

I.M.M.SALAMEH

BAU 2024

T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF EDUCATION

THE DEPARTMENT OF COMPUTER ENGINEERING

**PROGNOSIS PREDICTION IN ORAL SQUAMOUS CELL CARCINOMA
USING DEEP LEARNING**

MASTER'S THESIS

IBRAHIM SALAMEH

ISTANBUL 2024

T.C.
BAHCESEHIR UNIVERSITY
GRADUATE SCHOOL OF EDUCATION
THE DEPARTMENT OF COMPUTER ENGINEERING

**PROGNOSIS PREDICTION IN ORAL SQUAMOUS CELL CARCINOMA
USING DEEP LEARNING**

MASTER'S THESIS
IBRAHIM SALAMEH

THESIS ADVISOR
ASST. PROF. MUSTAFA UMIT ONER

ISTANBUL 2024



T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

...../...../.....

MASTER THESIS APPROVAL FORM

Program Name:	Big Data Analytics and Management (English - Thesis)
Student's Name and Surname:	Ibrahim Salameh
Name Of The Thesis:	PROGNOSIS PREDICTION IN ORAL SQUAMOUS CELL CARCINOMA USING DEEP LEARNING
Thesis Defense Date:	June 07, 2024

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

Assoc. Dr. Yücel Batu SALMAN

Institute Director

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

	Title/Name	Institution	Signature
Thesis Advisor's	Asst. Prof. Mustafa Umit Oner	Bahçeşehir University	
Member's	Asst. Prof. Ece Gelal SOYAK	Bahçeşehir University	
Member's	Prof. Tunca DOĞAN	Hacettepe University	

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname : Ibrahim Salameh

Signature :

ABSTRACT

PROGNOSIS PREDICTION IN ORAL SQUAMOUS CELL CARCINOMA USING DEEP LEARNING

Ibrahim Salameh

Master's Program in Big Data Analytics and Management.

Supervisor: Asst. Prof. Mustafa Umit Oner.

June 2024, 100 Pages

Prevalence of Oral Squamous Cell Carcinoma (OSCC) is increasing worldwide, and half of the cases have poor prognosis, i.e., having a recurrence in five years. Besides, there is a lack of objective and precise prognostic biomarkers, which is important for improving patient management. This study develops a Deep Learning (DL) system predicting prognosis in OSCC. The deep-learning system comprises segmentation, diagnosis, and prognosis models. The segmentation model achieves an accuracy of 0.991 on the test set, and it produces one segmentation map per OSCC slide. The diagnosis model accepts the segmentation data as an input and classifies the slides into tumor vs. normal. It achieves an area under the receiver operating characteristics curve value (AUROC) of 0.990 (95% confidence interval: 0.967 – 1.000). The prognosis model predicts poor prognosis vs. good prognosis, and it attained an AUROC value of 0.973 (95% confidence interval: 0.923-1.000). Moreover, an attribution map analysis is conducted. The analysis highlights the focus of the prognosis model in stroma regions.

Keywords: Deep Learning, Digital Pathology, Oral Squamous Cell Carcinoma, Semantic Segmentation, Multiple Instance Learning.

ÖZ

DERİN ÖĞRENME KULLANARAK ORAL SKUAMÖZ HÜCRELİ KARSİNOMDA PROGNOZ TAHMİNİ

Ibrahim Salameh

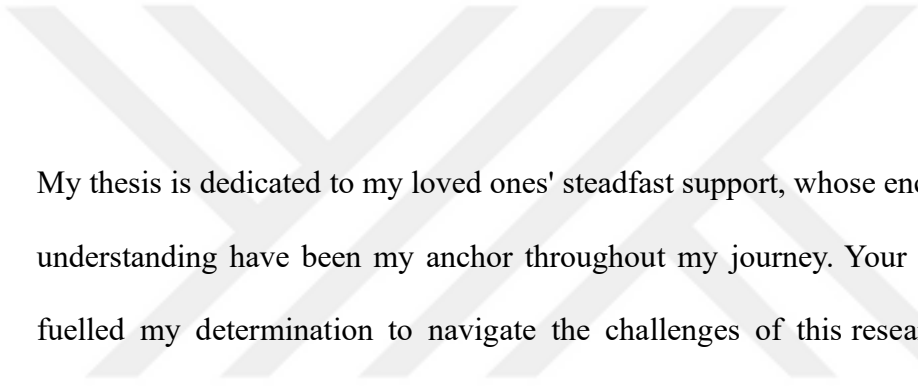
Büyük Veri Analizi ve Yönetimi Yüksek Lisans Programı

Tez Danışmanı: Dr. Öğrt. Üyesi Mustafa Ümit Öner

Haziran 2024, 100 Sayfa

Oral Skuamöz Hücreli Karsinomun (OSCC) prevalansı dünya çapında artmaktadır ve vakaların yarısında kötü prognoz vardır, yani beş yıl içinde nüksetme görülür. Ayrıca, hasta yönetimini iyileştirmek için önemli olan objektif ve kesin prognostik biyobelirteçlerin eksikliği vardır. Bu çalışma OSCC'de prognozu tahmin eden bir Derin Öğrenme (DL) sistemi geliştirmektedir. Derin öğrenme sistemi segmentasyon, teşhis ve prognoz modellerinden oluşur. Segmentasyon modeli, test setinde 0.991 doğruluk elde eder ve her OSCC slaydı için bir segmentasyon haritası üretir. Tanı modeli, segmentasyon verilerini girdi olarak alır ve slaytları tümör ve normal olarak sınıflandırır. Bu model 0.990 (%95 güven aralığı: 0.967 – 1.000) alıcı çalışma özellikleri eğrisinin altında kalan alan (AUROC) değerine ulaşmaktadır. Prognoz modeli, kötü prognozu ve iyi prognozu tahmin etmektedir ve 0.973 (%95 güven aralığı: 0.923-1.000) AUROC değerine ulaşmaktadır. Ayrıca atıf haritası analizi yapılmaktadır. Analiz, prognoz modelinin stroma bölgelerindeki odağını vurgulamaktadır.

Anahtar Kelimeler: Derin Öğrenme, Dijital Patoloji, Oral Skuamöz Hücreli Karsinom, Anlamsal Segmentasyon, Çoklu Örnek Öğrenme.



My thesis is dedicated to my loved ones' steadfast support, whose encouragement and understanding have been my anchor throughout my journey. Your belief in me has fuelled my determination to navigate the challenges of this research. Despite the hurdles, I am grateful for the opportunity to help advancement medical science. May this work serve as a tribute to the strength of endurance and triumph of the human spirit.

ACKNOWLEDGMENTS

I am writing to convey my deep appreciation to my supervisor, Mustafa Umit ONER, for his supervision, constructive perspectives, expertise in both practical and theoretical aspects, and vital contribution in successfully concluding this research. Furthermore, I acknowledge that he imparted a wealth of knowledge to me, encompassing personal and academic realms.

I want to extend my gratitude to our Hong Kong University (HKU) colleagues, particularly my colleague Chui Shan Chu, who studied and manually analyzed the biomarkers our research relies on and provided us with ground-truth data, the backbone behind this research's success.

I would also like to thank my colleague Sepideh KANI for her well-being during this journey and our insightful brainstorming sessions, which greatly enhanced our work.

Furthermore, I would like to express my heartfelt thanks to my family, who have always believed in me: my father, Mustafa SALAMEH, and my mother, Nuha SALAMEH, and my siblings, and to whom I owe continuous backing in every facet of every endeavor I have undertaken.

TABLE OF CONTENT

ETHICAL CONDUCT	iii
ABSTRACT	iv
ÖZ	v
DEDICATION	vi
ACKNOWLEDGMENTS	vii
LIST OF TABLES.....	xii
LIST OF FIGURES.....	xiv
Chapter 1	1
Introduction	1
1.1 Theoretical Framework	1
1.2 Statement of the Problem	2
1.3 Purpose of the Study.....	2
1.4 Significance of the Study	2
Chapter 2	3
Literature Review	3
2.1 Cancer Definition, Characteristics and Diagnosis.....	3
2.1.1 Definition of cancer	3
2.1.2 Hallmarks of cancer	3
2.1.3 Cancer diagnosis	5
2.2 Digital Histopathology	5
2.2.1 Whole slide imaging	6
2.2.1.1 Slide preparation	6
2.2.1.2 Slide scanning.....	6
2.2.1.3 Whole slide images.....	7
2.2.2 WSI viewers and annotation software	8
2.3 Artificial Intelligence.....	8

2.3.1 Machine learning hierarchy	8
2.3.2 Deep learning	9
2.3.2.1 Multilayer perceptron	10
2.3.2.2 Convolutional neural networks	11
2.3.3 Multiple instance learning	13
2.3.4 Vision transformers	14
2.3.5 Learning approaches	15
2.3.5.1 From scratch	15
2.3.5.2 Transfer learning	15
2.3.6 Explainable AI	16
2.4 Digital Histopathology Automation and Dataset Preparation	16
2.4.1 Digital histopathology automation	16
2.4.2 Dataset preparation	16
2.4.2.1 Slide annotating and patches cropping	17
2.4.2.2 Stain normalization	17
2.4.2.3 Data segregation	18
2.4.2.4 Data augmentation	18
2.5 Deep Learning Applications in Cancer Prognosis	19
Chapter 3	21
Methodology	21
3.1 Research Design	21
3.1.1 Prognosis prediction in OSCC	21
3.1.2 Effects of data segregation	21
3.2 Datasets (DSs)	21
3.2.1 HKU OSCC dataset	21
3.2.1.1 Ethics statement	22
3.2.1.2 Annotation mask creation	22
3.2.1.3 Patches cropping	23
3.2.1.4 Data segregation	25
3.2.1.5 Stain normalization	26
3.2.1.6 Balancing classes	26
3.2.2 Public CRC dataset	27

3.3 Prognosis Prediction.....	28
3.3.1 Segmentation model	28
3.3.1.1 Dataset preparation	28
3.3.1.2 Deep learning models	30
3.3.1.3 OTSU thresholding	31
3.3.1.4 WSI segmentation	32
3.3.1.5 WSI annotations file.....	32
3.3.2 Tumor vs. non-tumor WSI classification	33
3.3.2.1 Dataset preparation	33
3.3.2.2 Deep learning model	34
3.3.3 Good vs. poor prognosis prediction.....	34
3.3.3.1 Dataset preparation	34
3.3.3.2 Deep learning model	34
3.4 Effects of Data Segregation on Model's Performance	35
3.4.1 Dataset preparation	35
3.4.2 Deep learning models	37
3.5 Model and Data Availability.....	38
Chapter 4	39
Results.....	39
4.1 DL Model Segments Out OSCC Slides Successfully	39
4.1.1 DL models performance and segmentation model selection	39
4.1.2 Effects of stain normalization in the performance of segmentation models	41
4.1.2.1 Resnet50 segmentation model selection	41
4.1.2.2 ViT segmentation model selection.....	42
4.1.3 WSI segmentation map creation	44
4.1.4 WSI JSON annotations	45
4.2 DL Model Discriminates Tumor vs Non-Tumor	46
4.3 OSCC Prognosis Can Be Predicted from Brightfield and Confocal Images ...	46
4.4 Patient Level Segregation is Crucial in Deep Learning Model Development .	47
Chapter 5	51

Discussion	51
5.1 Stain Normalization Could Benefit WSIs Obtained by Different Scanners.....	51
5.2 CNNs and ViT Models for Tissue Microenvironment Analysis.....	51
5.3 CNN is Surpassing ViT in its Ability to Recognize Mixed Tissues	51
5.4 Segmentation Model’s Probabilities Provide Accurate Diagnosis Results	53
5.5 ECM Fibril Waveforms as Biomarkers to Predict OSCC Prognosis	54
5.6 Patient Level Segregation for Trustworthy DL Models’	55
5.7 Limitations.....	55
5.8 Future Work.....	56
5.9 Conclusion.....	56
REFERENCES.....	58

LIST OF TABLES

TABLES

Table 1 A Summary of the WSIs Employed at Different Phases of the Research.	22
Table 2 The Dataset's Components Aside with its Origin, the Number of Slides Used for Cropping, and the Overall Count of Cropped Patches Across All Classes.	27
Table 3 Patches Distribution Among the Classes for the Three DSs.	28
Table 4 Patches Distribution Between the Classes for the Three segregations, Where Different Patient has been Randomly Reserved for Each Split.	29
Table 5 Number of Parameters and Size in MB for DL Models Used on Medical Images Classification Tasks.	37
Table 6 The Models' Accuracies on Training, Validation and Testing DSs, for the Three Data Segregations Using the Stain-Normalized (SN) and the Non-Stain-Normalized (NSN) Histopathological Images.	40
Table 7 McNemar's Test Results Between Resnet50 Models Trained on The Same DS With Stain Normalization Applied (SN) and Without (NSN).	42
Table 8 McNemar's Test Results between ViT Models Trained on the Same Data with Stain Normalization Applied (SN) and Without (NSN).	42
Table 9 McNemar's Test Results Between Resnet50 and ViT Models trained on SEG2 with No Stain Normalization Applied (NSN).	43
Table 10 Resnet50 SEG2 Non-Stain Normalized (NSN) Model's Classification Report, Providing Key Metrics Including Precision, Recall, F1-score, and Support for Each Class in the Test Split.	43
Table 11 ViT SEG2 Non-Stain Normalized (NSN) Model's Classification Report, Providing Key Metrics Including Precision, Recall, F1-score, and Support for Each Class in the Test Split.	44

Table 12 The Classification Report for LeNet5 Diagnosis Model's Performance, Providing Key Metrics Including Precision, Recall, F1-score, and Support for Each Class in the Test Split.	46
Table 13 The Classification Report for the Three MIL Prognosis models' performances, Providing Key Metrics Including Precision, Recall, F1-score, and Support for Each Class in the Test Split.....	47
Table 14 McNemar's Test Results for the MIL Models Trained Using the Confocal (CF) and the Brightfield (BF) Images Solely and Combined.	47
Table 15 The Testing DS Accuracy Results of the Different Deep Learning (DL) Models with Different Data Segregations, Show the Impact of Patient-Level Segregation on the Model's Performance on Unseen Data.....	49

LIST OF FIGURES

FIGURES

- Figure 1.* The traditional and potentially new "Hallmarks of Cancer," are shown this figure. The hallmarks of cancer graphic have been adapted from (Hanahan & Weinberg, Hallmarks of cancer: the next generation, 2011). 4
- Figure 2.* On the left is a brightfield (BF) image obtained by scanning the glass-slide using a traditional light microscope, while confocal (CF) version of the image on the right, is the same glass-slide scanned by laser scanning confocal microscope (LSCM). 6
- Figure 3.* The pyramid structure of the WSI, each layer has a specific magnification level, representing a unique image with different dimensions and resolutions. The basic level is zero, which has the highest resolution. Each consecutive layer of the pyramid undergoes digital down sampling to decrease the size, resulting in lower resolution. The final layer is the image with the most down sampled dimensions and lowest resolution, located on the top of the pyramid. Image from (McClintock, Abel, & Cornish, 2022). 7
- Figure 4.* The Hierarchy of machine learning tasks. The tasks that fall in between are categorised as weakly supervised learning. Weakly supervised learning encompasses three forms of supervision: incomplete, inexact, and inaccurate. Multiple instance learning is a distinct kind of weakly supervised learning models and a prototypical illustration of weak supervision. Image from (ONER, 2021). 9
- Figure 5.* (A) The components of a single neuron include the input data, the weights, and the bias. The input values are multiplied with the weights and summed to a bias within the neuron, and the resulting value is sent to the activation layer to introduce non-linearity. The output of this layer becomes the input for the next layer. (B) A fully connected Artificial Neural Networks (ANNs) that forms a Multilayer Perceptron. 11
- Figure 6.* A (convolution layer) is a process where a kernel moves over an image of shape $[(W+2P), (H+2P), C]$ with a stride value of S , scanning from left to right and down. A single value in the output represents each group of values in the input and the kernel. Like the convolutional layer, the (Pooling Layer) applies a kernel that traverses across the array and performs a different function than in convolution; instead of calculating the sum, it applies a selected pooling function to the values within the

kernel and provides a single output value. The (Flattened Input) is the final intended output obtained after numerous iterations of convolution and pooling, depending on the level of complexity in the image features. This flattened input will be the input for the next FCL.....	13
<i>Figure 7.</i> A visual explanation showing the difference between Fully Supervised Learning and Weakly Supervised Learning in the way of handling tasks. In single-instance learning, each image is considered a unique data point with its label. However, in MIL, a group of images are assigned a single label. In this particular example, the label for a bag is set to 'negative' if at least one instance inside the group is the digit 0. Otherwise, the assigned bag label is 'positive'. Image from (ONER, 2021).	14
<i>Figure 8.</i> In the ViT model, an image is processed by dividing it into smaller patches. The attention mechanism in Transformers enables the model to selectively focus on specific tiles in an image that directly impacts the provision of a given label Image. Image from (Dosovitskiy, et al., 2020).	15
<i>Figure 9.</i> Following are two different histology slides of melanomas, each stained with hematoxylin and eosin, but with noticeably distinct appearances. The photos were acquired by the process of scanning the slides at a magnification of 20X using an Aperio Scans. Image from (Macenko, et al., 2009).	18
<i>Figure 10.</i> The annotations of a WSI as they appear after being loaded into QuPath, each color regions represent a specific tissue type.....	23
<i>Figure 11.</i> Cropping a patch over an annotations mask involves a sliding window that moves from left to right and down over the annotations mask. The displayed color percentage in each loop determines whether the patch will be considered or neglected. If the color percentage presented in the sliding window is greater than 95%, the patch will be extracted from the Brightfield WSI and saved alongside its label, which is obtained from a given color and labels key-value pairs.....	24
<i>Figure 12.</i> Eight samples for each class images, show the distinctions among the patches of each class.	25
<i>Figure 13.</i> The distribution of the classes among the splits of the first segregation slides, shows that the MUS class possesses the highest number of patches, whereas the LYM class exhibits the lowest number.	26
<i>Figure 14.</i> Under-sampling and Over-sampling methods for class imbalance issue handling.....	27

<i>Figure 15.</i> Two distinct DSs comprise three segregations with different patient groups within their splits. Each segregation was utilized to train, validate and test a two DL model. Segregations among the DSs are identical. (A) non-stain normalized, (B) stain-normalized.....	30
<i>Figure 16.</i> The picture on the left is a JPEG image extracted from the WSI at level 6, while the image on the right is the OTSU threshold mask results for the same image.	32
<i>Figure 17.</i> (A) The first stage involves creating a segmentation map using the best segmentation model. (B) The segmentation maps and probability files are then used to construct a 9D array for training a LeNet5 model on the Tumor Non-Tumor diagnosis task.	33
<i>Figure 18.</i> MIL model components, initially, we extracted and processed ROIs from BF and CF slides. Subsequently, we constructed bags of transformed images and used a fine-tuned Resnet50 model to extract the features of each instance within the bag. These extracted features are further processed by a linear layer which learns the mapping between the instances and determines the slide label whether good or poor prognosis.	35
<i>Figure 19.</i> The public dataset consists of three sources: the TCGA Data, copied three times and, for each replication, we segregated the data in distinct ways. The remaining China and Germany DSs were used to generate the external test dataset.....	36
<i>Figure 20.</i> A) Patient level segregation where each patient data only belongs to one split. B) Patch level segregation with partial intersection shows 44 patients shared among the splits, 17 only for training, and 12 for each validation and testing.	37
<i>Figure 21.</i> Initially, McNemar's test was used to compare the Resnet50 models trained on identical data segregation. Similarly, we conducted the same test for the ViT models. Finally, the same test was conducted between the models selected in the first round and one of the models was chosen to serve as the segmentation model for the subsequent phases.	41
<i>Figure 22.</i> A segmented WSI with colors showing different areas in the tissue generated by Resnet50 segmentation model trained on SEG2 non-stain normalized (NSN) DS.....	45
<i>Figure 23.</i> JSON annotations file obtained from the segmentation map and imported to QuPath.....	45

Figure 24. The accuracy results on the test sets for the different DL models trained using different DS segregations. FS: is From Scratch models and TL: refers to Transfer learning..... 48

Figure 25. A) The image is labeled as STR. The Resnet50 model correctly predicted it as STR, while the ViT model predicted it as NOR. B) Resnet50 predicted the MUS patch as MUS, but ViT provided a STR prediction. C) Resnet50 correctly identified the patch as LYM, while ViT attention was targeted on the TUM and ignored the majority area. 53

Figure 26. The original images and their attribution mask of the absolute contribution value show that the model tended to make predictions based on the fibril waveforms in the mixed tissue images. 54



LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
AI	Artificial Intelligence
BF	Brightfield
CAD	Computer-Aided Diagnosis
CF	Confocal
CNN	Convolutional Neural Network
CRC	Colorectal Cancer
DL	Deep Learning
DNN	Deep Neural Network
ECM	Extracellular Matrix
FCL	Fully Connected Layer
H&E	Hematoxylin and Eosin
LSCM	Laser Scanning Confocal Microscopes
ML	Machine Learning
MLP	Multi-Layer Perceptron
MIL	Multiple Instance Learning
NSN	Non-Stain Normalized
OSCC	Oral Squamous Cell Carcinoma
ReLU	Rectified Linear Unit
ROI	Region of Interest
SGD	Stochastic Gradient Descent
SN	Stain Normalized
WSI	Whole Slide Image

Chapter 1

Introduction

1.1 Theoretical Framework

Oral Squamous Cell Carcinoma (OSCC) is among the Head and Neck cancers, and it is turning into a serious problem in different countries. One in six deaths around the world result from Head and Neck cancers, with around 630.000 newly diagnosed patients, more than 50% of them dying yearly, and more than 90% of those cases are Squamous Cell Carcinoma (Vigneswaran & Williams, 2014).

