MATHÉMATIQUES ET INFORMATIQUE
Sciences
Université de Paris

University of Paris
UFR Mathematics and Computer Science

TER5 Project Final Report

Master 1 Vision and Machine Intelligent

Khoualdia Besma – Chikhi Mohammed Wassim
Supervised by Mr Camille Kurtz And Mr Nicolas Loménie

Academic year 2024-2025

Contents

1 Introduction 1
1.1 Background and Motivation : 1
1.2 Objectives : 1
1.3 Project Overview and Methodology : 2

2 Related works 3

3 Contribution 4
3.1 Dataset : 4
3.2 Convolutional Neural Networks (CNN) : 5
3.2.1 Defintion : 5
3.2.2 Models used : 6
3.3 Transformers : 6
3.3.1 Definiton : 6
3.3.2 Models used : 7
3.4 Foundation models : 8
3.4.1 Definiton : 8
3.4.2 Models used : 8
3.5 Results : 9
3.5.1 Convolutional Neural Networks : 9
3.5.2 Transformers : 11
3.5.3 Foundation Models : 13
3.5.4 Discussion : 14

4 Conclusion 15
4.1 Study limitations : 15
4.2 Future works : 16

References 17

A Tables 19

Chapter 1

Introduction

1.1 Background and Motivation :

Digital imaging is widely used by pathologists for creation of static images using microscope-dedicated optical cameras and, more recently, using smartphones. The introduction of whole slide imaging (WSI) in 1999 provided the opportunity of digitally converting the entire tissue on glass slide into a high-resolution virtual slide (VS). In the last two decades, we have witnessed an exponential growth in technology of acquiring virtual slide as well as its applications in various subspecialties of pathology [1]. WSI, also commonly referred to as "virtual microscopy", aims to emulate conventional light microscopy in a computer-generated manner. Practically speaking, WSI consists of two processes. The first process utilizes specialized hardware (scanner) to digitize glass slides, which generates a large representative digital image (so-called "digital slide"). The second process employs specialized software (ie, virtual slide viewer) to view and/or analyze these enormous digital files [2].

1.2 Objectives :

The primary objective of our research project is to use modern artificial intelligence techniques to accurately extract and learn biomarkers from WSI data, focusing on breast cancer diagnosis, then we will do a comparative study on the used techniques and models. A biomarker is a biological phenomenon that can be difficult to find, yet indicates a clinically significant outcome or interim consequence. Biomarker applications include identifying, characterizing, and monitoring diseases. Additionally, biomarkers can act as prognostic indicators, inform individualized treatment plans, and anticipate and manage negative medication reactions. Understanding the fundamental link between a biomarker and its clinical result is crucial for adequately appreciating its significance [3].

In recent times, convolutional neural networks have been the go to technique to process this kind of medical imagine, because of their ability to capture local spatial features effectively. However, CNNs struggle with large-scale contextual understanding, which is very important in processing WSI, where the relevant information can be across large areas of tissue. To remedy this struggle, the research community started to look into other ways to process and analyze such data. They explored the use of transformers, which were originally designed for natural language processing (NLP), but have demonstrated strong performance in capturing long-range dependencies in visual data as well. The researchers also looked into foundation models such as UNI, because they also shown good performances by offering generalized data representations that can be adapted to pathology.

1.3 Project Overview and Methodology :

This project presents a comparative study of 3 different state of the art techniques, including CNN's, vision transformers and foundation models, we evaluated their performances and their effectiveness in classifying WSI images of colon cancer and breast cancer tissues. The methodology will consist of taking the preprocessed large WSI images into smaller image patches, then processing them by extracting high dimensional feature embeddings, which will then be processed by one of the architectures we mentioned to produce a final prediction at the slide level. These embeddings may also serve as learned biomarkers, they can provide insights into pathology and clinical decision making.

This research was carried out using the Breast Histopathology images dataset that contains Invasive Ductal Carcinoma (IDC) slides, which is the most common subtype of all breast cancers. This provided us with a realistic testing ground for the models study.Development was conducted in Python, using PyTorch deep learning frameworks.

Chapter 2

Related works

Over the last few years, foundation models have significantly reshaped both computer vision and natural language processing. In vision, models such as CLIP and Segment Anything [4, 5] demonstrated that pretraining on large-scale image-text pairs or segmentation tasks can yield powerful general-purpose representations. Similarly, in NLP, models like BERT [6] and GPT [7] have established a new paradigm by using self-supervised objectives to extract rich semantic knowledge from unlabeled text.

This shift has inspired researchers to explore foundation models specifically tailored to medical imaging and histology. In this context, the UNI model [8] stands out as a general-purpose, self-supervised transformer for computational pathology. Trained on over 100 million patches from more than 100,000 diagnostic slides, UNI has shown state-of-the-art performance on various downstream tasks such as classification, retrieval, and segmentation—while significantly reducing the need for manual labels.

The broader foundation for this movement lies in the Transformer architecture itself. Introduced by the Google team in 2017, the Transformer [7] replaced traditional convolutional and recurrent neural networks with self-attention mechanisms, enabling efficient modeling of long-range dependencies. In 2020, the Vision Transformer (ViT) [9] brought this concept to image classification, treating image patches as tokens and applying the same attention-based processing used in NLP. ViT proved remarkably effective, especially on large-scale datasets, but came with its own limitations: high memory requirements, long training times, and sensitivity to data quantity and positional encoding strategies.

Several improvements followed to adapt transformers more effectively to medical imaging tasks. Among them, CTransPath and hybrid models sought to combine local information from CNNs with the global context modeling of transformers, improving performance on histopathological tasks that require multi-scale understanding.

Overall, the related literature suggests a strong trend toward self-supervised, transformer-based models in digital pathology. Our work builds on these developments by comparing traditional CNNs, attention-based models, and foundation models, in order to assess their strengths in classifying breast cancer slides from the IDC dataset.

Chapter 3

Contribution

This chapter presents the dataset and the techniques we used during our comparative study, which are : Convolutional Neural Networks, Transformers and Foundation models. We gave brief definitions of these concepts, described their architectures and defined the models we implemented in our project.

