VisTumor: Explainable Cancer Detection in Histopathology Images Using Grad-CAM++ and Saliency Mapping

Noah Hoang
University of Washington
nvhoang@cs.washington.edu

Huy Huynh
University of Washington
huyhuynh@cs.washington.edu

Abstract

Histopathology image analysis is critical for cancer diagnosis. Despite breakthroughs in image classification technology in deep learning models, they have yet to be adapted in medical applications because they often lack transparency in their decision-making processes. To address this limitation, we propose VisTumor, an explainable deep learning pipeline in whole slide histopathology images. Our system integrates several modern CNN architectures with Grad-CAM++ and saliency mapping techniques to provide visual explanations of model predictions. Using the CAMELYON16 dataset, we preprocess gigapixel WSIs into tissue-focused patches. VisTumor aims to bridge the gap between accuracy and explainability in medical AI and explores the potential of CNNs to be used as clinical decision support tools.

1. Introduction

1.1. Background & Motivation

The work from Liu et al. (2017), Selvaraju et al. (2017), and Tjoa & Guan (2020) provides the conceptual foundation for our work. We closely follow the experiment procedure from Liu et al. (2017), which demonstrated the efficacy of deep, pre-trained networks on histopathology images. Their approach has influenced our data pre-processing and model training process. To explore interpretability, we incorporated Grad-CAM++ (Chattopadhyay et al., 2017), a technique for generating visual explanations for model decision making. This research directly supports our goal of developing interpretable diagnostic tools. Lastly, the motivation behind our research is deeply rooted in the work done by Tjoa & Guan (2020) [1]. Their comprehensive investigation of the need for transparent and interpretable medical AI applications underscores the need for trustworthy and interpretable AI systems, especially in clinical contexts.

1.2. Related Work

Examination of histopathological whole slide images (WSIs) is a cornerstone in cancer diagnosis, however, it requires highly trained experts and is extremely labor-intensive. With the advent of convolutional neural networks (CNNs), we have seen significant progress that showcases how deep learning models can accurately detect metastases in lymph node tissue. However, most of the research done in this field has prioritized accuracy over explainability, which offers little insight into the reasons for a diagnosis. Grad-CAM++ [2] and saliency maps are popular methods for visualizing key points and can be effective for the evaluation of models, but few works have actually integrated these features.

1.3. Plan

We propose VisTumor, a deep learning pipeline for explainable tumor detection in histopathology images. Our system will test several CNN architectures trained on the CAMELYON16 [3] dataset in conjunction with Grad-CAM++ to generate heatmaps that highlight regions that contribute to the model's diagnosis. Our main goal is not only to evaluate classification accuracy, but also to test that our model correctly focuses on annotated tumor regions. Making a model accurate and interpretable is essential for trustworthy decision-making, especially with the gravity of a cancer diagnosis.

1. Pre-processing Data: Because of the resolution of WSIs, we must segment the full slides. Here, we will tune the patch and step size.

2. Choose Model Architecture: Our goal is to train a few models based on existing architectures that other researchers have already tuned. This allows us to verify which model architectures use the right features to classify.

3. Hyperparameter Tuning: Our goal is to use hyperparameters that have been found to already work in ex-

isting architectures, but we will verify and experiment if we can achieve better results.

4. Applying Grad-CAM++: After solidifying and tuning multiple model architectures, we will use Grad-CAM++ to create heatmaps of model attention.

5. Testing: We will evaluate the performance of classifications on our test patches, assessing the ability to generalize. Additionally, we will manually evaluate our saliency maps against professionally annotated slide images.

6. Compare Performance of Existing Models: Utilize existing models within the histopathology space and evaluate performance against our model in classification. Additionally, we will qualitatively compare saliency maps generated by our model versus existing models to assess interpretability.

1.4. Expected Challenges

1. Limited Storage and Compute Resources: As students, we are reliant on the use of Google Colab, which has some constraints. We have limited storage, which may present challenges regarding the dataset CAME-LYON16, since WSIs often exceed gigapixel resolutions, meaning we must load terabytes worth of data. Additionally, we have limited GPU hours, meaning hyperparameter searching must be carefully constructed to stay within usage limits.

2. Handling High Resolution WSIs: Whole slide images are extremely large, which makes them infeasible to input as raw pixels. We must divide each WSI into smaller patches to train our model, but this process limits spatial context.

3. Performance Trade-Off: Integrating Grad-CAM++ for explainability introduces extra computational overhead that may impact accuracy or inference speed. There will be a strong priority on balancing the benefits of visual explanations with the cost of overall performance and resource constraints.

