Artificial intelligence applications in oncology to augment data and support decisions

Summary

The application of Artificial Intelligence (AI) to solve medical tasks has seen remarkable progress in the recent years, yet a number of challenges persist concerning data augmentation and decision-making. In terms of data augmentation, the availability of large, high-quality unimodal and multi-modal clinical datasets is essential for deep learning but is often a significant obstacle. While traditional data augmentation methods such as flipping, rotating, zooming, etc. can increase the size of the dataset, they are insufficient to address data imbalance issues.

Generative Adversarial Networks (GANs) have shown great promise as a generative model for clinical data augmentation, capable of learning and imitating the distribution and features (e.g., shape, boundary, strength, texture, etc.) of each image to generate synthetic, but realistic data that does not exist in the original dataset. This can be used to generate simulated clinical data for training deep learning models. With regards to decision-making, the diversity of formats of medical data presents considerable challenges from macro image data to micro biomarkers. Different data sources have their own specific type and structure, which results in significant discrepancies in the format and content of the data. However, faced with different data and diseases, it is still necessary to design different models to deal with specific complex problems. In addition, the contribution of an AI-based diagnostic or prognostic model to clinical decision making still needs to be evaluated.

In this thesis, advances are made in AI applications in oncology for data augmentation and decision-making.

For the data augmentation, I mainly focus on the GAN model to generate clinical data as a data augmentation method. To further explore this, I tried to design an SRGAN (Super-Resolution GAN) for Ground Glass Nodule (GGN) generation, with two discriminators to capture global and local details during training processing. Qualitative visual Turing test (VTT) with clinicians and quantitative radiomics experiments were conducted to assess the similarity between GAN-generated and real lung lesions.

The results of these experiments indicate that GAN-generated data could be used to train and test junior doctors, particularly in hospitals without large datasets, established Picture Archiving and Communication Systems (PACS) or for privacy-preserving synthetic open datasets for research purposes. I also proposed a self-attention cycleGAN model to generate synthetic pneumonia from Radiation Pneumonitis (RP) and COVID-19 radiography datasets as a data augmentation method to solve data imbalance problems and create synthetic data with clearer details. A classification model was trained with generated data, and it had an excellent performance on several classification metrics indicating that the method can significantly improve the accuracy of the classification model compared with traditional data augmentation methods. Therefore, the potential of the proposed method as a general data augmentation tool to assist in overcoming the sample imbalance problem in medical image datasets, is demonstrated.

In terms of decision-making, AI diagnosis models with different modalities of medical data source were developed. Bootstrap samples and a prospective validation set were used to validate a model which demonstrated that radiomic features extracted from the CT and the radiotherapy dose matrices could assist doctors in predicting of Radiation Pneumonitis (RP). Furthermore, a comprehensive nomogram was built to support clinical decision-making and personalized treatment. A decision tree model was then developed to identify the correlation between preoperative circulating leukocytes (such as MO/LY, differentiation status, CA125 level, NE, ascites cytology, LY%, and age) and ovarian cancer survival.

Finally, machine learning models were built and evaluated to provide a rapid diagnosis prediction and assist clinicians in providing effective treatment advice for endometrial carcinoma, with comparisons between clinicians with and without the assistance of models demonstrating the contribution of AI-based models.

Overall, this thesis has confirmed the hypothesis that AI-based techniques can yield high-caliber medical data suitable for instruction of machine learning and deep learning models, with the ultimate goal of improving their classification performance.

Furthermore, AI decision models can leverage different sources of medical data for diagnoses and provide support to clinicians in the field of oncology.