Abstract

In this paper, we explore the application of machine learning and deep learning models for the classification of hyperplastic polyps (HPs) and sessile serrated adenomas (SSAs) in the colon. Accurately distinguishing between these polyp types holds immense clinical significance as it enables timely diagnosis and facilitates effective treatment strategies. Our study utilizes a comprehensive dataset comprising over 3000 fixed-size hematoxylin and eosin (H&E)-stained, formalin-fixed, and paraffin-embedded (FFPE) images. To achieve our objectives, we employ a convolutional neural network (CNN) and a random forest model. The random forest model shows promising performance, achieving a precision of 0.84, recall of 0.83, and F1-score of 0.83. These metrics indicate a good balance in accurately identifying HPs and SSAs while minimizing false positives and false negatives. On the other hand, the CNN demonstrates even higher performance, with a precision of 0.81, recall of 0.92, and F1-score of 0.86. The CNN's superior recall score suggests its effective ability to detect both HPs and SSAs, ensuring high sensitivity in correctly identifying these polyp types. Moreover, the precision score indicates a strong capability to reduce false positives, while the F1-score reflects a harmonious combination of precision and recall. This paper underscores the significance of machine learning and deep learning models as promising tools for the classification of HPs and SSAs in colon images. The successful application of these models opens up avenues for enhancing diagnostic accuracy and enabling timely intervention, leading to improved patient outcomes. Further advancements in this field can contribute to the development of automated systems that assist healthcare professionals in effectively identifying and treating these specific polyp types.

Keywords: Convolutional Neural Network, Random Forest, Image Classification, Binary Classification, Colorectal Polyps, Histopathology

Links to the Dataset & Extracted Features:

Original Dataset: https://bmirds.github.io/MHIST/

Introduction

In 2020, almost 2 million people were diagnosed with colorectal cancer and almost 1 million people died. The International Agency for Research on Cancer (IARC) estimates an increase of 56% from 2020 to 2040, which will lead colorectal cancer cases to more than 3 million per year. The IARC predicts an even bigger increase in death rate, supposedly of 69% [1]. To prevent this, it is important to detect and prevent first the subject area and detect the region affected by colorectal polyps, the field of study for examining tissues and detecting abnormal cells is called histopathology. The accurate and prompt categorization of polyps is key for directing clinical treatment and monitoring approaches, especially discerning between hyperplastic polyps (HPs) and sessile serrated adenomas (SSAs) as one is predominantly benign, whereas SSA can potentially develop into riskier forms or cancer. With the increasing complexity of histopathological images and the growing need for accurate and timely diagnosis, traditional human-driven approaches face significant challenges.

To address these challenges, deep learning, a subset of artificial intelligence and machine learning, offers powerful tools to help medical experts. By leveraging neural networks and large-scale datasets, deep learning algorithms can automatically learn to recognize intricate patterns and features in histopathology images, which may not be apparent to the human eye. This ability to identify and analyze complex patterns within the data can lead to improved diagnostic accuracy, early detection of diseases, and personalized treatment options.

As the topic is crucial, our project aims to help speed up the process of implementing AI tools for colon cancer prevention and diagnosis. The dataset we used comprises more than 3000 hematoxylin and eosin (H&E)-stained Formalin-Fixed Paraffin-Embedded (FFPE) fixed-size images (224 by 224 pixels) of colorectal polyps from the Department of Pathology and Laboratory Medicine at Dartmouth-Hitchcock Medical Center (DHMC).

The main drive of this project is to respond to the following **research question**:

"Can machine learning and deep learning enhance the classification accuracy between hyperplastic polyps (HP) and sessile serrated adenomas (SSA) in histopathology images?"

Objectives and organization

The primary goal of this project is to use machine learning and deep learning algorithms so we are able to precisely differentiate between two types of polyps in the colon texture area: Hyperplastic Polyps (HPs) and Sessile Serrated Adenomas (SSAs). Our approach entails the development of two distinct models: a Machine Learning Model utilizing Random Forest and a Deep Learning Model based on Convolutional Neural Networks (CNNs).

The Random Forest model implements k-means clustering for image segmentation, aiding in noise reduction. We additionally generate a series of other features that facilitate the prediction of the image class. This model represents our more traditional machine learning approach, involving a carefully designed feature extraction process.

