Machine Learning Write Up - Lung Cancer Feature Extraction

Urban Halpern

December 18, 2024

Github Repo: https://github.com/uhalpern/Lung_Cancer_Classification

Introduction

1. Problem

Lung cancer is one of the most deadly diseases in the world with a five year survival rate of 28% (Centers for Disease Control and Prevention, 2024). A notable technique for diagnosing cancer is Histopathology. This technique involves examining cells and tissues under the microscope and can be used to identify abnormalities. Furthermore, different types of lung cancer have contrasting responses to therapies so it is not only important to identify that cancerous cells are present, but also to generate an accurate and specific diagnosis. The manual process of diagnosis through histopathology can take up to two months in countries that lack personnel and laboratories (Masamba et. al., 2017). Deep learning can address these challenges through the use of CNNs that can take histopathological images of lung cancer and learn a function that returns a class label as output. Al-Jabbar et al. (2023) used this method to extract features from images and feed them to a fully connected network with a test-set accuracy of 99.64%. However, images are high dimensional and can be very computationally expensive to train on. Scikit-image is a python library that can extract numerical measurements from images to simplify the problem and filter out important features.

The purpose of this project is to determine if these extracted features can still be used to train an accurate classifier. First, a benchmark will be done to see what accuracy a CNN will get on the dataset with minimal training. Then, numerical features will be extracted from the images and trained on using a neural network to see if this benchmark can be reached.

2. Dataset

The L25000 dataset by Borkowski et. al. is composed of 25000 total 768x768 pixel images. The dataset was created from 250 histopathological images of each of the following classes: lung benign tissue, lung adenocarcinoma, lung squamous cell carcinoma, colon adenocarcinoma, colon benign tissue. These images were augmented to create 5000 images of each class. It is not clear what augmentations were done based on the documentation for the data. For this project, only the lung cancer images were extracted to create a final dataset of 15,000 lung cancer images. Furthermore, each of these images were processed using scikit-image to create a tabular dataset of shape 15,000 x 35.

3. Model Choice

To set a benchmark performance, a Convolution Neural Network was used. CNNs are a constrained version of a multilayer perceptron (MLP) that employ weight sharing and local connectivity. This is intuitive for image classification tasks since we can assume that pixels close together in an image are highly correlated. Furthermore, pooling can capture long range dependencies of pixels by reducing dimensionality. Images have desirable invariances since even though the pixel value of two images may be different, they may still convey the same semantic meaning. Weight sharing and local connectivity allows CNNs to generalize well to this invariance while a traditional MLP might give a different output for slightly different pixel values. The CNN model used for the

benchmark was Inception-v3 with a 42 layer deep. The objective function to optimize was the multiclass Categorical Cross Entropy loss.

For the second part of the project, a different neural network architecture was used to train on the tabular data. After extracting the features from each of the images, a lot of the benefits of using convolutional neural networks are lost. There is no longer a semblance of local connectivity with the input since there are no spatial dimensions to the input. A lot of the information is lost since the images are reduced to 35 1-D features instead of a 229x299x3 dimension. In order to effectively learn from tabular data, the TabNet model was used. This model was proposed by Arik & Pfister (2019) with tabular datasets in mind. It uses sequential-attention in which it dynamically selects the most relevant features for each input instance separately from thes rest of the dataset. The multiclass Categorical Cross entropy loss was optimized for this model.

Features and Preprocessing

For the benchmark mode, each image was cropped to 229x229 pixels and contains three channels for color. Furthermore, the images were normalized by mean=[0.485, 0.456, 0.406] and std=[0.229, 0.224, 0.225]. Each color channel intensity was subtracted by the mean and divided by its associated standard deviation. This represents the mean and standard deviation of the ImageNet dataset. However, the dataset for this project likely has a different distribution for color channel intensities.

For the tabular dataset, the features below were extracted using the skimage library. The first four represent statistics calculated from a grayscale representation of the image. The local binary patterns represent texture patterns found by assigning a LBP value to each pixel based on its neighboring pixels and binning them into frequent patterns. The morphological features of area, perimeter, eccentricity, solidity, and extent were extracted by creating a binary mask of grayscale pixel intensity using a threshold of 127. Then, characteristics of this mask were extracted.