The National Cancer Institute (NCI) defines cancer as a disease characterized by the uncontrolled growth and multiplication of abnormal cells, leading to the formation of tumours (NCI, 2024). It can start in various areas of the body, leading to different types of cancer. It can also spread to other body areas through the lymph and blood systems: this is called metastasis of cancer. Cancers are one of the most common causes of death worldwide causing around 10 million deaths in 2019 and it was ranked as the second-leading non-communicable disease worldwide in terms of deaths (Roser & Ritchie, 2024).

Early diagnosis is critical in lowering the cancer mortality rate. Pathology is the study of analyzing the slides of specimens for any abnormal cell growth. It is the primary method of accurate cancer diagnosis. As a subfield of pathology, digital pathology focuses on data management and, more specifically, the digitization of specimen slides as Whole Slide Images (WSIs). As WSIs emerged, pathologists can now store, view, interact with, and share specimens on digital screens, allowing laboratories worldwide to collaborate on cases. Moreover, WSIs enable the processing of digital slides using promising artificial intelligence models, which can help in the early diagnosis of cancer.

Recently, Artificial Intelligence (AI) models, namely Deep Learning (DL) models, have shown significant improvements in various digital pathology tasks including cancer diagnosis, such as prostate cancer diagnosis (Li, et al., 2019), and lymph node metastasis in OSCC using DL (Tang, et al., 2022) and achieved promising results. This thesis offers a DL system composed of different DL techniques to examine an

OSCC WSI. It offers multiple clinical outcomes, including slide segmentations, tumor vs. non-tumor diagnosis, and good vs. poor prognosis predictions.

1.2 Statement of the Problem

The prognosis of OSCC has been assessed based on cancer stage, tumor thickness and surgical margins. Various histological parameters have been examined in Hematoxylin and Eosin (H&E) stained tissue samples, including cervical lymph node metastasis (Eskander, et al., 2024) (Leusink, et al., 2012), extra-nodal extension (Zhu, et al., 2023) (Chai, et al., 2013), lymph-vascular invasion (Michikawa, et al., 2012) (Mascitti, et al., 2022), perineural invasion (Albergotti, et al., 2017) (Johnson, et al., 2020) and bone invasion (Wong, Keel, Glynn, & Varvares, 2000) (Fives, et al., 2017). Despite their use, they tend to be subjective, reliant on the operator and contentious, making it unlikely for them to provide accurate information about the condition of the disease. This highlights the urgent need for more objective and reliable prognostic assessment approaches.

1.3 Purpose of the Study

We conducted this study simultaneously with a study being conducted by our collaborators from Hong Kong University (HKU), who studied the Extracellular Matrix (ECM) fibril waveforms for prognosis prediction purposes in OSCC. The manual analysis of the ECM fibril waveforms has demonstrated their potential as biomarkers for the prognosis of OSCC. To automate the prognosis prediction process, we created a DL application that utilizes the same data with the manual analysis. This aims to accelerate the prognosis prediction process, reduce efforts, and improve the patient outcomes.

1.4 Significance of the Study

We have developed new algorithms to accurately predict the prognosis of OSCC, employing various DL models and methodologies. As far as we know, this work is the first to create a DL system processing OSCC confocal images to predict good vs. poor prognosis. The model has shown promising performance on predicting good vs. poor prognosis in OSCC patient cohort collected at Queen Mary Hospital, Hong Kong.

Chapter 2

Literature Review

2.1 Cancer Definition, Characteristics and Diagnosis

2.1.1 Definition of cancer

The human body comprises around 30 trillion cells that exist in a complex and mutually dependent manner, regulating each other's proliferation and maturation. Normal cells undergo reproduction only in response to signals from adjacent cells. Ongoing collaboration ensures that each tissue maintains an appropriate size and structure to meet the body's needs. Unlike normal cells, cancer cells deviate from this pattern due to genetic mutations that transform a normal cell into a cancerous cell. These mutations can be inherited, accumulate over time due to ageing and genetic degradation, or occur due to exposure to substances that cause damage to our genes, such as cigarette smoke (Vineis, et al., 2004), alcohol (Seitz, Pelucchi, Bagnardi, & La Vecchia, 2012), or ultraviolet radiation from the sun (Narayanan, Saladi, & Fox, 2010). They develop a reduced sensitivity to the typical mechanisms that regulate cell division and, instead, follow their internal blueprint for reproduction. Moreover, they possess a devastating attribute: the ability to move beyond their initial site, invading neighboring tissues and forming clusters in distant body regions. Malignant tumors gradually develop in aggressiveness and become lethal when they invade vital tissues and organs crucial for the survival of the entire body (Weinberg, 1996), (Phillips & Currow, 2010).

2.1.2 Hallmarks of cancer

Internal blueprints of cancer are the biological capabilities acquired by cancer cells to provide them survival advantages, which are also known as the hallmarks of cancer; (1) Supporting proliferative signaling, this hallmark is the ability of cancerous cells to provide the basic needs of life: power, metabolism, and space. (2) Avoiding growth suppressors: Cancer cells can avoid growth suppressors by negatively regulating cell proliferation. (3) Resisting cell death: Apoptotic cells automatically kill cells in their senescence stage, but cancer cells can work in different ways to limit or avoid apoptosis. (4) In cancer cells, growth and division are unlimited because they are able to pass the natural barriers of stopping this cycle; are senescence or cell death. (5) Cancerous cells generate tumor-associated neo-vasculature, to survive and induce

angiogenesis and metabolism. (6) E-Cadherin loss in cancerous cells enables them to formulate abnormal cell-cell contact, resulting in metastatic cells invade neighboring cells and results in tumors (Nives, 2003). (7) Cancerous cells adjust their energy metabolism, resulting in different growth needs than the normal cells. (8) The immune system should always be alert for new abnormal cells or diseases, but solid tumors could avoid detection (Hanahan & Weinberg, Hallmarks of cancer: the next generation, 2011); In order to get a deeper understanding of the cancer hallmarks, the researcher has chosen to concentrate on potential improvements to the conceptual model; he presents four more potential novel distinctive and facilitating attributes, which might eventually be included as fundamental elements of the conception of cancer hallmarks. The new parameters are (9) unlocking phenotypic plasticity, (10) non-mutational epigenetic reprogramming, (11) polymorphic microbiomes, and (12) senescent cells (Hanahan, Hallmarks of cancer: new dimensions., 2022), Figure 1 showing the hallmarks.

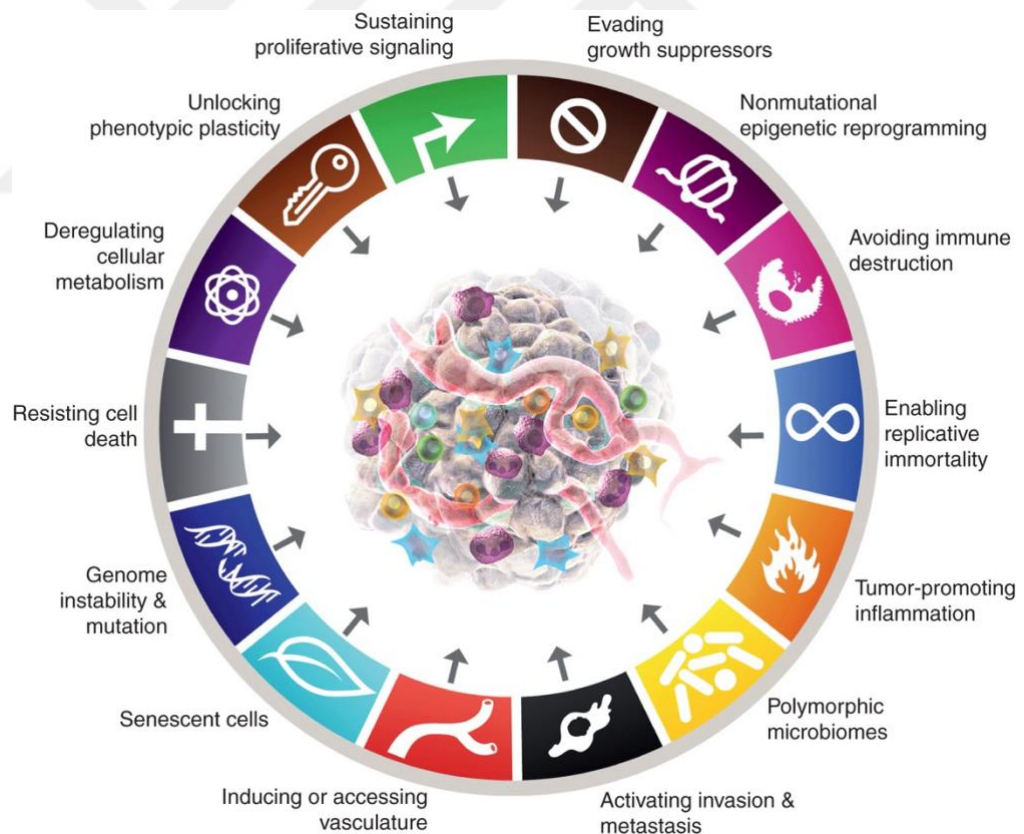


Figure 1. The traditional and potentially new "Hallmarks of Cancer," are shown this figure. The hallmarks of cancer graphic have been adapted from (Hanahan & Weinberg, Hallmarks of cancer: the next generation, 2011).

The above advantages of cancer make it a diverse group of diseases with various causes. Still, they all have one common feature: uncontrolled metastatic cells, which contribute to the complexity and time-consuming nature of diagnosing and treating the disease.

2.1.3 Cancer diagnosis

The cancer diagnosis process requires consultations with a doctor regarding appropriate screening procedures. These may include physical examinations, laboratory evaluations, imaging tests such as scans, and biopsies for a conclusive cancer diagnosis (Koning, et al., 2020) (Lester, et al., 2009). Tissue biopsy and histological analysis are the most accurate ways to diagnose Oral Squamous Cell Carcinoma OSCC (Carreras-Torras & Gay-Escoda, 2015) and other types of cancer due to their typical feature complexity. This method is used to detect anomalies and verify the existence of cancer by closely examining tissue specimens under a microscope for any abnormalities in cellular structure. This examination enables a conclusive diagnosis, which makes it essential to possess sufficient knowledge about the tissue specimen and allows pathologists to make decisions based on the tissue's morphology.

2.2 Digital Histopathology

Since the advent of microscope illumination for the examination of tissues in the early 1800s, the field of pathology has undergone substantial changes (Van, J. , & Taylor, 2010). The focus has evolved to prioritize tissues over organs, after which, to cells, gradually exploring more minor scales. This evolution has ultimately given rise to the discipline of histopathology; the development of histopathology as a discipline of study can be traced back to the combination of three Greek words: 'Histos' meaning tissue, 'pathos' meaning disease, and 'logos' meaning study (Mohan, 2010). Furthermore, the advent of slide scanners in late 1994, besides the improvements in software development in 1996, led to the emergence of virtual microscopy, which is the ability of pathologists to examine biopsy slides in a virtual environment as they are doing under a microscope; this was applicable through scanning the slide and producing the Whole Slide Image (WSI) that could be examined on a computer screen.

WSIs signified a notable transformation in histopathology and introduced a novel concept known as Digital Histopathology, which, in its basic definition, is the application of digital imaging technology in pathology (Montalto, 2016).

2.2.1 Whole slide imaging

2.2.1.1 Slide preparation

After extracting the biopsy from the patient, medical professionals perform a series of procedures to enhance the visibility of various regions within the tissues. Multiple techniques are available to create an examinable slide, with Hematoxylin and Eosin (H&E) being the most common. H&E staining refers to utilizing hematoxylin dye to impart a blue-purple color to cell nuclei. At the same time, eosin dye imparts a pink or red color to the cytoplasm ECM structures. This staining technique facilitates the recognition of distinct tissue constituents and cellular formations.

2.2.1.2 Slide scanning

After preparing the slide, or the “glass-slide” as known, it can be scanned using different scanners, which offer different morphology appearances as WSI. Traditional light microscope scanners which produce Brightfield (BF) images, and Laser Scanning Confocal Microscope (LSCM) scanners which produce Confocal (CF) images. The LSCM offers a significant advantage in optical sectioning, enabling the creation of a three-dimensional (3D) representation of a sample through high-resolution picture stacks offering detailed information. Figure 2 shows an example of the same slide scanned by each of the scanners.

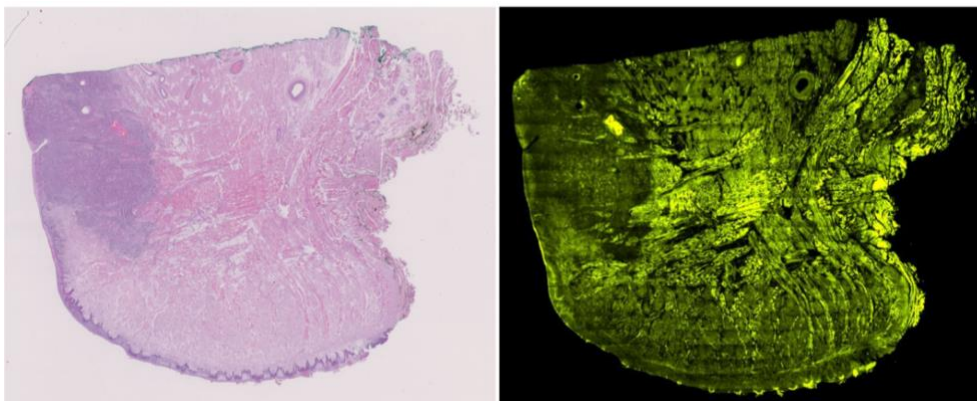


Figure 2. On the left is a brightfield (BF) image obtained by scanning the glass-slide using a traditional light microscope, while confocal (CF) version of the image on the right, is the same glass-slide scanned by laser scanning confocal microscope (LSCM).

2.2.1.3 Whole slide images

WSIs allow pathologists to have the same experience as a real microscope but on a computer screen. Pathologists can examine digital slides by navigating through several magnification levels, each with a particular resolution rate (Liu & Pantanowitz, 2019). The highest obtainable resolution is usually around 0.25 microns per pixel, which may be reached at the highest magnification level of 40X. Figure 3 illustrates the layered structure of the WSI, with the most significant and highest resolution layer at the base and subsequent layers displaying gradually reduced picture dimensions and resolution.

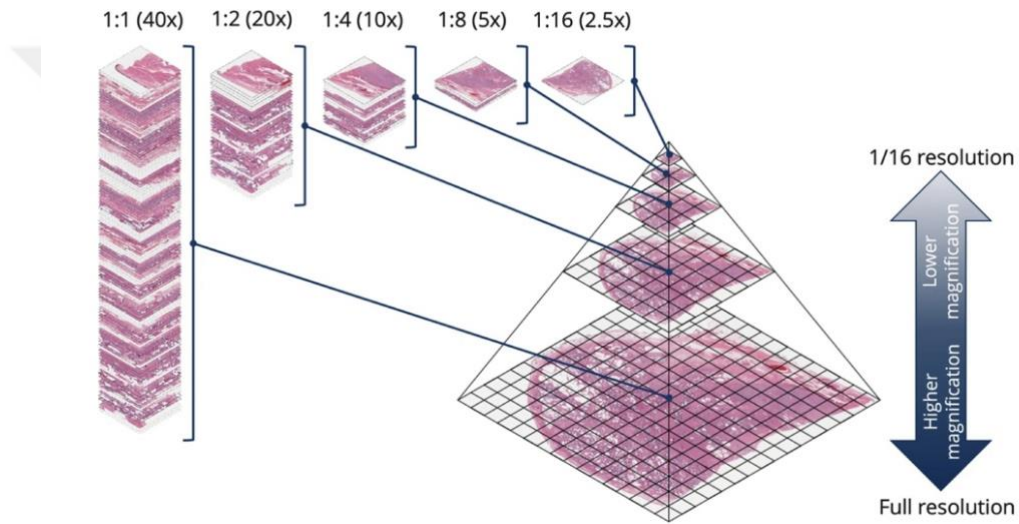


Figure 3. The pyramid structure of the WSI, each layer has a specific magnification level, representing a unique image with different dimensions and resolutions. The basic level is zero, which has the highest resolution. Each consecutive layer of the pyramid undergoes digital down sampling to decrease the size, resulting in lower resolution. The final layer is the image with the most down sampled dimensions and lowest resolution, located on the top of the pyramid. Image from (McClintock, Abel, & Cornish, 2022).

Pathology labs in the US and Europe have recently been using WSIs for primary diagnosis in surgical pathology. The increasing use of digital histopathology has significant potential to reduce pathologists' workloads, speed up the diagnosis process and enhance clinical processes via the computerized examination of tissue morphology. It has become possible through the remarkable progress in computer vision technologies, which now efficiently can assess cancer biomarkers and

histological features of tissue sample within a WSI, ensuring high accuracy and reliability. (Hamilton, et al., 2014), (Harder, et al., 2019) (Nearchou, et al., 2019), (Dimitriou, Arandjelović, Harrison, & Caie, 2018).

2.2.2 WSI viewers and annotation software

The advent of the WSI has necessitated the development of software that can efficiently handle and analyze these high-resolution digital images. Several apps were created to meet these specifications. The earliest virtual microscope was developed by Johns Hopkins from the University of Maryland on the period between 1996-1998 (Pantanowitz, et al., 2018), Another such program is QuPath (Bankhead, et al., 2017), developed by a team from the University of Edinburgh and designed to examine digital pathology images. It provides a user-friendly interface for pathologists and researchers to perform various functions, such as annotating and labeling Regions of Interest (ROIs) within the slides. In addition, it provides scripting capabilities in programming languages such as Groovy to enhance functions and streamline activities via automation.

2.3 Artificial Intelligence

Artificial Intelligence (AI) encompasses a range of technologies that aim to replicate human intelligence in acquiring knowledge, thinking logically, and comprehending tasks to initiate actions. Machine Learning (ML), a subdivision of AI, employs mathematical algorithms to identify patterns in data and generate predictions. On the other hand, Deep Learning (DL) takes inspiration from the brain's structure and utilizes multi-layered neural networks to make predictions. AI solutions are widely utilized in the healthcare industry to improve patient risk stratification, clinical care pathways, and diagnosis (Ahmad, Eckert, & Teredesai, 2018).

2.3.1 Machine learning hierarchy

In Machine Learning (ML) tasks, there are two primary methods of learning: supervised learning and unsupervised learning. Unsupervised learning addresses discovering relationships within the data using mathematical concepts, such as the K-Means clustering algorithm. Clustering involves grouping similar objects into one cluster (Yadav & Sharma, 2013), this method does not require human guidance. On the other hand, supervised learning aims to learn the relationship between the data and

their labels known as “ground-truth”, provided by experts, and used as baselines while training, validating and testing the model. Image classification is one example of supervised learning tasks.

Another method that falls between the previously mentioned methods is weakly supervised learning. In this method, the dataset includes weak supervision; depending on the level of supervision, we can categorize the weak supervision into three sub-types: incomplete, inexact, and inaccurate supervision. In incomplete supervision, accurate labeling applies to only a small subset of the data. In inexact supervision, the assignment of labels is limited to coarse-grained categories. In inaccurate supervision, the provided labels are noisy and cannot be considered as ground truth (Zhou, 2018), (ONER, 2021). Figure 4 shows the hierarchy of the machine learning tasks.

