UNIVERSITATEA POLITEHNICA DIN BUCUREȘTI
FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE
DEPARTAMENTUL DE CALCULATOARE

PROIECT DE DIPLOMĂ

Clasificarea Tipurilor de Cancer din Imagini cu Ganglioni Limfatici Folosind Învățarea Profundă

David Nicolae Mantu

Coordonator științific:

Șl.dr.ing. Radu Ioan Ciobanu

BUCUREȘTI

2023

UNIVERSITY POLITEHNICA OF BUCHAREST
FACULTY OF AUTOMATIC CONTROL AND COMPUTERS
COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

DIPLOMA PROJECT

Deep Learning-based Classification of Cancer Types from Lymphatic Nodes Images

David Nicolae Mantu

Thesis advisor:

Șl.dr.ing. Radu Ioan Ciobanu

BUCHAREST

2023

CONTENTS

SINOPSIS

Computer Vision are un impact din ce în ce mai mare în domeniul patologiei deoarece permite o întelegere mai buna a imaginilor patologice prin utilizarea învățării profunde și a învățării prin transfer. O aplicație importantă în acest domeniu este clasificarea nodulilor limfatici colorați cu hematoxilină și eozină, care necesită în mod normal o evaluare microscopică amănunțită din partea patologilor. Analiza probelor durează mult timp și implică un grad semnificativ de subiectivitate, deoarece în multe cazuri opiniile patologilor diferă, nivelul de incertitudine fiind foarte mare. Acest fapt este deosebit de îngrijorător, având în vedere mizele mari implicate în diagnosticarea, prevenirea și tratamentul cancerului. Această teză abordează problema diferențierii între diferite tipuri de cancer doar prin analizarea imaginilor cu ganglioni limfatici care conțin metastaze. Din câte stim, această probemă nu a mai fost înca abordată in literatura de specialitate. În anii trecuți, pentru a genera predicții pentru diferite tipuri de cancer în imaginile patologice, au fost utilizate arhitecturi CNN state-of-the-art, preantrenate pe seturi de date mari precum ImageNet. Această teză propune o metodă de combinare cu succes a seturilor de date cu ajutorul tehnicilor de procesare a imaginilor si arată impactul invățării prin transfer pentru clasificarea eșantioanelor de țesut colorate cu H&E din secțiuni de noduli limfatici care conțin metastaze sau cancer primar. Rezultatele pentru diferențierea între tipurile de cancer sunt foarte încurajatoare, obținându-se o acuratețe de 87.4%. Aceste rezultate pot avea un impact semnificativ în practica clinică, îndrumând patologii să ia o decizie mai bună în diferențierea limfomului de cancerul metastatic.

ABSTRACT

Computer Vision has an increasingly greater impact in the field of Pathology by enabling better understanding of tissue data through the use of deep learning and transfer learning. An important application in this field is the classification of hematoxylin and eosin stained lymph node sections, which normally requires extensive microscopic assessment by pathologists. This is highly time-consuming and involves a significant degree of subjectivity, since in a considerable amount of cases pathologists' opinions differ, having a high level of uncertainty. This is particularly concerning given the high stakes involved in cancer diagnosis, prevention and treatment. This thesis addresses the problem of differentiating between cancer types within lymphatic nodes. Specifically, we are addressing the difference between lymphoma and breast cancer metastasis. As far as we know, this problem has not yet been addressed in the literature. In the past years, to generate predictions for cancer type classification from histopathological images, state-of-the-art CNN's were used, after being fine tuned on large datasets such as ImageNet. This thesis proposes a method for combining datasets using image processing techniques and shows how to leverage transfer learning for the successful classification of HE stained tissue samples of lymph node sections containing metastatic or primary cancer. The results for distinguishing between cancer types are highly encouraging, obtaining a 87.4% accuracy. These results can have a significant impact into the clinical practice by helping pathologists make better decisions in differentiating lymphoma from metastatic cancer, with implications in cancer care and treatment.

ACKNOWLEDGMENTS

I express my gratitude Radu Ioan Ciobanu and Octavian Bucur (secondary advisor) for their invaluable guidance and support during the course of my research project. Their experience greatly contributed to the successful completion of my work.

1 INTRODUCTION

In this chapter we are providing an overview of the current state of pathology. We are identifying the most significant challenges encountered and discuss potential solutions to these problems, primarily focusing on the approach proposed in our paper.

1.1 Context

The application of computer vision in the field of pathology has grown in popularity in the recent years after showing great potential for improving the accuracy and efficiency of disease diagnosis: [20, 7, 4]. In pathology, a common practice for cancer identification in patients is the analysis of hematoxylin and eosin (H&E) stained tissue samples. These substances are used to highlight the cell structure and are required by pathologists to provide an accurate diagnosis. Computer Vision serves as a tool to improve this process. Deep learning models have the potential to discern intricate data relationships that might elude human observation, even with the assistance of high-powered microscopy. A general approach to this task is by fine tuning a state-of-the-art CNN on a particular dataset [5].

1.2 Problem

One of the main problems associated with this method is the limited availability of large, consistently labeled datasets. This means that results obtained on a particular dataset may not be easily reproducible on another dataset containing similar types of images acquired using the same methodology. The issue lies in the numerous conditions involved in capturing these images and the absence of a global standard for the process. This makes it impossible to set a standard in this field. Furthermore, the matter is further complicated by strict data protection regulations concerning patient confidentiality, which impose additional steps when acquiring images.

1.3 Objective

Despite these challenges, our objective is to process images from multiple, publically available datasets and extract relevant features that can effectively unify the datasets under a single representation. By doing so, we aim to develop a robust classification model that can accurately distinguish between the presence or absence of cancerous tissue and identify between

various types of cancers in a variety of histopathological images. The resulting dataset will be composed of images representing the same anatomical region (lymph nodes) and will contain multiple pathologies (lymphoma and breast cancer metastasis to the lymph node). This approach signifies a vital step towards mitigating the previously mentioned problems and may serve as a starting point for future investigations in this domain.

1.4 Proposed Solution

To address this problem, we leverage two datasets containing H&E stained tissue samples and process the images to unify them under a single dataset. Our approach seeks to facilitate multi-class classification tasks by attempting to assess not only if an image contains cancer but also determining its specific type, if present. The experiments conducted in this paper are the following:

- We develop a patch extraction algorithm to acquire images from whole slide images (WSI) slides. This method relies on the availability of the region of interest (ROI) coordinates. Positive and negative patches are selected efficiently at a 10x resolution to take into consideration the computational challenges that a higher resolution might introduce to the training process.
- We propose an image processing pipeline to perform patch selection, background elimination and color normalization to have a good starting point for feature extraction, clustering and image classification.
- We use a K-mean clustering algorithm to evaluate the separability of the images in the unified dataset.
- We perform transfer learning by fine tuning a pre-trained ResNet model on the Patch Camelyon dataset.
- We address the multiclass classification problem by focusing on distinguishing between breast cancer and primary lymph node cancer. We accomplish this by fine-tuning a ResNet model on the unified dataset.

