



Prediction of Lung and Colon Cancer Using Image Processing in Machine Learning

Rajesh Rangan and G Srimugambigai

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

June 12, 2022

PREDECTION OF LUNG AND COLON CANCER USING IMAGE PROCESSING IN MACHINE LEARNING

Rajesh.R¹, Srimugambigai.G²

Associate Professor¹, UGScholar²

Department of Computer Science and Engineering

IFET College of Engineering, Villupuram

Abstract - Medicine and healthcare have progressed dramatically over the last four decades. The true reasons of a range of infections were discovered during this time, new medical testing were invented, and new treatments were developed. Despite our achievements, diseases like cancer continue to affect us because we are still vulnerable to them. Cancer is indeed the second leading cause of death in the world, killing one out of every six people. Cancer is the leading cause of death worldwide, with the most common types being gastrointestinal and lung cancer. The first stage classifies the existence and absence of a tumor using endoscopic image global features, while the second stage uses CNN (deep convolutional network) segmentation. According to the findings, the framework can detect cancer cells rise to 96.33 %. This model will assist medical professionals in developing a fully automated and reliable system that can detect various types of lung as well as colon cancers.

Keywords: Lung and colon cancer, Random forest classifier, Image processing, CNN.

INTRODUCTION

Machine learning is a subset of artificial intelligence. Machine learning aims to understand data structure and fit that information into designs that people can understand and use. Many tests and counseling sessions with lung and colorectal cancer specialists are required in the existing lung and colorectal cancer diagnosis process. In this project, however, the user has entered a histopathological picture into the prototype and obtains a complete diagnosis of type of cell, malignant status, and type of malignancy. Cancer is a term that refers to a group of illnesses in which mutated lymphocytes develop inside the body due to random mutations. When these cells are formed, they divide uncontrollably and disperse throughout the organs. If left untreated, most types of cancer will eventually kill you.

OBJECTIVE

The primary goal of this study is to distinguish between five sorts of cancer photos: first, colon and lung cell images, then malignant and benign cells within these categories, and finally, particular types of malignant cells. The goal of this project is to use the Random Forest Classifier to predict various cancers based on image processing.

METHODOLOGY

A user would obtain and enter histological lung and colon pictures into the model in this project, and the model would then provide a thorough diagnostic type of cell, status of malignant, and kind of malignancy. By automating processes in the process, machine learning is able to provide a single comprehensive output diagnosis. Users do not have to be doctors; they might just be assist who communicate the results the primary physician for analytic purpose. As a result, medical experts might use this model as a "second opinion." The dataset was obtained using a Convolution Neural Network (CNN), and then classification was performed. The tensor is utilized after the picture of the dataset has been classified to obtain a better image of classification.

OVERVIEW

The present lung and colon cancer diagnosis process necessitates several tests, consultations with lung and colon specialists, and secondary opinions before a comprehensive diagnosis can be made. A user would obtain and enter a histological lung or colon picture into the model in our project, and the model would provide a complete diagnostic type of cell, malignant status, kind of malignancy. ML speeds up the process by automating intermediate phases, resulting in a single diagnostic. Users do not have to be doctors; they might just be assistants who communicate the results to the primary physician for analysis. As a result, our model serves