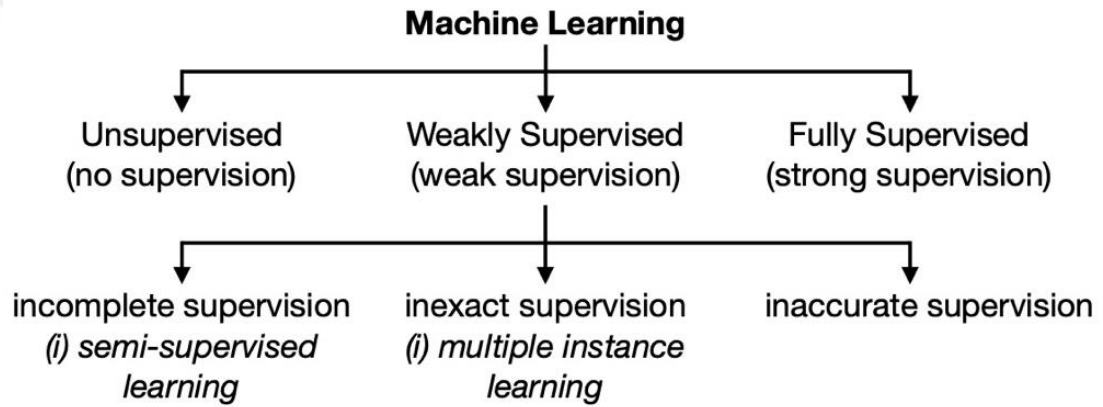


Figure 4. The Hierarchy of machine learning tasks. The tasks that fall in between are categorised as weakly supervised learning. Weakly supervised learning encompasses three forms of supervision: incomplete, inexact, and inaccurate. Multiple instance learning is a distinct kind of weakly supervised learning models and a prototypical illustration of weak supervision. Image from (ONER, 2021).

2.3.2 Deep learning

The human brain is an intricate network of neurons that allows humans to gain information and make decisions based on the gained knowledge. In recent decades, computer vision tasks have significantly improved by mimicking the human’s learning process mathematically. The new technique relies on a set of hyperparameters, some of them are to control the learning environment, others are trainable parameters called weights and biases. The training process involves adjusting the weight and bias values relying on a loss function, which calculates the error between the predicted and the ground-truth values. This process usually needs to be repeated multiple times, known

as epochs, to allow the model to discover the values that guarantee the global minimum point, where the loss value is minimized globally in the parameter space.

In some specific tasks, computers have surpassed humans in pattern recognition due to their ability to assess complicated multidimensional data with higher precision (Kühl, Goutier, Baier, Wolff, & Martin, 2022).

2.3.2.1 Multilayer perceptron

A Multilayer Perceptron (MLP) comprises three or more layers of fully linked Artificial Neural Network (ANN) arranged sequentially, MLP with deep architecture called Deep Neural Network (DNN). The layers consist of the input, hidden, and output layers. In the hidden layers of a neural network, the model weight and biases control the behavior of neurons. These parameters can either increase or decrease the likelihood of a neuron's connection to a specific neuron in the output layer. Each input value is multiplied by the weights and then added to the bias, the resulting sum is then passed through an activation function to introduce non-linearity, this process continues until the output layer and it's called the feedforward. The Sigmoid and Rectified Linear Unit (ReLU) are the most commonly used activation functions. Eq.1 shows a neuron output's calculation mathematically, in the feedforward process of training an MLP. During model training, the weights (W) and biases (b) are adjusted by calculating the loss value and performing backpropagation throughout the network in reverse; Figure 5 illustrates the fundamental elements of the MLP.

$$y = \sigma (W^T x + b) \quad \text{Eq.1}$$

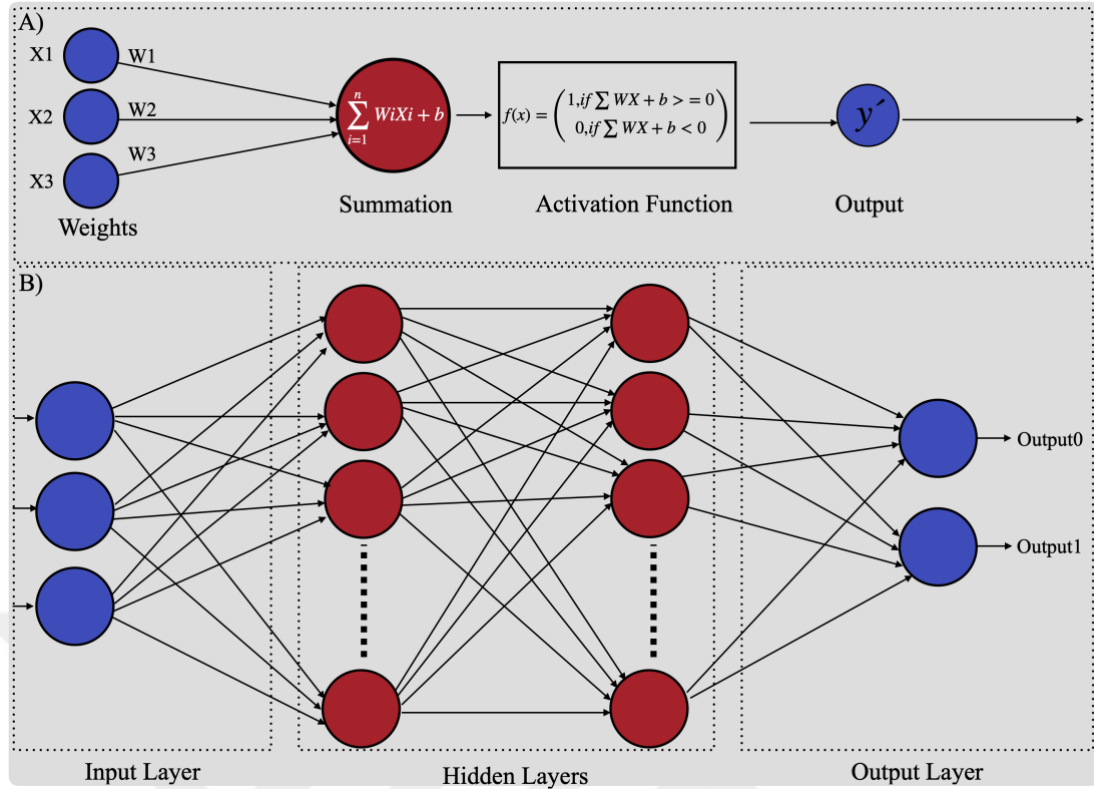


Figure 5. (A) The components of a single neuron include the input data, the weights, and the bias. The input values are multiplied with the weights and summed to a bias within the neuron, and the resulting value is sent to the activation layer to introduce non-linearity. The output of this layer becomes the input for the next layer. (B) A fully connected Artificial Neural Networks (ANNs) that forms a Multilayer Perceptron.

2.3.2.2 Convolutional neural networks

The integration between convolution layers and FCL introduced the Convolutional Neural Network (CNN), CNN models have significantly enhanced the efficacy of DL in image classification tasks, leading to advancements in computer vision for cancer diagnosis tasks (Gao, Yang, Lin, & Park, 2018).

The input is an image; the predicted class is the neuron with the highest probability in the output layer. CNN models use sequential group convolution and pooling layers to detect patterns inside an image while preserving the spatial relationships in grid-like data. Then, the flattened result is served to an FCL to make a prediction using the learned parameters. CNNs excel in these tasks because they can effectively analyze complex and multidimensional data to get spatially related data using two main components: convolution and pooling layers.

Convolution layer

The computer sees an image as organized multidimensional arrays with the shape of $[W^1, H^2, C^3]$. Each element of the array represents the color of the associated pixel. Channel arrangement and pixel encoding are crucial, especially in multidimensional RGB images. Therefore, flattening the data and directly inputting it into a fully connected layer for prediction purposes is impractical. By adopting this approach, we would fail to consider significant data that might reveal the spatial attributes of the input vector. Instead, we must utilize the input's dimensionality and maintain the integrity of the ordering and all characteristics, such as edges and colors, within the image. Therefore, the convolution layer, a linear transformation that preserves the concept of ordering, is adopted. The input-output relationship is characterized by sparsity, meaning that only a small number of input units contribute to a specific output unit. Additionally, parameter reuse occurs when the same weights are applied to several sites in the input.

The initial component in a convolutional layer is called the Kernel/Filter (K), which usually holds randomly defined values called weights, which are adjusted as the training progresses. The Kernel's primary function is to perform a convolution operation on an image. It moves from left to right and down the image, each step determined by a specified Stride (S) value. The stride value, along with the size of the Kernel, determines how many times the Kernel will shift over an image with shape of $[W, H]$. The image dimensions may increase with a padding (P) value, adapted by tasks where the exterior edges may be informative. Padding is basically adding a P dimensions zero array at the image matrix perimeter.

Each time the kernel moves, it performs a matrix multiplication between the kernel K_n and the corresponding stack of image patches I_n in the form of $([K_1, I_1]; [K_2, I_2]; [K_3, I_3])$. The results of these multiplications are then summed with the bias term to produce a single-depth channel of convoluted features as an output, CNNs use many filters among the convolution layers to effectively capture complex characteristics and identify diverse patterns in the input, generating a features map.

¹ Width

² Height

³ Channles

Pooling layer

Pooling layer is an essential part of CNNs, used to decrease the dimensionality of feature maps by summarizing subregions using different functions. Pooling is similar to convolution in certain aspects. Still, it replaces the linear combination of the kernel with other functions, such as calculating the average or maximum values inside the kernel. Figure 6 demonstrates the operations of the convolution and pooling layers, leading to the final flattened array that serves as the input for the FCL.

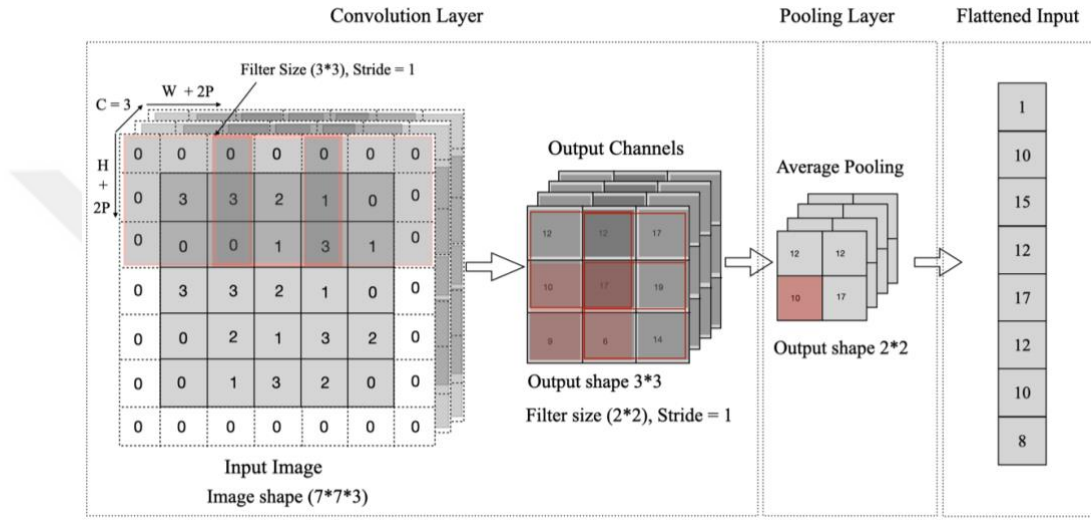


Figure 6. A (convolution layer) is a process where a kernel moves over an image of shape $[(W+2P), (H+2P), C]$ with a stride value of S , scanning from left to right and down. A single value in the output represents each group of values in the input and the kernel. Like the convolutional layer, the (Pooling Layer) applies a kernel that traverses across the array and performs a different function than in convolution; instead of calculating the sum, it applies a selected pooling function to the values within the kernel and provides a single output value. The (Flattened Input) is the final intended output obtained after numerous iterations of convolution and pooling, depending on the level of complexity in the image features. This flattened input will be the input for the next FCL.

2.3.3 Multiple instance learning

A significant challenge in digital pathology is the need for labeled data. Creating annotations that precisely depict particular tissue regions in an extensive collection of WSIs is laborious and time-consuming. Researchers can use imprecise labels by highlighting certain ROIs and assigning a single label to the annotated region as an alternative. Multiple Instance Learning (MIL) is a machine learning model that handles tasks involving weakly supervised learning, achieves this by effectively learning the mapping between a group of patches (instances) called bag and the inexact label

assigned to that bag. Figure 7 illustrates the difference between single-instance and multiple-instance learning models.

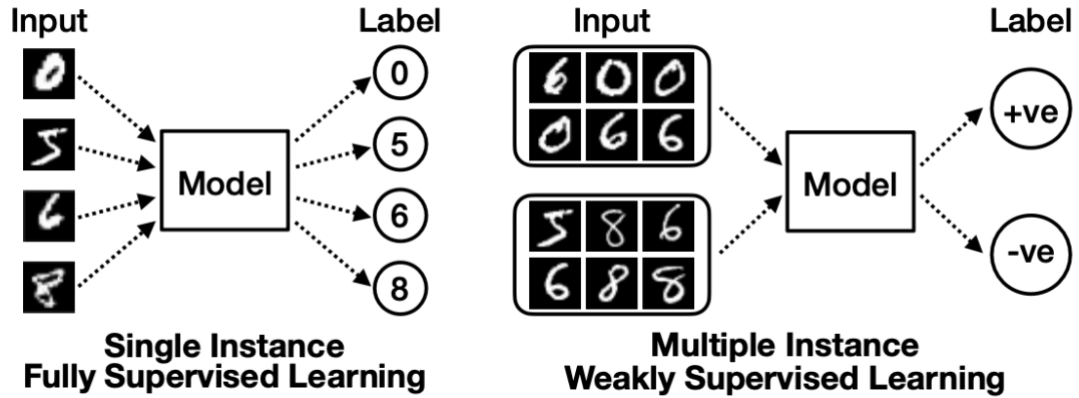


Figure 7. A visual explanation showing the difference between Fully Supervised Learning and Weakly Supervised Learning in the way of handling tasks. In single-instance learning, each image is considered a unique data point with its label. However, in MIL, a group of images are assigned a single label. In this particular example, the label for a bag is set to 'negative' if at least one instance inside the group is the digit 0. Otherwise, the assigned bag label is 'positive'. Image from (ONER, 2021).

2.3.4 Vision transformers

Transformers are a distinct DL architecture first created for tasks related to Natural Language Processing (NLP). They make use of the attention mechanism to enhance their computational capabilities. Transformers have a remarkable aptitude for understanding enduring connections within various texts and have become the leading technology in NLP. However, the attention mechanism design has been applied to pictures, resulting in the Vision Transformer (ViT) development, as described in work by (Dosovitskiy, et al., 2020). In Figure 8, we can see how the image is being processed as a sequence of tiles in ViT models.

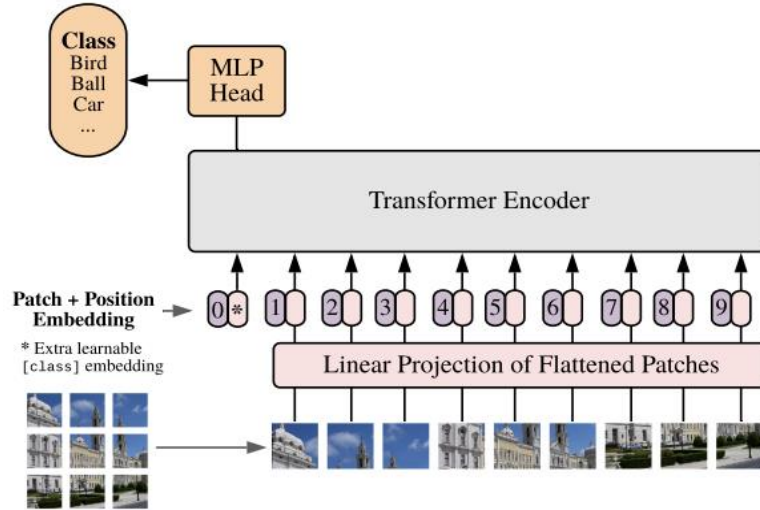


Figure 8. In the ViT model, an image is processed by dividing it into smaller patches. The attention mechanism in Transformers enables the model to selectively focus on specific tiles in an image that directly impacts the provision of a given label Image. Image from (Dosovitskiy, et al., 2020).

2.3.5 Learning approaches

2.3.5.1 From scratch

In the from scratch approach, a model is trained without any pre-existing knowledge; this entails initializing the model's parameters randomly, followed by training the model using a dataset relevant to the specific task. The model acquires the ability to identify and analyze unique features and patterns straight from the data by adjusting all parameters at every training epoch. Training from scratch requires a large amount of labeled data and high computational resources, especially for complex tasks or models.

2.3.5.2 Transfer learning

Transfer learning involves initializing the model with weights learned from a pre-trained model rather than starting the training process from scratch. Pretrained models are typically trained on extensive datasets such as ImageNet for purposes like image classification (Krizhevsky, Sutskever, & Hinton, 2017). Transfer learning can significantly reduce the need for computational resources when training a model. Transfer learning includes multiple techniques, including fine-tuning the pre-trained model on the new dataset; employing the pre-trained model as a fixed feature extractor; or adapting some components of the pre-trained model to the new task while keeping others frozen.

2.3.6 Explainable AI

ML models, particularly classification models, have been commonly referred to as black boxes due to their convoluted definitions. It is challenging to understand how the model arrived at a specific prediction. Explainable AI (XAI) has introduced novel techniques that can assist developers and researchers in explaining their models. These techniques, for example, can reveal the contribution of each pixel in an image to the final output in an image classification task. In the context of cancer diagnosis, certain features are crucial in determining whether to use a model in the clinic. A DL model leveraging features like the ones used by pathologists is more trustworthy and has a higher chance of being deployed for clinical use. To reveal feature attributions of our DL models, we use the Captum Pytorch library (Kokhlikyan, 2020), which has brought direct and easy-to-use functions for model interpretability.

2.4 Digital Histopathology Automation and Dataset Preparation

2.4.1 Digital histopathology automation

The increased capabilities of Digital Histopathology could enhance pathologists' productivity in addressing monotonous and repetitive tasks. Nevertheless, it is crucial to acknowledge that automation can assist, but in no way replace, the expertise of pathologists. The collaboration between pathologists and advancements in Digital Histopathology will enhance diagnostic precision and ultimately improve clinical decision-making in the future by utilizing Computer-Aided Diagnosis (CAD) tools such as DL (Pallua, Brunner, Zelger, Schirmer, & Haybaeck, 2020). To accomplish this, some researchers have used fully supervised learning, which works well with lots of labeled data, while others have used weakly supervised learning, which works well with a smaller amount of labeled data.

2.4.2 Dataset preparation

Digital Histopathology automation seeks to automate the decision-making process and diagnosis that pathologists typically perform. For a computer to do this task, it must acquire the same information pathologists possess in examining tissue morphology. Computers to obtain such information employ computer vision applications and DL models, which can extract distinctive characteristics and generate predictions based on the reached knowledge. Providing this information to a DL model necessitates a consistent dataset (DS). In the following subsections, we will thoroughly

explain the stages of preparing DS to effectively address the issues associated with WSIs in computer vision tasks.

2.4.2.1 Slide annotating and patches cropping

As acknowledged before, the involvement of a pathologist's knowledge is essential in automating diagnosis tasks. Pathologists must precisely annotate and label ROIs within a WSI. Subsequently, AI engineers can use these annotations to create a labeled DS fits for training DL models. WSIs are typically huge, measuring 100,000 pixels by 100,000 pixels. This size makes it impractical to simultaneously serve the entire image to a model. Thus, cropping extracts smaller patches from the image at a specific magnification level. These patches are then labeled based on pathologist annotations. This process aims to generate tiny patches to train a model to predict patches with similar features during the model's inference phase.

2.4.2.2 Stain normalization

Different staining methods and scanners produce slides with variations of color and intensity, as shown in Figure 9, which may negatively impact the performance and accuracy of CAD systems. The purpose of stain normalization techniques is to assist CAD systems by producing patches that have a consistent visual look for the many types of stains used in WSIs (Anghel, et al., 2019); one of the methods used to achieve this task is the Macenko method. The procedure includes converting RGB images into Optical Density (OD) values and removing data below a predetermined threshold. Subsequently, Singular Value Decomposition (SVD) is applied to the OD values, creating a plane based on the two most significant singular values. The data is then projected and normalized onto this plane. Angles are calculated, and robust extremes are identified. The extreme values are then reverted to OD values. Ultimately, the procedure outputs the most optimal stain vectors (Macenko, et al., 2009).

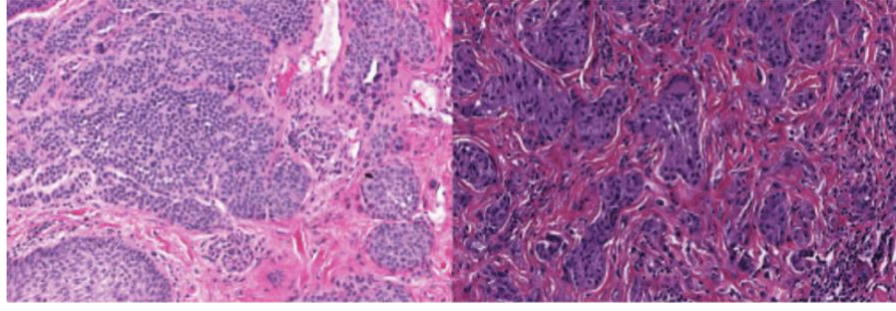


Figure 9. Following are two different histology slides of melanomas, each stained with hematoxylin and eosin, but with noticeably distinct appearances. The photos were acquired by the process of scanning the slides at a magnification of 20X using an Aperio Scans. Image from (Macenko, et al., 2009).

