



**DEVELOPMENT OF MACHINE LEARNING ALGORITHMS FOR THE  
AUTOMATED CHARACTERIZATION OF PROSTATE CANCER BASED ON  
HISTOPATHOLOGICAL IMAGES**

**Georgios Vlachos**

Biomedical Engineering, Aristotle University of Thessaloniki

A thesis submitted in fulfillment of the requirements for the  
Master of Science Degree

Thessaloniki, 2024



# **DEVELOPMENT OF MACHINE LEARNING ALGORITHMS FOR THE AUTOMATED CHARACTERIZATION OF PROSTATE CANCER BASED ON HISTOPATHOLOGICAL IMAGES**

**Georgios Vlachos**

Biomedical Engineering, Aristotle University of Thessaloniki

A thesis submitted in fulfillment of the requirements for the  
Master of Science Degree

SUPERVISOR: Professor Ioanna Chouvarda

THESIS COMMITTEE MEMBERS:

Professor Ioanna Chouvarda

Pofessor Anastasios Delopoulos

Dr Dimitrios Filos

Thessaloniki, 2024



**ΑΝΑΠΤΥΞΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΟΝ  
ΑΥΤΟΜΑΤΟΠΟΙΗΜΕΝΟ ΧΑΡΑΚΤΗΡΙΣΜΟ ΤΟΥ ΚΑΡΚΙΝΟΥ ΤΟΥ  
ΠΡΟΣΤΑΤΗ ΜΕ ΒΑΣΗ ΙΣΤΟΠΑΘΟΛΟΓΙΚΕΣ ΕΙΚΟΝΕΣ**

**Γεώργιος Βλάχος**

Βιοϊατρική Μηχανική, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Διπλωματική εργασία που υποβάλλεται για την απόκτηση μεταπτυχιακού  
διπλώματος

Θεσσαλονίκη, 2024

# Abstract

The pivotal role of early and precise cancer diagnosis cannot be overstated, especially given the complexity and variability inherent in oncological diseases. The traditional method of cancer diagnosis and characterization, histopathology, is based on the visual examination of human tissue under the microscope by specialized pathologists. This examination, while invaluable, can be time-consuming, subjective, and prone to diagnostic errors. In response to these challenges, this thesis explores the integration of computational strategies, specifically machine learning (ML) algorithms, to potentially augment and automate the diagnostic process. By providing a comprehensive review of existing methodologies in this area, the study highlights the progress and challenges faced in the automated analysis of cancer tissue images. It also presents a detailed summary of the available datasets of prostate cancer tissue images, which serve as a foundation for the development and validation of computational models. Central to this work is the exploration of deep learning, particularly convolutional neural networks (CNNs), and their fundamental principles. Leveraging state-of-the-art pre-trained architecture, this thesis introduces a customized CNN model designed to enhance the precision of prostate cancer characterization. The model employs a patch-based approach, wherein entire tissue images are dissected into smaller, manageable segments, allowing for a more detailed and comprehensive analysis. Utilizing the largest dataset available, the study demonstrates the model's capability to efficiently process and analyze a significant amount of histopathological data. The findings are meticulously analyzed, offering insightful conclusions, and proposing directions for future research to further refine and improve automated cancer characterization techniques. Through its analytical and methodological contributions, this thesis represents a significant step forward in the utilization of deep learning for the effective diagnosis and understanding of prostate cancer from histopathological images.

**Key-words:** Prostate cancer, histopathological image, Machine Learning, Deep Learning, Convolutional Neural Network, diagnosis, automation

# Περίληψη

Η έγκαιρη και ακριβής διάγνωση του καρκίνου του προστάτη αποτελεί ζωτικής σημασίας παράγοντα για την επιτυχή πρόγνωση και θεραπεία της νόσου. Η παραδοσιακή μέθοδος διάγνωσης, η ιστοπαθολογική εξέταση, βασίζεται στην οπτική αξιολόγηση των βιοψιών από εξειδικευμένους παθολόγους. Η διαδικασία αυτή, όμως, μπορεί να είναι χρονοβόρα, υποκειμενική και επιρρεπής σε διαγνωστικά σφάλματα. Η παρούσα διπλωματική εργασία εστιάζει στην ανάπτυξη αλγορίθμων μηχανικής μάθησης, και ειδικά βαθιάς μάθησης, για την αυτοματοποιημένη διάγνωση του καρκίνου του προστάτη με βάση ιστοπαθολογικές εικόνες. Σκοπός της εργασίας είναι η δημιουργία ενός αξιόπιστου και αντικειμενικού συστήματος υποβοήθησης της διάγνωσης, το οποίο θα δύναται να λειτουργήσει συμπληρωματικά προς την κρίση του παθολόγου. Αρχικά, η εργασία πραγματεύεται τη σημασία της έγκαιρης και ακριβούς διάγνωσης του καρκίνου του προστάτη, καθώς και τις δυσκολίες που σχετίζονται με την παραδοσιακή μέθοδο της ιστοπαθολογίας. Στη συνέχεια, παρουσιάζονται οι βασικές αρχές της μηχανικής μάθησης και της βαθιάς μάθησης, με έμφαση στα Συνελικτικά Νευρωνικά Δίκτυα (ΣΝΔ), τα οποία αποτελούν ιδανικά εργαλεία για την ανάλυση ιατρικών εικόνων. Ακολουθεί μια ανασκόπηση της βιβλιογραφίας σχετικά με την εφαρμογή της βαθιάς μάθησης στην αυτόματη διάγνωση του καρκίνου του προστάτη. Γίνεται αναφορά στις διαθέσιμες βάσεις δεδομένων ιστοπαθολογικών εικόνων, καθώς και στις υφιστάμενες μεθοδολογίες και τα επιτευχθέντα αποτελέσματα. Στο κύριο μέρος της εργασίας, παρουσιάζεται η μεθοδολογία που υιοθετήθηκε για την ανάπτυξη του ΣΝΔ. Περιγράφονται τα στάδια προεπεξεργασίας των εικόνων, η αρχιτεκτονική του δικτύου, η διαδικασία εκπαίδευσης και οι παράμετροι που βελτιστοποιήθηκαν. Ακολουθώντας, γίνεται ανάλυση των πειραματικών αποτελεσμάτων. Αξιολογείται η απόδοση του ΣΝΔ και συγκρίνονται τα αποτελέσματα με αντίστοιχες προσεγγίσεις. Στο τελευταίο κεφάλαιο, συζητούνται τα πλεονεκτήματα και οι περιορισμοί της προτεινόμενης μεθοδολογίας. Επισημαίνονται οι δυνατότητες βελτίωσης και οι μελλοντικές προοπτικές της έρευνας.

**Λέξεις-κλειδιά:** Καρκίνος προστάτη, ιστοπαθολογική εικόνα, μηχανική μάθηση, βαθιά μάθηση, Συνελικτικά Νευρωνικά Δίκτυα, διάγνωση, αυτοματοποίηση

# Contents

<b>Abstract .....</b>	<b>I</b>
<b>Περίληψη .....</b>	<b>II</b>
<b>Contents .....</b>	<b>III</b>
<b>1. INTRODUCTION.....</b>	<b>1</b>
<b>1.1 Background information: Prostate cancer &amp; histopathology .....</b>	<b>1</b>
<b>1.2 Problem statement .....</b>	<b>1</b>
<b>1.3 Limitations.....</b>	<b>2</b>
<b>1.4 Objectives and significance of the research .....</b>	<b>3</b>
<b>2. LITERATURE REVIEW .....</b>	<b>4</b>
<b>2.1 Deep Learning overview.....</b>	<b>4</b>
2.1.1 Categorization of Deep Learning methods .....	4
2.1.2 Deep Learning networks .....	5
2.1.3 Convolutional Neural Networks: Core components .....	6
<b>2.2 Histological image analysis with Deep Learning: An overview ..</b>	<b>10</b>
2.2.1 Image pre-processing & post-processing .....	10
2.2.2 Summary of Deep Learning methods in histological image analysis	13
2.2.3 Significant approaches: A synopsis .....	17
<b>2.3 Overview of available data sources .....</b>	<b>21</b>
<b>3. METHODOLOGY.....</b>	<b>24</b>
<b>3.1 Data acquisition .....</b>	<b>24</b>
3.1.1 Dataset overview.....	24
3.1.2 Labels - ISUP grading .....	25
3.1.3 Dataset characteristics .....	25
<b>3.2 Pre-processing.....</b>	<b>28</b>
<b>3.3 Model development .....</b>	<b>31</b>
3.3.1 CNN architecture.....	31
3.3.2 Implementation.....	33
<b>3.4 Training.....</b>	<b>36</b>
3.4.1 Data preparation .....	36
3.4.2 Loss function .....	36
3.4.3 Optimizer.....	37
3.4.4 Hyperparameters .....	38

3.5	Evaluation metrics and validation methods .....	39
3.5.1	Multiclass ROC AUC (Area Under the Receiver Operating Characteristic Curve).....	39
3.5.2	Quadratic Weighted Kappa (Cohen's Kappa).....	39
4.	RESULTS.....	41
4.1	Confusion matrix .....	41
4.2	Multiclass ROC AUC .....	42
4.3	Quadratic Weighted Kappa .....	42
5.	DISCUSSION.....	43
5.1	Confusion matrix interpretation .....	43
5.2	Performance factors .....	43
5.3	Noticeable attempts in model development and training.....	44
5.4	Approaches comparison.....	45
5.5	Conclusions .....	49
5.6	Future work or improvements .....	49
	References .....	50
	Appendices .....	55
	Appendix A1 .....	55
	Appendix A2.....	56
	Appendix A3.....	58

# **1. INTRODUCTION**

## **1.1 Background information: Prostate cancer & histopathology**

Prostate cancer is a significant health concern worldwide, being the second most common cancer among men and a leading cause of cancer-related deaths (Siegel, Miller, & Jemal, 2020). Early and accurate diagnosis is crucial for effective treatment and management, significantly impacting patient survival rates and quality of life. Traditionally, the diagnosis of prostate cancer involves a combination of clinical assessments, including prostate-specific antigen tests, digital rectal exams, and the histopathological examination of prostate tissue obtained through biopsy.

Histopathology, the study of tissues affected by disease under the microscope, is considered the gold standard for prostate cancer diagnosis. Pathologists examine tissue samples to identify cancerous cells, grading the cancer based on its appearance compared to normal prostate tissue. The most widely used grading system for prostate cancer is the Gleason score, which assesses the architectural patterns of cancer cells in the tissue. However, the process is time-consuming, subjective, and depends on the expertise of the pathologist, leading to potential variability in diagnosis outcomes (Carriaga & Henson, 1995; J. I. Epstein et al., 2016).

The importance of accurate diagnosis cannot be overstated, as it informs treatment decisions and prognostic evaluations. Misclassification or grading inaccuracies can lead to over- or under-treatment, affecting patient outcomes and quality of life. As such, enhancing the precision and efficiency of prostate cancer diagnosis, particularly through the use of automated and objective methods, is a critical area of research and development.

In recent years, the field of histopathology has been revolutionized by the advent of digital pathology, which involves scanning traditional glass slides to create high-resolution digital images. This innovation has facilitated a more detailed and accurate analysis of tissue samples. Moreover, it has opened the door to the application of advanced computational techniques, including machine learning, to assist pathologists in diagnosing and grading prostate cancer (Musumeci, 2014).

## **1.2 Problem statement**

Despite advances in medical imaging and diagnostic techniques, the histopathological examination of prostate tissue remains a challenge due to its inherent subjectivity and the complexity of cancer tissue structures. Variability in diagnostic interpretations among pathologists can lead to inconsistent grading, potentially affecting treatment decisions and patient outcomes. Furthermore, the increasing volume of prostate biopsy samples necessitates a considerable amount of time and resources for pathological assessment, highlighting the need for more efficient diagnostic processes (Van der Laak, Litjens, & Ciompi, 2021).



The integration of machine learning, particularly convolutional neural networks (CNNs), into the analysis of histopathology images presents a promising solution to the challenges the pathologists face. Machine learning algorithms can learn from vast amounts of data, identifying patterns that may not be immediately apparent to human observers. CNNs, a class of deep neural networks, are especially suited for image recognition tasks and have shown great potential in medical image analysis, including the classification and grading of cancer from histopathology images (Van der Laak et al., 2021).

This research addresses the specific problem of automating the characterization of prostate cancer in histopathological images. By leveraging Convolutional Neural Networks (CNNs), this study aims to develop an effective machine learning model capable of classifying prostate cancer from histopathological images with satisfactory accuracy and high efficiency, in terms of computational performance. The utilization of CNNs for image recognition and classification presents a promising solution to overcome the limitations of manual histopathological examination by providing a standardized, objective, and scalable method for diagnosing prostate cancer. The research focuses on the PANDA Challenge dataset from Kaggle (Bulten, Kartasalo, Chen, Ström, Pinckaers, Nagpal, Cai, Steiner, van Boven, & Vink, 2022; Bulten, Pinckaers, & Eklund, 2020), which offers a comprehensive collection of annotated and labelled histopathological images of prostate tissue, to train and validate the proposed CNN model. Through this approach, the study seeks to contribute to the improvement of diagnostic accuracy and potentially streamline the diagnostic workflow in clinical settings, by proposing an effective algorithm that utilizes the information efficiently, respecting the computational resources, for automated characterization of prostate histological images.

### **1.3 Limitations**

Technical limitations are inherent in any study involving advanced computational models. One of the key challenges is the selection and optimization of machine learning algorithms. While deep learning models, particularly CNNs, are well-suited for image analysis, determining the optimal architecture and parameters for the specific task of prostate cancer diagnosis and grading requires extensive experimentation and fine-tuning (Goodfellow, Bengio, & Courville, 2016).

A key factor that affects the ML model's performance is the quality of histopathology images. High-quality images are essential for accurate analysis, yet variations in scanning equipment, staining techniques, and image processing can impact image quality. Addressing these variations to ensure consistency across the dataset is a significant challenge. Additionally, the issue of pathologist non-agreement in histological images grading further complicates the training of CNNs. Inter-observer variability in the cancer characterization, due to subjective interpretations of histological patterns, can introduce inconsistencies in the training data, leading to decreased reliability of the CNN models. This inter-observer variability reflects the complexity of prostate cancer diagnosis (Ismail et al., 1989).

High-resolution images play a pivotal role in the analysis of histological images for prostate cancer diagnosis, serving as a double-edged sword in digital pathology. On one hand, high resolution is crucial for capturing detailed information at the cellular

level, where critical diagnostic features, such as the architectural and morphological characteristics indicative of cancer, are present. This level of detail is essential for accurate disease identification and grading, as it allows pathologists and systems to discern subtle features that may signify the presence of cancer. On the other hand, the very nature of these high-resolution images poses significant challenges for computational processing. Due to their large size, high-resolution histological images demand extensive computational resources for storage, processing, and analysis. Handling these images requires great CPU and GPU processing power, as well as considerable memory capacity, to efficiently process and analyze the data without compromising the speed or accuracy of the analysis. The limitation on these computational resources poses a big obstacle. Moreover, most deep learning models, including CNNs, are designed to process images of relatively lower resolutions, and scaling them to accommodate high-resolution inputs often results in increased complexity and computational costs. This scaling can exacerbate the already intensive requirements for training and inference, making it challenging to develop models that are both accurate and computationally efficient (Goodfellow et al., 2016).

