People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

Higher School of Computer Science

Artificial Intelligence and Data Science
Second Year Second Cycle

Project Theme:

Improving Breast Cancer Diagnosis Through Classification of Hematoxylin and Eosin Histopathological Images

---

Authors:
Abdelnour Fellah
Abderrahmane Benounene
Adel Abdelkader Mokadem
Meriem Mekki
Yacine Lazreg Benyamina

Supervisors:
Pr. Sidi Mohammed Benslimane
Dr. Nassima Dif

In collaboration with the
Anti-Cancer Center of Sidi Bel Abbes

Academic Year of 2023/2024

Acknowledgment

" First and foremost, we extend our sincerest gratitude to Allah for bestowing upon us the strength and perseverance to bring this project to fruition.

We would like to thank our supervisor, Dr. Nassima DIF, for her exceptional guidance and continuous support throughout the duration of this project. Her expertise and insights were invaluable to our work. Additionally, we would like to thank Pr. Djazia Asma BENCHOUK, for her invaluable knowledge and guidance on breast cancer. Her expertise greatly enriched our project and provided us with a deeper understanding of the medical aspects.

We also wish to thank Pr. Sidi Mohamed BENSLIMANE and the entire pedagogic staff of ESI SBA for providing us with a conducive environment for learning and growth. Their dedication and commitment to our education have been crucial to our academic development.

Lastly, we are profoundly grateful to our parents for their constant encouragement and support, which have been our foundation throughout this journey.

Thank you all for your contributions and support. "

Table of contents

1 Introduction . . . 6
2 Background . . . 8
2.1 Convolutional Neutral Networks . . . 8
2.2 Residual Networks . . . 9
2.3 Attention mechanisms . . . 11
2.4 Vision transformers . . . 11
2.5 Fine tuning & transfer learning . . . 12
2.6 Multi-instance learning . . . 13
3 State of the art . . . 14
4 BRACS Dataset . . . 16
4.1 Data collection and annotation process . . . 16
4.2 Dataset characteristics . . . 16
4.3 Dataset organization . . . 17
5 Approach . . . 19
5.1 Preparing Models For Feature Extraction . . . 20
5.2 Feature Extraction . . . 22
5.2.1 Grid-based feature extraction . . . 22
5.2.2 Feature extraction with patch selection . . . 23
5.3 WSI Classifiers . . . 24
5.3.1 Min-Max attention based classifier . . . 24
5.3.2 Attention-Challenging Multiple Instance Learning . . . 25
5.3.3 Hierarchical Image Pyramid Transformers . . . 27
6 Experiments and results . . . 30
6.1 Feature extractors fine tuning results . . . 30
6.2 WSI Classifiers results . . . 32
7 Information system . . . 35
7.1 Anti Cancer Center . . . 35
7.2 Objective . . . 35
7.3 Overview . . . 35
7.3.1 Web application . . . 35
7.3.2 Inference application . . . 37
8 Conclusion . . . 40
9 References . . . 42

List of Figures

1 Slide Scanning process . . . 7
2 Architecture of a traditional CNN [1]. . . . 8
3 Convolution layer [1]. . . . 8
4 Max-Pooling layer in action [1]. . . . 9
5 Shortcut connection [3]. . . . 10
7 Vision Transformer architecture . . . 11
8 The process of Fine tuning / transfer learning . . . 12
9 The organization of BRACS dataset folders [2]. . . . 18
10 Schematic representation of the general approach. . . . 19
11 Fine-tuning process on BRACS' ROI images. . . . 21
12 Grid-Based Feature Extraction (GFE) [15]. . . . 22
13 Clustering-constrained Attention Multiple instance learning . . . 23
14 Feature extraction with patch selection . . . 24
15 Min-Max attention based classifier (AC) [15]. . . . 24
16 An Overiew of the ACMIL Architecture[17]. . . . 25
17 HIPT Architecture [16]. . . . 28
20 Class Diagram of the information System . . . 35
21 Use Case diagram of the web application . . . 36
22 Home page . . . 36
23 Login page . . . 36
24 Dashboard page . . . 36
25 Patients list page . . . 36
26 Patient information form . . . 37
27 Patient's histology form . . . 37
28 Patient's immunohistochemistry form . . . 37
29 Archived patients page . . . 37
30 Use Case diagram of the inference application . . . 37
31 Login page . . . 38
32 Doctor's patients list page . . . 38
33 Patient's histologies list page . . . 38
34 Selecting a WSI page . . . 39
35 Feature Extraction on the WSI . . . 39
36 Prediction on the WSI . . . 39

List of Tables

1 Comparative table of state of the art methods 15
2 BRACS data distribution according to lesion types and subtypes. 17
3 WSI- and ROI-level split according to the lesion types. 17
4 WSI- and ROI-level split according to the lesion subtypes. 17
5 Data distribution of patches count across different splits and lesion types. 22
6 Data distribution of patches count across different splits and categories. 22
7 Hyperparameter Configurations for ResNet-18 30
8 Hyperparameter Configurations for ResNet-34 31
9 Fine-tuning feature extractors results (patches persepective) 31
10 Fine-tuning feature extractors results (soft-voting) 31
11 Fine-tuning feature extractors results (hard-voting) 32
12 Attention based classifiers results. 33
13 HIPT results. 33

Abstract

Breast cancer detection through the classification of histopathological images is a critical area of research with significant implications for improving diagnostic accuracy. This field has garnered substantial attention due to its potential to aid oncologists and pathologists and afford them the invaluable opportunity to optimize their time and efforts in the diagnosis and early detection of breast cancer.

Medical Artificial Intelligence, such as computer vision and deep learning techniques, play a pivotal role in analyzing and classifying histopathological images. These systems leverage convolutional neural networks (CNNs) and Vision Transformers (ViTs) to process images of breast tissue.

Overall, the classification of histopathological images through medical AI marks a crucial field in medical research, offering the transformative potential to automate diagnostic procedures. By using this technology, oncologists and pathologists can reclaim valuable time and effort, leading to enhanced efficiency and accuracy in diagnosing breast cancer. Moreover, as this field continues to advance, there is hope for even greater outcomes and advancements on the horizon, promising to revolutionize patient care and treatment effectiveness.

1 Introduction

Breast cancer is one of the most prevalent malignancies affecting women worldwide, with approximately 2.3 million new cases diagnosed annually (World Health Organization, 2021). It encompasses a heterogeneous group of diseases characterized by the uncontrolled growth of abnormal cells in the breast tissue. The majority of breast cancers originate in the ducts or lobules of the breast and are classified based on their histological characteristics and molecular profiles. Several factors contribute to the development of breast cancer, including genetic predisposition, hormonal influences, lifestyle factors, and environmental exposures. Early detection of breast cancer is critical for improving patient outcomes and reducing mortality rates. Screening methods such as mammography, clinical breast examination, and breast self-examination are commonly used for early detection. Additionally, biopsy plays a crucial role in the diagnostic process. During a biopsy, a small tissue sample is obtained from the breast, typically under imaging guidance such as ultrasound or stereotactic mammography. This tissue sample is then examined under a microscope to determine the presence of cancer cells and to characterize the type and nature of the tumor. The resulting histopathological images, which come from the biopsy tissue samples, are usually colorized using Hematoxylin and Eosin (H&E) staining, differentiating cellular structures and components within the tissue, aiding pathologists in making accurate diagnoses.Figure 1 depicted below illustrates the procedural steps involved in the digitization of a Whole Slide Image (WSI) originating from a biopsy specimen.

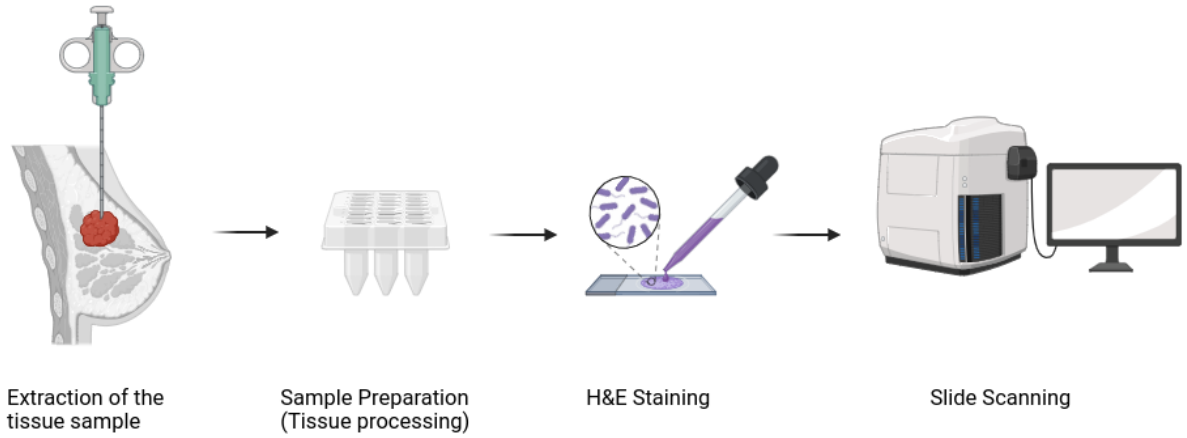


FIGURE 1: Slide Scanning process

This project aims to leverage computer vision and deep learning techniques to enhance the accuracy and efficiency of breast cancer diagnosis. By analyzing histopathological images of breast tissue samples, the goal is to develop a computer-aided diagnostic system capable of identifying malignant cells and distinguishing between different subtypes of breast cancer. Despite significant advancements in breast cancer diagnosis and treatment, several challenges persist in the field of medical AI. One general challenge is the annotation of histopathological images, which is a labor-intensive and time-consuming task requiring expert pathologists to accurately label the images. In our project, the main challenges were related to the size of the histopathological images (gigapixel images). These extremely high-resolution images result in substantial computational complexity, posing significant hardware challenges. Processing and analyzing such large images require considerable computational power and memory, which can be a limiting factor in developing efficient diagnostic systems. The primary objective of this project, undertaken as part of the multidisciplinary project of our 2nd year of the second cycle at ESI SBA and our internship at CAC SBA (Centre Anti Cancer de Sidi Bel Abbes), is to develop a computer-aided diagnostic system capable of accurately classifying histopathological images of breast tissue into three classes: malignant, atypical, and benign. By training deep learning models to recognize and classify different histological features associated with these breast cancer subtypes, evaluating their performance, and integrating the diagnostic system into a workflow system specifically designed for CAC SBA, which includes a database for saving the necessary information of patients, we aim to support pathologists in their decision-making process. We designed and developed an Information System (IS), specifically a workflow system, tailored for CAC SBA (Centre Anti Cancer of Sidi Bel Abbes), to complement the development of the computer-aided diagnostic system for accurately classifying histopathological images of breast tissue. This workflow system serves the dual purpose of facilitating the collection and storage of necessary patient data, as well as automating the process of retrieving patient information. By seamlessly integrating the diagnostic system with the workflow system, the aim was to streamline the diagnostic process and enhance overall efficiency in breast cancer diagnosis at CAC SBA.

2 Background

2.1 Convolutional Neutral Networks

Convolutional Neutral Networks (CNNs : Figure 2) are a type of deep learning architectures that made impressive achievements in various fields such as computer vision and natural language processing, a CNN architecture mainly consists of :

- **Convolutional layers:** Responsible for extracting features from the input data.
- **Pooling layers:** Reduce the size of the feature maps while retaining the most important features.
- **Fully connected layers:** Responsible for decision making, as it takes the feature vector generated by previously mentioned layers and outputs class probabilities.

.

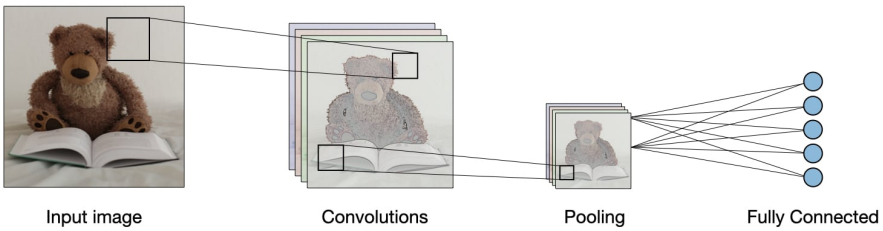


FIGURE 2: Architecture of a traditional CNN [1].

Convolutional layers

Convolutional layers provide a way to automatically extract features from the input data such as images. They consist of learnable filters (also called kernels) which have small widths and heights and the same depth as the input feature map (initially, it is the input image with a depth of 3 corresponding to the three channels R, G, and B).

- During the forward pass, the layer processes its input patch by patch using a sliding window with a width and height identical to the filter's.
- Each time the window moves, it shifts by a particular number of pixels called the stride.
- The output is the dot product between the kernel weights and the patch.
- After repeating this process for each filter, we will stack their results to get a new feature map with a depth equal to the number of filters.

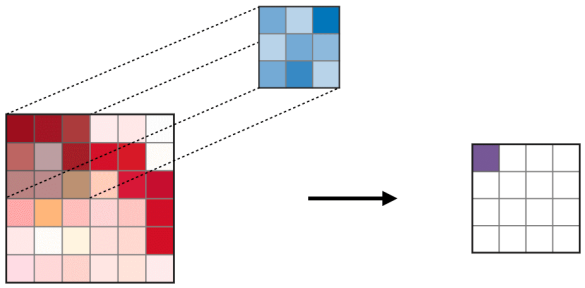


FIGURE 3: Convolution layer [1].

Pooling layers

Pooling layers with no learnable parameters responsible for reducing the spatial dimensions of the input feature map, in terms of width and height, while retaining the most important information, several types of pooling layers exists including: .

- **Max pooling:** takes the maximum of the region.
- **Average pooling:** takes the average of the region.
- **Global max pooling:** takes the maximum value over the entire feature map.
- **Global average pooling:** takes the average value over theentire feature map.
- **L2 pooling:** takes the L2 norm of each region.
- **Fractional max pooling:** takes the maximum value over a randomly chosen subset of the region.

The dimensions of the resulted feature map is calculated the same way as the convolutional layer,because pooling layers also process the input feature map using a moving window.

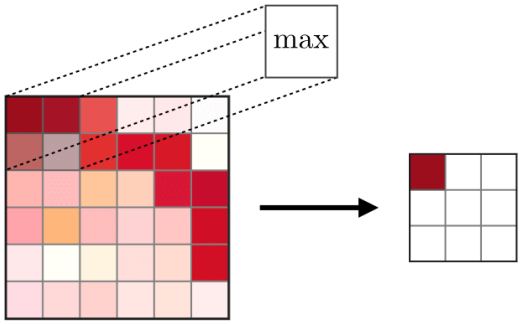
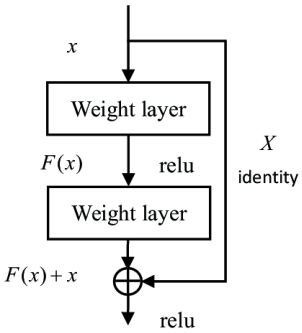


FIGURE 4: Max-Pooling layer in action [1].