2.4.2.3 Data segregation

In order to train a DL model, it is essential to divide the DS into three distinct subsets: training, validation, and testing. The training set serves as a directive to the training process, the validation set determines the most optimum model, and the testing set evaluates the model's performance on new and unseen data. In the context of cancer diagnosis, unseen data refers to tissue structures with features that vary from the ones used for training. Due to the stochastic nature of cancer development, it is possible to have two patients with the same cancer type, and their cells have different structures. Given the wide range of differences across patients, it is essential to consider each patient as an individual data point; thus, segregating data at the patient level is crucial, otherwise, data segregation will suffer from data leakage issues (Oner, Cheng, Lee, & Sung, 2020).

2.4.2.4 Data augmentation

Many labeled histopathological images are needed to build a robust model for multiple scenarios. Data Augmentation virtually increases the number of patches. It involves applying random pre-defined augmentation methods to an image within the DS before feeding it to the model during training. One commonly used augmentation method is vertically and horizontally flipping a tissue tile. This method is preferred because it does not result in any data loss in the tissue morphology. The benefit is that it presents the tissue tile in different directions, so the model can be aware of the other directions in case it appeared in future scans.

2.5 Deep Learning Applications in Cancer Prognosis

The DL models, with its remarkable precision and versatility, has emerged as a leading method for spotting intricate photo patterns, a feat that has left the scientific community in awe (He, Zhang, Ren, & Sun, 2016). Its pivotal role in developing several CAD systems for diverse types of cancer further underscores its significance. The model's unique ability to assist pathologists in decision-making across various cancer types, as evidenced by numerous studies, is a testament to its potential to revolutionize cancer diagnosis.

In their groundbreaking 2017 research, Song et al. discovered that CNNs outperformed DNNs in classifying lung nodules, achieving an impressive AUROC value of 0.916 compared to 0.877 for DNNs (Song, Zhao, Luo, & Dou, 2017).

A study by (Kather, et al., 2019), who created a CNN DL model that performs a multi classification task to analyze Colorectal Cancer (CRC) tissue microenvironment, and make accurate prognostic predictions by calculating the Tumor Stroma Ratio (TSR), revealed that the TSR is an autonomous prognostic factor for Overall Survival (OS). Their classification model achieved an accuracy of 0.9610.

In similar approach, (Zhao, et al., 2020) used several DSs to accurately calculate the Tumor-Stroma Ratio (TSR) in a WSI of a biopsy taken from a patient with CRC, their classification model achieved an accuracy of 0.9572 on the testing DS. Both studies above performed prognosis predictions using a segmentation map constructed using a CNN model trained to classify tissue tiles into specific classes. Each color in the segmentation map represents the predicted class of a tile. Subsequently, they applied image analysis techniques to these segmentation maps to extract the desired values.

Moreover, (Oner, et al., 2022) in their study, built a multi-resolution classification model that assist pathologists in detecting malignant glands in core needle biopsies of low-grade and low-volume cases in prostate cancer, their classification model achieved an area under the receiver operating characteristic curve (AUROC) value of 0.992 with a high confidence interval. Those studies have shown the high ability of CNN models in histopathological image classification tasks.

Furthermore, (Fu, et al.) developed a DL algorithm that autonomously identifies oral cavity squamous cell carcinoma (OCSCC) from histopathological images. This

algorithm, using cascaded CNNs, is able to detect OCSCC from pathological images with a high degree of accuracy. Its autonomous nature, from examining an oral photograph to producing a list of confidence scores for classifying patients with OCSCC and controls, instills confidence in its reliability and efficiency. This breakthrough in DL technology has the potential to significantly improve the speed and accuracy of cancer diagnosis, leading to earlier detection and more effective treatment.

To address the issues of limited availability for labeled data, (ONER, 2021) introduced in his research a Novel Multiple-Instance Learning Model (MIL) that could autonomously identify cancer histology without relying on explicitly labeled data, giving us the possibility of exploiting weak labels to achieve a prognosis task. However, it's important to note that this approach is still in its early stages and requires further validation and refinement before it can be widely adopted in clinical setting.

(Zeid, El-Bahnasy, & Abo-Youssef, 2021) Their research has shown the high potential for ViT models in performing multi-class classification for histopathology images; through training a ViT model on the same DSs used in Kather et al., their model was able to outperform the original paper results by getting an accuracy of 93.3%.

Chapter 3

Methodology

This chapter will present a comprehensive outline and workflow of the research. We will also discuss the various methodologies employed in the experiments to ensure the utmost accuracy and dependability of the results.

3.1 Research Design

3.1.1 Prognosis prediction in OSCC

The primary objective of this study is to showcase the potential of DL models in predicting prognosis (good vs. poor) in OSCC patients. To reach the goal above, we have two phases. Phase one: we developed a DL model that generates a segmentation map for a WSI, followed by another DL model that uses the segmentation map as an input to predict the tumor vs. non-tumor classification of the WSI. Phase two: we developed an MIL model that predicts prognosis (good vs. poor) for the tumor slides.

3.1.2 Effects of data segregation

In the clinic, cancer cells present diverse growth patterns in each patient, leading to unique patterns. Our model's ability to identify tumor areas in slides from patients never encountered before, is critical to our research. In this experiment, we have trained and inferred multiple DL models using various data segregation approaches and used external data sources to assess the impact of each segregation approach, to see which one could ensure the reliability of a DL model in replicating the traditional clinical workflow in personalizing the cancer diagnosis.

3.2 Datasets (DSs)

It is crucial to have a sufficient quantity of adequately labeled patches to effectively train, validate, and test the DL model's ability to classify patches. The datasets (DSs) utilized in these experiments were acquired from two distinct sources and will refer to them as HKU OSCC and the Public CRC DSs.

3.2.1 HKU OSCC dataset

Our collaborators from HKU provided us with a total of 470 WSIs, 278 BF images belong to 278 unique patients, in addition to 192 CF images for the same glass-slides

but scanned using LSCM, for participants at Queen Mary Hospital, Hong Kong between January 1, 2004 and December 31, 2022; those slides are categorized into three main classes: poor prognosis, good prognosis and non-tumor slides. Some BF slides have been used to crop patches for training, validating, and testing the segmentation model. The others were utilized by generating segmentation maps, each to serve as a data point for the phase one diagnosis model; in addition to that, we used them to obtain ROIs to build the phase two prognosis model. Table 1 displays a tabular count summary of the employed WSIs.

Table 1

A Summary of the WSIs Employed at Different Phases of the Research.

Slide label	Brightfield	Confocal	Total WSIs
Good Prognosis	89	89	178
Non-Tumor	103	103	206
Poor Prognosis	86	NA	86
Total	278	192	470

3.2.1.1 Ethics statement

This study complied with the ethical principles of the Declaration of Helsinki. Institutional review board approval was obtained for this study (HKU/HA HKW, UW21-602).

3.2.1.2 Annotation mask creation

Beside the slides, we received 186 annotations used as cropping guides and labeling engines. These annotations were diverse, comprising 50 for poor prognosis, 70 for good prognosis, and 66 for non-tumor annotation files. The categories of tissue that we considered include, Adipose Tissue (ADI), Red blood cells (DEB), Dysplasia (DYP), Gland (GLD), Hyperplasia (HYP), Immune cells (LYM), Muscle (MUS), Stroma (STR) and Tumor (TUM).

First, we loaded the WSIs into QuPath software and imported their corresponding annotation files (in qpdata format). Figure 10 displays WSI “qpdata” annotations

imported to QuPath. In addition to the existing annotations, we randomly annotated background areas from random slides to be cropped during the patch-cropping step. We added the Background (BAC) to our classes to ensure our model could recognize and differentiate it from other tissue types during WSI segmentation process. We validated these files and exported them as JavaScript Object Notation (JSON) files. JSON files contain the slide name, the coordinates for the ROIs, and the color and class name. We used them to create annotation masks for the slides by representing each ROI by a specific color, we generated the annotation masks at levels four and five based on the best available resolution for each slide.

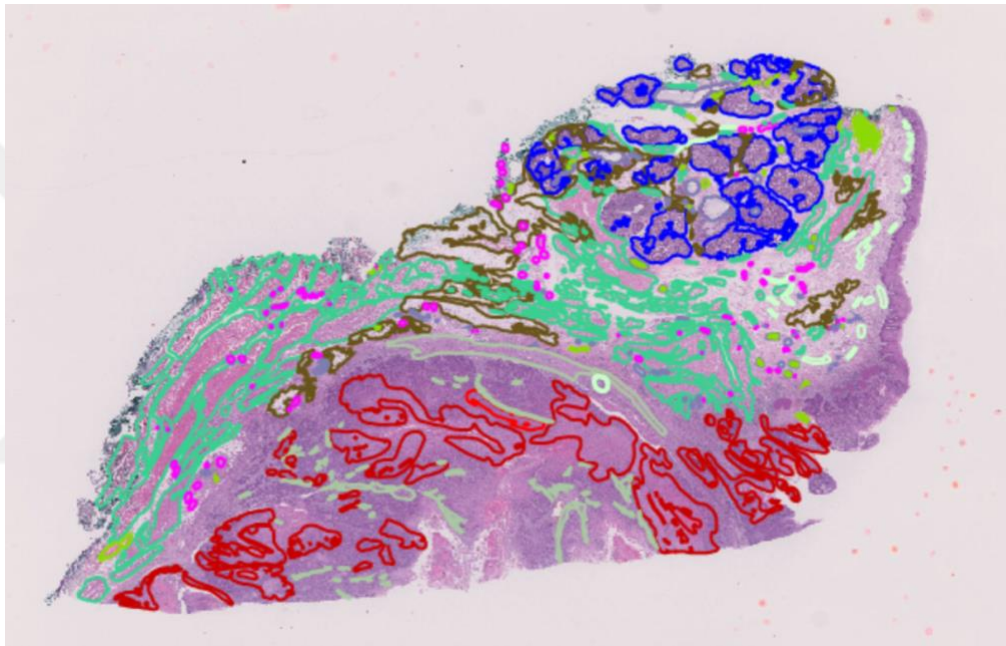


Figure 10. The annotations of a WSI as they appear after being loaded into QuPath, each color regions represent a specific tissue type.

3.2.1.3 Patches cropping

For each slide with an annotation mask, a sliding window was moved over the mask to crop patches of size 224 by 224 pixels at level one (0.50 micrometer per pixel) from the corresponding slide. Suppose the percentage of a tissue class within the window was greater or equal to 95%. The corresponding patch is extracted from the slide and labeled as the corresponding tissue type, Eq.2, Eq.3 and Eq.4 showing how we obtained sliding window size and stride over the mask.

$$\text{resolution ratio} = \frac{2^{\text{mask level}}}{2^{\text{patch level}}} \quad \text{Eq.2}$$

$$\text{mask window size} = \frac{\text{patch size}}{\text{resolution ratio}} \quad \text{Eq.3}$$

$$\text{mask stride value} = \frac{\text{patch level stride}}{\text{resolution ratio}} \quad \text{Eq.4}$$

We cropped all patches at level one with a 0.5 $\mu\text{m}/\text{px}$ and a 224 by 224 pixels tile size. The tiles for the classes with major annotated regions, namely ADI, BAC, GLD, MUS, and TUM, were distinct. However, since their annotated areas are relatively small, DEB, DYP, HYP, LYM and STR patches have overlapped to generate more patches. Figure 11 illustrates the cropping process over the annotation mask.

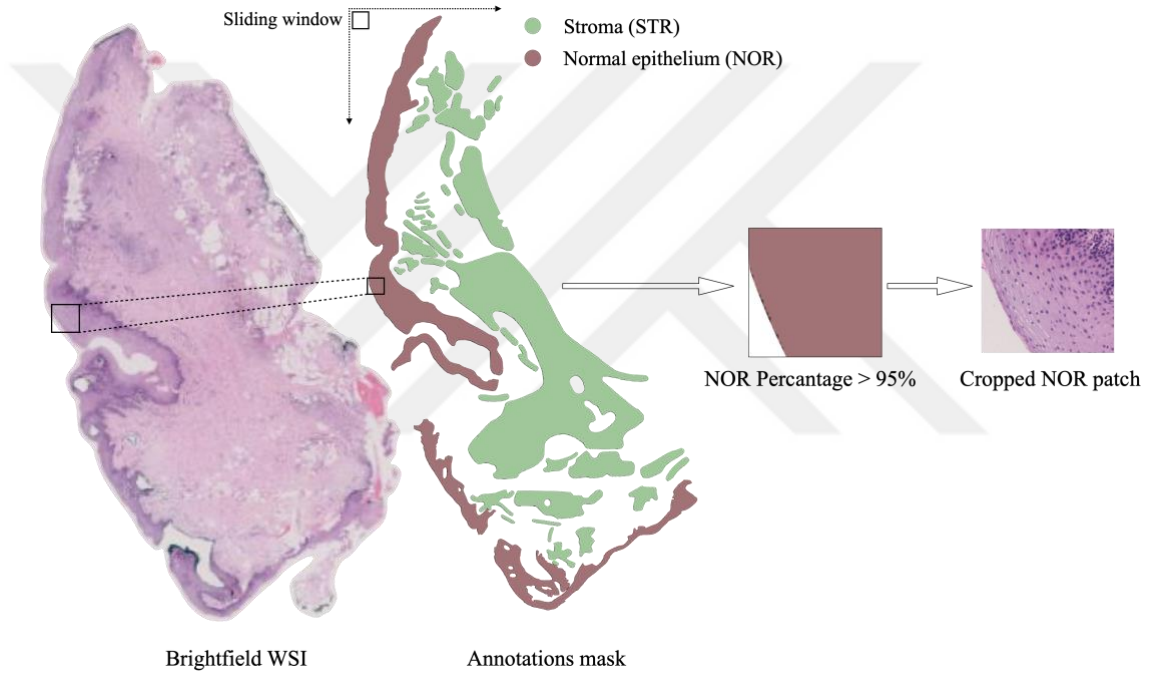


Figure 11. Cropping a patch over an annotations mask involves a sliding window that moves from left to right and down over the annotations mask. The displayed color percentage in each loop determines whether the patch will be considered or neglected. If the color percentage presented in the sliding window is greater than 95%, the patch will be extracted from the Brightfield WSI and saved alongside its label, which is obtained from a given color and labels key-value pairs.

The cropping process produced 267K tiles for all classes and patients, during the cropping process, we combined the HYP and DYP classes into one class, Normal epithelium (NOR), and we ended up with nine classes in total. Figure 12 shows example patches for each of the classes.

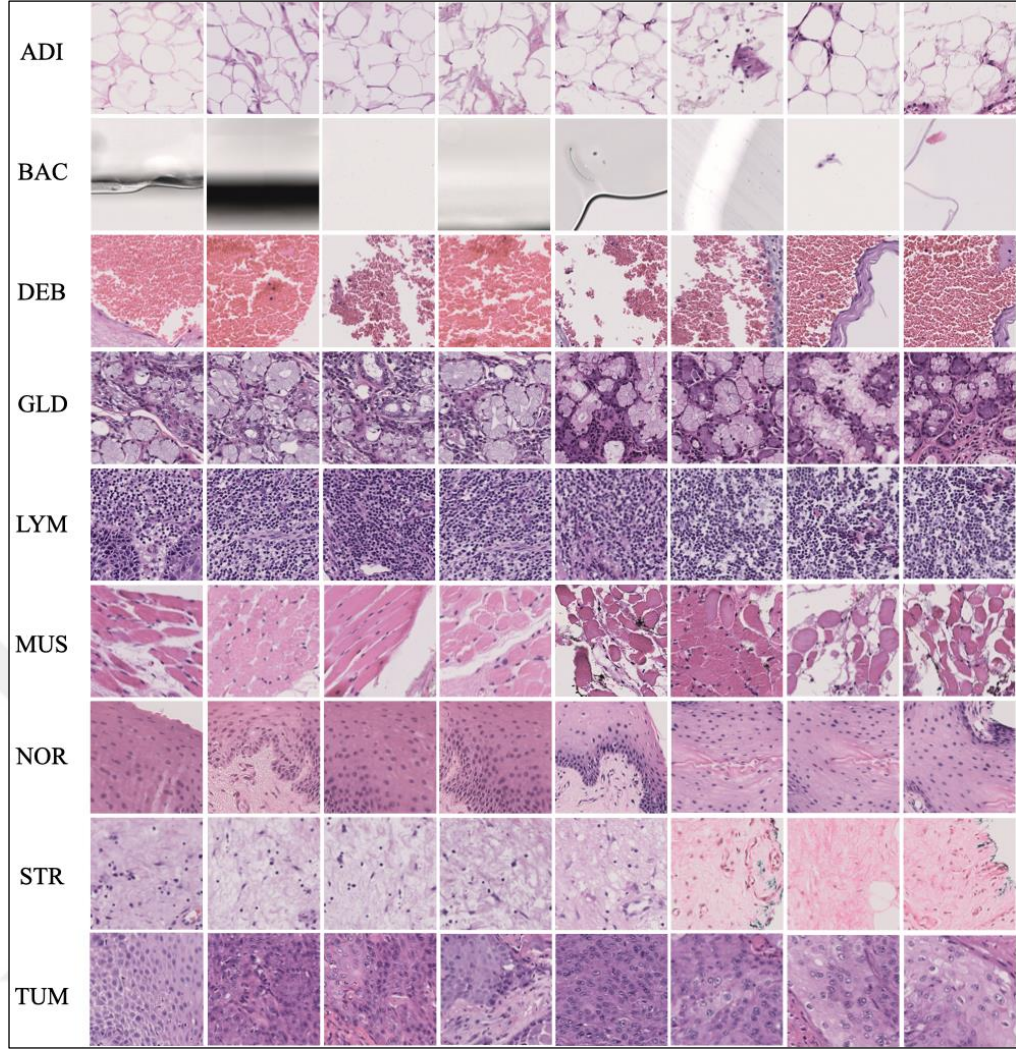


Figure 12. Eight samples for each class images, show the distinctions among the patches of each class.

3.2.1.4 Data segregation

The 267K cropped patches are distributed throughout 186 folders. Each folder has various sub-folders, ranging from one to nine, depending on the classes presented at that particular WSI. The data segregation stage aims to provide a well-balanced and patient-level segregated DS. The task is to partition these 186 folders into training, validation, and testing sets, keeping in mind the enforced separation at the patient level and a fair representation of classes throughout the splits. To achieve this, we have defined the percentage of each split as follows: 70% of the patients for training, 10% for validation, and 20% for testing. The approach iteratively picks folders randomly for the three divisions. Following that, the total number of patches in each division is calculated. If the difference between the patch count is 5% or less, and the total count matches the desired splitting ratio, the segregation process will finish by creating three

text files, each with the slide names that are going to be used in this split. We used this technique three times, where each iteration resulted in a three text files. The objective of this strategy is to create a robust model that is trained on a wide variety of distinctive characteristics. Figure 13 illustrates the distribution of the classes among the splits in the first segregation slides.

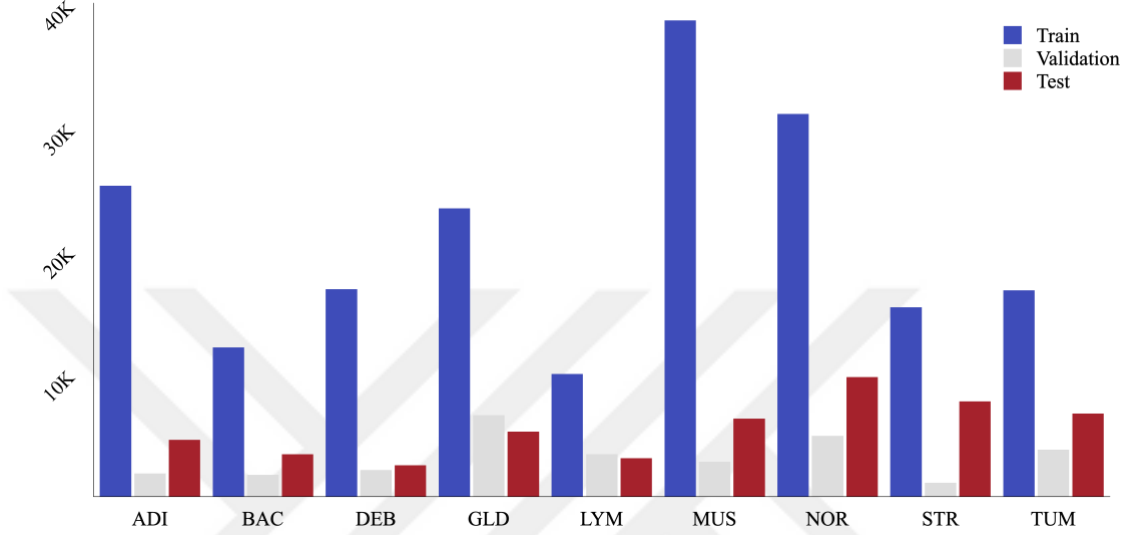


Figure 13. The distribution of the classes among the splits of the first segregation slides, shows that the MUS class possesses the highest number of patches, whereas the LYM class exhibits the lowest number.