3.1 Dataset :

Breast Histopathology Images Dataset represents the Invasive Ductal Carcinoma (IDC) which is the most common subtype of all breast cancers. To assign an aggressiveness grade to a whole mount sample, pathologists typically focus on the regions which contain the IDC. As a result, one of the common pre-processing steps for automatic aggressiveness grading is to delineate the exact regions of IDC inside of a whole mount slide. It consists of 162 whole mount slide images of Breast Cancer (BCa) specimens scanned at 40×. From these, 277,524 patches of size 50 × 50 were extracted (198,738 IDC negative and 78,786 IDC positive). Each patch's file name follows the format: `u_xX_yY_classC.png` — for example, `10253_idx5_x1351_y1101_class0.png`. Here, `u` is the patient ID (`10253_idx5`), `X` is the x-coordinate from where the patch was cropped, `Y` is the y-coordinate, and `C` indicates the class, where 0 represents non-IDC and 1 represents IDC [10].

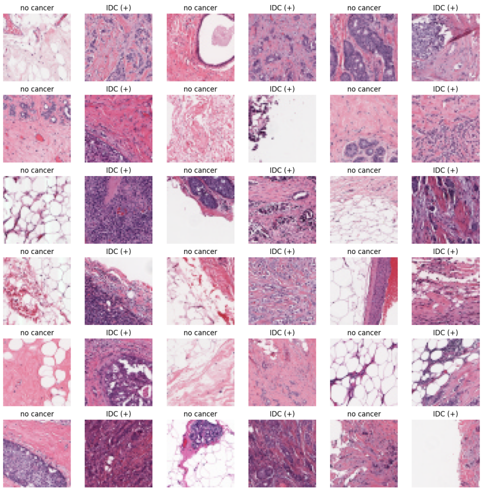


Figure 3.1: Sample from the breast Histopathology Dataset

3.2 Convolutional Neural Networks (CNN) :

In this section we defined the Convolutional Neural Network (CNN) architecture and we defined the two CNN based models we used in our research which are : EFFICIENT NET and RES NET.

3.2.1 Defintion :

In the field of Deep Learning, the CNN is the most famous and commonly employed algorithm. The main benefit of CNN compared to its predecessors is that it automatically identifies the relevant features without any human supervision. CNNs have been extensively applied in a range of different fields, including computer vision, speech processing, Face Recognition, etc. The structure of CNNs was inspired by neurons in human and animal brains, similar to a conventional neural network. More specifically, in a cat's brain, a complex sequence of cells forms the visual cortex; this sequence is simulated by the CNN. Goodfellow et al. identified three key benefits of the CNN: equivalent representations, sparse interactions, and parameter sharing. Unlike conventional fully connected (FC) networks, shared weights and local connections in the CNN are employed to make full use of 2D input-data structures like image signals. This operation utilizes an extremely small number of parameters, which both simplifies the training process and speeds up the network. This is the same as in the visual cortex cells. Notably, only small regions of a scene are sensed by these cells rather than the whole scene (i.e., these cells spatially extract the local correlation available in the input, like local filters over the input)[11].

An example of CNN architecture for image classification is illustrated in Figure 3.1.

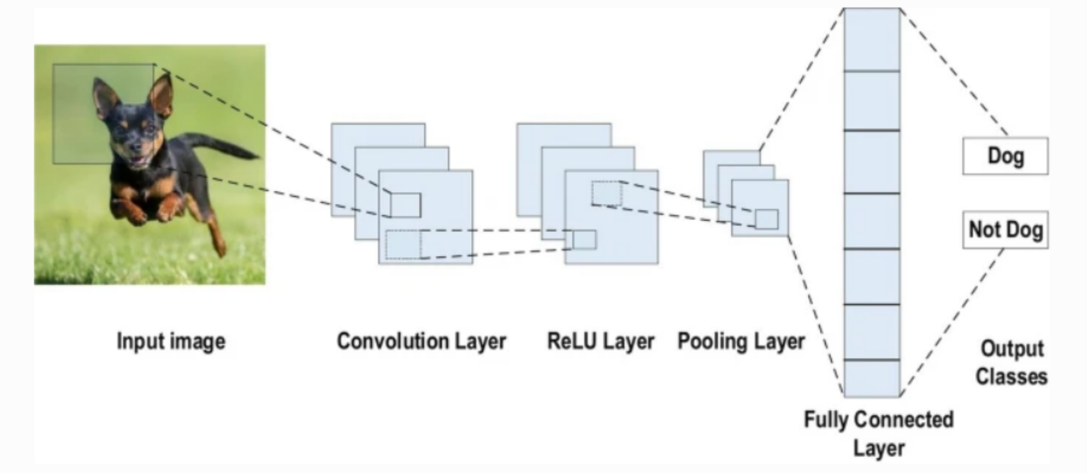


Figure 3.2: An example of CNN architecture for image classification[11]

A CNN is composed of a stacking of several building blocks: convolution layers, pooling layers (e.g., max pooling), and fully connected (FC) layers. A model's performance under particular kernels and weights is calculated with a loss function through forward propagation on a training dataset, and learnable parameters, i.e., kernels and weights, are updated according to the loss value through backpropagation with gradient descent optimization algorithm. ReLU, rectified linear unit[12].

3.2.2 Models used :

Here we present the two CNN based models we used for the classification of the Breast histopathology images Dataset which are : EFFICIENT NET and RES NET.

EFFICIENT NET :

EfficientNet is a convolutional neural network architecture (CNN) and scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient. Unlike conventional practice that arbitrary scales these factors, the EfficientNet scaling method uniformly scales network width, depth, and resolution with a set of fixed scaling coefficients. For example, if we want to use 2^N times more computational resources, then we can simply increase the network depth by α^N, width by β^N, and image size by γ^N, where α, β, and γ are constant coefficients determined by a small grid search on the original small model. EfficientNet uses a compound coefficient to uniformly scales network width, depth, and resolution in a principled way.

The compound scaling method is justified by the intuition that if the input image is bigger, then the network needs more layers to increase the receptive field and more channels to capture more fine-grained patterns on the bigger image[13].

RES NET :

ResNet, which is based on CNN architecture, is a breakthrough in the field of deep learning based on, particularly in tasks related to computer vision, and it has been widely adopted since its introduction. It mainly adds an architectural element known as a “skip connection” or “shortcut connection” in traditional CNN. Each layer feeds into the next one in a traditional deep neural network. However, in ResNet, the input to a layer is added to its output before being passed on to the next layer. Motivated by the problem of vanishing gradients, which can occur in intense networks and make them hard to train. ResNets found that it is easier to optimize the residual mappings than the original, unreferenced mappings. In practice, these shortcut connections allow the network to learn an identity function, ensuring that adding extra layers does not harm the network’s performance. This “residual learning” approach allows the training of much deeper networks than was previously possible. While networks before ResNet typically had a few dozen layers, ResNets can successfully train networks with 100, 200, or more layers, significantly improving their performance on benchmark tasks in image recognition and other areas[14].