1.5 Results

For the clustering task, we prepare the data by randomly selecting 5000 samples from each dataset following a 50/50 split between positive and negative labels. We rely on the features extracted by a pre-trained ResNet-18 model. The model is kept untouched apart from the last classifier layer which is removed. For clustering, the feature vectors are reduced to two dimensions using Principal Component Analysis (PCA). We performs multiple runs with different random initializations to increase the chances of finding an optimal solution. We get a final Accuracy score of 53.4%. This result highlights the characteristic of the pre-trained ResNet-18 model, which extracts features that exhibit a significant level of similarity within

their respective datasets. When these features are applied to a clustering algorithm, they tend to converge within a single space, as indicated by the achieved accuracy of 53%. This shows that despite originating from distinct datasets, the extracted features share a great level of overlap, enabling us to effectively combine all images under a single dataset. This method may be useful in the real world, considering the limited availability of large, labeled datasets. For binary classification, we trained the same ResNet-18 architecture on the PatchCamelyon [3] and DLBCL-Morphology [2] and get these individual accuracy scores: 86.5%, 95.5%. For the multiclass problem, we have 3 classes (one negative class and two positive classes, one for each cancer type) and get a combined accuracy of 87.4%. We will discuss the results of these experiements in greater detail in the Experiments and Results chapter.

1.6 Structure

This paper is divided into 7 chapters which provide a detailed description of our project. In this section, we will provide a concise overview of each chapter, describing the key topics they cover.

In Chapter 2 we describe the motivation for our project. We present a comprehensive examination of the current state of pathology and provide insights into the future prospects of the field. We define the most common problems that occur when working with medical images and give a potential solution to those problems.

In Chapter 3 we explore the concept of applying transfer learning to address the challenges in medical image classification. A significant portion of this chapter is dedicated to the exploration of the datasets used in our project. We describe the unique characteristics and challenges associated with each dataset, emphasizing how these factors significantly impact the training of convolutional neural networks (CNN's).

In chapter 4 we describe the key components of our classifiers and clustering algorithms.

Chapter 5 offers a detailed description of our implementation. We explain the essential algorithms employed for dataset preparation, the image processing techniques necessary, and how they facilitate the unification of multiple datasets into a more general, well developed dataset. We also define the training routine for the classification tasks.

Chapter 6 presents the results of our research. We explain how to evaluate the constructed model and show how well it predicts the presence and type of cancer in histopathological images of lymph node sections. We evaluate the performance of the model on both datasets, and explain the conditions necessary to replicate the same results.

In chapter 7 we present the final conclusions of our project.

2 MOTIVATION

The primary motivation of our project is to leverage the power of deep learning and advanced image analysis techniques to improve the accuracy, efficiency, and consistency of classifying medical images consisting of lymph nodes.

2.1 Current State of Pathology

In the context of pathology, deep learning has the potential to revolutionize the field by automating and augmenting various tasks. It can assist pathologists in accurate and efficient diagnosis, provide prognostic information, aid in tumor grading, and even guide personalized treatment decisions. Moreover, deep learning models can analyze large amounts of pathological data, uncover hidden patterns, and identify novel biomarkers that may have clinical significance [7].

2.2 Lymph Node Analysis

The morphological characteristics of lymph nodes in lymphoma and metastatic cancer from other organs can exhibit notable differences. Deep learning algorithms can be trained to differentiate between cancerous and non-cancerous lymph nodes and further distinguish between different types of cancer.

2.3 Proposed Challenge

Despite the great potential of deep learning in pathology, there are several challenges that need to be addressed. One major problem is the availability of large, high-quality, annotated datasets for training deep learning models [13, 11, 15, 9]. We aim to resolve this problem through the idea of unifying datasets. By combining multiple datasets and normalizing the images using image processing algorithms, such as Reinhard normalization, Macenko normalization, Vahadane Normalization [16, 14, 17], the images from different datasets can be brough to share a consistent representation. This normalization process helps to reduce the variability and discrepancies that may exist between the datasets. Our results highlight the potential of this method, enabling the training of deep learning models with improved generalization capabilities in pathology. Although the datasets used in our study contain images of lymph node sections, it is important to note that they represent different types of cancer.

In the first dataset, the images depict lymph nodes with breast cancer metastasis, which is the spread of cancer from a primary site to the lymph nodes. In contrast, the second dataset represents images of lymph nodes affected by primary lymph node cancer itself (1).

2.4 Other Challenges

Another challenge is the interpretability and explainability of deep learning models, as their decisions are often considered as "black boxes." Integration of deep learning into the existing pathology workflow and addressing regulatory and ethical considerations are additional issues that need attention.

3 RELATED WORK

The qualitative visual analysis of microscopic images has several limitations, including the absence of standardization, potential diagnostic errors, and the great amount of time associated with manually evaluating numerous cells across multiple slides during a pathologist's typical workload, according to [18] This has spiked an increase in research involving computational methods that can aid in the analysis of microscopic images in pathology.

3.1 H&E staining methodology

Historically, the histopathological examination of tissue has been most used method for cancer diagnosis worldwide. The staining of tissue samples with chemical dies, most notably hematoxylin and eosin (H&E) stand at the core of this procedure by enhancing cellular features for pathologists to interpret visually. The growing adoption of digital pathology has heightened the need for accurate stain quantification and standardization. Digital pathology involves scanning glass pathology slides using whole slide imaging systems to produce digital images. This technology has the potential of enhancing workflow and quality within pathology services with the help of artificial intelligence, However, despite being largely unchanged for over a century, the staining process exhibits well-known methodological variations which are commonly observed by examining samples from different laboratories. Ensuring consistency in the staining process is difficult to achieve which in turn makes is all the more difficult to use those images in the context of AI-based applications. One strategy for minimizing image variability involves reducing staining inconsistencies at the laboratory level through stringent quality control (QC) measures. Histopathology labs maintain strict protocols and regularly replenish reagents to limit variations, and most use automated stainers for enhanced consistency. Nonetheless, routine QC methods remain largely unchanged, relying on subjective assessments in both internal and external evaluations. Subjective human assessment is prone to bias and hinges on the staining of control tissues, which can themselves vary biologically between sections. Furthermore, variables prior to staining (e.g., fixation or section thickness) can also affect results. These limitations highlight the drawbacks of relying exclusively on tissue-based QC for assessing stain quality. Moreover, the processes involved can introduce challenges with color fidelity—an essential factor for defining thresholds in object and pattern detection, which in turn directly impacts the performance of classification algorithms [8, 12]. In this paper, we will focus on using various computational methods for stain normalization aimed at minimizing the bias introduced by differing methodologies across multiple institutions.

3.2 Patch Camelyon Dataset

This dataset is available at [3]. It was used in the Patch Camelyon Challenge in 2016 [1]. It contains H&E and IHC scans of lymph node sections (breast cancer) where each image of size 96x96px is annoted with a binary label indicating presence of metastasic tissue. A positive label means that the center 32x32px region of a patch contains at least one pixel of tumor tissue.

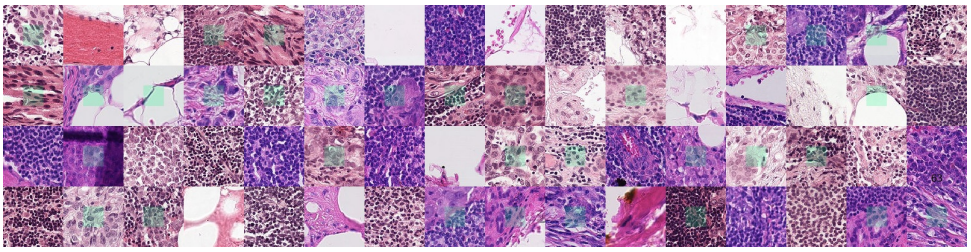

Figure 1: Example images from PCam. Green boxes indicate tumor tissue in the center region, which dictates a positive label. Source: [3]

The whole dataset is divided into a training set of 262.144 samples and a validation and test set of 32.768 samples each. To obtain images, the slides are undersampled at 10x to increase field of view. To prevent selecting background patches, slides are converted to HSV, blurred, and filtered out if the average value on each channel does not meet a certain threshold. It is worth mentioning that, for the sake of consistency throughout all datasets, only H&E images were included, therefore all immunohistochemistry (IHC) samples have been removed. This was done through a patch selection algorithm explained in detail in Chapter 5.