Additionally, a primary challenge is the limited diversity and number of available images, which can restrict the model's ability to learn the nuanced variations of prostate cancer appearances. This limitation often results in overfitting, where the CNN performs well on the training data but poorly on unseen data, reducing its generalizability and clinical applicability (Goodfellow et al., 2016).

Also, the reliance on a single dataset for training and validation may impact the model's generalizability to other datasets or real-world clinical settings. Future work will need to address these limitations by incorporating more diverse datasets, enhancing the current training.

## **1.4 Objectives and significance of the research**

The primary objective of this research is to develop an advanced CNN tailored for efficient grading of prostate cancer from high-resolution histopathological images. Recognizing that these images often contain vast areas of non-informative content alongside critical diagnostic information confined to small, localized regions, the workflow, code structure, and model design are crafted to prioritize computational efficiency. The approach involves appropriate preprocessing to retain the most information-rich parts of the images while discarding extraneous content, ensuring the model focuses on cellular-level details critical for accurate grading. This methodological novelty addresses the dual challenge of managing the high computational demands of processing large-scale images and the necessity of pinpointing diagnostically relevant features within them. Ultimately, the research aims to deliver a solution that not only enhances the accuracy of prostate cancer grading but also respects the constraints of computational resources, setting a benchmark for efficient analysis of histological images.

## 2. LITERATURE REVIEW

### 2.1 Deep Learning overview

#### 2.1.1 Categorization of Deep Learning methods

The categorization of Deep Learning (DL) (Alzubaidi et al., 2021; Goodfellow et al., 2016) methodologies can be primarily segmented into four distinct approaches: supervised, semi-supervised (or partially supervised), unsupervised, and deep reinforcement learning (DRL). Each classification reflects the nature of the data interaction and the learning paradigm employed.

#### Deep Supervised Learning

Deep Supervised Learning involves the utilization of labeled datasets, wherein the model is trained on a predefined set of inputs and their corresponding outputs  $(x_t, y_t) \sim \rho$ . This approach enables the learning algorithm to develop predictive capabilities by minimizing the discrepancy between the actual and predicted outputs through iterative optimization of the network parameters. The application of deep supervised learning spans across various architectures, including Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Deep Neural Networks (DNNs), with RNNs further encompassing Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks. The primary advantage of deep supervised learning lies in its efficacy in deriving insights and predictions from historical data, thereby facilitating accurate model training. However, a potential limitation emerges when the training dataset lacks representative samples from each class, leading to an overfitted model with reduced generalizability. Despite this, deep supervised learning remains a highly effective and straightforward approach for tasks where performance is paramount and labeled data is abundant.

#### Deep semi-supervised learning

Deep semi-supervised learning is a nuanced methodology that leverages datasets composed of both labeled and unlabeled data. This approach is particularly beneficial in scenarios where labeled data are scarce or expensive to obtain. It often incorporates generative adversarial networks (GANs) and deep reinforcement learning (DRL) techniques to enhance the learning process. Additionally, recurrent neural networks (RNNs), including Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTMs) networks, are utilized to facilitate partial supervision of the learning algorithm. A significant advantage of deep semi-supervised learning lies in its capacity to reduce the reliance on extensive labeled datasets, thereby lowering the overall resource requirements for model training. However, a potential limitation of this methodology is the risk of incorporating irrelevant input features into the training data, which can lead to inaccurate decision-making by the model.

## **Deep Unsupervised Learning**

Deep unsupervised learning facilitates the learning process without the need for labeled data. This method enables the algorithm to autonomously identify significant patterns, features, or underlying structures within the input data, thereby uncovering previously unrecognized relationships or classifications. It encompasses a variety of techniques, including but not limited to generative models, dimensionality reduction, and clustering, each playing a pivotal role in the exploration of data properties. Despite its advantages, deep unsupervised learning is not without its challenges. One of the main drawbacks is its potential inability to provide precise categorization or sorting of data, which can be a critical limitation for certain applications. Additionally, these algorithms can be computationally intensive, requiring significant resources for effective implementation. Among the various unsupervised learning methodologies, clustering stands out as a particularly prominent approach, offering a powerful means to group data based on similarity or other relevant criteria, thereby enabling meaningful insights to be drawn from unlabelled datasets.

## **Deep Reinforcement Learning**

Deep Reinforcement Learning (DRL) distinguishes itself by enabling algorithms to learn optimal behaviors through direct interaction with their environment, unlike supervised learning which relies on pre-labeled datasets. Originating from advancements made by Google DeepMind in 2013 (Alzubaidi et al., 2021), DRL has evolved significantly, giving rise to various sophisticated models. Unlike traditional supervised approaches, DRL navigates environments without a predefined loss function, relying instead on a trial-and-error process guided by a reward system. This method contrasts with supervised learning by not having explicit access to an optimization function and by the necessity of learning from sequential interactions influenced by past actions. Key motivations for employing reinforcement learning include its ability to discern actions yielding maximal long-term rewards, recognize situations necessitating action, devise strategies for substantial reward attainment, and provide a reward-based learning framework. Despite its strengths, reinforcement learning is not universally applicable, particularly where ample supervised data exists or where computational resources and time are limiting factors.

### **2.1.2 Deep Learning networks**

DL encompasses a variety of network architectures, each tailored to specific types of data and tasks. Among these, Recursive Neural Networks (RvNNs), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs) stand out due to their unique capabilities and applications (Alzubaidi et al., 2021; Goodfellow et al., 2016).

#### **Recursive Neural Networks (RvNNs)**

RvNNs are designed to process structured data, notably in hierarchical forms such as trees and graphs. They excel in applications where data exhibit a nested or recursive structure, making them particularly useful in Natural Language Processing (NLP) and syntactic analysis of sentences. RvNNs operate by creating compositional vector representations for parts of the structure (e.g., sentences, images divided into

segments) and then merging these vectors based on their hierarchical relationships. This process effectively captures the underlying structure of the data, allowing for sophisticated predictions and classifications. The ability of RvNNs to generate a fixed-width distributed representation from variable-sized recursive data structures is key to their utility in processing complex inputs.

## **Recurrent Neural Networks (RNNs)**

RNNs are distinguished by their ability to handle sequential data, making them ideal for tasks such as speech recognition, language modeling, and time-series analysis. By maintaining a form of short-term memory, RNNs can consider the context of previous inputs in making predictions about current or future inputs. This sequential processing capability allows RNNs to capture dynamic temporal behaviors within data. However, RNNs face challenges with gradient vanishing and exploding problems, which can hinder their ability to learn long-range dependencies. LSTM units, a variant of RNNs, address this issue by incorporating mechanisms that allow the network to remember or forget information over long durations, thus enhancing the model's ability to capture long-term dependencies within data.

## **Convolutional Neural Networks (CNNs)**

CNNs have revolutionized the field of computer vision by enabling highly efficient image processing and feature extraction without the need for manual feature engineering. Inspired by the visual cortex of animals, CNNs utilize a hierarchical structure of convolutional and pooling layers to process spatial data. These layers automatically detect and learn the most relevant features from the input data through the application of filters and down-sampling techniques. The architecture's use of shared weights significantly reduces the number of parameters, making CNNs more efficient and less prone to overfitting compared to fully connected networks. CNNs are versatile and have been successfully applied to a wide range of applications beyond image recognition, including video analysis, speech recognition, and medical image analysis.

### **2.1.3 Convolutional Neural Networks: Core components**

The architecture of CNNs draws inspiration from the neural structures of human and animal brains, particularly how visual cortex cells process parts of a scene (Alzubaidi et al., 2021). This bio-inspired mechanism is reflected in CNN's ability to capture spatial hierarchies in data through a series of convolutional, pooling, and fully connected (FC) layers, mimicking the local receptive fields of the visual cortex.

Core components of CNNs:

1. **Convolutional Layers:** Serve as the foundation of CNNs, employing filters (kernels) to perform convolution operations across the input image, generating feature maps that highlight important features.
2. **Pooling Layers:** Follow convolutional layers to perform down-sampling, reducing the dimensionality of feature maps while preserving essential

information, thereby reducing computational complexity, and mitigating overfitting.

3. **Activation Functions:** Non-linear functions (e.g., ReLU, sigmoid, tanh) applied after convolution operations to introduce non-linearity into the network, enabling it to learn complex patterns.
4. **Fully Connected (FC) Layers:** Typically located towards the end of the network, these layers integrate learned features from previous layers to perform classification tasks. Each neuron in a fully connected layer is connected to all activations in the previous layer, synthesizing the features to predict the final output.
5. **Loss Functions:** Utilized to quantify the difference between the predicted outputs and actual labels, guiding the network's training process through backpropagation.

Some of the key concepts and basic processes of CNNs are the regularization, the optimization, and the algorithms design (backpropagation).

### Regularization in CNNs

Regularization techniques are essential to prevent overfitting in Convolutional Neural Networks (CNNs), ensuring models generalize well to unseen data. Significant regularization strategies include:

1. **Dropout:** Randomly omits neurons during training, distributing feature detection across the network and encouraging the learning of independent feature representations. During inference, the full network is used.
2. **Drop-Weights:** Similar to dropout but drops connections (weights) between neurons instead of neurons themselves, promoting model robustness.
3. **Data Augmentation:** Artificially expands the training dataset using various transformations (e.g., rotation, scaling) to improve model robustness and prevent overfitting.
4. **Batch Normalization:** Normalizes the output of a layer to a unit Gaussian distribution, reducing internal covariate shift, stabilizing training, and speeding up convergence. It also has a slight regularization effect, potentially reducing the need for dropout.

### Optimizer Selection

Optimizers adjust network weights to minimize the loss function. Key points include:

- 🌈 **Learning Algorithm Selection (Optimizer):** Essential for defining how the network updates its weights during training to minimize the loss. Common choices include Gradient Descent and its variants, each with unique characteristics suitable for different scenarios.
- 🌈 **Gradient Descent:** Updates network parameters using the gradient of the loss function. Variants include:

1. **Batch Gradient Descent:** Updates parameters once per epoch using the entire training set, suitable for small datasets but may be impractical for large ones.
2. **Stochastic Gradient Descent (SGD):** Updates parameters for each training example, offering faster convergence but with more noise in the updates.
3. **Mini-batch Gradient Descent:** Combines the benefits of both, updating parameters for small batches of training data, offering a balance between computational efficiency and convergence stability.

#### Enhancements:

1. **Momentum:** Accelerates SGD by incorporating the direction of previous gradients to smooth out updates, helping to avoid local minima and speed up convergence.
2. **Adaptive Moment Estimation (Adam):** Combines ideas from Momentum and RMSprop, adjusting the learning rate based on first and second moments of the gradients, making it effective for deep learning applications.

## Backpropagation

Backpropagation is the fundamental algorithm for training neural networks, involving two main phases:

1. **Forward Pass:** Where inputs are passed through the network to generate an output.
2. **Backward Pass (Backpropagation):** Where the gradient of the loss function is computed with respect to each weight by the chain rule, effectively "propagating" the error backward through the network to update the weights and minimize the loss function.

This process is iteratively repeated, adjusting the network's weights to reduce the prediction error, thereby training the network to accurately model the target function.

The semantic advantages of CNNs:

1. **Weight Sharing:** Reduces the number of trainable parameters, enhancing generalization and reducing the risk of overfitting.
2. **Efficient Feature Learning:** Automatically identifies and learns important features from the input data without requiring manual feature extraction, making CNNs highly effective for tasks involving image data.
3. **Adaptability:** CNNs can be applied to a wide range of tasks beyond image recognition, including speech recognition and natural language processing, thanks to their ability to process data in two-dimensional form and capture spatial hierarchies.

The structure and benefits of CNNs underscore their significance in artificial intelligence applications, offering a powerful tool for automating feature extraction and enabling the development of sophisticated models for a variety of complex tasks.



## 2.2 Histological image analysis with Deep Learning: An overview

### 2.2.1 Image pre-processing & post-processing

As said, histopathological images are rich of information, due to their high-resolution quality and complex structure, but at the same time challenging for ML algorithms (Gurcan et al., 2009). For this reason, image pre-processing and post-processing operations (Table 1) have a crucial role in the accuracy of the results by preparing and making them suitable for the ML algorithms and the interpretation of outcomes (Salvi, Acharya, Molinari, & Meiburger, 2021).

- I. **Pre-processing:** The analysis of high-resolution histopathological images compounded by the complexity of the background and disruptive factors can slow down the process. To overcome this, image pre-processing algorithms should be employed. Also, pre-processing tasks are important for ensuring that the images are consistent and of high quality in specific size and resolution, and that the relevant features are preserved. Significant techniques that have been used:
  - a. **Image enhancement** (Maini & Aggarwal, 2010): Process of improving the visual quality (noise removal, sharpen, brighten, etc.) in order to make easier the key features identification. OpenCV functions (like `addWeighted`) (Kaggle.com, 2021) can be utilized for gray areas removal from Whole Slide Images (WSIs) or other contrast enhancement filters. Notable methods:
    - **Unsharp masking** (Kaur, Jindal, & Singh, 2021): Improves the details in high-frequency areas with the use of Gaussian filter.
    - **Histogram equalization** (Hsu & Chou, 2015): Method of contrast adjustment
  - b. **Color/Stain normalization** (Macenko et al., 2009; Ruifrok & Johnston, 2001): Employed to reduce the effects of variations in staining, scanning or illumination (color variability) (Figure 1).
  - c. **Region of interest (ROI) detection/Artifact detection** (Irshad, Veillard, Roux, & Racoceanu, 2013): Identifying-selecting the region of interest by eliminating areas with low content and noise (Hamilton et al., 1997).
  - d. **Patch selection/Patch-level predictions** (Kothari, Phan, Stokes, & Wang, 2013; Salvi et al., 2021): Integrating CNNs into WSIs analysis poses certain challenges. Firstly, significant information may be lost during the necessary down-sampling process. Secondly, CNNs may only learn from one of the many discriminative patterns in the image, leading to a lack of data efficiency. Also, unnecessary GPU's memory and time consumption can appear. To overcome these limitations, the key is to train the network on high-resolution patches of the image,

rather than the entire WSI (Whole Slide Images), and then make predictions based on the patch-level predictions. Like this, the large empty space areas in high-resolution WSIs are discarded and the small areas of concern become more obvious (<https://www.kaggle.com/iafoss/panda-16x128x128-tiles>, <https://www.kaggle.com/code/iafoss/panda-concat-tile-pooling-starter-0-79-lb/notebook>). It is crucial to choose representative patches for training to optimize the results.

- e. **Tissue/Nucleus segmentation** (Naylor, Laé, Reyat, & Walter, 2017; Pang et al., 2010): The segmentation of cell nuclei is a crucial pre-processing step in the analysis of histopathological images, as many disease characteristics, particularly in the case of cancer, are reflected in the nuclei. In fact, a significant amount of cytological and histopathological analysis relies solely on the analysis of nuclear features. Despite its importance, nuclei segmentation is a complex task, given that different tissue types, staining variations, and cell types each have distinct visual properties, making it difficult to create conventional image segmentation algorithms that can effectively handle these differences.

