Developing an Accurate AI Algorithm for Histopathologic Cancer Detection

Leah Ning

Abstract

This paper discusses the development of a machine learning algorithm that accurately detects metastatic breast cancer (the cancer has spread elsewhere from its origin part) in select images that come from pathology scans of lymph node sections. Being able to develop an accurate artificial intelligence (AI) algorithm would help significantly in breast cancer diagnosis since manual exanimation of lymph node scans is both tedious and oftentimes highly subjective. The usage of AI in the diagnosis process provides a much more straightforward, reliable, and efficient method for medical professionals and would enable faster diagnosis and, therefore, more immediate treatment. The overall approach used was to train a convolution neural network (CNN) based on a set of pathology scan data and using the trained model to binarily classify if a new scan were benign or malignant, outputting a 0 or a 1, respectively. The final model's prediction accuracy is very high, with 100% for the train set and over 70% for the test set. Being able to have such high accuracy using an AI model is monumental in regards to medical pathology and cancer detection. Having AI as a new tool capable of quick detection will significantly help medical professionals and patients suffering from cancer.

Introduction

Manual detection of cancer in scans is very tedious and is prone to human error, so it becomes paramount for AI to provide a tool that processes these images and accurately diagnoses the patient's condition. In this specific research project, we will be focusing on the lymph node scans of women with breast cancer, which is the most common cancer for women residing in the US, other than skin cancer. Research statistics show that about 1 in 8 women in the United States will develop invasive breast cancer throughout her life. The death rates from this type of cancer are higher than other types of cancers, excluding lung cancer, which makes early diagnosis essential to the treatment process. Because the manual detection time is very long, AI's role in detection becomes extremely important. Accurate detection algorithms would aid in earlier detection and potentially also help in decreasing death rates if these resulting earlier treatments could prove to be instrumental in the survival rates of many cancer patients.

Background

If AI is not utilized, currently, pathologists manually diagnose breast cancer, generally through a tissue-based diagnostic testing where a biopsy of the suspected tissue is performed for a histological examination. A standard dye, usually consisting of hematoxylin and eosin (H&E), is used on the tissue. The hematoxylin binds itself to DNA, staining the nuclei in the tissue purple, while the eosin binds itself to proteins, staining other structures in the tissue pink. The stained tissue is fitted into a glass slide and then examined by the pathologist under a microscope. This entire diagnosis process is tedious, time-consuming, and also prone to human error, with an average accuracy for professionals around 70%. Such accuracy can result in severe consequences for patients who are misdiagnosed. These limitations of human diagnosis can be addressed by the implementation of AI in the diagnosis and detection process.

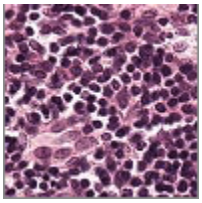
Dataset

The dataset used for this research project will be a modification of the PatchCamelyon (PCam) dataset which consists of color images (tissues stained by H&E) extracted from histopathologic scans of lymph node sections. The dataset had images that were either benign or malignant (negative or positive - 0 or 1, respectively). From Kaggle, I downloaded the 220,025 available images from the provided train dataset, along with the corresponding labels (0 or 1) from the provided train_labels.csv file. Images classified as positive have at least one pixel of tumor tissue out of the picture's entire size (96 by 96 pixels total).

After downloading the dataset, I proceeded to split the provided train dataset into two subsets, putting about ⅓ (33%) of the data into a training set and the remaining 67% into a test set. Because the correct labels for both of these subsets are available in the csv file, I am able to obtain an estimation of the error of the algorithm and adjust it by comparing the classifications my algorithm provides me with and the actually correct labels.

No data preprocessing was necessary since all of the images were rather uniform, with the same size; the backgrounds of the photos were also needed, and thus, were not preprocessed and removed.

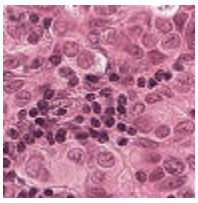
Here is a sample photo that is classified as negative/benign (0):



ID: 07ed972dc9add8af5ef375e8b5c2f415cd232474

Label: 0

Here is a sample photo that is classified as positive/malignant (1):



ID: 742cdd20802fed72b9021e31de9bdf530b5f341b.tif

Label: 1

Methodology/Models

Classifier Type	Classifier Overview
Logistic Regression	Each of the images' pixels is multiplied by a weight, and the products are all summed together. The end sum is within the range [0, 1], and the integer the sum is closest to will be the prediction the algorithm provides (0 for negative, 1 for positive).
KNeighbors	This classifier finds the images closest to the image the algorithm is currently trying to predict (the number of neighbors is based on the parameter given). From these images, the algorithm takes the labels and the majority of what the labels are is what it will predict for the new image.
Neural Network MLP (Multiple-layers perceptron)	Generally speaking, MLP takes the pixels as the input and the edge has weight, and goes through hidden layers to the final output layer. It takes images from the train set, goes through the network, and sees if the prediction is accurate or not. Then, based on the measurement of accuracy, it will change the network and train the data.
SVC (Support vector classifier)	SVC finds a line or a plane so that the points are split so that the ones above the line are one class and the points below are another class.
Decision Tree	Decision Tree has nested if statements so that eventually, a prediction of class 0 or 1 can be outputted.
Random Forest	Random Forest is a superset of the decision tree, so it has multiple decision trees. The end prediction is made by taking the result of what the majority of the decision trees output.

Results/Discussion

Currently, there is not enough memory for my Jupyter notebook to use all 220,025 images for the algorithm, so the following results are from the usage of the first 30,000 images from the dataset.

The following table shows the test and train set accuracies for all six classifier types used:

Classifier Type	**Test Set Accuracy**	**Train Set Accuracy**
Logistic Regression	0.5812	0.8677
KNeighbors	0.6396	0.7852
Neural Network MLP (Multiple-layers perceptron)	0.6046	0.5946
SVC (Support vector classifier)	N/A	N/A
Decision Tree	0.6475	1.0
Random Forest	0.7426	0.7921

The dataset used contained too many images, so it used up too much memory to try to run the program with all the images. The performance of my algorithm was evaluated by calculating the accuracy of how many images were correctly identified with the right corresponding label.

Conclusion

This paper addressed the issue that manual diagnosis of cancer has and the improvement in efficiency that AI provides. An algorithm was created to label pathological scans as accurately as possible. Six different classifiers were used, and the end accuracy results proved that the Random Forest classifier had the highest test set accuracy of 74.26% and the Decision Tree classifier had the highest train set accuracy of 100%. Future research would include improving the algorithm more and testing it on more images and datasets. Additionally, I would like to further advance my algorithm to be able to correctly diagnose pathological images for other types of cancers other than breast cancer. More improvements would be using a dataset that imports easier into my code and to try even more classifiers.

References

[1] *Breast cancer facts and statistics.* (n.d.). Retrieved September 18, 2022, from https://www.breastcancer.org/facts-statistics

[2] *Breast cancer statistics: How common is breast cancer?* American Cancer Society. (n.d.). Retrieved September 18, 2022, from https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

[3] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, M. Welling. "Rotation Equivariant CNNs for Digital Pathology". arXiv:1806.03962

[4] *Can artificial intelligence help see cancer in new ways?* National Cancer Institute. (2022, July 28, June 29, & June 23). Retrieved September 18, 2022, from https://www.cancer.gov/news-events/cancer-currents-blog/2022/artificial-intelligence-cancer-imaging

[5] Ehteshami Bejnordi et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA: The Journal of the American Medical Association, 318(22), 2199–2210. doi:jama.2017.14585

[6] Munien, C., & Viriri, S. (2021, April 9). *Classification of Hematoxylin and Eosin-Stained Breast Cancer Histology Microscopy Images Using Transfer Learning with EfficientNets*. Computational Intelligence and Neuroscience. Retrieved September 5, 2022, from https://www.hindawi.com/journals/cin/2021/5580914/

[7] Pham, H. H. N., Futakuchi, M., Bychkov, A., Furukawa, T., Kuroda, K., & Fukuoka, J. (2019, September 18). *Detection of lung cancer lymph node metastases from whole-slide histopathologic images using a two-step deep learning approach*. The American Journal of Pathology. Retrieved September 18, 2022, from https://ajp.amjpathol.org/article/S0002-9440(19)30718-7/fulltext