as a 'second opinion' for doctors. Steps Taken by Medical Professional using ML

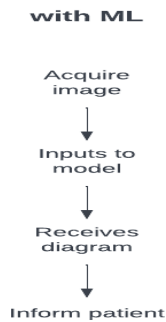


Figure1: With ML

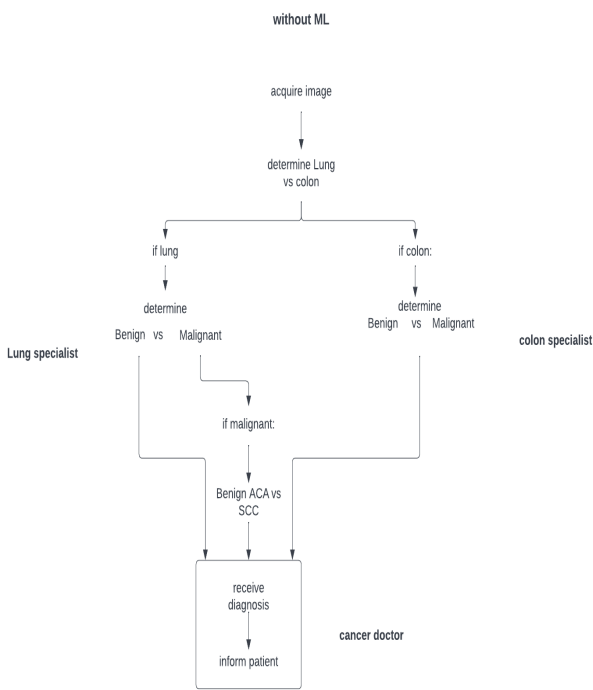


Figure 2: Without ML

1. Allows for image batching; can swiftly obtain several diagnoses
2. can provide a "second opinion" for clinicians;
3. automates a variety of decision-making processes into a single-input model
4. Convolutional filters can detect manual cell classification done visually.

Benefits of Using ML for Cancer Cell Classification

Because there is a lot of data to train and test our model on, we chose to classify lung and colon cells. Doctors aren't

always the ones who extract images from patients, so a classifier that can sort images by organ can help avoid misunderstanding. Furthermore, users can upload a batch of photos (each from a distinct patient) without having to manually filter and memorize organ kinds. Adenocarcinoma can attack both organs, adding to the complexity of the situation. We expect that by adding an organ differentiation phase, we can reduce the number of cases where cancer is correctly identified but the organ is misclassified. We intend to expand classification to other organs once we have enough data, but we believe the chosen organs are the most important.

SYSTEM ARCHITECTURE

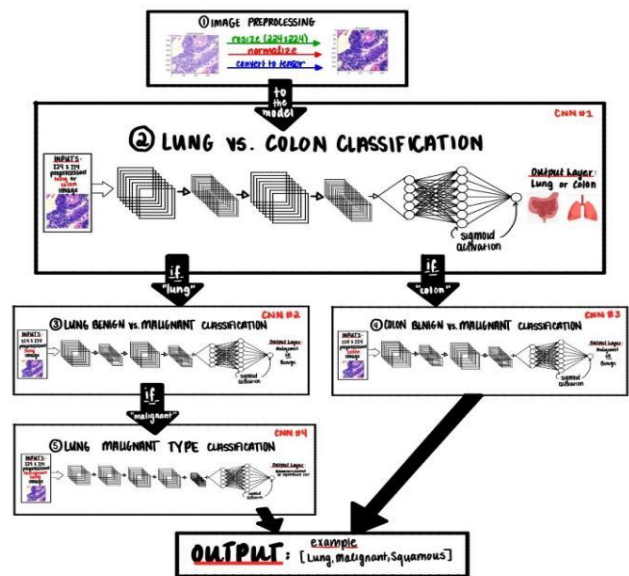


Figure 3: Project Illustration - Model Figures

BACKGROUND & RELATED WORK

ML is a rapidly developing technology that is proven to be increasingly beneficial in the diagnosis of cancer. For some of the most frequent kinds of the disease, such as the brain, prostate, breast, lung, bone, and skin, algorithms are being developed [2]. Recently, a new initiative was lauded for its capacity to detect malignant tumors with great accuracy [3]. Google Health and Imperial College London spearheaded the initiative, which aimed to leverage technology to improve breast cancer screening procedures [3]. The system was developed using a sample of 29,000 mammograms and tested against experienced radiologists' opinions [3]. When tested against a single radiologist, the system was found to be more effective than a two-person team [3]. The advantages of an algorithm like this one are particularly appealing because it

saves time and can help healthcare systems that lack radiologists [3]. This method should, in theory, be able to supplement one radiologist's viewpoint in order to get ideal results [3]. The purpose of this project is to use AI to make decisions on the existence of cancer in scans, which is similar to the goal of the previous one. The success of this breast cancer algorithm demonstrates that machine learning is capable of executing this task successfully.

DATA & DATA PROCESSING

There are five types of data: two types of colon cancer (benign and malignant adenocarcinoma) and three types of lung cancer (benign, malignant adenocarcinoma (ACA), and malignant squamous cell carcinoma (SCC)).

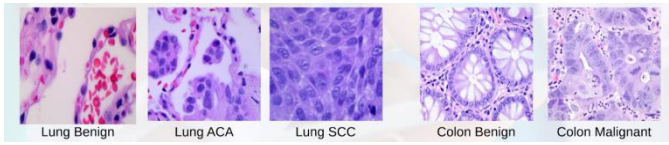


Figure 4: Data Visualization - Example from Each Class

This dataset contains 250 photos per class, which were pre-augmented to 5,000 images per class (for a total of 25,000 images) [4]. For uniformity and to reduce burden on our model, we normalized the pixel intensity of the photos to the [0,1] range and shrunk the images to 224x224 pixels. After then, the photos were converted to tensors.

