Generative Deep Learning-Based Data Augmentation Techniques via Adversarial and Diffusive Models for Enhanced Squamous Cell Carcinoma Histopathological Scan Diagnosis

Shah, Ashank (School: Sunset High School)

Cutaneous Squamous Cell Carcinoma (SCC) is the second most common form of skin cancer. Moreover, accurate, early diagnosis for pre-metastatic SCC is imperative; post-metastasis patient survival rate plummets from 99% to <50%. However, clinician-based interpretation of whole slide images is often inefficient and extensively time-consuming. Prior research has illustrated the efficacy of deep-learning approaches for autonomous classification of squamous cells for efficient SCC diagnosis; however, false negatives remain prominent due to bias against underrepresented, scarce malignant data. To circumvent the ramifications of imbalanced classes within datasets, research employed data augmentation techniques, but most fail to introduce new seed information and lead to model overfitting. This study developed two extensively fine-tuned novel generative machine learning architectures for SCC data augmentation and eventual classification enhancement: SquaGAN, a Conditional Generative Adversarial Network, and SquaDIFF, a Denoising Diffusion Probabilistic Model. Both models were trained on a SCC whole slide image patch dataset with severe malignant underrepresentation. To evaluate the success of the techniques, a VGG16 baseline metric classifier was developed and trained on three sets: the raw non-augmented dataset, a dataset amplified with existing data augmentation techniques, and finally a dataset amplified with the proposed generative data augmentation techniques. Both SquaGAN and SquaDIFF increased performance, with class-specific model accuracy jumps of +21% and +27%, respectively. Both models additionally demonstrated efficacy in a truncated evaluation, outperforming all existing counterparts. These results are promising in feasible autonomous diagnosis approaches in clinics.

Awards Won:

Third Award of $1,000