II. **Post-processing:** In some cases, the use of post-processing methods has been shown to significantly improve accuracy and performance ((Källén, Molin, Heyden, Lundström, & Åström, 2016; Ren, Sadimin, Foran, & Qi, 2017; Salvi et al., 2021). The post-processing operations are utilized in tasks related to classification, detection, and segmentation. Different techniques have been used to enhance the results of histopathological image analysis:

- i. Active contour differentiation
- ii. Color illumination
- iii. Image compression
- iv. Image restoration
- v. Zooming
- vi. Augmentation

In order to address the semantic segmentation issue, the algorithm discussed in (Zeng, Xie, Zhang, & Lu, 2019) requires intricate post-processing techniques. Also, techniques such as morphological operations can be employed to remove artifacts (Ren et al., 2017).

Figure 1: Example of color/stain normalization – A visual representation.

[Source: (Salvi et al., 2021)]

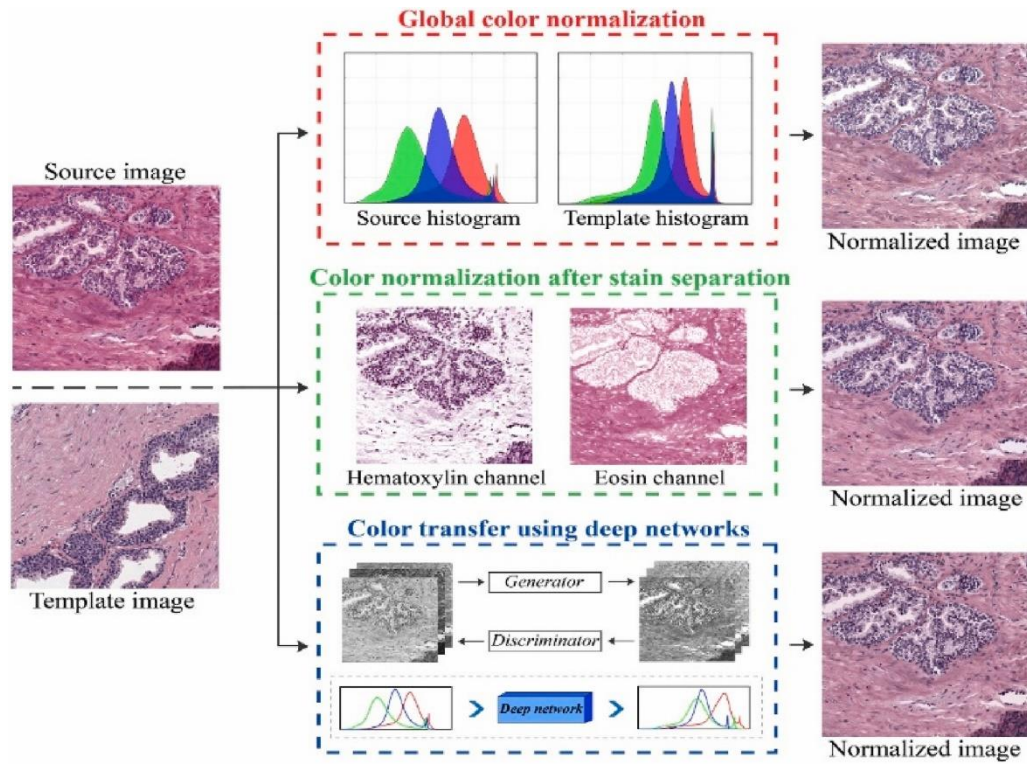



Table 1: Synopsis of histopathological image pre and post processing techniques.

<u>Pre-processing</u>	<u>Post-processing</u>
<ul style="list-style-type: none"> <li>• Image enhancing</li> <li>• Noise reduction</li> <li>• Sharpening</li> <li>• Tiling/Patch extraction</li> <li>• Normalization (Feature scaling)</li> <li>• Resolution reduction</li> </ul>	<ul style="list-style-type: none"> <li>• Active contour differentiation</li> <li>• Color illumination</li> <li>• Image compression</li> <li>• Image restoration</li> <li>• Zooming</li> <li>• Augmentation</li> </ul>

<u>Pre-processing</u>	<u>Post-processing</u>
<ul style="list-style-type: none"> <li>• <b>Stain/Color normalization</b></li> <li>• <b>ROI detection</b></li> <li>• <b>Morphological operation</b></li> <li>• <b>Tissue/Nucleus segmentation</b></li> </ul>	

## 2.2.2 Summary of Deep Learning methods in histological image analysis

Remarkable results have been recently reported using ML techniques, showing grading performance comparable to that of human pathologists (Bulten, Pinckaers, van Boven, et al., 2020; Nagpal et al., 2019; Ström et al., 2020). As previous mentioned, the focus of recent research has been on Deep Learning methods, with Convolutional Neural Networks (CNNs) being utilized as the primary tool for image processing and analysis (Litjens et al., 2016; Van der Laak et al., 2021). Their design helps to extract increasingly complex features and reduce the spatial dimensions of the input data, making it easier to identify patterns and relationships between pixels. By using multiple convolutional and fully connected layers, CNNs are able to learn extremely abstract representations of the input data, making them well-suited for tasks such as image classification, segmentation, and object detection (Szegedy et al., 2015). Deep learning techniques promise better outcomes compared to other traditional ML algorithms (Panigrahi & Swarnkar, 2020). Notable training methods that have been employed:

 **Transfer learning:** A ML approach (Pan & Yang, 2010; Talo, 2019) that leverages a pre-trained model for a new task, resulting in faster training and a need for less training data. It is especially useful when limited data is available for a new domain, as it allows for the transfer of knowledge from a large pre-existing data pool. This is particularly relevant in medical imaging due to small number of labeled data set in this field. Transfer learning can be divided into two strategies:

- Freezing the main network layers and using the pre-trained model as a feature extractor: In this approach, a pre-trained model is used as a feature extractor for the convolutional neural network. The last fully connected layer (the classifier layer) is removed, and the remaining layers are adapted for the new task. Instead of training the entire

network, only a new classifier is trained, significantly speeding up the process.

- Fine-tuning the entire network (Tajbakhsh et al., 2016): This strategy involves modifying the CNN's final layer and retraining all preceding layers via backpropagation, starting with low learning rates to utilize prior knowledge.





🌈 **End-to-end learning:** A DL technique (Julio Silva-Rodríguez, Colomer, Sales, Molina, & Naranjo, 2020) where the network is trained to perform a complex task from raw input to desired output, without explicit feature engineering or manual intervention at intermediate stages. The model learns all the steps between the input and output phases, allowing it to directly optimize the task-specific objective. All the parameters are trained simultaneously.




🌈 **Multi-task learning (MTL):** A ML technique (Ruder, 2017) that focuses on solving multiple learning tasks at the same time by utilizing similarities and differences between the tasks. This can lead to better learning efficiency and increased prediction accuracy for the individual tasks compared to training each model independently. According to Rich Caruana (Caruana, 1998), MTL improves generalization by utilizing the information present in the training data of related tasks. This is achieved by training all tasks in parallel while sharing a common representation. The training data from the additional tasks acts as an inductive bias, helping to improve learning. In MTL, a shared hidden layer is used to learn from all tasks at the same time, leading to improved performance in each task based on what is learned from the other tasks."

The design and complexity of CNNs are crucial. The state of the art of popular CNN architecture are:

1. **LeNet (Classical/First CNN architecture)** (LeCun, Bottou, Bengio, & Haffner, 1998)
2. **AlexNet** (Krizhevsky, Sutskever, & Hinton, 2017)
3. **VGG-16** (Simonyan & Zisserman, 2014)
4. **GoogleNet (Inception-v1)** (Szegedy et al., 2015)
5. **Inception-v3** (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016)
6. **ResNet** (He, Zhang, Ren, & Sun, 2016)
7. **UNet** (Ronneberger, Fischer, & Brox, 2015)
8. **MobileNet** (Howard et al., 2017)
9. **Graph convolutional networks (GCNs)** (Kipf & Welling, 2016)
10. **EfficientNet** (Hoang & Jo, 2021)
11. **DenseNet** (Zhu & Newsam, 2017)

A review of recent advances is presented:

-  **Region of interest-based models:** Accurate identification of small regions of interest in histopathological images can be a demanding piece of work. An approach that was intended to address this challenge was a ROI-based model that extracted feature maps for each class using ResNet (Li et al., 2018). The feature maps were then input into two heads: the Epithelial Proposal Head (EPH) and the Grading Network Head (GNH). The EPH determines if the image contains epithelial tissue, while the Region Proposal Network (RPN) identifies regions for the GNH to assign Gleason grades. If no epithelial cells are present, the whole image is considered stroma. In contrast, del Toro et al (del Toro et al., 2017) trained annotated datasets using various state-of-the-art architectures such as LeNet, AlexNet and GoogleNet. Each architecture was trained with the same patches obtained from the previous patch selection step at 40x resolution and underwent various image augmentations techniques focused on the region of interest. For higher Gleason grading cases, heatmaps and ROIs were used in the images.
-  **Segmentation based models:** The paper by Li et al. (Li et al., 2018) discusses the Grading Network Head (GNH) as a key component of segmentation task in nuclei. Others have used UNet or variations of it for nuclei segmentation. Zeng et al. (Zeng et al., 2019) used a Residual-Inception network to isolate local features, while Bulten et al. (Bulten, Pinckaers, van Boven, et al., 2020) extended UNet with normalization using CycleGan. Oskal et al. (Oskal, Risdal, Janssen, Undersrud, & Gulsrud, 2019) used a modified UNet for epidermal tissue segmentation. Ren et al. (Ren et al., 2017) used a CNN with encoding and decoding layers for binary classification, while Oda et al. (Oda et al., 2018) proposed a fully convolutional neural network for segmentation using semantic approach. BESNet could be considered as an improved version of UNet with two decoding paths for feature maps used for segmentation.
-  **Mobile architectures:** Arvaniti et al. (Arvaniti et al., 2018) used MobileNet as the base architecture in their study. The MobileNet model was employed to extract crucial spatial features from patches and detect patterns at the regional level, generating pixel-level probability maps for each Gleason grade. Patches were only predicted if the probability exceeded 0.8. They utilized class activation mapping techniques with a fully convolutional architecture. Additionally, the MobileNet model can be converted to a Tensorflow Lite model, making it simple to deploy on Android or Apple devices.
-  **Fusion of CNN and KNN:** Nagpal et al. (Nagpal et al., 2019) used a two-stage model. They first extracted features using ensembled Convolutional Neural Network and then used Kth-nearest-neighbor algorithm to group Gleason grades. The model was trained with region-level features and hard negative mining was applied to the entire training dataset using a partially trained model. In the second stage, labels were collected from pathologists and KNN was used to predict Gleason grades 1, 2, 3, or 4–5, which has been used before by other researchers for Gleason scoring (Wang, Chen, Lu, Baras, & Mahmood, 2020). Hatipoglu et al. (Hatipoglu & Bilgin, 2014) also used SVM (Support Vector Machine) and KNN for similar purposes.

-  **Transfer learning approach using state-of-the-art architecture:** Lucas et al. (Lucas et al., 2019) retrained Inception-v3 CNN to distinguish Gleason Grade 3 and 4( $\leq$ ) from other prostate histopathology tissues using CNTK. Källén et al. (Källén et al., 2016) used a deep Convolutional network pre-trained on a large image dataset with 22 layers to mimic manual Gleason scoring by pathologists and trained a random forest and SVM classifier. Asami et al. (Yonekura, Kawanaka, Prasath, Aronow, & Takase, 2017) proposed a Deep Convolutional Neural Network with 3 convolution layers, pooling layers, and a classifier to extract features from histopathological images and group them into grades. They used MxNet to build their model. Egevad et al. (Egevad et al., 2020) built an AI system to compare and solve Gleason grading complexities using an ensemble of two convolutional deep neural networks with 30 Inceptionv3 models pre-trained on ImageNet and a tailored classification layer.
-  **Novel CNN approach:** Anna et al. (Gummeson et al., 2017) proposed a novel CNN architecture to capture spatial details in images, with network outputs of four channels representing normal-tissue, Gleason grade 3, 4, and 5. They used dropout in the last two layers as regularization, reduced the learning rate for consistency, and implemented the architecture on MatConvNet11. Kwak et al. (Kwak & Hewitt, 2017) built a CNN architecture with six layers of ANN, including three convolution layers, two fully connected layers, and a SoftMax layer. Kumar et al. (Kumar et al., 2017) proposed a CNN architecture for predicting the distance transform of nuclei from images. Hatipoglu et al. proposed (Hatipoglu & Bilgin, 2014) a pyramidal CNN with 4 convolutional layers, 4 subsampling layers and a final dense neural network for classification, multiplied by adjustable weights and a bias term.
-  **Unique graph convolutional approach:** Wang et al. (Wang et al., 2020) proposed a unique method using a Graph Convolution Network (GCN). They represented nuclei or cells as nodes in a graph and accumulated feature vectors from neighboring nodes via iteration. The GCN was able to learn representations for morphological features, texture features, and contrastive predictive coding in each node. The network works similar to a CNN and produced a unique graph representation through max pooling.



### 2.2.3 Significant approaches: A synopsis

Table 2: Summary of different deep learning-based prostate cancer detection & Gleason grading architectures and the techniques used .