2.2 Residual Networks

For complex tasks, deeper networks are preferable, but they present their own challenges, mainly computational complexity, which increases as more layers are added, and the ease of training, as it is proven that deeper networks are generally harder to optimize due to the problem of vanishing or exploding gradient descent.

Residual Networks [2] (ResNet) are a family of architectures that have been proven successful in image recognition tasks. They make use of the concept of deep residual learning, which is achieved using "shortcut connections" (Figure 5) the building block for this type of architecture to ease the training and optimization of deeper networks with less computational complexity.

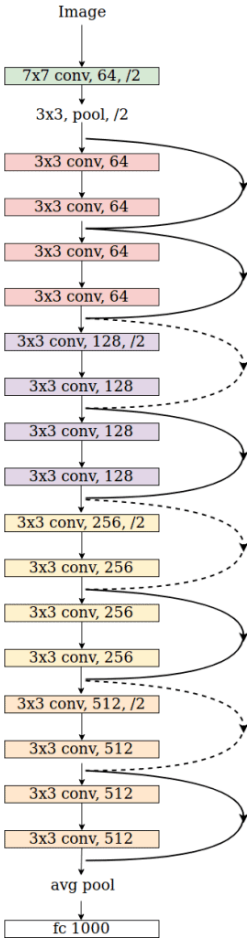


A shortcut connection skips some of the layers in the neural network and feeds the output of one layer as the input to the next layers. In the case of ResNet, the shortcut connections simply perform identity mapping, and their outputs are added to the outputs of the stacked layers. It is defined as:

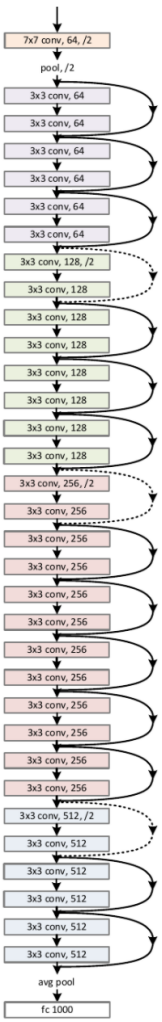
$$y = \mathcal{F}(x, \{W_i\}) + x$$

FIGURE 5: Shortcut connection [3].

ResNet18 (Figure 6a) and **ResNet34** (Figure 6b) are two state-of-the-art residual networks for image classification tasks,they consist of 18 and 34 layers respectively including convolutional layers, batch normalization, and ReLU activation functions and one fully connected layer responsible for outputting classes probabilities.



(A) ResNet-18 Architecture [3]



(B) ResNet-34 [3].

2.3 Attention mechanisms

Attention mechanisms are deep learning techniques,emerged initially to improve computer vision and the encoder-decoder-based neural machine translation system,it is basically a dynamic weight adjustment function based on an attention function $g(x)$ and an input feature map x that is superimposed between the convolutional layers. its role is to tell the next layer of the deep network which features are more or less important, this can be formulated as :.

$$Attention = f(g(x), x)$$

Here g is responsible for generating attention which corresponds to the process of attending to the discriminative regions. $f(g(x), x)$ means processing input x based on the attention $g(x)$ to get more information.

2.4 Vision transformers

Building on the concept of attention, the transformer architecture [4] was introduced,they rely entirely on self-attention mechanisms to process input data, dispensing with the recurrent structure of RNNs. This allows transformers to handle long-range dependencies more efficiently and in parallel, leading to significant improvements in performance for tasks such as machine translation, text summarization, and beyond. Transformers have become the backbone of state-of-the-art models like BERT, GPT, and T5.

Vision transformers [5] (ViTs Figure 7) are encoder-only transformers adapted for computer vision tasks. The idea is to break down an image of size $W \times H$ into a series of patches of size $P_w \times P_h$. These patches are then flattened, resulting in a two-dimensional matrix of dimensions $S \times D$, where $S = \frac{W}{P_w} \times \frac{H}{P_h}$ and $D = P_w \times P_h$. A linear projection layer is used to transform the individual flattened patches to a lower-dimensional vector, resulting in a matrix of dimension $S \times D'$. Finally, positional embedding is applied.

The output of patching and positional embedding is then fed to a regular transformer encoder,the outputs of the encoder are then passed to a Multi-Layer Perception to output the classes probabilities.

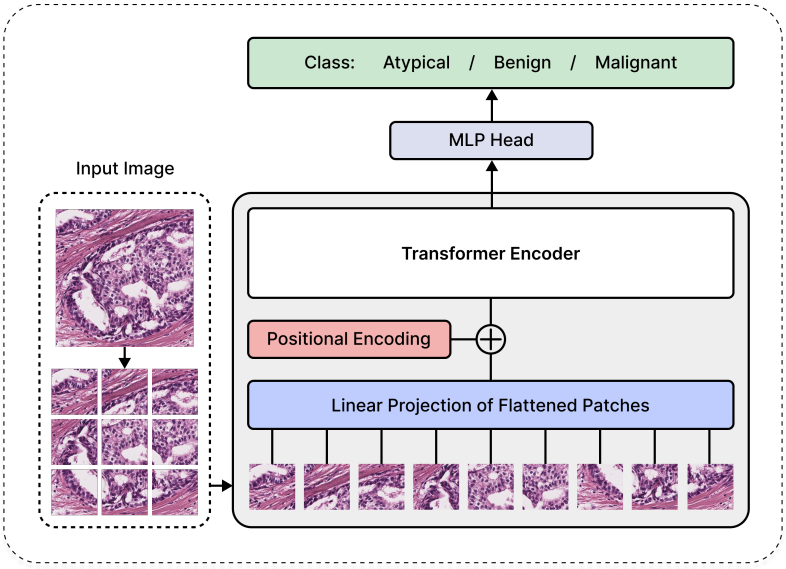


FIGURE 7: Vision Transformer architecture

2.5 Fine tuning & transfer learning

Fine-tuning and transfer learning are particular techniques in deep learning that allow reusing the knowledge learned from training a model on a certain task for other tasks. Both work by adapting a pre-trained model, which was trained on a more general task or another similar task, by retraining it on the new task.

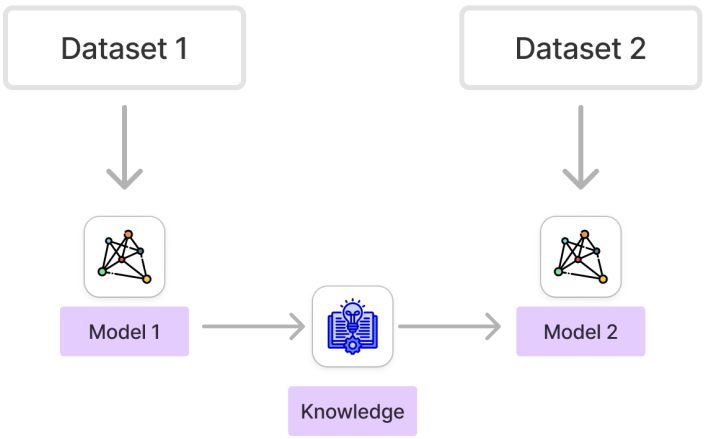


FIGURE 8: The process of Fine tuning / transfer learning

Transfer learning utilizes the pre-trained model as a feature extractor, which means that the layers responsible for feature extraction, such as convolutional layers in computer vision, are frozen, while the decision-making layers, such as the last fully connected layers in CNNs, are modified to match the task's nature and requirements, and only these decision-making layers are trained. Transfer learning is suitable when limited labeled data or computational resources are available, and there is a high similarity between the tasks.

In fine-tuning, we do not completely rely on the features learned from the source task but further "fine-tune" them to enhance the representation of our data and overcome potential differences between the source and the target task. This can be achieved by unfreezing some or all of the layers responsible for feature extraction, usually those closest to the decision-making layers, as they are responsible for extracting more fine-grained features, while earlier layers are responsible for extracting more low level features that similar tasks usually share,therfore fine tuning represents a more flexible way to re-use model as the number of layers to unfreeze have an impact on the model's performance and can be tweaked in function of the anount of available data,the similarity between the tasks and the computational resources.

2.6 Multi-instance learning

Multi-instacne learning is a sub-type of supervised learning,where instead of associating a label with each data point,we associate a single label Y with a set of data points called a bag $X = \{x_1, x_2, ..., x_K\}$, where K is the size of the bag that could vary for different bags, in MIL it is assumed that labels $\{y_1, y_2, ..., y_K\}$ for individual instances exists, but they are unknown and that a bag may be labeled negative if all the instances in the bag are negative, or positive if there is at least one positive instance in the case of a binary classification problem, the assumption behind MIL can be formulated as :

$$Y = \begin{cases} 0 \text{ if } \sum_{i=1}^{K} y_i = 0 \\ 1 \text{ otherwise} \end{cases}$$

MIL is particularly useful in scenarios where labeling individual instances is challenging or expensive, but labeling groups (bags) is feasible. Common applications include:

- Medical Imaging: Identifying whether a medical image (bag) contains any instances of a disease (instance) without pinpointing the exact location of the disease.
- Text Categorization: Classifying documents (bags) based on whether they contain certain topics (instances), with each document being a collection of text segments or sentences.
- Drug Activity Prediction: Determining whether a molecule (bag) has a particular activity based on its conformations (instances).

There are several approaches to solving MIL problems, each with its own strengths and weaknesses. Some common methods include:

- Instance-based Approaches: These methods focus on the instances within the bags. A common technique involves assuming the presence of at least one positive instance in positive bags.
- Bag-based Approaches: These methods treat the entire bag as a single entity and attempt to classify it based on its overall characteristics.
- Embedded Space Approaches: These methods transform the bags into a different space where traditional machine learning algorithms can be applied.

3 State of the art

This section aims to provide an overview of the current state-of-the-art in deep learning-based approaches for breast cancer diagnosis through multi-classification of histology images. Several studies have proposed CNN-based approaches [6] for breast cancer histology image classification. Attention models [7] and Vision Transformers [5] have been explored as well.
In [8] a CNN Class Structure-based approach was used on the Breast Cancer Histopathological (BreakHis) dataset [9], in order to leverage the hierarchical feature representation. At the image level, an accuracy of about 92% for all magnification factors was achieved in respect of a multi-classification in eight classes, and an approximate accuracy of 96% was achieved for the binary classification. Another approach was presented in [10], where the performance was evaluated on the dataset provided by the Bioimaging 2015 challenge [11]. A CNN and a combination of a CNN and an SVM were proposed, achieving accuracies of 77.8% for the four class test and 83.3% for the carcinoma/non-carcinoma test. [12] presented a supervised method for the multi-classification of breast cancer histology images based on the fine-tuning strategy of a ResNet model [3]. The classification of a breast cancer histology image has been obtained by combining three configurations of the ResNet with a different number of layers according to the maximum probability rule. The method has been tested on the Grand Challenge on Breast Cancer Histology Images (BACH) dataset [13], and on the dataset provided by the Bioimaging 2015 challenge. Different experiments have been performed, using different training modalities of ResNet. The best accuracy was obtained by applying the fine-tuning strategy, training all the layers of the networks. The accuracy achieved was 97.3% for the multi-classification in four classes and 98.7% for carcinoma/non-carcinoma test on the BACH dataset. To reduce the training parameters and reduce the risk of model over-fitting, [14] designed a small SE-ResNet model based on the combination of residual module and Squeeze-and-Excitation block. Compared to the bottleneck SE-ResNet module and basic SE-ResNet module, the parameters of the small SE-ResNet module is reduced to 29.4% and 33.3%, respectively. The authors also proposed a new learning rate scheduler named Gaussian error scheduler which can get excellent performance without complicatedly fine-tuning the learning rate. Additionally they designed a novel CNN network based on small SE-ResNet module, pooling layer, and fully connected layer. This model has been tested on the BreakHis dataset for binary classification and multi-class classification with competitive experimental results. Accuracies of 93.74%, 93.81%, 92.22%, and 90.66% have been acheived on the multi-classification task. [15] proposed a framework that consists of two stages: a Grid-based Feature Extraction (GFE) and an Attention-based Classifier (AC). The GFE applies a CNN to extract patch-wise features and aggregate feature vectors in a compact grid representation according to the spatial location of the corresponding patches in the WSI. The AC implements both the min- and max-attention mechanisms separately on the input grid-based feature map and produces two different sets of attention maps. The attention maps are then used for classification. [16] proposed a framework named Hierarchical Image Pyramid Transformer (HIPT). This work aimed to adapt ViTs [5] for slide-level representation learning to capture the hierarchical structure of WSIs. ViT256 -16, ViT4096 -256 and ViTWSI -4096 have been proposed and utilized to aggregate visual tokens at different levels to form slide representations. The [CLS] tokens from ViT256 -16 are used as the input sequence for ViT4096 -256, with the [CLS] tokens from ViT4096 -256 subsequently used as the input sequence for ViTWSI -4096.

Below a comparative table the mentioned approaches:

Reference	DL Technique	Dataset	Results	Multiple Instance Learning
[8]	Class structure-based deep convolutional neural network (CSDCNN) for the classification task.	BreakHis	- multi classification accuracy: 92% - binary classification accuracy: 96%	No
[10]	Convolutional Neural Networks (CNNs) for feature extraction + SVM for classification.	Bioimaging 2015 challenge dataset	- multi classification accuracy: 77.8% - binary classification accuracy: 83.3%	Yes
[12]	Fine tuning strategy on ResNet model	- BACH: Grand Challenge on Breast Cancer Histology Images. - Bioimaging 2015 challenge dataset	- multi classification accuracy on Bioimaging 2015 challenge dataset: 97.2% - multi classification accuracy on BACH dataset: 97.3%	No
[14]	- small SE-ResNet - learning rate scheduler named Gaussian error scheduler - novel CNN network based on small SE-ResNet module	BreakHis	- multi classification accuracy: 90.66% and 93.81% - binary classification accuracy: 98.87% and 99.34%	Yes
[15]	- a Grid-based Feature Extraction (CNN) - Attention-based Classifier	- Camelyon16 - TUPAC16	- Prediction of Metastasis presence AUC: 0.711 - Prediction of Tumor Proliferation Speed, spearman correlation coefficient: 0.662-0.653	Yes
[16]	- Adapting ViTs. - Hierarchical Self-Supervised Learning	TCGA	- 0.821-0.874 AUC on BRCA subtyping	Yes

TABLE 1: Comparative table of state of the art methods

4 BRACS Dataset

BReAst Carcinoma Subtyping (BRACS) dataset [2], a large cohort of annotated Hematoxylin and Eosin (H&E)-stained images to facilitate the characterization of breast lesions.