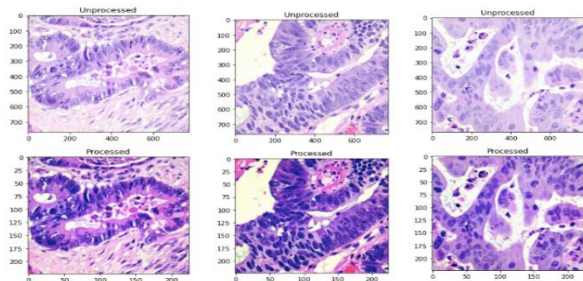


Figure 5: Data Visualization - Processed vs Unprocessed

Because the datasets we used had already been significantly preprocessed, the data splitting and sorting were our primary processing duties. Because our classifier is made up of four linked CNNs, we have to make sure that:

1. Each CNN had a dataset that was appropriate for its classification task.
2. The datasets for all CNNs were balanced.
3. A subset of the same training set was used to train all CNNs.

4. Each CNN was evaluated independently on a portion of the same testing set that they had never seen before.

5. The entire linked CNN model was thoroughly evaluated on a fresh collection of data that had never been seen before in training, validation, or individual testing.

To do so, we divided the dataset into 70:15: 7.5: 7.5 training, validation, individual testing, and overall testing sets. Individual model datasets were constructed from them, as illustrated below:

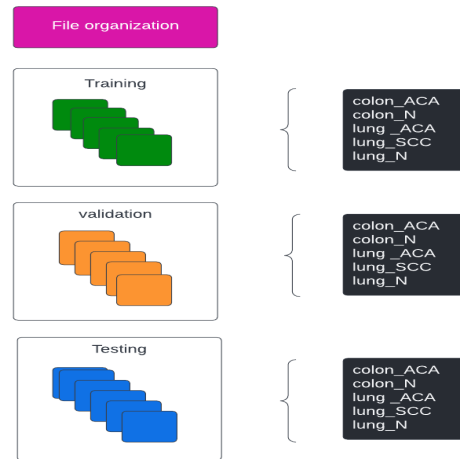


Figure 6: Data Split per Model

MODEL ARCHITECTURE

The architecture is made up of four binary CNNs. All CNNs take in a single preprocessed 224x224 square image and output a zero or a one based on the model's classification. The first CNN distinguishes between scans of the lungs and scans of the colon. If the image is determined to be a lung scan, it is sent to CNN #2, which can identify between cancerous and benign lung cells. If the image is determined to be a colon scan, it is sent to CNN #3, which can identify between cancerous and benign colon cells. The 4th CNN is used to distinguish between adenocarcinoma as well as squamous cell carcinoma, two types of lung cancer.

Because of its invariance properties, the CNN architecture was chosen [5]. We need a model that can recognize more complicated patterns that may be present because there is a lot of variation between photos, such as differences in cell size, orientation, and position [5].

	Batch size	Learning rate	#of epochs	#of convolution layers	#of pooling layers	#of fully connected layers
CNN1 : lung vs. colon	256	0.001	14	2	2	2
CNN2 : lung benign vs. malignant	150	0.01	9	2	2	2
CNN3 : colon benign vs malignant	256	0.001	14	2	2	2
CNN4 : lung SCC vs. ACA	64	0.0065	13	4	1	2

