

## Article

# Skeleton Segmentation on Bone Scintigraphy for BSI Computation

Po-Nien Yu <sup>1</sup>, Yung-Chi Lai <sup>2</sup>, Yi-You Chen <sup>1</sup> and Da-Chuan Cheng <sup>1,3,\*</sup> 

<sup>1</sup> Department of Biomedical Imaging and Radiological Science, China Medical University, Taichung 404, Taiwan; lu09200016@gmail.com (P.-N.Y.); nses123789@gmail.com (Y.-Y.C.)

<sup>2</sup> Department of Nuclear Medicine, Feng Yuan Hospital Ministry of Health and Welfare, Taichung 420, Taiwan; daniellai999@hotmail.com

<sup>3</sup> Center of Augmented Intelligence in Healthcare, China Medical University Hospital, Taichung 404, Taiwan

\* Correspondence: dccheng@mail.cmu.edu.tw; Tel.: +886-4-2205-3366

**Abstract:** Bone Scan Index (BSI) is an image biomarker for quantifying bone metastasis of cancers. To compute BSI, not only the hotspots (metastasis) but also the bones have to be segmented. Most related research focus on binary classification in bone scintigraphy: having metastasis or none. Rare studies focus on pixel-wise segmentation. This study compares three advanced convolutional neural network (CNN) based models to explore bone segmentation on a dataset in-house. The best model is Mask R-CNN, which reaches the precision, sensitivity, and F1-score: 0.93, 0.87, 0.90 for prostate cancer patients and 0.92, 0.86, and 0.88 for breast cancer patients, respectively. The results are the average of 10-fold cross-validation, which reveals the reliability of clinical use on bone segmentation.

**Keywords:** Mask R-CNN; Double U-Net; Deeplabv3 +; bone segmentation; bone scintigraphy



**Citation:** Yu, P.-N.; Lai, Y.-C.; Chen, Y.-Y.; Cheng, D.-C. Skeleton Segmentation on Bone Scintigraphy for BSI Computation. *Diagnostics* **2023**, *13*, 2302. <https://doi.org/10.3390/diagnostics13132302>

Academic Editor: Yanwu Xu

Received: 19 June 2023

Revised: 2 July 2023

Accepted: 5 July 2023

Published: 6 July 2023



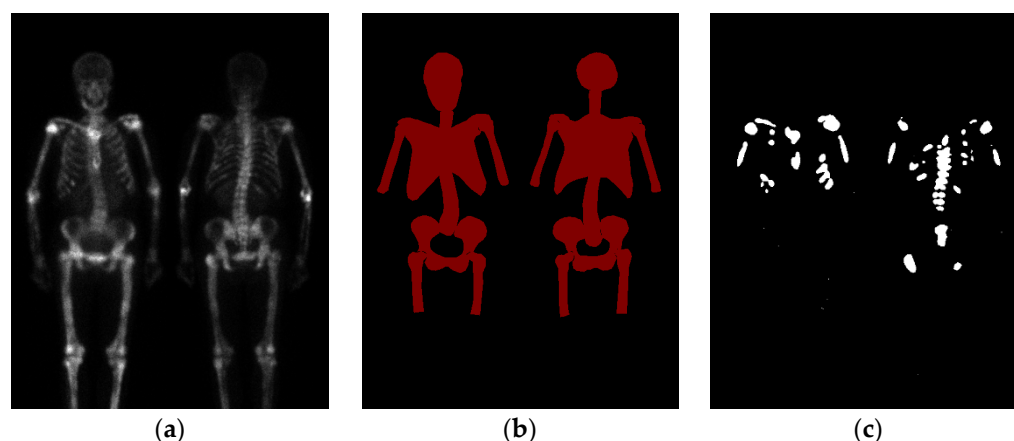
**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Bone is the most common targeted site for metastatic cancer, especially in the advanced and later phases of cancer progression—notably breast, prostate, and lung cancers, with the highest incidence rates [1]. Bone metastases can severely impact patients' daily activities and quality of life due to severe pain and associated major complications. The protracted clinical course of bone metastasis poses significant challenges to treatment. Per a 2022 report published in the Taiwan National Health Insurance Research Database [2], prostate cancer ranked sixth among the leading causes of cancer death among Taiwanese men. In contrast, breast cancer ranked second among the leading causes of cancer death among Taiwanese women. Diagnostic techniques for bone metastasis currently include bone scintigraphy (BS), X-ray imaging, computed tomography (CT), and magnetic resonance imaging (MRI), while BS serves as the most cost-effective early screening method. BS can diagnose bone metastasis earlier than CT or X-ray, within 3 to 6 months [3].