This dataset poses several challenges for computer-based analysis, including variations in tissue appearance, presence of artifacts and background, and the need to accurately identify metastatic cancer cells within the lymph node tissue.

3.3 DLBCL-Morphology Dataset

The dataset, available at [2] contains TMA slides on lymph nodes affected by primary lymph node cancer and matching ROI's for metastasic tissue presence in the form of geometric coordinates. It contains 42 digital high-magnification scans of tissue microarrays (TMAs) from cases at Stanford Hospital.

Images of the same 96x96px size are obtained following the methodology described by the providers of the PCam dataset in order to maximize uniformity between the two datasets. The geometric points are used to define areas where positive patches can be extracted. Consequently, the remaining areas are used to extract negative samples from each slide.

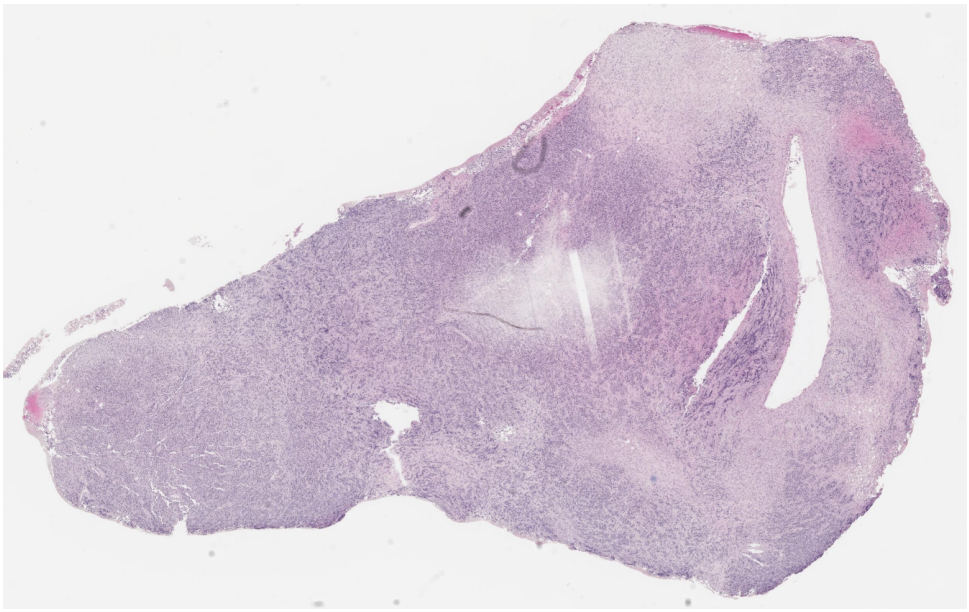

Figure 2: Tissue slide image of lymph node. Source: [2]

patient_id	tma_id	stain	xs	ys	xe	ye
17658	TA255	HE	44533	20283	47118	22924
17673	TA255	HE	4001	7152	6594	9777
17660	TA255	HE	31177	15345	33787	17954
17642	TA255	HE	17158	33174	19727	36014
17665	TA255	HE	13408	11292	15744	14034
17642	TA255	HE	12677	33400	15411	36070
17665	TA255	HE	18015	11191	20532	13784
17674	TA255	HE	3991	2985	6534	5619
17662	TA255	HE	13227	15654	15569	18205
17672	TA255	HE	14933	8284	15524	8860
17668	TA255	HE	44271	11728	46952	14425
17668	TA255	HE	39816	11337	42633	13314
17649	TA255	HE	40331	24262	42966	27043
17652	TA255	HE	13482	23981	15955	26583
17666	TA255	HE	26977	10946	29554	13625

Figure 3: ROI coordinates

3.4 Cancer Detection from Lymph Nodes

As reported in [21], fine-tuning encoders on histopathology data yielded superior performance, reflected in higher AUC values than architectures relying solely on transfer learning. This outcome supports the idea that domain-specific encoders are more effective for extracting histopathology features than those pre-trained on ImageNet. The authors advocate using publicly available weights for histopathology tasks and fine-tuning state-of-the-art CNNs. They achieved an 84.5 AUC by fine-tuning a ResNet-18 model on the PatchCamelyon dataset.

We extend this transfer learning approach to our two datasets as a preliminary step in our multiclass classification framework, obtaining a slightly higher score than the reported result. Nonetheless, our performance remains significantly below the 98% AUC attained by the Camelyon Grand Challenge 2016 winners [7], who employed a GoogLeNet [6] architecture. Moreover, neither of these methods addresses model generalization, suggesting that matching this level of accuracy on a similar dataset containing lymph node images with various cancer types is unlikely.

In light of this, a model trained for multiclass classification, which leverages information across all classes, may deliver better results and generalize more effectively in real-world applications compared to separate binary classifiers.

4 PROPOSED METHOD

In this chapter we present the technologies and architectural components used in our project.

4.1 PyTorch Lightning

PyTorch Lightning is a streamlined wrapper for PyTorch that simplifies code organization. It allows developers to focus on model flexibility while maintaining performance at scale. In recent years, it has gained significant popularity and is quickly becoming a standard in the research community. Its advanced features include automated checkpointing, logging, distributed training, and modularization.

Throughout development, we also used well-established machine learning libraries such as NumPy, scikit-learn, OpenCV, Matplotlib, and OpenSlide (for handling whole slide images).

4.2 Full Slide Images

Digital pathology often involves working with full slide images (WSI or TMA) that can exceed several gigabytes in size, with resolutions around 100k $\times$ 100k pixels. Feeding such large images directly into a neural network is infeasible due to memory limitations. A common strategy is to break down these slides into smaller patches that can be processed and labeled individually.

Following the methodology recommended by the authors of PatchCamelyon, we use the geometric coordinates available from the DLBCL-Morphology dataset (Figure [3]) to extract 96 $\times$ 96 px patches labeled as positive or negative. We also implement a threshold-based patch selection algorithm to discard patches with excessive background or significantly poorer quality than the rest. Such extreme outliers, if included, can disrupt the training process by causing the model's weights to adjust incorrectly. While this step substantially improves the dataset's overall quality, it reduces its size to around 10,000 samples (50/50 split). In the following section, we provide more details about the algorithms used to prepare this dataset.

4.3 Residual Neural Network

ResNet [10] is a well-known Convolutional Neural Network (CNN) architecture, primarily recognized for its use of skip connections. These connections allow gradients to bypass certain

layers, thereby alleviating the vanishing gradient problem, which can hinder the training of very deep networks. ResNet is composed of multiple blocks, each containing several layers. Its robustness and performance have been extensively validated in image recognition challenges, cementing its reputation as a reliable architecture for a range of image-based tasks.

In this work, we employed both ResNet-18 and ResNet-50 because they are relatively compact and simpler to train, without sacrificing too much accuracy.