The dataset was built through the collaboration of the National Cancer InstituteScientific Institute for Research, Hospitalization and Healthcare (IRCCS) Fondazione G. Pascale, the Institute for High Performance Computing and Networking (ICAR) of National Research Council (CNR) and International Business Machines (IBM) ResearchZurich. The dataset was acquired from patients between 2019 and 2020, by board-certified clin-icians of the Department of Pathology at the National Cancer InstituteIRCCS Fondazione G. Pascale in Naples (Italy).

The BRACS dataset comprises both Whole Slide Images (WSIs) and Regions of Interest (ROIs). WSIs are digital images of entire histological slides, capturing the complete tissue section at high resolution (Gigapixels), and they have been obtained by scanning slides that were selected by a biologist of the pathological anatomy department. Regions of Interest (ROIs), on the other hand, represent carefully selected and annotated image patches that were extracted from the Whole Slide Images (WSIs), focusing on specific areas of interest within the vast expanse of the WSIs.

4.1 Data collection and annotation process

The samples were obtained from hematoxylin and eosin (H&E) stained breast tissue biopsy slides, with the selection based on the diagnostic reports of the patients. The age distribution of the patients ranged from **16** to **86** years, with approximately **61%** of the patients falling within the **4060** age range and only a few patients under **20** or over **80** years old.

A Whole Slide Image (WSI) typically includes several lesions of varying subtypes, including Normal **(N)**, Pathological Benign **(PB)**, Usual Ductal Hyperplasia **(UDH)**, Flat Epithelial Atypia **(FEA)**, Atypical Ductal Hyperplasia **(ADH)**, Ductal Carcinoma in Situ **(DCIS)**, and Invasive Carcinoma **(IC)**. To ensure a high level of reliability in the annotations, three expert pathologists were involved in annotating both the WSIs and Regions of Interest (ROIs).

Initially, each pathologist inspected the WSIs, assigning the corresponding label according to the most aggressive tumor subtype they detected within the image. After that, all annotations were collectively reviewed by the three pathologists, and those with disagreement were further discussed and re-annotated when consensus was reached or discarded otherwise.

A subset of the annotated WSIs was split into three disjoint subsets, each of which was assigned to one of the pathologists. Each pathologist extracted a set of ROIs from their assigned subset, while ensuring that at least one ROI with the same subtype as the WSI was selected and that the set of extracted ROIs maintained a balanced distribution across the classes.

4.2 Dataset characteristics

The BRACS dataset comprises a total of **547 WSIs** obtained from **189 different patients**. The dataset also includes **4,539 ROIs** that were extracted from **387 WSIs** collected from **151 patients**. All slides in the dataset were scanned using an **Aperio AT2** scanner, which helps to capture images at a high resolution of **0.25** μm/pixel. The scanning process used a magnification factor of

40x, which ensured detailed imaging of the tissue samples. Table 2 reports the distribution of WSIs (with and without ROIs) and ROIs according to the lesion types and subtypes.

	Types			Subtypes						
Data	Benign	Atypical	Malignant	N	PB	UDH	FEA	ADH	DCIS	IC
WSIs with ROIs	149	75	163	17	77	55	34	41	51	112
WSIs without ROIs	116	14	30	27	70	19	7	7	10	20
WSIs	265	89	193	44	147	74	41	48	61	132
ROIs	1837	1263	1439	484	836	517	756	507	790	649

TABLE 2: BRACS data distribution according to lesion types and subtypes.

4.3 Dataset organization

BRACS dataset provides pre-defined WSI- and ROI-level splits in train, validation and test sets. Data split was generated such that all the WSIs extracted from a patient belong to the same set. Similarly, all the ROIs extracted in a given WSI are assigned to the same split. Table 3 reports the number of WSIs and ROIs included in the train, validation and test splits according to the lesion types.

	WSI-level split				ROI-level split			
Data	Benign	Atypical	Malignant	Total	Benign	Atypical	Malignant	Total
Train	203	52	140	395	1460	1011	1186	3657
Validation	30	14	21	67	135	90	87	312
Test	32	23	32	85	242	162	166	570
Total	265	89	193	**547**	1837	1263	1439	**4539**

TABLE 3: WSI- and ROI-level split according to the lesion types.

While Table 4 reports the number of WSIs and ROIs included in the train, validation and test splits according to the lesion subtypes.

	WSI-level split							ROI-level split						
Data	N	PB	UDH	FEA	ADH	DCIS	IC	N	PB	UDH	FEA	ADH	DCIS	IC
Train	27	120	56	24	28	40	100	357	714	389	624	387	665	521
Validation	10	11	9	6	8	9	12	46	43	46	49	41	40	47
Test	7	16	9	11	12	12	20	81	79	82	83	79	85	81
Total	44	147	74	41	48	61	132	484	836	512	756	507	790	649

TABLE 4: WSI- and ROI-level split according to the lesion subtypes.

The dataset is organized as follows. The WSIs are stored in the 'Whole Slide Images Set' folder, which is further partitioned into train, validation, and test data subsets. Each of these subsets is divided into Benign (BT), Atypical (AT), and Malignant (MT) folders, each containing folders corresponding to specific lesion subtypes. The WSIs are saved as '.svs' files, adhering to a consistent naming convention.
On the other hand, the ROIs are stored in the Region of Interest Set folder, which follows the same structure as the WSI set. The ROI files are stored in png format. Finally, a summary file in '.xlsx' format is included, which lists for each WSIs its label, data split assignment

(train/validation/test), associated patient ID, and the number of corresponding ROIs, if there is any. Figure 9 highlights the folder organization of BRACS dataset.

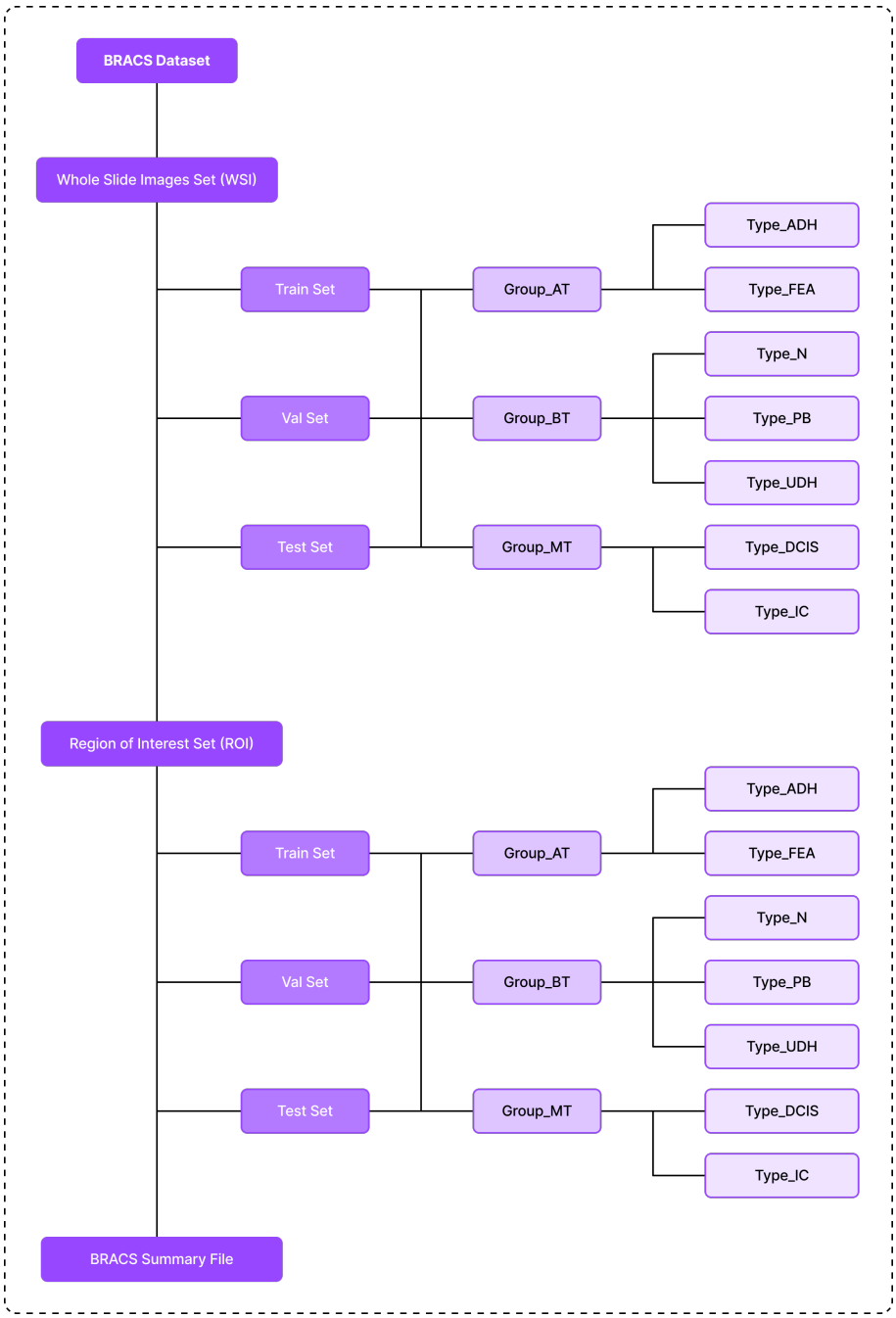


FIGURE 9: The organization of BRACS dataset folders [2].

5 Approach

Due to the computational challenges posed by gigapixel images and limited hardware and GPU RAM, a two-step approach was adopted to overcome these challenges. In the first step, a feature extraction process takes place, where we process the WSI using the **OpenSlide** library to partially load the image and then process it through a pre-trained neural network to produce a more compact representation of the WSI that can be easily processed by the attention models. The second step involves training the attention models using the generated tensors from the first step. Two architectures were chosen: a **Min-Max Attention Classifier** [15] and a **Multi-Branch Attention Challenging Classifier**[17]. In the feature extraction step, we experimented with four different models: **ResNet-18, ResNet-34** (fine-tuned on BRACS' regions of interest), **ResNet-50** (trained on Kather100k Dataset) and a **ViT-S/16** pre-trained using DINO on a substantial collection of 36,666 WSIs. Figure 10 shows a schematic representation of the general approach.

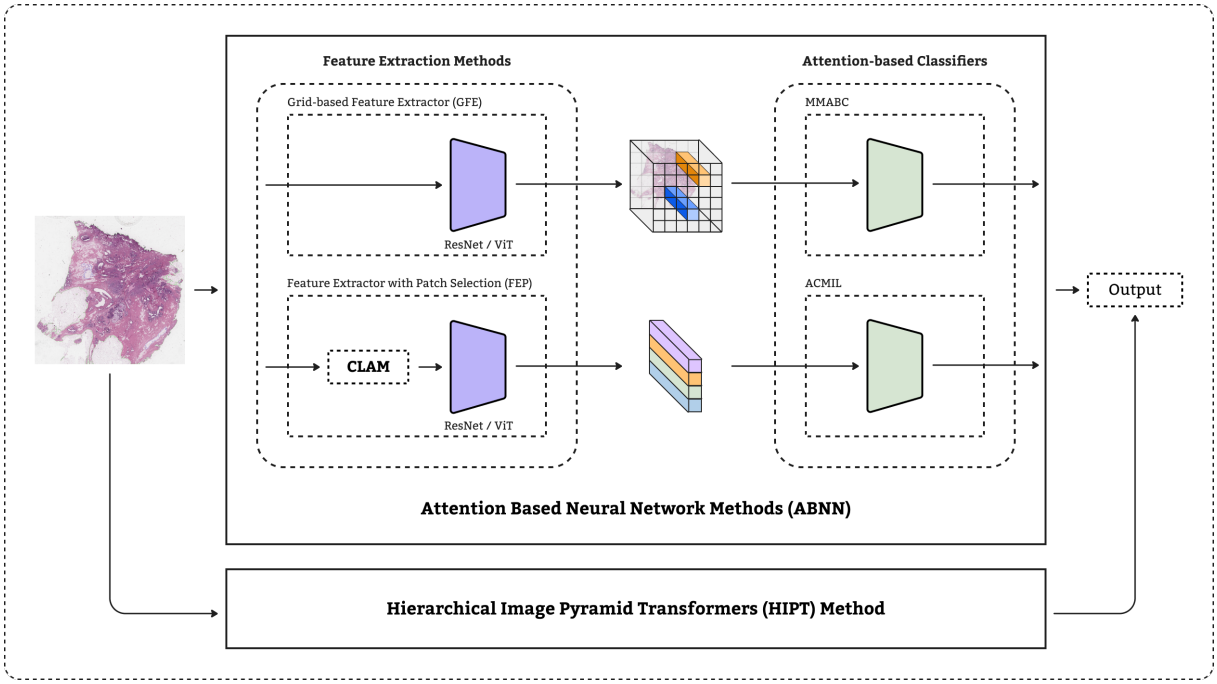


FIGURE 10: Schematic representation of the general approach.

A high-level overview of the process is as follows :

- Dividing the WSI into a set of non-overlapping patches.

$$X = \{x_1, x_2, ..., x_K\}$$

- Pass the patches to a pre-trained neutral network to transform the patch into a low-dimensional embedding

$$h_i = f(x_i)$$

.

- Use an attention mechanism to generate a bag-level representation of the WSI .

$$\begin{cases} z = \sum_{i=1}^{K} a_i * h_i \\ a_i = \sigma(h_n) \end{cases}$$

.

- Generate the WSI prediction from the bag-level representation.

$$\hat{Y} = g(z)$$

Similarly, fine-tuning the **HIPT** model was divided into two steps due to computational limitations. As explained earlier, the standard approach when applying fine-tuning or transfer learning is to load a pre-trained model's weights and initialize your architecture with those weights. Then, you can optionally freeze a number of layers and retrain the model on your own task. However, since it is impossible to load the entire WSI into the GPU's memory, and the time it takes for it to propagate through the network's layers, even if partial loading was used to overcome the memory challenge, a similar approach was adopted. We processed the entire dataset using the frozen layers and saved their output to the disk, which would be the input to the part of the network we wished to train. The only drawback is that we cannot apply data augmentation to the input on-the-fly, and due to the spatial and time complexity challenges that the dataset possesses, iterating multiple times through the dataset and saving an augmented version of it is not an option. To overcome this drawback, we applied on-the-fly data augmentation on the generated tensors instead. Since the tensors have a shape of $W \times H \times C$, some of the usual data augmentation techniques used in computer vision were used, but re-implemented to work on an arbitrary number of channels.