Conversely, for our Deep Learning Model, we feed raw images directly into the CNN, allowing the neural network to identify and learn features autonomously through several preprocessing methods. This model embodies a more modern, end-to-end learning approach that embraces the power of deep learning.

Finally, we assess their performance using a separate test dataset. By analyzing various metrics including accuracy, precision, recall, and F1 score, we evaluate the models' capability to accurately classify the polyps.

Related work

AI's ability to recognize characteristics and patterns beyond human visual detection enhances the process of identifying cancer. The usage of machine learning with histopathology not only streamlines the diagnostic process but also potentially saves lives by offering a greater chance of detecting cancer at its earliest and most treatable stages. In recent years, machine learning approaches in the analysis of histopathology images have seen significant growth and have been extensively explored using diverse methodologies.

Among these studies, the work by Xu et al. (2017) has been particularly influential, providing valuable insights into histopathology diagnosis and the use of deep learning techniques in histopathology image analysis. In their paper, Xu et al. propose a framework for large-scale tissue histopathology image

classification, segmentation, and visualization, using pre-trained CNN on ImageNet to extract activation features from the cell images [2].

Another noteworthy study was conducted by Peng et al. (2011), who developed a computer-aided method to segment individual glandular structures in prostatic adenocarcinoma [3]. They used a k-means clustering and region-growing method for segmentation and extracted quantitative features like average gland size, spatial gland density, and average gland circularity. The methodology was effective in accurately identifying prostatic adenocarcinoma based on these features.

Conceptual Framework

1. Histopathology

Histopathology is an essential tool in diagnosing and treating colorectal polyps, which is a small clump of cells that forms on the lining of the colon [4]. It is important to differentiate between benign hyperplastic polyps (HPs) and precancerous sessile serrated adenomas (SSAs). HPs possess unique characteristics such as superficial serrated architecture and elongated crypts, which set them apart from other lesions. On the other hand, SSAs exhibit broad-based crypts with complex structures and heavy serration, providing key differentiating factors. This distinction, while challenging, is crucial as SSAs require more vigilant management due to their potential cancerous transformation, while HPs are predominantly benign. By providing a detailed view of cellular structures, histopathology guides precise diagnosis and treatment planning. Our project leverages machine learning to augment this process, aiming for efficient and accurate colorectal polyp classification.

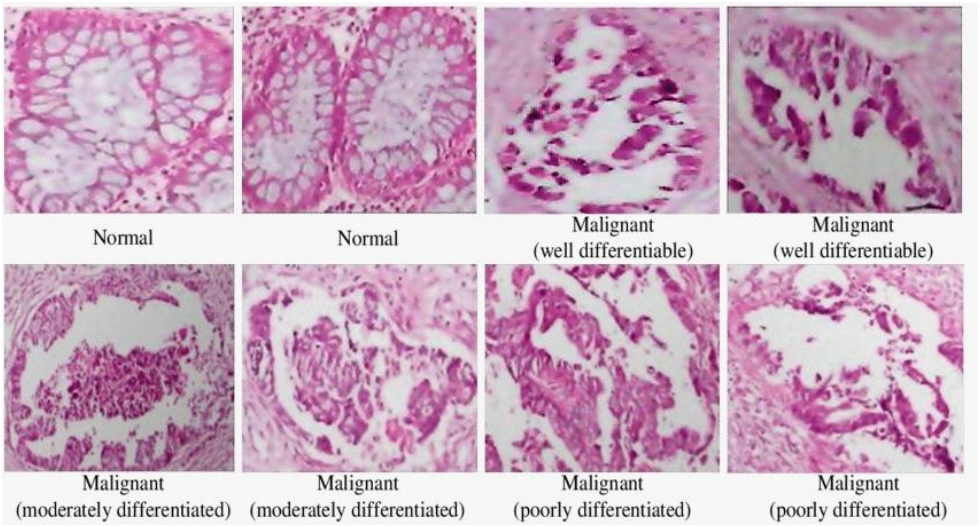

Figure 1
Examples of malignant colorectal biopsies. Source: Scientific Figure on ResearchGate

2. Model development

Convolutional Neural Network

Interpreting and understanding visual data, such as images or videos, is a complicated task known as visual processing. This is integral to computational models for functions like object identification or event spotting. One of the most proficient methods to manage this process is through the use of Convolutional Neural Networks (CNNs).

CNNs are a specialized deep learning model developed for processing structured grid data, such as images. They recognize the spatial correlation between pixels by learning image attributes using small squares of input data. CNNs are uniquely capable of learning spatial hierarchies of features automatically and adaptively, making them exceptionally efficient for a variety of visual processing tasks. These range from simple image recognition to complex tasks like medical image analysis, where they can help identify abnormalities or diseases.

Image Segmentation

Image segmentation is a common technique in digital image processing and analysis to partition an image into different regions, usually based on the characteristics of the pixels in the image. Typical usages of Image segmentation are separating foreground from background or cluster pixels based on color or shape similarities. [5] One effective way to perform image segmentation tasks is through the use of Random Forests (RFs).

Random Forests is a type of machine learning algorithm that operates by constructing a multitude of decision trees during the training phase and outputting the most common class for classification or mean prediction for regression of the individual trees. RFs are highly flexible and capable of handling a large amount of data with high dimensionality.

In the context of image segmentation, Random Forests can be trained to recognize and classify each pixel or group of pixels in an image, thereby distinguishing different regions based on learned features. This makes RFs particularly effective for tasks such as medical image segmentation, where they can help identify and delineate regions of interest, such as tumors or specific organs.

Methodology

1. Dataset Privacy

To access and utilize the MHIST dataset for this project, our team had to comply with a comprehensive set of privacy guidelines specified by Dr. Hassanpour's research laboratory, the data publisher. These guidelines ensure the ethical use of the dataset and safeguard the de-identified information it contains. The full details of the Research Use Agreement we adhered to can be found at the following link given below: https://bmirds.github.io/MHIST/Dataset%20Research%20Use%20Agreement.pdf. We are allowed to make a direct copy of the MHIST dataset for non-commercial research use within the guidelines of this agreement, which we shared with the CBS faculty for the purpose of fulfilling the requirements of our project. However, distribution, publishing, or reproduction of any portion or all of the MHIST dataset to others without explicit prior written permission from the data publisher is not permitted.

2. Dataset Description

The Colorectal Polyps Histology Image Dataset consists of 3,152 hematoxylin and eosin (H&E)-stained Formalin Fixed Paraffin-Embedded (FFPE) images of colorectal polyps. The dataset was obtained from the Department of Pathology and Laboratory Medicine at Dartmouth-Hitchcock Medical Center (DHMC) and is intended for research purposes in the field of digital pathology.

The images in the dataset have a fixed size of 224 by 224 pixels and represent various histological patterns of colorectal polyps. The dataset has been de-identified and released with permission from Dartmouth-Hitchcock Health Institutional Review Board (IRB).

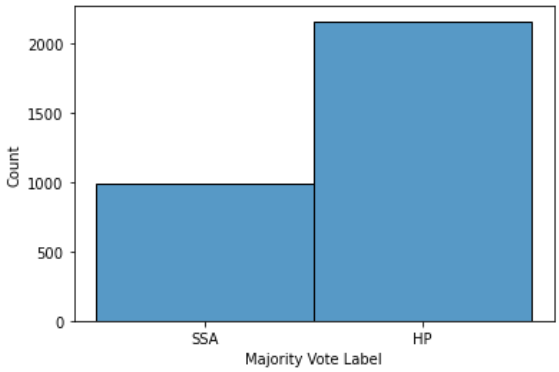

Figure 2: Majority Vote Label

The dataset focuses on a binary classification task called the MHIST Binary Classification Task, which aims to distinguish between two clinically important histological patterns: Hyperplastic Polyp (HP) and Sessile Serrated Adenoma (SSA). Each image is labeled based on the majority opinions of seven pathologists from the Department of Pathology and Laboratory Medicine at DHMC. A threshold of 4 is set,

and an image is classified as a SSA if four or more annotators vote in favor of SSA. Conversely, if fewer than four annotators assign the image as SSA, it is labeled as HP. Their expertise was used to determine the type of colorectal polyp represented in each image.

Initially, approximately 67% of the pictures in the dataset are labeled as HP, indicating the predominant histological pattern, while the remaining 33% are labeled as SSA (see Figure 2). This distribution reflects the majority-vote labels provided by the seven pathologists involved in the annotation process.