Bone metastasis typically affects the central skeletal system and the proximal regions of the upper and lower limbs. The central skeletal system contains red bone marrow, which may contribute to the formation of bone metastasis due to its physiological characteristics [4]. Physicians often perform a whole-body bone scan (WBBS) to diagnose the presence of bone metastasis. <sup>99m</sup>Tc-MDP is the radiopharmaceutical injected into a patient's vein, which can enter the bone cells and deposit with mineral components in four hours. Consequently, Tc-99m MDP tends to accumulate in areas of active bone formation in the affected region, resulting in localized increased radiopharmaceutical activity that appears as a "hot spot" on BS, allowing physicians to identify bone metastasis [5]. However, BS may suffer from ambiguity owing to impacts such as bone injury, arthritis, and degenerative changes, and causes interpretation challenges. Inexperienced clinical physicians may struggle to make accurate judgments or even misinterpret images.

Bone scan index (BSI) is an imaging biomarker used to quantify the extent of bone metastasis in cancers [6]. BSI is calculated as the ratio of “the number of bone lesions indicating bone metastasis” to “the number of regions with a high incidence of bone metastasis” [7–9], as shown in Figure 1. With artificial intelligence, machine learning, and big data, BSI calculation has become more objective, accurate, and diagnostically efficient. BSI’s most attractive application is monitoring treatment and prognosis, providing significant clinical value. Armstrong et al. from Duke University introduced the automated bone scan index (aBSI) as an objective imaging parameter [10], which can evaluate the prognosis of metastatic castration-resistant prostate cancer (mCRPC) patients undergoing systemic treatment in clinical trials. In [11,12], manual and automated BSI measurements were highly correlated ( $\rho = 0.80$ ), and automated BSI scoring demonstrated reproducibility, eliminating the subjectivity of clinical judgment while retaining the same clinical significance as manual BSI scoring. Furthermore, some studies confirmed the utility of aBSI in mCRPC patients [13–15], while other studies have begun to explore its application and refinement in other tumors [16].



**Figure 1.** (a) represents a WBBS, (b) depicts the skeleton regions with a high incidence of bone metastasis, and (c) indicates the areas where bone metastasis is present. The BSI in (a) corresponds to the area ratio of (c) to (b).

Generally, computer-assisted diagnosis (CAD) systems that utilize machine learning or neural network (NN) framework for calculating BSI on WBBS images can be divided into two parts: lesion segmentation and skeleton segmentation, which respectively reflect the numerator and denominator of the BSI value [17–20]. Recently, numerous studies [21,22] and related patents [23,24] on lesion segmentation using the NN framework have been conducted. However, the performance of the lesion pixel-wise segmentation has not been thoroughly and rigorously investigated. Similarly, research on skeleton segmentation using deep learning and NN models is scarce in previous studies [20,25] despite the mention of its skeleton segmentation approach in [20], which lacks comparison with other NN models. Although [25] compared its performance with U-Net, it remained confined to traditional semantic segmentation network architectures. Thus, the field of skeleton segmentation using NN remains insufficiently explored. This paper uses different NN models for skeleton segmentation on WBBS images and investigates their results. Additionally, we have built a website platform for online skeleton segmentation of WBBS images [Appendix A], which provides effective skeleton segmentation data for further evaluation of BSI.