5.1 Preparing Models For Feature Extraction

In the context of our project, it involves using pre-trained models and fine-tuning them on ROI images, which are specific areas of WSIs that contain significant diagnostic information selected by experts. By fine-tuning on these carefully selected regions, the models learn to identify critical patterns and features associated with cancerous tissues. This specialized training enables the models to become highly adept at recognizing the subtle distinctions necessary for accurate cancer detection. The fine-tuned models are then used as feature extractors for the entire WSIs, allowing for a more efficient and specialized analysis. This approach maximizes the utility of annotated data, enhances the model's sensitivity to relevant pathological features, and ultimately improves the accuracy and reliability of cancer diagnosis in large-scale histopathological images. Figure 11 illustrates the fine-tuning process on ROI images.

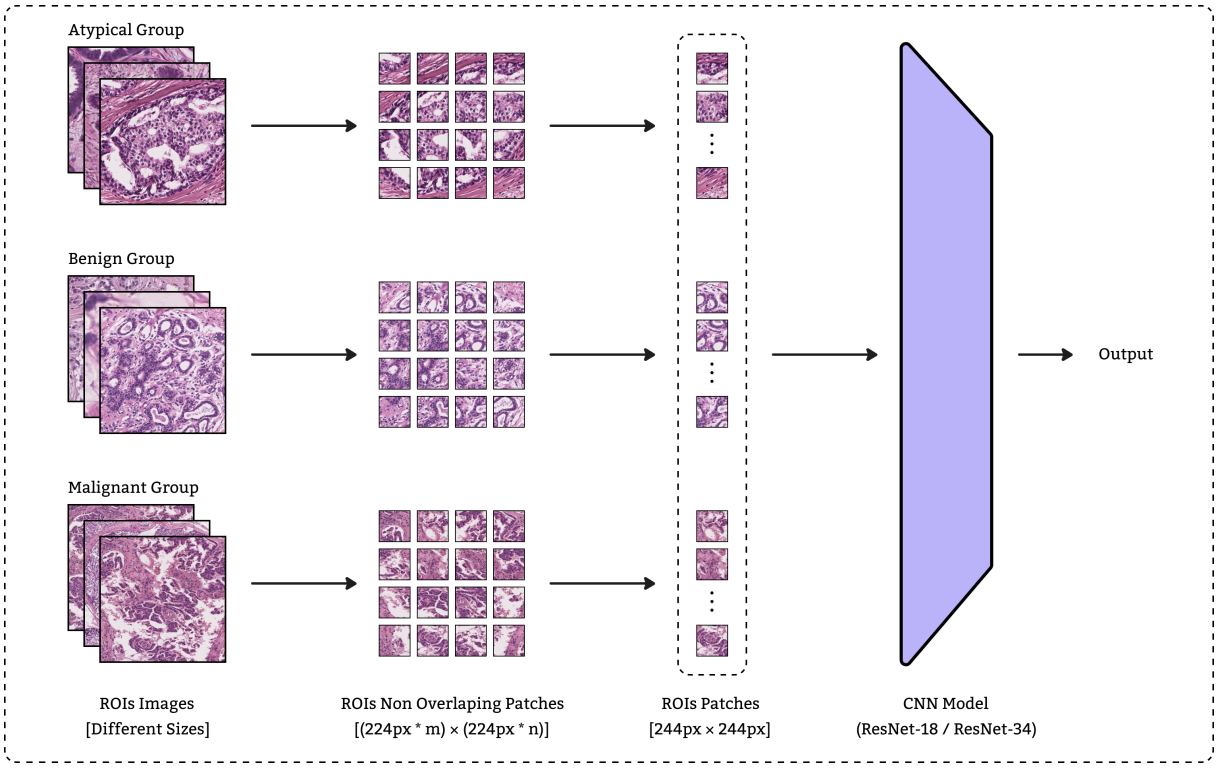


FIGURE 11: Fine-tuning process on BRACS' ROI images.

Before fine-tuning the models on ROI images some preprocessing must be made so they can be fed to the ResNet models which require an input of 224 x 224 pixels. However, The ROIs in the dataset have varying dimensions, i.e they can not be divided into equal patches, which poses a challenge for direct input of our models. To address this issue, we had to choose one of two options:

- Using overlapping 224 x 224 pixels patches to cover the whole image.
- Resizing the height and width to the closest multiple of 224 (could be bigger or smaller than the current value).

We opted for the latter because the first option could produce too many patches, which would drastically elevate the training time. By resizing each image to the closest multiple of 224, we ensured compatibility with the image patching process, allowing the resized image to be covered entirely by 224 x 224 pixel patches.

After resizing, we employed an image patching algorithm, splitting each ROIs image into multiple non-overlapping patches of size 224 x 224 pixels, while attributing the label of the original image to the label of the patch (weak patch labeling) and keeping the same original train, validation and test splits they belonged to originally. Furthermore we ensured that the resizing process does not cause any loss of information in the original image,Tables Table 5 and Table 6 reports the distribution of the patches according to their types and sub-types.

Data	Benign	Atypical	Malignant	Total
Train	141578	33898	157798	333274
Validation	8151	4472	18635	31258
Test	13137	6191	20564	39892
Total	162866	44561	196997	**404424**

TABLE 5: Data distribution of patches count across different splits and lesion types.

Split	N	PB	UDH	FEA	ADH	DCIS	IC	Total
Train	27470	94454	19654	16821	17077	68491	89307	333274
Validation	3267	3018	1866	2793	1679	5986	12649	31258
Test	3501	5509	4127	2499	3692	5303	15261	39892
Total	34238	102981	25647	22113	22448	79780	117217	**404424**

TABLE 6: Data distribution of patches count across different splits and categories.

A comparison between Table 3 and Table 5 shows that Malignant became the predominant class, and that is due to the fact that type IC has the ROIs with the highest resolution across the dataset, which means it produces more patches (Table 6).

5.2 Feature Extraction

5.2.1 Grid-based feature extraction

The goal of the Grid-based Feature Extraction stage (GFE) is to create a more compact representation of the original image that can be processed by the Attention-based Classifier. To achieve this goal, the GFE takes as an input the original gigapixel Whole Slide Image (WSI) and transforms it into a lower dimensional feature space while preserving local spatial relationships. The WSI is partitioned into a set of non overlapping patches. These patches are then mapped into feature vectors by applying deep learning models, such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). The resulted feature vectors are rearranged in a grid format to maintain the spatial proximity information of the original patches in the WSI. Figure 12 shows an illustrated representation of the GFE stage.

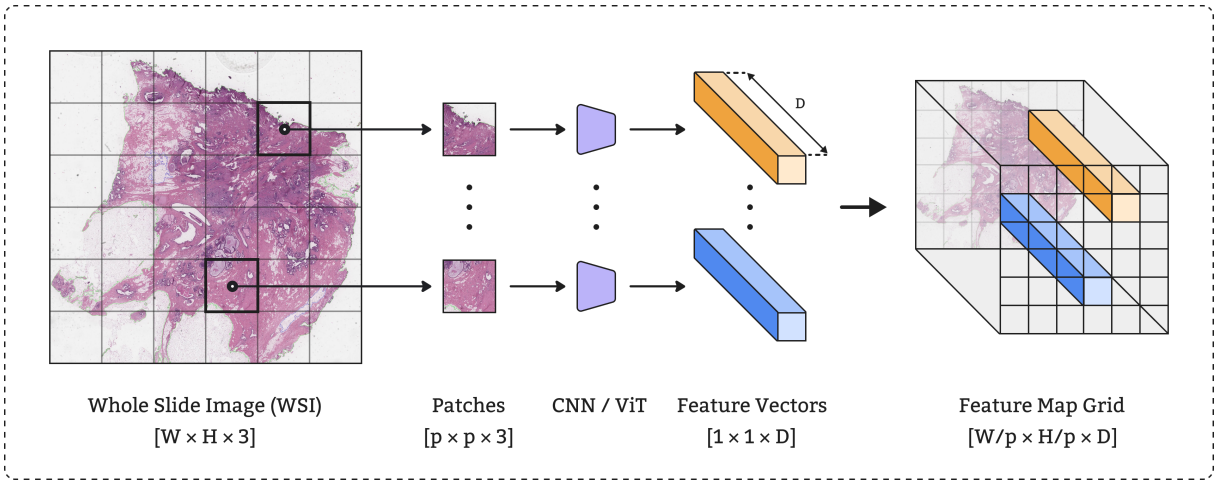


FIGURE 12: Grid-Based Feature Extraction (GFE) [15].

For more details, lets $W \in R^{W \times H \times 3}$ represents the input image, where W and H are the width and the height of the Whole Slide Image with three color channels (RGB). The GFE divides the input image into a set of non-overlapping patches $X = \{x_{i,j}\}$, and that can be done by sampling W along the i^{th} row and j^{th} column according to a uniform grid of size $p \times p \times 3$. Each of these patches is then independently fed to a CNN or a ViT, which maps it into a $1 \times 1 \times D$ feature vector obtained from the network's global average pooling layer. These patch-wise feature vectors are then assembled into a three-dimensional compressed representation organized in a grid $G \in R^{W \times H \times D}$ where $W' = \frac{W}{p}$ and $H' = \frac{H}{p}$, so that the spatial arrangement of the vectors in G corresponds to the original positioning of their respective patches in the WSI.

5.2.2 Feature extraction with patch selection

Contrary to grid based feature extraction which takes as input the entire WSI image, in this method the WSI first goes through CLAM for patch selection before doing any feature extraction. CLAM (Clustering-constrained Attention Multiple instance learning) is a method proposed in [18]. It is a high-throughput and interpretable method for data efficient WSI classification using slide-level labels without any ROI extraction or patch-level annotations, and is capable of handling multi-class subtyping problems. One of its main functionalities is segmentation and patching, in the context of feature extraction with patch selection it is used as a pre-processing step to locate tissue regions in each WSI, i.e removing the background. CLAM returns equal patches coordinates for a given WSI (see Figure 13), these coordinates are then used to select which patches to be fed to the feature extraction model (ResNet-18, ResNet-34 or a ViT-S/16), each patch produces a feature vector, the vectors are then assembled to produce a compact representation of the WSI and finally saved to the disk as tensors. Figure 14 shows a schematic representation of the overall process.

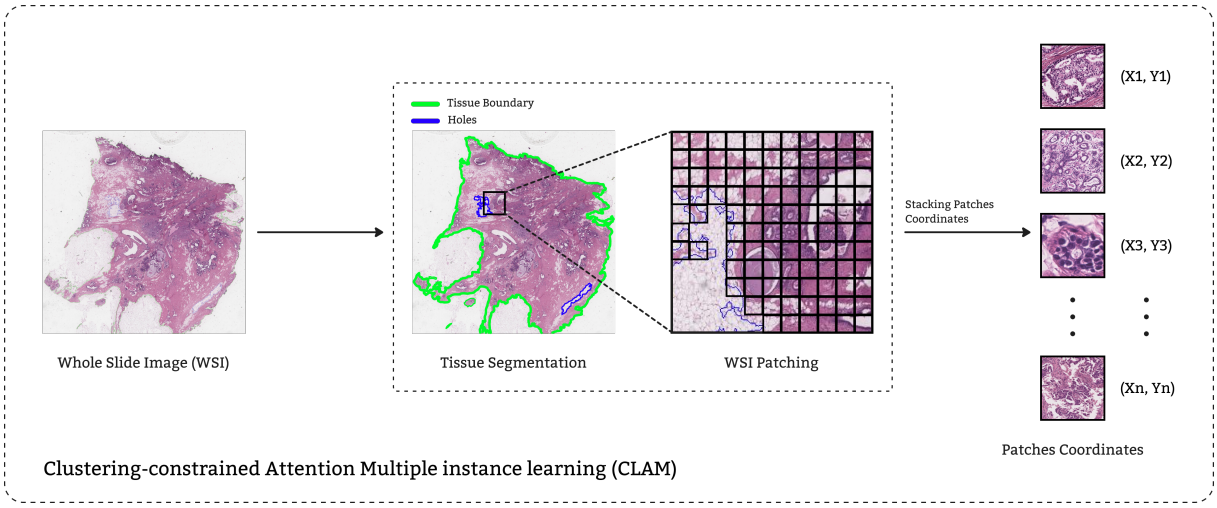


FIGURE 13: Clustering-constrained Attention Multiple instance learning

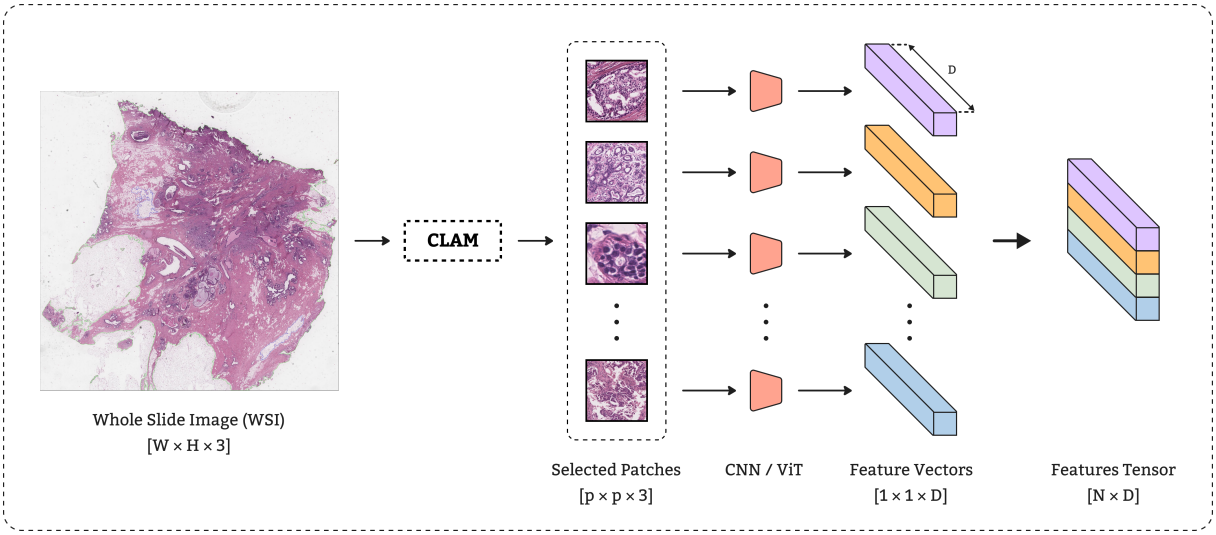


FIGURE 14: Feature extraction with patch selection

5.3 WSI Classifiers

5.3.1 Min-Max attention based classifier

Min-Max attention based classifier (MMABC)[15] is architecture taht exploits 3D convolutional layers and min-max attention mechanisms to extract relevant patterns from the compressed WSI,without ignoring the relative location between the patches' embeddings.

the model operates on the outputs of grid-based feature extraction technique (GFE) which means it receives an input $G \in R^{W' \times H' \times D}$,where W' and H' are the width and the height of the compressed WSI and relies on a 3D convolutional layer for feature extraction then those features are fed to two different attention layers (min & max attention layers) and the generated feature vectors from both layers are concatenated then passed to a softmax layer to output the probabilities. Figure 15 shows a schematic representation of the proposed attention network.