3.3 Transformers :

In this section we defined the Transformers architecture and the two Transformer based models we employed in our research which are : Vision Transformer (VIT) and CTransPath.

3.3.1 Definiton :

The transformer is a neural network with a specific structure that includes a mechanism called self-attention or multi-head attention. Attention can be thought of as a way to build contextual representations of a token’s meaning by attending to and integrating information from surrounding tokens, helping the model learn how tokens relate to each other over large spans [15].In our research we focus on vision transformers.

3.3.2 Models used :

Here we present the two Transformer based models we used for the classification of the Breast histopathology images which are : Vision Transformer (VIT) and CTransPath.

Vision Transformer (VIT) :

Vision Transformers (ViT) is an architecture that uses self-attention mechanisms to process images. The Vision Transformer Architecture consists of a series of transformer blocks. Each transformer block consists of two sub-layers: a multi-head self-attention layer and a feed-forward layer.

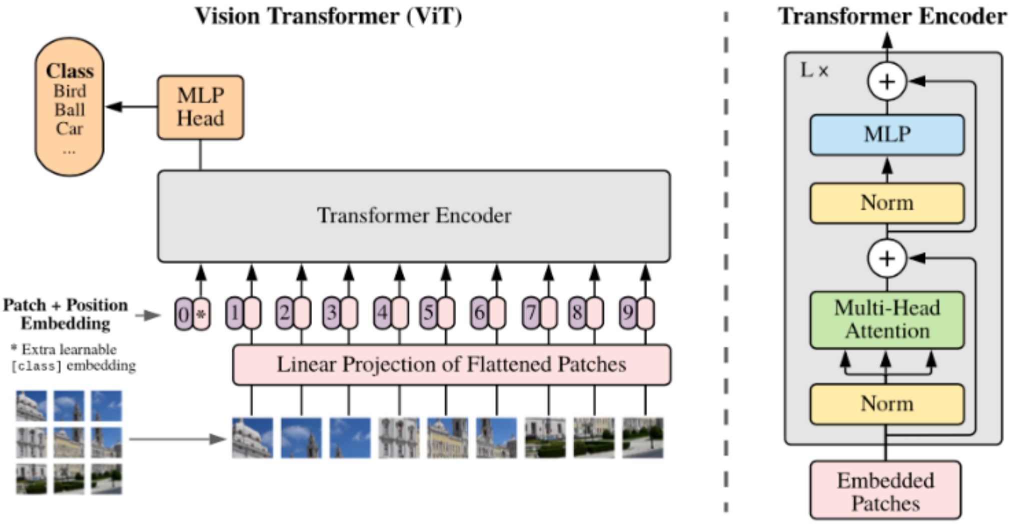


Figure 3.3: Vision Transformer ViT Architecture [16]

The overall structure of the vision transformer architecture consists of the following steps:

- Split an image into patches (fixed sizes).
- Flatten the image patches.
- Create lower-dimensional linear embeddings from these flattened image patches.
- Include positional embeddings.
- Feed the sequence as an input to a SOTA transformer encoder.
- Pre-train the ViT model with image labels, then fully supervise on a big dataset.
- Fine-tune the downstream dataset for image classification[16].

CTransPath :

CTransPath adopts a hybrid architecture combining a convolutional neural network (CNN) and a multi-scale Swin Transformer facilitating a collaborative local-global feature extraction. CTransPath is pre-trained on a large unlabeled dataset of pathology images from TCGA and PAIP, comprising approximately 15 million image patches cropped from over 30,000 WSIs by leveraging the semantically relevant contrastive learning (SRCL)[17].

3.4 Foundation models :

In this section we define the Foundation Model general architecture and we present the Foundation based model we employed in our research which is : the UNI model.

3.4.1 Definiton :

Current foundation models are large-scale artificial intelligence models, i.e., deep learning models, trained on a large amount of broad, typically unlabeled data. They cover well-known models popularized in the press, such as GPT-3, Dall-E 2, Florence and, widely adapted, early models, such as BERT . They arguably constitute the next milestone in the evolution of the main branch of AI, although they still face many diverse challenges and come with substantial risks spanning many sociotechnical concerns. In turn, interdisciplinary research is needed. For illustration, Stanford university alone has established the Center for Research on Foundation Models covering more than ten departments and more than 100 researchers contributing to this topic. Foundation models can serve as the foundation for many downstream applications, and they can be adjusted (or fine-tuned) with limited labeled training data for a specific task. Surprisingly, they can also be used to address tasks not explicitly trained for. They permit "in-context learning", i.e., during the training phase, the model extracts a rich set of patterns and broad skills from the diverse training data. In turn, a (language) model can perform downstream tasks simply by providing a prompt, i.e., a description of the task in natural language and, possibly, a few examples[18].

3.4.2 Models used :

Here we present the Foundation based model we used for the classification of the Breast histopathology images which is : the UNI model.

UNI :

UNI is a general-purpose, self-supervised vision encoder for anatomic pathology based on the Vision Transformer architecture, achieving state-of-the-art performance across 33 clinical tasks in anatomic pathology[19].

a) Slide distribution of Mass-100K, a large-scale and diverse pretraining dataset of 100 million tissue patches sampled from over 100,000 diagnostic whole-slide images across 20 major organ types.

b) UNI is pretrained on Mass-100K using the DINOv2 self-supervised training algorithm, which consists of: a mask image modeling objective and a self-distillation objective.

c) UNI outperforms other pretrained encoders on 33 clinical tasks in anatomical pathology (average performance of the 8 SegPath tasks reported).