[Source: (Linkon, Labib, Hasan, Hossain, & Jannat, 2021)]

Reference	Application	Deep learning architecture	Training	Patch size (pixels)	Accuracy	Dataset
Li et al. (Li et al., 2018)	Two-step training methods for the identification of epithelial cells and Gleason grading simultaneously	R-CNN	end-to-end	512 × 512	Epithelial cells detection accuracy 99.07% Gleason grading (overall pixel) accuracy 89.40%	Privately owned
Nagpal et al. (Nagpal et al., 2019)	Gleason scoring of whole slide images of prostatectomies	CNN	end-to-end	911 × 911	70% accuracy on Gleason scoring task	TCGA Dataset
Arvaniti et al. (Arvaniti et al., 2018)	Heterogeneous Gleason grading	MobileNet	Transfer learning	750 × 750	65% accuracy on evaluation	Privately owned
del Toro et al. (del Toro et al., 2017)	Using transfer learning methods to classify whole slide images considering the region of interest	GoogleNet	Transfer learning	128 × 128, 256 × 256	40% of the patches classified as high grading among 36 out of 46 images with a 78.2% accuracy	TCGA-PRAD Dataset
Lucas et al. (Lucas et al., 2019)	Differentiating GP 3 and GP ≥ 4 from non-atypical tissue	Inception-v3	end-to-end	299 × 299	• Non-atypical and malignant (GP ≥ 3) accuracy 92% •	Privately owned



Reference	Application	Deep learning architecture	Training	Patch size (pixels)	Accuracy	Dataset
					Differentiation between $GP \geq 4$ and $GP \leq 3$ accuracy 90% (sensitivity 77% and specificity 94%)	
Källén et al. (Källén et al., 2016)	Classification of both patches and whole slide images	Deep CNN	end-to-end	$87 \times 87$ , $167 \times 167$	The architecture was designed for Gleason, scoring from 3 to 5. The model provides an accuracy of 81.1% among the patches. Classification of entire images provides an accuracy of 89.2%.	TCGA Dataset
Yonekura et al. (Yonekura et al., 2017)	Disease stage classification method for Glioma histopathological images	Deep CNN	end-to-end	$28 \times 28$	Training and classification accuracies were 98.2% and 87.2%	TCGA Dataset
Bulten et al. (Bulten, Pinckaers, van Boven, et al., 2020)	Automated Gleason grading from prostate biopsies comparing with the pathologists' annotations	UNet	end-to-end	–	AUC for benign vs malignant is 0.990, 95% CI (0.982–0.996)	Radboud University Medical Center Dataset

Reference	Application	Deep learning architecture	Training	Patch size (pixels)	Accuracy	Dataset
Egevad et al. (Egevad et al., 2020)	Identification of areas of Gleason scoring difficulties and comparison of AI system with pathologists' image	InceptionV3	Transfer learning	2048 × 2048 (Whole image)	Gleason score of 3 + 3 = 6, 3 + 4 = 7, 4 + 3 = 7, 4 + 4 = 8, and 9–10 in 13.9% (5), 25.0% (9), 33.3% (12), 19.4% (7), and 8.3% (3) respectively	Pathology Imagebase (ISUP) Dataset
Zeng et al. (Zeng et al., 2019)	Nuclei segmentation using residual inception channel attention UNet architecture	RIC-UNet	end-to-end	224 × 224	The average DICE coefficient is 0.8624	TCGA Dataset
Oskal et al. (Oskal et al., 2019)	Epidermal tissue segmentation	UNet	end-to-end	512 × 512	<ul style="list-style-type: none"> <li>Training dataset mean Positive Predictive Value <math>0.84 \pm 0.28</math> (Sensitivity <math>0.97 \pm 0.07</math>)</li> <li>Test dataset mean Positive Predictive Value <math>0.89 \pm 0.16</math></li> </ul>	University of British Columbia Virtual Slidebox ( <a href="https://pathology.ubc.ca/virtual-collections/">https://pathology.ubc.ca/virtual-collections/</a> )
Ren et al. (Ren et al., 2017)	Computer-aided Imaging method for Gleason pattern 3 and 4 classification using nuclei segmentation technique	UNet	end-to-end	480 × 360	The sensitivity, specificity, and accuracy were $0.70 \pm 0.15$ , $0.89 \pm 0.04$	Privately owned

Reference	Application	Deep learning architecture	Training	Patch size (pixels)	Accuracy	Dataset
					and $0.83 \pm 0.03$	
Kwak et al. (Kwak & Hewitt, 2017)	Localization of epithelial nuclei in prostate tissue microarrays	CNN	end-to-end	$128 \times 128$	AUC of 0.974 with 95% Confidence Interval	Tissue microarray research program at the National Institutes of Health dataset
Kumar et al. (Kumar et al., 2017)	Prostate cancer recurrence prediction using prostate tissue images	CNN	end-to-end	$51 \times 51$	0.81 under AUC curve	CPCTR Dataset
Oda et al. (Oda et al., 2018)	Detection and semantic segmentation of cells using enhanced boundary detection architecture images	UNet	end-to-end	$1636 \times 1088$ (Whole image)	89.5% of the detection rate	Nagoya University Hospital Dataset
Hatipoglu et al. (Hatipoglu & Bilgin, 2014)	Classification of biopsy images with spatial dependencies of cell and non-cell pixels evaluation	CNN, SVM, KNN	end-to-end	$896 \times 768$ (Whole image)	The best accuracy was 86.91%	Yale Tissue Microarray Facility Dataset
Gummeson et al. (Gummeson et al., 2017)	Gleason grading using small convolutional network into grade 3, 4, and 5	CNN	end-to-end	$106 \times 106$	92.7% accuracy	Dataset from PathXL, Belfast, and Beaumont Hospital in Dublin
Wang et al. (Wang et al., 2020)	The spatial organization of cells as a graph to classify into Gleason grades	GCN	end-to-end	$256 \times 256$	88.61% accuracy and AUC of 0.9659	BACH dataset (Aresta et al., 2019)

## 2.3 Overview of available data sources

This section highlights the primary, available data sources of prostate cancer histological images and their most important characteristics.

### **PANDA challenge**

The PANDA (Prostate cANcer graDe Assessment) (Bulten, Kartasalo, Chen, Ström, Pinckaers, Nagpal, Cai, Steiner, van Boven, Vink, et al., 2022) challenge was initiated through a collaboration between the Computational Pathology Group at Radboud University Medical Center and the Department of Medical Epidemiology and Biostatistics at Karolinska Institute. It is the largest publicly available whole-slide image dataset to date. The dataset encompasses a collection of 10621 whole-slide images (WSIs), which are histologically stained with Hematoxylin and Eosin (H&E) and derived from biopsy specimens. These specimens were collected from patients across a time span ranging from 2012 to 2017. In order to facilitate detailed examination and analysis, each image was scanned at a magnification level of 20x and subsequently converted into the TIFF (Tagged Image File Format) for standardized access. The dataset is partitioned into two primary components: a training set and a combined public and private test set, the latter comprising a total of 800 images. Each image within the dataset corresponds to an individual biopsy case; however, it is noteworthy that multiple cases may originate from a single patient, highlighting the dataset's depth in capturing patient-specific pathological variations. Furthermore, the development-training set is characterized by the presence of label noise, which introduces an additional layer of complexity in the analysis and interpretation of the data.

### **TCGA - PRAD**

The TCGA-PRAD (The Cancer Genome Atlas - Prostate Adenocarcinoma) (Zuley, 2016) dataset represents a significant compilation of case studies focused on prostate adenocarcinoma, sourced from a diverse set of 480 cases. These cases have been collected from a variety of geographical locations across the globe, encapsulating a broad spectrum of patient demographics and disease presentations. This global collection effort has led to a heterogeneous dataset, characterized by considerable variability in the images due to the use of different scanner modalities, manufacturing differences, and processing protocols.

### **GLEASON 2019 challenge**

The GLEASON 2019 challenge ("Gleason2019 data | biomedical imaging and artificial intelligence," 2019), introduced as a segment of the grand challenges in pathology at the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2019 conference, focuses on the critical area of prostate cancer grading through the analysis of tissue microarray (TMA) images. This competition provides a structured dataset comprising TMA images specifically aimed at evaluating and enhancing the methodologies used in the automated grading of prostate cancer based on the Gleason scoring system. The dataset presented for this challenge is divided into two main components: a training set and a test set. The training set includes a total of 244 Prostate Tissue Microarray (TMA) images, while the test set comprises 87 TMA

images. All images within this dataset were collected from the Vancouver Prostate Center, ensuring consistency and high-quality across the samples.

## SICAPv2 - Prostate Whole Slide Images with Gleason Grades Annotations

The SICAPv2 database (Spanish Initiative for Pathology Archive with Clinical Annotations, version 2) (J. Silva-Rodríguez, 2020; Julio Silva-Rodríguez et al., 2020) comprises of 155 biopsy samples from 95 patients. These samples were sliced, stained, and digitized at 40x magnification to obtain WSIs. The slides were analyzed by a team of expert urogenital pathologists at the Hospital Clínico of Valencia, and a combined Gleason score was assigned per biopsy. In case of uncertainty, the label was assigned through consensus of all expert pathologists to minimize inter-observer variability. The primary Gleason grade in each biopsy is distributed as follows: 36 non-cancerous regions, 40 samples with Gleason grade 3, 64 with Gleason grade 4, and 15 with Gleason grade 5.

A synopsis of the datasets is presented (Table 3):

*Table 3: Prostate cancer image datasets and their significant characteristics*

Name	Description	Sample Size	Magnification	Annotation	Access
<b>PANDA Challenge</b>	Whole-slide images of H&E-stained biopsy specimens. The largest prostate cancer dataset at the moment. Data collected 2012-2017 from two Institutes. Noisy training data.	10621 WSIs	20x	Gleason score/ISUP grade, label masks	Open access:  <a href="https://www.kaggle.com/competitions/prostate-cancer-grade-assessment/data">https://www.kaggle.com/competitions/prostate-cancer-grade-assessment/data</a>
<b>TCGA-PRAD</b>	Part of The Cancer Genome Atlas, contains prostate adenocarcinoma cases with genomic, transcriptomic, and radiologic data. Collected from various locations. Charecterized by heterogeneity.	480 cases, 771 WSIs	Varied	Gleason score, genomic data	Open access:  <a href="https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=6884022">https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=6884022</a>  <a href="https://www.cbioportal.org/">https://www.cbioportal.org/</a>

Name	Description	Sample Size	Magnification	Annotation	Access
					study/summary?id=prad_tcga
<b>GLEASON 2019 Challenge</b>	Tissue microarray (TMA) images, enhancing diagnostic models based on the Gleason scoring system. Primary features are the consistency and high-quality. Part of MICCAI 2019.	331 TMA images (244 training set- 87 test set)	Not specified	Gleason score	Open access:  <a href="https://gleason2019.grand-challenge.org/Register/">https://gleason2019.grand-challenge.org/Register/</a>
<b>SICAPv2</b>	Whole-slide images of prostate biopsies that analyzed by a group of expert urogenital pathologists, small inter-observer variability.	95 patients, 155 samples: 36 non-cancerous samples, 40 with Gleason grade 3, 64 with Gleason grade 4 and 15 with Gleason grade 5	40x	Gleason score	Open access:  <a href="https://data.mendeley.com/datasets/9xxm58dvs3/2">https://data.mendeley.com/datasets/9xxm58dvs3/2</a>

## 3. METHODOLOGY

### 3.1 Data acquisition

#### 3.1.1 Dataset overview

The dataset that used for processing and the development of the algorithm is the PANDA challenge (Prostate cANcer graDe Assessment) (Bulten, Kartasalo, Chen, Ström, Pinckaers, Nagpal, Cai, Steiner, van Boven, & Vink, 2022; Bulten, Pinckaers, & Eklund, 2020). It comprises 10621 whole-slide images (WSIs) of digitized Hematoxylin and Eosin (H&E) stained biopsies. The fact that the collection is sourced from two distinct centers, significantly enhancing its diversity and representativeness. With a total size of approximately 411.9 GB of “.tiff” images, it stands as the largest public WSI dataset available for prostate cancer. Each “tiff” file consists of three layers of magnification. The full resolution layer and thus the highest level of detail is in 20x magnification. Each image has a corresponding label that reflects the grade of biopsy cancer. The labels are provided through a “.csv” file, where the ID of the image matches with the proper label value. There are six label values (0-5). Along with the images, label masks (Figure 3) are provided to indicate regions of interest (ROIs) that contributed to the determination of the label value. The grading of label masks by the data providers, namely the Radboud University Medical Center and the Karolinska Institute, follows specific conventions in pixel values:

Radboud University Medical Center

- **0:** Background (non-tissue) or unknown
- **1:** Stroma (connective tissue, non-epithelium tissue)
- **2:** Healthy (benign) epithelium
- **3:** Cancerous epithelium (Gleason 3)
- **4:** Cancerous epithelium (Gleason 4)
- **5:** Cancerous epithelium (Gleason 5)

Radboud's masks provide a detailed view of the tissue, distinguishing between different types of epithelium and stroma, thus offering a granular approach to understanding tissue composition.

Karolinska Institute

- **0:** Background (non-tissue) or unknown
- **1:** Benign tissue (stroma and epithelium combined)
- **2:** Cancerous tissue (stroma and epithelium combined)

### 3.1.2 Labels - ISUP grading

The International Society of Urological Pathology (ISUP) (Jonathan I Epstein et al., 2016) grading system for prostate cancer is a refinement and update of the traditional Gleason scoring system, aimed at providing a more standardized and prognostically relevant assessment of prostate cancer severity. The ISUP grading system categorizes prostate cancer into 5 grade groups based on the Gleason scores, with 1 being the least aggressive and 5 the most aggressive form of prostate cancer. Here is what each ISUP Grade Group means:

- **ISUP Grade Group 1:** Gleason score  $\leq 6$  (Gleason pattern 3 + 3). This group represents the least aggressive form of prostate cancer, with a good prognosis. Tumors in this group are well differentiated.
- **ISUP Grade Group 2:** Gleason score 7 (Gleason pattern 3 + 4). In these tumors, the majority of the cancer pattern is still well differentiated, but there is a significant component that is more aggressive and poorly differentiated.
- **ISUP Grade Group 3:** Gleason score 7 (Gleason pattern 4 + 3). Here, the more aggressive, poorly differentiated pattern 4 cancer makes up the majority of the tumor, indicating a worse prognosis than Grade Group 2.
- **ISUP Grade Group 4:** Gleason score 8 (Gleason patterns 4 + 4, 3 + 5, or 5 + 3). Tumors in this group are poorly differentiated and more aggressive, with a significant risk of adverse outcomes.
- **ISUP Grade Group 5:** Gleason scores 9-10 (Gleason patterns 4 + 5, 5 + 4, or 5 + 5). This group includes the most aggressive and poorly differentiated tumors, with the highest risk of metastasis and poor outcomes.

### 3.1.3 Dataset characteristics

The notable characteristics of the dataset are:

1. **High-Resolution Images:** Individual images within the dataset boast very high resolutions, with dimensions such as (27648, 29440) and file sizes up to 60 MB each. Such high resolutions are necessary to capture the detailed histological structures of prostate tissue but present significant data handling challenges.
2. **Variability in Image Dimensions:** No consistent shape or size ratio exists across the dataset, requiring flexible preprocessing and analysis pipelines to handle the variability.
3. **Presence of Empty Content:** A significant portion of the images consists of whitespace (non-tissue areas), which can complicate the identification of relevant tissue regions for analysis (Figure 2).
4. **Imperfect labels:** Despite the structured guidelines provided by systems like the Gleason score and ISUP grading, individual pathologists may interpret the same slide differently due to several factors. Thus, the dataset includes labels



with inherent inaccuracy. Also, not all images have corresponding segmentation masks.

5. **Stain/color differentiations:** The dataset presents an added layer of complexity due to the inclusion of data from two different institutions, each potentially using slightly different staining protocols or slide preparation techniques. This variability introduces differences in stain appearance.
6. **Unbalanced classes of labels:** Unbalanced classes refer to the uneven distribution of images across distinct categories or grades of prostate cancer severity, as defined by the ISUP grading system.
7. **Image channels:** The original histological images consist of three-color channels, while the label masks of one channel.

