Machine Learning Algorithm for the Classification of Appendiceal Cancer

Preliminary Design Review

December 4, 2020
Fall 2020
EGR 314

Team Member Names

By signing, students consent to have reviewed the contents of this report and to agreeing with the content presented.

Gabrielle Prichard : Gabrielle Prichard
Margaret Nyamadi : Margaret Nyamadi
Bryan Bennett : Bryan Bennett
Ethan Cooley : Ethan Cooley
Hao Tong : Hao Tong

Faculty Coach Name(s)

Dr. Olga Pierrakos
Dr. Melissa Kenny
TA: Josh Copus

Table of Contents

EXECUTIVE SUMMARY

The lack of established risk factors and the prevalence of tumor diversity associated with appendiceal cancer contributes to the high incidence of appendiceal cancer misdiagnosis. This project seeks to address the issue of appendiceal cancer misdiagnosis through the use of a machine-learning algorithm to classify appendiceal cancer specimens.

The primary stakeholders involved in this project are Dr. Votanopoulos, pathologists at Wake Forest Baptist Health who will utilize this technology in appendiceal cancer diagnosis, and patients who suffer from appendiceal cancer. To develop the system requirements of the machine learning algorithm, the team must identify primary attributes associated with an appendiceal cancer diagnosis, must investigate machine learning models used for image processing, benchmark existing solutions, and examine codes and regulations. Appendiceal cancer is typically diagnosed following an appendectomy and an analysis of a histopathologic specimen. During analysis, pathologists look for the presence of tumors, tumor size, and tumor location. After researching various machine learning models, the team found that Neural Networks are effective in medical image classification and analysis. Various platforms and toolkits are available for use, all of which will be experimented with to determine which one is best fitted to our project. The team has identified various benchmarked solutions that illustrate the efficacy of using machine learning for medical diagnosis and disease classification. Some examples outlined in this paper include the identification of melanoma, diagnosis and classification of diabetic retinopathy, and the diagnosis and classification of prostate cancer.

According to the Food and Drug Administration (FDA), when developing and implementing machine learning algorithms for medical diagnosis, there are various guidelines in place to ensure that the algorithms developed for medical diagnoses are safely and effectively used. This project uses clinical data to make conclusions that may impact classification and treatment decisions. As a result, an Institutional Review Board (IRB) is required to ensure the project and its methods are ethical and clinically significant. Once the IRB approval is secured, the next steps include concept generation and selection. Concept generation involves investigating key functions of the system to assist in identifying potential solutions to yield an optimal result. Given project constraints and available data, the appropriate concept or combination of concepts will be selected and development and testing will follow. Following the development and production of one concept, the project will branch out and take on various other tasks including linking patient clinical data for prognosis and analyzing time-series data. In addition, the relationship between the Extracellular Matrix (ECM) profiles and appendiceal cancer subtype will likely be investigated following the completion of this project.

1 PROBLEM DISCOVERY

1.1 Problem Statement

Appendiceal cancer is rare, with an age-adjusted incidence of 0.12 cases per million people per year (McCusker et al., 2002). Appendiceal cancer is an understudied region of oncology, and risk factors, treatment, and histopathology are areas where additional consensus is needed. This contributes to high misclassification rates among appendiceal cancer patients, often leading to the improper use of expensive and cytotoxic treatments. This project seeks to address the issue of appendiceal cancer misdiagnosis through the use of software tools (i.e. image processing, machine learning, etc). The goal of this project, which will last a total of 9 months, is to create a software tool based on histological samples of appendiceal cancer specimens to more accurately classify appendiceal cancer subtype. Ultimately, the primary stakeholders of this project are the physicians who will benefit from this technology and future patients whose finances and well-being will benefit from more targeted treatments.

1.2 Mission Statement and Broader Impacts

The subtypes of appendiceal cancer are often misclassified due to subjective and overlapping diagnostic criteria. Treatment for appendiceal cancer varies with subtype, so an incorrect misdiagnosis can result in inappropriate treatment and prognosis. A tool that aids pathologists in accurately classifying appendiceal cancer would reduce false positive and false negative diagnoses, increasing overall survival rates. This approach would also contribute to the medical community's understanding of appendiceal cancer subtype classification, and can serve as a stepping stone for similar models. The framework for this project can be utilized by other researchers to develop models for diagnostic purposes. This project is applicable to almost all conditions that are diagnosed primarily via biopsies and/or medical images.

1.3 Stakeholder Analysis

The primary stakeholders associated with this project are Dr. Votanopoulos, pathologists at Wake Forest Baptist Health, patients affected by appendiceal cancer, and insurance companies. Dr. Votanopoulos is the primary stakeholder associated with this project, as he is a surgical oncologist at Wake Forest Baptist Health who works closely with appendiceal cancer and tumors. After consulting with Dr. Votanopoulos, it is clear that the diagnosis of appendiceal cancer has proven an extremely challenging area. Very little research exists surrounding the diagnosis and classification of appendiceal cancer. As a result, there is a lack of consensus on how to properly classify appendiceal cancer subtype, which can lead to misdiagnosis. Dr. Votanopoulos has expressed interest in working closely with pathologists to address this challenge. Ultimately, this machine learning model will serve as a complementary tool for pathologists at Wake Forest Baptist Health to utilize for appendiceal cancer diagnosis and classification. The model will support the pathologists, which in turn supports the physician in providing a patient with a prognosis.

Patients suffering from appendiceal cancer will benefit from this technology as well. Appendiceal cancer patients are often misdiagnosed and receive unnecessary treatment for their misdiagnosis. This product will assist a pathologist in detecting and classifying appendiceal cancer, therefore benefiting the patient. Insurance companies will ultimately benefit from the product as well. Cases of cancer misdiagnosis cost insurance companies more money, so decreasing the occurrence of misdiagnosis of appendiceal cancer will save insurance companies money in the long term.

1.4 Background Research: Literature Review

Histopathology

In classifying appendiceal cancer, pathologists analyze several different features on a macroscopic and microscopic scale. On a macroscopic scale, the tumor site, tumor size, and tumor configuration are analyzed, and on a microscopic scale, individual attributes of each subtype are used in the classification (College of American Pathologists, 2006). Developing a machine learning algorithm that can accurately determine if a given image contains signs of appendiceal cancer requires background knowledge regarding what visual characteristics pathologists look for when making a diagnosis. Although classification criteria have changed over time, uniform guidelines from the World Health Organization (WHO) now exist that facilitate the histopathological interpretation of appendiceal epithelial neoplasms (abnormal growth of cells within the epithelium).

The WHO guidelines released in 2019 specify five main subtypes of appendiceal cancer. They are: mucinous neoplasms, adenocarcinoma, neuroendocrine neoplasms (NEN), serrated lesions and polyps, and goblet cell adenocarcinoma (Nagtegaal et al., 2020). Each category has its own various subtypes, which are grouped together based on their similarities in histopathology, pathogenesis, and prognosis. Subtypes are often categorized as low-grade or high-grade, sometimes on a scale from 1 (low-grade) to 4. Increasing grade means the cells within a tumor are less differentiated and appear in a more uniform and organized pattern. The higher the grade, the more dangerous the cancer often is. H&E stains will be used to discern structures within the slides. This staining method causes cell nuclei to turn blue and other tissues to turn pink.

The first subtype is mucinous neoplasms, which are characterized by excessive mucin production that oftentimes extends outside the tumor and into surrounding tissues and organs. Large pools of mucin (a glycoprotein that is used in the production of mucus) may exist within and around the tumors, and may even penetrate the appendix wall into the peritoneal cavity. Two subtypes exist within this classification, low-grade appendiceal mucinous neoplasms (LAMN) and high-grade appendiceal mucinous neoplasms (Nagtegaal et al., 2020).

LAMN is categorized by mucinous epithelial proliferation that disrupts the typical appendiceal mucosa (mucous membrane). Tumors often contain a large volume of mucin in vacuoles (compartments that hold fluid) that cause the nuclei within the tumor to compress. Because LAMN is a low-grade cancer, the cells within the tumor are well-differentiated (relative to HAMN and other high-grade cancers) and display relatively mild levels of dysplasia (see figure 12 in appendix). The wall of the appendix oftentimes becomes translucent, calcified, or fibrotic (scarred) (Nagtegaal et al., 2020). See figures 1(a) and 1(b) below.

HAMN is less common than LAMN and is similar in terms of the production and presence of mucin within and around tumors, nucleus stratification, and fibrosis. The main difference here is the presence of high-grade features, namely that cells within the tumor are poorly differentiated and appear in an organized, uniform pattern.

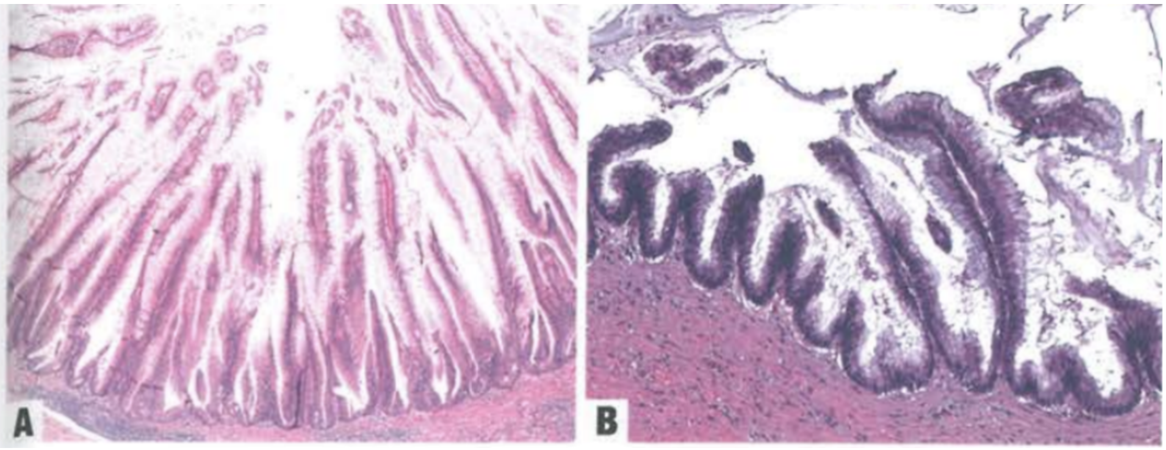


Figure 1(a) and 1(b): Two cases of low-grade appendiceal mucinous neoplasms (LAMN). Figure 1(a) shows the normal epithelium replaced by a well-differentiated tumor with extracellular mucin buildup. 1(b) illustrates the tendency for epithelial cells to elongate. Extracellular mucin buildup is also apparent (Nagtegaal et al., 2020).

The second main subtype of appendiceal cancer is adenocarcinoma, which is characterized by infiltrative invasion and a desmoplastic stromal response (the surrounding tissue has a fibrotic or scarring appearance). Adenocarcinoma is generally a more aggressive cancer. The subtypes of adenocarcinoma include mucinous adenocarcinoma, signet ring cell adenocarcinoma, and colonic-type adenocarcinoma.

Mucinous adenocarcinoma is differentiated from HAMN (another high-grade mucinous neoplasm) and LAMN in its pattern of invasion and spread. LAMN and HAMN tend to expand broadly with a tumor front, whereas mucinous adenocarcinoma is focally invasive. In mucinous adenocarcinoma, extracellular mucin pools compose more than half of the tumor, and chunks of epithelial cells or mucoceles often float within the pools (Nagtegaal et al., 2020). See figure 2(a).

Signet ring cell adenocarcinoma is another high-grade adenocarcinoma, as the presence of signet ring cells usually always constitutes a high-grade classification. Signet ring cells comprise over half the tumor and often invade surrounding tissues (Nagtegaal et al., 2020). See figure 2(b) below.

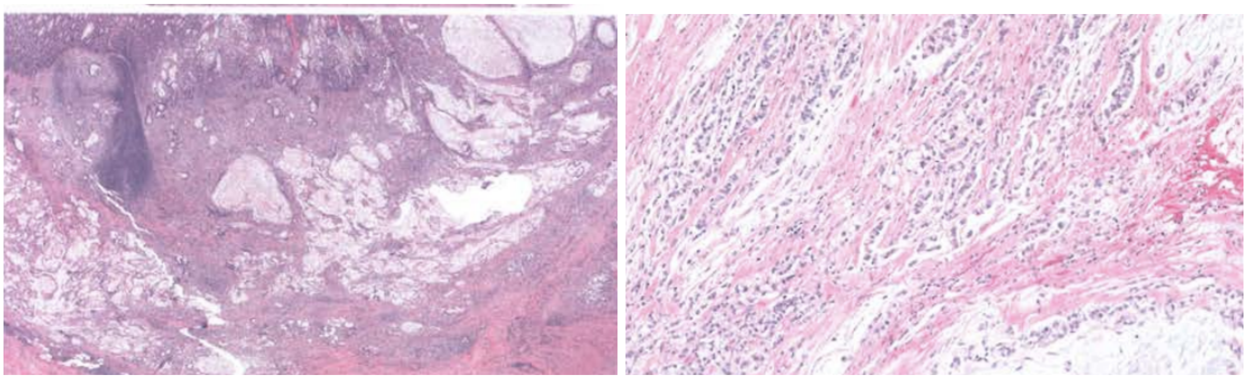


Figure 2(a) and 2(b): Figure 2(a), left, is mucinous adenocarcinoma in low-power view. Notice the extracellular pools of mucin. 2(b), right, is signet ring cell adenocarcinoma in low-power view, illustrating infiltration of the appendiceal wall and extracellular mucin (Nagtegaal et al., 2020).

Neuroendocrine Neoplasms (NENs) are the third subtype this project will focus on, and consist of neuroendocrine tumors (NETs) and neuroendocrine carcinomas (NECs). NETs are well-differentiated, small (mean diameter < 1cm), localized tumors that correspond with high levels of localized inflammation, particularly in the lymph nodes. Desmoplastic stromal response (the surrounding tissue has a fibrotic or scarring appearance) is common, and the tumors often group together with localized glandular development. Most NETs are grade 1 or grade 2 (low grade), but grade 3 (high grade) can also occur. Three-quarters of appendiceal NETs occur in the distal appendix, and they often enlarge the nodular wall or submucosa. Interestingly, some studies estimate NETs to account for 80% of all appendiceal tumor specimens, but only 11% of malignant appendiceal tumors (Leonards et al., 2017). Neuroendocrine carcinomas are poorly differentiated and are classified as either large-cell or small-cell NECs. NECs are rare and will likely have little relevance to this project (Nagtegaal et al., 2020).

