Identifying Cancer Types in Microscope Images of Lung and Colon Cells

Adam Sayed

Abstract

This program seeks to reach a high accuracy of detecting types of cancer in addition to whether or not cells are benign. The model categorizes the cell images into 5 classes: benign colon, benign lung, cancerous lung adenocarcinoma, cancerous colon adenocarcinoma, and cancerous lung squamous cell carcinoma. Two models are then trained with the images: a random forest classifier, and a deep neural network using convolutional layers. Finally a third model predicts which classifier is more likely to make the correct prediction for a given image, and that classifier's prediction is used. This tactic allows the models to learn different patterns, but still be effective at predicting all classes of images.

Introduction

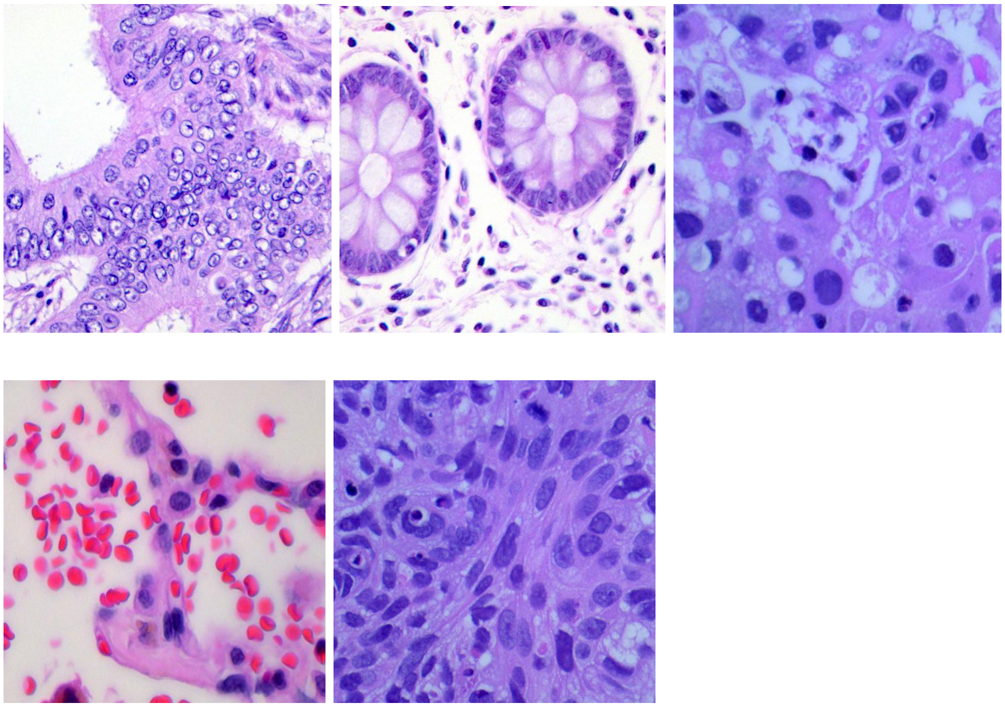
Lung adenocarcinoma cells make up large proportions of lung cancer cases and colon cancer is one of the most prevalent cancers in the United States.[1] Although many different medicines have been recently developed to attack these cancers, the most effective way to stop it is early detection. Early diagnoses can lead to minimally invasive surgeries or minor radiation treatments, which can be significantly less harmful to patients, and can remove tumors before

[1] https://www.nature.com/articles/s41419-017-0063-y
https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html

they metastasize to other parts of the body. Developments in creating accurate artificial intelligence models to identify cancer can assist current manual analyses by doctors to improve the rate of correct diagnoses in patients. Other AI models have attempted to diagnose cancer in cells, but are not advanced enough yet to be consistently accurate.[2]

Dataset

The dataset for this model consists of 25000 images of human cells, divided into 5 sets of 5000 of the following types: benign colon cells, benign lung cells, squamous cell carcinoma lung cells, adenocarcinoma lung cells, and adenocarcinoma colon cells[3]. Each image is a square 764 by 764 pixels, but this resolution is resized during the preprocessing stage.