*Figure 2: Sample image in low resolution*

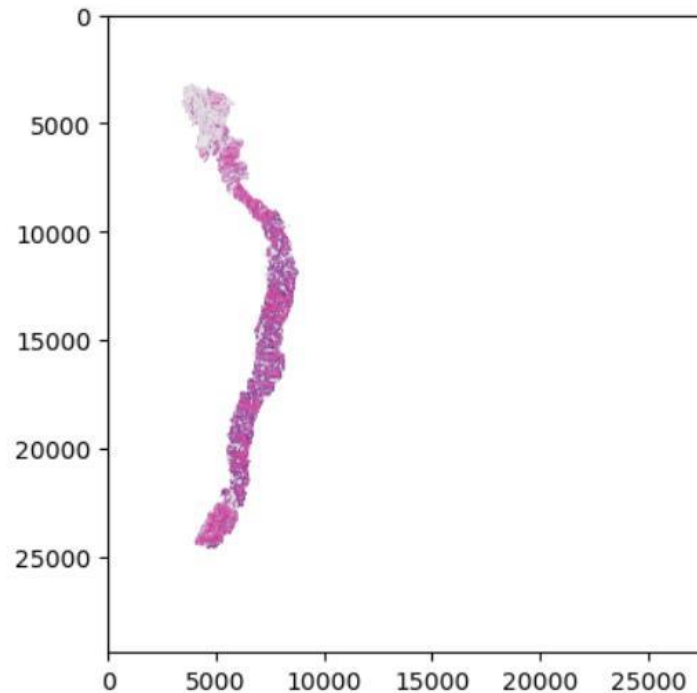
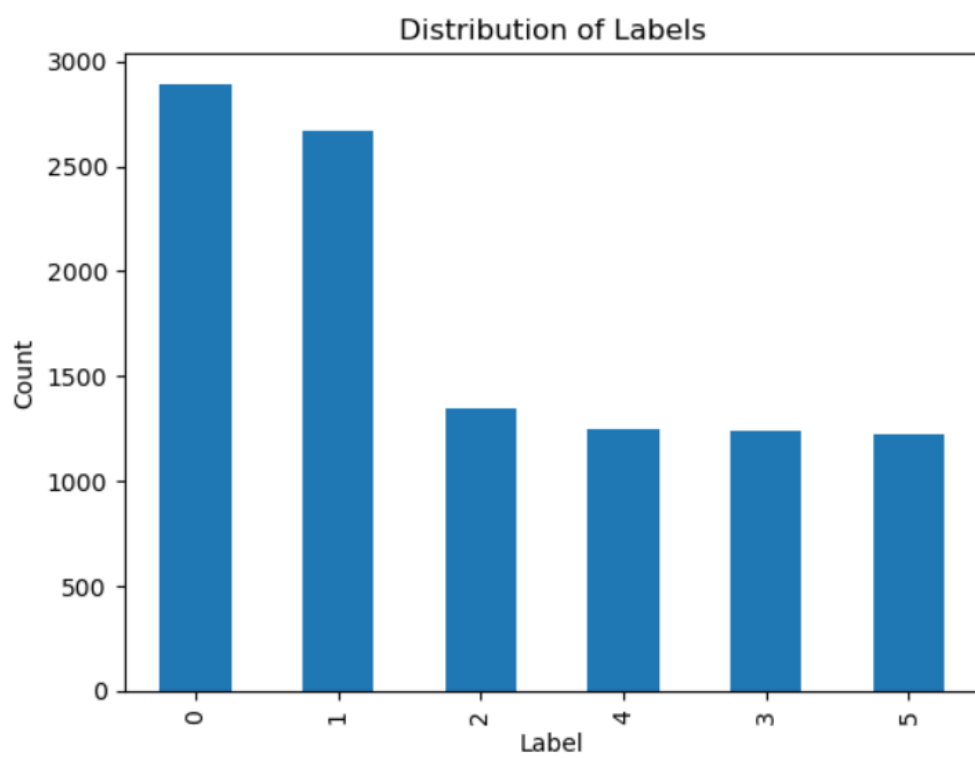


Figure 3: Corresponding label mask



Figure 4: Labels distribution



## 3.2 Pre-processing

The chosen preprocessing approach and algorithm try to address some of the inherent, previous mentioned, dataset challenges:

1. High-resolution images
2. Variability in image dimensions
3. Presence of empty content
4. Focus on cellular-level features

The main goal was to divide the images into smaller regions, retaining essential information (Figure 5) and eliminating what is unnecessary for the classification, based on the relevant approach of 'Concat tile pooling' (<https://www.kaggle.com/code/iafoss/panda-16x128x128-tiles/notebook>). By dividing large, high-resolution images into smaller patches, the computational models can process and analyze the data more efficiently. The initial, images are too large to fit into memory, but smaller patches are manageable. Also, patches allow for focusing on details and for cellular structures examination, which are crucial for understanding disease mechanisms. A considerable drawback is the context loss. Discarding parts of the image can lead to a loss of larger context and spatial relationships between different regions of the tissue, which can affect the diagnosis.

Thus, the main function of the pre-processing algorithm (Appendix A1) performs as follows:

**Inputs:** a. Histological image, b. the size of patch, c. the number of patches

➤ **Step 1: Calculate padding**

The function begins by calculating how much padding is needed to make the image dimensions evenly divisible by the defined patch size. This ensures that all patches will have uniform dimensions. It calculates both horizontal (width) and vertical (height) padding.

➤ **Step 2: Pad the image**

Next, the image is padded with the calculated padding values. Padding is applied equally on both sides of each dimension (top-bottom for height, left-right for width), ensuring the image remains centered. The padding is filled with a constant white value (255 for each channel in the RGB image).

➤ **Step 3: Reshape and rearrange the image into patches**

This step transforms the padded image into smaller, uniformly sized segments. The padded image starts as a 3D array with dimensions (height, width, channels). Then the image is reshaped into a 5D array. The reshaping adjusts the array to dimensions that reflect a grid layout along with the patch size and the color channels. The padded image is reshaped to (grid\_height, patch\_height, grid\_width, patch\_width, channels). This structure groups the image data into a 5D array where:

- grid\_height and grid\_width correspond to the number of patches vertically and horizontally across the image.
- patch\_height and patch\_width are the dimensions of each patch.
- channels remain the same, indicating the color depth.

Then the array is rearranged to bring the grid dimensions next to each other and keep the patch dimensions together, resulting in (grid\_height, grid\_width, patch\_height, patch\_width, channels). The array is finally reshaped again to flatten the grid, creating a 4D array where the first dimension indexes each patch. The final structure is (num\_patches, patch\_height, patch\_width, channels), where num\_patches is the total number of patches, calculated as grid\_height \* grid\_width. Each patch is now a small image of the specified size.

➤ **Step 4: Ensure a minimum number of patches**

If the reshaped image results in fewer patches than the specified minimum number (n\_patches), the array of patches is padded with additional patches filled with the constant padding value (white). This ensures that there are always exactly n\_patches patches, even if some of them are just blank (white) patches.

➤ **Step 5: Select the most informative patches**

The function then flattens each patch to a single vector and calculates the sum of its pixel values. The assumption here is that patches with a lower sum (darker patches) are more informative than patches with a higher sum (lighter or blank patches). It sorts the patches based on this sum and selects the top n\_patches most informative patches. This step prioritizes patches with more content and less whitespace.

➤ **Step 6: Package patches with indices**

Finally, each selected patch is packaged into a dictionary with its corresponding index (idx), which serves as a unique identifier for the patch within the image. This index will be used for tracking and referencing patches after processing.

**Output:** List of dictionaries

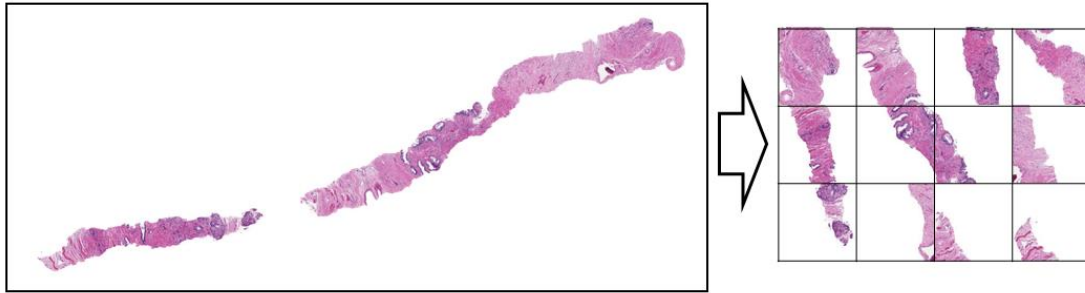
The result is a list of dictionaries, each representing a patch from the original image. Each dictionary contains the patch image and its index.

The algorithm input values are defined based on the accuracy of final results and the size of the output, across the different tries. The selected values are:

- Number of patches: 14
- Patch size: 256

*Figure 5: Patch extraction.*

[Source: <https://www.kaggle.com/code/iafoss/panda-16x128x128-tiles/notebook>]



## 3.3 Model development

### 3.3.1 CNN architecture

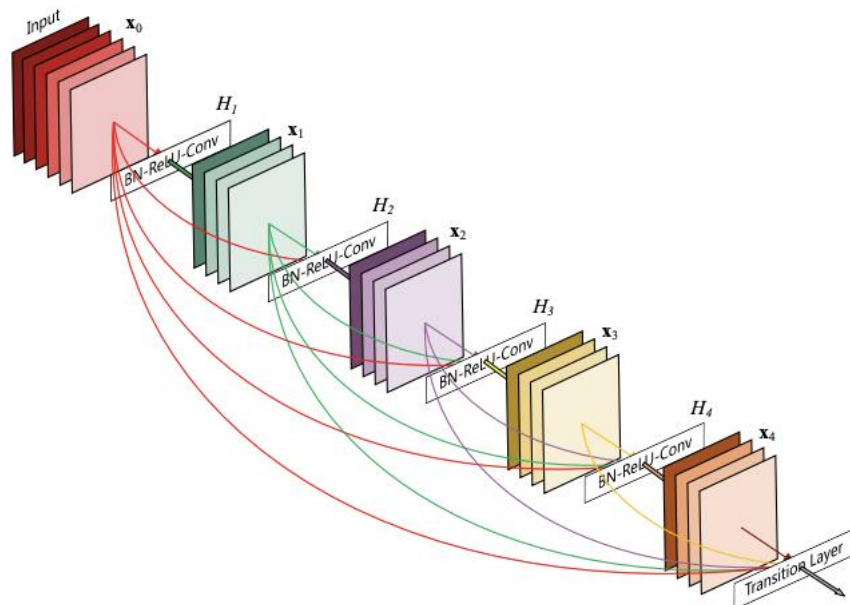
#### CNN Components

##### Convolutional Layers

The model leverages a pre-trained DenseNet201 as backbone, which is integrated into the DenseNetModel class. DenseNet201 is known for its densely connected convolutional networks (Figure 6), where each layer is connected to every other layer in a feed-forward fashion. For feature extraction, the encoder is built by extracting all layers except the final classification layer from the pre-trained DenseNet201. This structure allows the model to benefit from hierarchical feature representation, capturing both low-level features (edges, colors) and high-level features (textures, patterns) from histological images.

Figure 6: Visual representation of DenseNet architecture

[Source: [https://pytorch.org/hub/pytorch\\_vision\\_densenet/](https://pytorch.org/hub/pytorch_vision_densenet/)]



#### Activation Function

The Mish activation function (Misra, 2019) is a novel, non-monotonic function that has been demonstrated to improve model performance and generalization in deep learning architectures. It is defined as  $x \cdot \tanh(\ln(1 + e^x))$ , where  $x$  is the input to the function. This function is particularly used in the model to introduce non-linearities, allowing the network to learn complex patterns in the data more effectively than traditional functions like ReLU. Mish aims to overcome the vanishing gradient problem by ensuring a

smooth gradient flow. It has properties that help in preventing the loss of information during transmission through deep networks, making it suitable for tasks that require preserving intricate details, such as in histological image analysis.

## Pooling layers

- 🌈 **Adaptive Average Pooling:** This layer calculates the output feature map of a specified size (e.g.,  $H' \times W'$ ) by dividing the input feature map into sub-regions and computing the average of each sub-region. The adaptive nature allows it to handle any input size and adaptively select the pool size to ensure the output dimensions are consistent. Conceptually, this means it can take an input feature map of any dimension and produce an averaged output of a predetermined size, which simplifies downstream processing and makes the architecture more flexible. In this instance, it operates similarly to Global Average Pooling and outputs the feature average value across all the spatial dimensions of an individual patch.
- 🌈 **Adaptive Max Pooling:** Similar to adaptive average pooling, but instead of averaging the elements within each sub-region, it selects the maximum element. This approach is particularly useful for emphasizing the most prominent features within a feature map, ensuring that the most critical signals are not diluted during the pooling process. In this instance, it operates similarly to Global Max Pooling and outputs the feature max value across all the spatial dimensions of an individual patch.

## Patch aggregator class

A custom-designed component responsible for aggregating features from individual patches of histological images into a cohesive representation that can be used for classification. It incorporates an attention mechanism that allows the model to focus on the most informative parts of the image patches. Its concept and functionality:

- 🌈 **Focus on informative features:** The attention mechanism enables the model to weigh the features extracted from different patches differently, giving more importance to the features that are more relevant for the classification task.
- 🌈 **Learning to attend:** The attention mechanism learns to assign these weights through the training process, effectively learning which parts of the image are most important for making accurate classifications. This is achieved by a learnable set of layers that takes the features from each patch as input and outputs a value of weight or score that indicate the importance of each patch's features. This value ranges from 0 to the number of patches extracted from the same image.
- 🌈 **Aggregation of patch features:** Takes the features extracted from the individual patches and combines them into a single, comprehensive feature vector or tensor that represents the entire image by calculating the weighted mean for each feature across all correlated patches.

## Classification head

The custom classification head is a tailored component designed to finalize the decision-making process of the model by interpreting the aggregated features from the entire image or significant patches and producing a prediction regarding the presence of cancer. It consists of:

- 🎨 **Fully connected layers:** Dense layers that process the aggregated feature vector from the previous stages of the model. These layers are crucial for learning non-linear combinations of the high-level features that are indicative of the target classes.
- 🎨 **Batch normalization:** This technique normalizes the output, reducing internal covariate shift and improving the stability of the neural network. It can lead to faster convergence and improved overall performance.
- 🎨 **Mish activation function:** Operates as described previously.
- 🎨 **Dropout and Regularization:** To prevent overfitting, especially given the high dimensionality of the feature vectors and the complex nature of histological images, dropout layer randomly drop units from the fully connected layers during training.
- 🎨 **Output Layer:** The final fully connected layer of the classification head returns an output size matching the number of target classes (e.g., binary for cancerous vs. non-cancerous).

### 3.3.2 Implementation

The process (Figure 7) that a set of correlated patches follows through the network involves a series of steps designed to handle the absence of the WSI itself and defined through the main script (Appendix A2). Here is an overview of how the network handles a WSI as a combination of patches respecting its architecture (Appendix A3).

1. **Receiving a Batch of Sets:** The network starts by receiving a batch of sets (tensors) where each set contains patches extracted from the same histological image.
2. **Reshaping for individual processing:** Before processing, the patches are reshaped as necessary to match the input requirements of the encoder. This step ensures that each patch can be individually processed by the convolutional layers of the network, allowing for the extraction of detailed features from each patch.
3. **Passing individual patches through the encoder:** Each patch is then passed through the encoder and feature maps for the individual patches are generated.
4. **Pooling operations on patch features:** After feature extraction, pooling operations are applied to the features of each patch. These include both average pooling and max pooling, which reduce the dimensionality of each feature map while retaining essential information.
5. **Aggregating patch features - Unifying feature maps:** The features from all patches of a given image are then aggregated into a single representation. This aggregation is performed from the patch aggregator module. The goal is to



combine the information from all patches to create a comprehensive feature set that represents the entire image.

6. **Classification head for final classification:** Finally, the aggregated features are passed to the custom classification head. This component of the network makes the final classification decision. It outputs the probability that the image contains cancerous tissue, based on the learned patterns and features from the training process.