Table 1: Finalized Model Hyper parameters for Each Convolutional Neural Network

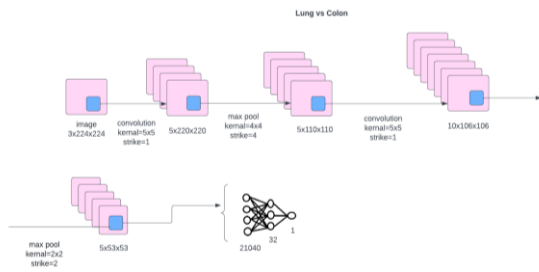


Figure 7: CNN #1 - Lung vs. Colon Architecture

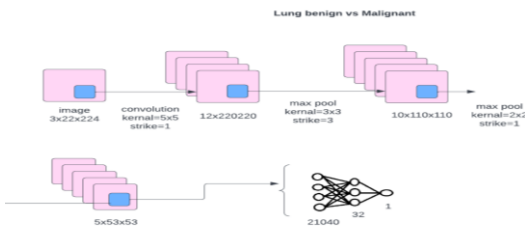


Figure 8: CNN #2 - Lung Benign vs. Malignant Architecture

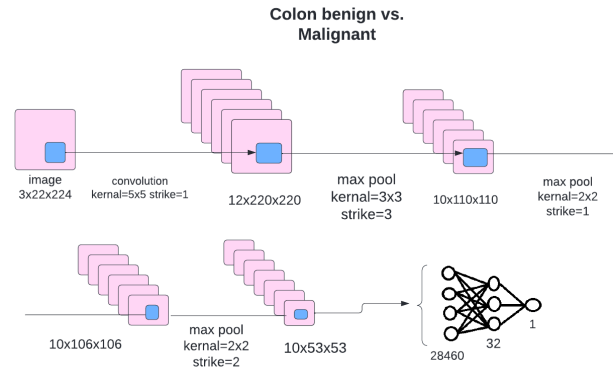


Figure 9: CNN #3 - Colon Benign vs. Malignant Architecture

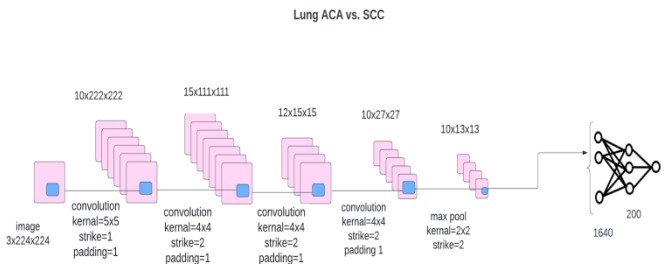


Figure 10: CNN #4 - Lung Malignant SCC vs. ACA Architecture.

BASELINE MODEL

The Random Forests Classifier [12] was used as the baseline. Over 7,000 photos from the training set were trained using 1,000 estimators. The classifier was then used to classify 1,000 photos from the validation collection. The model had an accuracy of 80.6 percent. This model had a number of flaws, including incorrectly identifying 19.7% of colon ACA photos as benign and incorrectly categorizing 25.0 percent of lung SCC images as lung ACA. Overall, the distinction between the two malignant lung subtypes was often muddled, and the model's high false negative rates rendered it useless.

PREDICTION	LABEL				
	Colon ACA (213 Images)	Colon Benign (177 Images)	Lung ACA (127 Images)	Lung Benign (213 Images)	Lung SCC (Images Images)
Colon ACA	167	16	0	2	0
Colon Benign	41	161	0	18	0
Lung ACA	4	0	109	5	50
Lung Benign	1	0	12	195	12
Lung SCC	0	0	30	5	136
CLASS ACCURACY	78.4%	91.0%	78.7%	86.9%	69.0%
FALSE POSITIVE (classify non-cancerous as cancerous)		9% as colon ACA		4.7% as lung SCC or lung ACA or colon ACA	
FALSE NEGATIVE (classify cancerous as non-cancerous)	19.7% as colon benign or lung benign		6.1% as lung benign		6.0% as lung benign
INCORRECT CANCER TYPE (classify ACA as SCC or vice versa)			15.2% as lung SCC		25.0% as lung ACA
WRONG ORGAN (classify colon as lung or vice versa)	2.3% as lung			9.4% as colon	

Table 2: Baseline Model Confusion Matrix Analysis

QUANTITATIVE RESULTS

PREDICTION	LABEL (373 IMGs/CLASSES)				
	Colon ACA	Colon Benign	Lung ACA	Lung Benign	Lung SCC
Colon ACA	359	72	0	0	0
Colon Benign	8	301	0	0	0
Lung ACA	7	0	347	5	49
Lung Benign	0	0	4	368	0
Lung SCC	0	0	22	0	324
CLASS ACCURACY	96.0%	80.7%	93.0%	98.7%	86.9%
FALSE POSITIVE (classify non-cancerous as cancerous)		19.3% (as colon ACA)		1.3% (as lung ACA)	
FALSE NEGATIVE (classify cancerous as non-cancerous)	2.1% (as colon benign)		1.1% (as lung benign)		
INCORRECT CANCER TYPE (classify ACA as SCC or vice versa)			5.9% (as lung SCC)		13.1% (as lung ACA)
WRONG ORGAN (classify colon as lung or vice versa)	1.9% (as lung ACA) *				

* Cancer type correct but organ incorrectly classified.