Feature	Description
contrast	Measure of the intensity contrast between a pixel and its neighbor over the whole image. High contrast indicates a lot of variations (sharp edges), while low contrast indicates smooth variations
homogeneity	Represents how often pixel pairs with specific gray level values occur at a certain distance and orientation
energy	Sum of squared elements in the image.
correlation	Measure of how correlated a pixel is to its neighbor over the whole image
local binary pattern (0-25)	Extracts distinct patterns of pixel intensity variations that represent textures.
area	Number of pixels within mask
perimeter	Length of boundary around connected region

eccentricity	Elongation of shape
solidity	Compactness of shape distribution.
extent	Proportion of bounding box occupied by connecting region

Furthermore, some of the extracted features were heavily skewed. For features with two tails, the square root of the data was taken. For heavily skewed distributions, the natural log was taken along with min and max scaling.

Data Splits

For both datasets, the 15,000 instances were randomly shuffled and then a 60/20/20 train/validation/test split was done using train_test_split from sklearn.model_selection. In each split, the proportion of classes was kept equal. This resulted in 9000 samples for training, 3000 for validation, and 3000 for test. The training set was used for learning the function mapping from inputs to outputs by updating the models weights and biases along with early stopping criteria for the TabNet model. The validation set was used for hyperparameters selection and the test set was used to evaluate the model's ability to generalize. The model with the highest validation accuracy at the last training epoch was chosen to evaluate on the test set.

Hyperparameter Search Space for TabNet Classifier

Hyperparameter	**Search Space Range (inclusive)**	**Description**
Learning Rate	[5e-3, 1e-2, 2e-2]	Controls the magnitude of the weight and bias update after each batch is processed.
Epochs	[75, 100]	How many complete passes of the training dataset are taken.
Batch Size	[32, 100]	Number of training examples fed forward through the network during one training iteration.
Weight Decay	[1e-3, 5e-3, 1e-2]	L2 regularization to prevent weights from getting too large and push them to 0.
Optimization Algorithm	Adam	Optimization algorithm used for updating weights and biases.
Early Stopping	Patience=5	How many epochs to wait without

		seeing a minimum increase of performance on the validation set before stopping training early.
n_d (Decision Width)	[7,8,9]	Width of decision layers (number of neurons)
n_a (Attention Width)	[7, 8, 9]	Width of the attention layers. More features can be attended to with a higher width.
n_steps (Number of Steps)	[2, 3, 4]	Number of decision steps or iterations the model performs. In each step, the attention mechanism selects which features to focus on and the model makes predictions.
gamma (regularization for feature reuse)	[1.2, 1.3, 1.5]	Higher gamma discourages the model from using the same features at each decision step
Lambda_sparse (Sparsity Regularization)	[5e-5, 1e-4, 2e-4]	Penalizes the attention mechanism for relying on too many features at each step

Hyperparameter Optimization

The hyperparameters in the search space were not tested exhaustively. A random search was conducted using the defined search spaces above. Some hyperparameters were kept constant such as the optimization algorithm and patience. A total of 30 models were trained using random selections from the search space. Each model was saved with its weights and biases preserved. The training losses for all the epochs and the final validation accuracy were also logged. At the end of the training iterations, the model with the highest validation accuracy at the last epoch was saved.

Benchmark Model Hyperparameters:

Hyperparameter	**Value For Model 1**
Learning Rate	1e-3
Epochs	2

Batch Size	32
Optimization Algorithm	Adam
Unfrozen Layers	Last 2
Early Stopping	Patience=2 min_delta= 0.005

Best TabNet Model From Training Iteration 3:

Hyperparameter	**Value**
Learning Rate	1e-2
Epochs	75, stopped at 18
Batch Size	100
Weight Decay	1e-3
Optimization Algorithm	Adam
Early Stopping	5
n_d (Decision Width)	8
n_a (Attention Width)	8
n_steps (Number of Steps)	2
gamma (regularization for feature reuse)	1.5
Lambda_sparse (Sparsity Regularization)	5e-5

Evaluate Model on Clean Test Set

The main evaluation metric used for this project was accuracy which follows the formula correct predictions / all predictions. Furthermore, since the problem was classifying lung cancer, the metrics of precision, recall, and specificity were also reported on the test set.