## 2. Materials and Methods

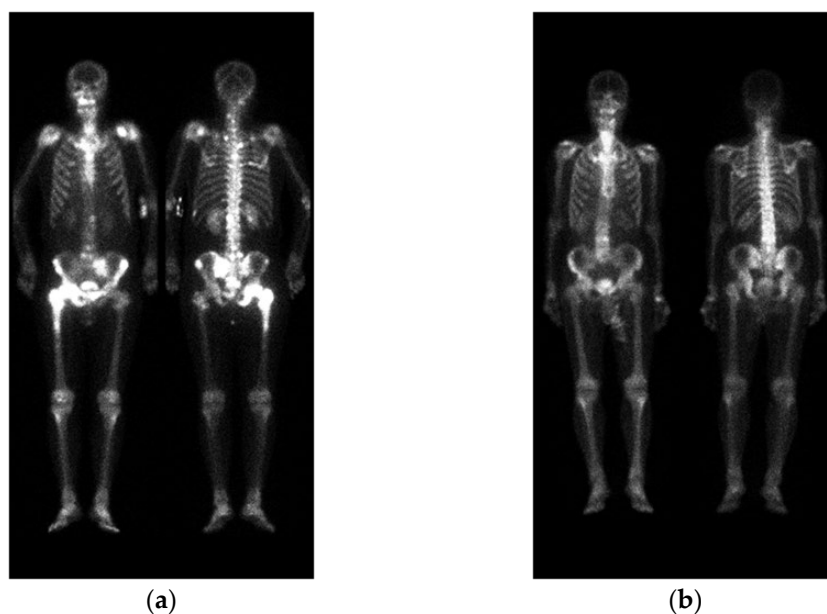
### 2.1. Materials

In this retrospective study in collaboration with the Department of Nuclear Medicine at China Medical University Hospital, 196 WBBS images of patients with prostate cancer were collected. Among the 196 patients, 110 patients had bone metastasis, and 86 patients

had no evidence of bone metastasis. We also collected 163 WBBS images of patients with breast cancer. All of them had bone metastasis. The study was approved by the Institutional Review Board (IRB) and the Hospital Research Ethics Committee (CMUH106-REC2-130) of China Medical University.

The radiopharmaceutical used for WBBS was Tc-99m MDP, and the imaging was performed 4 h after the vein injection. A Gamma camera (Millennium MG, Infinia Hawkeye 4, or Discovery NM/CT 670 system; GE Healthcare, Waukesha, WI, USA) was used for planar bone scanning, with a low-energy high-resolution or general-purpose collimator, a matrix size of  $1024 \times 256$ , a photon energy centered on the 140 keV peaks, and a symmetric 20% energy window. The collected bone scan images were in DICOM format, with a spatial resolution of  $1024 \times 512$  pixels (composed of anterior-posterior (AP) and posterior-anterior (PA) views), and the intensity information of each pixel was saved in 2-byte (uint16). The images were preprocessed using the dedicated GE Xeleris workstation (GE Medical Systems, Haifa, Israel; version 3.1) before being uploaded to PACS.

A standard WBBS image contains two views: anterior and posterior. The original DICOM images were first converted to PNG format after removing any identifiable information. Following the approach described in [22], pre-processing was performed by normalizing the image size and intensity. Afterwards, the anterior and posterior views were cropped into a single image with a size of  $950 \times 512$ , without any scaling or geometric transformations, as shown in Figure 2.



