Synthetic Data in Machine Learning for Pathology

Richard J Chen1,2,3,4, Ming Y Lu1,3,4, Tiffany Y Chen1,3, Drew F K Williamson1,3, Faisal Mahmood1,2,3,4,
1Department of Pathology, Brigham and Women’s Hospital and Harvard Medical School, Boston, Massachusetts
2Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts
3Cancer Program, Broad Institute of Harvard and MIT, Cambridge, Massachusetts
4Cancer Data Science Program, Dana-Farber Cancer Institute, Boston, Massachusetts

Introduction

Large and heterogeneous datasets are necessary to develop and refine best practices in evidence-based medicine involving AI. To overcome the paucity of annotated medical data in real-world settings, we used Generative adversarial networks (GANs) to generate photorealistic histopathological images of Renal Cell Carcinoma (RCC) pathology as a synthetic data augmentation strategy for improving model generalization.

Methods

- Curation of 10,000 real images of each RCC subtype from TCGA
- Synthetic data generation was performed using the Conditional Progressive Growing GAN (PG-GAN) network architecture
- We used PG-GAN to generate 10,000 synthetic images of clear cell, papillary, and chromophobe renal cell carcinoma pathology, followed by Visual Turing Tests and interpolation to explore the latent space
- We conducted an ablation study comparing the performance of a Convolutional Neural Network (CNN) trained with only real data in RCC subtyping, with that of a CNN trained with both real and synthetic data generation

Latent Space Interpolation

Results

RCC Subtyping AUC

On a held-out, external cohort of 1,661 patches from BWH, we demonstrate that CNNs trained with synthetic data as data augmentation outperform CNNs trained with only real data on macro-averaged AUC performance. The synthetic images generated by the GAN finely mimic the characteristic thin-walled ‘chicken wire’ vasculature of the clear-cell carcinoma subtype and the unique features of the other two subtypes.

Conclusion

By closely mimicking real-world observational data, synthetic data could transform interoperability standards in the sharing of health data, improving reproducibility. In lieu of revealing actual patient data, synthetic datasets that accurately capture the original distribution of the data would substantially lessen patient privacy concerns and could be freely shared. Training generative models with multi-institutional datasets that capture a larger diversity of clinical phenotypes and outcomes can improve model generalization and reduce biases, and larger training datasets naturally lead to more robust algorithms.