3.2.1.5 Stain normalization

To minimize the variation in color intensity between patches and achieve a more uniform color distribution, we utilized the Macenko stain normalization method (Macenko, et al., 2009). We carefully selected a specific TUM patch from test split, with a wide range of colors as the target image. Then we applied the stain normalization for all patches except the DEB, ADI, and BAC patches since they don't include tissue to be stained, also staining them in this way is going to cause loss in the cells structure.

3.2.1.6 Balancing classes

Ensuring an equal number of patches for each class during training is recommended to enhance model generalizability. There are two methods to achieve this: under-sampling and over-sampling. Under-sampling involves randomly selecting the same number of instances from the majority classes as in the minority classes. However, this method leads to a loss of data. On the other hand, the over-sampling

method considers the patch count of the majority class and virtually duplicates the patches from minority classes to be equal to the majority class count. We conducted our experiments using the over-sampling method, utilizing around 35K patches per class. Figure 14 illustrates how each method works to achieve the balancing.

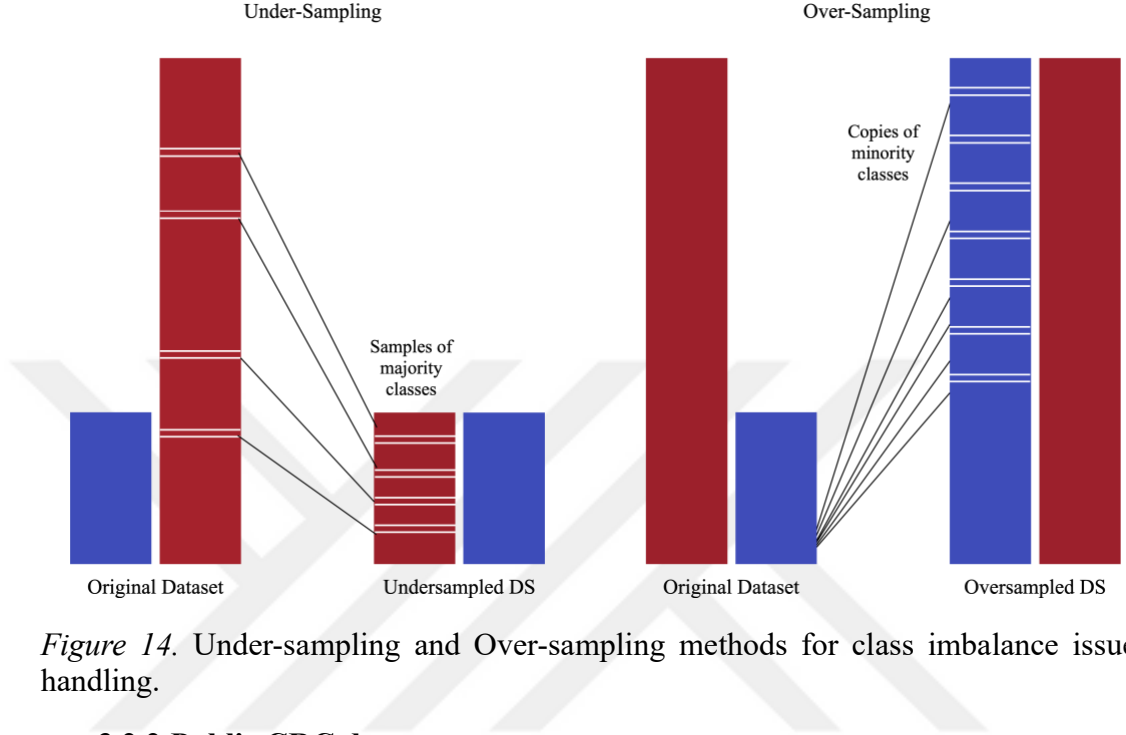


Figure 14. Under-sampling and Over-sampling methods for class imbalance issue handling.

3.2.2 Public CRC dataset

This DS is a set of non-overlapping image patches extracted from H&E stained Colorectal Cancer (CRC) slides, it comprises five folders derived from the initially provided data NCT-CRC-HE-100K and CRC-VAL-HE-7K, made available in the publication by (Kather, et al., 2019). Additionally, it includes TSR-CRC-Training-set-part1, TSR-CRC-Training-set-part3, and TSR-CRC-Test-set2 from the publication by (Zhao, et al., 2020), DSs were obtained from various countries, including the USA, Germany, and China. The patches have been classified into nine different categories.

Table 2 provides information about the source of the DSs, the number of slides, and the count of cropped patches for each folder.

Table 2

The Dataset's Components Aside with its Origin, the Number of Slides Used for Cropping, and the Overall Count of Cropped Patches Across All Classes.

Dataset	Source	# Slides	# Patches
---------	--------	----------	-----------

NCT-CRC-HE-100K	NCT and UMM, Germany	86	88.9K
CRC-VAL-HE-7K	NCT and UMM, Germany	50	6.3K
TSR-CRC-Training-set-part1	TCGA ⁴ , USA	85	88.5K
TSR-CRC-Training-set-part3	GPP's ⁵ Hospital, China	106	105.1K
TSR-CRC-Test-set2	Yunnan Cancer Hospital, China	48	22.5K

We have categorized them according to their origins and will refer to them as TCGA DS, Germany DS, and China DS. In Table 3, we present the distribution of classes within each DS.

Table 3

Patches Distribution Among the Classes for the Three DSs.

Dataset / Class	ADI	BAC	DEB	LYM	MUC	MUS	NOR	STR	TUM
TCGA DS	9710	10736	10603	2337	9398	12798	9962	10011	12899
Germany DS	11745	11413	11851	12191	9931	14128	9504	10867	15550
China DS	12500	25065	12499	8331	13237	12500	15868	15084	12500

3.3 Prognosis Prediction

We employed a series of DL models to predict prognosis in OSCC. We developed the segmentation model to examine the OSCC slide and generate a semantic segmentation map. The first model identifies tumor slides to be examined later by the subsequent prognosis model. We will thoroughly explain all aspects in the below sections.

3.3.1 Segmentation model

3.3.1.1 Dataset preparation

The segmentation model plays a significant role in the subsequent diagnosis model. since the produced segmentation map's data serves as the main input for the

⁴ The Cancer Genome Atlas.

⁵ Guangdong Provincial People's.

diagnosis model. We carefully divided our data and ran several experiments to develop a model that predicts classes accurately for unseen patients, thereby ensuring generalization. We considered various factors during our analysis, such as the stain normalization in patches and distinct feature identification. We showed how to address each factor through multiple experiments to determine the best method to create a more comprehensive model.

We generated three random segregations at the patient level, which we will refer to them as SEG1, SEG2, and SEG3. Each segregation contains different patients in its training, validation, and testing splits. Table 4 displays the number of patches for each class in the three segregations. During the training, we applied the over-sampling method to virtually duplicate patches, in this way, 34150 images of each patch were used to train the models in SEG1, while 38588 in SEG2 and 34168 for SEG3.

Table 4

Patches Distribution Between the Classes for the Three segregations, Where Different Patient has been Randomly Reserved for Each Split.

SPLIT	ADI	BAC	DEB	GLD	LYM	MUS	NOR	STR	TUM
SEG1									
Train	25182	12090	16287	22148	11611	29780	34150	18484	19173
Valid	2722	1767	2286	5993	259	8508	1658	3219	2259
Test	3761	3441	2917	7098	4595	9425	9806	2474	5801
SEG2									
Train	25198	12090	16806	23365	9940	38588	31007	15338	16708
Valid	1871	1767	2138	6600	3422	2810	4915	1119	3803
Test	4596	3441	2546	5274	3103	6315	9692	7720	6722
SEG3									
Train	22550	12090	14797	27721	14545	34168	26434	17935	18585
Valid	3627	1767	674	2452	1425	1609	4433	3756	6421
Test	5488	3441	6019	5066	495	11936	14747	2486	2227

This technique has provided three unique DSs, each used with two versions of the images: stain-normalized and non-stain-normalized, which gave us six different DSs. Figure 15 displays the segregations utilized to train the segmentation models.

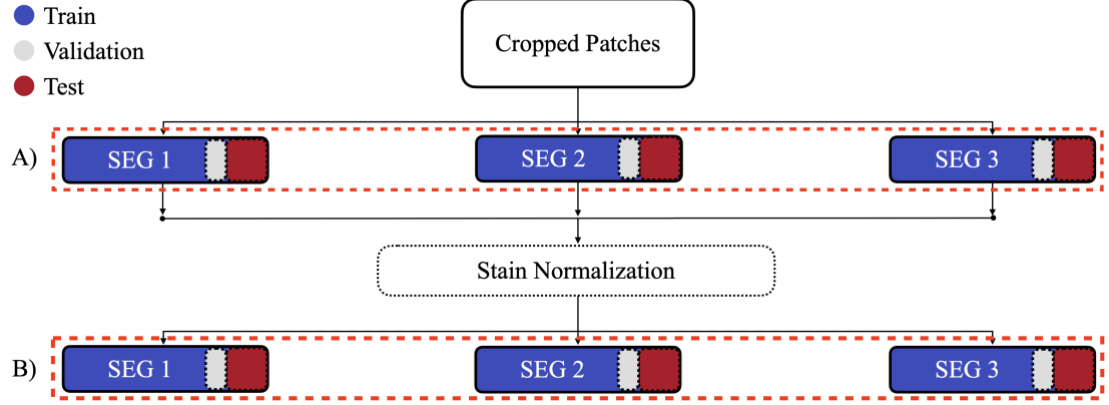


Figure 15. Two distinct DSs comprise three segregations with different patient groups within their splits. Each segregation was utilized to train, validate and test a two DL model. Segregations among the DSs are identical. (A) non-stain normalized, (B) stain-normalized.

3.3.1.2 Deep learning models

Our research involved training a Resnet50 CNN and ViT model for each of the 12 previously described DSs. This comprehensive approach resulted in a total of 12 models. We then employed various comparison methods to identify the optimal model with the highest ability for distinguishing between the classes. A detailed analysis of these comparisons will be presented in Chapter 4.

We followed the from scratch learning approach for the CNN models and the transfer learning approach to train our ViT models, we obtained the CNN model architecture from PyTorch hub and the for the ViT, we used the pretrained model (Wu, et al.). For hyperparameters, the batch size was 64, learning rate of 0.0001. The optimization technique employed was Stochastic Gradient Descent (SGD) with a momentum value of 0.9, and the loss function was Cross-Entropy Loss. Training epochs were 50 for all models. We trained and inferred CNN models on a desktop workstation with NVIDIA GeForce RTX 3080 GPUs and a workstation with NVIDIA GeForce RTX 3090 GPUs for ViT models.

To evaluate the classification performance, we calculated the accuracy on the testing split, and the area under the Receiver Operating Characteristic curve (AUROC), which measures the overall ability of a model to distinguish between positive and negative

classes, and the area under the Precision – Recall curve (AUPRC), which is useful for imbalanced DSs and measures the trade-off between precision (positive predictive value) and recall (sensitivity). Then, the McNemars's test (McNemar, 1947) was used to select the best model, ensuring the precision of our results.

Moreover, we have calculated the best model's classification report, which shows important metrics to assess the model performance are; (1) Precision, a measure of the model's positive prediction accuracy, is calculated by dividing the number of true positives (TP) by the sum of true positives and false positives (TP + FP). (2) Recall, also known as sensitivity, indicates the model's ability to detect positive instances accurately, it's determined by the ratio of true positives (TP) to the sum of true positives and false negatives (TP + FN). (3) F1-Score, a metric that indicates high precision and recall when it has a high value, is derived by multiplying the precision and recall values, then multiplying the result by two, and finally dividing it by the sum of precision and recall. On the other hand, (4) Support denotes the actual occurrences of each class in the test split.

3.3.1.3 OTSU thresholding

We used OTSU thresholding to create masks separating the foreground and background over the WSIs (OTSU, 1979). We first obtained the slide at level 6, the output image shape determined the mask dimensions. Then, we applied various image-processing techniques using the OpenCV (Bradski, 2000). Initially, we transform the cropped picture into grayscale. We subsequently applied different noise reduction methods, such as Gaussian blurring. This mask accelerated the diagnostic procedure by decreasing the time required to build the segmentation map. In the next part, we will explain how we use this mask to improve the speed of the segmentation process by avoiding cropping background tiles. Figure 16 shows an example of an OTSU mask generated for a WS

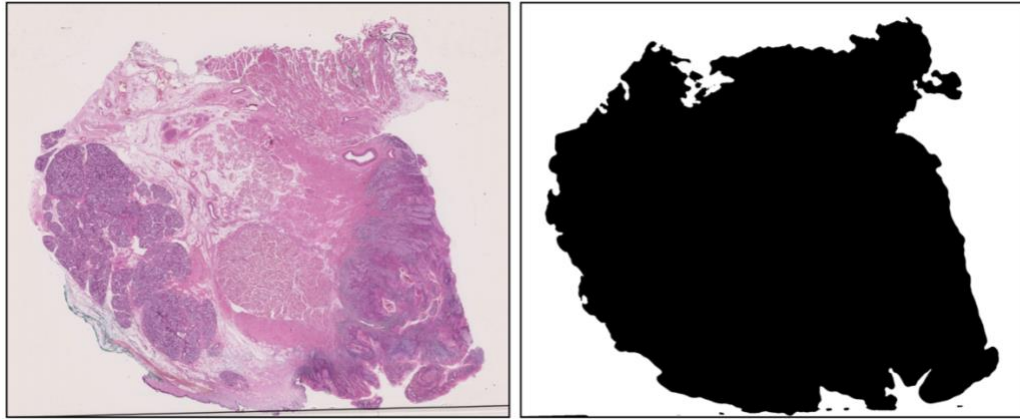


Figure 16. The picture on the left is a JPEG image extracted from the WSI at level 6, while the image on the right is the OTSU threshold mask results for the same image.

3.3.1.4 WSI segmentation

We began by creating an empty (zeros) segmentation array of a shape equals to the WSI's dimensions at level 6. The process involves capturing tiles from the WSI using a sliding window of size 224*224 pixels that moves from left to right and down over its corresponding OTSU mask. Once the highlighted area within the window is all black, the tile is cropped from the WSI and transformed. Then, we used the segmentation model to predict the probabilities for the cropped tile. The highest probability is then used to substitute the corresponding area values in the WSI segmentation map's array with a specific color value. We also logged the predicted probabilities with their corresponding coordinates at both the WSI and the OTSU mask's levels after each prediction, the probabilities will be used in the subsequent model's DS generation.

3.3.1.5 WSI annotations file

Utilizing the segmentation maps, we generated JSON annotation file for each WSI, it holds the geographical coordinates, class names, and colors information for each segmented tile within the WSI. We used the findContour function from the OpenCV library and implemented custom functions to adjust the coordinates to match the WSI at level 0. The objective of this file is to reduce the workload on pathologists by providing a comprehensive annotation for the slide, eliminating the need for manual detection and annotation of the ROIs. The straightforward structure of the file allows pathologists to immediately integrate it into QuPath for further validation or in-depth examination of a particular ROI.

3.3.2 Tumor vs. non-tumor WSI classification

After generating the segmentation maps of all the slides, we utilized them to generate a new DS to train this diagnosis model on classifying a WSI into tumor or non-tumor. The segmentation maps data and their patch probabilities served as the primary input for this model. Figure 17 illustrates the entire process: (A) receiving a slide and generating the segmentation map and probabilities file, (B) classifying the slide as tumor or non-tumor.

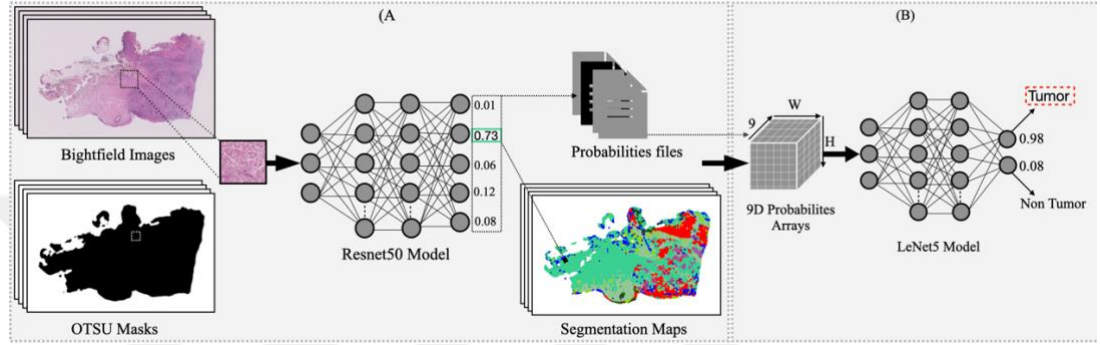


Figure 17. (A) The first stage involves creating a segmentation map using the best segmentation model. (B) The segmentation maps and probability files are then used to construct a 9D array for training a LeNet5 model on the Tumor Non-Tumor diagnosis task.

3.3.2.1 Dataset preparation

We created this DS using the segmentation maps and the probability files. Segmentation maps were basically used to get the dimensions to initialize a zeros array with the shape of $[W/3, H/3, 9]$. Taking into account the sliding window stride, which was 96 pixels in the segmentation map generation process, we adjusted the dimensions by dividing it by three. This strategy helped us minimize redundant data while overwriting the zeros with the highest probability value. We applied zero padding to the segmentation maps to align them with the largest width and height map to ensure that all elements in the DS have the same shape and can be fed to a model as batches.

We produced 278 probability arrays for the 278 WSI and assigned poor and good prognosis slides for the tumor class, and tumor-free slides went under the non-tumor class. Subsequently, we divided the data into three sets, considering that the slides were used to train the segmentation model, will also be used to train this diagnosis model, and so on for the other splits. The base segregation used was the one that has been used for the best segmentation model, ensuring we will not confront any data leakage. Since

the number of slides used to train the segmentation model differs, we added the previously used slides to their corresponding splits then we randomly distributed the slides that did not appear in the base segregation between the splits.

3.3.2.2 Deep learning model

LeNet5 architecture (Lecun, Bottou, Bengio, & Haffner, 1998) is utilized in this stage because it's light and can cover the complexity of the segmentation map data. We fine-tuned the hidden layer parameters to match the new input size by implementing a function to dynamically calculate them based on the received input size.

The model underwent 500 training epochs using the SGD optimizer and a batch size of 4 and learning rate of 0.0001. We calculated the loss using the Cross-Entropy loss function. To regularize the network, we used a weight decay of 0.00001 to penalize large weights and a node drop-out probability of 0.2 after each layer to mitigate overfitting. The AUROC and AUPRC were used to evaluate the classification performance.

3.3.3 Good vs. poor prognosis prediction

3.3.3.1 Dataset preparation

We have received 384 "qpdata" annotation files from our collaborators at HKU, a total of 192 files for BF images and an equal number for CF images. Within the annotations, the pathologist manually marked certain ROIs that may contain characteristics that aid in distinguishing good and poor prognosis slides. We transformed the "qpdata" into JSON annotations and developed a Python method to crop the regions from both BF and CF slides. We then applied a center crop of 512*512 pixels to the images and stored them in folders where each slide folder holds its cropped ROIs.

The 192 slides were divided into three splits 70% training, 10% validation, and 20% for testing, and again, following the same segregation that has been used for the best segmentation model, to prevent any data leakage.

3.3.3.2 Deep learning model

Considering the nature of our data, where the slide label is used as the ground truth for a group of images, we adopted a weakly supervised learning technique to build our

DL model. Each slide in our scenario represented a bag, and the images of the slide represented the instances of that bag.

We developed an MIL DL model consisting of three main components. The starting component is the feature extractor, a Resnet50 model pre-trained on ImageNet DS obtained from the PyTorch hub library. We sat the model in the evaluation mode and used it to extract spatial features of size $[1,1024]$ from each image. The following component is a linear layer that receives input of shape $[\# \text{ instances}, 1024]$ and learns the mapping between the bag's features. The last component is a classification layer to classify the bag into good or poor prognosis. The model underwent 500 training epochs, with a batch size of 1, a learning rate 0.0001 and a weight decay of 0.00001 to prevent the model from overfitting. Figure 18 illustrates the DS preparation process and the MIL prognosis model components. The AUROC and AUPRC were used to evaluate the classification performance.