**Figure 2.** Two WBBS, (a) has bone metastasis and (b) has no metastasis.

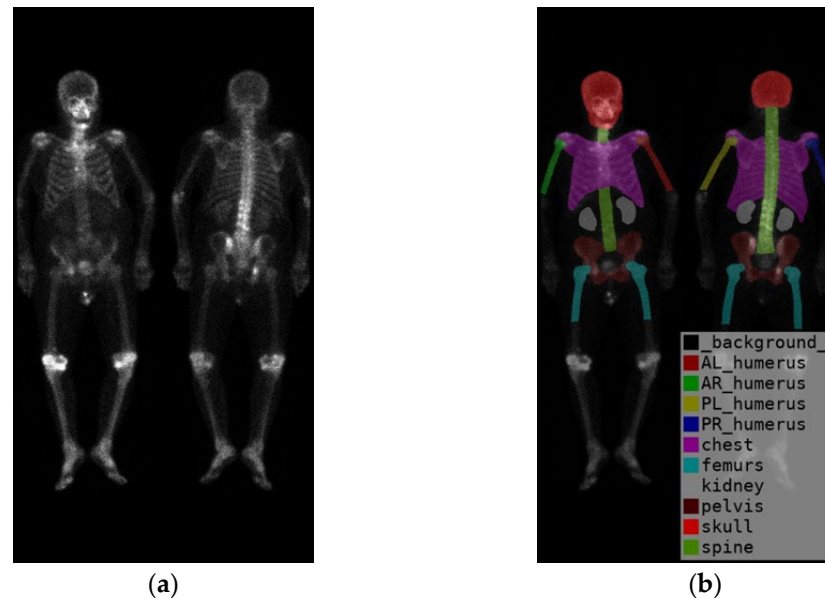
## 2.2. Region Definition

To identify the skeletal regions where bone metastases occur most frequently, we consulted with two experienced nuclear medicine physicians and established standards. The standards require the approval of these two board-certified nuclear medicine physicians. The regions are the skull, spine, chest (including ribs, scapula, and clavicle), humerus (proximal to midshaft of the femurs), femurs (proximal to midshaft of the humerus), and pelvis.

The positions of the humerus on images differ significantly, as shown in Figure 2. Different from only one category on femurs, we categorize humerus into four categories, i.e., the left and right humerus in the anterior and posterior views separately. The reason for doing so will be addressed in the discussion. Furthermore, Tc-99m MDP undergoes renal metabolism, which can result in the kidneys appearing as high signal areas. In some

situations, the kidney will be misclassified as metastasis. To alleviate this problem, we created an extra kidney category to exclude ambiguity.

In summary, there are in total ten categories (Figure 3), including the skull, spine, chest (including ribs, scapula, and clavicle), anterior right humerus (AR), anterior left humerus (AL), posterior right humerus (PR), posterior left humerus (PL), femurs (proximal to midshaft of the humerus), pelvis, and kidney.

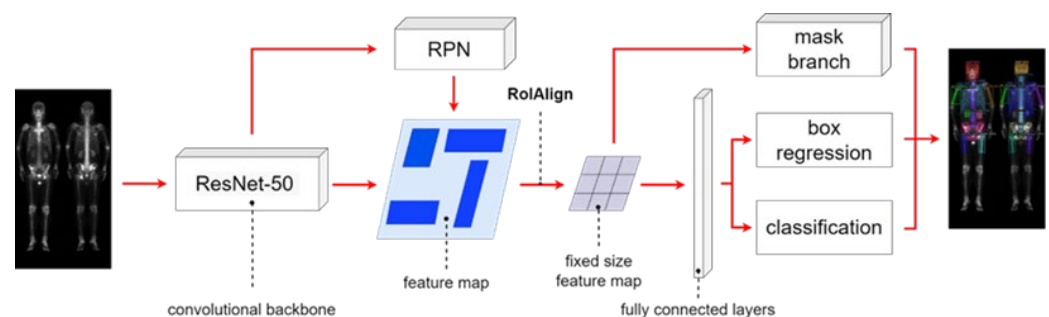


**Figure 3.** (a) shows the original image, while (b) displays the ground truth of the bone metastasis-prone regions and 10 (+1 background) categories with different colors.