4. Performance Verification: Although we can verify model diagnosis accuracy through testing, we cannot validate the accuracy of our heat maps or attention visualizations. Those would require analysis and validation from expert pathologists, which goes beyond the scope of this project/class.

1.5. Expected Outcome

Our project aims to develop a deep learning model that not only accurately detects cancerous regions within WSIs of lymph nodes, but also increments on the interpretability through visual explanations. We expect our model to classify whether an individual image patch contains cancerous tissues. Additionally, by using saliency techniques, we aim to generate heatmaps that display the most influential regions in influencing the model's predictions as a means of transparency. The hope is that such heatmaps are used in aiding in clinical interpretations of WSIs. If our model performance is favorable, we may seek expert feedback from pathologists to validate the relevance of said heatmaps.

The project's pipeline is foundational for other deep learning projects with partially annotated data, such as the WSIs being used. We hope that the performance and interpretability of our proposed model can justify a potential clinical use.

2. Methods

2.1. Data

The dataset being used for this project is CAME-LYON16. This dataset is a benchmark dataset designed for the development and evaluation of deep learning models on cancer detection of metastases in histopathological whole slide images (WSIs) of lymph node tissue. This is a small dataset consisting of 400 WSIs split into 270 training examples and 130 testing examples. Each slide is taken from extremely high-resolution microscope imaging (some over 100,000 x 100,000 pixels). This dataset presents some unique challenges because of its high resolution, but it serves as an excellent test bed for our project, which focuses on accurate and interpretable cancer detection.

2.2. Preprocessing

2.2.1 Patch Extraction

To focus on more relevant anatomy within our dataset, we began by applying automated segmentation masks on the dataset to locate regions of interest.

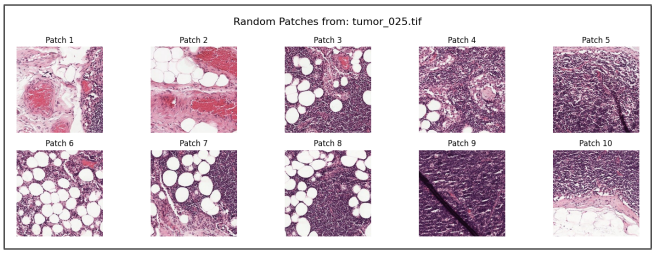

Figure 1. Patches taken during preprocessing without grey patches (non-tissue).

The first step in this process was sampling patches of 512 x 512 pixels. Although our models require input sizes of 224 x 224, we decided to use larger patches and downsample to capture greater spatial context. Additionally, because

each slide contains a large amount of whitespace, which would have created noise in our training data, we decided to threshold each image patch to ensure it contains enough viable tissue (60%).

While we could split healthy slides using thresholding alone, we had to follow a different approach for splitting slides with cancer. WSIs with cancer are primarily composed of healthy cells, except for select regions. To account for this, we had to reference annotated masks from CAMELYON16. Our strategy for sampling patches was to scan both the WSI and the annotated mask simultaneously, and only sample patches composed of at least 5% cancer tissue (a heuristic we found online to ensure our models can detect even small instances of metastases).

For our experiment, we divided ≈60 WSIs, around a quarter of the full CAMELYON16 training dataset. We extracted ≈20,000 image patches, evenly split between healthy and metastasized examples (Figure 1). Since we used a limited portion of the dataset, we decided to leverage data augmentation. We applied horizontal/vertical flips and rotations to enforce position invariance. We also applied color/hue variations to account for samples coming from different lab setups and staining variation. Finally, we normalized our data to ensure consistent distributions during training.

2.2.2 Train/Test Data Splitting

Because the test set is hidden by the Grand Challenge platform, we used a portion of our training patches as a test set. We split 90% of our patches to use as training data, and 10% to use as testing data. This left us with 18,000 examples to use for training and 2,000 examples reserved for testing benchmarks. It is important to note that we are not training on all 20,000 data points and treating those 2,000 samples from the training set as a true test set.

2.3. Model Architectures

In our project, we will be evaluating the performance and interpretability of 3 modern CNN architectures: ResNet50 [4], DenseNet201 [5], and EfficientNetB7 [6]. While these models are robust and widely used in many image classification tasks, we will be leveraging these existing architectures to see if CNNs are strong candidates for future medical imaging applications, in particular.