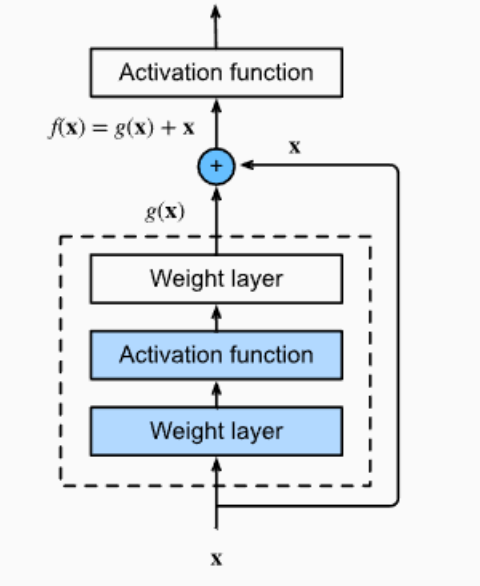

Figure 4: RestNet block[1]

4.4 Clustering Method

To determine whether our two datasets can be merged, we use the K-means algorithm provided by scikit-learn and evaluate the results with the metrics listed below. K-means clusters similar samples based solely on their features, without relying on any predefined labels.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Accuracy measures the number of correct predictions out of the total number of instances. In this case, a high accuracy value would represent a high degree of separation between the clusters, which is far from ideal. Our aim is to get an accuracy score as close to 50% as possible, which would validate the the two datasets have been successfully combined.

$$Silhouette(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2}$$

[1]Source: `https://d2l.ai/chapter_convolutional-modern/resnet.html`

If the silhouette score is close to 0, it suggests that the clusters are overlapping or very close to each other. The silhouette score measures the difference between the average distance within clusters (cohesion) and the average distance to the next nearest cluster (separation).

$$ARI = \frac{\text{Index} - \text{ExpectedIndex}}{\text{MaxIndex} - \text{ExpectedIndex}} \tag{3}$$

The Adjusted Rand Index is a measure that quantifies the similarity between two data clusterings. It is used to compare the output of a clustering algorithm to a ground truth. The ARI has a range of [-1, 1]. An ARI of 1 means that the clusterings are identical, while a 0 indicates the result is no better than random, and a negative value means the result is even worse than random. In this case, we aim for a value as close to 0 as possible.

4.5 Classification Methods

For our binary classification task, we build upon ResNet-18, modifying its final layer to output a single neuron with a Sigmoid activation (Equation 7) to map predictions to 0 or 1. We use pre-trained ImageNet weights to leverage general patterns already learned, then fine-tune the model separately on the PatchCamelyon and DLBCL-Morphology datasets. Our loss function is BCELoss (Equation 4), which penalizes the model for incorrect predictions:

$$\text{BCELoss}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \tag{4}$$

This loss function is commonly used in binary classification problems. It penalizes the model for wrong predictions by comparing the predicted probability ($\hat{y}$) with the ground truth (y). N is the total number of samples.

$$\text{CrossEntropyLoss}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{5}$$

For the multiclass experiment, the last layer is replaced with 3 output neurons referring to all possible classes: negative, positive for breast metastasis and positive for lymphoma. We change the loss function to CrossEntropyLoss which automatically adds a Softmax activation over the last layer ensuring that the sum of the class probabilities predicted by the model is equal to 1 (100%). The predicted output is the label with the greatest probability.

5 IMPLEMENTATION DETAILS

This chapter explains our implementation in detail, highlighting the challenges encountered while preparing the combined dataset. The results of the following experiments are thoroughly discussed in the next section.

5.1 Image Processing

The image processing pipeline comprises several stages. One key component is color-based normalization. We employ two normalization techniques to minimize differences in stain distribution between our two datasets. Following the respective authors' guidelines, we implemented and evaluated these approaches on our datasets.

5.1.1 Normalization Techniques

- **Macenko Normalization:** Proposed by Macenko et al. [14], this method targets histopathology images, aligning the stain intensities between a reference and a target image by estimating their respective stain matrices. These matrices capture the color information for different stains, such as H&E. After deriving these matrices, each image is transformed based on the reference matrix, thereby reducing color variations produced by different staining protocols.
- **Vahadane Normalization:** Vahandane et al. [17] introduced a variation of Macenko Normalization, adding steps to refine color normalization for histopathology images. It lessens the influence of background regions by applying an adaptive thresholding method that identifies and excludes non-tissue areas. This helps achieve more accurate color normalization in H&E-stained images.
- **Reinhard Normalization:** Reinhard et al. [16] compute the mean and standard deviation of each color channel in both the source and target images. The source image channels are first normalized with respect to their own mean and standard deviation, followed by applying the target image's mean and standard deviation. This process synchronizes the color distribution of the source images to the target domain.

Each approach strives to reduce color inconsistencies introduced during the staining process and image acquisition in histopathology. Figures (5, 6) show how these methods enhance the quality of our images.

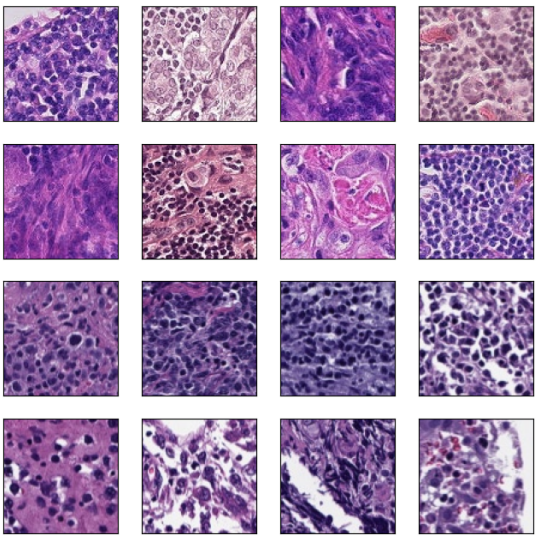

Figure 5: Samples from both datasets before preprocessing

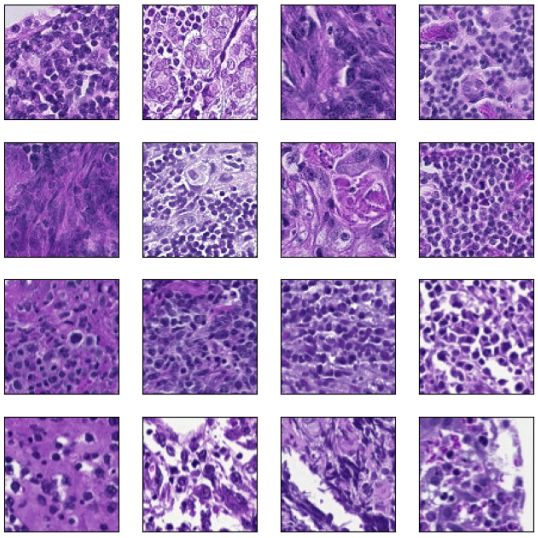

Figure 6: Samples from both datasets after preprocessing

5.1.2 Patch Extraction

Additionally, our pipeline involves removing immunohistochemistry (IHC) images from Patch Camelyon and extracting patches from DLBCL-Morphology. These steps improve the overall dataset quality. Below is a brief overview of the relevant algorithms.

Since we aim to unify the datasets, several modifications to Patch Camelyon are needed. Images come in zipped HDF5 format, and we exclude IHC images because our other dataset solely contains H&E stains. We switch each image to HSV color space, smooth it using Gaussian Blur, then compute the mean values of hue, saturation, and brightness to classify positive patches. The threshold values that gave the best results are within these ranges SATURATION_THRESHOLD = [0.3, 0.4], BRIGHTNESS_THRESHOLD = [0.5, 0.8], HUE_THRESHOLD = [0.5, 0.6]. We discover these thresholds by a grid search over a subset of 1000 samples, which speeds up the process. Patches meeting these criteria are saved in JPEG format.

For DLBCL-Morphology, we read the CSV annotations file (3) and focus on H&E-stained slides. We keep a dictionary of positive regions for each slide and generate 96x96px patches from these regions. Similar thresholding is applied to eliminate sections with substantial background or compromised quality 8. The values that give best results are SATURATION_THRESHOLD $>$ 0.07 BRIGHTNESS_THRESHOLD $<$ 0.8. Any patch that meets these thresholds is retained and saved in JPEG format.