### 2.3. Neural Network Architectures

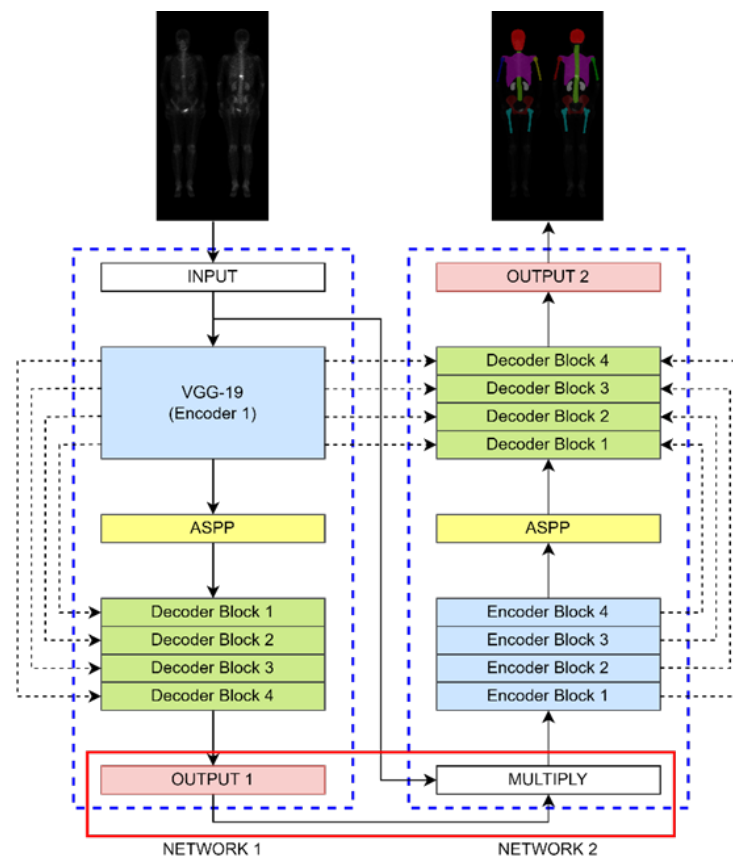
Three different neural network architectures were tested, including Mask R-CNN [26], Double U-Net [27], and Deeplabv3 plus [28]. We used similar hyperparameters on these three models to conduct experiments to compare their performances.

The Mask R-CNN architecture shown in Figure 4 comprises four main parts: backbone architecture, RPN, RoIAlign, and head architecture. We used ResNet-50 as the backbone. The hyperparameters hold the same learning rate of 0.005, batch size of 4, and 100 epochs.



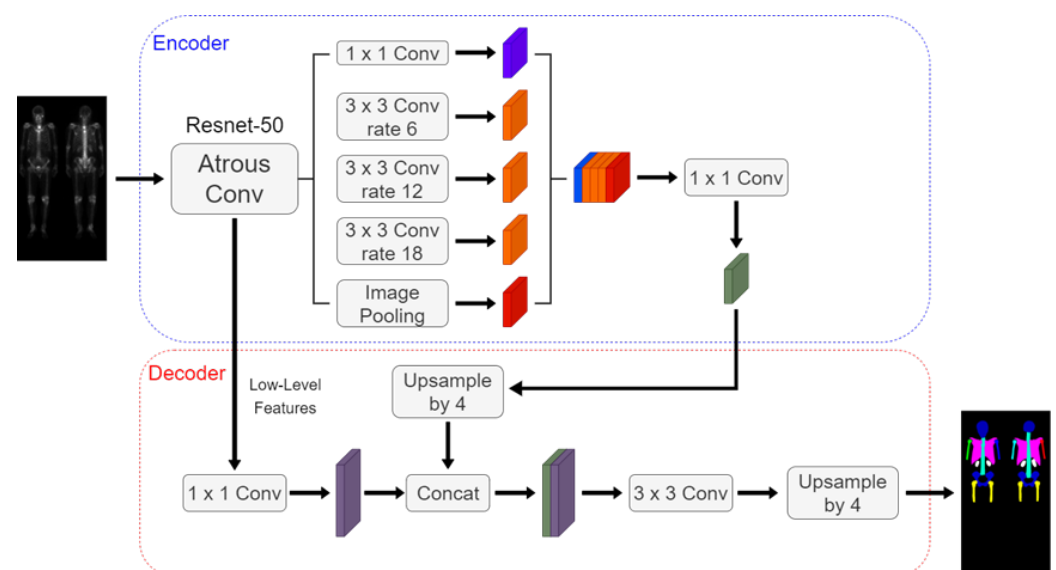
**Figure 4.** The architecture of Mask R-CNN having multi-class classification.

