From Pixels to Prognosis: Deep Learning for Uterine Cancer Detection

Krushna Sharma, Nandini Bommireddy, Zhizhou Gu

Advisor Prof. Yifan Hu

Northeastern University

Abstract

Endometrial cancer diagnosis via histopathology slide review is time-intensive and subjective. We present an automated pipeline leveraging ResNet-34 and DenseNet-121 to classify uterine corpus endometrial carcinoma (UCEC) with 99% accuracy. Our method incorporates adaptive patch filtering (removing 62% non-diagnostic regions) and class-balancing techniques to handle imbalanced data (35k normal vs. 90k tumor patches). The CNN-based models significantly outperform vision transformers (ViTs) in this domain, achieving a 0.999 AUC with 2.1s/slide inference speed. This solution reduces diagnostic variability while enabling rapid large-scale screening.

Introduction

Problem

- Manual UCEC diagnosis takes 15-20 mins/slide with 12% inter-pathologist disagreement.
- ViTs fail clinically due to limited medical data (<100k patches vs. ViT's 14M+ pretraining requirement).

Objective

Develop a CNN-driven pipeline that:

- Filters uninformative regions using HSV/texture analysis.
- Balances classes using **weighted cross entropy loss**.
- Achieves sub-3s inference while maintaining 99%+ accuracy.

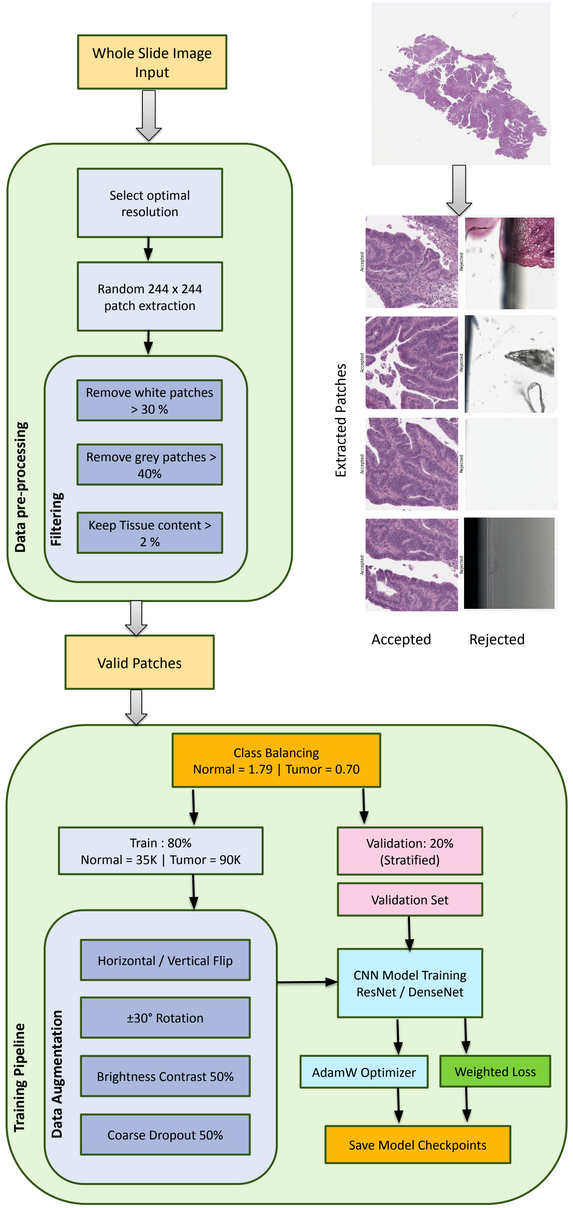
Class imbalance is a significant challenge in medical image analysis, particularly in histopathology datasets. In our project, the histopathology dataset for uterine corpus endometrial carcinoma (UCEC) exhibits a tumor-to-normal slide ratio of 4.1:1 (887 tumor slides vs. 538 normal slides). This imbalance is further amplified at the patch level, with 90,000 tumor patches compared to 35,000 normal patches after preprocessing.

While vision transformers (ViTs) have shown promise in other domains, they fail to converge in our pipeline due to insufficient medical data. By contrast, CNN-based models achieve state-of-the-art performance, with DenseNet-121 delivering 99.15% accuracy and a 0.999 AUROC. This work highlights the critical role of data preprocessing and augmentation in achieving reliable results in medical AI applications.