ResNet introduces residual skip connections that mitigate vanishing gradients and allow it to ignore layers that do not contribute to better results. DenseNet leverages dense connection patterns between layers that allow for better feature reuse and gradient flow through the network. EfficientNet utilizes a compound scaling (depth, width, resolution) method to increase the accuracy and efficiency of CNNs. All of these are viable architectures for our experiment, but we aim to explore whether CNN pipelines are viable for creating explainable models.

Because of our limitations in computing and storage, we will follow the approach used by Liu et. al. in Google's cancer detection pipeline [7]. Each of our models will be pre-trained on ImageNet, and we will augment and train the last layer on CAMELYON16 to fit our application. The final fully connected layer will be the only layer trained on the dataset, and will be adjusted to produce binary classification outputs as opposed to the 1000 ImageNet classes. We expect these models to perform well enough to assess how well CNN architectures adapt to histopathological cancer detection and how effectively they highlight relevant features. The purpose is not to create new benchmarks or SOTA models.

2.4. Grad-CAM++

The method of explainability we have chosen for our pipeline is Grad-CAM++ [2]. With Grad-CAM++, we aim to create localization maps by computing the gradient of the final convolutional feature map. The resulting heatmaps will highlight regions that strongly influence the model's prediction. As we explore model explainability, this is a useful technique to qualitatively measure model interpretability. While Grad-CAM++ is great for this experiment, there are several key drawbacks we must take note of. The heat maps may potentially highlight clinically irrelevant regions and may not be able to capture low-resolution or fine-grained features. This is why having strong quantitative results as well as qualitative assessments is important for explanation fidelity.

3. Experiment Outline

3.1. Training Setup

We used the same setup for each of our architectures. Each model measures a Cross-Entropy Loss and uses the Adam optimizer. For the sake of consistency and simplicity, we did not explore optimizing different loss functions and optimizers. Each model was pre-trained on ImageNet, with only the final classification layer being unfrozen for training. We used a batch size of 32 and ran for 5 epochs. After preliminary experiments, we found that loss and training accuracy began to plateau around 5 epochs (Figure 2). While we could have trained for longer for marginal performance gains, we decided the extra performance would not be relevant for visual interpretability.

3.2. Hardware and Environment

Our code base was in Google Colab, and we split the workload between hosted runtimes and a local GPU. Because we did not have the local disk space to store 100GB of WSI data, we stored the raw data in Google Drive. We

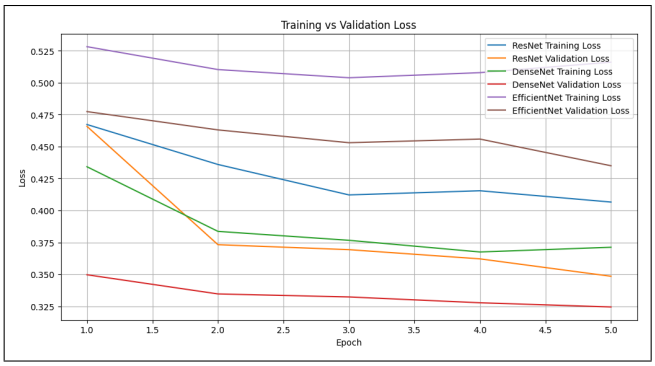

Figure 2. Training and Validation Loss Curves

utilized a Google hosted runtime for image pre-processing. After pre-processing the raw data, we downloaded the image patches (which were significantly smaller in storage size) on our local machine, where we could leverage a local GPU for training. Because we did not have access to an NVIDIA GPU compatible with CUDA, we used DirectML in an Anaconda environment. While a Google runtime may have been faster/more optimized, our local GPU gave us unlimited usage and flexibility.

3.3. Evaluation Metrics

For quantitative evaluation, we measured test accuracy, false-positive, and false-negative classification rates. Since this is a pipeline designed for clinical use, it's important to determine what kind of misclassifications each model makes during inference. As for qualitative evaluation, we compared Grad-CAM++ saliency maps to the ground truth annotation to see if the areas showed similarity.

3.4. Hyperparameter Selection

Since our goal is not to create a state-of-the-art model and push the boundary for benchmarks, we did not perform grid or random search and cross-validation for hyperparameters. Instead, we chose empirically effective hyperparameters. As outlined earlier, we trained on a batch size of 32 over 5 epochs, with Cross-Entropy Loss and the Adam optimizer. We used a learning rate of $1 \times 10^{-}3$, which provided stable convergence for all models.

4. Results

4.1. Performance Comparison and Benchmarks

After training all models, we tested each architecture on the partitioned test set consisting of malignant and normal patches. Our primary quantitative metrics were accuracy, false-positive rate (FPR), and false-negative rate (FNR), given the importance of minimizing misdiagnoses in clinical contexts.

