1. Introduction

Bone metastasis refers to the spread of cancer cells from a primary organ to bone tissue and is one of the major causes of serious complications in cancer patients. This condition can trigger various clinical symptoms such as severe pain, impaired mobility, pathological fractures, spinal cord compression, cranial nerve palsy, nerve root injury, hypercalcemia, and disrupted hematopoietic function due to bone marrow involvement. Bone metastases are most commonly found in patients with breast and prostate cancer, where their presence significantly influences treatment strategy choices and patient prognosis [1], [2]. Therefore, early detection of bone metastases is crucial for improving therapy effectiveness and extending patient survival.

One commonly used imaging method for detecting bone metastases is the Whole-Body Bone Scan (WBBS), which remains a clinical standard due to its relatively low cost and competitive sensitivity compared to other modalities such as Magnetic Resonance Imaging (MRI) [3].

Although WBBS is effective in detecting metastatic lesions, the interpretation of the images still heavily depends on the expertise and experience of clinicians. Challenges arise as some non-neoplastic conditions, such as osteomyelitis, arthropathy, or fractures, can produce imaging patterns similar to metastasis. Additionally, patients without metastases may still exhibit increased tracer uptake (hotspots), raising the risk of misdiagnosis [4]. Manual analysis is not only subjective but also time-consuming and labor-intensive.

To overcome these limitations, various artificial intelligence-based approaches have been developed, particularly in the field of medical image analysis [5]–[8]. The use of deep learning models in analyzing WBBS images offers potential to improve accuracy, efficiency, and consistency in detecting bone metastases. AI-based systems enable more objective diagnosis processes and help reduce the workload of clinicians in daily practice.

With the advancement of technology, the application of deep learning in medical image analysis has shown promising results, especially in enhancing diagnostic accuracy and efficiency. Several previous studies have developed bone metastasis analysis systems using different approaches. For instance, Papandrianos et al. [5] developed a CNN model for classifying bone scintigraphy images of prostate cancer patients. The dataset used consisted of 778 images classified into three categories: normal, degenerative, and malignant. The developed CNN model achieved a classification accuracy of 91.61.

Moving toward segmentation approaches, Cao et al. [6] proposed a self-defined five-layer U-Net-based model, tested on 260 SPECT images. This model achieved a Dice Coefficient (DSC) score of 0.6556. Meanwhile, Shimizu et al. [7], a system was built for bone segmentation and hotspot extraction, followed by automatic BSI measurement. The system used butterfly-type networks for segmentation and extraction, and was tested using 246 images with three-fold cross-validation. The results showed an average DSC score of 0.8892 and a cross-correlation score for BSI calculation of 0.9337.

Huang et al. [8] introduced the BS-80K dataset, a large scale, open-access dataset consisting of 3,247 pairs of anterior and posterior images from patients at West China Hospital. They evaluated the performance of several object detection models—namely Faster R-CNN, Cascade R-CNN, and RetinaNet—using this dataset. The models were trained with 5-fold cross-validation and employed the WarmupMultiStepLR schedule for learning rate adjustment. Using a ResNet-101 backbone, the Average Precision (AP) at an IoU threshold of 0.5:0.95 was 0.2423 for Faster R-CNN, 0.2484 for Cascade R-CNN, and 0.1381 for RetinaNet. The AP50 scores (IoU $\geq$ 0.5) for the same models were 0.6189, 0.6128, and 0.1381, respectively. When using a ResNet-50 backbone, the AP scores were 0.2417 for Faster R-CNN, 0.2457 for Cascade R-CNN, and 0.1334 for RetinaNet. The corresponding AP50 scores were 0.6095, 0.6038, and 0.3830.

Meanwhile, one-stage detection models such as You Only Look Once (YOLO) have also demonstrated competitive performance. For example, Khanam et al. [9] compared the performance of YOLOv5, YOLOv8, and YOLOv11 in detecting defects in solar panels. YOLOv11 achieved an mAP50 score of 93.4% with an inference time of only 7.7 ms per image, demonstrating 3–4 times higher computational efficiency compared to previous models. Another study by Li et al. [10] used YOLOv11 with the addition of Efficient Channel Attention (ECA) to detect miners in underground environments. The model was tested under various lighting and blur conditions and recorded an mAP50 score of 0.958, outperforming other models.

In addition to experiments on industrial datasets [9], [10], recent studies has extended the application of YOLOv11 into medical domains. For example, Tariq & Choi [11] utilized YOLOv11 enhanced with attention modules (GAM, ResNet GAM, SE BLOCK) to detect pediatric wrist fractures in X-rays from the GRAZPEDWRI-DX dataset, achieving mAP50 scores up to 64.3% at 1024×1024 resolution. Meanwhile, Ferdi [12] developed G-YOLOv11, a lightweight version utilizing ghost convolution that reduced inference time to only 2.4 ms with mAP@0.5 of 0.535, significantly lowering model size while maintaining acceptable accuracy. Comparisons with other YOLO variants show consistent improvements: YOLOv9, for instance, raised mAP50-95 from 42.2% to 43.7% on the same dataset.

In this study, we aim to develop an automatic detection system for bone metastasis hotspots using the YOLOv11 model [13] with the BS-80K dataset. The choice of YOLOv11 is based on its advantages in detection speed and accuracy, making it a potential solution for accelerating and simplifying the WBBS image analysis process, which currently still heavily relies on manual evaluation by clinicians. This study also

explores various hyperparameter combinations to obtain the best training configuration. As a comparison, a Faster R-CNN model was also built and evaluated on the same dataset to assess the relative performance of both models.

The structure of this paper is organized as follows. Section II describes the BS-80K dataset and the detection model architectures used. Section III presents the experimental results and evaluative analysis. Section IV concludes the findings and provides directions for future research.