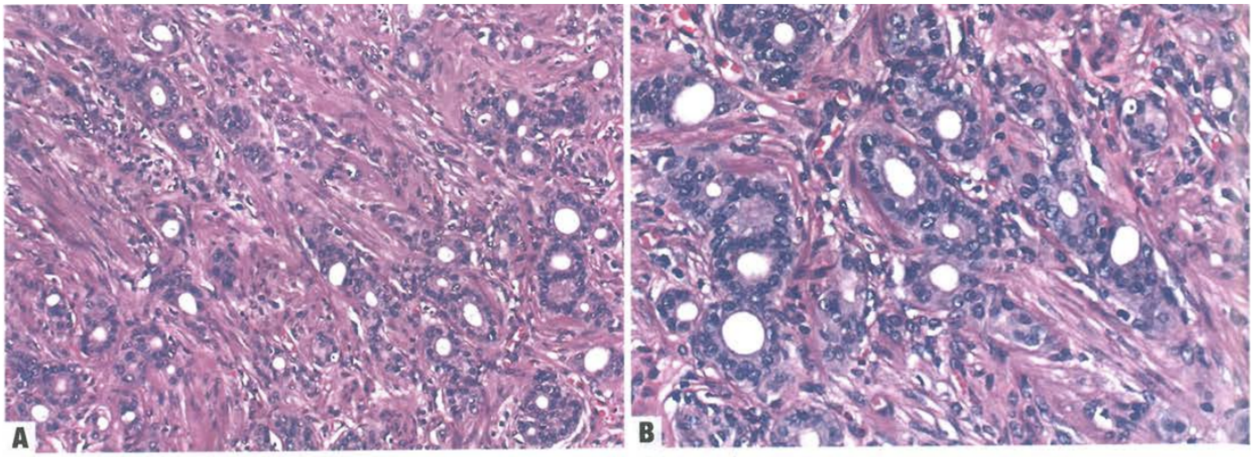


Figure 3(a) and 3(b): Neuroendocrine tumor in low-power (a) and high-power views (b) (Nagtegaal et al., 2020, #).

Appendiceal serrated lesions and polyps is the fourth subtype of appendiceal cancer. These are mucosal epithelial polyps that are often characterized by serrated (sawtooth or stellate) architecture of the crypt lumen and can occur throughout the appendix. Serrated lesions without dysplasia feature crypts that are serrated with abnormal architecture, crypt dilatation, and varying L shapes and inverted T shapes. There could also be mild cytological atypia with dystrophic goblet cells and possible mitotic figures. In addition, there may be an abundance of luminal mucin. Serrated lesion with dysplasia can take 3 forms:

I. adenoma-like dysplasia
II. serrated dysplasia
III. traditional serrated adenoma-like dysplasia

In a single polyp, multiple morphological patterns of dysplasia may be present. Serrated dysplasia maintains the architecture of the crypts, but the crypts are lined by cuboidal to low columnar cells with hyperchromatic enlarged nuclei, reduced cytoplasmic mucin, and increased mitosis. Typically, adenoma-like dysplasia develops a villous growth pattern with elongated, hyperchromatic nuclei with pseudostratification, increased mitosis, and apoptotic bodies.

The fifth subtype of appendiceal cancer is the goblet cell adenocarcinoma, which is an amphicrine tumor composed of goblet-like mucinous cells and endocrine cells and Paneth-like cells, normally arranged as tubules resembling intestinal crypts. It is usually located away from the center of the appendix. Signet-ring cell adenocarcinoma shows a greater than 50% disorganized growth of signet-ring cells with high-grade cytological features which differs from high-grade goblet cell adenocarcinoma by the absence of a recognizable low-grade goblet cell adenocarcinoma component.

Low-grade goblet cell adenocarcinoma features tubules of goblet-like mucinous cells, endocrine and paneth-like cells with granular eosinophilic cytoplasm, mild nuclear atypia, and infrequent mitoses, extracellular mucin, and circumferential involvement of the appendix wall by tumor cells, without stromal reaction.

High-grade goblet cell adenocarcinoma features tumor cells infiltrating as single mucinous or non-mucinous cells, complex anastomosing tubules, cribriform masses, sheets, or large aggregates or goblet-like or signet-ring cells with high-grade cytological features, numerous mitoses with atypical mitotic figures and necrosis desmoplastic stromal response.

Regardless of the sub-classification of appendiceal cancer, most cases share common macroscopic characteristics. This may include swelling, fluid buildup in the peritoneal cavity, and enlarged omentum due to metastatic infiltration. The presence of mucin, signet ring cells, tumor cell differentiation (grade), tumor size, and pattern of invasion will prove essential features in determining subtype for this project.

Machine Learning

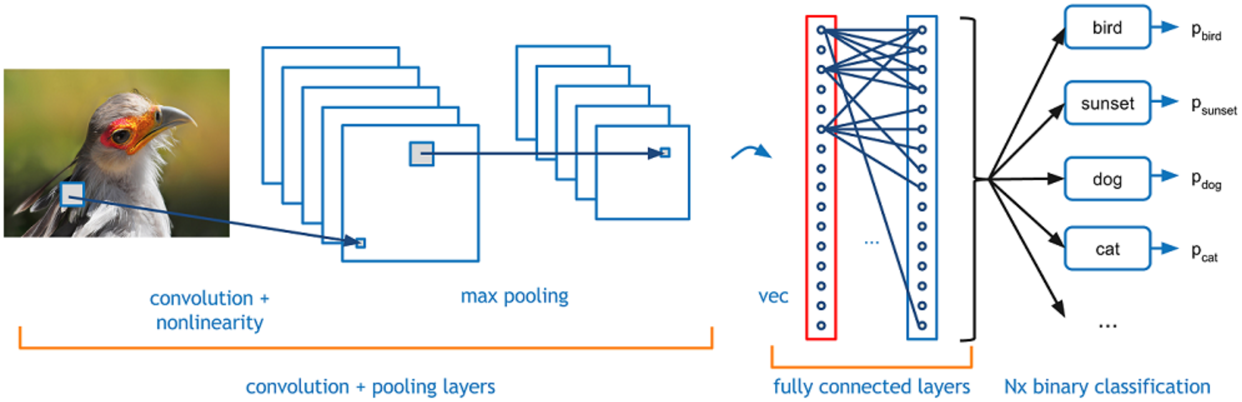


Figure 4. Overview of the convolutional neural network involving the convolution layer for feature extraction and the vector layer used to make predictions. (Convolutional Neural Network, 2019)

At a high level, machine learning is a way for computers to identify patterns in a given dataset and make conclusions from those patterns. There are various types of machine learning, however, the most common are supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, multi-task learning, ensemble learning, neural network learning, and instance-based learning. Each type of machine learning utilizes a different approach to handling data. For instance, the specific algorithms that utilize a neural network learning framework possess a layer that takes input, one or more layers that process the input, and a layer that displays the output (Dey, n.d.). For the purposes of this project, it is imperative to identify an algorithm that can process images.

The most common machine learning technique used in image classification is the convolution neural network (CNN). In image recognition, CNN utilizes a two or three-dimensional structure that analyzes grayscale or color images. It analyzes black and white layers of grayscale images, and the red, blue, and green layers for colored images. CNN ultimately recognizes and extracts important features to classify images (*Convolutional Neural*

Network: How to Build One in Keras &PyTorch, n.d.). The appendiceal cancer specimens are colored via H&E staining, therefore the algorithm will use a three-dimensional structure.

Examining Figure 4, the first part of the algorithm is the convolution phase, where many filters are applied to the image and each filter is used to detect some kind of features such as vertical edges or horizontal edges. Once the image has gone through the filters, the output response then undergoes an activation function to ensure the nonlinearity of the output. The nonlinearity is important because it ensures that each layer of the model learns something new, a linear output of all layers would cause any additional layers to be useless as they can be compressed into a single liner layer. That output is then used as input for the second step, max pooling, which essentially downsamples the image by half. The pooling stage results in a summary of the most important information in the image. The convolution step can be repeated multiple times to achieve the best selection and optimization of features. Once the convolution portion of the algorithm is complete, the final output will be converted to a vector of values that represent the image based on those features. This vector is then fed into the fully connected layers or the more traditional neural networks that make predictions based on the input. The output prediction is then compared to the true label of the input image, and a loss function is calculated. Using the loss function, the program goes back to each layer of the network to adjust values such as weight vector, bias, and the filter values following the gradient descent. Once all the values have been calculated through this backpropagation step, the new values are employed to make the new predictions, and more adjustments will be made to those adjustable weights. This cycle will be conducted many times until an accurate model is produced. In practice, machine learning algorithms have been utilized in medical diagnoses such as lung cancer and diabetic retinopathy.

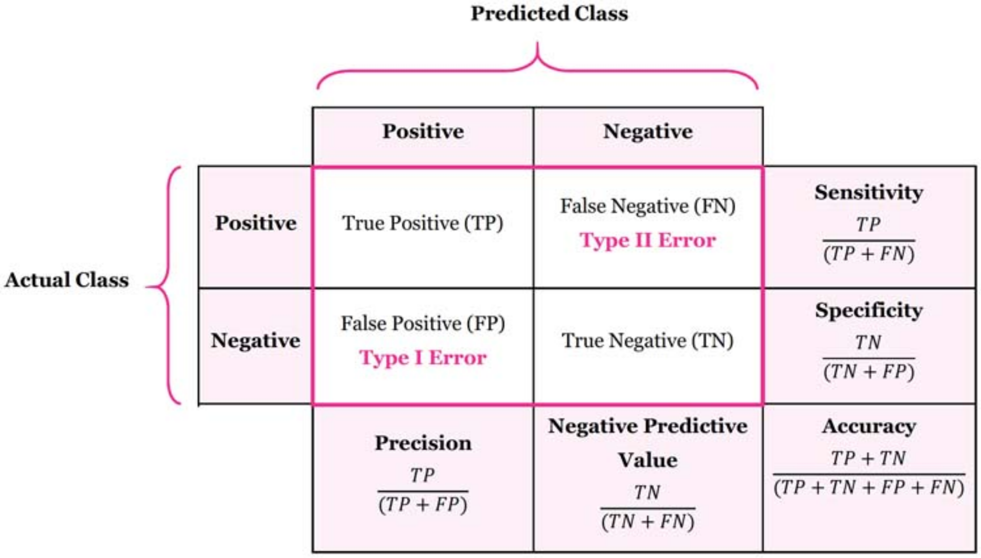
As stated above, feature extraction is an essential component of deep learning. For whole slide image feature analysis and extraction, models must have the capability to extract low-level, mid-level, and high-level features. Examples of low-level features include vertical and horizontal lines, edges, and corners. Mid-level features are more complex and are typically either geometric shapes or patterns, and high-level features include identifiable objects. When appropriate features have been extracted from images (either through convolution layers in a deep learning network or through an image processing algorithm), machine learning models can be utilized to interpret the relationships between the features and image tags. A machine learning model will learn how to associate certain features with a given tag or image classification based on certain patterns and trends in the data. The loss function is a method used to determine how well a machine learning model performs when making a classification or prediction, and allows the model to perform optimization to improve prediction accuracy. A low loss function value would indicate that the model is making accurate predictions. The mean squared error function is a type of loss function that measures the average squared difference between the actual data and the predictions that the model makes (Parmar, 2018). The mean squared error can be calculated using the following equation:

$$Mean\ Squared\ Error\ =\ \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}$$

Accuracy is an essential component to consider when analyzing the results of a machine learning algorithm. In the context of this project, it is imperative that any machine learning component has the ability to classify appendiceal cancer subtype within a certain accuracy threshold. When machine learning models learn to associate certain visual features with specific tags, accuracy can be measured by the number of false-positives and false-negatives. In addition, precision, recall/sensitivity, and specificity are all essential to consider in the context of machine learning. Precision indicates what percentage of indicated positive labels by the model were actually true positives. Recall (also known as sensitivity) indicates what percentage of all positive conditions were predicted correctly as positive. Specificity is the proportion of actual negatives which are classified as true negatives by the model. An F1 score is often used to ensure that there is a balanced relationship between precision and recall, and is calculated using the formula.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

A confusion matrix is a very useful tool that outlines the performance of a machine learning model. For a binary classification system (i.e. labeling something as either "positive" or "negative,") the following confusion matrix can be implemented to outline system performance:



		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	**Negative Predictive Value** $\frac{TN}{(TN + FN)}$	**Accuracy** $\frac{TP + TN}{(TP + TN + FP + FN)}$

Figure 5. The figure above shows how each performance metric is calculated in a binary classification case (What is Confusion Matrix and Advanced Classification Metrics?, 2019).

For a multi-class classification system, a confusion matrix with additional rows and columns for each label would need to be implemented. An ROC curve, or a receiver operator curve, is often used to show how good the model is predicting in a binary classification situation while varying the threshold values. Based on Figure 6, the blue dotted line represents the accuracy achieved by a model that guesses randomly, which is 50% accurate. The orange line represents the model produced by the algorithm and the closer to the top left, the lower the false positive rate, the higher true positive rate and the better the model. From this same graph, AUC or area under the ROC curve, it can be calculated by existing algorithms once the ROC is constructed. It is an aggregate measure of a model's performance as it takes into account all threshold values and makes a score representing how the model performs under different conditions. A score of 1 means the best performance and highest accuracy, while a score of 0 means worst performance where no prediction is correct under any conditions (Narkhede, 2018).

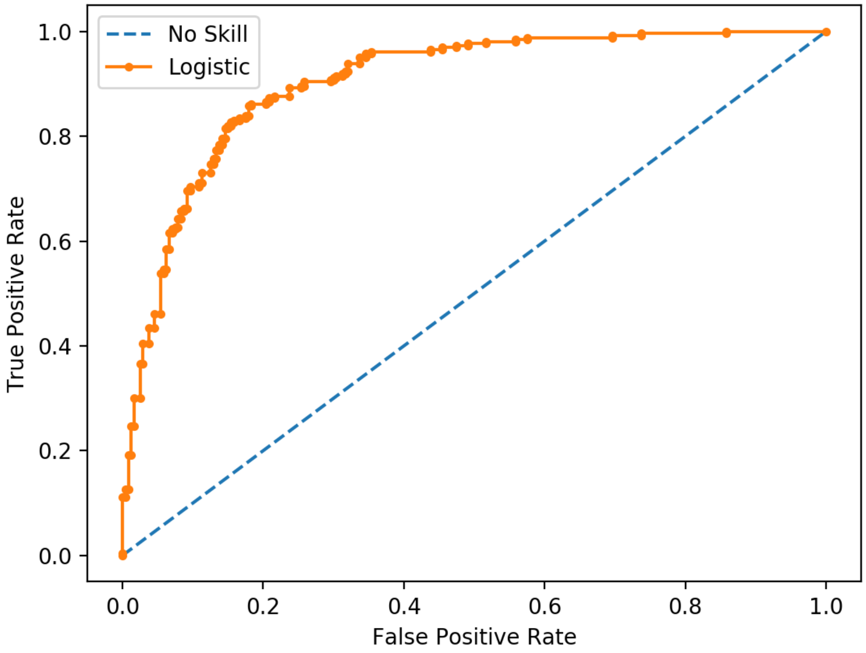


Figure 6. The figure above shows an example of what a potential ROC curve would look like in comparison to a random guess which is the blue line giving a 50% accuracy (Brownlee, 2020).