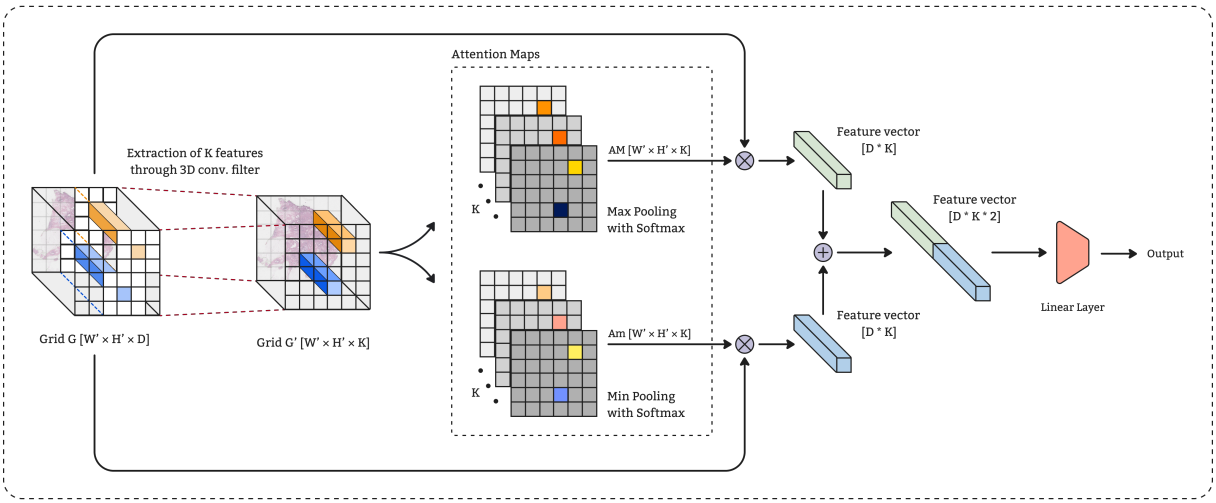


FIGURE 15: Min-Max attention based classifier (AC) [15].

Processing the compressed WSI with a 3D convolutional layer involves using a 3D convolutional layer to extract K filters from G, resulting in a tensor $G' \in \mathbb{R}^{W' \times H' \times K}$. The idea behind this is to extract features based on the relative location between patches, as each patch was treated independently during the GFE. The min attention mechanism divides each filter of the input features G' into smaller non-overlapping regions of dimensions $S \times S$, then only the minimum value from each region is kept while the others are set to zero. The softmax function is applied to the flattened output of the previous stage, which results in K attention maps $A_{\text{min}} \in \mathbb{R}^{W' \times H' \times K}$.

The max attention operates similarly, but instead, only the maximum value from each region is kept, outputting M attention maps $A_{\text{max}} \in \mathbb{R}^{W' \times H' \times K}$. Finally, the attention maps of each attention layer independently are multiplied with each channel of the original input to result in two feature vectors, each of dimension $D \times K$, one corresponding to the min-layer and the other to the max-layer. The two feature vectors are then concatenated to get a feature vector of dimension equal to $D \times K \times 2$. The feature vector outputted by the min-max attention mechanism is fed to a linear layer with a softmax activation function to generate the probabilities of each class.

5.3.2 Attention-Challenging Multiple Instance Learning

Attention-Challenging Multiple Instance Learning [17] (ACMIL : Figure 16) is an attention-based architecture that was proved successful in reducing the overfitting for various cancer datasets including BRACS Dataset, by introducing a new regularization technique called STKIM(Stochastic Top-K Instance Masking),as well as a new loss function that insures the diversity of the extracted features.

The ACMIL model operates on an input $x \in R^{N \times D}$ representing the embeddings of flattened patches, where N is the number of patches and D is the embedding's dimension.

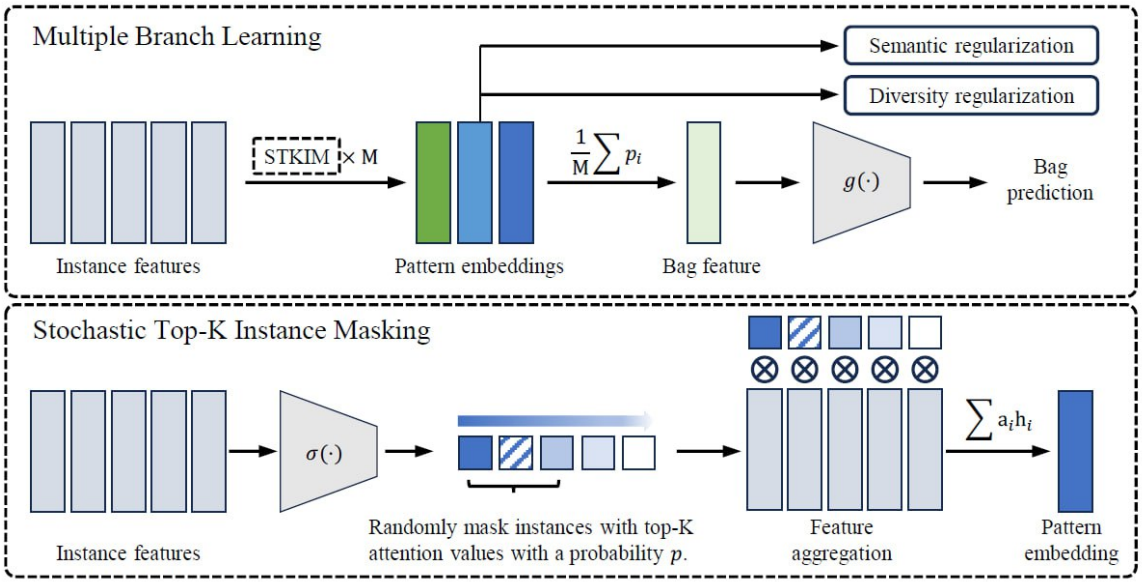


FIGURE 16: An Overiew of the ACMIL Architecture[17].

First of all the input (a compressed WSI) is past to a dimentiality reduction layer is a regular MLP network that maps the dimension of the input patches features of dimension D to a lower dimenion D'.

$$X' = DimReduction(X)$$

The the output X' is then processed by the multi-branch attention block,to better understand the multi-branch attention mechanism we first need to understand the attention gated mechanism,The attention gated mechanism allows the model to focus on different parts of the of the input by computing a score for each patch embedding.

An attention gated layer is composed of three matrices $U \in R^{D' \times D_{hidden}}$,$V \in R^{D' \times D_{hidden}}$ and $W \in R^{D_{hidden} \times 1}$:

- the matrix U reduces the dimension of the input tensor to D_{hidden} followed by a $tanh$ activation function resulting in matrix $U_{output} \in R^{N \times D_{hidden}}$.
- the matrix V is used in conjunction with the $sigmoid$ function to control the flow of the information in the network by computing values between 0 and 1, resulting in matrix $V_{output} \in R^{N \times D_{hidden}}$.
- the matrices V_{output} and U_{output} are multiplied element wise.
- finally the matrix W followed by a transpose operation then the $softmax$ function is used to compute a score for each patch, resulting in an attention filter of dimenions $A \in R^{1 \times N}$.

$$\begin{cases} a = ((tanh(X' \times U) * sigmoid(X' \times V)) \times W)^T \\ A = softmax(a) \end{cases}$$

The multi-Branch attention modifies the attention gated mechansim by not only computing one score per patch but multiple ones,this can be acheived by modifying the dimensions of the matrix W to be $D_{hidden} \times K$,where K is the number of branches,resulting in attention filter $A \in R^{K \times N}$,which can be seen as having multiple branches (Gated Attention layers) that process the input simultaneously.

Stochastic top-k instance masking

Stochastic top-k instance masking (STKIM) is a regularization technique similar to dropout that randomly masks (sets to zero) the top-k attention scores with probability of p, this technique deals with one of the main reasons of overfitting in whole slide images,which is the model relying on small number of patches in the decision making by randomly masking a portion of the patches with the higher scores,we encourage the model to extract patterns from the other patches.

STKIM is applied an the outputs of the attention layer before applying the softmax function to ensure that scores always add up to one, and just like dropout STKIM is deactivated in inference mode.

The loss function

To insure that the model is extracting meaningfull and diverse patterns from the training data a new loss function had to be introduced,it is composed of three sub-losses :

Diversity Loss :

To ensure that the branches are not learning similar patterns a diversity loss is introduced,it is simply the average cosine similarity between the vectors representing the attention scores (before applying the softmax function) given to the patches by each
branch :

$$L_d = \frac{2}{K * (K-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} cos(a_i, a_j)$$

Semantic Loss :

Diversity loss alone is not very usefull as the model can learn diverse but non-relevant patterns,so a semantic loss is added,by calculating the outputs of the model based on the attention scores of each branch separately,then calculating the average cross entropy loss based on those outputs and the true labels,this requires an MLP for each branch :

$$\begin{cases} L_s = -\frac{1}{K} \sum_{i=1}^{M} Y log(\hat{Y}_i) + (1-Y) log(1-\hat{Y}_i) \\ \hat{Y}_i = softmax(MLP_i(A_i \times X')) \end{cases}$$

The bag classifier's loss function :

For the final prediction is made by calculating the attention values for each branch are averaged,multiplied by the input matrix then fed to an MLP network to generate the output,these quality of the prediction are assessed using the regular cross entropy loss :

$$\begin{cases} L_b = -(Y log(\hat{Y}) + (1-Y) log(1-\hat{Y})) \\ \hat{A} = \frac{1}{K} \sum_{i=1}^{K} A_i \\ \hat{Y} = softmax(MLP(\hat{A} \times X')) \end{cases}$$

The overall loss function is simply the sum of the three losses :

$$L = L_s + L_d + L_b$$

5.3.3 Hierarchical Image Pyramid Transformers

Hierarchical Image Pyramid Transformers (HIPT) is a Vision Transformer (ViT) architecture that was proposed in [16], designed for analyzing gigapixel whole-slide images (WSIs) in computational pathology. HIPT leverages the inherent hierarchical structure of WSIs to learn high-resolution image representations through self-supervised learning. Pretrained on a large dataset covering 33 cancer types and evaluated across multiple slide-level tasks, HIPT demonstrates superior performance in cancer subtyping and survival prediction. Adapting the HIPT architecture to the BRACS Dataset is a novel experimentation which has not been done before.

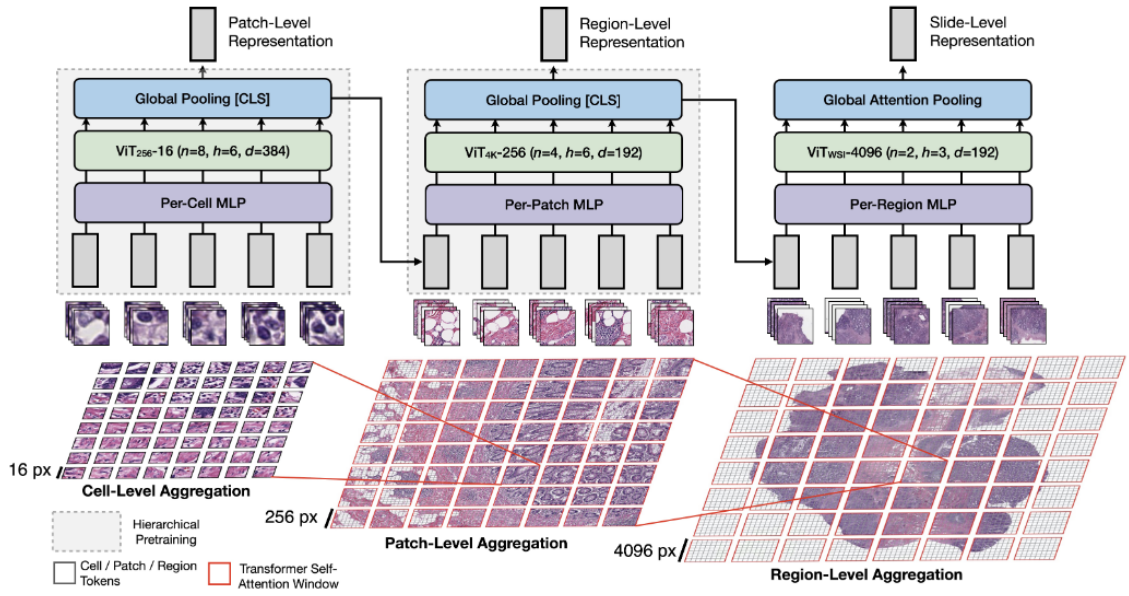


FIGURE 17: HIPT Architecture [16].

As depicted in Figure 17, HIPT uses a hierarchical structure to process WSIs through multiple scales of visual tokens, systematically aggregating information from smaller patches to larger regions and finally to the entire slide. This hierarchical processing is crucial for capturing the complex and multi-scale patterns present in pathology images, we distinguish four different levels :

Cell-Level (16Œ16 pixels): 16Œ16 pixel patches are tokenized and processed using a Vision Transformer (ViT256-16) to learn fine-grained detailed cell-level representations.

Patch-Level (256Œ256 pixels): Tokens representing 256Œ256 pixel patches are further aggregated using another Vision Transformer (ViT4096-256), capturing interactions within a larger local clusters of cell interactions.

Region-Level (4096Œ4096 pixels): Tokens for 4096Œ4096 pixel regions are aggregated using a final Vision Transformer (ViTWSI-4096), forming a macro-scale representation of the WSI.

Slide-Level: The aggregated tokens from the region-level are combined to integrate information from all hierarchical levels and provide a comprehensive representation of the entire WSI.