Figure 7: CNN architecture – Network process

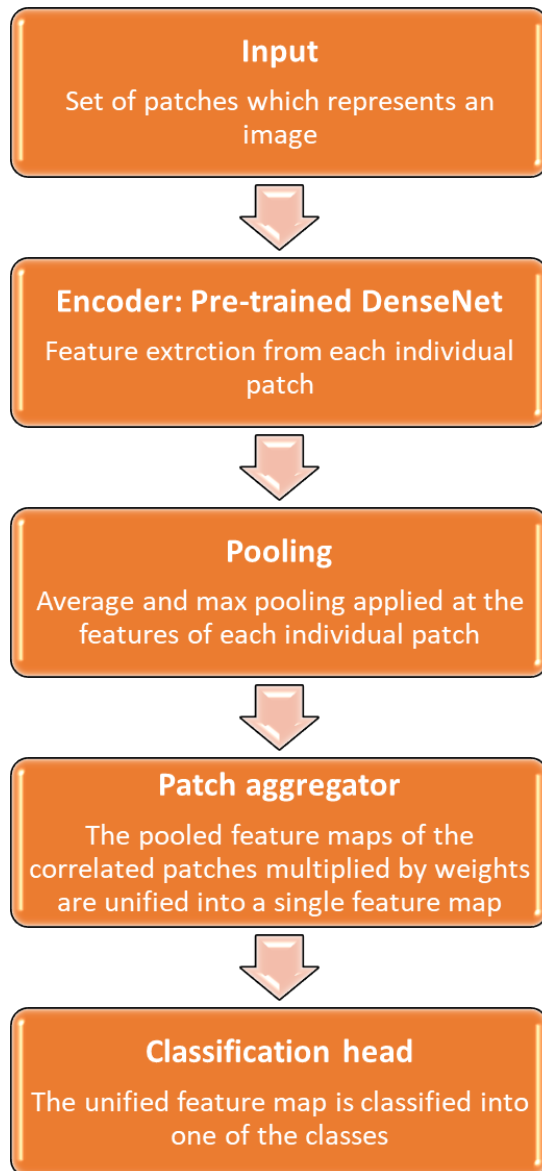
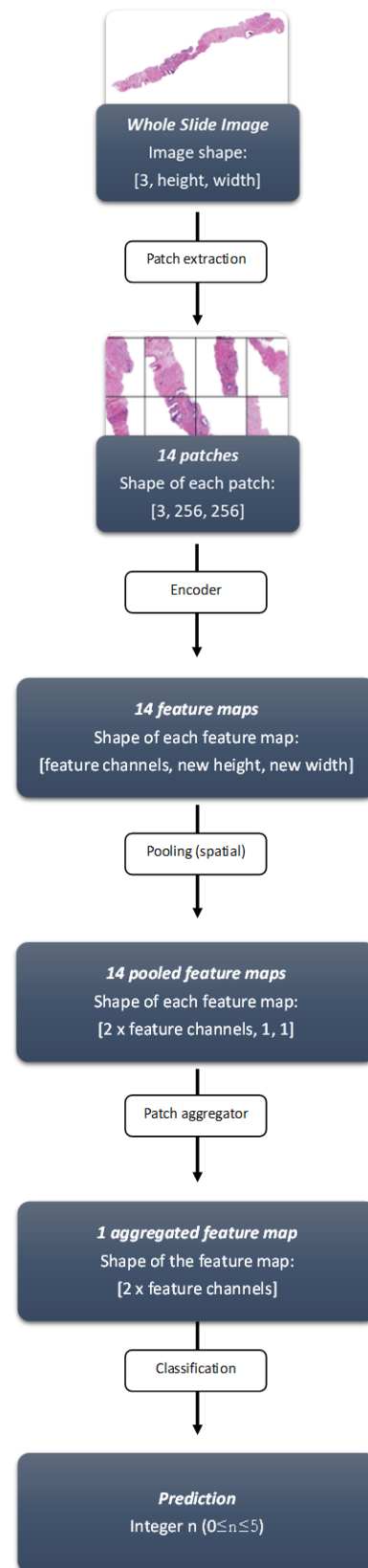


Figure 8: Whole processing of the WSI – An example



## 3.4 Training

### 3.4.1 Data preparation

The data preparation for CNN feeding was implemented through the following steps:

1. **Loading dataset labels:**

The preparation script starts by reading the dataset labels from a CSV file, which includes image IDs and corresponding classes or grades.

2. **Filtering and sorting files:**

Then, the available image files based on their IDs are filtered and sorted in order to be ensured they match the labels.

3. **Train/Validation/Evaluation split:**

A stratified split based on the labels creates the training, validation, and evaluation sets. This helps maintain the distribution of classes across both sets. The evaluation set is used to evaluate the model performance on unseen images after the training and consists of 10% of the initial dataset. The training and validation sets consist of 85% -15% of the remaining set.

4. **Data Augmentation:**

Augmentations like random flips, rotations, and cropping are applied to image patches. This enhances the diversity of the training set and helps the model generalize better to unseen data.

5. **Dynamic padding/cropping:**

Patches are dynamically padded to ensure a uniform size, followed by a conversion from image to tensor.

6. **Image normalization:**

Image patches are normalized using predefined mean and standard deviation values, aligning pixel value distributions, and facilitating model convergence.

7. **Batch preparation:**

A custom collate function is used to assemble batches of processed patches, making them ready for training or validation through the model.

### 3.4.2 Loss function

A custom loss function is used as a variation of the Focal Loss and designed to address class imbalance in classification tasks. Focal Loss is an enhancement over the classic Cross-Entropy Loss, introducing a modulating factor to focus learning more on hard, misclassified examples and less on easy examples. This approach helps in improving

model performance on imbalanced datasets. Here is a detailed breakdown of its implementation and purpose within the context of this model:

### Focal Loss components

1. **Cross-Entropy Loss:**

The base of the Focal Loss, which measures the difference between the predicted class probabilities and the actual class labels.

2. **Modulating Factor:**

A scaling term added to the Cross-Entropy Loss, dependent on the correct classification probability. For each sample, it is calculated as  $(1-p_t)^\gamma$ , where  $p_t$  is the model's estimated probability for the class with the true label, and  $\gamma$  is a focusing parameter.

3. **Class weights (alphas):**

Pre-computed weights for each class to address class imbalance. These weights are applied to the loss of each sample, further helping to focus the model's learning on under-represented classes.

### Implementation in the script


- 🔗 **Alphas Calculation:** The script calculates class weights (alphas) based on the inverse frequency of each class in the dataset. These weights are then normalized and scaled to ensure that the model pays more attention to less frequent classes, by setting bigger penalties to the model predictions on underrepresented classes through the loss function.
- 🔗 **Focal Loss function:** Defined as a custom module, the function takes the model's predictions, the true labels, and the pre-computed alphas as inputs. It computes the Cross-Entropy Loss for each sample, modulates it based on the sample's classification probability and class weight, and applies the focusing parameter  $\gamma$  to adjust the impact of each sample on the loss based on how well it was classified.
- 🔗 **Gamma parameter:** The focusing parameter  $\gamma$  is set to adjust the rate at which easy examples are down-weighted. Higher values of  $\gamma$  increase the effect, pushing the model to focus more on hard-to-classify examples.
- 🔗 **Reduction:** The loss values across all samples are aggregated using a reduction method (mean in this case), which determines how individual sample losses contribute to the overall loss value that the model optimizes during training.

### 3.4.3 Optimizer

The Ranger optimizer is a synergistic combination of RAdam (Rectified Adam) and Lookahead, integrating their respective advantages to improve the training process.

- 🔗 **RAdam (Rectified Adam):** RAdam is an enhancement over the traditional Adam optimizer. It introduces a rectification term to adjust the adaptive learning rate, aiming to stabilize and improve the training phase in the initial epochs when the variance of the adaptive learning rate can be high. This rectification

helps in mitigating the warm-up phase issue seen in Adam, leading to more stable and consistent training performance from the start.

 **Lookahead:** Lookahead is a mechanism that periodically updates the weights of the neural network by interpolating between the current weights and the weights from a number (defined by the parameter  $k$ , default:  $k=6$  steps) of steps ahead in the gradient descent path. This process helps in exploring a wider area of the weight space, leading to potentially better and more robust solutions.

The combination of Lookahead with RAdam in this case, provides both the benefits of refined gradient descent steps (through RAdam) and a broader exploration of solution spaces (through Lookahead), enhancing the model's ability to find optimal solutions.

### 3.4.4 Hyperparameters

The hyperparameters defined after some trials:

1. **Learning Rate:** Affects how much the model weights are updated during training. The script employs a learning rate finder to identify an optimal maximum learning rate. Also, a (FastAI's) tool is employed to dynamically adjust the learning rate during the training.
2. **Number of epochs:** 12-15 on this occasion. This determines how many complete passes the model makes over the entire training dataset.
3. **Batch size:** 8-10
4. **Gamma (in Focal Loss):** 2

## 3.5 Evaluation metrics and validation methods

The script employs two key metrics for evaluating model performance: the Multiclass ROC AUC and Cohen's Kappa.

### 3.5.1 Multiclass ROC AUC (Area Under the Receiver Operating Characteristic Curve)

✚ **Purpose:** Measures the model's ability to distinguish between classes across all thresholds, providing a single measure of performance regardless of class imbalance.

✚ **How It Works:**

- **ROC Curve:** Plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- **AUC (Area Under the Curve):** The area under the ROC curve; a higher AUC indicates better model performance, with 1.0 being perfect discrimination between classes and 0.5 representing no better than random chance.
- **Multiclass adaptation:** For multiclass problems, the ROC AUC is calculated for each class against all others and then averaged, either through a one-vs-rest approach or by considering each class pair in a one-vs-one strategy.

✚ **Benefits:**

- Effective for imbalanced datasets as it evaluates model performance across all classification thresholds.
- Provides a comprehensive overview of model's discriminative ability.

### 3.5.2 Quadratic Weighted Kappa (Cohen's Kappa)

✚ **Purpose:** Quantifies the agreement between the predicted and actual classifications, adjusting for the agreement that could occur by chance.

✚ **How It Works:**

– **Calculation:**

- Confusion Matrix:** First, a confusion matrix  $O$  is computed, which contains the actual vs. predicted ratings.
- Weight Matrix:** A weight matrix  $W$  is created, where each element  $W_{i,j}$  is the square of the difference between the actual and predicted ratings, normalized by the maximum possible difference. This quadratic weighting penalizes disagreements more if they are further apart.

$W_{i,j} = \left(\frac{i-j}{N-1}\right)^2$  where  $i$  and  $j$  are the actual and predicted ratings, respectively, and  $N$  is the number of ratings.

- iii. **Expected Ratings Matrix:** An expected ratings matrix  $E$  is calculated under the hypothesis of random chance agreement.
- iv. **Kappa Calculation:** Finally, the kappa score is calculated as:

$$\kappa = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}}$$

- **Interpretation:** A Kappa of 1 indicates perfect agreement; 0 indicates no agreement beyond chance; negative values indicate agreement less than chance, which is a sign of serious misclassification issues.

#### **Benefits:**

- i. Useful in cases where the ratings are ordinal and can take on a range of values. It compensates for the agreement occurring by chance.
- ii. The "weights=quadratic" parameter indicates that the differences between classes are treated quadratically, meaning that the metric penalizes disagreements between the ratings more severely as the disagreement becomes larger.

## 4. RESULTS

### 4.1 Confusion matrix

A  $n \times n$  confusion matrix,  $n \geq 2$ , is a specific table layout that allows visualization of the performance of a classification algorithm. Each row of the matrix represents the instances in an actual class, while each column represents the instances in a predicted class. The elements of the  $n \times n$  confusion matrix, denoted as  $C_{ij}$ , represent the number of observations known to be in group  $i$  but predicted to be in group  $j$ . Here,  $i$  and  $j$  index the actual and predicted classes, respectively. For a classification problem with  $n$  classes:

- The diagonal elements  $C_{ii}$  (where  $i=1,2,\dots,n$ ) represent the number of correct predictions for each class, indicating instances where the predicted class matches the actual class.
- The off-diagonal elements  $C_{ij}$  (where  $i \neq j$ ) represent the misclassification errors: instances where the actual class is  $i$  but the model predicts class  $j$ .

Specifically, in a  $n \times n$  confusion matrix:

- The sum of the elements in row  $i$ ,  $\sum_{j=1}^n C_{ij}$ , gives the total number of instances actually belonging to class  $i$ .
- The sum of the elements in column  $j$ ,  $\sum_{i=1}^n C_{ij}$ , provides the total number of instances predicted to belong to class  $j$ .
- The sum of all diagonal elements,  $\sum_{i=1}^n C_{ii}$ , gives the total number of correct predictions made by the model.
- The overall accuracy of the model can be calculated as the ratio of the sum of the diagonal elements to the total number of instances,  $\frac{\sum_{i=1}^n C_{ii}}{\sum_{i=1}^n \sum_{j=1}^n C_{ij}}$ .

The confusion matrix thus allows for the easy identification of confusion between classes, with the main diagonal values representing correct predictions and the off-diagonal values indicating errors. In this case, the confusion matrix that is produced by the predictions of the model on the evaluation set, which contains a total of 1,062 (10% of the initial dataset) unseen images is:

```
[[271 13 4 0 1 0]
 [ 52 181 22 8 4 0]
 [ 7 42 64 13 5 3]
 [ 9 3 25 63 17 7]
 [ 9 3 6 23 74 10]
 [ 6 3 6 13 16 79]]
```



## **4.2          Multiclass ROC AUC**

The score of the multiclass ROC AUC on the evaluation set is: 0.91

## **4.3          Quadratic Weighted Kappa**

The Quadratic Weighted Kappa on the evaluation set is: 0.83

## 5. DISCUSSION



### 5.1 Confusion matrix interpretation

Based on the result of the confusion matrix, the following interpretations could be drawn for each class:

- **Class 1 performance:** The model performs best at classifying Class 1 instances, with the highest number of true positives (271) and relatively few misclassifications. This could indicate that the features for Class 1 are well-defined or significantly different from other classes, making it easier for the model to distinguish. That means that the model can identify with high accuracy the presence of cancer and it could potentially perform well in a binary classification (cancerous tissue or benign) having considerable number of TP (True Positives) in class 1 prediction.
- **Class 2 performance:** There are noticeable misclassifications with Class 2 instances being predicted as Class 1 (about 19%). This suggests that there might be some overlap in the feature space or similarities between Class 1 and Class 2 that confuse the model.
- **Class 3 performance:** Class 3 seems the most challenging for the model, with a significant number of instances misclassified as Class 2 (about 31%). This could mean that the features defining Class 3 are not as distinct or that the images of this class contain noise or that the model requires more data to improve its predictions for this class.
- **Classes 4,5 performance:** Classes 4 and 5 have a mixed performance. While there are significantly more true positives than any other single category, the spread of misclassifications across other classes suggests that the model may have a discriminatory ability for these, but not too strong.
- **Class 6 performance:** Class 6, while having a few misclassifications, seems to have a relatively high number of true positives (79) compared to false positives and false negatives. This suggests that while the model has a good ability to correctly identify Class 6, there is room for improvement.