Another important factor to consider when using deep learning networks is to ensure the model is generalizing the patterns it extracts between all classes instead of memorizing the training data. This is called overfitting, where the model has excellent performance with the training data but poor performance with the validation set and real-world data. This is because the model has effectively memorized the training data instead of learning the general patterns associated with each classification. To remedy this, various approaches exist such as "dropout". Dropout randomly alters weights within the network causing the layers to re-wire information flow (Srivastava et al., 2014). It can prevent the network from over-learning a specific set of images. An important part of testing for overfitting is ensuring there are no overlapping images in the training and validation set. The best way to do this is to have the training and validation images come from different institutions, annotated by different pathologists under different imaging conditions (Shah, 2017).

Another way to ensure the model is generalizing patterns associated with each subtype is to add variation to color, brightness, and other image attributes prior to model uptake. This would be done in an image pre-processing pipeline, which ensures the training set will adequately represent real-world variations in images. For this project in particular, there is a risk the model only learns to classify images with specific staining and imaging attributes because the training set will consist of images from a select number of institutions. It therefore may not be able to classify images by subtype if they come from an outside institution with different staining or imaging techniques. To ensure the model is generalizing all patterns within each subtype so that it works for images of outside institutions, an image preprocessing pipeline can prevent the model from learning attributes that are specific to the training set.

The next factor to consider is time and space complexity. The process of training a neural network has a significant computational cost regardless of whether the model is trained locally or on the cloud. This is especially true with deep learning convolutional neural networks, because the multiple layers of computational execution and large data throughput from high-resolution images lead to a significant number of required operations. Popular GPU's (Graphics processing units) often cost several dollars per hour and algorithm execution can easily reach into the thousands price range if not developed efficiently. Predicting the time and space complexity of our model's training will allow for a deeper understanding of the feasibility and guide the optimal decisions to make regarding platform, network architecture, and input data volume (Justus et al., 2019).

One approach for predicting and understanding the complexity of a ML model is to compare it to similar models that have known performance. While these predictions still provide only rough estimates, it is a common approach to planning out model architectures (Justus et al., 2019). As the team develops our architecture, such comparisons will be made to benchmark the efficiency of our solution and ensure the software runs efficiently.

Finally, the last factor to consider when developing algorithms or models is the user experience. While it is not the most important factor, it is essential for the user to be able to interact with the interface easily and effectively. A certain level of abstraction is necessary to prevent unnecessary complexity for the user. Abstraction is a principle in computer science that emphasizes hiding lower levels of algorithmic complexity from the user to prevent cognitive overload. The user simply needs to know how to use the algorithm and not how it works. The following are principles of good user interface design to consider (Adobe, 2019):

(a) Place users in control - A good algorithm places the user in control by ensuring that the interface is easy to navigate, provides useful results or feedback, and contains visual cues to direct a user through the steps to take to achieve a particular outcome. In addition, the user interface can indicate the status of a system to show the user potentially how long it will take to get results.

(b) Encourage a comfortable interaction with Interface - The user interface should avoid jargon that could confuse users unfamiliar with AI-specific terms and the algorithm should be easily accessed from a laptop or desktop. An important factor to consider error handling. The algorithm should include proper error handling for user-entered data and avoid crashing, breaking, or looping infinitely.

(c) Reduce cognitive overload - It is important to ensure that the user is not overwhelmed by the interface. There should not be excess information that clouds the judgment or understanding of the user. Instead, the focus should be on visual clarity and an easy to understand interface.

(d) Apart from a graphical user interface, a simple API call can be used that sends an image to a server and retrieves a subtype classification. This still requires the development of the "front end," but does not involve a GUI.

1.5 Benchmarking of Existing Solutions

The first solution to benchmark is the current manual method of subtype classification by pathologists. According to a quality analysis of reports and referral data for appendiceal neoplasms conducted by surgical oncologists Dr. Konstantinos Votanopoulos and Dr. Christopher Mangieri, 48.4% of pathology reports were misaligned with final pathologic findings. Along with that, low-grade disease was misdiagnosed 36.06% of the time and high-grade disease was misdiagnosed 62.7% of the time. This analysis covers 375 index operative reports over a 27-year period, making it the most comprehensive study of appendiceal cancer diagnosis accuracies currently available (Votanopoulos & Mangieri, 2020). It is important to note that due to the rarity of appendiceal cancer, many of these pathologists will often go their entire careers without seeing a single case. Therefore it makes sense why these numbers are so high. A machine learning or image processing model that can accurately predict the diagnosis of histopathological slides and serve as a complementary tool to pathologists will help to lower this rate of misdiagnosis thus avoiding discordant treatment. The team will use this information regarding misclassification rate to determine an appropriate accuracy threshold for our model. Based on the 48.4% misclassification rate, the model should have an accuracy percentage of at least 51%. Accuracy is a percentage of how many images were predicted correctly in their respective classes out of the total predictions made. It does not deal well with the class imbalance in which the accuracy of one class dominates the total accuracy.

Looking at models that have been developed in relevant biomedical fields can highlight the importance of developing the first known machine learning model for automated analysis of appendiceal cancer. The following model concerns the prediction of CT-images for lung cancer. Researchers from Charles Stuart University in Sydney, Australia, the National Institute of Technology in Haryana, India, and Walden University in the United States used CT-scan image processing and machine learning to detect lung cancer (Makaju et al., 2018). Makaju *et al.* prepared a training database using pre-classified CT-images containing benign and malignant tumors. The researchers then implemented an algorithmic machine and deep learning model to detect and classify cancer nodules for lung cancer classification purposes. Prior to implementing the model, which consisted primarily of convolutional neural networks, researchers performed pre-processing to remove noise, smooth, and segment the images. In addition, the pre-processing allowed for feature extraction as input training data for the classifier. Researchers properly identified patterns and classified cancer nodules as either benign or malignant after sending the pre-processed training data into the classifier. Researchers used a total of 1018 cases from seven academic centers and eight medical imaging companies, then subsequently developed their model in MATLAB using a machine learning toolbox. Their detection model was 92% accurate and their classification model was 86.6% accurate (Makaju et al., 2018). The results of sending in five CT scan images of lung cancer nodes into the classifier can be seen in the table below, where 15 nodules were detected and 13/15 of them were classified accurately. Overall, while this solution focused on detecting and classifying lung cancer and not appendiceal cancer, its methods and apparatuses are still highly relevant to this project.

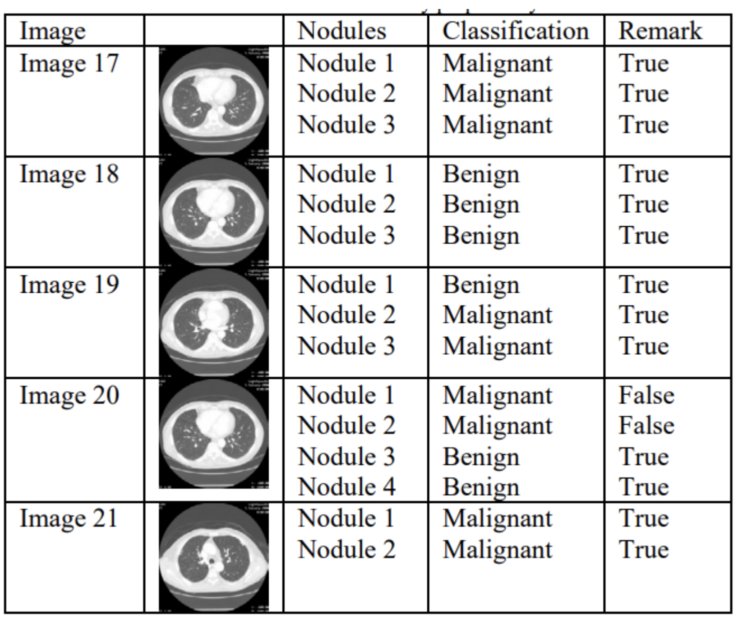


Image		Nodules	Classification	Remark
Image 17		Nodule 1 Nodule 2 Nodule 3	Malignant Malignant Malignant	True True True
Image 18		Nodule 1 Nodule 2 Nodule 3	Benign Benign Benign	True True True
Image 19		Nodule 1 Nodule 2 Nodule 3	Benign Malignant Malignant	True True True
Image 20		Nodule 1 Nodule 2 Nodule 3 Nodule 4	Malignant Malignant Benign Benign	False False True True
Image 21		Nodule 1 Nodule 2	Malignant Malignant	True True

Figure 7. Five images of lung cancer nodes used to test a machine learning algorithm. 13 out of the 15 nodules detected were classified correctly giving it an accuracy of 86.6% (Makaju et al., 2018).

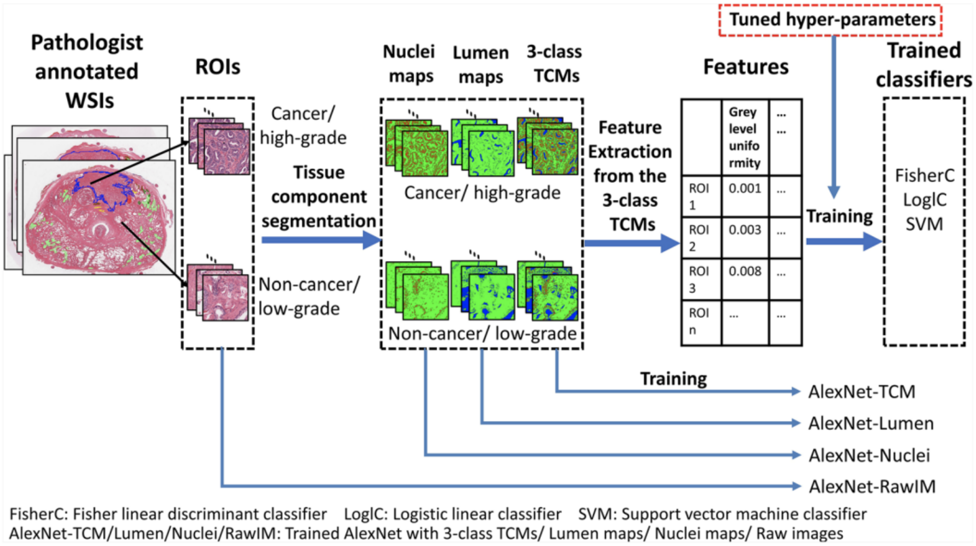
The use of machine learning for medical diagnosis purposes has proven effective in additional studies. Researchers from the Department of Computer Science and Engineering at Annamalai University, India conducted a study to detect and classify diabetic retinopathy using various machine learning techniques (Priya & Aruna, n.d.). Prita *et al.* used a Probabilistic Neural Network (PNN), Bayesian Classification, and Support Vector Machine (SVM) within the study to classify non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR). Prior to implementing the models, the images were pre-processed using techniques to adjust grayscale, adjust image size, and reduce noise. Following image pre-processing, the researchers extracted key features within the images that could provide insight into the presence of diabetic retinopathy (i.e. blood vessels, hemorrhages, etc). The researchers used a total of 250 images for training purposes, and 100 images for model validation. Listed below is a summary of the model outcomes. The team should consider these result metrics when evaluating the accuracy of the appendiceal cancer subtype classification model.

Models	**True Positive**	**True Negative**	**False Positive**	**False Negative**	**Sensitivity**	**Specificity**	**Accuracy**
PNN	180	44	6	20	90	88	89.6
Bayes	190	46	4	10	95	92	94.4
SVM	196	48	2	4	98	96	97.6

Figure 8. This table illustrates the results of the three classifier models used: Probabilistic Neural Network (PNN), Bayesian Classification, and Support Vector Machine (SVM).

Medical machine learning models have proven useful for skin cancer as well. Researchers from Stanford University developed a CNN that can diagnose melanoma with normal images or dermoscopy images. The team used a GoogleNet Inception v3 CNN architecture with similar or better prediction accuracy compared to dermatologists. In the task of three-classes disease partition, the model had an overall accuracy of 72.1%, while the dermatologists attained an average accuracy of around 66% (Esteva et al., 2017). Other machine learning algorithms have used other metrics to predict certain diseases. Jonathan G. Richens, Ciarán M. Lee & Saurabh Johri developed an algorithm that looked into the causal relationship between the symptoms presented and the potential diseases that caused it. The results of the experiment showed that the algorithm could predict better than 75% of the clinicians, making it in the ranks of experts. There is an increase in the performance of predicting rare and very rare diseases, placing it in an even higher rank (Richens et al., 2020).

One final study focused on utilizing various machine learning models to detect and grade cancerous regions on radical prostatectomy (RP) specimens. Researchers analyzed 286 whole slide images from 68 patients; researchers consulted a pathologist to identify key regions of interest on the slides (Han et al., 2020). The slides were annotated by a pathologist, and the researchers identified key regions of interest. Subsequently, the researchers labeled individual pixels within the regions of interest as either nuclei, lumina, and stroma to create tissue component maps. Next, researchers trained seven different machine learning algorithms to determine whether the slide was cancerous or non-cancerous and to identify the grade.


Tuned hyper-parameters
Pathologist annotated WSIs
ROIs
Cancer/ high-grade
Tissue component segmentation
Non-cancer/ low-grade
Nuclei maps
Lumen maps
3-class TCMs
Cancer/ high-grade
Non-cancer/ low-grade
Feature Extraction from the 3-class TCMs
Features
Training
Trained classifiers
FisherC
LogIC
SVM
Training
AlexNet-TCM
AlexNet-Lumen
AlexNet-Nuclei
AlexNet-RawIM
FisherC: Fisher linear discriminant classifier   LogIC: Logistic linear classifier   SVM: Support vector machine classifier
AlexNet-TCM/Lumen/Nuclei/RawIM: Trained AlexNet with 3-class TCMs/ Lumen maps/ Nuclei maps/ Raw images

	Grey level uniformity	… …
ROI 1	0.001	…
ROI 2	0.003	…
ROI 3	0.008	…
ROI n	…	…

Figure 9. The figure above illustrates the process of annotating slides, identifying regions of interest, creating tissue component maps, and training the model based on feature extraction (Han et al., 2020).

