CANCER GENOMICS CLOUD
SEVEN BRIDGES

VELSERA

Cloud-based machine learning for enhanced tumor classification in cancer genomics: an end-to-end solution for whole slide imaging data

#6730

Jovana Babic, Nevena Nikolic, Milan Kovacevic, Tariq Khoyratty, Zelia Worman
Velsera Inc.

Introduction

The digitization of histological slides into high-resolution Whole Slide Images (WSIs) has transformed pathology workflows, enabling automated and AI-driven analysis for tasks such as tumor classification, subclassification, and grading. Machine learning (ML) models can now extract diagnostic insights with improved accuracy, reproducibility, and efficiency.

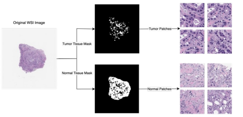
To make advanced ML more accessible, Velsera has developed a **comprehensive solution for classifying WSIs by disease or tumor subtype**, based on the morphological characteristics of the tissue. Building on the Cancer Research Data Commons' (CRDC) cloud infrastructure, we take advantage of hosted data from the Imaging Data Commons (IDC) and Human Tumor Atlas Network (HTAN) paired with the computational resources available in the **Seven Bridges - Cancer Genomics Cloud (SB-CGC)**, powered by Velsera, to host a reproducible WSI ML solution. This solution includes data preparation and preprocessing, model training, evaluation, and predictions, within an interactive analysis environment. The analysis harnesses the computational capabilities of the SB-CGC platform to efficiently process large-scale datasets. To facilitate data preparation, we integrated tools previously developed on the SB-CGC to extract regions of interest (ROIs) from WSIs, enabling the creation of expanded datasets by generating tumor tissue patches for model training.

Preprocessing notebook

The **Preprocessing Notebook** offers a fast, easy-to-use solution for transforming raw whole-slide images (WSIs) into high-quality datasets for **further downstream analysis**, such as classification or tumor grading. Built on Velsera's WSI ROI Extraction Workflow, it integrates segmentation, patch extraction, stain normalization, and data augmentation–all in a single automated run.

By leveraging the computing capabilities of SB-CGC this notebook enables users to:
- Automatically detect and segment tumor regions in liver, colon, or breast slides
- Extract high-quality tumor and normal tissue patches
- Normalize staining across slides for consistent visual representation
- Augment data to increase variability and improve model robustness

Beyond simplifying WSI processing, this notebook offers an automated, scalable solution for analyzing entire datasets, while allowing easy tuning of input parameters that influence segmentation quality and ensure the extracted patches are suitable for downstream tasks like classification.



Classification notebook

This notebook implements a deep learning pipeline for image patch classification, adaptable to any image classification task. It is demonstrated here as an end-to-end solution for tumor subtype classification from histopathology patches. Patches can be generated using the accompanying preprocessing notebook or sourced from any external workflow. The main steps are outlined below:

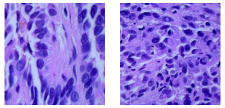
- **Input Pipeline**
A custom data pipeline loads images from path-label pairs, resizes and normalizes them, applies random augmentations, and prepares the data for efficient training through shuffling, batching, and prefetching.
- **Model Definition and Training**
Three pretrained architectures–InceptionV3, ResNet101V2, and DenseNet121–are used with custom classification heads tailored for this task. Early stopping and learning rate scheduling are applied to improve generalization and training efficiency. Performance is monitored using MLflow and tracked through accuracy and loss metrics.
- **Ensemble via Soft Voting**
A soft voting ensemble combines predictions from all three models, averaging class probabilities to produce final outputs. This approach improves robustness and leverages complementary strengths of individual models.
- **Evaluation and Visualization**
Model performance is evaluated using confusion matrices, precision-recall curves, and ROC curves with AUC scores to provide insight into class-wise performance and overall model reliability.

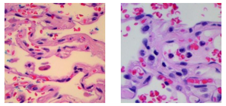
Results

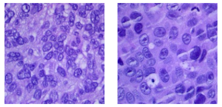
Kaggle Lung Histology Dataset

To validate the classification pipeline, we first applied the model to a publicly available Kaggle dataset consisting of 15,000 pre-processed histopathological image patches (768×768 px) representing three lung tissue classes: benign, adenocarcinoma, and squamous cell carcinoma (5,000 images per class).

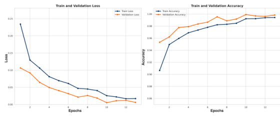
Example histopathological image patches from each lung tissue class used in model training: benign, adenocarcinoma, and squamous cell carcinoma.


Lung adenocarcinoma tissue

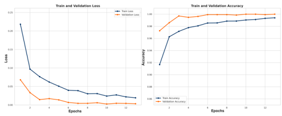

Lung benign tissue


Lung squamous cell carcinoma tissue

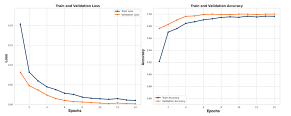
The models were trained on histopathological image patches using separate training, validation, and test subsets. InceptionV3, DenseNet121, and ResNet101V2 architectures were each trained independently. Their performance during training is reflected in the accuracy and loss trends shown below.