[2] https://www.jmir.org/2021/3/e23483

[3] https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images

Fig 1: *Pictured left to right and top to bottom - adenocarcinoma colon cells, benign colon cells, adenocarcinoma lung cells, benign lung cells, squamous cell carcinoma lung cells*

In some cases such as the benign lung cells with generally different color patterns including bright pink spots, the class is easy to detect, but in others, more extensive learning is required for models to pick up on the patterns between cell types.

Methods and Models

The images are loaded into an array and a label array is split into the five classes of images. The data is then split into 80 percent training images and 20 percent testing images. The training data is used to teach the models patterns, and the testing data is saved for later to get an unbiased perspective on the accuracy of the model.

This model combines the predictions of a random forest classifier and a deep neural network using two sets of convolutional layers. The random forest classifier is a set of binary trees that determine which of the pixels in the image are most important for classification, and then create a binary tree. Because the classes of images contain distinct clusters of pixels, patterns can be effectively learned by binary trees. The random forest in the model classifier uses 100 binary trees to analyze different patterns within the training set. The large number of binary trees combine to make the first half of this model by detecting a large majority of the easier images that present more visible patterns. The trees prioritize the easy patterns because they quickly result in the highest accuracy, leaving the harder images to be classified from the second half of the model.

The other predictions are made by the deep neural network which attempts to pick up the more complex patterns in the harder to identify cells. It does this by training on the images that the random forest classifier predicted incorrectly, allowing for two different types of patterns to be learned by the model. The network uses a rectified linear activation function on a convolutional layer and then pools that output. This process is repeated twice to pick up complex patterns within the images and then is flattened to the correct dimension. Then 40% of the edges are removed from the network using dropout to remove chances of overfitting within the network.[4] The final network uses the same training images as the first half of the network, but its labels are based on whether that training image was predicted correctly or incorrectly by the first half of the network. When it is tested, it predicts which network is more likely to understand the pattern that would best identify the image, and uses that network's prediction of the given image. Using this combined tester network tactic, the final model can focus on where both models are accurate, and avoid a majority of the mistakes caused by harder to identify patterns.

Results

When trained across all 25,000 images the final classifier had an accuracy of (final code has not been trained on all images). The random forest classifier on its own had an accuracy of 80.4%. This was improved when combined with the 65% accuracy of the deep neural network trained on the images predicted incorrectly by the former network. Other classifiers such as MLP, KNN, and AdaBoost failed for various reasons. MLP and AdaBoost were not able to pick up the patterns as accurately as the binary tree due to the importance of specific pixels in the dataset, so they were not as useful to combine into the final model. KNN classifiers were not able to capture complex enough patterns to distinguish between the classes of cells. Compared to other models that attempted to classify this dataset, this is the only one that combined multiple classifiers to

[4] https://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf

create the final result instead of using a singular deep neural network. This model still is most error prone in distinguishing benign and malignant colon cells, as they appear more similar than cancerous and normal lung cells. Future models can attempt to preprocess the images further to filter out irrelevant pixels and help the model examine the distinguishing features for each class within an image.

Conclusion

The combined model approach resulted in a noticeable improvement from both models being examined separately. Using two distinct models allowed for a close understanding of the underlying patterns of cancer diagnosis within each classifier. Further developments will allow artificial intelligence to be used as a primary care tool, and improve accuracy in treating and diagnosing potential cancer patients.

References

Denisenko, T.V., Budkevich, I.N. & Zhivotovsky, B. Cell death-based treatment of lung adenocarcinoma. *Cell Death Dis* 9, 117 (2018). https://doi.org/10.1038/s41419-017-0063-y

https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html

Jones O, Calanzani N, Saji S, Duffy S, Emery J, Hamilton W, Singh H, de Wit N, Walter F
Artificial Intelligence Techniques That May Be Applied to Primary Care Data to Facilitate Earlier Diagnosis of Cancer: Systematic Review
J Med Internet Res 2021;23(3):e23483
URL: https://www.jmir.org/2021/3/e23483 DOI: 10.2196/23483

https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images
Srivastava, Nitish, et. al, Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research 15*, 1929 (2014).
https://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf