0.52 + [DCAF 13 = (0.2 - 0.4]] * 0.56 + [GPR160 = (0.6 - 0.8]] * 0.43 + [GPR160 = (0.8 - 1])] * -1.14 + [VPS4A = (0 - 0.2]] * 0.52 + [VPS4A = (0.6 - 0.8]] * 0.56 + [PYCR2 = (0 - 0.2]] * 1.85 + [C8orf55 = (0.2 - 0.4]] * 0.4 + [FAM 49B = (0 - 0.2]] * -0.52 + [FAM49B = (0.2 - 0.4]] * 0.26 + [TPPP3 = (0.2 - 0.4]] * -0.64 + [MST4 = (0 - 0.2]] * -0.41 + [MST4 = (0.4 - 1])] * 0.67 + [UHRF1BP1 = (0.6 - 0.8]] * -1.19 + [WSB2 = (0 - 0.2]] * 1.21 + [C16orf62 = (0.2 - 0.4]] * -0.83 + [C16orf62 = (0.4 - 0.6]] * 1.12 + [C3orf14 = (0 - 0.2]] * 0.29 + [C3orf14 = (0.2 - 0.4]] * -0.46 + [TXNDC15 = (0 - 0.2]] * -0.76 + [TXNDC15 = (0.2 - 0.4]] * 0.27 + [CRISPLD2 = (0 - 0.18]] * -0.73 + [CRISPLD2 = (0.54 - 0.72]] * -0.48 + [C9orf69 = (0 - 0.22]] * 0.79 + [C9orf69 = (0.61 - 1])] * -0.29 + [SLC25A25 = (0.22 - 0.42]] * -0.28 + [SLC25A25 = (0.61 - 0.81]] * -1.88 + [SLC25A25 = (0.81 - 1])] * 0.67 + [AGAP4 = (0.6 - 1])] * -0.71 + [SLC38A10 = (0.6 - 0.8]] * -1.11 + [SLC38A10 = (0.8 - 1])] * 0.81 + [CCDC12 = (0.42 - 0.61]] * 0.39 + [CCDC12 = (0.61 - 1])] * -1.19 + [ANKS6 = (0.6 - 1])] * -0.24 + [STAC3 = (0 - 0.2]] * 0.3 + [SLC25A30 = (0 - 0.18]] * -0.48 + [SLC25A30 = (0.18 - 0.35]] * 0.63 + [IAH1 = (0.2 - 0.4]] * -0.24 \quad (0.14)$$

Class Recurrence :

$$0.92 + [PSA = (0.04 - 0.11]] * 0.28 + [PSA = (0.11 - 1)] * 0.91 + [EMP2 = (0.6 - 0.8]] * -0.73 + [ITPR1 = (0.2 - 0.4]] * 1.22 + [PCCA = (0.6 - 0.8]] * -0.55 + [RGS2 = (0.4 - 1)] * -0.67 + [SRGAP2 = (0.2 - 0.4]] * -0.78 + [AIDA = (0.4 - 1)] * 0.58 + [SLC25A30 = (0 - 0.2)] * 1.22 + [ZFHX3 = (0.8 - 1)] * -0.51 + [WDR5 = (0.8 - 1)] * -1.22 + [NFIB = (0.6 - 0.8]] * -1.38 \quad (0.15)$$

Class (0 - 24) :

$$
\begin{aligned}
&— 0.47 + [PSA = (0 - 0.1]] * -0.26 + [PSA = \\
&(0.1 - 0.38]] * 0.15 + \\
&[EMP2 = (0.6 - 0.8]] * -0.66 + \\
&[ITPR1 = (0.2 - 0.4]] * 1.37 + \\
&[PCCA = (0.6 - 0.8]] * 0.2 + \\
&[RGS2 = (0.4 - 1)] * 0.19 + \\
&[SRGAP2 = (0.2 - 0.4]] * 1.11 + \\
&[AIDA = (0.4 - 1)] * 0.49 + \\
&[SLC25A30 = (0 - 0.2)] * 0.58 + \\
&[ZFHX3 = (0.8 - 1)] * 0.59 + \\
&[WDR5 = (0.8 - 1)] * 2.87
\end{aligned}
\tag{0.16}
$$

Class (24-inf) :

$$0.47 + [PSA = (0 - 0.1]] * 0.26 + [PSA = (0.1 - 0.38]] * -0.15 + [EMP2 = (0.6 - 0.8]] * 0.66 + [ITPR1 = (0.2 - 0.4]] * 1.37 + [PCCA = (0.6 - 0.8]] * -0.2 + [RGS2 = (0.4 - 1)] * -0.19 + [SRGAP2 = (0.2 - 0.4]] * -1.11 + [AIDA = (0.4 - 1)] * -0.49 + [SLC25A30 = (0 - 0.2)]] * -0.58 + [ZFHX3 = (0.8 - 1)] * -0.59 + [WDR5 = (0.8 - 1)] * -2.87 \quad (0.17)$$