There are several important aspects to note about this specific study. Additional pathologists confirmed the annotations made on the whole slide images to ensure accuracy. In addition, the researchers ensured that the training and validation sample sets did not contain slides from the same patient. Training and testing datasets should have no overlapping images to ensure that the model generalizes similarities between subtypes instead of memorizing patterns specific to individual images. This ensures the variability between data and allows the model to learn what images within each subgroup have in common, despite inconsistencies in staining, grade, etc (Han et al., 2020).

2 SYSTEM REQUIREMENTS

2.1 Methods & Analysis: Deriving System Requirements

Image Processing Requirements

Based on literature review, it is clear that image classification through machine learning begins with image pre-processing and analysis. After consulting with technical experts and performing the necessary research, it is clear that the team needs to prioritize obtaining quality whole slide images at a certain magnification from a pathologist. While it is always important to have a large number of images to train the machine learning model, the quality of the images will train the model more accurately and in turn produce more significant predictions. Using images that are annotated, segmented, and cropped rather than just identified as cancerous or non-cancerous will help the algorithm learn to detect cancer nodules, extract features, and classify subtypes of appendiceal cancer. The images must have a certain quality factor to ensure model accuracy. The images will be very large, so the team will need to download the jpeg images on a hard disk, rather than on a cloud-based server. Furthermore, the pathologist, Dr. Coldren, will annotate the images with bounding boxes to identify regions of interest that indicate cancerous regions.

Algorithm Requirements

Many algorithm requirements have been derived from conversations with technical experts, and evidence drawn from benchmarked solutions. Dr. Pauca, a professor of Computer Science at Wake Forest University, recommended that the team use a transfer learning approach to develop a beta-model. Given the time constraint of the project, transfer learning will allow the team to utilize a pre-existing model to begin training the data. Transfer learning works well with smaller datasets and saves an immense amount of time. Dr. Metin Nafi Gurcan, the director of the Center for Biomedical Informatics and the Clinical Image Analysis Lab at Wake Forest Baptist Health, suggested avoiding unsupervised learning, as this approach may yield insignificant outcomes due to misinterpreted data clustering.

The team has researched existing solutions consisting of machine learning algorithms with similar biomedical applications. By observing which features these effective machine learning (ML) image-processing algorithms have in common, the team can determine more system requirements such as statistical significance, reproducibility, and time and space complexity. Furthermore, the team can note how these models were developed, updated, and validated to determine specific methods and apparatuses the team can implement into our own product. For example, in a machine learning algorithm developed by researchers from India, Australia, and the USA, a model obtained an accuracy of 86.6% and a training time of 5.93 seconds (Makaju et al., 2018). These are numbers the team should aim to match, if not improve upon, in the development of their product. From the model used to predict melanoma, it is on average 6% more accurate than dermatologists and achieves over 91% area under the curve. The sensitivity and specificity of the model perform better than almost all dermatologists.

Deriving system requirements related to the pathology and diagnosing of appendiceal cancer will rely heavily on interviews with specialists at Baptist Health as well as extensive background research. Codifying the pathologist's approach to diagnosing appendiceal cancer will allow the team to determine which system requirements are essential in developing a successful algorithm. Understanding what makes an appendiceal cancer diagnosis accurate with the current approach will enable the team to narrow down the essential attributes of our machine learning model. Methods may include recording the process pathologists use to classify appendiceal cancer, as well as the most common errors made that lead to misdiagnosis. Deriving system requirements related to the visual characteristics of appendiceal cancer will largely involve translating information collected from literature reviews and interviews with specialists into overlapping characteristics seen in all successful diagnoses.

2.2 Standards, Codes, Regulations

Machine Learning Algorithms (MLAs) as a tool for medical diagnosis has become increasingly popular in the last few decades (Dey, n.d.). As with any emerging technology to be used in medical diagnosis, certain standards and regulations must be adhered to. In the case of MLAs, there are currently no existing harmonious systems or regulations governing the use of MLAs in medical diagnosis (Johner Institute GmbH, 2020). However, the FDA has introduced modifications that provide pre-specifications for MLAs known as "Software as a Medical Device" or SaMD (FDA, n.d.).

According to the FDA and the International Medical Device Regulators Forum (IMDRF), SaMDs can be placed into categories depending on the state of healthcare condition and the significance of the information provided by SaMD to healthcare decisions (Dey, n.d.). In addition, many modifications can be made to MLAs and they are generally found in three general categories: Performance, Inputs, and Intended use. These categories are not mutually exclusive and help inform what type of MLA the team will develop and implement. The modifications to consider include:

1. Performance modifications but without changing the intended use or input type.
2. Input modifications but without changing the intended use.
3. Modifications to the intended use of the SaMD.

For this design project, there will likely be no modifications to the intended use of the SaMD or to the input type. However, there will be modifications to performance as more data is added to the algorithm to increase accuracy. An important factor in the development and implementation of MLAs in medicine is the Total Product Lifecycle (TPLC), which allows for continuous improvement while maintaining effective safeguards. The FDA proposes the following general principles that aim at balancing the benefits and the risks of a safe and effective SaMD (FDA, n.d.). These principles have been modified to be specific to this design project.

1. The design team should establish clear expectations on good machine learning practices
2. Demonstrate reasonable assurance of effectiveness and establish clear expectations of the SaMD
3. Establish proper usage of the SaMD

When researching with human subjects, the IRB outlines specific training and practices to ensure that human subjects are not abused and that safety measures are taken to protect both the privacy of the human subject and the researchers. According to the IRB, the definition of human subject applies to identifiable private information as well. Therefore, cancer specimens of patients are considered human subjects (UCSF Institutional Review Board, n.d.). The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule determines the conditions under which protected health information can be used for research purposes (Health and Human Services, 2018). These privacy rules apply mostly to federally funded research and some privately funded research. Documentation of IRB approval which should include the IRB or Privacy Board that approved the waiver and a brief description of what type of data is being used, among other things. Any protected health information used should not pose any risk to the privacy of patients by ensuring that there is an adequate plan to de-identify data and prevent improper use of sensitive patient information. Images used in the algorithm should be correctly identified in such a way that prioritizes the privacy of the patient. If any other information is added in addition to the images, they need to be properly de-identified. The project proposer, Dr. Votanopoulos, is securing IRB approval from the WFU IRB for the design team, faculty coaches, TA, and the external technical experts. This approval should be received in 4 to 6 weeks from October 8th, 2020 and as a result, the team will not be able to access any actual appendiceal cancer specimen images until November at the earliest.

The College of American Pathologists (CAP) is the world's leading organization governing the protocols of pathological examination. According to the CAP, there is no uniform grading system for appendiceal carcinoma. Instead, CAP recommends the WHO's criteria, explained in the background research section (College of American Pathologists, 2006). In the case of appendiceal cancer, there is a protocol checklist that applies to all invasive carcinomas of the appendix, excluding carcinoid tumors and related lesions, lymphomas, and sarcomas (College of American Pathologists, 2006). The protocol checklist assesses histologic grade, macroscopic properties, and microscopic properties among others, to determine the classification of an image specimen. With these standards, the system requirements are prioritized and justified in the next section.

2.3 System Requirements Prioritized and Justified

The prioritization and justification of our system requirements are based on background research, benchmarking, and recommendations from external technical experts (ETEs). The foundation for accuracy depends on the team obtaining high resolution, annotated images and non-annotated images that come directly from health institutions that have a tissue bank for appendiceal cancer specimens. These images, used as training data for the detection algorithm or for feature extraction in image processing, will be critical for ensuring a high accuracy in our classification of the subtypes. Another important factor to consider is the quantity of training data. An accurately annotated and extensive dataset is necessary to ensure proper training of a machine learning model or for identifying features in an image processing model. For a machine learning model, if there is not enough data to train the model, the model will not yield an accurate result. On the other hand, with an image processing model, the number of images does not have to be as large. Additionally, as the algorithm is developed, it must have statistical significance that is similar, if not better, than existing solutions. On this same note, the algorithm should have low thresholds for acceptable false positive and negative percentages and should be reproducible under a variety of conditions.

The next phase of prioritization follows the classification of subtypes of appendiceal cancer. The algorithm should be able to quantify and display understandable visual characteristics of these subtypes. If an image processing model is used, it should be able to detect relevant features such as signet ring cells or the nuclei of cells. Finally, the machine learning model or the image processing model should have a user friendly and intuitive interface, such that a majority of people will be able to use it. To improve efficiency, the team will prioritize time and space complexity, making it so that the algorithm has a practical training time and does not take up an unreasonable amount of computer memory and storage.

Rank	System Requirement	Justification	Parameter
1	Accurately Determines Subtype	The most important function of the system is the accuracy of the results it outputs. The current approach has an accuracy of 49% (Votanopoulos & Mangieri, 2020) so in order for the solution the team is developing to be used, the accuracy needs to surpass those of the current approach.	Accuracy: > 51%
2	Perform feature extraction and image preprocessing from images	It is important to extract as many features as possible so that the model would have a wide variety of features for subtype prediction. Usually, the more features, the more accurate the predictions.	Number of high-level extracted features: > 5
3	Model generalizes common patterns within each subtype, regardless of intra-sample variability	The model should learn the general patterns that all images of each subtype have in common including unseen images. This ensures accuracy in training transfer to testing in the real world.	Difference between test and validation set accuracy: < 10%
4	Time and Space Complexity	The model has to be trained with a limited time and resources, since we have a time restriction for the project as well as a budget restriction. Resource cost should be less than our budget and less than a day for training.	Time required for training: < 30-60 min.
5	User Interface (Front End)	The user has to input data to our system, and it is important to be as easy as possible since most users would be medical professionals.	User ratings on prototype ease of use (5 being very easy): > 3 on a scale of 1-5

Figure 10: Solution-Independent System Requirements.

3 CONCEPTUAL DESIGN AND ANALYSIS

3.1 Concept Generation

Prior to developing a morphological matrix to serve as a concept generation tool, it is important to first identify the key functions of the system. The entire system has many functions related to image preprocessing, feature analysis, and subtype classification. The primary functions that have been identified are as follows:

A. Convert pixel-level data into a machine-readable binary matrix
   a. Subfunction: Loop through 1-dimensional RGB byte array within an image file
   b. Subfunction: Store each subsequent byte as an RGB value. Store every three bytes as a pixel. Place these pixels in an [N x M] matrix of pixels, where N and M are the vertical and horizontal resolutions, respectively.

B. Modify the color, exposure, contrast, saturation, and other attributes of training data to simulate real-world differences in H/E stained slides.
   a. Subfunction: Determine range using clustering algorithm for colors, contrast, exposure, etc. These ranges are denoted "*X*". Frequency analysis can also be performed to ensure certain clusters are not disproportionately represented.
   b. Subfunction: Duplicate images and separate them into two sets. Original and transformed.
   c. Subfunction: Map *X* → *Y,* where *Y* is the range of desirable color, contrast, and exposure distributions (in the transformed set).
   d. Subfunction: Rotate, reflect, scale, and zoom images using affine transformation (in the transformed set).

C. Downsample image matrix to matrices that only include desirable features
   a. Subfunction: Extract low-level features from images. Low-level features include vertical and horizontal lines, edges, and corners.
   b. Subfunction: Extract mid-level features from images. Mid-level features include geometric shapes or simple patterns.
   c. Subfunction: Extract high-level features from images. High-level feature extraction is based on low and mid-level feature extraction. It involves the identification of distinguishable objects (such as a water bottle, a stop sign, or a tumor).

D. Categorize feature matrices based on similarity and diagnostic relevance. Output these findings to subtype determination.

E. Interpret the relationship between features and tags. This involves setting the output to the corresponding subtype and reading in the extracted features.

The functions outlined above can be grouped into three main components: *image preprocessing, feature analysis,* and *subtype determination*. The team utilized a system architecture model to separate these three primary components of the system, and to assign each function to a component. The first component of the system is image preprocessing, which involves converting whole slide images into pixel-level data that can be interpreted at a low level. This step is imperative so that image attributes such as color and contrast can be altered to simulate variability in H&E stained images. The second component of the system is feature analysis, which involves the extraction of low-level, mid-level, and high-level features. Once these features have been extracted, individual feature matrices will be pooled together and categorized. The final component will utilize the features extracted during the feature analysis component, and associate these features with a subtype for training purposes.

System Architecture

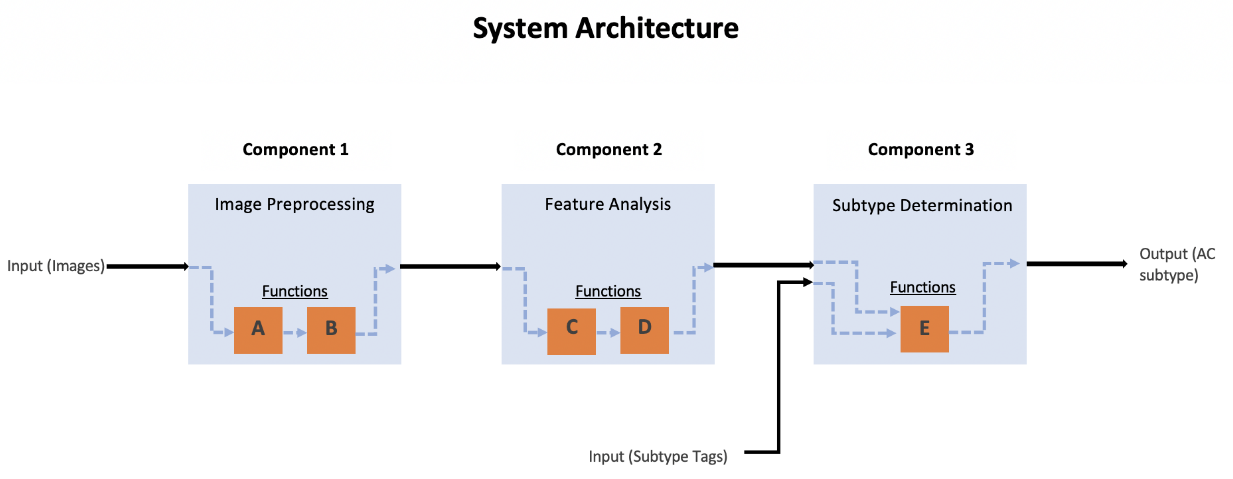


Figure 11: Illustrations of System Architecture.