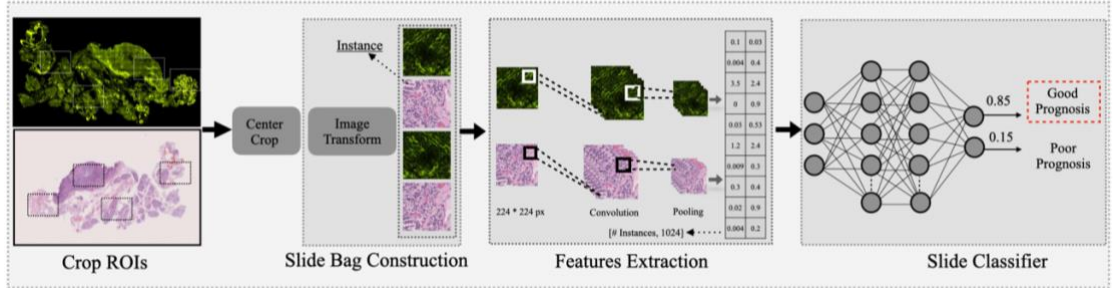


Figure 18. MIL model components, initially, we extracted and processed ROIs from BF and CF slides. Subsequently, we constructed bags of transformed images and used a fine-tuned Resnet50 model to extract the features of each instance within the bag. These extracted features are further processed by a linear layer which learns the mapping between the instances and determines the slide label whether good or poor prognosis.

3.4 Effects of Data Segregation on Model's Performance

3.4.1 Dataset preparation

We utilized the Public CRC DS for this experiment to show the effect of different segregation methods on the final model results. We used TSR-CRC-Training-set-part1 DS as the basis for creating three separate DS versions generated using various separation methods. By obtaining a unique list of the patient information from each patch's name within the DS, we got a list of 85 individual patients, we used them to create three versions of the DS by segregating the data based on Patch Level with

complete intersection, Patch Level with partial intersection, and Patient Level. Then, three subsets, train, validation, and test sets, with a split ratio of 70:15:15, respectively, were created for each segregation. We used Germany and China DSs as external tests to evaluate the performance of each model; in Figure 19, we illustrate a hierarchy of the DS used in the experiments.

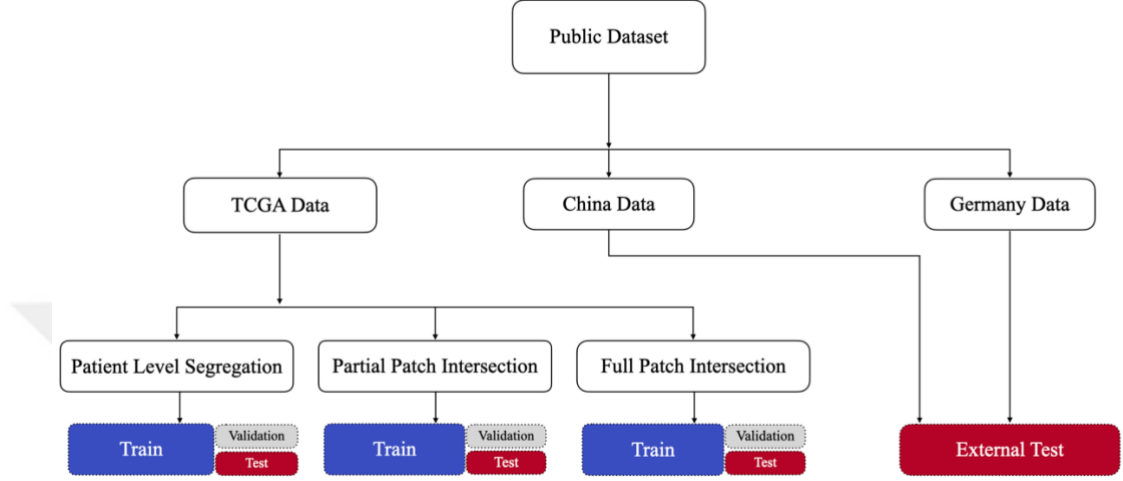


Figure 19. The public dataset consists of three sources: the TCGA Data, copied three times and, for each replication, we segregated the data in distinct ways. The remaining China and Germany DSs were used to generate the external test dataset.

The first approach used a method called 'Patch Level Segregation' with complete intersection. In this method, the sets were formed by randomly assigning patches from each tissue type. Importantly, in this approach, each patient's patches are present throughout all splits. The second approach is also a form of 'Patch Level Segregation', but with partial intersection. In this particular situation, there were 44 patients in all splits; 17 patients were present in the train set, and 12 were for each validation and test set. Ultimately, the third approach used a different method called 'Patient Level Segregation' without any overlap between patients. In this method, we selected a set of patients in each split. More precisely, 61 patients were used for the training set, whereas 24 patients were assigned equally to the validation and test sets. Figure 20 illustrates the areas of overlap between these sets.

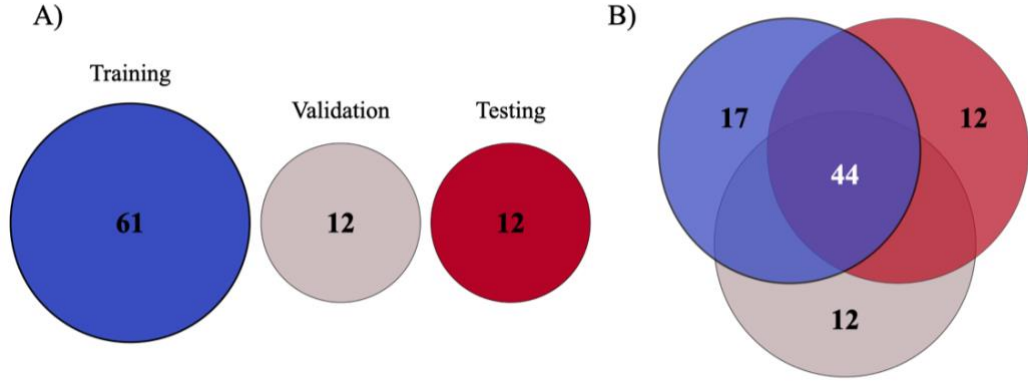


Figure 20. A) Patient level segregation where each patient data only belongs to one split. B) Patch level segregation with partial intersection shows 44 patients shared among the splits, 17 only for training, and 12 for each validation and testing.

3.4.2 Deep learning models

For each segregation mechanism, we trained five models: two CNN models, one from scratch and one following the transfer learning approach; models are VGG19 (Simonyan & Zisserman, 2014) and ResNet50 (He, Zhang, Ren, & Sun, 2016), in addition to a Vision Transformer (ViT) model trained following the transfer learning approach by training only the classification layer weights, owing to computational limitations. We obtained the model architectures from the PyTorch hub. After that, we fine-tuned them by setting the number of output classes in the last classification layer to nine. Table 5 compares the models we have used in this study in terms of size and number of parameters.

Table 5

Number of Parameters and Size in MB for DL Models Used on Medical Images Classification Tasks.

Model	Size (MB)	Number of Parameters
VGG19	549	143.7 M
ResNet50	98	25.6 M
ViT	345.70	85.5 M

The incorporation of diverse models was intended to emphasize the primary influence of the data segregation mechanism in deciding the model's performance on never seen data. We will discuss the results in detail in Chapter 4.

We trained the models using the same hyper-parameters, including a batch size of 32 and a fixed learning rate of 0.0001 for 30 epochs. The optimization technique employed was SGD with a momentum value of 0.9, and the loss function used was a Cross Entropy Loss. We performed the training on a desktop workstation with two NVIDIA GeForce RTX 3080 GPUs.

3.5 Model and Data Availability

The code and DS used in this thesis are made publicly available to ensure the reproducibility of the results presented. The segmentation, diagnosis, and prognosis models code can be accessed at the following GitHub repository: https://github.com/ibsalameh95/DL_NOVEL_PROG_OSCC.

Additionally, the DS utilized in this research was massive; thus, we couldn't upload the WSIs. Instead, we provided the cropped patches to train the models, which are publicly accessible from <https://zenodo.org/records/10658626>. The pre-processing scripts required to prepare the data for training and evaluation are provided in the same GitHub repository.

Chapter 4

Results

4.1 DL Model Segments Out OSCC Slides Successfully

We used two kinds of histopathological images, with stain-normalization and without stain-normalization, over three segments, SEG1, SEG2, and SEG3, to train two DL model architectures: Resnet50 and ViT. This process yielded 12 segmentation models, from which we selected the best one to be used in the subsequent phases. The objective was to classify the images into nine distinct categories. The Resnet50 models achieved a mean accuracy of 97.5% across all nine classes on the test splits, while the ViT models had an average accuracy of 92.8% on the test splits. The Resnet50 model trained using SEG2 without stain normalization achieved an AUROC value of 0.9999 with a 95% confidence interval (CI) of [0.9998-0.9999] and an AUPRC of 0.9992 with a 95% confidence interval (CI) of [0.9998-0.9999] for TUM class, while the ViT model achieved an AUROC value of 0.9988 with a 95% CI of [0.9986-0.9991] and an AUPRC of 0.9944 with a 95% CI of [0.9986-0.9991] for the same class on the same DS.

4.1.1 DL models performance and segmentation model selection

Table 6 presents the accuracies of the 12 models, with the CNN models demonstrating superior performance in training and testing splits compared to ViT models among the whole segregations. This performance difference can be attributed to the ViT models' general requirement for a wider range of images (Dosovitskiy, et al., 2020). The differences in performance between similar models trained on different datasets, can be attributed to the unique tumor features present in each patients' group, that were not included in the training set used for that segregation.

Table 6

The Models' Accuracies on Training, Validation and Testing DSs, for the Three Data Segregations Using the Stain-Normalized (SN) and the Non-Stain-Normalized (NSN) Histopathological Images.

SEGREGATION	S. Normalization	Train %	Valid %	Test %
Resnet 50				
SEG 1	SN	0.9991	0.9919	0.9788
	NSN	0.9985	0.9915	0.9789
SEG 2	SN	0.9992	0.9946	0.9762
	NSN	0.9986	0.9905	0.9910
SEG 3	SN	0.9991	0.9965	0.9546
	NSN	0.9985	0.9924	0.9714
ViT				
SEG 1	SN	0.9694	0.9356	0.9330
	NSN	0.9657	0.9386	0.9253
SEG 2	SN	0.9669	0.9489	0.9172
	NSN	0.9626	0.9577	0.9603
SEG 3	SN	0.9669	0.9785	0.8973
	NSN	0.9632	0.9785	0.9315

In the following subsection, we will analyze the performance of the models by doing the McNemar's test (McNemar, 1947) on the 12 models. The purpose is to identify the most optimal model to be utilized in the subsequent phases. Figure 21 illustrates the flow of the segmentation model selection.

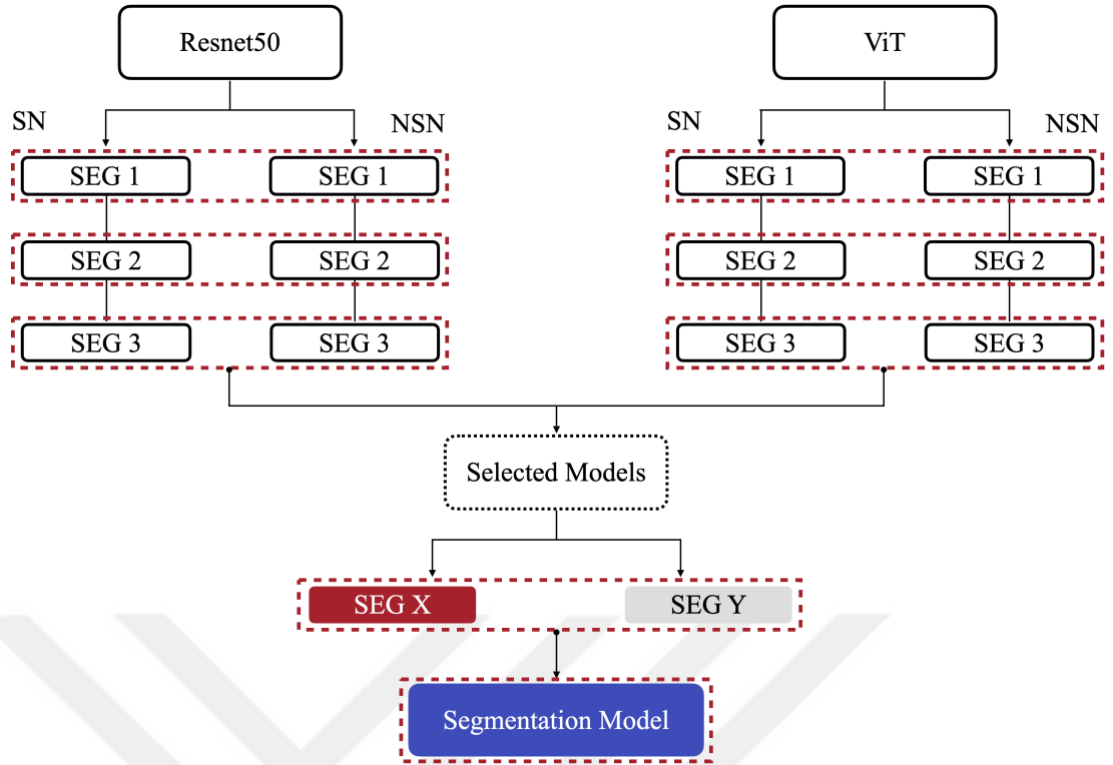


Figure 21. Initially, McNemar's test was used to compare the Resnet50 models trained on identical data segregation. Similarly, we conducted the same test for the ViT models. Finally, the same test was conducted between the models selected in the first round and one of the models was chosen to serve as the segmentation model for the subsequent phases.

4.1.2 Effects of stain normalization in the performance of segmentation models

To evaluate the effect of stain normalization on the model's ability to detect the TUM classes, we employed McNemar's test (McNemar, 1947). between the models trained on the stain normalized (SN) and non-stain normalized (NSN) of the identical segregations. We assessed three classes: NOR, STR, and TUM. We considered the patch as correctly classified if the actual label matched the predicted label. While, misclassified patches comprised TUM patches that were predicted as another tissue type, as well as NOR and STR patches that were predicted as TUM.

4.1.2.1 Resnet50 segmentation model selection

The p-values and the number of misclassified patches for the test results between Resnet50 models are shown in Table 7. SEG3 results showing the worst outcomes, due to the absence of essential tumor characteristics in their training DS. A closer examination of the Confusion Matrices (CMs) and Receiver Operating Characteristics (ROCs) in Appendices section **Error! Reference source not found.**, for the three

segregations (SEG1, SEG2, and SEG3) reveals that models faced challenges in identifying TUM more than other classes. As a result, we selected the SEG2 NSN model, since it has shown the lowest number of misclassified patches among the other segregations and a significant difference compared to SEG2 SN.

Table 7

McNemar's Test Results Between Resnet50 Models Trained on The Same DS With Stain Normalization Applied (SN) and Without (NSN).

Model A	Model B	P Value	Both Incorrect	A is Incorrect	B is Incorrect	Both Correct
SEG1 SN ⁶	SEG1 NSN ⁷	7.488e-06	33	97	43	17087
SEG2 SN	SEG2 NSN	1.587e-31	108	253	49	23607
SEG3 SN	SEG3 NSN	9.338e-07	937	287	180	18038

4.1.2.2 ViT segmentation model selection

The McNemar's test results in Table 8 indicate that the ViT models exhibit similar behavior with Resnet50. Yet, the models trained on the NSN DS demonstrate a significant difference in their ability to discriminate TUM patches.

Table 8

McNemar's Test Results between ViT Models Trained on the Same Data with Stain Normalization Applied (SN) and Without (NSN).

Model A	Model B	P Value	Both Incorrect	A is Incorrect	B is Incorrect	Both Correct
SEG1 SN	SEG1 NSN	0.000515	460	201	278	15687
SEG2 SN	SEG2 NSN	0.002528	242	181	127	23287
SEG3 SN	SEG3 NSN	1.7491e-05	1261	390	278	17274

Considering the findings above and the test accuracy results presented in Table 6, which indicate a slightly higher accuracy of the ViT SEG2 NSN model, we have selected the SEG2 NSN model to be compared with the chosen Resnet50 model. Subsequently, we used McNemar's test between the selected models. Table 9 displays the test results between those models.

⁶ Stain Normalized.

⁷ Non- Stain Normalized.

Table 9

McNemar's Test Results Between Resnet50 and ViT Models trained on SEG2 with No Stain Normalization Applied (NSN).

Model A	Model B	P Value	Both Incorrect	A is Incorrect	B is Incorrect	Both Correct
SEG2 NSN	SEG2 NSN	5.2406e-23	39	114	321	23481

Table 10 and Table 11 also provide a comprehensive overview of those models' classification performances. They reveal that the Resnet50 model metrics were slightly higher than the ViT model trained on the same DS among all classes. Additionally, we have provided the ROCs and CMs for all of the models in the Appendices section **Error! Reference source not found.** for further analysis.

Table 10

Resnet50 SEG2 Non-Stain Normalized (NSN) Model's Classification Report, Providing Key Metrics Including Precision, Recall, F1-score, and Support for Each Class in the Test Split.

Class	Precision	Recall	F1-Score	Support
ADI	0.99	0.99	0.99	4596
BAC	1.00	1.00	1.00	3441
DEB	1.00	0.99	0.99	2546
GLD	0.98	1.00	0.99	5274
LYM	0.99	0.99	0.99	3103
MUS	0.99	0.99	0.99	6315
NOR	1.00	1.00	1.00	9692
STR	0.99	0.99	0.99	7720
TUM	0.99	0.98	0.99	6722
Accuracy			0.99	49409

Table 11

ViT SEG2 Non-Stain Normalized (NSN) Model's Classification Report, Providing Key Metrics Including Precision, Recall, F1-score, and Support for Each Class in the Test Split.

Class	Precision	Recall	F1-Score	Support
ADI	0.97	0.97	0.97	4596
BAC	1.00	1.00	1.00	3441
DEB	0.94	0.93	0.94	2546
GLD	0.98	0.98	0.98	5274
LYM	0.98	0.96	0.97	3103
MUS	0.95	0.89	0.92	6315
NOR	0.98	0.96	0.97	9692
STR	0.91	0.98	0.94	7720
TUM	0.97	0.97	0.97	6722
Accuracy			0.96	49409

After thoroughly analyzing McNemar's test findings, classification metrics, ROCs, PRCs and CMs, we have carefully chosen the Resnet50 model, which has been trained on SEG2 NSN DS, to generate the segmentation maps for our acquired WSIs.

4.1.3 WSI segmentation map creation

We employed the selected model to examine each WSI at the magnification level one, with crops of size 224*224 to meet the model requirements; the resolution rate at that level is 0.5 micro/pixel, while the corresponding OTSU mask's resolution rate is 16 micro/pixel. Therefore, the sliding window must have different widths over the mask and the WSI. Based on the given information, we can calculate the resolution rate of 32 according to Eq.2 which means that each 7*7 pixel in the mask represents 224*224 pixels in the WSI according to Eq.3. Considering the sliding window stride which was 96 px, we calculated the mask level correlation of 3 based on Eq.4. The correlation value was used to reduce the images dimensions during the diagnosis model's DS generation. Figure 22 illustrates a WSI is segmented into nine classes.

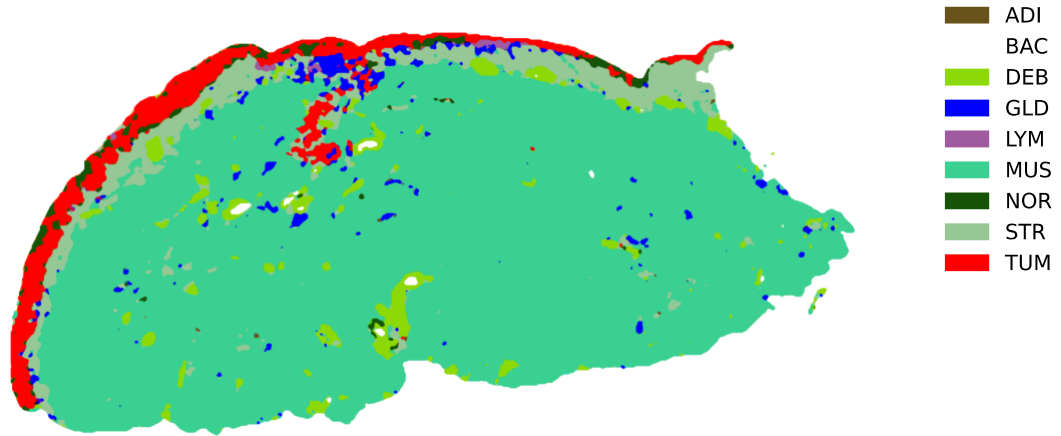


Figure 22. A segmented WSI with colors showing different areas in the tissue generated by Resnet50 segmentation model trained on SEG2 non-stain normalized (NSN) DS.

4.1.4 WSI JSON annotations

Figure 23 displays the WSI JSON annotations file as it appears after importing it to QuPath. The segmentation file was created by utilizing the segmentation map generated by the model, and we applied the previously mentioned image processing techniques to extract the region coordination and turn it into a JSNO file format readable by QuPath.