3. Data Analysis Process

The distribution depicted in Figure 3 represents the number of votes from annotators assigned to each classification. Notably, a substantial portion of the classifications falls within the range distinguishing between HP and SSA (3 & 4). This observation underlines the inherent challenge for domain specialists to confidently differentiate between HP and SSA cases.

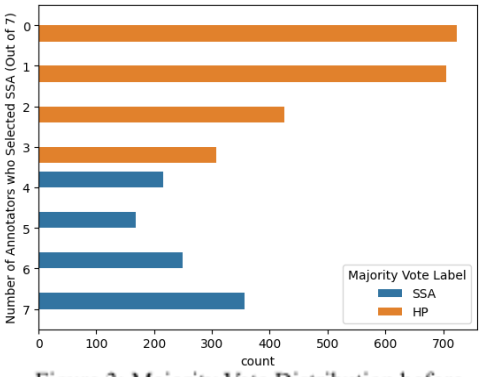

Figure 3: Majority Vote Distribution before the Pre-Process

4. Data Pre-Processing

When the dataset was initially examined, it became apparent that there was a substantial overrepresentation of HP labeled pictures, accounting for approximately 67% of the total images. Such an imbalance posed a challenge for accurately training and evaluating the model. Consequently, a deliberate effort was made to curtail the number of HP labeled pictures while simultaneously augmenting the count of SSA labeled pictures, ultimately paving the way for a more well-rounded and balanced dataset.

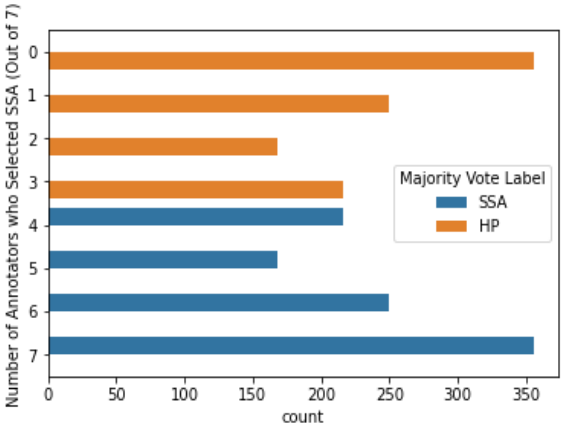

Figure 4: Majority Vote Distribution after the Pre-Process

This endeavor involved meticulously selecting and eliminating rows with lower agreement levels, strategically prioritizing those instances where the labeling agreement was less pronounced. By removing such instances, the dataset's composition was gradually transformed, tilting it towards a more equitable distribution between HP and SSA labeled images.

The objective behind this data preprocessing step was twofold. Firstly, it aimed to rectify the imbalance within the

dataset, allowing the model to acquire a better understanding and discernment between the two classes. Secondly, by achieving a more balanced distribution, the model's performance would be more accurate and reliable when confronted with real-world scenarios or new, unseen data.

Figure 4 shows the transformed image distribution after preprocessing. It is noticeable that the distribution has improved significantly and has become much more uniform compared to the previous state.