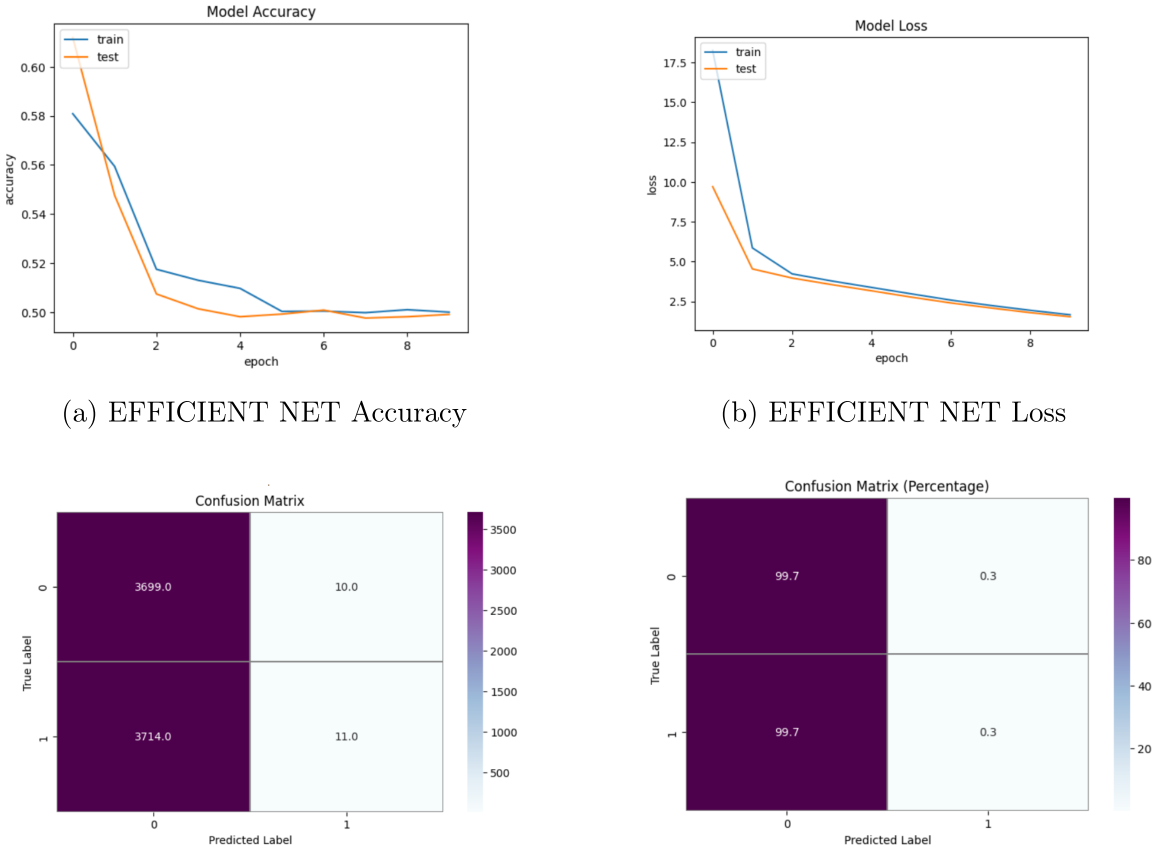
d) The evaluation tasks are comprised of ROI-level classification, segmentation, retrieval, prototyping, and slide-level classification tasks.

3.5 Results :

In this section we show the results (Accuracy, Loss) obtained from training the 5 different models on the Breast Histopathology images Dataset.

3.5.1 Convolutional Neural Networks :

EFFICIENT NET :



(a) EFFICIENT NET Accuracy

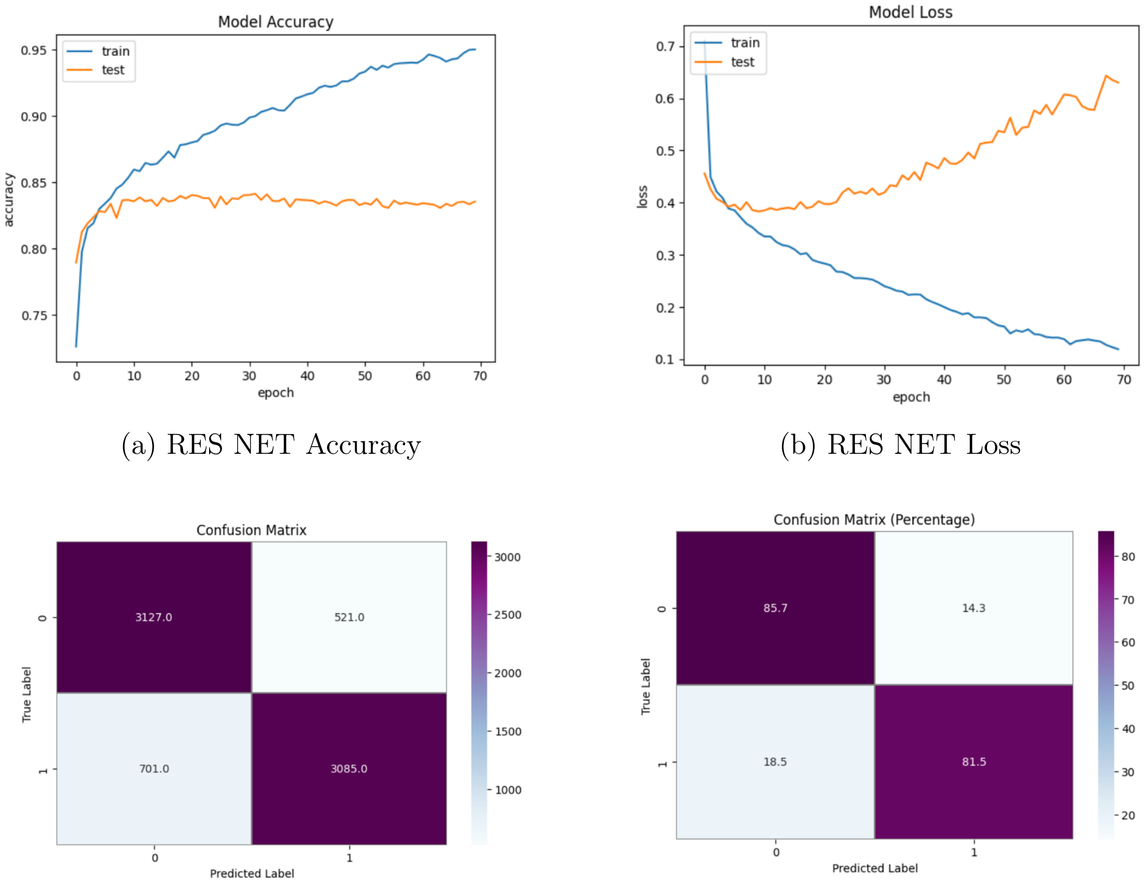
(b) EFFICIENT NET Loss

(c) EFFICIENT NET Confusion matrix

(d) EFFICIENT NET Confusion matrix %

Figure 3.4: Results of the EFFICIENT NET model

RES NET :