### 5.2 Performance factors

In this section, the pivotal factors that significantly influenced the performance of the model are delineated. The following parameters were methodically evaluated to optimize the model's predictive accuracy:

-  **Magnification/Resolution layer selection:** The initial histological images, stored in the multi-layered ".tiff" format, comprised three magnification levels. An empirical analysis was conducted to ascertain the impact of these varying resolutions on the model's efficacy. It was discerned that the intermediate magnification layer was most conducive to extracting representative features from the histological images, while keeping the output within manageable limits, which, in turn, enhanced the classification results.
-  **Patch number and resolution optimization:** A series of experiments was undertaken to evaluate the optimal number and resolution of image patches for

training the CNN. Various configurations were tested, encompassing 40 patches of 62x62 pixels, 16 patches of 128x128 pixels, 18 patches of 192x192 pixels, 16 patches of 224x224 pixels, and 14 patches of 256x256 pixels. The experiments demonstrated that the configuration utilizing 14 patches at a resolution of 256x256 pixels outperformed the alternative patch configurations. This finding underscores the critical nature of patch resolution in capturing the nuanced histological patterns necessary for accurate classification.

🌈 **Training epoch threshold:** The determination of an adequate number of training epochs was essential to ensure the model's convergence and generalization capabilities. It was found that a minimum of 8 epochs was necessary to achieve stable and reproducible results. Training the CNN for fewer epochs led to underfitting, indicating the model's inability to fully learn and generalize from the training data.

In conclusion, the selection of an appropriate magnification layer for patch extraction, coupled with the optimization of patch number and resolution, were contributory in enhancing CNN's performance. Furthermore, establishing a minimum epoch requirement was crucial in ensuring the robustness of the model's classification capabilities.

### 5.3 Noticeable attempts in model development and training

Within the scope of this study, several methodological approaches were hypothesized to enhance the performance of the model for the final classification. However, upon testing, certain strategies did not yield the anticipated improvements and were subsequently deemed non-contributory to the model's efficacy. This section provides an analytical overview of these approaches.

**1. Incorporation of label masks into training:** Attempts were made to integrate label masks into the training process to provide the CNN with spatial context that could potentially aid in classification. Two distinct methodologies were employed: (i) image enhancement within regions of interest (ROIs) as delineated by label masks, such as through contrast augmentation, and (ii) a dual-input model architecture wherein a parallel branch processed features derived from the label masks. The classification head was designed to consider features from both the original images and the label masks. Contrary to expectations, these methods did not result in measurable performance gains, indicating that the inclusion of spatial context via label masks did not align synergistically with the model's learning mechanisms.

**2. Stain/Color normalization:** Considering the variability in histological staining procedures, stain normalization was postulated to standardize the input data and thus benefit CNN's performance. Stain normalization was implemented using the torchstain toolkit (<https://github.com/EIDOSLAB/torchstain>). Despite the theoretical rationale, this approach did not improve the model's classification accuracy.

**3. Application of image enhancement filters:** A series of image preprocessing filters, such as sharpening, were evaluated with the intention of accentuating image features, and aiding the CNN's feature extraction processes. Finally, these preprocessing steps did not confer an advantage. On

the contrary, they may have introduced artifacts or exaggerated noise, thereby obscuring the histological patterns crucial for accurate classification.

In summary, the aforementioned strategies, while conceptually sound, did not translate into improved performance for the CNN model in the context of histological image classification. This underscores the complexity of model development where certain intuitive interventions may not align with the intricate nature of deep learning models.

## 5.4 Approaches comparison

In this section, approaches on the same dataset and their findings will be presented. The work that is used as reference (<https://www.kaggle.com/code/iafoss/panda-concat-tile-pooling-starter-0-79-lb/notebook>) is derived from the highest ranked notebook, as of the current date, in terms of votes among all submissions to the PANDA Challenge competition, that implements a full model training and evaluation. It is the third-highest ranked notebook, based on voter preference. Additionally, it ranks second in terms of the number of comments received. This approach has served as a foundational methodology for numerous competitors, including those whose solutions were ranked near the top positions. The methodology from this work has been utilized as a basis for the research presented in the current thesis.

### The findings-results of the reference work

- **Quadratic Weighted Kappa: 0.776**
- **Confusion matrix:**

```
[[2511 250 6 11 42 53]
 [ 512 1629 237 88 97 53]
 [ 96 348 382 224 175 116]
 [ 68 81 58 340 309 370]
 [ 88 77 20 66 607 387]
 [ 70 21 11 31 164 918]]
```

The reference model under consideration was assessed using the entirety of the dataset, encompassing both the training and validation subsets. Consequently, its outcomes are not directly comparable to the results produced by the model developed in the current thesis. For a direct comparison, a confusion matrix that encompasses the entire dataset would be requisite. However, implementing such a matrix for the full dataset may not represent a methodologically robust approach.

### Main difference in methodology

The primary difference lies within the CNN architecture employed by the two approaches. While both strategies process individual patches through their respective encoders, they diverge significantly in their method of combining feature maps before the classification head. The approach of the reference work, following feature extraction, concatenates the features from correlated

patches and performs spatial pooling. This pooled feature map is then directed towards the classification head. Conversely, the strategy of current thesis work conducts spatial pooling on each individual patch immediately after feature extraction. Subsequently, it combines these pooled features across all patches by calculating weighted means, which takes into account the importance of each patch. This nuanced approach towards feature aggregation emphasizes the distinct consideration of patch significance on the final prediction before the collective feature map proceeds to the classification head.

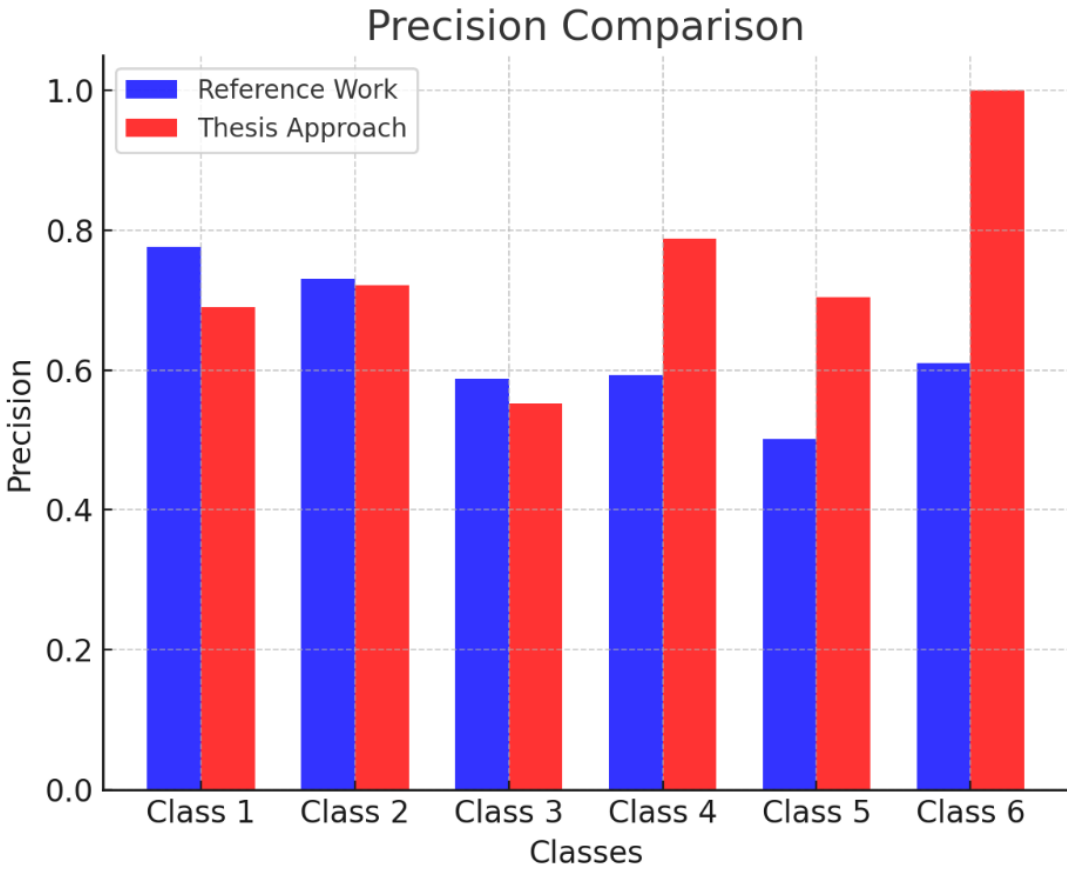
#### **Comments and observations**

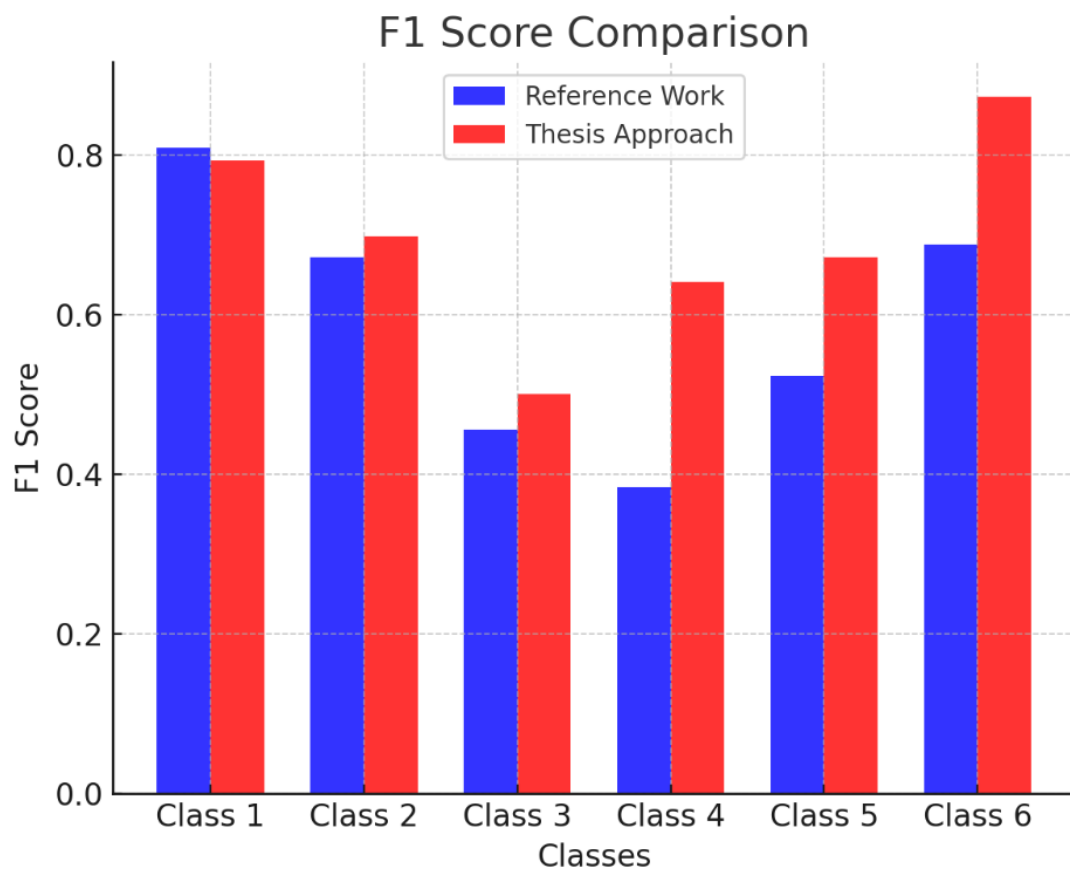
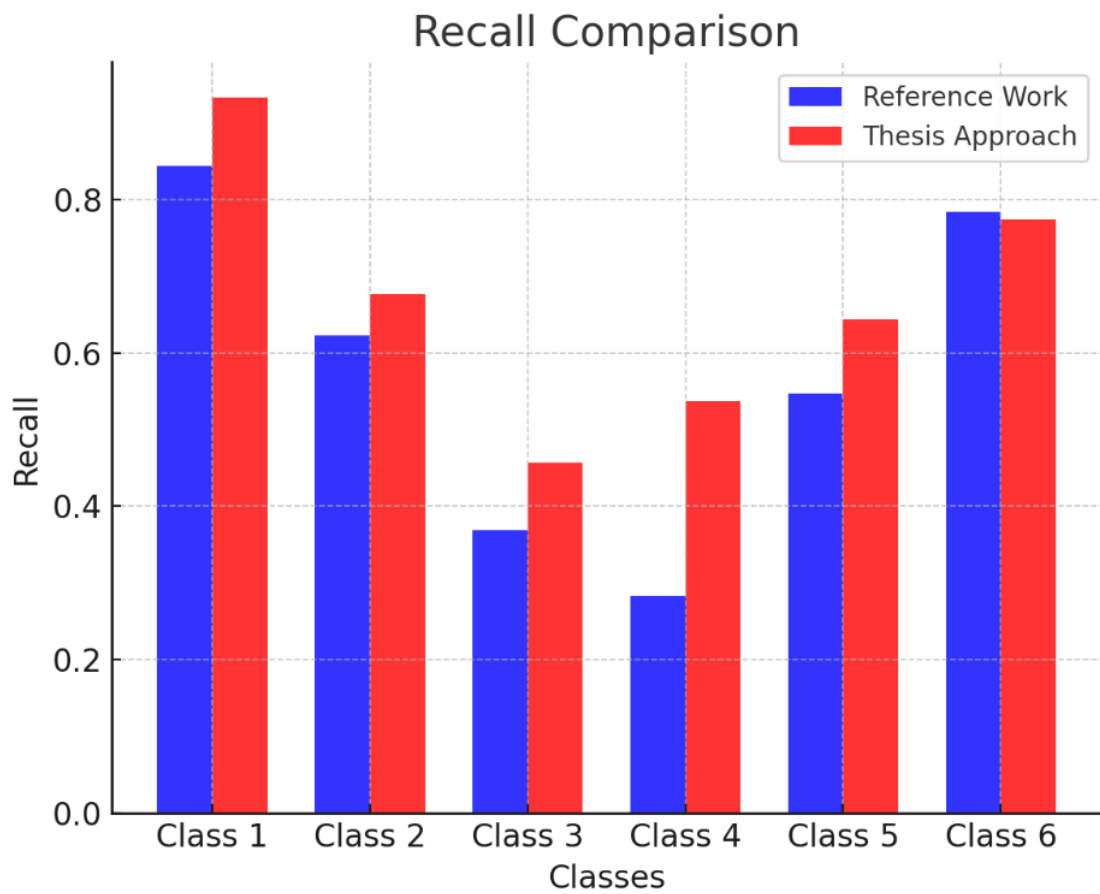
- The Quadratic Weighted Kappa is clearly higher in the thesis approach.
- For the reference work, the precision is generally lower except for the first class where it is quite high. This may indicate that the model is overfitting to the data it was trained on or affected by the class distribution imbalance.
- The thesis approach shows higher precision for most classes, suggesting better generalization, but the recall is lower in some classes, indicating some true positives are being missed.
- The F1 scores are mixed, with the thesis approach having generally higher scores for some classes but also the reference work having higher scores. The reference work has a notably lower F1 score in the fourth class, which suggests poor performance on that particular class when using the entire data set.

The use of the entire dataset for both training and validation can lead to overfitting, where the model performs well on its training data but may not generalize well to unseen data. This is evident in the higher recall rates for the reference work confusion matrix, suggesting it is better at detecting positive samples, but possibly at the expense of correctly identifying negative samples (lower precision).

In contrast, the thesis approach confusion matrix shows more balanced precision and recall, which could indicate a model that generalizes better to new, unseen data. However, this comes with a trade-off as the recall is generally lower, which means it misses more positive samples. It is important to consider that the absolute numbers of samples in each class affect these metrics. If one class has very few samples, the metrics for that class can be more volatile and less indicative of overall performance.