Morphological Matrix

Solutions to Component 1:

Option I: MATLAB Image Processing Toolbox (custom algorithm)

- Image Processing Toolbox provides functions that can be called from a program. These functions can perform clustering analyses and image augmentations.
- A program will be generated that performs the necessary image preprocessing steps using these function calls. The functions used will depend on the initial set of images provided.

Option II: Pre-existing image processing software (VisioPharm, QuPath, etc.)

- Image processing software such as VisioPharm has built-in image analysis and augmentation programs, which will automatically perform the required adjustments given specific instructions.

Option III: Open-source data augmentation packages

- Various data augmentation packages have been thoroughly developed and made publicly available. Some of these include DeepAugment, CLoDSA, and Codebox/Image Augmentor.

- The packages include methods that scan a directory containing image files and generate new images by performing a specified set of augmentation operations on each file that it finds. This process multiplies the number of training examples that can be used when developing a neural network, and should significantly improve the resulting network's performance, particularly when the number of training examples is relatively small.

Solutions to Component 2:

Option I: Deep Learning CNN

- Uses neural networks to extract relevant features based on their relationship with the training set.
- This CNN may not necessarily extract features that appear pathognomic to AC, but the model has learned they are strongly correlated.
- Convolution involves manipulating input matrices by some function that results in a matrix of extracted features.
- Pooling layers are used to down-sample input matrices to generalize the location of specific features.
- A deep learning CNN uses a combination of many convolutional, pooling, and fully-connected layers to return features that are most correlated with subtype.

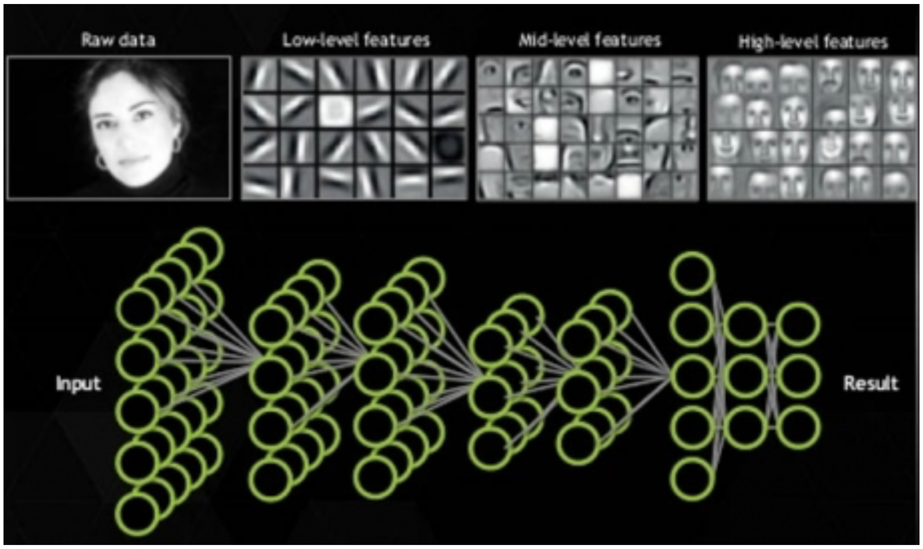


Figure 12: Illustration of how a CNN extracts features (Worrell, 2017).

Option II: Image Processing Algorithm

- Use image processing algorithm containing cell segmentation algorithm to measure cell morphology; Develop a cell map of the sample
- Utilize the algorithm to classify and detect different types of cells that might be indicative of appendiceal cancer

Solutions to Component 3:

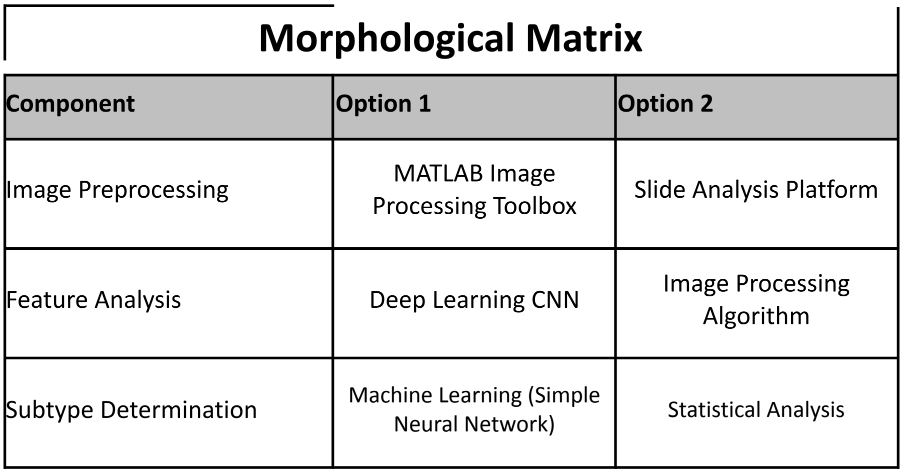
Option I: Machine Learning Model (Simple Neural Network)

- Model will learn to associate visual features with subtype by recognizing patterns in the features

Option II: Statistical Analysis

- Utilize a traditional statistical model to associate certain visual features with subtype
- Model will utilize training data to make a prediction on test images

A morphological matrix was created to outline all potential options for each system component. This facilitates solution generation, as all options within each condensed category are just various implementations that will not change with the solution.



Morphological Matrix

Component	Option 1	Option 2
Image Preprocessing	MATLAB Image Processing Toolbox	Slide Analysis Platform
Feature Analysis	Deep Learning CNN	Image Processing Algorithm
Subtype Determination	Machine Learning (Simple Neural Network)	Statistical Analysis

Figure 13: Morphological Matrix with generalized options. This is used for concept generation.

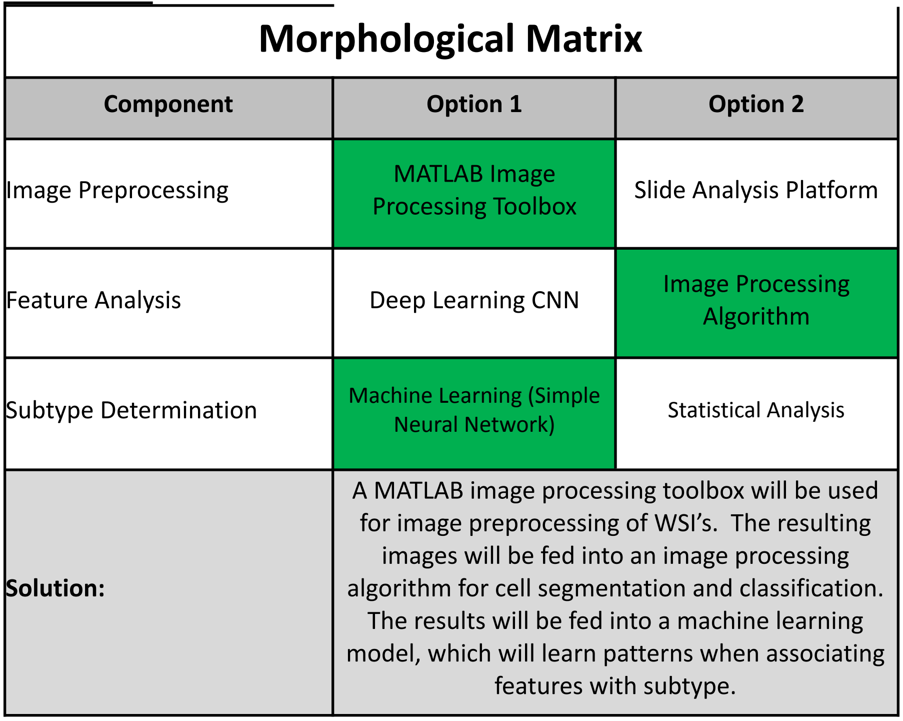
Since there are three components, there are 8 possible combinations. Some combinations do not work, such as using a deep learning CNN for feature extraction and statistical analysis for subtype determination. The three most likely solutions are illustrated below.

Solution 1:

Image Preprocessing: MATLAB Image Processing Toolbox
Feature Analysis: Image Processing Algorithm
Subtype Determination: Statistical analysis

Solution description: A MATLAB image processing toolbox will be used for image preprocessing of WSI's. The resulting images will be fed into an image processing algorithm for cell segmentation and classification. A statistical model will associate extracted features with subtype.

Solution 2:



Morphological Matrix

Component	Option 1	Option 2
Image Preprocessing	[✓ MATLAB Image Processing Toolbox]	Slide Analysis Platform
Feature Analysis	Deep Learning CNN	[✓ Image Processing Algorithm]
Subtype Determination	[✓ Machine Learning (Simple Neural Network)]	Statistical Analysis
Solution:	A MATLAB image processing toolbox will be used for image preprocessing of WSI's. The resulting images will be fed into an image processing algorithm for cell segmentation and classification. The results will be fed into a machine learning model, which will learn patterns when associating features with subtype.	

Figure 14: Illustration of solution 2.

Solution 3:

Image Preprocessing: MATLAB Image Processing Toolbox
Feature Analysis: Deep Learning CNN
Subtype Determination: Machine Learning

Solution description: A MATLAB image processing toolbox will be used for image preprocessing of WSI's. The resulting images will be fed into a deep learning CNN which will extract features that might be associated with AC. The results will be fed into a machine learning model, which will learn patterns when associating features with subtype.

All three concept alternatives will satisfy the two most important system requirements: the ability to determine subtype, and the ability to extract features from images. Solutions 1 and 2 involve using a MATLAB image processing toolbox that has capabilities including edge detection, image orientation, image segmentation, noise reduction, deblurring, and color segmentation. All of these capabilities are necessary for image preprocessing, and ultimately play a fundamental role in preparing for feature analysis. Solutions 1 and 2 also implement an image processing algorithm. The capabilities of an image processing algorithm vary based on the specific algorithm, but many have cell segmentation capabilities for feature extraction. The goal of an image processing algorithm will quantify components of images as shown in Figure 17. The main difference between Solution 1 and Solution 2 is that Solution 1 implements a statistical method for subtype classification, whereas Solution 2 implements a machine learning model for subtype classification.

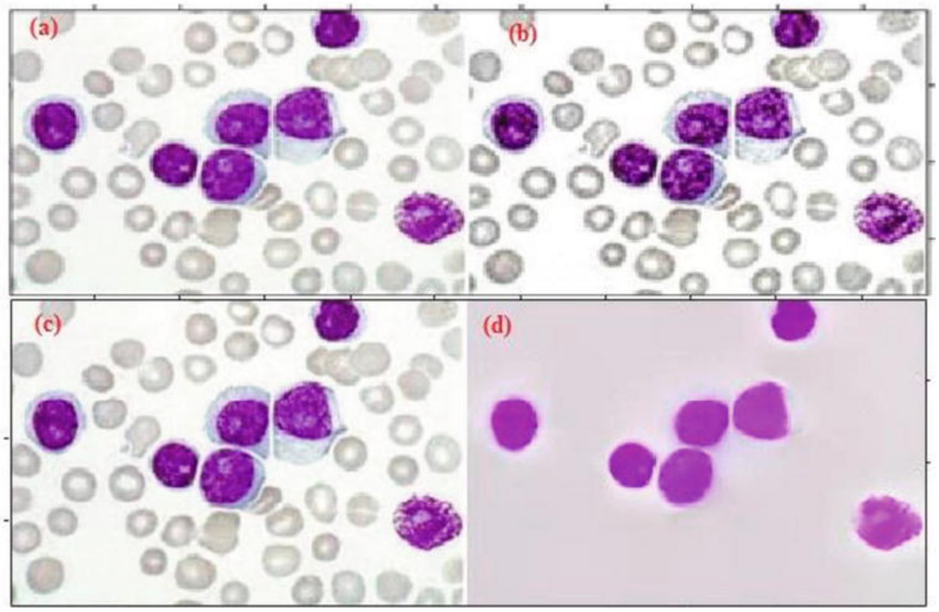


Figure 15: This shows an image processing algorithm that uses contrast to extract leukemia cells from images (Abdallah, 2019).

Solution 3 involves using a MATLAB image processing toolbox for image preprocessing. Unlike Solutions 1 and 2, Solution 3 implements a deep learning model. The team reflected on the benchmarking analysis and determined that a deep learning model would be an effective way to extract features. AC is an understudied region of oncology due to its rarity, and it is challenging to identify the key features indicative of subtype. A deep learning model might have the ability to locate and identify features that might not be obvious or visible to the human eye.

3.2 Concept Evaluation

The first criteria that will be used to evaluate each concept is the amount of training data available, due to the nature of the two methods of extracting features from images, the amount of images available would be the main determining factor of which methods to go. The second criteria would be the time that it takes to develop each method, given the timeline the team has for the project, especially the time it takes to have access to all the images. The third criteria is the accuracy of the selected methods, which is the main measurement for the success of the project outcome. The next criteria is the accessibility and price, what kind of hardware is accessible for us, and how much the team would pay for the hardware. The ease of development of each method would also be considered in company with the timeline of the project. Our final criteria is the ability for our project to be reused for future groups working on the same project since our project is just a small piece of the overarching project being developed by the client Dr. Votanopoulos. Dr. Votanopoulos is spearheading a much larger project that includes research on appendiceal cancer and colorectal cancer, that will continue several years after this capstone project is done. Therefore, it is important that our product can be readily incorporated into other aspects of this much larger project that will continue after this capstone project.

Criteria	Image processing + statistical method *(Solution 1)*	Image processing + machine learning *(Solution 2)*	Deep Learning (CNN) *(Solution 3)(datum)*
	Score	**Score**	**Score**
Training Data	+	+	**0**
Timeline	+	+	**0**
Accuracy	-	-	**0**
Accessibility/ Price	**0**	**0**	**0**
Ease of Development	-	-	**0**
Future Development	**0**	**0**	**0**
Total	**0**	**0**	**0**
Rank	1	1	1

Figure 16: Pugh Chart of the Concepts Generated.

	Amount of Images	**Timeline**	**Accuracy**	**Accessibility/Price**	**Ease of Development**	**Future Development**	**Total**
Weight	6	4	6	3	3	3	
Image processing + statistical method (Solution 1)	7	7	5	7	4	7	**154**
Image processing + machine learning (Solution 2)	7	7	7	7	6	7	**172**
Deep learning (CNN) (Solution 3)	5	5	8	7	7	7	**161**

Figure 17: Weighted Decision Matrix of Concepts Generated.