Modeling, Methods and Tools

1. Random Forest Classifier Model

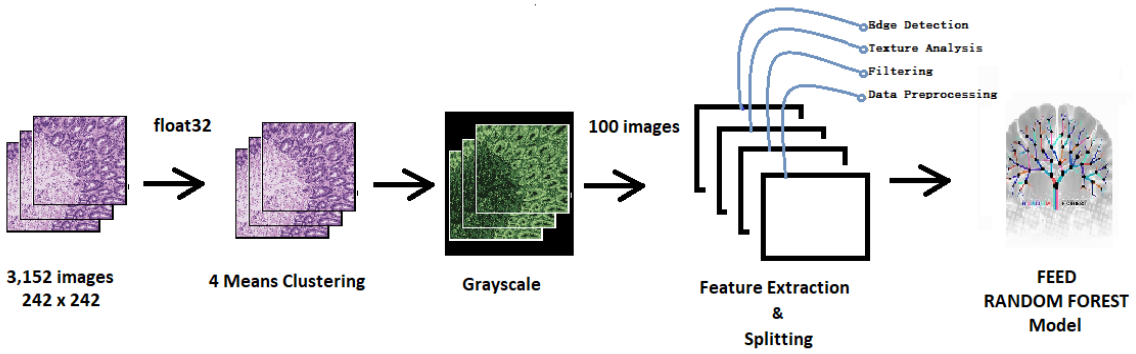

Figure 5: Random Forest Model Workflow

From the SSA vs HP extensive voting column, we saw that for some cases, even expert pathologists can disagree to a certain degree to diagnose a patient case. This is specifically because of the' unique features of benign and malignant tissues. Although formulated as a binary classification task, a gland can be diseased based on the diagnosis of benign background benign gland malignant background , malignant gland and also the arrangement of epithelial cells inside of the gland area. [6] We therefore formulate a classification problem, in which we distinguish four classes in our image data. We take the k means clustering approach to segregate these parts of a cell in accordance to their color intensity.

For these reasons, acknowledging the fact that the silhouette score, and curves in the histogram of pixel distribution may suggest otherwise, we limited our picture to segment between 4 cluster so that when we

convert the images to grayscale, the gland outlines, the insides of the glands, the mucous and the blank areas are clearly defined and we are able to reduce additional noise in color distribution of the pixels.

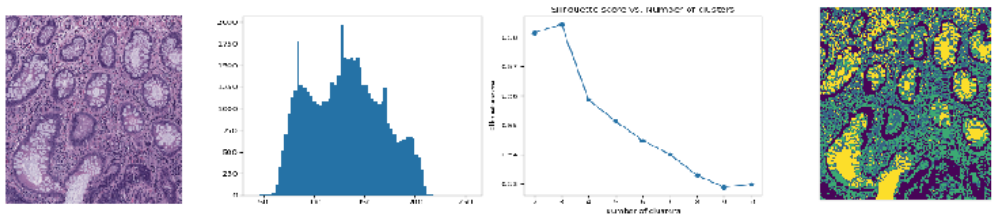

Figure 6: Clustering Approach for Segmentation

All the segmented images by 4-means clustering are stored in a different folder for further preprocessing. After that we take the segmented images, convert them into grayscale and perform feature extraction. It is to be noted that, in our implementation, we have only taken 50 random images from SSA and 50 from non SSA to train our random forest model because of lack of resources for using all of the images in the folder for training. Each feature has been extracted for an image, and the pixels with their corresponding feature values have been stored into a dataframe to train our model. The features extracted are described below :

1. **Gabor Kernels:** The transformed images obtained from the Gabor transformation highlight different aspects of the original image. For example, some kernels might enhance edges in specific directions, while others might enhance textures or patterns of a particular scale or orientation. By combining the responses from multiple Gabor kernels, a more comprehensive representation of the image's features can be obtained.
2. **Edge and Shape Detectors:** The transformed images obtained by various edge and shape detection approaches like Roberts edge, Sobel, Scharr and Prewitt highlights different aspects of edges and shapes.
3. **Filtering:** The filtered image by two different degrees of Gaussian blur, Median blur and Variance filter represents different visual features, such as smoothing, noise reduction, or texture analysis.

After storing these pixel values for the original grayscale image and all of these filters we get a pandas dataframe of 5017600 rows which we split into the training and testing set and fit into our model. We ran

our random forest model and obtained an overall accuracy of 84 percent where the precision in detection of HP is 86 percent and the precision in SSA is 81 percent.

The algorithmic complexity of our model is $O(n^5)$ as a resulting factor of the five nested for loops in the gabor kernel features extraction. It is to be mentioned that, as a sensitive medical diagnostics task, the features have not been shortened with respect to their significance to the model.

2. CNN Model Architecture and Iterations

A baseline model was constructed, consisting of three convolutional layers with the *Rectified Linear Unit* (ReLU) activation function. The convolutional layers are responsible for learning local patterns in the input image through the use of convolutional filters. The number of filters, or kernels, increases in each successive convolutional layer, allowing the network to learn increasingly complex features. The first layer uses 32 filters, the second uses 64, and the third uses 128. Each filter has a size of 3x3 pixels, a commonly used size in CNNs.

Batch normalization is used to normalize the activations of the neurons in a layer. It helps accelerate the training process and stabilize the learning process by reducing internal covariate shift.

Max pooling is performed after each convolutional layer, reducing the spatial dimensions (width and height) of the input. Pooling provides a form of translation invariance and reduces computational complexity by down-sampling the input's size.