HIPT employs self-supervised learning, specifically using a technique called student-teacher knowledge distillation (DINO) [19], to pretrain each hierarchical level. In the student-teacher framework, two networks (student and teacher) are trained simultaneously. The teacher network, updated more slowly using an exponential moving average of the student's parameters, stabilizes the learning process. Multi-view learning is employed where different augmented views of the same image are used, and the student network processes these views to produce embeddings that are aligned with those of the teacher network, encouraging invariant feature learning. The self-distillation loss matches the students embeddings to the teachers embeddings, ensuring the model learns robust and generalizable features.

HIPT involves two main stages of pretraining, each focusing on different scales of the input data to capture both fine-grained and coarse-grained morphological features essential for diagnostic tasks,In this stage, 256 Œ 256 pixel patches of histology images are used for pretraining a Vision Transformer (ViT256-16). This process employs the DINO framework, which includes a student network and a teacher network. The student network is trained to match the probability distribution output of the teacher network using a cross-entropy loss function. Data augmentation plays a critical role here, with the introduction of local views (96 Œ 96 crops) and global views (224 Œ 224 crops) to encourage the model to learn local-to-global correspondences. This approach helps the model capture detailed cellular structures and their interactions within the local tissue environment, which are crucial for understanding cellular morphology in pathology, The second stage involves a more extensive scale, using 4096 Œ 4096 pixel regions. Here, the pretrained ViT256-16 model is reused to tokenize these larger regions into sequences of 256 tokens. These tokens are then used as input to another Vision Transformer model (ViT4096-256) for further pretraining. This stage also follows a similar DINO framework, with data augmentations adjusted to match the scale of the larger patches, such as local-global crops of 6 Œ 6 and 14 Œ 14. This stage is crucial for capturing broader spatial organizations and macro-scale interactions within the tissue, such as tumor invasion patterns and lymphocytic infiltration, which are significant for understanding the overall tissue architecture and for making accurate diagnostic and prognostic assessments .

6 Experiments and results

In this section, we will showcase and analyze the results of the trained models. Firstly, we will highlight the outcomes of fine-tuning the feature extractors on The ROIs Dataset, followed by an examination of the attention classifiers' performance on the tensors extracted using those models and additional ones.

The experiments were conducted on an NVIDIA GeForce RTX 3050 Ti Laptop GPU, so consequently, the reported time corresponds to that particular device.

6.1 Feature extractors fine tuning results

As mentioned earlier, we experimented with fine-tuning **ResNet18** and **ResNet34** on BRACS' ROIs, and for each model, we experimented with different configurations (illustrated in Table 7 and Table 8) that ranged from changing hyper-parameters such as the batch size, learning rate, and weight decay, to changing the number of layers to fine-tune, experimenting with balanced sampling, trying different optimizers, and changing the task to adapt the model on.

Model	**Hyperparameters**	**Config 1**	**Config 2**
ResNet-18	Initial weights	ImageNet	ImageNet
	Batch Size	256	64
	Learning rate	0.001	0.001
	Optimizer	ADAM	SGD
	Sampler	Random	Balanced
	Weight decay	None	0.001
	Decay rate	None	None
	Dropout	None	None
	Depth (Fine-tuning)	2	3
	Epochs	20	10
	Average Epoch time (min)	14	16.5

TABLE 7: Hyperparameter Configurations for ResNet-18

Model	Hyperparameters	Config 1	Config 2	Config 3
ResNet-34	Initial Weights	ImageNet	ImageNet	kather100k
	Batch size	128	64	64
	Learning rate	0.001	0.001	0.0001
	Optimizer	ADAM	SGD	ADAM
	Sampler	Random	Balanced	Balanced
	Weight decay	None	0.1	0.1
	Decay rate	None	None	None
	Dropout	None	None	None
	Depth (Fine-tuning)	2	3	3
	Epochs	20	10	10
	Average Epoch time (min)	22	27	27

TABLE 8: Hyperparameter Configurations for ResNet-34

Before diving into the next step which is feature extraction, the fine-tuned models must be evaluated on regular metrics (Accuracy, Precision, Recall, F1-score), we consider the macro F1-score as the primary metric since our dataset is unbalanced, however since our preprocessing involved splitting the ROIs into multiple patches,we evaluated the models on two different tasks : *(i)* Patches Classification and *(ii)* ROIs Classification Classification were we used two different methods to aggregate the predictions of the patches : soft voting and hard voting, soft voting involves computing the average of the predicted probabilities across all patches. The class with the highest average probability is then assigned as the final prediction for that ROI image,meanwhile hard voting considers "the hard predicted labels" from all patches instead of the probabilities and the most frequent class is considered as the label of ROI. The results are shown in details in tables 9,10 and 11.

Model	Configuration	accuracy	precision	recall	f1 score
ResNet-18	Config 1	0.579	0.492	0.459	0.460
	Config 2	***0.632***	***0.563***	***0.559***	***0.557***
ResNet-34	Config 1	0.555	0.460	0.405	0.382
	Config 2	0.622	0.548	0.534	0.536
	Config 3	0.592	0.518	0.500	0.495

TABLE 9: Fine-tuning feature extractors results (patches persepective)

Model	Configuration	accuracy	precision	recall	f1 score
ResNet-18	Config 1	0.461	0.543	0.470	0.396
	Config 2	***0.624***	***0.651***	***0.640***	***0.617***
ResNet-34	Config 1	0.340	0.442	0.370	0.231
	Config 2	0.540	0.614	0.566	0.523
	Config 3	0.543	0.570	0.522	0.487

TABLE 10: Fine-tuning feature extractors results (soft-voting)

Model	Configuration	accuracy	precision	recall	f1 score
ResNet-18	Config 1	0.463	0.535	0.471	0.397
	Config 2	***0.621***	***0.632***	***0.631***	***0.612***
ResNet-34	Config 1	0.340	0.421	0.370	0.233
	Config 2	0.557	0.610	0.578	0.541
	Config 3	0.542	0.551	0.520	0.480

TABLE 11: Fine-tuning feature extractors results (hard-voting)

The obtained results show how sampling techniques can impact performance in the case of imbalanced datasets. We notice that the models trained using random sampling not only suffered from relatively low performance across all metrics compared to those trained using balanced sampling, but they also showed a significant gap in performance between the tasks of classifying individual patches and classifying ROIs a problem that is less noticeable when using balanced sampling.

The low performance on the test set can be attributed to the dataset's imbalance and the distribution differences between the train and test sets, while the performance gap can be interpreted as a result of the distribution differences between the patches dataset and the ROIs dataset.

The results also shows that there's a little difference to note between soft and hard voting approaches in the task of classifying ROIs,and in that in general ResNet-18 outperforms ResNet-34 but the best model of each architecture will be used in the feature extraction step.

6.2 WSI Classifiers results

We trained three different architectures on the task of classifying WSIs (MMABC, ACMIL, HIPT-WSI). MMABC and ACMIL were trained on the compressed WSIs generated using different feature extractors: **ResNet-18**, **ResNet-34** (fine-tuned on BRACS' regions of interest), **ResNet-50** (trained on the Kather100k Dataset), and a **ViT-S/16** pre-trained using DINO on a substantial collection of 36,666 WSIs. **HIPT-WSI** was trained on the features generated from **HIPT-4096**.

MMABC models were trained using a learning rate of **0.0001**, a weight decay of **0.001**, and a dropout rate of **0.2**. The ACMIL models were trained with a learning rate of **0.00001**, a weight decay of **0.001**, a dropout rate of **0.2**, and cosine learning rate decay. HIPT was trained with a learning rate of **0.00001**, a weight decay of **0.001**, and a dropout rate of **0.35**.

When training **MMABC**, data augmentation was performed on the compressed WSIs. It involved randomly choosing one of the following operations: left, right, up, down shifting, horizontal or vertical flip, or rotating the image by 90 or 270 degrees. We also experimented with the same data augmentation when training **HIPT**.

Table 12 and Table 13 summarizes the obtained results accross five different metrics : AUC,Accuracy,F1 Score,Precision and Recall.

Model	Feature Extractor	LR Decay	AUC	F1 Score	Accuracy	Precision	Recall
MMABC	ResNet-18	no	0.773	0.503	0.620	0.539	0.566
	ResNet-34	no	0.816	0.539	**0.666**	0.563	0.608
	ResNet-50	no	0.731	0.465	0.550	0.586	0.516
	ViT-S/16	no	**0.864**	0.518	0.632	0.509	0.577
ACMIL	ResNet-18	no	0.763	0.520	0.551	0.518	0.524
		yes	0.794	0.486	0.632	0.430	0.573
	ResNet-34	no	0.777	0.557	0.597	0.555	0.566
		yes	0.779	0.441	0.574	0.382	0.520
	ResNet-50	no	0.678	0.399	0.517	0.359	0.468
	ViT-S/16	no	0.821	0.567	0.632	0.562	0.589
		yes	0.847	**0.653**	**0.666**	**0.653**	**0.658**

TABLE 12: Attention based classifiers results.

Model	Data Augmentation	AUC	F1 Score	Accuracy	Precision	Recall
HiPT	No	0.77	0.56	0.62	0.56	0.58
	Yes	0.81	**0.68**	**0.72**	**0.73**	**0.68**

TABLE 13: HIPT results.

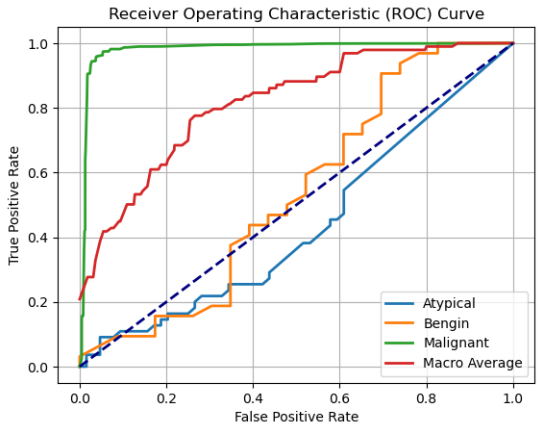
When analyzing ABNN models, ViT-S/16 demonstrated a substantial lead in terms of AUC when employed with both MMABC (0.864) and ACMIL (0.847). Furthermore, ACMIL-ViT, trained with learning rate decay, achieved an F1 score of 0.653, the highest overall among attention-based classifiers. It also shared the best accuracy (0.666) with MMABC-ResNet34, highlighting the effective data compression provided by ViT-S/16 despite its lower embedding dimension of 384 compared to ResNet-18 and ResNet-34's 512 and ResNet-50's 2048.

ResNet-50, trained on the Kather100k dataset, showed the poorest performance among all feature extractors, with MMABC-ResNet50 having the worst F1 score among MMABC models and ACMIL-ResNet50 showing the lowest results across all metrics.

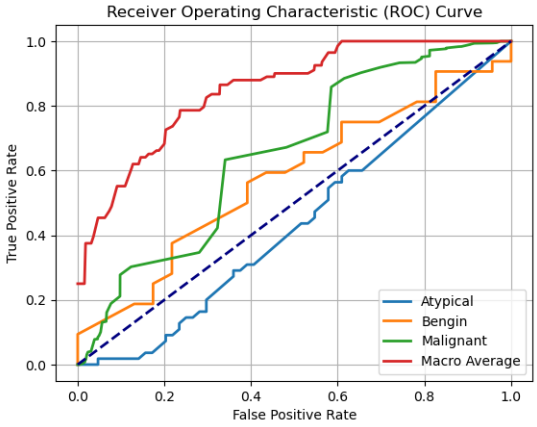
Focusing on MMABC alone, ResNet-34 achieved the best F1 score (0.53) and the best accuracy with a noticeable lead. ResNet-18 and ResNet-34 shared similar results in terms of these two metrics. For ACMIL models, ResNet-34 without learning rate decay had a better F1 score than ResNet-18, while ResNet-18 with learning rate decay achieved better accuracy.

The HIPT model demonstrated improved performance when data augmentation was applied. With data augmentation, the HIPT model achieved an AUC of 0.81, an F1 Score of 0.68, and an accuracy of 0.72. In contrast, without data augmentation, the performance metrics were lower, with an AUC of 0.77, an F1 Score of 0.56, and an accuracy of 0.62.

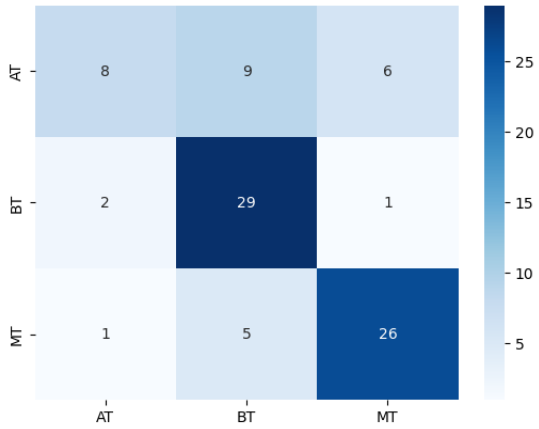
Comparing both models, it is evident that ABNN models, particularly those employing ViT-S/16, achieved superior performance in terms of AUC and accuracy. However, the HIPT model with data augmentation showed competitive results, especially in F1 score and accuracy. The data augmentation significantly enhanced the HIPT model's performance, emphasizing the importance of this technique in improving model robustness.



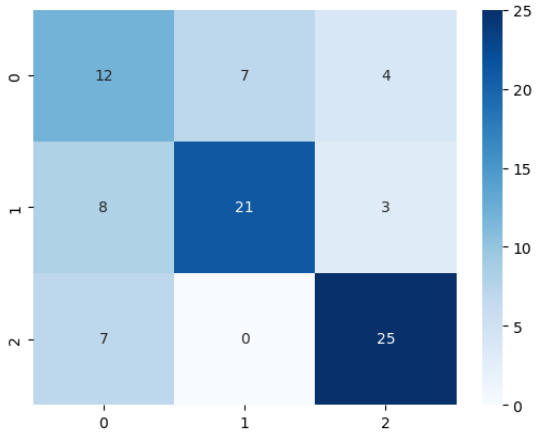
(A) Receiver Operating Characteristic Curve (HIPT)



(B) Receiver Operating Characteristic Curve (ACMIL)



(A) Confusion matrix (HIPT)



(B) Confusion matrix (ACMIL)

7 Information system

7.1 Anti Cancer Center

The Anti Cancer Center of Sidi Bel Abbés (CAC SBA) is one of 25 centers across the country, which are dedicated to the diagnosis and treatment of cancer, it was established in 2017 under the direct supervision of the Algerian Ministry of Health and Population.

7.2 Objective

The objective of this collaboration with CAC SBA, is to a create an information system for the purpose of automating data collection and the digitization of patient files. Since everything in their existing system is done manually, introducing an automated system yields many benefits such as:

- Saving a considerable amount of time for both patients and medics.
- Allowing the medical staff to do statistics and analyse the data very easily.
- Facilitating data sharing with other centers.
- Enabling the creation of datasets to develop dedicated AI models.

7.3 Overview

The system is divided into two parts: a web application in which the doctors manage patients files, and an inference desktop application in which the doctor uploads a WSI of a breast pathology and makes a prediction using the best AI model amongst our experimentations. Figure 20 represents the class diagram of the information system.

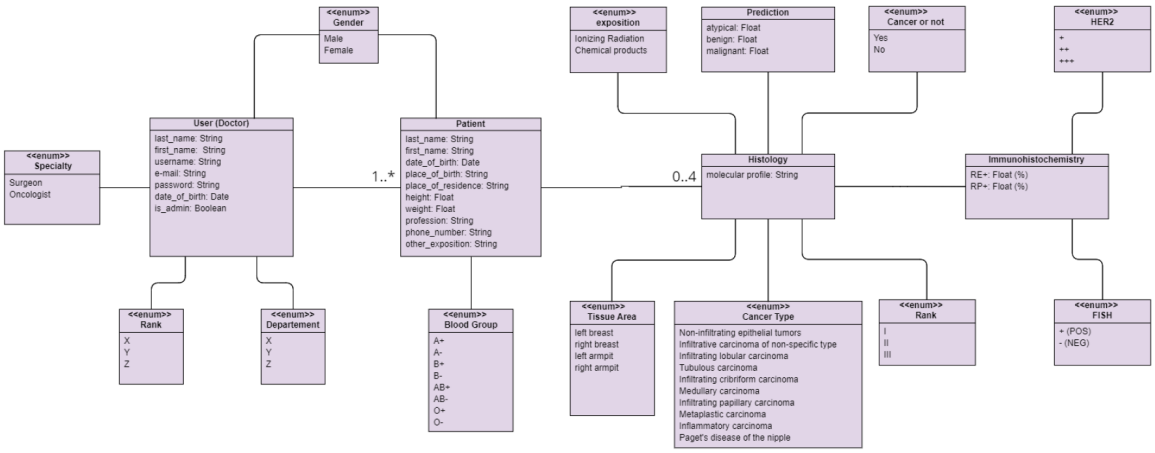


FIGURE 20: Class Diagram of the information System

7.3.1 Web application

The web application serves as the primary interface for data collection and managing patient information, it has two principal actors: an Admin and a Doctor.

Admin: The Admin is responsible for managing the system's users, specifically the doctors. Admin tasks include creating, modifying, deleting, and restoring doctors' passwords when necessary.

Doctor: The Doctor is the main user of the system, responsible for managing patient files. They can create, modify, and archive patient records, fill in patient information, view AI model prediction results for patients, and access statistical data. The overall use case diagram is shown in Figure 21.

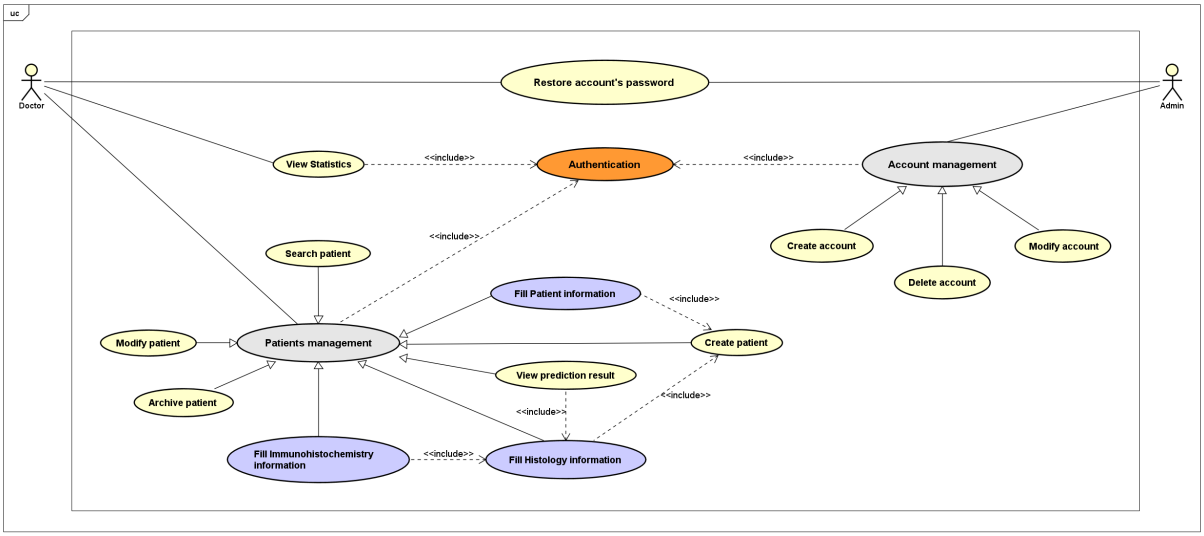


FIGURE 21: Use Case diagram of the web application

FIGURE 22: Home page

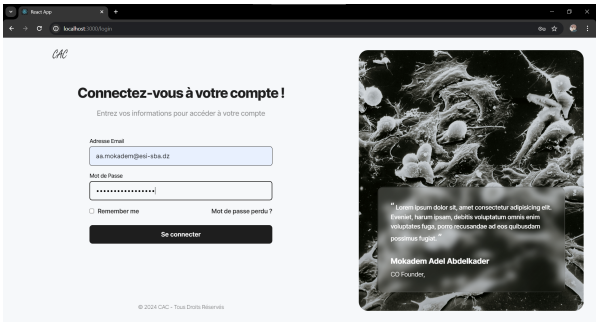


FIGURE 23: Login page

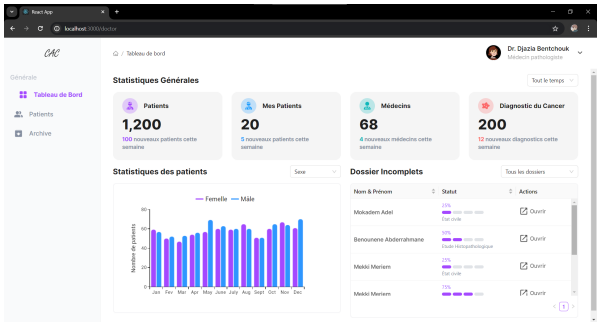


FIGURE 24: Dashboard page

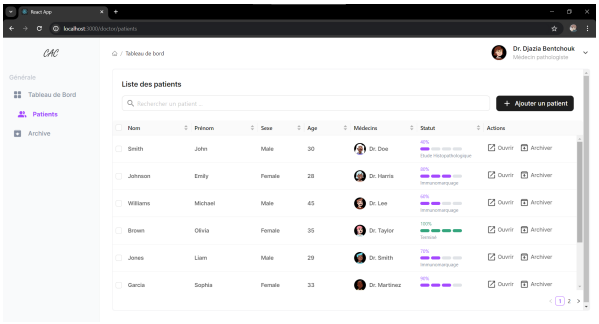


FIGURE 25: Patients list page

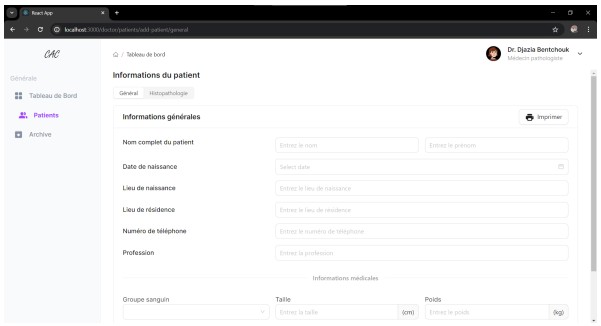


FIGURE 26: Patient information form

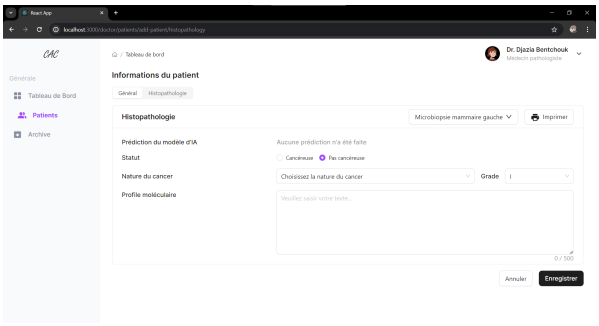


FIGURE 27: Patient's histology form

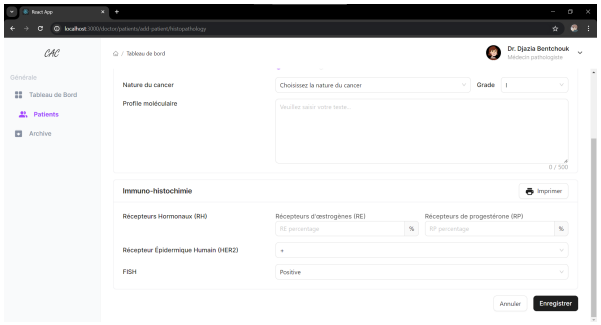


FIGURE 28: Patient's immunohistochemistry form

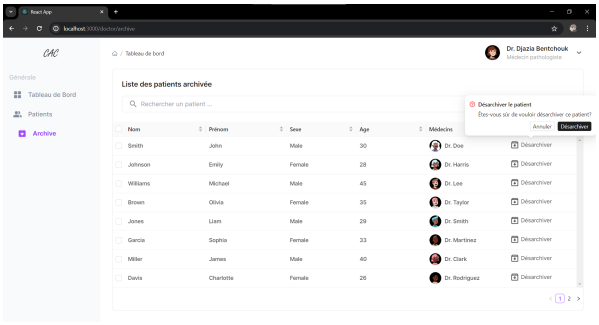


FIGURE 29: Archived patients page

7.3.2 Inference application

The inference application is designed for making predictions on histology biopsies of specific patients. Doctors can search for a patient, select an available histology, and then upload a WSI to obtain a prediction.