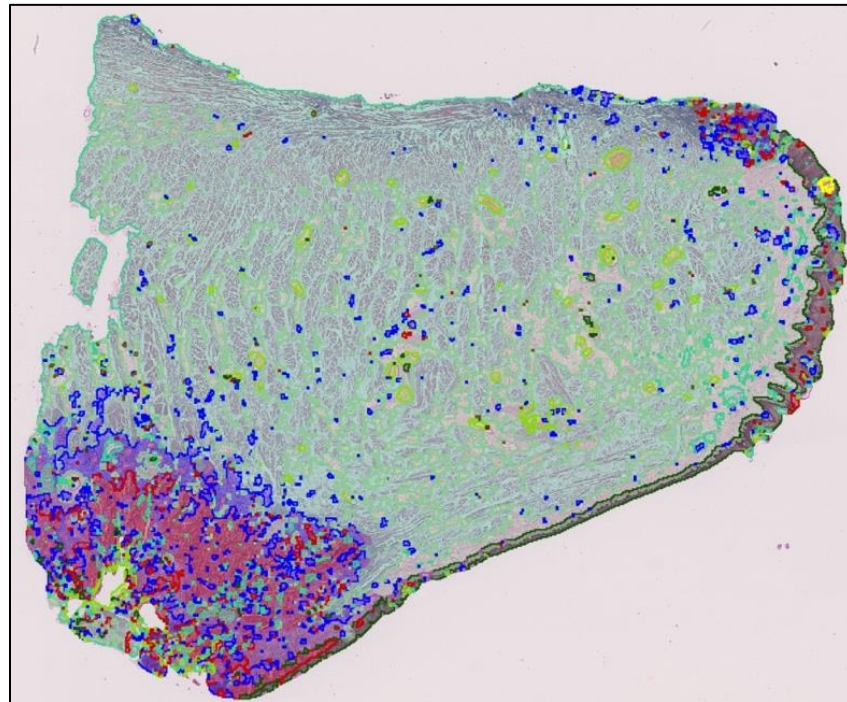


Figure 23. JSON annotations file obtained from the segmentation map and imported to QuPath.

4.2 DL Model Discriminates Tumor vs Non-Tumor

The segmentation model selected in the previous step has been utilized to generate the DS to train this model. We acquired the labels from the original slide labels provided by the pathologist. Our LeNet5 diagnosis model has demonstrated remarkable performance distinguishing between tumor and non-tumor OSCC slides, highlighting our segmentation model's capability in accurately identifying TUM tissue types. The model, with its outstanding precision, exhibited an AUROC of 0.990 with a 95% CI of [0.967-1.000], and an AUPRC of 0.982 with a 95% CI of [0.967-1.000] on the testing split. Table 12 presents the classification report for this diagnosis model. In Appendices section **Error! Reference source not found.**, we added the ROC, PRC and the CM for more details of the model's performance.

Table 12

The Classification Report for LeNet5 Diagnosis Model's Performance, Providing Key Metrics Including Precision, Recall, F1-score, and Support for Each Class in the Test Split.

Class	Precision	Recall	F1-Score	Support
Tumor	0.94	0.97	0.96	34
Non-Tumor	0.94	0.89	0.91	18

4.3 OSCC Prognosis Can Be Predicted from Brightfield and Confocal Images

We generated three models trained on three kinds of data: the first model was trained using the ROIs obtained from CF slides, the second model was trained using the ROIs obtained from BF slides, and finally, we combined both image types to train the third model. The model trained using CF images had an AUROC of 0.942 with a 95% CI of [0.848-1.000], and an AUPRC of 0.913 with a 95% CI of [0.848-1.000]; the BF images showed an area of improvement in the model's performance with AUROC of 0.950 with a 95% CI of [0.861-1.000], and an AUPRC of 0.963 with a 95% CI of [0.861-1.000], while CF and BF combined model achieved the highest accuracy with AUROC of 0.973 with a 95% CI of [0.923-1.000], and an AUPRC of 0.971 with a 95% CI of [0.923-1.000], Table 13 shows the classification report for each of the models. In Appendices section **Error! Reference source not found.**, we added the ROCs, PRCs and CMs for further analysis and understanding.

Table 13

The Classification Report for the Three MIL Prognosis models' performances, Providing Key Metrics Including Precision, Recall, F1-score, and Support for Each Class in the Test Split.

Class	Precision	Recall	F1-Score	Support
Confocal Model				
Good Prognosis	1.00	0.64	0.78	22
Poor Prognosis	0.71	1.00	0.83	20
Brightfield Model				
Good Prognosis	0.94	0.73	0.82	22
Poor Prognosis	0.76	0.95	0.84	20
Confocal & Brightfield Model				
Good Prognosis	0.91	0.91	0.91	22
Poor Prognosis	0.90	0.90	0.90	20

McNemar's test was applied to analyze the models' performance. Although none of the models showed a significant difference, as listed in Table 14, the model trained on BF and CF images generated fewer incorrect predictions than those trained solely on BF and CF images.

Table 14

McNemar's Test Results for the MIL Models Trained Using the Confocal (CF) and the Brightfield (BF) Images Solely and Combined.

Model A	Model B	P Value	Both Incorrect	A is Incorrect	B is Incorrect	Both Correct
CF & BF	CF	0.288844	2	2	6	32
CF & BF	BF	0.371093	3	1	4	34
CF	BF	1.0	4	4	3	31

4.4 Patient Level Segregation is Crucial in Deep Learning Model Development

A total of 15 DL models resulted from training three DL architectures: VGG19, Resnet50, and ViT, using two learning approaches: from-scratch and transfer learning,

with three data segregations: Full Patch Intersection, Partial Patch Intersection and Patient-Level Segregation. Among the models trained on patient-level segregated data, those trained on full and partial patch intersections exhibited higher accuracy on the internal test split. Figure 24, highlights the variations in models' performances between different segregations. The variation in tumor characteristics among various patients has significantly influenced the models' performances. More specifically, the models trained using patient-level segregation which have the lowest accuracy on the internal test set, tended to make more misclassifications between NOR and TUM patches. In contrast, their performance in other classes seems more effective, as illustrated in the CMs shown in Appendices section **Error! Reference source not found..**

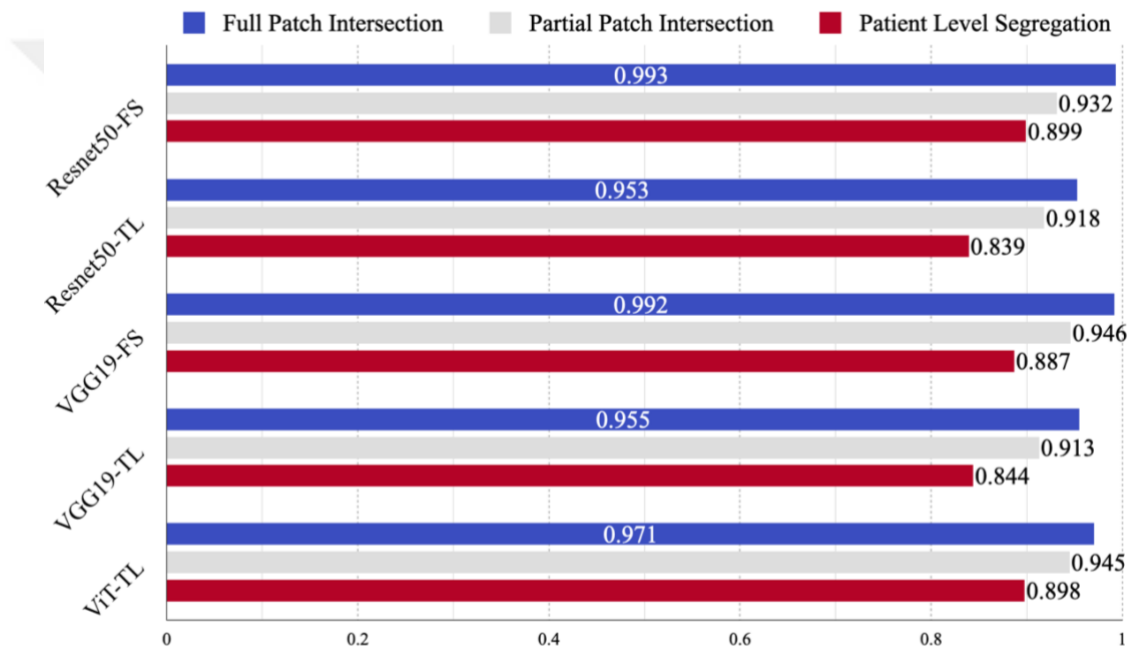


Figure 24. The accuracy results on the test sets for the different DL models trained using different DS segregations. FS: is From Scratch models and TL: refers to Transfer learning.

Subsequently, we derived conclusions based on the external DSs. The results shown in Table 15, indicate that all models achieved close results when tested on previously unseen data. There is a significant disparity between the internal test split and the external DSs results for Full Patch, Partial Patch Intersections and patient level segregation. In Appendices section **Error! Reference source not found..**, we provided the training and validation accuracies, loss vs. epoch for each model and the CMs for the models' inference results on the internal test set, China DS and Germany DS.

Table 15

The Testing DS Accuracy Results of the Different Deep Learning (DL) Models with Different Data Segregations, Show the Impact of Patient-Level Segregation on the Model's Performance on Unseen Data.

MODEL	Method	Train	Valid	Test	Ex1 ⁸	Ex2 ⁹	Ex3 ¹⁰	Ex4 ¹¹
Full Patch Intersection								
Resnet50	FS ¹²	99.9	99.36	99.31	94.65	84.94	84.46	81.86
Resnet50	TL ¹³	94.48	95.57	95.3	90.35	78.24	79.03	75.34
VGG19	FS	99.82	99.23	99.21	96.46	83.28	85.86	82.05
VGG19	TL	98.13	95.53	95.52	90.07	75.33	79.48	71.78
ViT	TL	96.48	96.26	97.07	94.28	83.64	84.93	83.52
Partial Patch Intersection								
Resnet50	FS	99.92	97.36	93.15	94.31	83.21	83.17	81.01
Resnet50	TL	94.73	93.65	91.83	90.41	77.83	78.4	73.85
VGG19	FS	99.8	97.47	94.61	95.98	83.47	86.19	81.84
VGG19	TL	98.29	93.53	91.29	90.04	75.62	79.89	71.85
ViT	TL	96.75	94.33	94.48	93.85	83.87	84.82	82.95
Patient Level Segregation								
Resnet50	FS	99.92	90.87	89.93	94.43	83.07	82.84	82.32
Resnet50	TL	94.99	87.06	83.94	89.03	78.1	78.35	74.25
VGG19	FS	99.84	91.46	88.7	94.46	82.73	84.4	80
VGG19	TL	98.39	83.33	84.41	87.9	73	76.73	68.8
ViT	TL	96.94	91.22	89.78	93.69	83.94	85.46	83.16

How well these models recognize tumors in new patients is essential for determining whether they will work in the real world or couldn't be trusted. Any departure from this standard raises the risk of overfitting, a condition in which models are best at identifying patterns in images they previously observed. The models' performance on

8 External DS 1 (China DS).

9 External DS 2 (Germany DS original stain normalized).

10 External DS 3 (Germany DS non-stain normalized).

11 External DS 4 (Germany DS stain normalized on patches from the internal testing DS).

12 From Scratch.

13 Transfer Learning.

the external DSs in Table 15 demonstrates the fact that all models regardless of their segregation method, are having almost similar performance when it comes to never seen patients.



Chapter 5

Discussion

5.1 Stain Normalization Could Benefit WSIs Obtained by Different Scanners

Our findings have considerable implications for using stain normalization in image analysis and machine learning. The Resnet50 models trained on the NSN DS showed a significant difference from those trained on the same DS that underwent SN. This difference suggests that SN does not necessarily enhance the model's ability to detect TUM classes; in contrast, it may negatively affect the model's performance, especially if the slides were scanned using the same scanner. According to (Macenko, et al., 2009) there may be variations between different scanners. However, the use of stain normalization can assist in minimizing this variance.

5.2 CNNs and ViT Models for Tissue Microenvironment Analysis

Our CNN models have achieved a remarkable accuracy in analyzing the tissue microenvironment, aligning with other studies which have shown the ability of CNN models to analyze and segment WSIs and derive prognosis outputs according to the models outcomes, such as the study of (Kather, et al., 2019) and (Zhao, et al., 2020) who used a CNN model to analyze tissue slides and make prognosis based on the analyzed area and achieved remarkable results. In similar approach, our segmentation model has achieved a high classification accuracy of 0.9910 compared to 0.9610 and 0.9572 testing accuracy achieved by of the previously mentioned studies respectively.

Moreover, the ViT model trained by (Zeid, El-Bahnasy, & Abo-Youssef, 2021) using the same DSs used in (Kather, et al., 2019), highlighted the potential of ViT in analyzing the tissue microenvironment, their ViT classification model achieved an accuracy of 0.9330. In similar approach, our study has provided a ViT classification model for OSCC tissue classification, and the best model has achieved an accuracy of 0.9603.

5.3 CNN is Surpassing ViT in its Ability to Recognize Mixed Tissues

Our DS with its patches of mixed tissues, presented a significant challenge for our models. The task of recognizing the class based on primary tissue type within the area

was crucial. The CNN models, with their superior accuracy in training, validation, and testing sets in SN and NSN versions, outperformed the ViT models. This difference in performance can be attributed to the unique architecture of each model in extracting features from the images. While CNNs excel at capturing local spatial features from the images through the receptive fields, the self-attention mechanism in ViT, though efficient in considering the relationship between all pairs of positions in the input simultaneously, proved to be less effective in capturing important local spatial features, such as in our task to classify the histopathological images.

Figure 25 presents three intriguing instances where the ViT model misidentified patches due to the presence of diverse tissue types. These misclassifications, while unexpected, provide valuable learning opportunities. (A) The Resnet50 model correctly predicted this image as STR, but the ViT model classified it as NOR. Upon examining the attribution mask, we discovered that the ViT model based its prediction on a minor image section corresponding to a NOR area. (B) The Resnet50 model accurately classified the MUS patch as MUS, but the ViT model predicted it as STR. This misclassification, although expected due to the fibrous architecture shared by MUS and STR, occurred because the ViT model overlooked crucial areas for accurate classification. (C) Resnet50 correctly classified the patch as LYM, while ViT focused on different regions within the TUM zones and disregarded the dominant area.

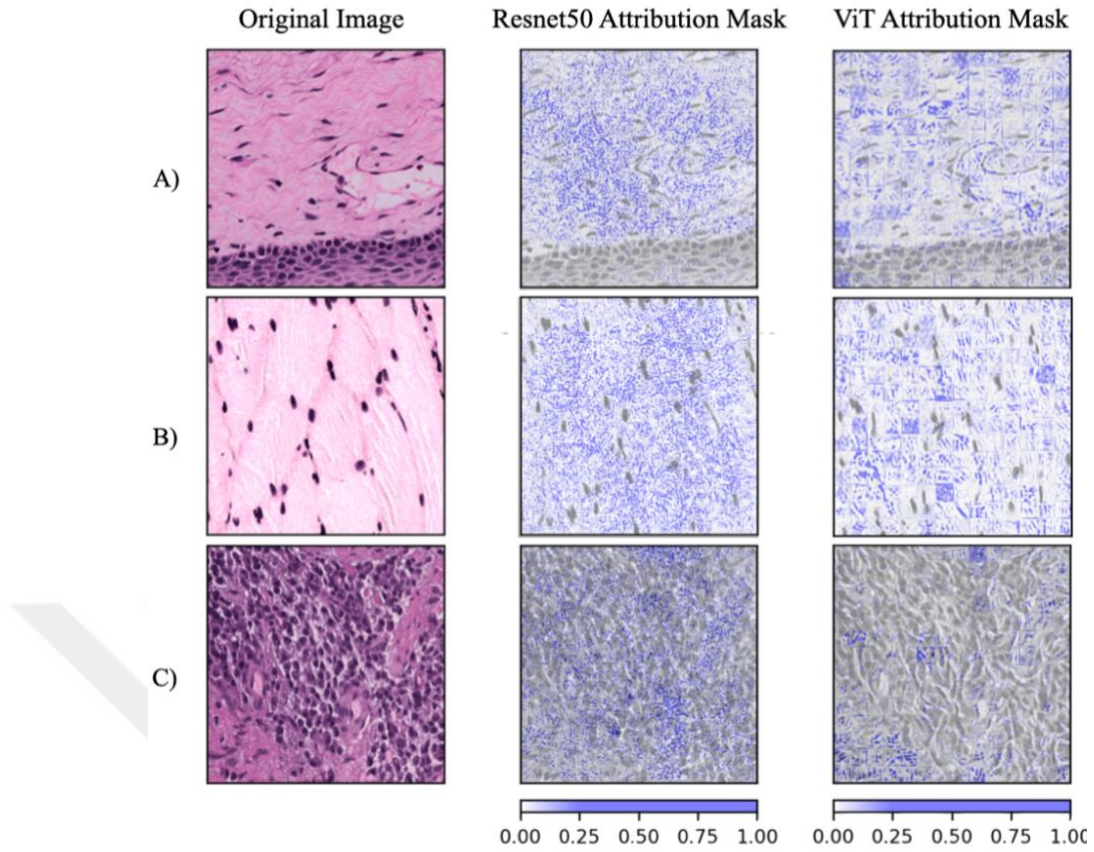


Figure 25. A) The image is labeled as STR. The Resnet50 model correctly predicted it as STR, while the ViT model predicted it as NOR. B) Resnet50 predicted the MUS patch as MUS, but ViT provided a STR prediction. C) Resnet50 correctly identified the patch as LYM, while ViT attention was targeted on the TUM and ignored the majority area.

5.4 Segmentation Model's Probabilities Provide Accurate Diagnosis Results

To provide a diagnosis and prognosis prediction for the WSI, researchers (Kather, et al., 2019) and (Zhao, et al., 2020) utilized a segmentation map generated by a DL segmentation model and applied several image processing to calculate specific areas within the map. Despite its ability to provide helpful information, it may have hidden data that we can utilize. The model's probabilities can be utilized to provide more accurate predictions; in our method, we employed every single probability made by the model while segmenting out the WSI to train the diagnosis model. This approach proved its ability to discriminate between tumor and non-tumor classes and achieved a remarkably high AUROC of 0.990, demonstrating the model's potential for practical application. Taking into account the imbalanced DS, we evaluated the performance using Precision-Recall Curve and our model achieved an AUPRC of 0.982 on the test split.

5.5 ECM Fibril Waveforms as Biomarkers to Predict OSCC Prognosis

Our prognosis model, which effectively distinguishes between slides with good and poor prognosis, was designed to extract and analyze features from CF and BF ROIs. When comparing models' performances on the three sets of images (CF, BF and CF&BF), there was no significant difference in the scope of prediction results but a slight performance improvement noticed in the CF&BF combined; this suggests that both the CF and BF ROIs may include valuable information to strengthen the prognosis prediction. Our hypothesis was validated by examining attribution maps obtained using the Integrated Gradients method (Sundararajan, Taly, & Yan, 2017). We visualized the attribution maps for images from the test split. The attribution maps in Figure 26 highlight the stroma regions, incorporating ECM fibril waveforms as well, which provide clear evidence on how the model made the predictions. This process ensures the scientific integrity of our research, by showing that our model was highly depending on the fibril waveforms within the ECM, to predict the prognosis.

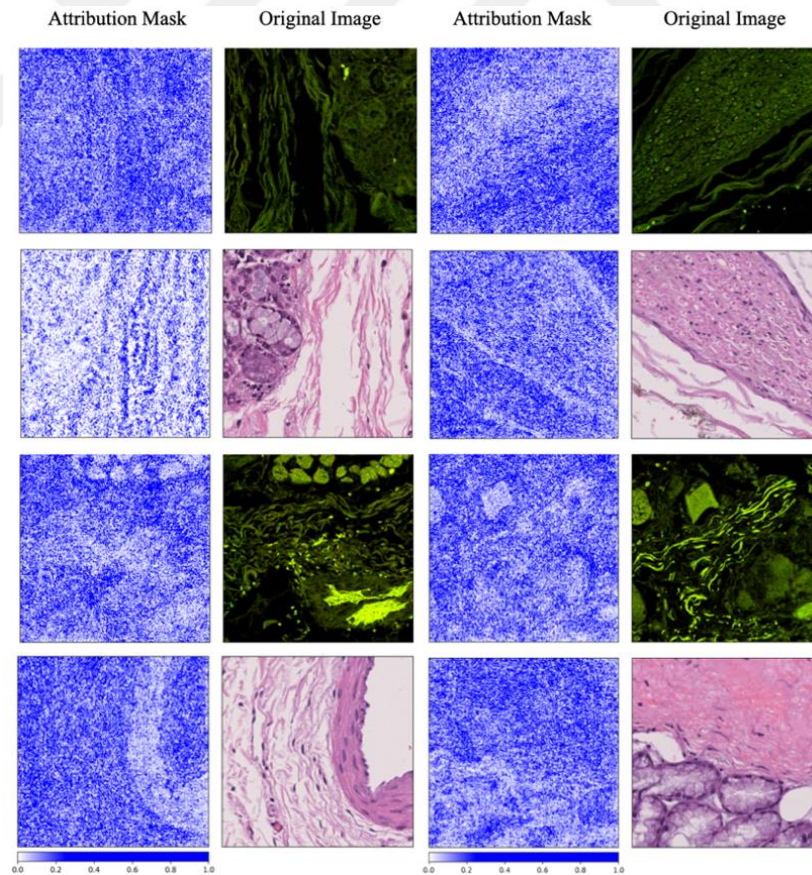


Figure 26. The original images and their attribution mask of the absolute contribution value show that the model tended to make predictions based on the fibril waveforms in the mixed tissue images.