Precision $= \frac{TP}{TP + FP}$ Proportion of true positive predictions among all positive predictions

Recall $= \frac{TP}{TP + FN}$ Proportion of actual positives correctly identified by the model

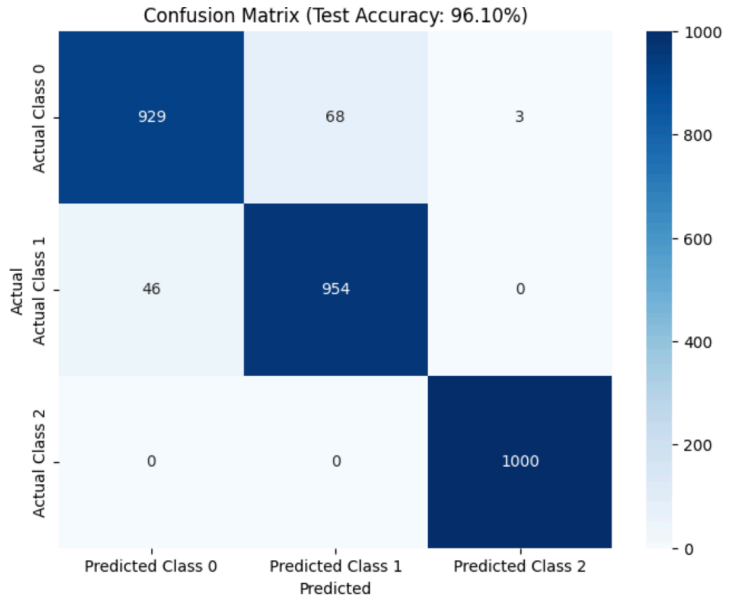
Specificity $= \frac{TN}{TN + FP}$ Proportion of actual negatives correctly identified by the model

Benchmark Model Evaluation:

Metric	**Dataset**	**Score**
Accuracy	Validation	95.62%
Accuracy	Test	96.10%

Benchmark Model Precision, Recall, Specificity for Each Class on the Test Set

Class	**Precision**	**Recall**	**Specificity**
lung_aca	0.95	0.95	0.95
lung_scc	0.93	0.93	0.93
lung_n	0.997	1.00	0.999



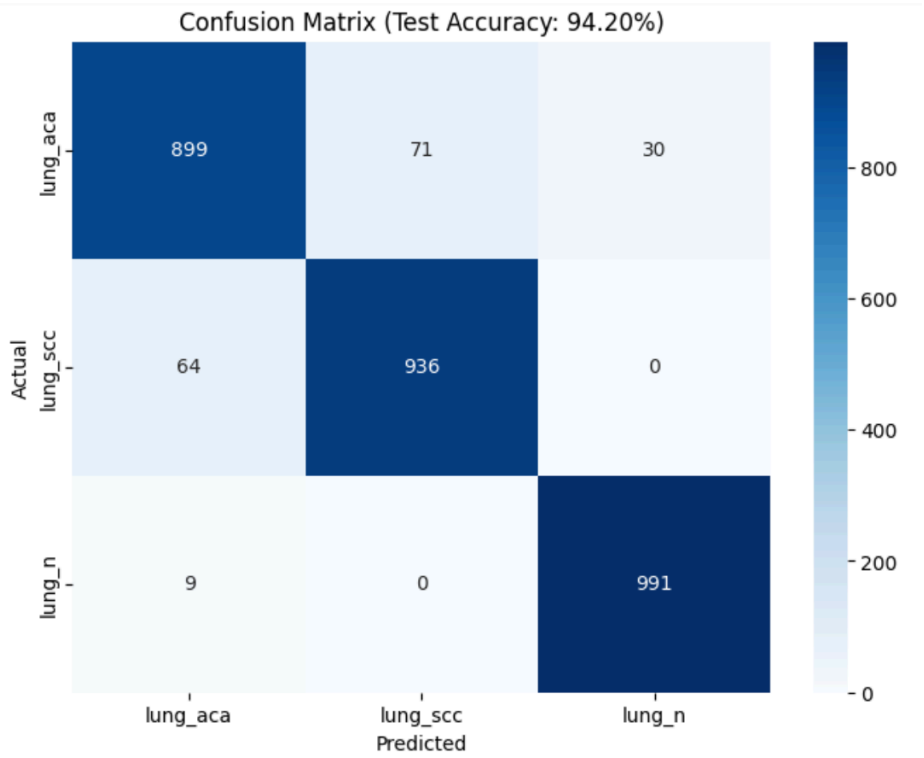
Where class 0 is lung adenocarcinoma, class 1 is lung squamous cell carcinoma, and class 2 is lung benign tissue