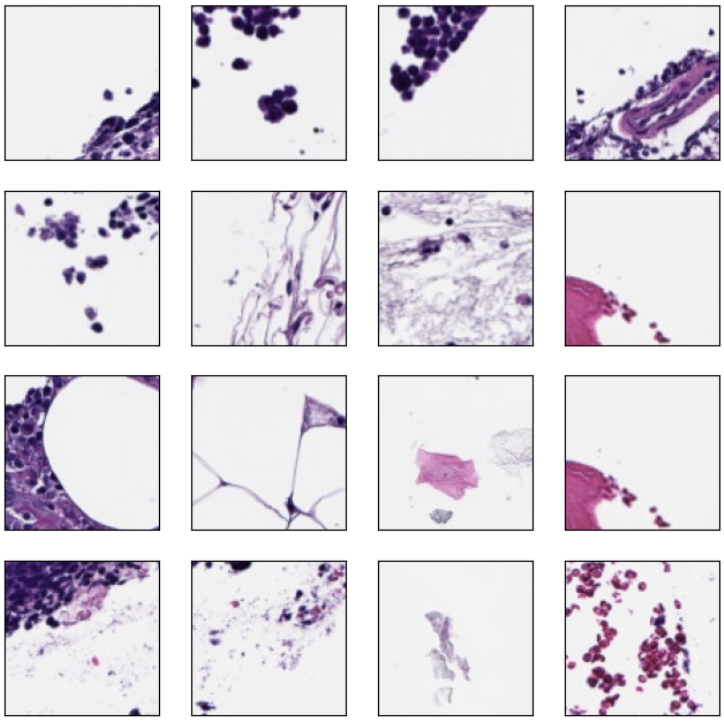

Figure 7: Examples of discarded images from DLBCL-Morphology

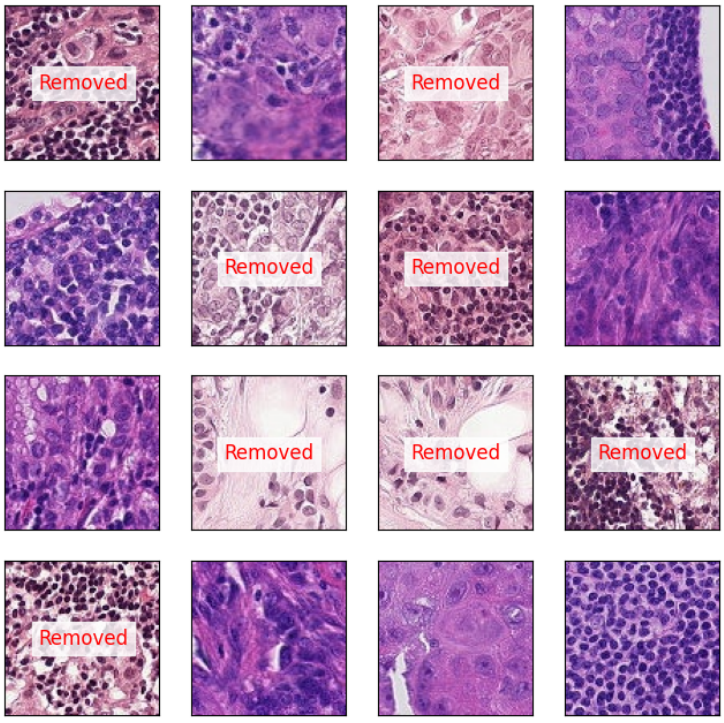

Figure 8: Examples of discarded IHC images from Patch Camelyon

Negative patches are taken from portions of the TMA slide not marked as containing cancer. To make this step more efficient, we skip areas that cannot fit at least one patch. We save the labels in a JSON file to ensure consistency across both datasets.

5.2 ResNet-18 Architecture

Here we describe how ResNet-18 is adapted for feature extraction during clustering. ResNet-18 contains several layers, each contributing to its capacity for representation learning:

- **Input Layer:** Receives RGB images of size 224x224px, but the PyTorch implementation can accept alternative image sizes (like our 96x96px patches).
- **Convolutional Layers:** Extract features from the input by modifying size and channel depth.
- **Pooling Layers (Max / Average):** Decrease spatial dimensions by taking the maximum or average value inside the pooling window, retaining critical features.
- **Residual Blocks:** ResNet-18 has 4 blocks [4], each containing shortcut connections that facilitate gradient flow, mitigating vanishing gradients. The Rectified Linear Unit (6) acts as the activation function, and Batch Normalization follows each convolution and fully connected layer to expedite training.

$$ReLU(x) = \max(0, x) \tag{6}$$

- **Output Layer:** Typically ends with a Softmax activation for ImageNet (1000 classes). Since we only require feature embeddings, we discard this final layer. Thus, the last convolutional layer's output (`batch_size, 512`) forms our feature vectors for clustering.

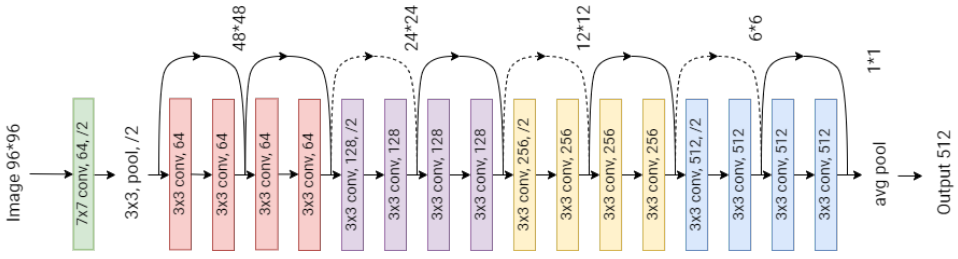

Figure 9: Feature extractor based on ResNet-18 architecture. Image adapted from: [10]

5.3 Clustering Experiment

We now detail our clustering experiment. Only negative samples from each dataset are used to ensure fairness and reduce possible dataset-related biases. The following steps outline the procedure:

- **Data Loading and Preparation:** Gather negative samples from both DLBCL-Morphology and Patch Camelyon. A label dictionary tracks each image's origin.
- **Shuffling:** Randomly shuffle the data to remove inherent ordering effects.
- **Stain Normalization:** Apply the above-described normalization methods. Macenko normalization yields the best outcomes.
- **ResNet-18 Feature Extraction:** Process images through ResNet-18 (minus its last layer) to obtain feature vectors.
- **PCA:** Use Principal Component Analysis to reduce the dimensionality of extracted features, leveraging the `sklearn` implementation.
- **KMeans Clustering:** Use the `KMeans` class in `sklearn`. The algorithm uses the `k-means++` strategy for centroid initialization. The number of initializations depends on input size.
- **Label Assignment:** Assign each point to the nearest cluster center. Use the Hungarian algorithm to match these clusters with the original dataset labels, minimizing total assignment cost.
- **Evaluation:** Measure Accuracy, Silhouette Score, and ARI. Finally, we visualize scatter and density plots to interpret the results.

5.4 Classification Architectures

Our main model class extends the `LightningModule` from PyTorch Lightning. ResNet-18 and ResNet-50 with pretrained ImageNet weights serve as the backbone.

5.4.1 Binary Classification

For binary classification, we replace the final layer with a single neuron and a Sigmoid activation. This predicts whether an image contains cancerous tissue. We run this experiment on the entire Patch Camelyon dataset to benchmark performance:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{7}$$

5.4.2 Multiclass Classification

For the combined dataset, we assign three classes: 0 (negative), 1 (positive for breast metastasis), and 2 (positive for lymphoma). The final layer now has 3 output neurons, alongside a Softmax activation to compute probabilities:

$$\text{softmax}(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{8}$$

We present the results for both tasks in the next section.