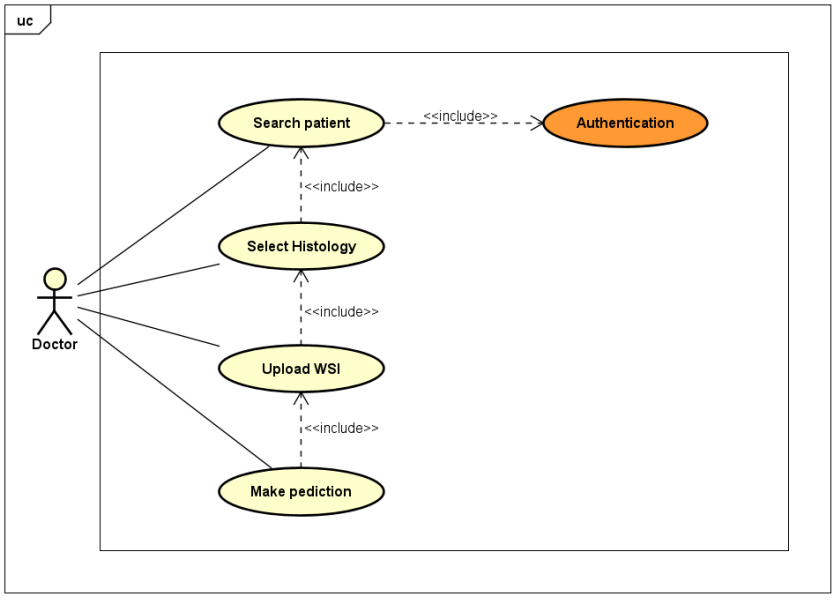


FIGURE 30: Use Case diagram of the inference application

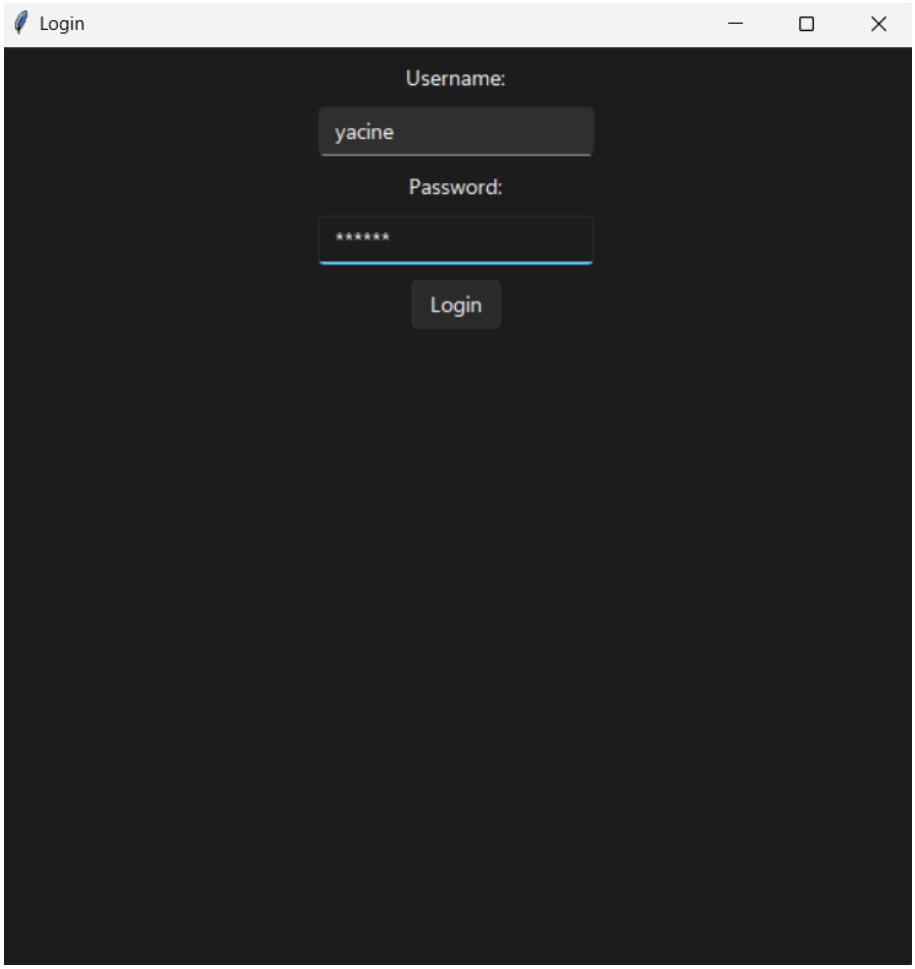


FIGURE 31: Login page

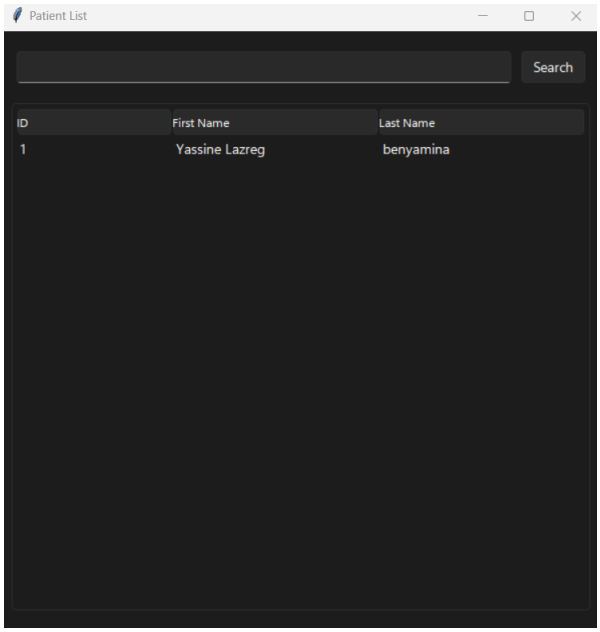


FIGURE 32: Doctor's patients list page

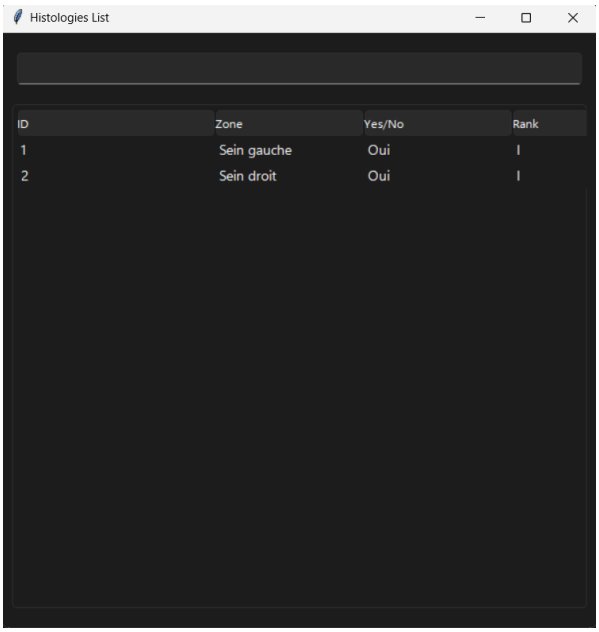


FIGURE 33: Patient's histologies list page

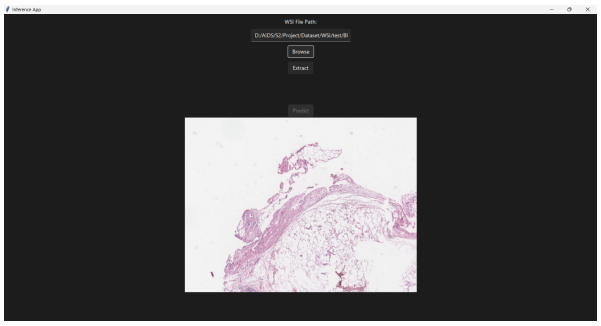


FIGURE 34: Selecting a WSI page

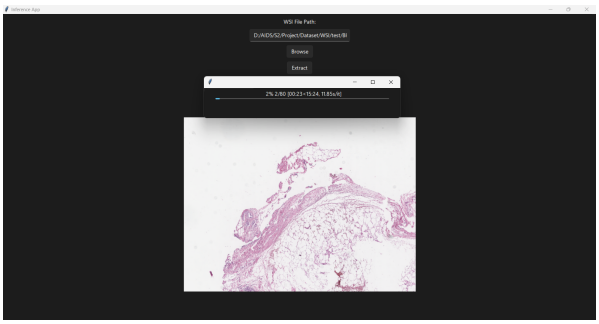


FIGURE 35: Feature Extraction on the WSI

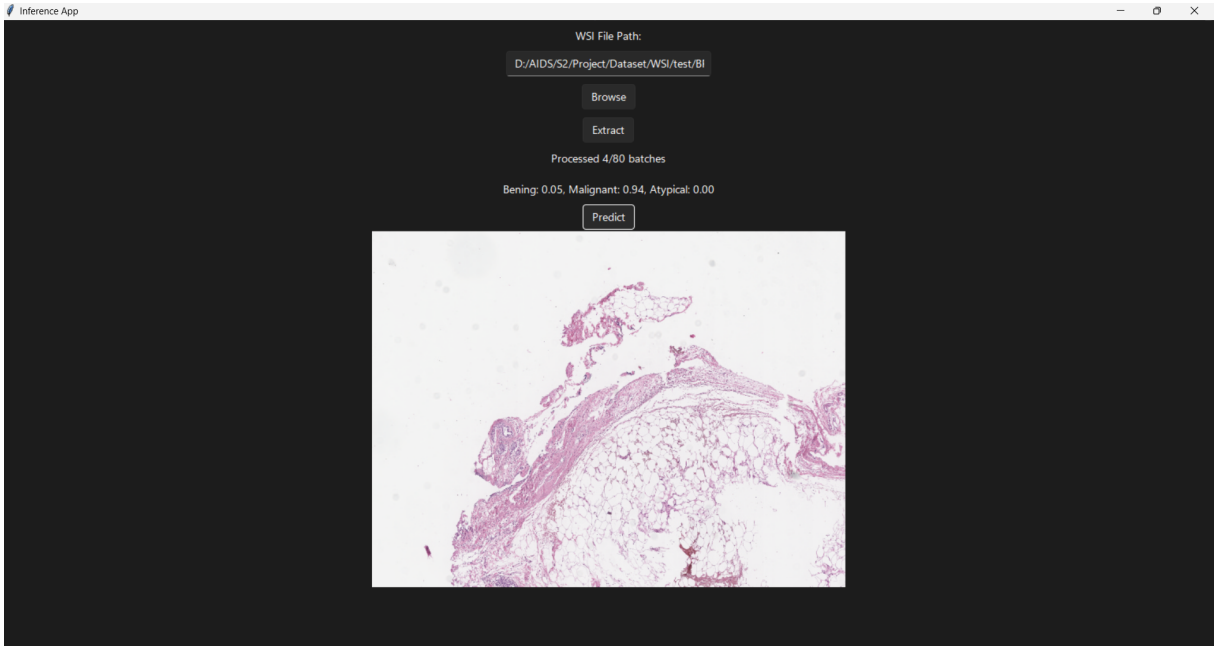


FIGURE 36: Prediction on the WSI

8 Conclusion

This multidisciplinary project aimed to develop an effective and robust system for breast cancer detection in whole slide histopathological images (WSIs), which are complex and high-resolution (gigapixels). A comprehensive approach was employed, utilizing advanced deep learning techniques, cutting-edge model architectures, and data preprocessing and augmentation strategies.

A key component was the use of OpenSlide, a powerful open-source library for processing and manipulating high-resolution WSIs, exploiting the hierarchy of svs files. This tool enabled efficient handling and analysis of the large image files, facilitating the extraction of relevant features for subsequent model training and evaluation.

To enhance feature extraction effectiveness, a fine tuning strategy was employed for the feature extractors. By fine tuning pretrained models on the ROIs from the BRACS dataset, the feature extractors could capture the most salient and discriminative characteristics of the histopathological images, leading to improved performance when applied to the larger WSIs.

The project explored and evaluated diverse deep learning architectures and techniques. Attention-based neural networks, such as those used with ResNet-18, ResNet-34, and the state-of-the-art Vision Transformer (ViT), were employed to leverage the power of self-attention mechanisms in capturing long-range dependencies and spatial relationships within the WSIs.
Additionally, the Hierarchical Image Processing Transformer (HIPT) model, designed to handle the multi-scale nature of WSIs, was explored. By incorporating data augmentation techniques like rotations, shifts, and flips, the HIPT model exhibited remarkable performance, outperforming the other models across all evaluation metrics except for the AUC.

The project's success can be attributed to the synergistic combination of advanced deep learning architectures, effective data preprocessing and augmentation strategies, and fine tuning of feature extractors. The results demonstrate the potential of these techniques in addressing the challenges of breast cancer detection in histopathological images, paving the way for more accurate and reliable diagnostic tools, taking the load off doctors.

Deploying such a system in real-world clinical settings would require rigorous validation and testing to ensure robustness and reliability across diverse patient populations and imaging protocols. Collaboration with domain experts, such as pathologists and oncologists, would be crucial in refining the system and addressing any potential biases or limitations.

Ultimately, the successful implementation of this breast cancer detection system could have far-reaching implications, empowering healthcare professionals with a powerful tool for early detection and accurate diagnosis, potentially leading to improved patient outcomes and more effective treatment strategies.

Moreover, the collaboration with the Anti Cancer Center of Sidi Bel Abbés (CAC SBA) demonstrates the critical role of a well-structured information system in healthcare. By automating data collection and digitizing patient files, the system not only enhances efficiency but also enables comprehensive data analysis, facilitates data sharing with other centers, and supports the development of AI models. This integrated approach represents a significant step forward in modernizing cancer diagnosis and treatment, promising considerable benefits for both medical staff and patients.

However, we do not intend to stop here. In the future, we aim to leverage Vision Mamba [20], a new architecture that is said to replace transformers due to its performance and computational efficiency. By exploring this novel and potentially revolutionary architecture, we strive to remain at the forefront of technological advancements, continuing to improve our breast cancer detection system. This proactive approach ensures that we stay ahead of the curve and contribute significantly to the field of medical AI.

References

[1] *Convolutional Neural Networks cheatsheet*. https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks. Accessed: 2024-06-05.

[2] Nadia Brancati et al. *BRACS: A Dataset for BReAst Carcinoma Subtyping in H&E Histology Images*. 2021. arXiv: 2111.04740 [q-bio.QM].

[3] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *arXiv e-prints*, arXiv:1512.03385 (Dec. 2015), arXiv:1512.03385. DOI: 10.48550/arXiv.1512.03385. arXiv: 1512.03385 [cs.CV].

[4] Ashish Vaswani et al. "Attention Is All You Need". In: *arXiv e-prints*, arXiv:1706.03762 (June 2017), arXiv:1706.03762. DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762 [cs.CL].

[5] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *arXiv e-prints*, arXiv:2010.11929 (Oct. 2020), arXiv:2010.11929. DOI: 10.48550/arXiv.2010.11929. arXiv: 2010.11929 [cs.CV].

[6] Zewen Li et al. "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.12 (2022), pp. 6999–7019. DOI: 10.1109/TNNLS.2021.3084827.

[7] Meng-Hao Guo et al. "Attention Mechanisms in Computer Vision: A Survey". In: *arXiv e-prints*, arXiv:2111.07624 (Nov. 2021), arXiv:2111.07624. DOI: 10.48550/arXiv.2111.07624. arXiv: 2111.07624 [cs.CV].

[8] Zhongyi Han Benzheng Wei Yuanjie Zheng Yilong Yin Kejian Li Shuo Li. *Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model*. 2017. URL: https://www.nature.com/articles/s41598-017-04075-z.

[9] Fabio A. Spanhol; Luiz S. Oliveira; Caroline Petitjean; Laurent Heutte. *BreaKHis Breast Cancer Histopathological Database*. https://ieeexplore.ieee.org/document/7312934. 2015.

[10] Teresa Araújo; Guilherme Aresta; Eduardo Castro; José Rouco; Paulo Aguiar; Catarina Eloy; António Polónia; Aurélio Campilho. *Classification of breast cancer histology images using Convolutional Neural Networks*. 2017. URL: https://journals.plos.org/plosone/articleid=10.1371/journal.pone.0177544#:~:text=Images%20are%20classified%20in%20four,nuclei%20and%20overall%20tissue%20organization..

[11] *BIOIMAGING challenge dataset*. http://www.bioimaging2015.ineb.up.pt/dataset.html. 2015.

[12] Nadia Brancati; Daniel Riccio; Maria Frucci. *Multi-classication of Breast Cancer Histology Images by Using a Fine- Tuning Strategy*. https://link.springer.com/chapter/10.1007/978-3-319-93000-8_87. 2018.

[13] *BATCH dataset, ICIAR 2018 Grand Challenge on BreAst Cancer Histology Images*.

[14] Yun Jiang; Li Chen; Hai Zhang; Xiao Xiao. *Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module*. 2019. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0214587.

[15] Nadia Brancati et al. "Gigapixel Histopathological Image Analysis Using Attention-Based Neural Networks". In: *IEEE Access* 9 (2021), pp. 8755287562. ISSN: 2169-3536. DOI: 10.1109/access.2021.3086892. URL: http://dx.doi.org/10.1109/ACCESS.2021.3086892.

[16] Richard J. Chen et al. "Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning". In: *arXiv e-prints*, arXiv:2206.02647 (June 2022), arXiv:2206.02647. DOI: 10.48550/arXiv.2206.02647. arXiv: 2206.02647 [cs.CV].

[17] Yunlong Zhang et al. *Attention-Challenging Multiple Instance Learning for Whole Slide Image Classification*. 2024. arXiv: 2311.07125 [cs.CV].

[18] Ming Y. Lu et al. *Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images*. 2020. arXiv: 2004.09666 [eess.IV].

[19] Mathilde Caron et al. *Emerging Properties in Self-Supervised Vision Transformers*. 2021. arXiv: 2104.14294 [cs.CV].

[20] Lianghui Zhu et al. *Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model*. 2024. arXiv: 2401.09417 [cs.CV].

[21] *What Is the Difference Between Fine-Tuning and Transfer-Learning?* https://www.geeksforgeeks.org/what-is-the-difference-between-fine-tuning-and-transfer-learning/. Accessed: 2024-06-04.

[22] *Transfer learning and fine-tuning*. https://www.tensorflow.org/tutorials/images/transfer_learning. Accessed: 2024-06-04.

[23] *Introduction to Convolution Neural Network*. https://www.geeksforgeeks.org/introduction-convolution-neural-network/. Accessed: 2024-06-05.

[24] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles". In: *Artificial Intelligence* 89.1 (1997), pp. 31–71. ISSN: 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(96)00034-3. URL: https://www.sciencedirect.com/science/article/pii/S0004370296000343.

[25] James Keeler, David Rumelhart, and Wee Leow. In: *Advances in Neural Information Processing Systems*. Ed. by R.P. Lippmann, J. Moody, and D. Touretzky.