Table 3: Overall Model Confusion Matrix Analysis

Overall, the model performed well, with the best results in the lung benign, lung adenocarcinoma, and colon cancer scan classes. The overall accuracy of the model was 91.05 percent. When it came to distinguishing between lung and colon images, the model performed well. The accuracy in determining whether a lung tumor was malignant or benign was likewise excellent. Because unique CNNs were employed for each phase of the classification process, each had hyper parameters and architecture that were tailored to the classification task at hand. The more board convolutional neural networks (CNN #1 and CNN #2) have very high accuracies. The CNNs in the outskirts (CNN #3 and CNN #4) made the most mistakes. This meant that error propagation was kept to a minimum.

	Training accuracy	Validation accuracy	Testing accuracy
CNN1: lung vs. colon	99.9%	99.99%	99.99%
CNN2: lung benign vs. malignant	99.9%	99.5%	99.3%
CNN3: colon benign vs malignant	100%	94.8%	96.1%
CNN4: lung SCC vs. ACA	96.0%	90.1%	89.5%

Table 4: Training, Validation, & Testing Accuracies for Each Convolutional Neural Network

False negative findings for both the colon and the lung were also minimal, showing that the model would only overlook a malignant scan in rare cases. This is critical because erroneous negative results could lead to the patient seeking therapy much later, when the cancer has progressed. This critical error is avoided by our model. One disadvantage is that it has a tendency to generate false positive results on colon imaging. Healthy colon scans were commonly mistaken as cancerous by the model. This could lead to a healthy patient having to undergo a second scan or biopsy, which could be costly and inconvenient. Although less serious than erroneous negative classification, it has the potential to be harmful. In addition, the model misidentified malignant lung subtypes on occasion, albeit much less frequently than the baseline model. This could lead to the patient receiving an ineffective treatment. This is why, at this time, this model is only meant to be used in conjunction with a doctor; while the findings are promising, it is not a substitute for medical advice.



Table 5: Error/Loss Training Curves for Each Convolutional Neural Network

QUALITATIVE SAMPLE









QUALITATIVE SAMPLE TESTING				
CNN #1	 Ex: lungacc1010.jpeg	CNN #1 → LUNG 10/10 Sample Predictions Correct	 Ex: colonn1010.jpeg	CNN #1 → COLON 10/10 Sample Predictions Correct
CNN #2	 Ex: lungaca1010.jpeg	CNN #2 → MALIGNANT 10/10 Sample Predictions Correct	 Ex: lungn1010.jpeg	CNN #2 → BENIGN 10/10 Sample Predictions Correct
CNN #3	 Ex: colonica1010.jpeg	CNN #3 → MALIGNANT 10/10 Sample Predictions Correct	 Ex: colonn150.jpeg	CNN #3 → BENIGN 9/10 Sample Predictions Correct
CNN #4	 Ex: lungaca1012.jpeg	CNN #4 → ACA 9/10 Sample Predictions Correct	 Ex: lungacc102.jpeg	CNN #4 → SCC 9/10 Sample Predictions Correct

Table 6: Qualitative Sample Testing

Table 6 shows that the model performs the worst on benign colon scans, as we expected given that CNN #3 performed poorly in sample testing when it came to categorizing benign pictures. Furthermore, the model correctly detects benign lung pictures. This makes sense because these images only pass through CNN #1 and #2, which were tested and sampled to near-perfect accuracy. Benign lung scans, on the other hand, are rarely identified as malignant, and if they are, they are simply classed as ACA. Only lung ACA cells, on the other hand, are wrongly labelled as benign (false negative). some lung ACA samples exhibit red organelles, pink colors that are similar to benign images, and negative space, whereas SCC images are congested with dark blue cells.