5.5 Data Loader

Training efficiency is enhanced via PyTorch's `Dataset` and `DataLoader` classes. Each `DataLoader` manages batching, shuffling, and parallel loading. We have distinct DataLoaders for training, validation, and testing. Our custom class also accepts transformation functions (random flips, rotations, normalization) to augment data, converting images to PIL format before applying transformations and then back to tensors. We normalize each subset using the training dataset's mean and standard deviation, helping keep gradients in a stable range.

5.6 Adam Optimizer

We optimize the network weights using the Adam optimizer, which is known for its adaptive learning rate and broad applicability. Our chosen hyperparameters are a learning rate of 1×10^{-4} and a weight decay of 1×10^{-3}. Weight decay adds a penalty to minimize overfitting.

5.7 Learning Rate Scheduler

We monitor validation loss to trigger a decrease in the learning rate by a factor of 0.1 if performance remains unchanged for five epochs (patience $= 5$). This approach fine-tunes the model more effectively as training progresses.

5.8 Callbacks

We define two callbacks based on validation accuracy and loss, saving the top-performing model state in a `.ckpt` file. This allows us to resume training without losing progress and is particularly important for longer training sessions. We also incorporate early stopping (patience $= 20$) based on the validation loss.

5.9 Training

For both classification tasks, we reserve 80% of the data for training. We augment that data as described earlier to increase dataset diversity, using a batch size of 64. We train for up to 100 epochs, starting with a 1×10^{-4} learning rate, which is reduced every 10 epochs if progress stalls. Early stopping can interrupt training prematurely. The DataLoader leverages GPU parallelization for faster processing.

5.10 Logging

We rely on PyTorch Lightning's built-in logging features for tracking model performance. These metrics can be visualized in TensorBoard.

5.11 Evaluation

In binary classification, we compute Accuracy (1) and the ROC curve, measuring AUC-ROC to gauge performance across varying thresholds. The ROC curve compares the true positive rate (TPR) and the false positive rate (FPR).

For multiclass classification, we also focus on Precision and Recall for each individual class. Achieving a high Precision for both positive classes (breast cancer metastasis and lymphoma) is crucial, although Recall is paramount to minimize missed cancer diagnoses.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{9}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{10}$$

$$\text{True Positive Rate} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{False Positive Rate} = \frac{\text{FP}}{\text{N}} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{12}$$

6 EXPERIMENTS AND RESULTS

In this chapter we present the outcomes of our experiments, along with a discussion of encountered problems and how we resolved them.

6.1 Clustering Results

K-means runs several centroid initializations, selecting the one with the lowest inertia. The number of initializations depends on the number of input samples. Since K-means is sensitive to outliers and limited to linear boundaries, we linearize images and apply PCA to reduce the input to two dimensions.

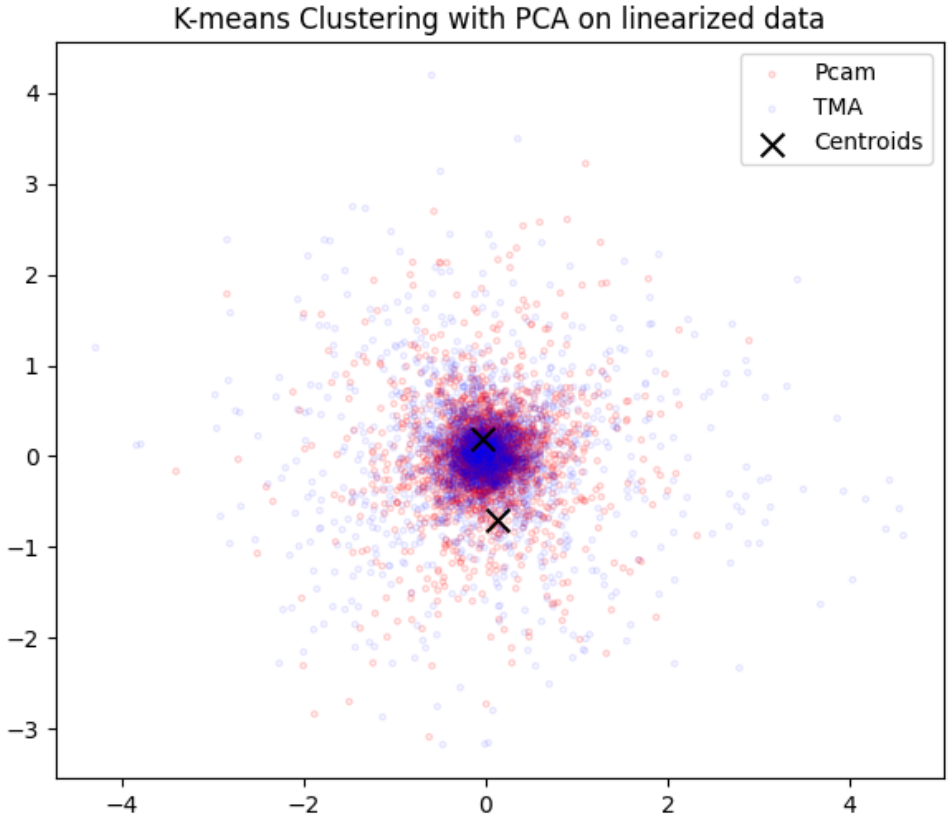

Figure 10: K-means Clustering on linearized images

We compute Accuracy, Silhouette Score, and ARI. For Silhouette Score, a value near zero indicates overlapping clusters, and for ARI, values close to zero show random-level similarity. An Accuracy near 50% also suggests successful data blending. We achieve:

$$\text{ARI: } 0.009, \quad \text{Accuracy: } 0.512.$$

Initially, these values imply our pipeline may have integrated the data effectively. However, because deep learning rarely uses raw features, we repeat the experiment using ResNet-18 embeddings. This yields:

$$\text{ARI: } 0.198, \quad \text{Accuracy: } 0.723.$$

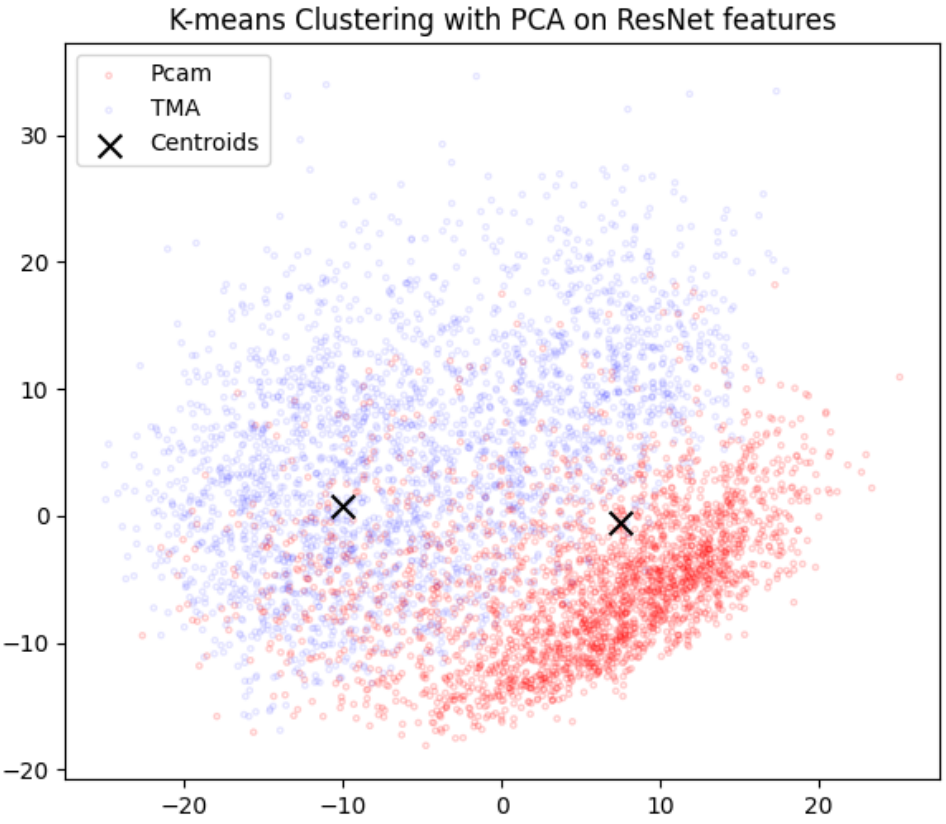

Figure 11: K-means Clustering with PCA on ResNet-18 features

Upon inspection, many correctly classified DLBCL-Morphology samples were from TMA edges, containing large background areas. While excluding background might help, certain tissue surroundings can hold useful context for classification. Nonetheless, extreme outliers with excessive artifacts or background should be removed as they strongly shift the cluster centers.

Differences in how the datasets were obtained can also affect feature extraction. Applying the stain normalization techniques notably improves clustering, producing a nearly 20% drop in Accuracy and confirming that color discrepancies and poor-quality images greatly impacted the extraction process:

$$\text{ARI: } 0.004, \quad \text{Accuracy: } 0.534.$$

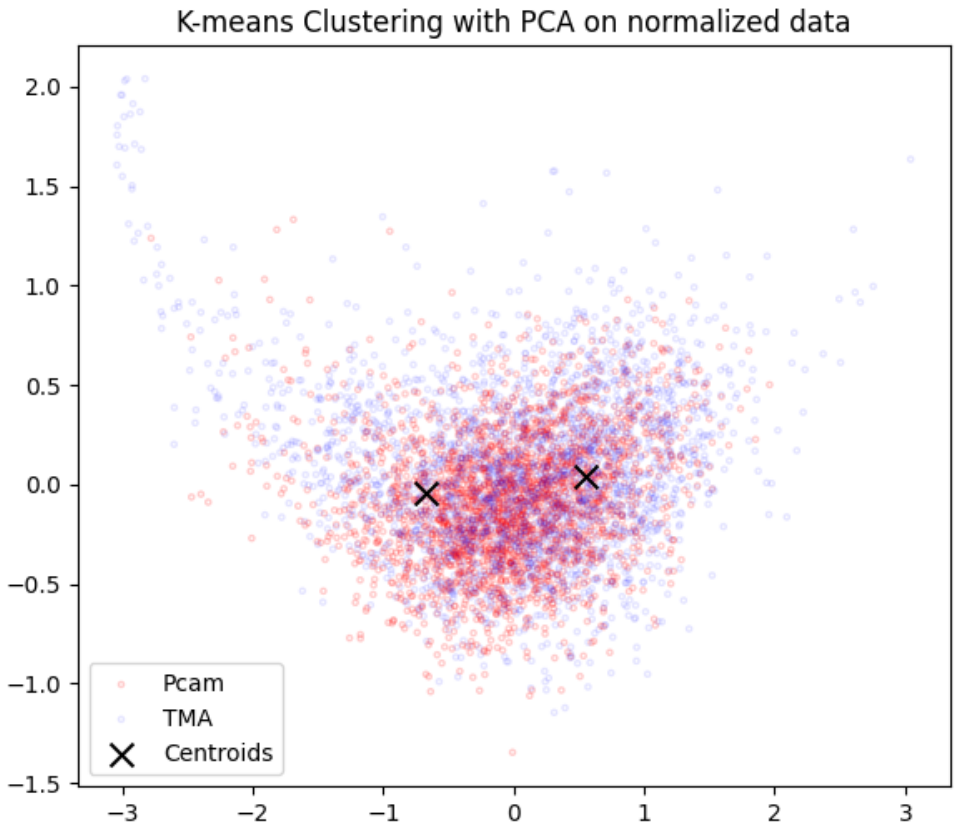

Figure 12: K-means Clustering with PCA on ResNet-18 features extracted from processed images

Future enhancements include using domain adaptation strategies. For instance, adding gradient reversal layers and a domain-specific classifier could enforce domain-invariant features, making images from different sources indistinguishable.

6.2 Classification Results

The clustering outcomes suggest that our feature extractor cannot reliably differentiate the negative samples from the two datasets, implying minimal dataset bias and a suitable environment for developing a combined dataset. As a result, we construct a dataset with three classes (negative, breast cancer, lymphoma) totaling 20,000 images. We compare two standalone binary classifiers for each dataset and our multiclass approach. Table 1 and Table 2 show the results.

Given our hardware constraints, we tested smaller models. According to [21], more extensive networks may reach similar performance with enough time. On Patch Camelyon, ResNet-18 and ResNet-50 achieve 86.5% and 82.5% accuracy, respectively, with generally strong precision, recall, and F1 scores for the positive class. Meanwhile, ResNet-18 excels on DLBCL (95.5% accuracy), though its smaller size casts doubt on representativeness for real-world performance.

Data	Model	Positive P/R/F1	Negative P/R/F1	Accuracy
Camelyon	ResNet-18	94.9/77.2/85.1	80.8/95.8/87.7	86.5
Camelyon	ResNet-50	91.7/71.6/80.4	76.7/93.5/84.3	82.5
DLBCL	ResNet-18	92.5/99.0/95.7	98.9/92.0/95.3	95.5
DLBCL	ResNet-50	78.1/99.4/87.5	99.2/72.2/83.6	85.8

Table 1: Results for Binary Classification

Data	Model	Breast P/R/F1	Lymphoma P/R/F1	Negative P/R/F1	Accuracy
Mixed	ResNet-18	96.5/89.4/92.8	74.1/97.4/84.2	92.5/81.4/86.6	87.4
Mixed	ResNet-50	98.0/86.2/91.7	60.4/97.4/74.6	89.4/67.4/76.9	79.6

Table 2: Results for Multiclass classification

For the multiclass dataset, ResNet-18 displayed higher accuracy on detecting lymphoma. Although the binary classifier for lymphoma surpasses the multiclass model in recall, the multiclass approach excels in breast metastasis detection, indicating that it can learn both transferable and domain-specific features. This approach also streamlines workflows by eliminating the need for multiple separate binary classifiers.

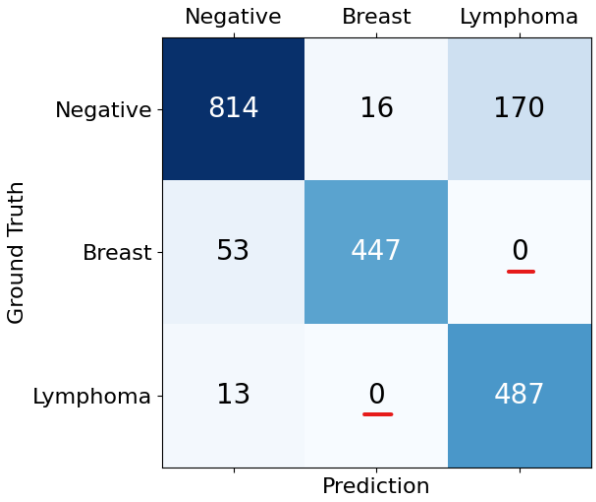

Figure 13: Multiclass Confusion Matrix

The confusion matrix offers additional evidence of the multiclass model's effectiveness. Notably, there are no instances of lymphoma being misclassified as breast cancer or vice versa, indicating that the model successfully captures the distinct features of each cancer type.