A *fully connected part* (dense layers) is included after the convolutional and pooling layers. The output from the last pooling layer is flattened to a one-dimensional array and then fed into a dense layer with 128 neurons. A dropout rate of 50% is used here to prevent overfitting by randomly setting half of the input units to 0 during training. The final dense layer has two neurons, one for each class, and uses the softmax activation function to provide probabilities for each class.

The model is compiled using the *Adam optimizer*, a widely used algorithm that adapts the learning rate during training. The learning rate's optimal value is found using a learning rate range test. The model also uses the 'categorical_crossentropy' loss function, which is suitable for multi-class classification problems.

The baseline performed poorly on validation, peaking at 75% accuracy with high volatility and non-converging with training over 25 epochs. To achieve satisfactory performance behaviors: convergence, steady accuracy increase and plateauing trend, several model iterations were tested.

Given the complexity of the classification task at hand, model V3 was *deepened* with the addition of a convolutional layer with 256 filters to help capture more subtle features and patterns. This iteration showed higher, more consistent performances.

Model V4 was built using the same architecture as V3 but was trained on the pre-processed data. This model yielded mediocre results (54% accuracy on testing set).

To improve validation performances and improve convergence, model V5 implements *data augmentation* techniques such as rotation, width and height shifts, and horizontal and vertical flips are used to diversify the training data. *Learning rate scheduling* is implemented with the use of the ReduceLROnPlateau callback function from TensorFlow's Keras API. This function adjusts the learning rate dynamically during training, specifically when the model's performance plateaus or stops improving. This iteration showed more stable metrics hovering around 75%.

To accommodate class imbalance in the dataset, model V6 implements a class weighting function to increase the underrepresented SSA class's weight in respect of the inverse of its frequency. This iteration further improved performance stability and convergence with an increasing trend at 40 epochs.

The last iteration of the model, V7, implements an *initial learning rate optimiser*, computing an optimum value around 10e-2.5, associated with the sharpest decrease in validation loss. Training and testing split was changed from .8/.2 to .7/.3 respectively. Epochs were increased to 80 to ensure a holistic overview of performances. This iteration showed promising prospects with stable, converging and plateauing performances around 77%-80% with some overfitting.

In its final version (V7), the CNN model presents $O(NMK^2F^2E)$ complexity where:

- N is the number of input samples.
- M is the number of output channels or filters in the convolutional layer.
- E is the number of iterations or epochs for training the model.
- F is the size of the input image.
- K is the size of the filter in the convolutional layer.

Note that time complexity is also subjected to the data augmentation technique applied, adding computation overhead to the image size factor.

Results

The Random Forest classifier and Convolutional Neural network yielded comparable performances against testing sets. The CNN displayed satisfactory behavior with stable accuracy and loss curves.

	Precision	Recall	F1-score
RF	0.84	0.83	0.83
CNN	0.81	0.92	0.86

Table 1: Precision, Recall & F1-Score

Fine tuning CNN hyperparameters constituted the bulk of the coding work. Since gridsearch proved too computationally heavy, only primary parameters such as class weights, training and testing sets size, initial learning rate, number of epochs and network complexity (number of layers) were tuned throughout the workflow.

The Random Forest Model gives an 84 percent of accuracy. Nonetheless, the accuracy can change for different interactions of the images. The model also is prone to overfit the data and the degree of overfitting is hard to find unlike deep learning approaches. However, in comparison to other machine learning classifiers, Random Forest gives a good amount of fitting for binary classification. Although the number of trees can dominate the results of the model, finding an optimum number is not difficult like deep learning methodologies.

Random Forest classifiers can also be a great tool to classify or extract features that integrate with CNN or other deep learning methodologies. It is a very good model to pick for a lower quantity of images. For example, we could have a few images with SSA polyp regions annotated by experts, we could use them to draw in more regions in all the images. It is also to be mentioned that training the model with more data available could have yielded better results. Considering the computational limitations encountered, these models yielded satisfactory results and show promising prospects for further hyper-parameter optimization.

Limitations

When examining our project results, it is crucial to take into account the various limitations we encountered throughout the study. By recognizing and addressing these constraints, we can enhance the validity and reliability of our conclusions. Therefore, we will below discuss the most relevant limitations we encountered during our research process.