All models achieved relatively high performance on the test set, with DenseNet achieving the best overall performance. DenseNet achieved the highest overall accuracy of 86%, with a false-positive rate of 16%, and a false-negative rate of 12%. However, this was marginally better than the other architectures, which all had relatively similar accuracies, FPRs, and FNRs.

Given that we did not tune our architectures for maximal performance and trained on limited data, the results show that the models have sufficient performance to support visual interpretations using Grad-CAM++. There is potential for model improvement with increased time and scale, but it was not our primary focus.

Architectures	Accuracy	FPR	FNR	Recall
ResNet50	83.58%	16.15%	16.70%	83.30%
DenseNet201	86.32%	15.69%	11.60%	88.40%
EfficientNetB7	80.80%	11.68%	26.98%	73.02%

4.2. Grad-CAM++ Analysis

We applied Grad-CAM++ to visualize the attention of each model during inference to understand whether CNNs are focusing on medically relevant regions when making predictions. We compared our generated saliency maps against the pixel-level annotations from the CAMELYON16 dataset. Since there are no annotations for the normal WSIs, our analysis is will be based primarily on true positive and false negative classifications of malignant patches.

During our experiment, we found that ResNet had the most promising saliency maps and highlighted regions spatially consistent with the ground truth annotations. In the true positive case we analyzed, ResNet's attention was distributed over areas closely matching the annotated malignant regions (Figure 3). This is a strong indication that the model is not only making the correct predictions, but is also learning relevant clinical features of pathology slides.

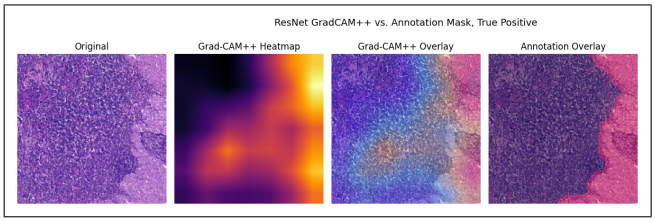

Figure 3. True positive saliency map and ground truth annotation comparison.

In the false negative case, we also saw that the model's attention closely followed the annotated region (Figure 4). Despite this, the model still misclassified the patch as benign. It's hard to pinpoint the exact reason for this behavior, but it shows promise that scaling this architecture

could yield improved performance while maintaining interpretability.

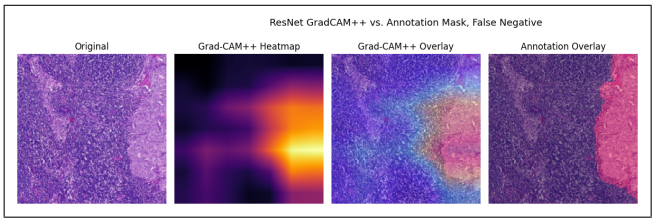

Figure 4. False negative saliency map and ground truth annotation comparison.

In contrast, despite DenseNet's higher accuracy, both DenseNet and EfficientNet presented saliency maps with less similarity to the annotations (Figure 5). Their attention seemed to be less focused on relevant regions, suggesting weaker localization ability. This may be due to a variety of reasons including potential underfitting because of layer freezing or ineffective feature hierarchies for capturing histopathology imaging specifically.