6.3 Feature Visualization

Grad-CAM [19] helps interpret the model's predictions by highlighting which regions of an image it deems most significant. It applies the gradients of the final convolutional layer to

create a heatmap, indicating higher relevance in red and lower in blue. We use this to gauge how our multiclass model focuses on different features.

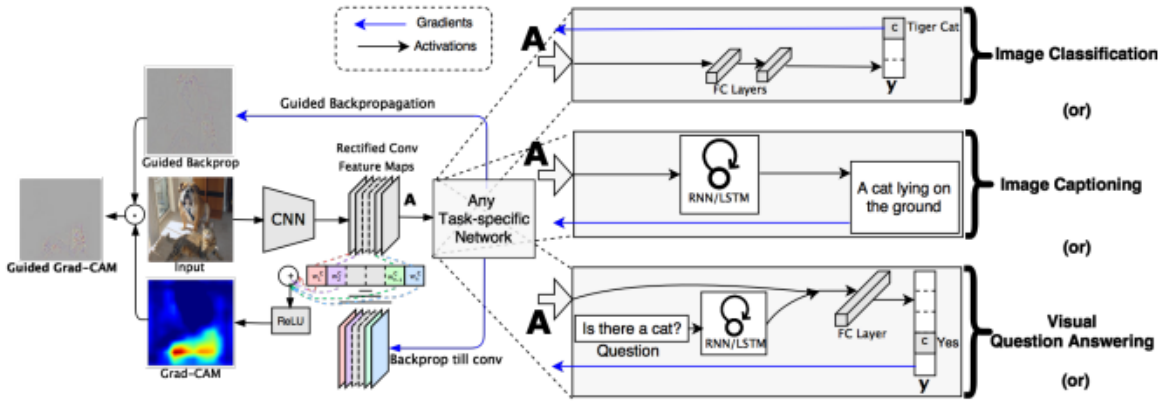

Figure 14: Grad-CAM overview. Figure taken from [19]

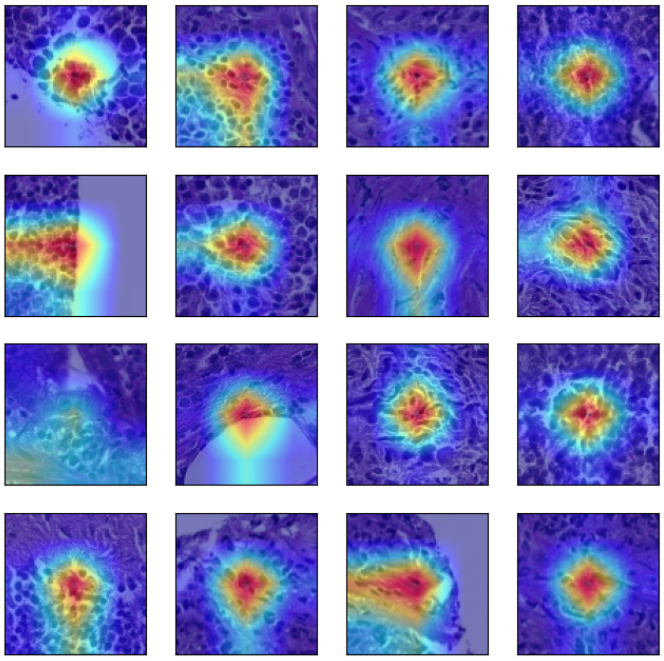

Figure 15: Grad-CAM feature maps for FP and FN classified images

As visible from the GradCam activation mappings, the network focuses primarily on the central region for PatchCamelyon, which contains a 32×32 px tumor area. In contrast, the DLBCL-Morphology images do not follow this pattern, potentially explaining why the binary classifier outperforms the multiclass model in this domain.

6.4 Future Work

Firstly, we intend to investigate the GradCam results in greater detail to address the model's tendency to concentrate on a specific patch location. By understanding how these activation maps differ between datasets, we aim to correct the sources of underperformance and improve our models' capacity to localize critical features. In doing so, we can refine our network architectures and training procedures, driving progress toward more accurate cancer classification and better patient outcomes.

Another goal is to benchmark several state-of-the-art deep learning architectures side by side on the combined dataset. Comparing larger and more advanced models in a controlled manner will contribute to the robustness and reliability of the results, offering a clearer perspective on which architectures excel under various conditions.

BIBLIOGRAPHY

[1] Camelyon grand challenge 2016. https://camelyon16.grand-challenge.org/Results.

[2] Dlbcl-morphology. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=119702520.

[3] Patchcamelyon (pcam). https://github.com/basveeling/pcam.

[4] Ramon Pires Flavia Vasques Bittencourt Sandra Avila Afonso Menegola, Michel Fornaciali and Eduardo Valle. Knowledge transfer for melanoma screening with deep learning. *IEEE*, 2017.

[5] Swarat Chaudhuri Chris Jermaine Arkabandhu Chowdhury, Mingchao Jiang. Few-shot image classification: Just use a library of pre-trained feature extractors and a simple classifier. *ICCV*, 2021.

[6] Y. Jia P. Sermanet S. Reed D. Anguelov D. Erhan V. Vanhoucke C. Szegedy, W. Liu and A. Rabinovich. Going deeper with convolutions. *CVPR*, 2015.

[7] Rishab Gargeya Humayun Irshad Andrew H. Beck Dayong Wang, Aditya Khosla. Deep learning for identifying metastatic breast cancer. 2016.

[8] Bejnordi BE et al. Stain specific standardization of whole-slide histopathological images, 2016.

[9] Greenspan et al. Deep learning and convolutional neural networks for medical imaging and clinical informatics. *JAMA*, 2016.

[10] He et al. Deep residual learning for image recognition. *IEEE*, 2016.

[11] Hirsch et al. The need for large-scale annotation of biomedical data sets. *Journal of Pathology Informatics*, 2018.

[12] Howard FM et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Google Scholar*, 2021.

[13] Litjens et al. Deep learning in medical image analysis: Challenges and applications. *PubMed*, 2017.

[14] Macenko et al. A method for normalizing histology slides for quantitative analysis. *IEEE International Symposium on Biomedical Imaging*, 2009.

[15] Madabhushi et al. Challenges and opportunities in machine learning and deep learning in pathology. *JAMA*, 2019.

[16] Reinhard et al. Color transfer between images. *IEEE Computer Graphics and Applications*, 2001.

[17] Vahadane et al. Structure-preserving color normalization and sparse stain separation for histological images. *PubMed*, 2016.

[18] P. A. Carney B. M. Geller T. Onega A. N. Tosteson H. D. Nelson M. S. Pepe K. H. Allison S. J. Schnitt et al. J. G. Elmore, G. M. Longton. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA*, 2015.

[19] Abhishek Das Ramakrishna Vedantam Devi Parikh Dhruv Batra Ramprasaath R. Selvaraju, Michael Cogswell. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IJCV*, 2019.

[20] Fiona Ryan Zachary Beaver Jan Freyberg Jonathan Deaton Aaron Loh Alan Karthikesalingam Simon Kornblith Ting Chen Vivek Natarajan Mohammad Norouzi Shekoofeh Azizi, Basil Mustafa. Big self-supervised models advance medical image classification. *ICCV*, 2021.

[21] Sana Syed Donald E. Brown Yash Sharma, Lubaina Ehsan. Histotransfer: Understanding transfer learning for histopathology. *IEEE*, 2021.