(a) RES NET Accuracy

(b) RES NET Loss

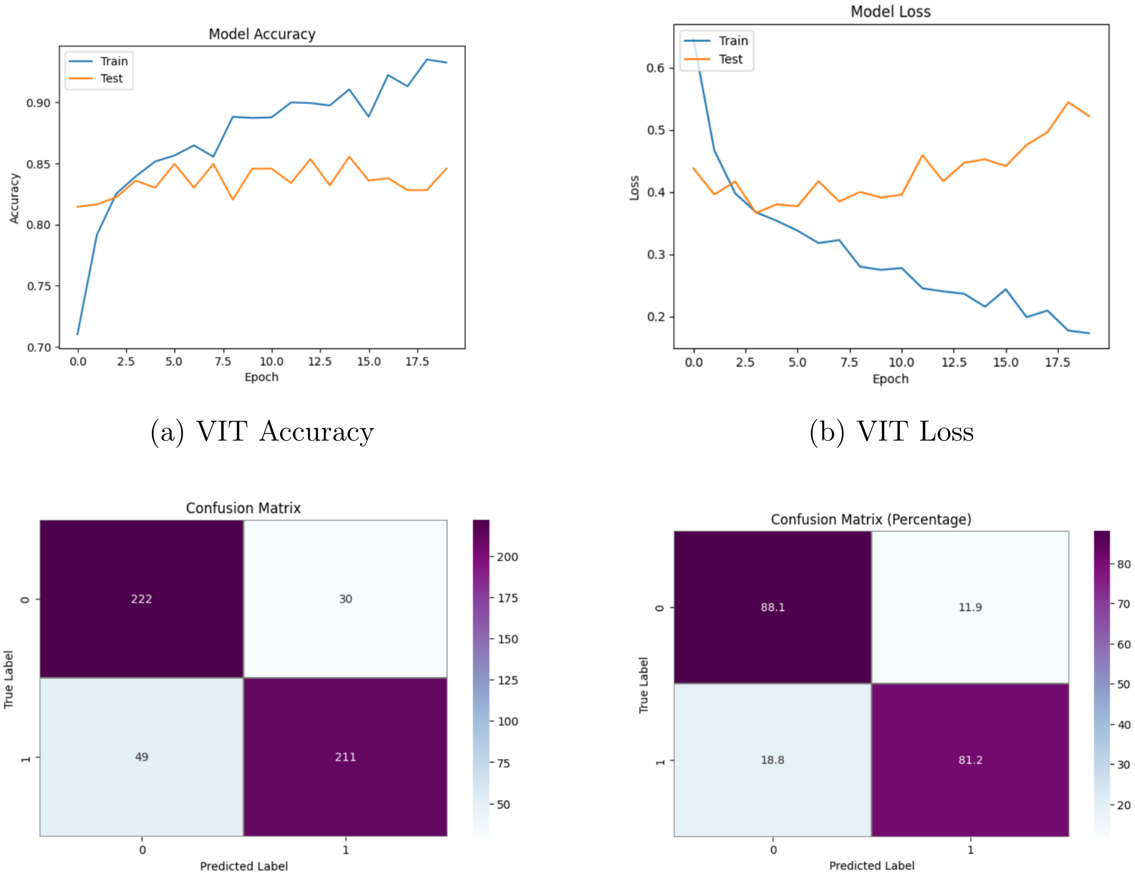
(c) RES Confusion matrix

(d) RES Confusion matrix %

Figure 3.5: Results of the RES NET model

3.5.2 Transformers :

Vision Transformer (VIT) :



(a) VIT Accuracy

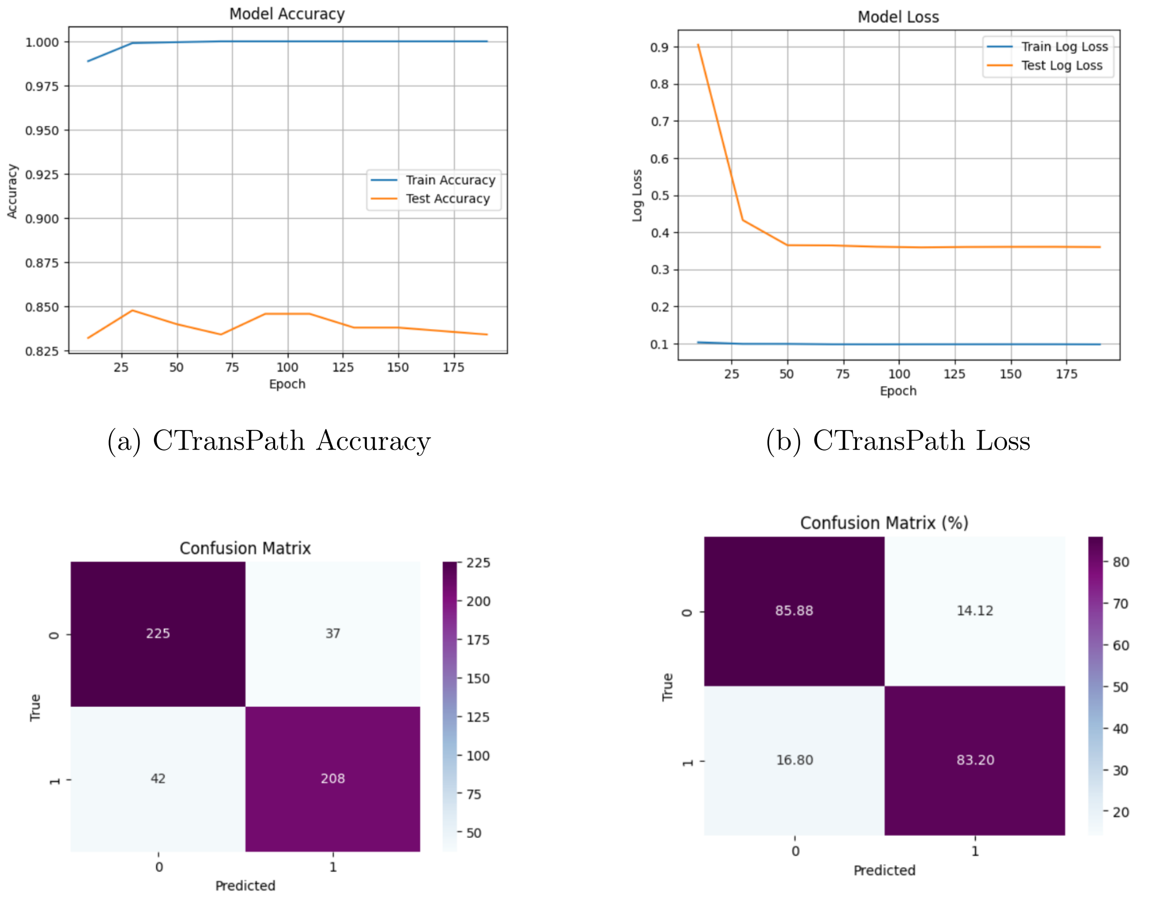
(b) VIT Loss

(c) VIT Confusion matrix

(d) VIT Confusion matrix %

Figure 3.6: Results of the VIT model

CTransPath :