1. Class Imbalance

Determining whether a class imbalance is greater or lesser depends on the context and problem at hand. While a class distribution of 67% and 33% does not significantly affect the performance of the model in certain cases, it could have adverse consequences in other cases, especially if the minority class contains important or critical information. Therefore, defining the significance and relevance of the minority class is subjective and prone to interpretation.

There are other approaches and variations that were not mentioned in our project, and their effectiveness may vary depending on the data set and the problem. For example, one such approach is the use of cost-sensitive learning. In this technique, different misclassification costs are assigned to different classes, emphasizing the importance of accurately predicting instances from the minority class. By incorporating the misclassification cost into the learning process, cost-sensitive learning provides a tailored solution to the imbalance between classes.

Another variation is the combination of oversampling and undersampling techniques known as hybrid sampling. This approach aims to take advantage of both undersampling and oversampling by selectively applying them to different regions of the feature space or at specific stages of the learning process. Hybrid sampling methods provide flexibility and adaptability in dealing with class imbalances based on the features of the dataset.

2. Computational Overhead

Throughout this work, computing capability proved to be the prominent limiting factor. The models designed required substantial computing resources for training, optimization thus intractable on consumer grade laptops. To accommodate these limitations, we turned to Google Colab A100 GPU. This allowed us to complete our model's training more efficiently, reducing the overall time required for experimentation and iteration. Opting for A100 GPUs did not alleviate the computation capability

problem all together. Model optimization was still limited, especially holistic methods such Gid Search were still intractable.

Although Google Colab GPU offers access to GPUs, it is essential to acknowledge the associated costs. While the platform provides no free GPU usage, as such, it was necessary for us to consider the financial implications of utilizing Google Colab GPU for our project.

Conclusion & Future Work

"Can machine learning and deep learning enhance the classification accuracy between hyperplastic polyps (HP) and sessile serrated adenomas (SSA) in histopathology images?"

The project's results positively answer the above question, demonstrating that both machine learning and deep learning can effectively help in the classification of hyperplastic polyps (HP) and sessile serrated adenomas (SSA) in histopathology images.

With the traditional machine learning approach, our model utilizing Random Forest and k-means clustering for image segmentation has exhibited its robustness in isolating pertinent features from complex data, thereby enhancing the classification between HP and SSA, giving a superior accuracy while training and testing the model.

On the other hand, using the deep learning methodology, our Convolutional Neural Network (CNN) model leveraged the power of end-to-end learning. By processing raw images directly and learning essential features autonomously, this model has shown significant accuracy in differentiating between HP and SSA.

With more advanced models, improved data, and the integration of multiple state-of-the-art DL models, we believe there is room for even greater precision in classification tasks. We believe that improved classification accuracy in the identification of colorectal polyps could lead to more timely and appropriate treatment strategies. This, in turn, could potentially hold the risk of these polyps progressing to colon cancer, contributing to a decrease in mortality rates associated with this disease, which is expected to rise over the next 17 years. Moving forward, we see a bright horizon for the application of machine learning in medical diagnostics. We are excited to witness the evolution of the medical field and see how AI can support healthcare and patient prognosis, as the potential is unlimited.

Bibliography

[1] International Agency for Research on Cancer. (2022). Colorectal Cancer Awareness Month 2022. https://www.iarc.who.int/featured-news/colorectal-cancer-awareness-month-2022/. Accessed May 17, 2023.

[2] Xu, Y., Jia, Z., Wang, L. B., et al. (2017). Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinformatics, 18, 281.

[3] Peng, Y., Jiang, Y., Eisengart, L., Healy, M., Straus, F., & Yang, X. (2011). Computer-aided identification of prostatic adenocarcinoma: segmentation of glandular structures. Journal of Pathology Informatics, 2(1), 33. DOI: 10.4103/2153-3539.83193.

[4] Mayo Foundation for Medical Education and Research. (n.d.). Colon polyps - Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/colon-polyps/symptoms-causes. Accessed May 17, 2023.

[5] The MathWorks, Inc. (n.d.). Image Segmentation - MATLAB & Simulink. https://se.mathworks.com/discovery/image-segmentation.html. Accessed May 17, 2023.

[6] Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization https://peerj.com/articles/3874/ Accessed: 2023-05-17.