InceptionV3 model performance

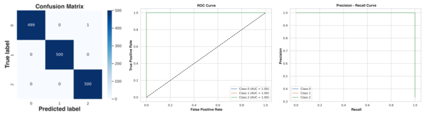


ResNet101V2 model performance



DenseNet121 model performance

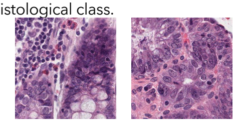
To improve prediction robustness, outputs from the three individual models were aggregated using soft voting, where class probabilities were averaged to form a consensus. This ensemble model was evaluated on a held-out test set consisting of unseen image patches. Its performance is summarized using a confusion matrix, ROC curves, and precision-recall plots.

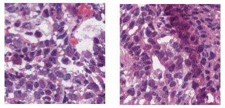


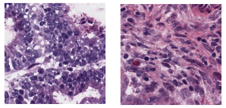
Class 0 – adenocarcinoma; Class 1 – benign; Class 2 – squamous cell

TCGA COAD colon dataset

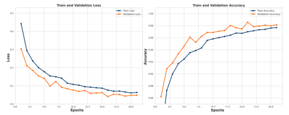
Using a custom preprocessing notebook, we generated image patches from TCGA-COAD whole slide images spanning three histological classes: *adenomas and adenocarcinomas*, *cystic, mucinous, and serous neoplasms*, and *complex epithelial neoplasms*. Patches sized 500×500 px were created for the first two classes, while 299×299 px patches were extracted from the limited complex epithelial cases. Data augmentation was applied to balance class sizes, resulting in 7000, 7126, and 6902 patches per class, respectively. Below, we showcase representative patch examples from each histological class.


Colon cystic, mucinous and serous neoplasms

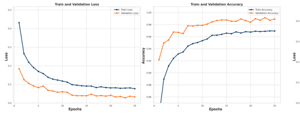

Colon complex epithelial neoplasms


Colon adenomas and adenocarcinomas

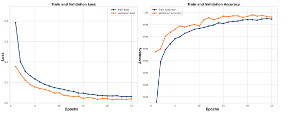
The same model architectures–InceptionV3, DenseNet121, and ResNet101V2–were used, with training performed over more epochs and a longer early stopping patience. Since this dataset was generated from raw WSIs, training required more careful tuning. The resulting accuracy and loss trends are shown below.



InceptionV3 model performance

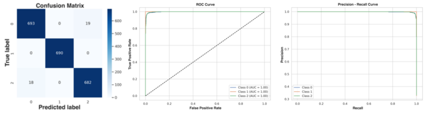


ResNet101V2 model performance



DenseNet121 model performance

While all models performed well individually, their individual strengths varied slightly. Soft voting helped balance these differences by averaging predictions into a more robust ensemble. Though results were slightly lower than with pre-curated dataset, performance remained strong on the held-out test set, as shown by confusion matrix, ROC, and precision-recall curves.



Class 0 – cystic, mucinous and serous neoplasms; Class 1 – complex epithelial neoplasms; Class 2 – adenomas and adenocarcinomas

Conclusion

This work demonstrates the power of coupling AI-ready harmonized imaging data with advanced ML techniques, showcasing an end-to-end ML workflow on the SB-CGC. By leveraging best practices in data preparation, model development, and consensus-driven accuracy improvement, our solution enhances accessibility and reproducibility in histopathology analysis. Reducing technical barriers enables researchers to apply these workflows to their own datasets, driving deeper insights into cancer diagnostics and treatment.

To achieve strong classification performance on histopathology images–even when access to large annotated datasets is limited–we leveraged state-of-the-art pretrained CNN architectures with custom classification heads. These models were fine-tuned with carefully selected hyperparameters, using early stopping and learning rate scheduling to support stable convergence and strong generalization.

When starting analysis from raw WSIs, it's essential to extract high-quality image patches that clearly capture tumor regions. Our preprocessing notebook supports this by allowing flexible tuning of segmentation and patch extraction settings. For datasets with class imbalance, data augmentation can help create a more balanced and representative training set–ultimately contributing to more reliable model outcomes.

References

1. Borkowski AA, Thomas LB, Wilson CP, DeLand LA, Mastorides SM. Lung and Colon Cancer Histopathological Image Dataset (LC25000). arXiv:1912.12142v1 [eess.IV], 2019
2. Khened, M., Kori, A., Rajkumar, H. *et al.* A generalized deep learning framework for whole-slide image segmentation and analysis. *Sci Rep* **11**, 11579 (2021).
3. Ahmed, S., Shaikh, A., Alshahrani, H., Alghamdi, A., Alrizq, M., Baber, J., & Bakhtyar, M. (2021). Transfer Learning Approach for Classification of Histopathology Whole Slide Images. *Sensors*, *21*(16), 5361.

Contact

Tariq Khoyratty, DPhil
tariq.khoyratty@velsera.com

Zelia Worman, PhD
zelia.worman@velsera.com