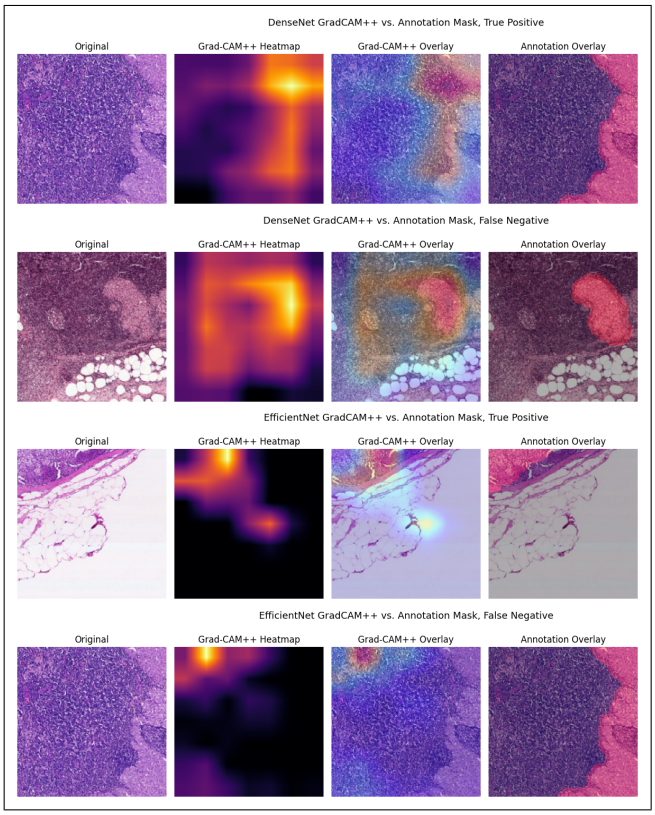

Figure 5. Saliency maps and ground truth annotations for DenseNet and EfficientNet.

While our qualitative analysis was limited to one true positive and one false negative per model, the results still offer valuable insight into how different architectures learn different spatial features in histopathological patches. Grad-CAM++ highlighted ResNet's ability to recognize medically relevant regions which enhances interpretability and confidence. It also simultaneously exposed limitations in DenseNet and EfficientNet. Even with high accuracy benchmarks, these architectures failed to consistently localize regions significant to diagnosis. Our results emphasize and reiterate how quantitative metrics alone are insufficient for robust medical AI systems.

5. Discussion

5.1. Experiment Summary

This experiment presents an exploration of explainable deep learning models for histopathology image classification, focusing on using Grad-CAM visualizations as an interpretability tool for tumor annotations. Our qualitative analysis reveals that saliency alignment varies significantly across architectures.

ResNet demonstrated the strongest correlation between ground truth annotations and saliency maps, in both true positive and false negative cases. While this alone is not enough for industry use, it's an encouraging signal that CNNs like ResNet may be learning clinical features, even when misclassifying. Contrastingly, DenseNet and EfficientNet produced misaligned and inconsistent saliency maps compared to annotated regions, even though DenseNet had the highest predictive performance. This accentuates the notion that higher performance does not imply better interpretability.

Our findings maintain the conclusion that quantitative metrics alone are not sufficient when evaluating medical AI systems. Models must perform well but also provide transparent justifications for their decisions. Grad-CAM++ was one way for us to expose architectural differences in attention behavior that would have remained hidden under standard metrics like accuracy.

5.2. Broader Impact

The advent of this type of technology could completely reshape the practices of imaging and pathology across the world before the end of the decade. AI systems that can reliably flag suspicious regions and support human diagnoses could drastically improve efficiency and accuracy. These tools could democratize access to high-quality care in regions where trained pathologists are scarce and reduce the pressure on overworked clinicians.

However, this also comes with the possibility of a negative impact. As AI reduces the time, labor, and expertise required for imaging tasks, it may also diminish the demand for human pathologists. While this shift might benefit the greater good and improve efficiency, it highlights the re-

sponsibility engineers hold to build systems that support rather than replace. The ideal role for AI in pathology is not to replace, but rather to assist and reduce fatigue-based errors.

6. Future Works

6.1. Pixel-Level Segmentation

While our current approach is patch-level classification with Grad-CAM++ or interpretability, real-world diagnoses require much more precise localization of tumors than our saliency maps provide. Future work could involve pixel-level segmentation using architectures such as U-Net, which learn to map each pixel to a class label and output explicit prediction masks, unlike Grad-CAM++, which produced attention post hoc. These would offer much more interpretable and detailed information behind model predictions and clinical utility.

6.2. Tumor Grading and Cancer Severity Estimation

Pathology and oncology are much more complex than simply detecting the presence of cancer or not. The severity and progression of a tumor has important implications for the kind of care and treatments a patient will receive. A direction to explore is extending our models to predict tumor grade or stage, or identifying specific subtypes. This would provide clinicians with much richer diagnostic insight and move medical AI closer to real-world applicability.

6.3. Clinical Integration and Human Verification

As engineers, are scope is limited to model training and performance. We may have a general understanding of the science behind pathology, but we certainly aren't well informed about the day-to-day work of a pathologist. Future work must explore human evaluation to assess how well visual explanations or segmentation maps would support decision-making and workflow in the field. These experiments would give valuable direction on what kind of tools pathologists prioritize and whether our current approaches add value in practical settings.

References

[1] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," 2021.

[2] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," 2018.

[3] B. E. Bejnordi, M. Veta, P. J. van Diest, *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," 2017.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016.

[5] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2017.

[6] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2019.

[7] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado, J. D. Hipp, L. Peng, and M. C. Stumpe, "Detecting cancer metastases on gigapixel pathology images," 2017.