論　文　の　要　旨

題　目　　Research on Adversarial Attacks and Enhancing Defense Capabilities in Medical Image Deep Learning Systems

（医療画像ディープラーニングシステムにおける敵対的攻撃と防御能力の強化に関する研究）

氏　名　　LI　YANG

With the widespread application of deep learning technology in the medical field, it has demonstrated excellent performance in areas such as medical image analysis, drug development, and clinical decision support. Deep learning systems used for medical image analysis can learn effective feature representations from large-scale datasets, providing accurate and rapid diagnoses. Adversarial attacks, by adding slight perturbations to the original images, can cause models to make severe classification errors. However, medical images often contain significant amounts of noise, making them potential targets for adversarial attacks. Such attacks are particularly critical in the fields of life and health, potentially leading to incorrect diagnoses and treatment decisions, thereby threatening patient safety and health. Therefore, it has become a key research focus to test the security of deep learning systems based on medical images and to develop effective defense mechanisms to enhance their security and reliability.

This paper investigates adversarial attacks and defenses in deep learning systems based on medical images. We tested the security vulnerabilities in deep learning systems for medical images using both white box and black box adversarial attacks. Subsequently, we proposed more effective defense methods to defend against these attacks. Specifically, we first used white box attack algorithms to attack different deep learning systems based on medical images, testing the security of the medical image deep learning systems. We found that medical image deep learning systems are vulnerable to attacks, leading to misclassification of medical images. Subsequently, we proposed better defense method to defend against these white box attacks. Furthermore, to better simulate real-world scenarios, we proposed a black box attack algorithm and tested it on different medical image deep learning systems. This algorithm demonstrated superior attack capabilities. Finally, to defend against black box attack, we proposed a defense method with improved performance, thereby enhancing the security of medical image deep learning systems.