(a) CTransPath Accuracy

(b) CTransPath Loss

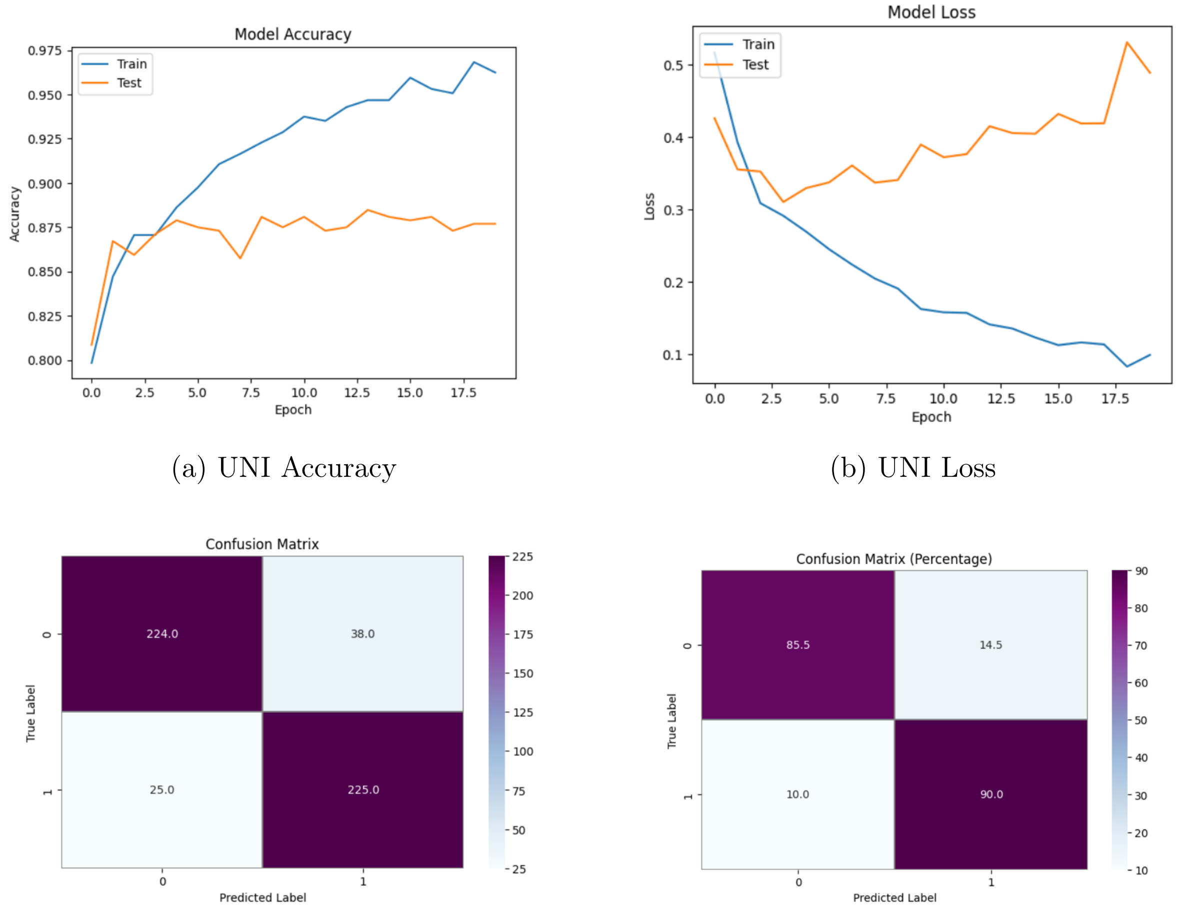
(c) CTransPath Confusion matrix

(d) CTransPath Confusion matrix %

Figure 3.7: Results of the CTransPath model

3.5.3 Foundation Models :

UNI :



(a) UNI Accuracy

(b) UNI Loss

(c) UNI Confusion matrix

(d) UNI Confusion matrix %

Figure 3.8: Results of the UNI model

Model	Train Accuracy (%)	Test Accuracy (%)	Train Loss	Test Loss
EfficientNet	65.00	58.00	18.00	10.00
ResNet	95.20	84.40	0.73	0.65
ViT	94.00	85.00	0.65	0.45
CTransPath	98.00	85.00	0.90	0.10
UNI	**97.50**	**87.50**	**0.54**	**0.43**

Table 3.1: Train and test accuracy and loss of the five evaluated models

3.5.4 Discussion :

Throughout this project, we set out not only to compare models, but to understand how different types of architectures behave when applied to real medical data—and what that means for future clinical applications.

Starting with Convolutional Neural Networks (CNNs), we appreciated their simplicity and efficiency. EfficientNet showed limited performance on this dataset, highlighting its difficulty in generalizing under low-resource settings, and ResNet proved stable and robust. These models are well-suited when computation is limited, or when interpretability and training time are important. But their local perspective makes them less adapted to the spatial complexity of histological slides, where relevant features may lie far apart across the tissue.

With Transformers, especially ViT and CTransPath, we observed an immediate gain in the ability to model this spatial complexity. ViT opened up new levels of context awareness, showing that attention mechanisms can outperform convolution when the task requires reasoning across the entire image. CTransPath refined that idea by merging local and global representations in a way that seemed particularly tailored to pathology. It was the first time in our experiments where the model felt aligned with the structure of the problem.