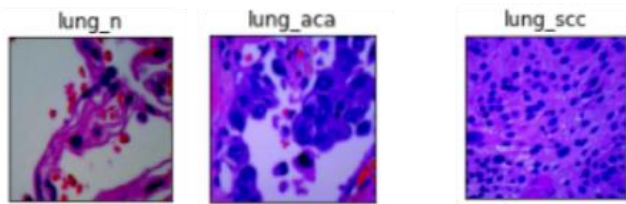


Figure 11: Data Visualization - Lung Samples

Lung malignant subtypes are frequently misidentified, while SCC pictures are more likely to be classified as ACA. This is supported by the following example, which shows lung congestion and a lack of negative space:

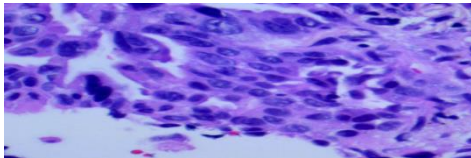


Figure 12: Data Visualization - lungaca1123.png

Our model misclassified colon ACA as lung ACA on multiple occasions, despite the fact that lung vs. colon categorization is nearly perfect. In general, colon and lung

cells appear to be highly distinct, however some samples with no defined cell borders appear to be identical.

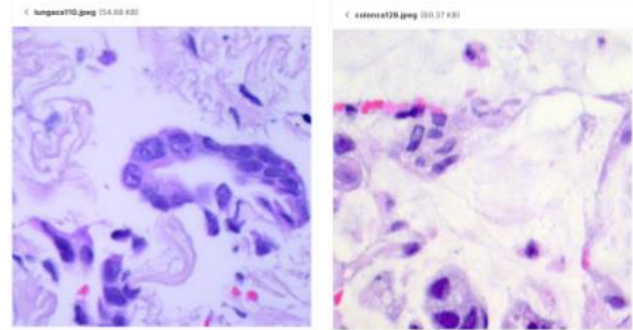


Figure 13: Data Visualization - Lung ACA Sample vs. Colon ACA Sample.png

MODEL EVALUATION

A demo set of ten photos was utilized to demonstrate the model's performance on new data. Because it was difficult to obtain fully new data due to the nature of this situation, we were limited to a subset of the original photographs (holdout set).

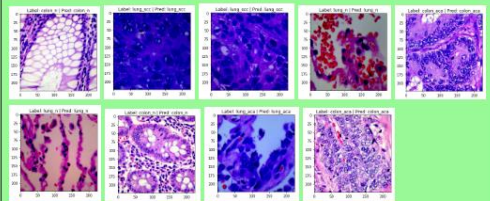
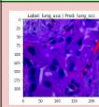
CORRECTLY CLASSIFIED	
INCORRECTLY CLASSIFIED	

Table 7: Classification of Dataset Images

The fact that the dataset has been enhanced from 1250 original photographs is an issue when employing images from the same dataset. Because we can't tell which photographs are enhanced, it's possible that the images used in the tests weren't fully unique. Although augmentation is sufficient to assure the model does not identify the image, the style and attributes of the photos are highly similar, limiting our ability to determine the model's performance on completely new data. To further validate our model, we used Google Images to find 1 image of each class [6][7][8][9]. Two of the five samples were appropriately classified: the lung and colon benign pictures. The model was able to correctly distinguish between lung and colon images 80% of the time, but struggled with the individual classes. We were unable to ascertain the technique by which the scans were taken and produced due to the difficulty in locating genuine

photos. However, the process utilized to generate the photos in our collection is likely to have been different, as seen by the large variations in these photographs from our originals. The model's performance on appropriately prepared photos demonstrated that it could categorize fresh data if it was prepared in the same way as the data in the dataset. Because our model was trained on a single dataset, it has learned to recognize only one way of scanning and processing, limiting its performance. However, if more reliable data is made accessible, the architecture and principles mentioned could be generalized.