It's crucial to mention that we couldn't compare our prognosis model with any existing state-of-the-art model because we have employed new features to make the prognosis prediction, in addition to using information from the confocal images not previously used in any OSCC prognosis tasks.

5.6 Patient Level Segregation for Trustworthy DL Models'

Our data segregation experiments have provided results that support the idea of segregating data at the patient level. These findings, which are consistent with (Oner, Cheng, Lee, & Sung, 2020), show that using patient-level segregation can lead to a performance drop on the test set, indicating a data leakage. To ensure our models can provide accurate and personalized diagnosis findings, we have implemented the patient-level segregation method in our segmentation, diagnosis, and prognosis models.

5.7 Limitations

Our segmentation map generation method for a WSI uses a procedure that divides the image into smaller patches. The process is achieved by employing the sliding window that moves from left to right and down. However, this practical approach may lead to capturing mixed tissues, that model have never encountered, and may confuse it. To tackle this problem, we have adjusted the stride to 96 pixels, enabling us to display the image from various perspectives. Although this approach can decrease the probability of inaccurate classifications, it cannot eliminate them.

A significant challenge has marked our journey towards a fully automated DL diagnosis system. We've been striving to leverage the segmentation map to accurately detect fibral structures within the segmented WSI. Once identified, these fibral regions can be automatically cropped and forwarded to the diagnosis model. However, we encountered a major hurdle in this phase: the registration of the BF and CF slides to indicate the corresponding regions in the CF slides using the BF segmentation map.

Our DL system has exhibited remarkable performance and promoted using ECM fibril waveforms as biomarkers for OSCC diagnosis. However, our model has not been evaluated using external datasets as we cannot access relevant external data.

5.8 Future Work

The limitations above underscore the need for further research and development to provide a more trustworthy OSCC diagnosis and prognosis DL system. We believe that by working together, we can overcome these challenges and make significant strides in the field. In future works, we plan to overcome the mixed tissue problem faced in the semantic segmentation method by applying different segmentation methods, such as instance segmentation. Additionally, we plan to obtain OSCC WSIs from different cohorts to serve as external validation DS to evaluate our models' performance on never-seen data of the same cancer type.

Finally, to provide a fully automated diagnosis and prognosis model, we are considering using the segmentation maps to help recognize and crop the ROIs from both BF and CF images. This will significantly reduce the work needed from the pathologists to perform the prognosis, as they will only need to review the segmented regions rather than the entire image, thereby increasing efficiency and reducing the risk of human error.

5.9 Conclusion

Using new biomarkers, we have developed a DL system that provides prognosis prediction for OSCC WSIs. Our DL system comprises multiple DL models are segmentation model, diagnosis model, and finally the prognosis model for OSCC WSIs. Our data indicates that the diagnosis model can discriminate between the tumor and non-tumor slides with an AUROC of 0.990 using data obtained by the segmentation model. In addition, the prognosis prediction model has an AUROC of 0.973. Our results show the potential of considering the ECM fibril waveforms as a biomarker to predict OSCC prognosis.

Some limitations in our study prevented us from fully automating the system, which is that the BF and CF images were scanned from different angles and needed to be registered before being able to exploit the BF slide segmentation map in obtaining the needed ROIs for the phase two prognosis model. Although our model has not been tested on external DSs, we cannot claim that it can replace the manual work. However, it could reduce the time and effort needed in the annotation tasks and recognizing ROIs, and it is promising in accelerating the prognosis process.

The prognosis model in the current workflow requires a pathologist to manually annotate the ROIs of the ECM from the BF and CF slides. These annotations are then used to generate the prognosis prediction. The need for automation becomes apparent when considering the time, effort, and domain knowledge involved in this manual process. To address this, we proposed a method that can automatically identify and extract those ROIs by utilizing the BF segmentation map to crop the ROIs from the BF and CF slides.



REFERENCES

- Ahmad, M. A., Eckert, C., & Teredesai, A. (2018, August). Interpretable machine learning in healthcare. In Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics, 559-560.
- Albergotti, W. G., Abberbock, S., Ferris, R. L., Johnson, J. T., Duvvuri, U., & Kim, S. (2017). Defining the Prevalence and Prognostic Value of Perineural Invasion and Angiolymphatic Invasion in Human Papillomavirus-Positive Oropharyngeal Carcinoma. *JAMA otolaryngology - head & neck surgery*, 143(12), 1236–1243. <https://doi.org/10.1001/jamaoto.2017.2019>.
- Anghel, A., Stanisavljevic, M., Andani, S., Papandreou, N., Rüschoff, J. H., Wild, P., . . . Pozidis, H. (2019). A High-Performance System for Robust Stain Normalization of Whole-Slide Images in Histopathology. *Frontiers in medicine*, 6, 193 <https://doi.org/10.3389/fmed.2019.00193>.
- Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., . . . Hamilton, P. W. (2017). QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1), 16878 <https://doi.org/10.1038/s41598-017-17204-5>.
- Bradski, G. (2000). The OpenCV Library. Dr. Dobb's, *Journal of Software Tools*.
- Cancer.Net. (2022, 11). Genes and Cancer. Retrieved from Cancer.Net: <https://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genes-and-cancer>
- Carreras-Torras, C., & Gay-Escoda, C. (2015). Techniques for early diagnosis of oral squamous cell carcinoma: Systematic review. *Medicina oral, patologia oral y cirugia bucal*, 20(3), e305–e315. <https://doi.org/10.4317/medoral.20347>.
- CDC. (2022, December 8). Cancer Characteristics, Definitions, and Recent Investigations. Retrieved from Centers for Disease Control and Prevention: <https://www.cdc.gov/ncch/cancer-environment/guidelines/cancer-characteristics.html>

- Chai, R. L., Rath, T. J., Johnson, J. T., Ferris, L., R., Kubicek, G. J., . . . Branstetter, B. F. (2013). Accuracy of computed tomography in the prediction of extracapsular spread of lymph node metastases in squamous cell carcinoma of the head and neck. *JAMA otolaryngology - head & neck surgery*, 139(11), 1187–1194. <https://doi.org/10.1001/jamaoto.2013.4491>.
- Chen, S. W., Zhang, Q., Guo, Z. M., Chen, W. K., Liu, W. W., Chen, Y. F., . . . & Song, M. (2018). Trends in clinical features and survival of oral cavity cancer: fifty years of experience with 3,362 consecutive cases from a single institution. *Cancer management and research*, 10, 4523–4535. <https://doi.org/10.2147/CMAR.S171251>.
- Dimitriou, N., Arandjelović, O., Harrison, D. J., & Caie, P. D. (2018). A principled machine learning framework improves accuracy of stage II colorectal cancer prognosis. *NPJ digital medicine*, 1(1), 52.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2020). An image is worth 16x16 words. *arXiv preprint*, 2010.11929.
- Dumoulin, V., & Visin, F. (2018). A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv*, 1603, 07285.
- E., P. P. (2003). The World Oral Health Report 2003: continuous improvement of oral health in the 21st century--the approach of the WHO Global Oral Health Programme. *Community dentistry and oral epidemiology*, 31(1), 3-23. <https://doi.org/10.1046/j..2003.com122.x>.
- Eskander, A., Dziegielewska, P. T., Patel, M. R., Jethwa, A. R., Pai, P. S., Silver, N. L., . . . St John, M. (2024). Oral Cavity Cancer Surgical and Nodal Management: A Review From the American Head and Neck Society. *JAMA otolaryngology - head & neck surgery*, 150(2), 172-178. <https://doi.org/10.1001/jamaoto.2023.4049>.
- Ferlay, J., Parkin, D. M., & Steliarova-Foucher, E. (2010). Estimates of cancer incidence and mortality in Europe in 2008. *European journal of cancer*, 46(4), 765–781.

- Fives, C., Nae, A., Roche, P., O'Leary, G., Fitzgerald, B., Feeley, L., & Sheahan, P. (2017). Impact of mandibular invasion on prognosis in oral squamous cell carcinoma four centimeters or less in size. *The Laryngoscope*, 127(4), 849–854. <https://doi.org/10.1002/lary.26211>.
- Fu, Q., Chen, Y., Li, Z., Jing, Q., Hu, C., Liu, H., & Xiong, X. (n.d.). A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: A retrospective study. *EClinicalMedicine*, 27.
- Gao, J., Yang, Y., Lin, P., & Park, D. S. (2018, March 4). Computer Vision in Healthcare Applications. *Journal of Healthcare Engineering*, 2018, 5157020.
- Hamilton, P. W., Bankhead, P., Wang, Y., Hutchinson, R., Kieran, D., McArt, D. G., & Salto-Tellez, M. (2014). Digital pathology and image analysis in tissue biomarker research. *Methods*, 70(1), 59-73.
- Hanahan, D. (2022). Hallmarks of cancer: new dimensions. *Cancer discovery*, 8(1), 31-46.
- Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5), 646-674.
- Harder, N., Schönmeier, R., Nekolla, K., Meier, A., Brieu, N., Vanegas, C., & Schmidt, G. (2019). Automatic discovery of image-based signatures for ipilimumab response prediction in malignant melanoma. *Scientific reports*, 9(1), 7449.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.
- Johnson, D. E., Burtress, B., Leemans, C. R., Lui, V. W., Bauman, J. E., & Grandis, J. R. (2020). Head and neck squamous cell carcinoma. *Nature reviews Disease primers*, 6(1), 92.
- Kühl, N., Goutier, M., Baier, L., Wolff, C., & Martin, D. (2022). Human vs. supervised machine learning: Who learns patterns faster? *Cognitive Systems Research*, 76, 78-92.

- Kather, J. N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C. A., & Halama, N. (2019). Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1), e1002730.
- Kokhlikyan, N. a. (2020). Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.
- Koning, d., J., H., Aalst, v. d., M., C., Jong, d., A., P., . . . Oo. (2020). Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *The New England journal of medicine*, 382(6), 503–513. <https://doi.org/10.1056/NEJMoa1911793>.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90.
- Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. in *Proceedings of the IEEE*, 86(11), 2278–2324. [10.1109/5.726791](https://doi.org/10.1109/5.726791).
- Lester, S. C., Bose, S., Chen, Y. Y., Connolly, J. L., de Baca, M. E., Fitzgibbons, P. L., . . . Pathologists, C. o. (2009). Protocol for the examination of specimens from patients with invasive carcinoma of the breast. *Archives of pathology & laboratory medicine*, 133(10), 1515–1538. <https://doi.org/10.5858/133.10.1515>.
- Leusink, F. K., van Es, R. J., de Bree, R., Baatenburg de Jong, R. J., van Hooff, S. R., Holstege, F. C., . . . Takes, R. P. (2012). Novel diagnostic modalities for assessment of the clinically node-negative neck in oral squamous-cell carcinoma. *The Lancet. Oncology*, 13(12), e554–e561. [https://doi.org/10.1016/S1470-2045\(12\)70395-9](https://doi.org/10.1016/S1470-2045(12)70395-9).
- Li, W., L. J., Sarma, K. V., Ho, K. C., Shen, S., Knudsen, B. S., . . . Arnold, C. W. (2019). Path R-CNN for Prostate Cancer Diagnosis and Gleason Grading of Histological Images. *IEEE transactions on medical imaging*, 38(4), 945–954. <https://doi.org/10.1109/TMI.2018.2875868>.

- Liu, Y., & Pantanowitz, L. (2019). Digital pathology: Review of current opportunities and challenges for oral pathologists. *Journal of Oral Pathology & Medicine*, 48(4), 263-269.
- Ma, W., Wang, J., Yu, L., Zhang, X., Wang, Z., Tan, B., & Cheng, Y. (2012). Tumor-Stroma Ratio Is an Independent Predictor for Survival in Esophageal Squamous Cell Carcinoma. *Journal of Thoracic Oncology*, 7(9), 1457-1461.
- Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., & Guan, X. (2009). A method for normalizing histology slides for quantitative analysis. *IEEE International Symposium on Biomedical Imaging*, 10, 1107.
- Mascitti, M., Togni, L., Caponio, V. C., Zhurakivska, K., Bizzoca, M. E., Contaldo, M., . . . Santarelli, A. (2022). Lymphovascular invasion as a prognostic tool for oral squamous cell carcinoma: a comprehensive review. *International journal of oral and maxillofacial surgery*, 51(1), 1–9. <https://doi.org/10.1016/j.ijom.2021.03.007>.
- McClintock, D. S., Abel, J. T., & Cornish, T. C. (2022). Whole Slide Imaging Hardware, Software, and Infrastructure. . *Whole Slide Imaging: Current Applications and Future Directions*, 23-56.
- McNemar, Q. (1947). Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2), 153-157.
- Michikawa, C., Uzawa, N., Kayamori, K., Sonoda, I., Ohyama, Y., Okada, N., . . . Amagasa, T. (2012). Clinical significance of lymphatic and blood vessel invasion in oral tongue squamous cell carcinomas. *Oral oncology*, 48(4), 320–324. <https://doi.org/10.1016/j.oraloncology.2011.11.014>.
- Mohan, H. (2010). *Textbook of pathology*. SPRINGER.
- Montalto, M. C. (2016). An industry perspective: An update on the adoption of whole slide imaging. *Journal of pathology informatics*, 7(1), 18.
- Narayanan, D. L., Saladi, R. N., & Fox, J. L. (2010). Ultraviolet radiation and skin cancer. *International journal of dermatology*, 49(9), 978–986. <https://doi.org/10.1111/j.1365-4632.2010.04474.x>.

- NCI. (2024, May 5). What Is Cancer? Retrieved from cancer: <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>
- Nearchou, I. P., Lillard, K., Gavriel, C. G., Ueno, H., Harrison, D. J., & Caie, P. D. (2019). Automated analysis of lymphocytic infiltration, tumor budding, and their spatial relationship improves prognostic accuracy in colorectal cancer. *Cancer immunology research.*, 7(4), 609-620.
- NIH. (2023, January 17). How Cancer Is Diagnosed. Retrieved from National Cancer Institute: <https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis>
- Nives, S.-P. (2003). Tumor suppressor gene E-cadherin and its role in normal and malignant cells. *Cancer Cell Int.*, 3(1), 17. <https://doi.org/10.1186/1475-2867-3-17>.
- ONER, M. U. (2021, 7 30). NOVEL MULTIPLE INSTANCE LEARNING MODELS FOR DIGITAL HISTOPATHOLOGY. ScholarBank@NUS Repository.
- Oner, M. U., Cheng, Y.-C., Lee, H. K., & Sung, W.-K. (2020). Training machine learning models on patient level data segregation is crucial in practical clinical applications. *medRxiv*, doi:10.1101/2020.04.23.20076406.
- Oner, M. U., Ng, M. Y., Giron, D. M., Chen Xi, C. E., Yuan Xiang, L. A., Singh, M., . . . Lee, H. K. (2022). An AI-assisted tool for efficient prostate cancer diagnosis in low-grade and low-volume cases. *Patterns (New York, N.Y.)*, 3(12), 100642. <https://doi.org/10.1016/j.patter.2022.100642>.
- OTSU, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62-66.
- Pallua, J. D., Brunner, A., Zelger, B., Schirmer, M., & Haybaeck, J. (2020). The future of pathology is digital. *Pathology-Research and Practice*, 216(9), 153040.
- Pantanowitz, L., Sharma, A., Carter, A. B., Kurc, T., Sussman, A., & Saltz, J. (2018). Twenty Years of Digital Pathology: An Overview of the Road Travelled, What is on the Horizon, and the Emergence of Vendor-Neutral Archives. *Journal of pathology informatics*(9), 40. https://doi.org/10.4103/jpi.jpi_69_18.

- Phillips, J. L., & Currow, D. C. (2010). Cancer as a chronic disease. *Collegian*, 17(2), 47-50.
- Roser, M., & Ritchie, H. (2024, March 12). Cancer. Our World in Data, <https://ourworldindata.org/cancer>. Retrieved from Our World in Data.: <https://ourworldindata.org/cancer#:~:text=The%20Global%20Burden%20of%20Disease,deaths%20worldwide%20were%20from%20cancer>.
- Seif, G. (2018, November 19). Handling Imbalanced Datasets in Deep Learning. Retrieved from Towards Data Science: <https://towardsdatascience.com/handling-imbalanced-datasets-in-deep-learning-f48407a0e758>
- Seitz, H. K., Pelucchi, C., Bagnardi, V., & La Vecchia, C. (2012). Epidemiology and pathophysiology of alcohol and breast cancer. *Alcohol and alcoholism* (Oxford, Oxfordshire), 47(3), 204–212. <https://doi.org/10.1093/alcalc/ags011>.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*, 1409-1556.
- Song, Q., Zhao, L., Luo, X., & Dou, X. (2017). Using deep learning for classification of lung nodules on computed tomography images. *Journal of healthcare engineering*.
- Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International conference on machine learning*, (pp. 3319-3328. PMLR.).
- Tang, H., Li, G., Liu, C., Huang, D., Zhang, X., Qiu, Y., & Liu, Y. (2022). Diagnosis of lymph node metastasis in head and neck squamous cell carcinoma using deep learning. *Laryngoscope investigative otolaryngology*, 7(1), 161–169. <https://doi.org/10.1002/lio2.742>.
- van Dijk, M. R., Steyerberg, E. W., & Habbema, J. D. (2008). A decision-analytic approach to define poor prognosis patients: a case study for non-seminomatous germ cell cancer patients. *BMC medical informatics and decision making*, 8(1), <https://doi.org/10.1186/1472-6947-8-1>.

- Van, d. T., J. , G., & Taylor, C. R. (2010). A brief history of pathology: preface to a forthcoming series that highlights milestones in the evolution of pathology as a discipline. *Virchows Archiv*, 457, 3-10.
- Vigneswaran, N., & Williams, M. D. (2014). Epidemiologic trends in head and neck cancer and aids in diagnosis. *Oral and maxillofacial surgery clinics of North America*, 26(2), 123-141.
- Vineis, P., Alavanja, M., Buffler, P., Fontham, E., Franceschi, S., Gao, Y. T., . . . Doll, R. (2004). tobacco and cancer: recent epidemiological evidence. *Journal of the National Cancer Institute*, 96(2), 99–106. <https://doi.org/10.1093/jnci/djh014>.
- Weinberg, R. A. (1996, September). How Cancer Arises. *Scientific American*, 275(3), 62–70.
- WHO. (2022, February 3). Cancer. Retrieved from World Health Organization: <https://www.who.int/news-room/fact-sheets/detail/cancer>
- Wong, R. J., Keel, S. B., Glynn, R. J., & Varvares, M. A. (2000). Histological pattern of mandibular invasion by oral squamous cell carcinoma. *The Laryngoscope*, 110(1), 65-72.
- Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., & Vajda, P. (n.d.). Visual transformers: Token-based image representation and processing for computer vision. (p. preprint arXiv:2006.03677.). arXiv.
- Wu, J., Liang, C., Chen, M., & Su, W. (2016). Association between tumor-stroma ratio and prognosis in solid tumor patients: a systematic review and meta-analysis. *Oncotarget*, 7(42), 68954.
- Yadav, J., & Sharma, M. (2013). A Review of K-mean Algorithm. *Int. J. Eng. Trends Technol*, 4(7), 2972-2976.
- Zeid, M. A., El-Bahnasy, K., & Abo-Youssef, S. E. (2021). Multiclass Colorectal Cancer Histology Images Classification Using Vision Transformers. *Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, 224-230 doi:10.1109/ICICIS52592.2021.9694125.

- Zhang, Y. L., Zhang, Z. S., Wu, B. P., & Zhou, D. Y. (2002, February). Early diagnosis for colorectal cancer in China. *World journal of gastroenterology*, 8(1), 21-25.
- Zhao, K., Li, Z., Yao, S., Wang, Y., Wu, X., Xu, Z., & Liu, Z. (2020). Artificial intelligence quantified tumour-stroma ratio is an independent predictor for overall survival in resectable colorectal cancer. *EBioMedicine*, 61.
- Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, 5(1), 44–53.
- Zhu, V. L., Rand, D. R., Arnold, K. E., Pagedar, N. A., Bayon, R., & Buchakjian, M. R. (2023). Association of Core Biopsy With Extranodal Extension in Surgically Treated Human Papillomavirus-Associated Oropharyngeal Squamous Cell Carcinoma. *JAMA otolaryngology - head & neck surgery*, 149(11), 955–960. <https://doi.org/10.1001/jamaoto.2023.1897>.