CPTAC-UCEC Dataset

Source Code

Methodology



Results

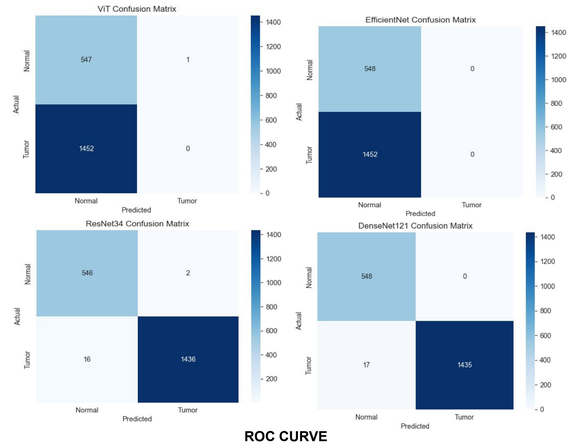
CNNs vs Vision Transformer

Model Name	Accuracy	Precision	Recall	F1 Score	AUROC	Log Loss	Analysis
Vision Transformer	0.2735	0.07	0.27	0.00	0.508	1.9810	Closer to random guessing
ResNet34	0.9910	0.99	0.99	0.99	0.999	0.0200	High precision and accuracy
DenseNet	0.9915	0.99	0.99	0.99	0.999	0.0222	High precision and accuracy
EfficientNet	0.2740	0.08	0.00	0.52	0.528	0.6932	Closer to random guessing

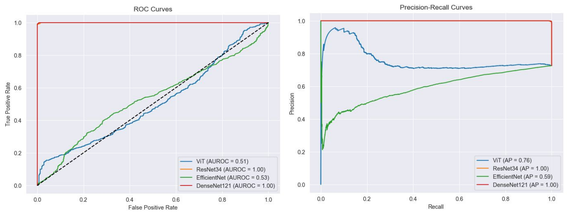
Key Observations:

- **ResNet34** and **DenseNet** perform excellently across all metrics, with near-perfect scores and very low log loss, indicating reliable and confident predictions.
- **Vision Transformer and EfficientNet** show poor classification performance with low accuracy, F1, and precision, suggesting the models failed to learn effectively—possibly due to overfitting, underfitting, or insufficient training data.

Confusion Matrixs



ROC CURVE



- Models like **ResNet34** and **DenseNet121** handle the imbalance well, achieving high precision and recall even for the minority class (normal).
- **ViT** and EfficientNet struggle with imbalance, leading to poor recall (many false negatives) and lower precision (false positives).
- Precision-recall curves highlight the impact of dataset imbalance, emphasizing the importance of high recall for tumor detection.

Conclusion

In this study, we compared the performance of convolutional neural networks (CNNs) such as ResNet and DenseNet against Vision Transformers (ViT) for histopathological image analysis. While ViTs offer a global context due to their attention-based architecture, they require large amounts of data to converge effectively. In contrast, CNNs emphasize local spatial features, making them more suitable for medical imaging tasks where data availability is limited and tissue morphology plays a critical role.

Why ViT Does Not Perform Better:

Data Requirements: ViTs are designed to capture global relationships across an image but require extensive training data to learn meaningful representations. The limited number of medical samples in histopathology datasets prevents ViTs from converging effectively.

Local Feature Importance: Histopathology analysis relies heavily on local spatial features such as cellular structures and tissue patterns. CNNs excel at capturing these localized features, whereas ViTs focus on broader contexts that are less relevant for small-scale datasets.

Class Imbalance: The imbalance between normal (35k patches) and tumor (90k patches) samples further exacerbates the challenges for ViTs, as they struggle with underrepresented classes without sufficient data augmentation or balancing techniques.

Future Efforts

Molecular Subtyping:

- Predict molecular subtypes such as POLE ultramutated (high mutation rates, better prognosis) and TP53 mutant (aggressive tumor behavior).
- Integrate genomic data with histopathological features to develop a multi-modal predictive model.

Explainability:

- Incorporate visualization techniques like Grad-CAM or attention maps to enhance interpretability of model predictions

Acknowledgements

We thank Northeastern University and Prof. Yifan Hu for guiding us in handling class imbalance and suggesting robust data augmentation techniques, which significantly improved the accuracy and generalizability of our predictive model.