DISCUSSION

Using machine learning to supplement medical diagnoses is a difficult and dangerous process. False-positive outcomes can lead to more mistakes, but false-negative predictions can be devastating to a patient's health and well-being. In the instance of CNN #2 and #3, for example, misclassifying a malignant picture as benign in practice could have disastrous consequences for cancer survivability. The focus of most ML models is accuracy, as it is with this one. The system achieved 91.05 percent accuracy on a test dataset of 1865 photos, which is a good result when compared to the baseline. However, 8.95 percent of the test set, or 167 photos, were incorrectly identified, which is extremely important in medicine. Because even the tiniest error can have a significant impact, any machine learning model should only be used in conjunction with the experience of a medical practitioner at this time. Perhaps, when our system is combined with doctors' knowledge, we will be able to minimize misdiagnoses to a true 0%. Given our unique approach to the problem, it's critical to discuss the advantages of a four-binary neural-network technique over a single multi-class CNN. To support our judgement, we created a multi-class classifier as part of our project inquiry. The multi-class classifier has various flaws, including memory and time requirements for training (given that it would work with much more photos separated into five groups), as well as poorer accuracies (having only achieved 65 percent initially). With four separate binary networks, each stage of the algorithm may be fine-tuned individually. Furthermore, having a high-accuracy model that distinguishes across organs as the first stage lowers misdiagnosis of tumours of comparable types (Lung ACA vs. Colon ACA). A model tree also allows for future development flexibility. CNN #1 might be changed to include the classification of other important body organs, and CNN #4 could contain photos of small cell lung cancer (SCLC) in addition to NSCLC images (ACA and SCC). Finally, numerous models allow users to "activate" classification functions on a case-by-case basis to best suit their needs. We've learned the importance of having preprocessed data as we've built this repository. While

prepared data may be ideal for training or testing a model, it will not perform as well on data that has been processed in a different way (prepared with different dyes etc.). Furthermore, aggregate accuracy scores are frequently deceptive, and false negative/positive values are a more exact way to assess accuracy. detect the flaws in a model Finally, we've admitted that a typical approach to a problem isn't always the ideal, as independent binary models outperformed a multi-class classifier in our scenario.

CONCLUSION

We used an image enhancing technique called unsharp masking to preprocess the data. For image classification, three feature sets were extracted. The features were then combined to form a combined feature set that was given into the machine learning algorithm. The performance of our proposed approach will also be examined on different histological pictures of colon and lung cancer to determine its efficacy.

REFERENCES

- [1] A. Szöllösi, "Pigeons classify breast cancer images," BCC News, 20-Nov-2015. [Online]. Available: <https://www.bbc.com/news/science-environment-34878151>. [Accessed: 10-Aug-2020].
- [2] N. Savage, "How AI is improving cancer diagnostics," Nature News, 25-Mar-2020. [Online]. Available: <https://www.nature.com/articles/d41586-020-00847-2>. [Accessed: 13-Jun-2020].
- [3] F. Walsh, "AI 'outperforms' doctors diagnosing breast cancer," BBC News, 02-Jan-2020. [Online]. Available: <https://www.bbc.com/news/health-50857759>. [Accessed: 13-Jun-2020].
- [4] Borkowski AA, Bui MM, Thomas LB, Wilson CP, DeLand LA, Mastorides SM. Lung and Colon Cancer Histopathological Image Dataset (LC25000). [Dataset]. Available:<https://www.kaggle.com/andrewmvd/lung-and-colon-cancer-histopathological-images> [Accessed: 18 May 2020].
- [5] S. Colic, Class Lecture, Topic: "CNN Architectures and Transfer Learning." APS360H1, Faculty of Applied Science and Engineering, University of Toronto, Toronto, Jun., 1, 2020
- [6] Histology of the Lung. Youtube, 2016.
- [7] J. Stojšić, "Precise Diagnosis of Histological Type of Lung Carcinoma: The First Step in Personalized Therapy," Lung Cancer - Strategies for Diagnosis and Treatment, 2018.

[8] V. S. Chandan, "Normal Histology of Gastrointestinal Tract," *Surgical Pathology of Non-neoplastic Gastrointestinal Diseases*, pp. 3–18, 2019.

[9] Memorang, "Colon Cancer (MCC Exam #3) Flashcards," Memorang. [Online]. Available: https://www.memorangapp.com/flashcards/92659/Colon_Cancer/. [Accessed: 10-Aug-2020].

[10] J. Voigt, "The Future of Artificial Intelligence in Medicine," *Wharton Magazing*. [Online]. Available: <https://magazine.wharton.upenn.edu/digital/the-future-of-artificial-intelligence-in-medicine/>. [Accessed: 10-Aug-2020].

[11] "Lab_2_Cats_vs_Dogs," APS360, Faculty of Applied Science and Engineering, University of Toronto, Toronto, summer 2020. [.ipynb file]. Available: https://q.utoronto.ca/courses/155423/files/7477635/download?download_frd=1.

[12] F. Boyles, "Using Random Forests in Python with Scikit-Learn," Oxford Protein Informatics Group, 26-Jul-2017. [Online]. Available: <https://www.blopig.com/blog/2017/07/using-random-forests-in-python-with-scikit-learn/>. [Accessed: 10-July-2020].