Then came UNI. What was impressive was not just the performance—which was the highest among all tested models—but the elegance of its approach. Trained on more than 100 million patches using a self-supervised method, While UNI achieved the best results overall, the performance gap with ViT and CTransPath was narrower than expected—suggesting that for some tasks, fine-tuned Transformers may be sufficient alternatives. The model's embeddings were so expressive that we could plug them into simple classifiers and still obtain excellent accuracy. That shift—from building models from scratch to leveraging pre-trained generalists—feels like a paradigm change in medical AI.

More than just numbers, what we learned is how scale, pretraining, and design philosophy impact outcomes. Training a CNN from scratch teaches you the mechanics; using a Transformer shows you the power of context; but using a Foundation Model like UNI teaches you how far we've come—and what may soon become standard in clinical pipelines.

This journey also highlighted the practical trade-offs. Not all hospitals have access to powerful GPUs, and not all datasets are large enough for Transformers. But with the right infrastructure and open-access pretrained models, the door is now open for accurate, scalable, and even explainable AI tools in pathology.

Chapter 4

Conclusion

This work explored the classification of histopathology images using five state-of-the-art deep learning models across three architectural families: CNNs, Transformers, and Foundation Models. Evaluating their performance on a breast cancer dataset, helped us to gain valuable insights on how each type of model processes medical imaging data.

Throughout this comparative study, we found that while CNNs offer efficient and reliable results, they failed in modeling global context—an essential aspect in pathology. Transformers addressed this with greater flexibility and better contextual understanding. But it was the foundation model, UNI, that stood out, achieving the highest performance with minimal fine-tuning thanks to its massive pretraining.

Overall, the project not only benchmarked model performance, but also helped us appreciate the trajectory of AI in medicine. From local patch processing to scalable, generalized embeddings, the future of computational pathology appears increasingly tied to the use of pretrained, context-aware, and self-supervised models.

The following sections detail the practical limitations we faced, and outline several directions for future work.

4.1 Study limitations :

Despite the promising results we have obtained from this comparative study, we have encountered few significant limitations throughout the project. One of the primary issues we faced was the sizes of the whole slide images datasets, there were a large volume of high-resolution images to process, which posed storage and processing difficulties. Being unable to load them in order to import the different models, led to slower data handling and longer training time reaching up to 10 hours per code. In addition, our machines were limited in both GPU memory and processing speed which became another challenge in training these deep learning models on such large datasets. This constrained the type of the models we could use and the batch sizes we could handle, forcing us to make a trade off in both model complexity and training time. Facing these limitations affected our ability to run extensive hyper parameters tuning, experiment with more complex models or use bigger datasets.

4.2 Future works :

In order to improve our research, get better classification results and address the limitations we have encountered, in the future we can do few changes in the project methodology, such as: gaining access to more powerful hardware, with high performance GPU or cloud based computing platforms. Additionally, we can incorporate large and more diverse datasets, branching out to other types of cancers, which will help improve the generalization and reduce overfitting, we can also explore a way to reduce the sizes of the dataset with more efficient patching strategies, that can help reduce the volume of redundant and less informative data within the WSI. Finally, we can incorporate self-supervised or semi supervised learning techniques that could help process unlabeled slides and improve robustness.

Kaggle Notebooks

All experiments and implementations were documented and executed via Kaggle notebooks. These are publicly accessible:

- **ResNet :** https://www.kaggle.com/code/besmakhoualdia/resnet
- **EfficientNet :** https://www.kaggle.com/code/besmakhoualdia/efficient-net
- **ViT :** https://www.kaggle.com/code/besmakhoualdia/vit-model-base-2
- **CTransPath :** https://www.kaggle.com/code/besmakhoualdia/ctranspath-base-2
- **UNI :** https://www.kaggle.com/code/besmakhoualdia/uni-base2

References

[1] Wright AM, Smith D, Dhurandhar B, Fairley T, Scheiber-Pacht M, Chakraborty S, Gorman BK, Mody D, and Coffey DM. *Digital slide imaging in cervicovaginal cytology: a pilot study.* 2013. DOI: 10.5858/arpa.2012-0430-OA. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7522141/ (page 1).

[2] Gifford AJ, Colebatch AJ, Litkouhi S, and al. *Remote frozen section examination of breast sentinel lymph nodes by telepathology.* 2012. DOI: 10.1111/j.1445-2197.2012.06191.x. URL: https://pubmed.ncbi.nlm.nih.gov/22924988/ (page 1).

[3] Allegra C.J. and Jessup J.M.and Somerfield M.R.and Hamilton S.R.and Hammond E.H.and Hayes D.F.and McAllister P.K.and Morton R.F.and Schilsky R.L. *American Society of Clinical Oncology Provisional Clinical Opinion: Testing for KRAS Gene Mutations in Patients With Metastatic Colorectal Carcinoma to Predict Response to Anti–Epidermal Growth Factor Receptor Monoclonal Antibody Therapy.* 2009. DOI: 10.1200/JCO.2009.21.9170. URL: https://ascopubs.org/doi/abs/10.1200/JCO.2009.21.9170 (page 1).

[4] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, and al. *Learning transferable visual models from natural language supervision.* 2021. URL: https://proceedings.mlr.press/v139/radford21a (page 3).

[5] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, and al. *Segment anything.* 2023. DOI: 10.1109/ICCV51070.2023.00371. URL: https://ieeexplore.ieee.org/document/10378323 (page 3).

[6] Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers).* 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/ (page 3).

[7] Brown Tom, Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan, Jared D, Dhariwal, Prafulla, Neelakantan, and al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems.* Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf (page 3).

[8] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. "Towards a general-purpose foundation model for computational pathology". In: *Nature Medicine* 30.3 (2024), pp. 850–862. DOI: 10.1038/s41591-024-02857-3. URL: https://pubmed.ncbi.nlm.nih.gov/38504018/ (page 3).

[9] Yingzi Huo, Kai Jin, Jiahong Cai, Huixuan Xiong, and Jiacheng Pang. "Vision Transformer (ViT)-based Applications in Image Classification". In: May 2023, pp. 135–140. DOI: 10.1109/BigDataSecurity-HPSC-IDS58521.2023.00033 (page 3).

[10] Janowczyk A and Madabhushi A. *Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases*. 2016. DOI: 10.4103/2153-3539.186902. URL: https://www.ncbi.nlm.nih.gov/pubmed/27563488 (page 4).

[11] Alzubaidi L, Zhang J, Humaidi A J, and al. "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions." In: *J Big Data* 8.53 (2021). DOI: 10.1186/s40537-021-00444-8. URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8#article-info (page 5).