Tabnet Model Evaluation

Metric	**Dataset**	**Score**
Accuracy	Validation	93.97%
Accuracy	Test	94.20%

Tabnet Model Precision, Recall, Specificity for Each Class on the Test Set

Class	**Precision**	**Recall**	**Specificity**
lung_aca	0.92	0.93	0.97
lung_scc	0.93	0.94	0.96
lung_n	0.97	0.99	0.96



Differences Between Datasets

One reason we can expect the models to generalize well is because each split has the same proportion of classes. The dataset is not imbalanced which means the model will not learn to always predict one class. Furthermore, the images are from the same original set of 750 images. It is not clear what augmentations were done to create 5,000 images from 250 but a common practice with histopathological images is to segment the high resolution raw whole-slide image into subsections. If that is the case, the datasets may be from the same unsegmented image which removes variations in scan quality between different slides and ensures consistency in how the tissue features are captured across all samples. This consistency helps the model avoid learning noise related to scan discrepancies and improves its ability to generalize with this dataset. However, that does mean that model may fail to generalize well to a separate set of histopathological lung cancer images that were taken with different equipment.

Conclusion

Based on the results, the benchmark model and the model trained on the extracted features were able to generalize well to the held out test dataset with improvements in accuracy for both. Deep learning continues to be an effective aid in the domain of histopathology. Furthermore, both models performed exceptionally well in separating lung cancer cases from the control. This highlights the importance of balanced classes as some cancer classification models have been trained with an overrepresentation of negative samples. Most of the confusion appears to be between lung adenocarcinoma and lung squamous cell carcinoma. The benchmark model performed better in most metrics however a stand out metric for the TabNet classifier was an increased performance on the lung squamous cell carcinoma class with a higher precision, recall, and specificity. Other than the performance on that class, the loss in information was highlighted by a decrease in most metrics. The benchmark model was also not trained extensively while the best hyperparameters were searched for with the TabNet classifier. For future work, there could be more efficient ways to find the best hyperparameters for the TabNet classifier. The extracted features could still be proven to be more effective by using a traditional tabular model like boosted decision trees. Finally, the extracted features can also be used as additional variables to be trained on along with the images in a CNN for a possible boost in performance.

References

Al-Jabbar, M., Alshahrani, M., Senan, E. M., & Ahmed, I. A. (2023). Histopathological Analysis for Detecting Lung and Colon Cancer Malignancies Using Hybrid Systems with Fused Features. *Bioengineering* (Basel, Switzerland), 10(3), 383. https://doi.org/10.3390/bioengineering10030383

Arik, S. Ö., & Pfister, T. (2019, May). Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI conference on artificial intelligence*, 35(8), 6679–6687. https://doi.org/10.48550/arXiv.1908.07442

Borkowski, A. A., Bui, M. M., Thomas, L. B., Wilson, C. P., DeLand, L. A., & Mastorides, S. M. (2019). Lung and Colon Cancer Histopathological Image Dataset (LC25000). *arXiv*. https://arxiv.org/abs/1912.12142

Centers for Disease Control and Prevention (2024). U.S. Cancer Statistics Lung Cancer Stat Bite. *U.S. Department of Health and Human Services*. https://www.cdc.gov/united-states-cancer-statistics/publications/lung-cancer-stat-bite.html

Masamba, L. P. L., Mtonga, P. E., Kalilani Phiri, L., & Bychkovsky, B. L. (2017). Cancer Pathology Turnaround Time at Queen Elizabeth Central Hospital, the Largest Referral Center in Malawi for Oncology Patients. *Journal of global oncology*, 3(6), 734–739. https://doi.org/10.1200/JGO.2015.000257