3.3 Concept Selection

The concept evaluation section features both a pugh chart and a decision matrix for the concepts generated, using the criteria discussed above. The criteria determined in the concept evaluation section was used to develop both the pugh chart and the decision matrix in order to weigh each concept generated and justify the weights given to each concept. In the pugh chart, each criteria is given a row on the left and each concept is given a column on the right. The results of the pugh chart are inconclusive so a decision matrix is used to supplement the concept evaluation and selection. The criteria are: amount of images, timeline, accuracy, accessibility and price, ease of development, and future development. Below, the criteria and concepts are evaluated.

The first concept (Solution 1) is the combination of an image processing model and a statistical analysis method. For this concept, it is not necessary to have large amounts of image data because the image processing model will only need to extract features from the images and then a statistical analysis method will be used to determine the correlation of the features extracted to a particular subtype. This first concept will not require large amounts of training data to train an algorithm to learn the features which is why it received a high score in that category. As a result, the image processing model will be developed using existing toolboxes and libraries in Python or MATLAB, which would work for the timeline given for the duration of this project. In addition, these libraries and toolboxes are easily accessible because python is open source and Wake Forest has purchased a MATLAB license. With regards to future implementation, this first concept will not require any further development. The accuracy can be determined using equation one below and has been explained earlier in the system requirements.

The second concept (Solution 2) is the combination of the image processing software and a machine learning algorithm. This concept will also not require large amounts of training data because the algorithm will not be learning the features of each individual subtypes but rather classify the images according to subtypes based on the features extracted from the image processing software. As a result, this concept gets a high score in the amount of data category. Furthermore, this concept also gets a high score in the timeline and accuracy categories because the use of the combination of an image processing model and image classification algorithm will not require the development of a machine learning algorithm that will have to learn the features of each subtype and require large amounts of training data. The image processing model will extract the features and the model will simply classify them rather than the model doing both tasks. This second concept will also be relatively easy to develop and overall accessible because there are libraries and toolboxes to support image processing in Python and MATLAB, as well as existing convolutional neural networks that can be developed and tweaked for the purposes of this project. Finally, this concept can more readily be adapted for future applications because the feature extraction task and image classification task are kept separate, allowing for the algorithm or the model to be tweaked independently of one another. This is the concept that had the highest score in the weighted decision matrix.

The third concept (Solution 3) is a deep learning neural network that will take a whole slide image, perform segmentation and other image processing sequences, and then classify the image according to a subtype. This concept requires large amounts of training data to allow the algorithm to be able to decide what features it finds to be most important for each subtype, so it scores low in this category. In addition, the development and implementation of this concept will be extremely time and labor-intensive, which might be challenging given the timeline of one academic year. On the other hand, this concept will be the most accurate because it will be trained with images that have an associated tag that is known to be true. This gives us a ground truth with which the model will use to learn the relevant features. Along with this, the deep learning model has the potential to identify features and characteristics that are not currently standardized by pathologists. However, this does run the risk of identifying and correlating irrelevant characteristics, hence why the team will need a large data set. Overall, the deep learning concept will be implemented and developed under the conditions that the team has substantial data for each subtype, as well as data for healthy slides.

The pugh chart and weighted decision matrix indicate the best solution depends on the amount of data the team receives. Based on preliminary estimates from the client, the team has agreed that Solution 2 is the best option as it requires fewer images. However, this could possibly change as Solution 3 could be advantageous if more images are received than expected. After meeting with the ETEs during the PDR presentation review and gathering feedback on the solutions presented, everyone agreed that Solution 2 is practical, feasible, and meets the timeline restrictions. Furthermore, the entire group has access to MATLAB, and many Python packages and libraries are open-source. Therefore, the team will go into the embodiment design phase with Solution 2 as the concept to be fully developed and implemented.

Below are some of the equations guiding the image processing models and accuracy analysis.

$$\frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

Equation (1) is the calculation for the accuracy which simply sums the true positives and negatives and divides by the total sum of the true positives, true negatives, false positives, and false negatives (*What is Confusion Matrix and Advanced Classification Metrics?*, 2019).

$$LoG(x,y) = -\frac{1}{\pi\sigma^4}\left[1 - \frac{x^2 + y^2}{2\sigma^2}\right] e^{\frac{-x^2+y^2}{2\sigma^2}} \qquad (2)$$

Equation (2) is the Laplacian of Gaussian (LoG) operation. The LoG is a two-step process that is used to in detecting areas of rapid change i.e. edges in image processing models. The first step is the Gaussian filter that is used to smooth the image and the second step is the Laplacian derivative filter that is used to find areas of rapid change i.e. edges in images. The first step is necessary because derivative filters are sensitive to noise (Mathys, 2001). These equations have already been built into the toolboxes and libraries of the image processing models in MATLAB and in Python.

4 MANAGEMENT

4.1 Team Organization

The team consists of two product managers, a project manager, a practice manager, and a team manager. Although many of the same functions are shared between these roles, there are independent responsibilities each team member must take on. Ultimately, it is a cohesive effort within and between each individual to make sure system requirements are met, plans are followed, and communication is efficient.

Product Managers:

- Focus on the system/product being designed and ensuring the system requirements are being met each step of the way.
- Responsibilities include product/design requirements and specifications, managing product backlog, prototyping of design, and communication with client and technical expert(s).
  - Gabrielle Prichard
    - Gabrielle will focus mainly on product requirements and prototyping of design. Her coding knowledge and internship experience in machine learning reinforce her qualifications for this role.
  - Bryan Bennett
    - Bryan will focus primarily on the histopathologic component of the project, namely developing a visual understanding of the subtypes and ensuring the model is identifying the correct features. Bryan will also communicate with the client and most technical experts. His knowledge of computer science and internship experience in NLP will enable him to bridge the medical and ML components of the project.

Project Manager:

- Considers the team's processes in meeting deliverables and project milestones
- Responsibilities include resource planning, scheduling, documentation, and communication with capstone coordinators.
  - Margaret Nyamadi
    - Margaret will utilize her knowledge of Matlab, Python, and product design in ensuring the proof of concept developed will meet the needs of the client.

Practice Manager

- Focuses on optimizing the team's tools, technologies, and resources
- Responsibilities include ensuring best practices, identifying cost budgets, and managing and guaranteeing product success.
  - Hao Tong

        - Hao specializes in deep learning algorithms and will avail the team of his knowledge and skills in developing the system architecture of the algorithm.

Team Manager

- Optimizes performance of the team and ensures the team has the right expertise to tackle all aspects of the project.
- Responsibilities include communicating with the Faculty Coach and managing team effectiveness, conflict, and team productivity.
    - Ethan Cooley
        - Ethan will employ his skills in Python, JavaScript, data science, and user interface design and development for the successful completion of the project.

4.2 Project Organization

At a high level, the project will occur in 4 phases. Phase 1 is the derivation of the system requirements known as the discovery design phase. In this phase, the design team is conducting multiple interviews with the client, technical experts in image processing and machine learning, and with the engineering librarian. The team is also engaging in research to understand appendiceal cancer and its subtypes, understand requirements for image processing, understanding creating and developing neural networks, and learning how to adopt and modify existing algorithms to our project's needs. The deliverable for this phase is the SRR and the design presentation.

Phase 2 is the conceptual design and occurs in two steps: concept generation and concept selection. Concept generation entails ideation of multiple solutions using techniques such as morphological matrices, brainstorming, reverse engineering, and functional analysis in addition to solution benchmarking, that helped inform the concepts generated. The concept generation techniques used in this design project are functional analysis and morphological matrices. The team researched image processing in MATLAB, Python, as well as CNNs as a medical diagnosis tool from the UNET library. Afterwards, a decision matrix was developed using criteria that was determined from the concept evaluation step and the system requirements. The deliverable for this phase is the Preliminary Design Review (PDR) which details the concept selection process and the justification for which concept was selected to move forward into the embodiment design phase.

Phase 3 is the embodiment design phase and it is centered on the physical form of the system. For software systems, the embodiment design will focus on mapping and developing the system architecture of the concept selected, as well as debugging and training the algorithm using appendiceal cancer specimen images. The four Fs of the embodiment design phase are: Function, Fit, Form, and Finish. These Fs are systematized into 5 essential elements i.e. Architecture, Configuration, Modeling, Prototyping, and Testing. The features of various appendiceal cancer subtypes will be extracted using MATLAB image processing, and then a machine learning algorithm will correlate the features to subtypes using a deep learning convolutional neural network. In addition, various artifacts will be created along the way to display the group's progress such as pseudocode, flow charts, use cases, class diagrams, and source code. The group will employ model-driven development by utilizing an integrated set of diagrams known as Unified Modeling Language (UML) to visualize, construct, and document the artifacts mentioned.

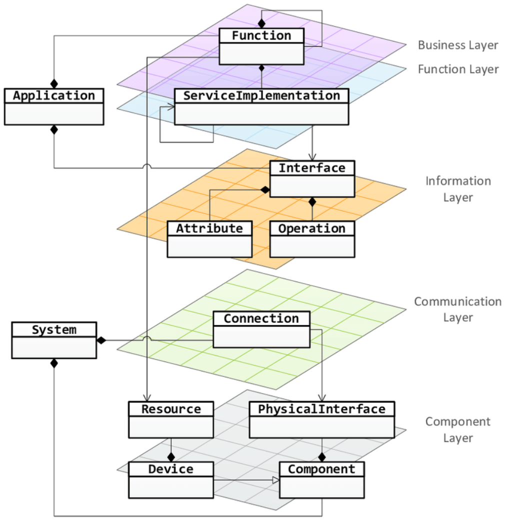


Figure 18: Example implementation of Unified Modeling Language (UML) (Andrén et al., 2017)

Lastly, the algorithm will go through stages of debugging and testing to assess quality of performance, perform risk assessments, and finalize a budget. The deliverables for this section are the Critical Design Review (CDR) and the Test Readiness Review (TRR). The former will certify that the prototype(s) have met all the requirements and the latter will review the design team's plan for formal testing.

Phase 4, the last phase, will focus on optimization and performance verification of the algorithm, further assessment of failure modes, and a detailed, systematic maintainability of the algorithm. The deliverable for this phase is the Production Readiness Review (PRR). The PRR will entail the design with proper documentation and ensure the readiness of the design for manufacture by meeting all requirements. The project timeline is shown in the Gantt chart and Work Breakdown Structure in the appendix.

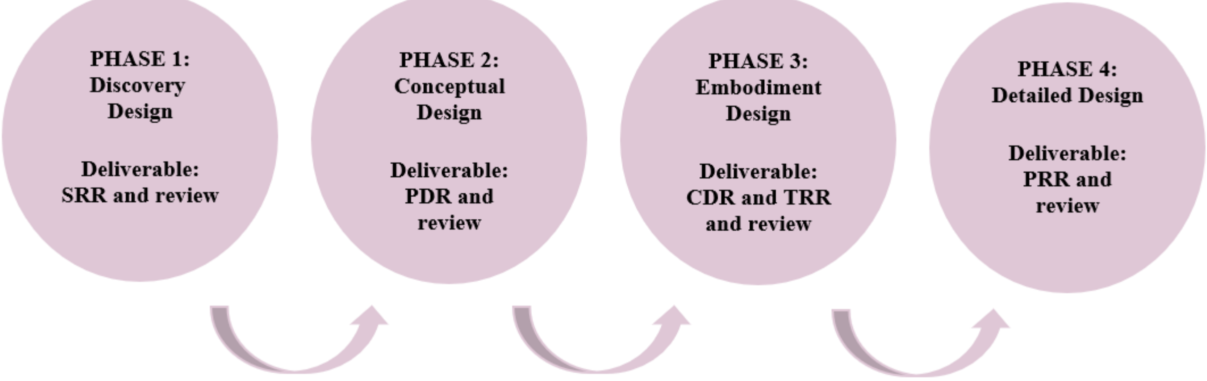


Figure 19: Overview of the Phases of the Design Process

4.3 Preliminary Plan for Next Design Review (Critical Design Review)

The next review is the critical design review, this will involve developing a proof of concept model that can accurately classify grades of cancer using histopathological image data sets that are publicly available. The team will be using publicly available images until IRB approval is obtained and access to the images is granted. The model will be developed using a combination of existing open-source image augmentation software, processing algorithms, object detection networks, and classification algorithms. The team will compare different platforms and perform a concept selection process to choose the platform that best matches the system requirements based on performance metrics such as accuracy in classification. The team will come up with the best system architecture and a prototype of the design for determining the baseline performance of the system. The team will perform another iteration of the risk assessment to reevaluate how the risks have changed and what other risks have developed since the preliminary design review.

4.4 Risk Analysis, Challenges, and Roadmap to Success

There are several methods that can be used in risk assessment of design projects such as fault-tree analysis, failure modes and effects analysis, a risk breakdown structure, and risk analysis and likelihood matrix. The risks have been broken down into four general categories: organizational and/or management risks, external risks, technical risks, and other risks. Organizational and management risks assess internal factors such as team conflict and workload distribution, external risks assess factors outside of the team's control that can inhibit the progression of the project such as how quickly whole slide images can be procured from collaborating institutions and sent to the team. Technical risks assess factors such as the accuracy of the model and integration of image processing software and the image classification algorithm.

At this phase of the project, the technical and external risks are the biggest factors affecting the development of the project. The team will be receiving whole slide images from collaborating institutions that have been contacted by the client, after IRB approval is submitted to both the Wake Health IRB and the IRBs of the other institutions. On the technical aspect, there is some ambiguity with how the image processing model will be integrated with the image classification model. The image processing model will likely be developed using MATLAB but the classification algorithm is still up for debate. As a result, the team will need to determine the best way to handle communication between both components and transference of image data in a way that preserves the quality of the images.

Other risks that have been identified for the duration of the project are the limited timeline that the team has, and the limiting amount of images. The complete risk analysis matrix for this design project and the chart used are in the appendix (Figure 25).

4.5 Changes Implemented Since PDR Initial Draft

- Problem Statement
  - Condensed problem statement
- Mission Statement
  - Expanded impacts to include other groups of people
- Background and Benchmarking
  - Added background research to support system requirements. Much of the information from the previous system requirements has been moved here and incorporated into paragraph form to produce flow
  - Updated benchmarking to include sections from background research
- System Requirements
  - Condensed the justification of each requirement.
- Concept Evaluation
  - Concept Generation
    - Removed excess morph matrices and consolidated solutions to include just relevant concepts
  - Concept Evaluation
    - Elaborated on the criteria for future development of the project
    - Explained the overarching project client is working on and how our capstone project fits into client's larger project
    - Updated pugh charts and decisions matrices to reflect feedback received from both ETEs and course instructors
  - Concept Selection
    - Selected and elaborated on Solution 2
    - Expanded the justification for our decision to choose Solution 2
- Project Organization
  - Elaborated phase 3 section
  - Added in software artifacts and UML diagram

  - Added high quality visual, google docs previously compressed the visual and there was a loss in quality. Took a screenshot instead which fixed this problem.
- Bibliography
  - Updated bibliography with new citations

5 BIBLIOGRAPHY

Abdallah, Y. (2019, June 24). *Research in Medical Imaging Using Image Processing Techniques*. Medical Imaging - Principles and Applications. Retrieved 11 20, 2020, from https://www.intechopen.com/books/medical-imaging-principles-and-applications/research-in-medical-imaging-using-image-processing-techniques

Adobe. (2019, October 7). *The 4 Golden Rules of UI Design: Adobe XD Ideas*. Ideas. Retrieved October 23, 2020, from https://xd.adobe.com/ideas/process/ui-design/4-golden-rules-ui-design/

Andrén, F. P., Strasser, T. I., & Kastner, W. (2017, March). *Engineering Smart Grids: Applying Model-Driven Development from Use Case Design to Deployment*. ResearchGate. https://www.researchgate.net/publication/315117587_Engineering_Smart_Grids_Applying_Model-Driven_Development_from_Use_Case_Design_to_Deployment

Brownlee, J. (2019, 08 08). *Confidence Intervals for Machine Learning*. machine learning mastery. https://machinelearningmastery.com/confidence-intervals-for-machine-learning/

Brownlee, J. (2020, 08 22). *How to Use ROC Curves and Precision-Recall Curves for Classification in Python*. Machine learning mastery. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/

College of American Pathologists. (2006, July 1). *Surgical Pathology Cancer Case Summary (Checklist)*. CAP.org. Retrieved October 23, 2020, from https://webapps.cap.org/apps/docs/committees/cancer/cancer_protocols/2006/appendix06_pw.pdf

Confusion matrix. (n.d.). sk learn. https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html

Convolutional Neural Network. (2019, 2 24). towards data science. https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529

Convolutional Neural Network: How to Build One in Keras &PyTorch. (n.d.). MIssingLInk.ai. Retrieved October 23, 2020, from https://missinglink.ai/guides/neural-network-concepts/convolutional-neural-network-build-one-keras-pytorch/

Dewis, R., & Gribbin, J. (2009, February). Breast Cancer: Diagnosis and Treatment: An Assessment of Need. *National Collaborating Centre for Cancer*, *3*(Epidemiology). https://www.ncbi.nlm.nih.gov/books/NBK61914/

Dey, A. (n.d.). Machine Learning Algorithms: A Review. *7*(3). https://ijcsit.com/docs/Volume%207/vol7issue3/ijcsit2016070332.pdf

Esteva, A., Kuprel, B., & Novoa, R. A. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*, 115–118. https://www.nature.com/articles/nature21056

FDA. (n.d.). *Proposed Regulatory Framework For Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software As A Medical Device (SaMD)*. U.S. Food and Drug Administration. Retrieved October 23, 2020, from https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf

Han, W., Johnson, C., Gaed, M., Gómez, J. A., Moussa, M., Chin, J. L., Pautler, S., Bauman, G. S., & Ward, A. D. (2020). Histologic tissue components provide major cues for machine learning-based prostate cancer detection and grading on prostatectomy specimens. *Scientific Reports*, *10*(1). 10.1038/s41598-020-66849-2

Haque, I. R. I., & Neubert, J. (2020, January 26). Deep learning approaches to biomedical image segmentation. *Informatics in Medicine Unlocked*. Retrieved October 23, 2020, from https://www.sciencedirect.com/science/article/pii/S235291481930214X#!

Health and Human Services. (2018, June 13). *Research*. HHS.gov. Retrieved October 23, 2020, from https://www.hhs.gov/hipaa/for-professionals/special-topics/research/index.html

Johner Institute GmbH. (2020, October 8). *Regulatory Requirements for Medical Devices with Machine Learning*. Johner Institut GmbH. Retrieved October 23, 2020, from https://www.johner-institute.com/articles/regulatory-affairs/and-more/regulatory-requirements-for-medical-devices-with-machine-learning/

Justus, D., Brennan, J., Bonner, S., & McGough, A. S. (2019, January 24). *Predicting the Computational Cost of Deep Learning Models*. IEEE XPlore. Retrieved October 23, 2020, from https://ieeexplore.ieee.org/document/8622396/authors#authors

Kelly, K. J. (2015, December). Management of Appendix Cancer. *Clin Colon Rectal Surg*, *28*(4), 247-255. 10.1055/s-0035-1564433

The leading training data platform for data labeling. (n.d.). Labelbox. Retrieved 23, October, from https://labelbox.com/

Leica Biosystems. (2019, August 14). *H&E Staining Overview: A Guide to Best Practices*. Leica Biosystems. Retrieved October 23, 2020, from https://www.leicabiosystems.com/knowledge-pathway/he-staining-overview-a-guide-to-best-practices/

Leonards, L. M., Pahwa, A., Patel, M. K., Petersen, J., Nguyen, M. J., & Jude, C. M. (2017). Neoplasms of the Appendix: Pictorial Review with Clinical and Pathologic Correlation. *RadioGraphics*, *37*(4), 1059. 10.1148/rg.2017160150

Makaju, S., Prasad, P.w.c., Alsadoon, A., Singh, A.k., & Elchouemi, A. (2018). Lung Cancer Detection using CT Scan Images. *Procedia Computer Science*, *125*, 107-114. ScienceDirect. 10.1016/j.procs.2017.12.016

Mathys, D. (2001, February 14). *LoG Filters*. LoG Filters. Retrieved November 21, 2020, from https://academic.mu.edu/phys/matthysd/web226/Lab02.htm#:~:text=Laplacian%20filters%20are%20derivative%20filters,of%20Gaussian%20(LoG)%20operation.

McCusker, M. E., Coté, T. R., Clegg, L. X., & Sobin, L. H. (2002, June 15). Primary malignant neoplasms of the appendix: a population-based study from the surveillance, epidemiology and end-results program, 1973-1998. *Cancer*, *94*(12), 3307-3312. 10.1002/cncr.10589

Mohajon, J. (2020, 05 26). *Confusion Matrix for Your Multi-Class Machine Learning Model*. towards data science. https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

Nagtegaal, I. D., Odze, R. D., Klimstra, D., Paradis, V., Rugge, M., Schirmacher, P., Washington, K. M., Carneiro, F., Cree, I. A., & The WHO Classification of Tumours Editorial Board. (2020). The 2019 WHO classification of tumours of the digestive system. In *Histopathology* (4th ed., pp. 182-188). The World Health Organization.

Narkhede, S. (2018, 06 28). *Understanding AUC - ROC Curve*. towards data science. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

National Cancer Institute. (n.d.). *NCI Dictionary of Cancer Terms*. National Cancer Institute. Retrieved October 23, 2020, from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/low-grade

National Cancer Institute. (n.d.). *NCI Dictionary of Cancer Terms*. National Cancer Institute. Retrieved October 23, 2020, from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/high-grade

National Cancer Institute (NIH). (n.d.). *NCI Dictionaries*. National Cancer Institute. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/dysplasia

Neural Networks for Image Recognition: Methods, Best Practices, Applications. (n.d.). MissingLink.ai. Retrieved October 23, 2020, from https://missinglink.ai/guides/computer-vision/neural-networks-image-recognition-methods-best-practices-applications/

Parmar, R. (2018, September 02). *Common Loss functions in machine learning*. Towards Data Science. Retrieved December 04, 2020, from https://towardsdatascience.com/common-loss-functions-in-machine-learning-46af0ffc4d23

Pretrained Deep Neural Networks. (n.d.). Mathworks. https://www.mathworks.com/help/deeplearning/ug/pretrained-convolutional-neural-networks.html

Priya, R., & Aruna, P. (n.d.). Diagnosis of diabetic retinopathy using machine learning techniques. *ICTACT Journal on soft computing*, *4*(4), 563-575.

Richens, J. G., Lee, C. M., & Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun*, *11*(3923). https://www.nature.com/articles/s41467-020-17419-7

Shah, T. (2017, 12 6). *About Train, Validation and Test Sets in Machine Learning*. towards data science. https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., & University of Toronto. (2014, June). *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. Journal of Machine Learning Research. https://www.cs.toronto.edu/~hinton/absps/JMLRdropout.pdf

UCSF Institutional Review Board. (n.d.). *Medical Record Review*. Medical Record Review | UCSF Institutional Review Board. Retrieved October 23, 2020, from https://irb.ucsf.edu/medical-record-review

Votanopoulos, K. I., & Mangieri, C. W. (2020, November 12). *Quality analysis of operative reports and referral data for appendiceal neoplasms with peritoneal dissemination*.

ScienceDirect. Retrieved November 20, 2020, from https://www.sciencedirect.com/science/article/abs/pii/S0039606020306772

Wang, P., Fan, E., & Wang, P. (2020, August 1). *Comparative Analysis of Image Classification Algorithms Based on Traditional Machine Learning and Deep Learning*. ScienceDirect. Retrieved October 23, 2020, from https://www.sciencedirect.com/science/article/pii/S0167865520302981

What is Confusion Matrix and Advanced Classification Metrics? (2019, 4 29). data science and machine learning. https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html

Worrell, J. (2017, May 10). *Nvidia to train 100,000 developers on deep learning AI*. Fudzilla. https://www.fudzilla.com/news/43610-nvidia-to-train-100-000-developers-on-deep-learning-ai-in-2017

6 APPENDIX

6.1 Appendix 1: Graphs, Figures and Charts

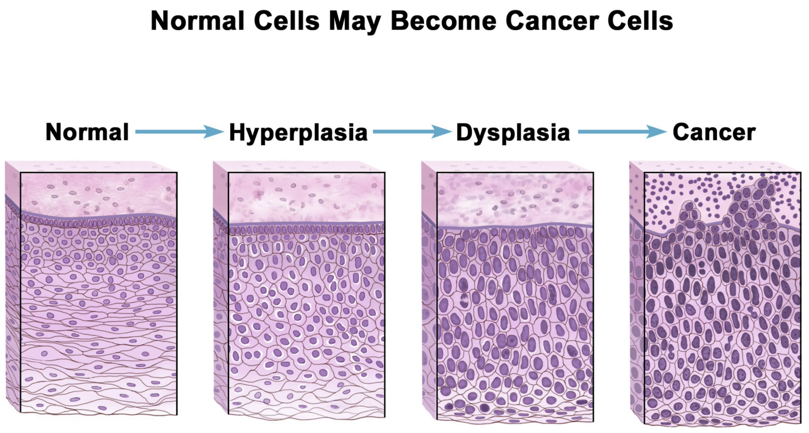
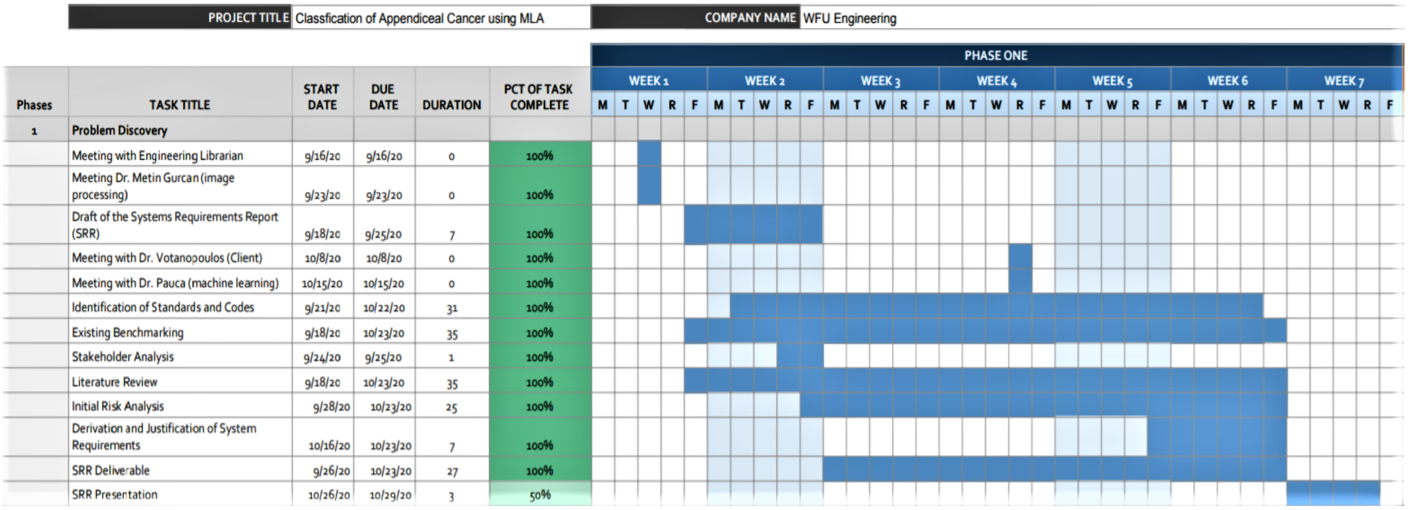


Figure 20: This illustrates the visual difference between normal tissue, dysplasia, and cancer. Notice cancer looks very similar to dysplasia, but has begun to spread into other tissues. (National Cancer Institute (NIH), n.d.)

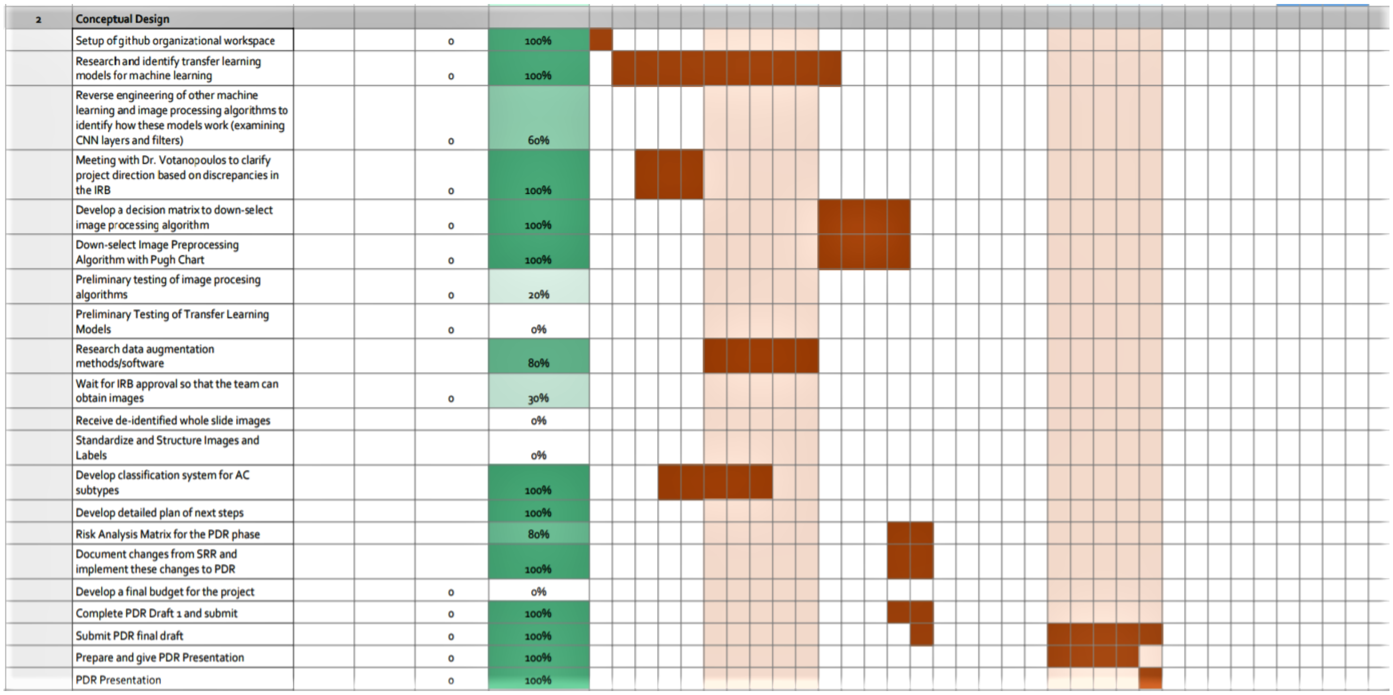


PROJECT TITLE Classfication of Appendiceal Cancer using MLA

COMPANY NAME WFU Engineering

Phases	TASK TITLE	START DATE	DUE DATE	DURATION	PCT OF TASK COMPLETE
1	**Problem Discovery**				
	Meeting with Engineering Librarian	9/16/20	9/16/20	0	100%
	Meeting Dr. Metin Gurcan (image processing)	9/23/20	9/23/20	0	100%
	Draft of the Systems Requirements Report (SRR)	9/18/20	9/25/20	7	100%
	Meeting with Dr. Votanopoulos (Client)	10/8/20	10/8/20	0	100%
	Meeting with Dr. Pauca (machine learning)	10/15/20	10/15/20	0	100%
	Identification of Standards and Codes	9/21/20	10/22/20	31	100%
	Existing Benchmarking	9/18/20	10/23/20	35	100%
	Stakeholder Analysis	9/24/20	9/25/20	1	100%
	Literature Review	9/18/20	10/23/20	35	100%
	Initial Risk Analysis	9/28/20	10/23/20	25	100%
	Derivation and Justification of System Requirements	10/16/20	10/23/20	7	100%
	SRR Deliverable	9/26/20	10/23/20	27	100%
	SRR Presentation	10/26/20	10/29/20	3	50%

Figure 21: This Gantt chart shows the current timelines for the project from phase 2.



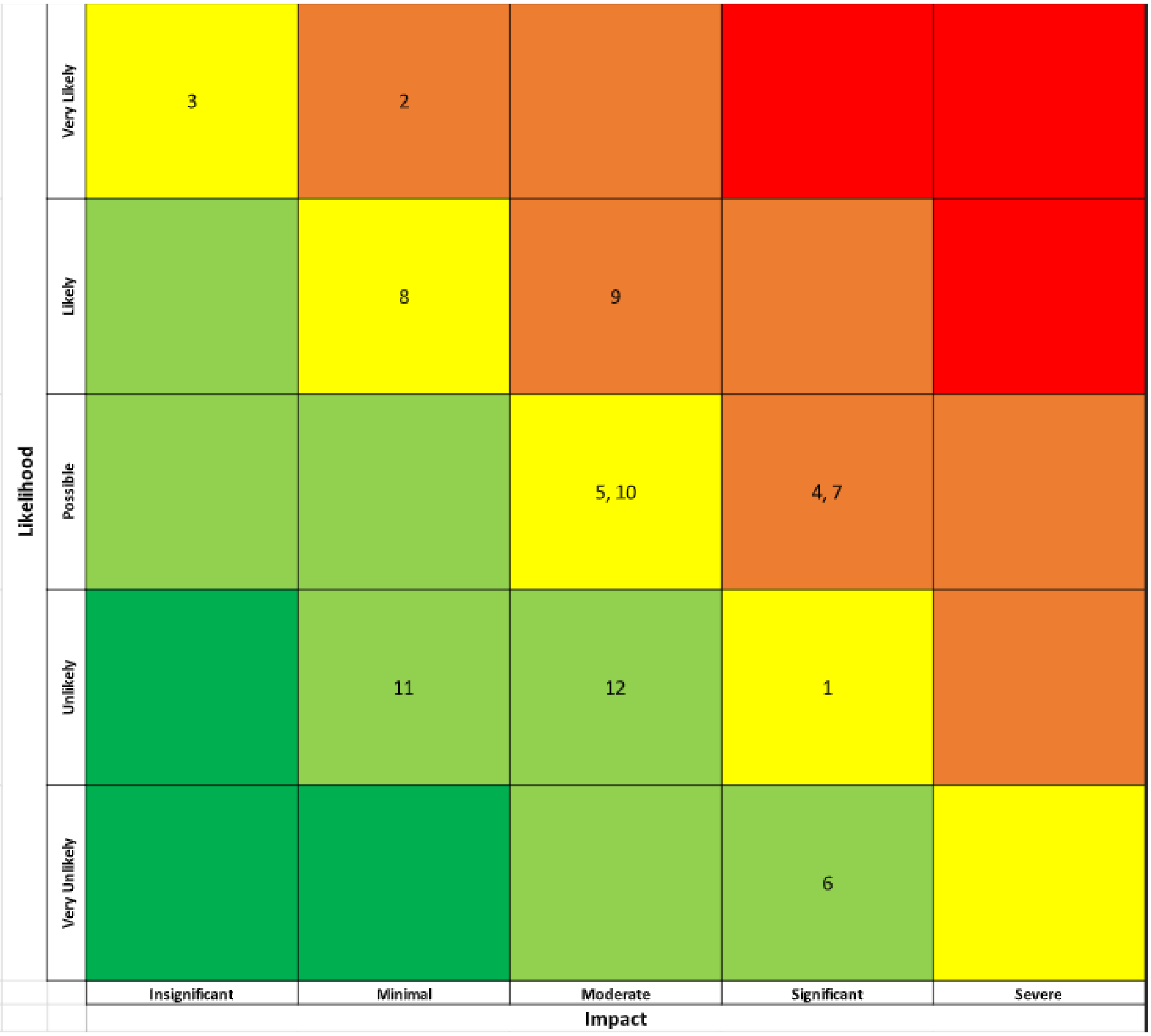
Phases	TASK TITLE	START DATE	DUE DATE	DURATION	PCT OF TASK COMPLETE
2	**Conceptual Design**				
	Setup of github organizational workspace			0	100%
	Research and identify transfer learning models for machine learning			0	100%
	Reverse engineering of other machine learning and image processing algorithms to identify how these models work (examining CNN layers and filters)			0	60%
	Meeting with Dr. Votanopoulos to clarify project direction based on discrepancies in the IRB			0	100%
	Develop a decision matrix to down-select image processing algorithm			0	100%
	Down-select Image Preprocessing Algorithm with Pugh Chart			0	100%
	Preliminary testing of image procesing algorithms			0	20%
	Preliminary Testing of Transfer Learning Models			0	0%
	Research data augmentation methods/software				80%
	Wait for IRB approval so that the team can obtain images			0	30%
	Receive de-identified whole slide images				0%
	Standardize and Structure Images and Labels				0%
	Develop classification system for AC subtypes				100%
	Develop detailed plan of next steps				100%
	Risk Analysis Matrix for the PDR phase				80%
	Document changes from SRR and implement these changes to PDR				100%
	Develop a final budget for the project			0	0%
	Complete PDR Draft 1 and submit			0	100%
	Submit PDR final draft			0	100%
	Prepare and give PDR Presentation			0	100%
	PDR Presentation			0	100%

Figure 22: This Gantt chart shows the current timelines for the project from phase 2.

Work Breakdown Structure	
Phase 1	Discovery Phase ➢ Literature Review

	➢ Meeting with Engineering Librarian ➢ Meeting with Dr. Metin Gurcan (image processing) ➢ Draft of the Systems Requirements Report (SRR) ➢ Meeting with Dr. Votanopoulos (Client) ➢ Meeting with Dr. Victor Pauca (machine learning) ➢ Meeting with Dr. Coldren classifying Appendiceal Cancer ➢ Identification of Standards and Codes for MLA in medical diagnoses ➢ Existing Benchmarking of MLAs for medical application ➢ Stakeholder Analysis ➢ Initial Risk Analysis ➢ Draft Budget ➢ Derivation and Justification of System requirements Deliverable: SRR
Phase 2	Conceptual Design ➢ Setup of github organizational workspace ➢ Research and identify transfer learning models ➢ Reverse engineering of other machine learning and image processing models ➢ Clarify Project direction based on IRB ➢ Develop classification system for AC subtypes ➢ Generate concepts using functional analysis and reverse engineering ➢ Develop decision matrix and pug chart for concepts generated ➢ Preliminary testing of image processing models and classification algorithms ➢ Recieve de-identified whole slide images ➢ Standardize and Structure Images and Labels ➢ Risk Analysis Matrix for PDR phase ➢ Document changes from SRR in PDR based on feedback Deliverable: PDR
Phase 3	Embodiment Design ➢ Map system architecture of algorithm selected ➢ Prototyping/debugging of algorithm ➢ Training the algorithm using multiple orientation of images ➢ Bill of materials and obtaining of materials ➢ Performance assessment of algorithm ➢ Risk Assessment Deliverable: CDR and TRR
Phase 4	Detailed Design ➢ Performance assessment and verification of results ➢ Optimization of algorithm ➢ Analysis of failure mode(s) ➢ A systematic approach to maintainability of algorithm ➢ Formal Testing of the algorithm Deliverable: PRR

Figure 23: The Work Breakdown Structure (WBS) shows the overall direction of the project.



	Insignificant	Minimal	Moderate	Significant	Severe
Very Likely	3	2			
Likely		8	9		
Possible			5, 10	4, 7	
Unlikely		11	12	1	
Very Unlikely				6	

Figure 24: Risk Analysis Chart

Category of Risk	Risk Label (likelihood)	Impact	Risk Description	Proposed Mitigation Plan
Organizational and Management	Unlikely	Moderate	Different sections of written assignments are difficult to make "cohesive."	The team will meet before a final writing assignment is due to proof-read the entire document. It is important to ensure that the writing remains formal and cohesive.
	Unlikely	Significant	Divergence from long-term schedule/objectives	Set short term goals/schedules that aligns with long term schedule/objectives to better monitor our process
External	Possible	Significant	Use of product may make sense to the team, but once it is supplied to clients the	Spend a substantial amount of time on UI and front-end transparency

			operation of our product is not simple	
	Possible	Significant	Product succeeds in accurately classifying appendiceal cancer, but does not provide additional value to client beyond the capabilities of a pathologist	Continually communicate with Dr. Votanopoulos to ensure product is aligned with his vision and system requirements
	Possible	Moderate	Doctors and Pathologists cannot establish trust in our product	Prove model accuracy through experimental testing. Justify results if necessary.
	Possible	Moderate	Data the team receives from Wake Health does not provide sufficient information to train the model (such as information about grade, subtype, etc)	Proper communication with pathologist and Dr. Votanopoulos about what information the team needs to train the model should prevent or remedy this problem
	Possible	Significant	Image data set is does not represent all stages and subtypes, so model becomes biased	Collect data from the collaborating institutions with all subtypes represented.
	Possible	Significant	Certain time delays might hinder the progress of the project (i.e. processing images will take a significant amount of time, marking up images will take a significant amount of time)	The team will plan ahead to anticipate time delays;The design team will regularly schedule meetings with Dr. Coldren and pathologists so that the team is aware of certain timelines. The team will also try to develop model using images that are not annotated if that process takes too long.
Technical	Possible	Significant	Model is not statistically significant and therefore cannot be used in practice by pathologists.	In order to mitigate this risk, the team will start developing a model with a few images for low grade and high grade appendiceal cancer. As the team continue with the project,the design team will add more images to improve the model accuracy and improve the statistical significance score
	Likely	Significant	It will be very difficult to interpret the low grade and high grade appendiceal cancer images.	The team will meet regularly with Dr. Coldren and other pathologists to ensure that the team is identifying important features to extract.
	Likely	Significant	Model does not work well with predicting external data	Make sure the data that the team are using are representative of all data, making sure the model can be generalized on different data when building the model
	Likely	Significant	High inaccuracy in prediction of results	Develop different models in parallel to reduce the chance that all models are

				inaccurate
	Unlikely	Moderate	It might be difficult to code certain aspects of the machine learning model	The team will utilize resources available (LinkedIn Learning, Department of Computer Science) to develop our technical coding skills related to machine learning
	Likely	Significant	Our understanding of pathology is not sufficient to develop the algorithm ourselves	Spending time with the pathologist will give the team information/intuition the team likely could not develop from literature reviews
	Possible	Significant	Integration of the image processing model with the image classification algorithm	Utilize resources from literature review and advice from ETEs to rectify this issue.
Other	Likely	Moderate	Unable to schedule meetings with physicians and stakeholders	The team will send out multiple time slots to stakeholders several days ahead of schedule so they can be available to meet with the team and/or for design review presentations.
	Likely	Significant	IRB Approval takes longer than anticipated and hinders the team's ability to progress	The team has submitted the names of everyone involved in the project including ETEs, faculty coaches, and the TA to Dr. Votanopoulos, who will secure IRB approval for the team. It could take 4 to 6 weeks, however, the team has not begun developing the algorithm yet because the team is still in the problem discovery phase.

Figure 25: Risk Analysis Matrix