[12] Yamashita R, Nishio M, Do R K G, and al. "Convolutional neural networks: an overview and application in radiology". In: *Insights Imaging* 9 (2018), pp. 611–629. DOI: 10.1007/s13244-018-0639-9. URL: https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9 (page 5).

[13] Mingxing Tan and Quoc V Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *CoRR* abs/1905.11946 (2019). arXiv: 1905.11946. URL: http://arxiv.org/abs/1905.11946 (page 6).

[14] Nan Luo, Xiaojing Zhong, Luxin Su, Zilin Cheng, Wenyi Ma, and Pingsheng Hao. "Artificial intelligence-assisted dermatology diagnosis: From unimodal to multimodal". In: *Computers in Biology and Medicine* 165 (2023), p. 107413. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2023.107413. URL: https://www.sciencedirect.com/science/article/pii/S0010482523008788 (page 6).

[15] Daniel Jurafsky and James H Martin. "The transformer". In: *Speech and Language Processing* (2025). URL: https://web.stanford.edu/~jurafsky/slp3/9.pdf (page 7).

[16] Boesch Gaudenz. ""Vision Transformers (ViT) in Image Recognition: Full Guide". In: *viso.ai* (2023). URL: https://viso.ai/deep-learning/vision-transformer-vit/ (page 7).

[17] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. "Transformer-based unsupervised contrastive learning for histopathological image classification". In: *Medical Image Analysis* 81 (2022), p. 102559. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2022.102559. URL: https://www.sciencedirect.com/science/article/pii/S1361841522002043 (page 8).

[18] Johannes Schneider. *Foundation models in brief: A historical, socio-technical focus*. 2022. arXiv: 2212.08967 [cs.AI]. URL: https://arxiv.org/abs/2212.08967 (page 8).

[19] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H. Song, Muhammad Shaban, Mane Williams, Anurag Vaidya, Sharifa Sahai, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Walt Williams, Long Phi Le, Georg Gerber, and Faisal Mahmood. *A General-Purpose Self-Supervised Model for Computational Pathology*. 2023. arXiv: 2308.15474 [cs.CV]. URL: https://arxiv.org/abs/2308.15474 (page 8).

Appendix A

Tables

In this section we gathered tables containing related works of the 3 architectures we employed in our research, this gave us an insight on the models, datasets used and the results we can expect.

Convolutional Neural Networks (CNN) :

Year	**Dataset(s)**	**Domain**	**Models used**	**Objective**	**Outcome**	**Link**
2019	Kaggle Diabetic Retinopathy (DR) dataset, skin images with artificial biomarkers	Domaine médical	CNN, GNN	Localizing biomarkers	AUC = 0.397	Springer Link
2022	The Cancer Genome Atlas (TCGA), HEROHE (European Congress on Digital Pathology challenge dataset)	Digital Pathology: Prediction of molecular tumor biomarkers from H&E histopathology images	CNNs, Transformers, GNNs, GANs	Predicting molecular biomarkers from H&E using self-supervised learning and domain adaptation techniques	Improved performance on external cohorts	MDPI Link

Year	Dataset(s)	Domain	Models used	Objective	Outcome	Link
2021	West China Hospital, Sichuan University dataset of 25 H&E-stained IDC slides	Breast Cancer, Digital Pathology	CNN	Predict pathological complete response (pCR) to neoadjuvant chemotherapy (NAC) in breast cancer	Accuracy = 0.853	Springer Link
2024	TCGA, Clinical Proteomic Tumor Analysis Consortium (CPTAC)	Digital Pathology, Precision Medicine	CNNs, ResNet34, Autoencoder	Predict multi-omic biomarkers (genomic, transcriptomic, proteomic) from routine H&E images	50% models: AUC 0.644, 25% models: AUC 0.719	Nature Link

Transformers :

Year	Dataset(s)	Domain	Models used	Objective	Outcome	Link
2023	BLCA, BRCA, GBMLGG, LUAD and UCEC datasets	Domaine médical	Neural network termed pattern-perceptive survival transformer (Surformer)	Cancer survival prediction from WSIs	Accuracy 0.67	ScienceDirect
2023	TCGA (The Cancer Genome Atlas)	Digital pathology	Multimodal transformer (PathOmics)	Prédiction de la survie liée au cancer du côlon et du rectum	—	Springer Link

Year	Dataset(s)	Domain	Models used	Objective	Outcome	Link
2024	Deux ensembles de données sur le cancer colorectal	Digital Pathology	Modèle basé sur Whole Slide Image (WSI) avec techniques de prompting et grands modèles de langage (LLMs)	Prédiction des biomarqueurs génétiques dans le cancer colorectal	AUC = 91.49%	Springer Link
2023	HER2 pour le cancer du sein	Digital pathology	Vision Transformer (ViT)	Détection de régions d'intérêt (ROI) pour l'analyse des Whole Slide Images (WSI) en pathologie, avec un focus sur le grade HER2 dans le cancer du sein	Accuracy = 99%	Nature Link

Foundation Models :

Year	Dataset(s)	Domain	Models used	Objective	Outcome	Link
2025	Non spécifiée, déployée dans plusieurs hôpitaux à travers le monde	Pathologie computationnelle	Deep Learning (modèles agnostiques aux biomarqueurs), STAMP (Solid Tumor Associative Modeling in Pathology)	Prédiction des biomarqueurs directement à partir des WSI	—	Nature
2024	Dataset of 587,196 whole slide images WSI	Histopathology, specifically H&E-stained slides	PRISM (Slide-level foundation model)	Cancer detection	AUC entre 0.95 et 0.97	arXiv
2024	423 whole-slide images (WSIs) with 8% mismatch repair (MMR) deficiency cases	Digital pathology	Self-supervised learning (SSL), weakly supervised learning	Predicting key colorectal cancer biomarkers	AUROC = 0.9466	Springer
2024	335,645 whole-slide images (WSIs), 423,122 synthetic captions from multimodal AI copilot	Computational Pathologie	TITAN, a multimodal whole slide foundation model	Rare disease retrieval and cancer prognosis	TITAN outperforms PRISM (balanced accuracy +121.9%)	arXiv

Author Contributions

All preprocessing, model implementation, training, visualization, and result analysis were conducted by the authors. This includes the design of the pipeline, selection of architectures, and evaluation. The project was done as part of a collaborative TER with supervision by Mr. Camille Kurtz and Mr. Nicolas Loménie.