Figure 9,10,11: Comparative analysis of performance metrics - Reference Work vs. Thesis Approach across multiple classes





## 5.5 Conclusions

The current work proposes a robust CNN model tailored for the classification of prostate cancer histopathological images. Utilizing the PANDA challenge dataset as a foundational benchmark, this approach circumvents the constraints imposed by publicly available computational resources, negating the dependency on privately maintained high-performance computing systems. At the core of this methodology is a novel patch-based strategy coupled with a customized CNN architecture that underpins the model implementation. This framework introduces a weighting mechanism that evaluates the relative significance of each histological patch and its contribution to the characterization of the whole slide image. Preliminary outcomes suggest that this nuanced consideration of patch relevance could potentially improve the performance of patch-based methods for cancer histopathological images grading.

## 5.6 Future work or improvements

Future endeavors in refining the present research should prioritize the strategic enhancement of patch selection criteria, with an emphasis on increasing the volume of extracted patches. This progression calls for the allocation of more robust computational resources, which is anticipated to substantively elevate the accuracy of the results. Moreover, the integration of label masks into the training stage—by implementing a pre-training phase focused on individual patch predictions—may significantly sharpen the CNN's proficiency in identifying cancerous patterns within histological images. Additionally, broadening the training spectrum by incorporating images from auxiliary datasets stands to imbue the model with a richer, more generalizable understanding of varied pathological presentations. Lastly, meticulous attention to the denoising process promises to further refine the model's performance. These advancements collectively signify a concerted effort to improve the model's diagnostic acumen.



# References

- Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., . . . Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data*, 8, 1-74.
- Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S. S., Safwan, M., Alex, V., . . . Donovan, M. (2019). Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56, 122-139.
- Arvaniti, E., Fricker, K. S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., . . . Claassen, M. (2018). Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific Reports*, 8(1), 12054.
- Bulten, W., Kartasalo, K., Chen, P.-H. C., Ström, P., Pinckaers, H., Nagpal, K., . . . Vink, R. (2022). Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature medicine*, 28(1), 154-163.
- Bulten, W., Kartasalo, K., Chen, P.-H. C., Ström, P., Pinckaers, H., Nagpal, K., . . . the, P. c. c. (2022). Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature medicine*, 28(1), 154-163. doi:10.1038/s41591-021-01620-2
- Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., . . . Litjens, G. (2020). Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*, 21(2), 233-241.
- Bulten, W., Pinckaers, S., & Eklund, K. (2020). The PANDA challenge: Prostate cANcer graDe Assessment using the Gleason grading system. *MICCAI challenge*.
- Carriaga, M. T., & Henson, D. E. (1995). The histologic grading of cancer. *Cancer*, 75(S1), 406-421.
- Caruana, R. (1998). *Multitask learning*: Springer.
- del Toro, O. J., Atzori, M., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., . . . Müller, H. (2017). *Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score*. Paper presented at the Medical Imaging 2017: Digital Pathology.
- Egevad, L., Swanberg, D., Delahunt, B., Ström, P., Kartasalo, K., Olsson, H., . . . Humphrey, P. A. (2020). Identification of areas of grading difficulties in prostate cancer and comparison with artificial intelligence assisted grading. *Virchows Archiv*, 477, 777-786.
- Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R., & Humphrey, P. A. (2016). The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. *Am J Surg Pathol*, 40(2), 244-252. doi:10.1097/pas.0000000000000530
- Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R., Humphrey, P. A., & Committee, G. (2016). The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. *Am J Surg Pathol*, 40(2), 244-252.
- Gleason2019 data | biomedical imaging and artificial intelligence. (2019). Retrieved from <https://bmiai.ubc.ca/research/miccai-automatic-prostate-gleason-grading-challenge-2019/gleason2019-data>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*: MIT press.
- Gummesson, A., Arvidsson, I., Ohlsson, M., Overgaard, N. C., Krzyzanowska, A., Heyden, A., . . . Aström, K. (2017). *Automatic Gleason grading of H and E stained microscopic prostate images using deep convolutional neural networks*. Paper presented at the Medical Imaging 2017: Digital Pathology.

- Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., & Yener, B. (2009). Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2, 147-171.
- Hamilton, P. W., Bartels, P. H., Thompson, D., Anderson, N. H., Montironi, R., & Sloan, J. M. (1997). Automated location of dysplastic fields in colorectal histology using image texture analysis. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 182(1), 68-75.
- Hatipoglu, N., & Bilgin, G. (2014). *Classification of histopathological images using convolutional neural network*. Paper presented at the 2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.
- Hoang, V.-T., & Jo, K.-H. (2021). *Practical analysis on architecture of EfficientNet*. Paper presented at the 2021 14th International Conference on Human System Interaction (HSI).
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., . . . Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- Hsu, W.-Y., & Chou, C.-Y. (2015). Medical image enhancement using modified color histogram equalization. *Journal of Medical and Biological Engineering*, 35, 580-584.
- Irshad, H., Veillard, A., Roux, L., & Racoceanu, D. (2013). Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. *IEEE reviews in biomedical engineering*, 7, 97-114.
- Ismail, S. M., Colclough, A. B., Dinnen, J. S., Eakins, D., Evans, D., Gradwell, E., . . . Newcombe, R. G. (1989). Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia. *British Medical Journal*, 298(6675), 707-710.
- Kaggle.com. (2021). Let's enhance the images. Retrieved from <https://www.kaggle.com/debanga/let-s-enhance-the-images>
- Källén, H., Molin, J., Heyden, A., Lundström, C., & Åström, K. (2016). *Towards grading gleason score using generically trained deep convolutional neural networks*. Paper presented at the 2016 IEEE 13th International symposium on biomedical imaging (ISBI).
- Kaur, K., Jindal, N., & Singh, K. (2021). Fractional derivative based Unsharp masking approach for enhancement of digital images. *Multimedia Tools and Applications*, 80, 3645-3679.
- Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- Kothari, S., Phan, J. H., Stokes, T. H., & Wang, M. D. (2013). Pathology imaging informatics for quantitative analysis of whole-slide images. *Journal of the American Medical Informatics Association*, 20(6), 1099-1108. doi:10.1136/amiajnl-2012-001540
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84-90.
- Kumar, N., Verma, R., Arora, A., Kumar, A., Gupta, S., Sethi, A., & Gann, P. H. (2017). *Convolutional neural networks for prostate cancer recurrence prediction*. Paper presented at the Medical Imaging 2017: Digital Pathology.
- Kwak, J. T., & Hewitt, S. M. (2017). Nuclear architecture analysis of prostate cancer via convolutional neural networks. *Ieee Access*, 5, 18526-18533.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

- Li, W., Li, J., Sarma, K. V., Ho, K. C., Shen, S., Knudsen, B. S., . . . Arnold, C. W. (2018). Path R-CNN for prostate cancer diagnosis and gleason grading of histological images. *IEEE transactions on medical imaging*, 38(4), 945-954.
- Linkon, A. H. M., Labib, M. M., Hasan, T., Hossain, M., & Jannat, M.-E. (2021). Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: An extensive study. *Informatics in Medicine Unlocked*, 24, 100582. doi:<https://doi.org/10.1016/j.imu.2021.100582>
- Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., . . . van der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports*, 6(1), 26286. doi:10.1038/srep26286
- Lucas, M., Jansen, I., Savci-Heijink, C. D., Meijer, S. L., de Boer, O. J., van Leeuwen, T. G., . . . Marquering, H. A. (2019). Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. *Virchows Archiv*, 475, 77-83.
- Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., . . . Thomas, N. E. (2009). *A method for normalizing histology slides for quantitative analysis*. Paper presented at the 2009 IEEE international symposium on biomedical imaging: from nano to macro.
- Maini, R., & Aggarwal, H. (2010). A comprehensive review of image enhancement techniques. *arXiv preprint arXiv:1003.4053*.
- Misra, D. (2019). Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*.
- Musumeci, G. (2014). Past, present and future: overview on histology and histopathology. *J Histol Histopathol*, 1(5), 1-3.
- Nagpal, K., Foote, D., Liu, Y., Chen, P.-H. C., Wulczyn, E., Tan, F., . . . Wren, J. H. (2019). Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ digital medicine*, 2(1), 48.
- Naylor, P., Laé, M., Rey, F., & Walter, T. (2017). *Nuclei segmentation in histopathology images using deep neural networks*. Paper presented at the 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017).
- Oda, H., Roth, H. R., Chiba, K., Sokolić, J., Kitasaka, T., Oda, M., . . . Mori, K. (2018). *BESNet: boundary-enhanced segmentation of cells in histopathological images*. Paper presented at the Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11.
- Oskal, K. R., Risdal, M., Janssen, E. A., Undersrud, E. S., & Gulsrud, T. O. (2019). A U-net based approach to epidermal tissue segmentation in whole slide histopathological images. *SN Applied Sciences*, 1, 1-12.
- Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 1345-1359.
- Pang, B., Zhang, Y., Chen, Q., Gao, Z., Peng, Q., & You, X. (2010). *Cell nucleus segmentation in color histopathological imagery using convolutional networks*. Paper presented at the 2010 Chinese Conference on Pattern Recognition (CCPR).
- Panigrahi, S., & Swarnkar, T. (2020). Machine learning techniques used for the histopathological image analysis of oral cancer-a review. *The Open Bioinformatics Journal*, 13(1).
- Ren, J., Sadimin, E., Foran, D. J., & Qi, X. (2017). *Computer aided analysis of prostate histopathology images to support a refined Gleason grading system*. Paper presented at the Medical Imaging 2017: Image Processing.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-net: Convolutional networks for biomedical image segmentation*. Paper presented at the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th

- International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18.
- Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.
- Ruifrok, A. C., & Johnston, D. A. (2001). Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4), 291-299.
- Salvi, M., Acharya, U. R., Molinari, F., & Meiburger, K. M. (2021). The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. *Computers in Biology and Medicine*, 128, 104129.
- Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA Cancer J Clin*, 70(1), 7-30. doi:10.3322/caac.21590
- Silva-Rodríguez, J. (2020). SICAPv2 - Prostate Whole Slide Images with Gleason Grades Annotations (Publication no. doi: 10.17632/9xxm58dvs3.2). from Mendeley Data
- Silva-Rodríguez, J., Colomer, A., Sales, M. A., Molina, R., & Naranjo, V. (2020). Going deeper through the Gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer methods and programs in biomedicine*, 195, 105637.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Ström, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D. M., . . . Humphrey, P. A. (2020). Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology*, 21(2), 222-232.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). *Going deeper with convolutions*. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). *Rethinking the inception architecture for computer vision*. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.
- Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5), 1299-1312.
- Talo, M. (2019). Automated classification of histopathology images using transfer learning. *Artificial intelligence in medicine*, 101, 101743.
- Van der Laak, J., Litjens, G., & Ciompi, F. (2021). Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5), 775-784.
- Wang, J., Chen, R. J., Lu, M. Y., Baras, A., & Mahmood, F. (2020). *Weakly supervised prostate tma classification via graph convolutional networks*. Paper presented at the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI).
- Yonekura, A., Kawanaka, H., Prasath, V. S., Aronow, B. J., & Takase, H. (2017). *Improving the generalization of disease stage classification with deep CNN for glioma histopathological images*. Paper presented at the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).
- Zeng, Z., Xie, W., Zhang, Y., & Lu, Y. (2019). RIC-Unet: An improved neural network based on Unet for nuclei segmentation in histology images. *Ieee Access*, 7, 21420-21428.
- Zhu, Y., & Newsam, S. (2017). *Densenet for dense flow*. Paper presented at the 2017 IEEE international conference on image processing (ICIP).
- Zuley, M. L., Jarosz, R., Drake, B. F., Rancilio, D., Klim, A., Rieger-Christ, K., Lemmerman, J. (2016). The Cancer Genome Atlas Prostate Adenocarcinoma Collection (TCGA-PRAD) (Version 4) [Data set]. The Cancer Imaging Archive.

(Publication no. <https://doi.org/10.7937/K9/TCIA.2016.YXOGLM4Y>).  
<https://portal.gdc.cancer.gov/projects/TCGA-PRAD>

# Appendices

## Appendix A1

*Code Snippet 1: Pseudo-code of the get\_patches - the main function for parch extraction*

```
algorithm get_patches

  input: Image img with dimensions (height, width, channels),
         Patch size patch_size,
         Number of patches n_patches,
         Padding mode mode
  output: List of patches with their corresponding indices

  Initialize an empty list result for storing patches

  Calculate image dimensions: h, w, c from img

  Calculate necessary padding for height (pad_h) and width
  (pad_w) to make the image dimensions evenly divisible by
  patch_size, adjust based on mode

  Apply padding to img to get img_padded with constant values
  (255 for white)

  Reshape img_padded into a 5D array considering patch rows,
  patch columns, and patch dimensions

  Transpose and reshape the 5D array to a 4D array (number of
  patches, patch height, patch width, color channels)

  If the number of reshaped patches is less than n_patches, pad
  the array with additional white patches

  Sort patches based on the sum of their pixel values to
  prioritize by brightness or another criterion

  Select the top n_patches based on the sorting criterion

  For each selected patch, store the patch and its index in the
  result list

  return result
```

## Appendix A2

### *Code Snippet 2: Pseudo-code of the model development script*

```
algorithm ProstateCancerGradeAssessment

    input: Dataset with images and labels, patch_size,
batch_size, n_patches
    output: Trained model, predictions, evaluation metrics

    Initialize libraries and constants

    Load and preprocess dataset
        - Read labels and image paths
        - Split dataset into train, validation, and evaluation
sets

    Define data preparation functions
        - get_x: Returns path for image patches
        - get_y: Returns label for an image
        - open_images_eval and open_images: Load and preprocess
patches for evaluation and training
        - collate: Custom collation function for DataLoader
        - custom_splitter: Function to split data into training
and validation

    Setup DataLoaders for training, validation, and evaluation

    Define model architecture
        - Mish activation function
        - PatchAggregator: Aggregates features from image patches
        - DenseNetModel: Defines the neural network model using
DenseNet

    Define loss function and metrics
        - FocalLoss: Custom loss function
        - MulticlassRocAuc: Metric for ROC AUC score

    Initialize and configure Learner
        - Setup model, optimizer, loss function, and metrics
        - Apply mixed-precision training and gradient clipping

    Train the model
        - Use fit_one_cycle with early stopping and model
checkpoint callbacks

    Evaluate the model
        - Generate predictions on evaluation dataset
```

- Calculate `and` print evaluation metrics (Kappa Score, ROC AUC, Confusion Matrix)



## Appendix A3

*Code Snippet 3: CNN architecture - The custom layers of the model*

```
(aggregation): PatchAggregator(
  (attention): Sequential(
    (0): Linear(in_features=3840, out_features=128, bias=True)
    (1): Mish()
    (2): Linear(in_features=128, out_features=1, bias=True)
  )

  (head): Sequential(
    (0): Linear(in_features=3840, out_features=512, bias=True)
    (1): BatchNorm1d(512, eps=1e-05, momentum=0.1, affine=True,
track_running_stats=True)
    (2): Mish()
    (3): Dropout(p=0.5, inplace=False)
    (4): Linear(in_features=512, out_features=6, bias=True)
  )
)
```