The Double U-Net architecture shown in Figure 5 comprises two sub-networks, dilated convolution, spatial pyramid pooling, and an SE block. It was originally designed for binary classification. Here we modified it to make the multi-class classification. We changed the output layer of Network 1 to have a SoftMax activation function to enable multi-class classification. The hyperparameters were set to be a learning rate of 0.0005, batch size of 4, 200 epochs (without data augmentation), or 20 epochs (with data augmentation).



**Figure 5.** The architecture of Double U-Net is comprised of two sub-networks. To enable multi-class classification, we modified the output of Network 1 and the input of Network 2.

The Deeplabv3 plus architecture shown in Figure 6 includes an encoder, decoder, dilated convolution, and depth-wise separable convolution. We used ResNet-50 as the encoder backbone. The hyperparameters were set to be a learning rate 0.0005, batch size of 4, and 200 epochs.



**Figure 6.** The architecture of the DeeplabV3 plus using ResNet-50 as the backbone.

The learning rate is a hyperparameter used in various machine learning algorithms, particularly in gradient-based optimization. It determines the step size at which the model

updates its parameters during training. The choice of learning rate depends on the specific problem. Typically, every model has its own suggested learning rate. In this study, we choose a balance between accuracy and training speed. For this task, Mask R-CNN uses a learning rate of 0.005, while Double U-Net and DeeplabV3 plus use a learning rate 0.0005.

#### 2.4. Image Pre-Processing

The input matrix size for Mask R-CNN was  $950 \times 512$ . Double U-Net and Deeplabv3 plus's input matrix size was adjusted to  $960 \times 512$  by padding with zeros due to their restriction. The labels were saved in PNG format with integers ranging from 0 to 10.

Augmentation included rotations ( $-3^\circ$ ,  $0^\circ$ ,  $3^\circ$ ) with step  $1^\circ$ , scaling (0.9, 1, 1.1) with step 0.1, and brightness adjustments (0.8, 0.93, 1.06, 1.19, 1.32, 1.45, 1.58, 1.7 times). The augmented images had the same matrix size as the original images, resulting in a total rate of 63 times increase. The augmentations were only used in training.

#### 2.5. Evaluations

In this study, the terms true positive (TP), false positive (FP), true negative (TN), and false negative (FN) were defined in pixel scale. The evaluation metrics used in the experiment were precision (Equation (1)) and sensitivity (Equation (2)), and the overall model evaluation was based on the F1 score (Equation (3)).

$$\text{Precision} = (\text{True positive}) / (\text{True positive} + \text{False positive}), \quad (1)$$

$$\text{Sensitivity} = (\text{True positive}) / (\text{True positive} + \text{False negative}), \quad (2)$$

$$\text{F1 score} = 2(\text{Precision} \times \text{Sensitivity}) / (\text{Precision} + \text{Sensitivity}), \quad (3)$$

### 3. Results

#### 3.1. 10-Fold Cross-Validation

In this study, validations on these three models used 10-fold cross-validation. Two datasets comprised 196 prostate cancer WBBS images and 163 breast cancer WBBS images, respectively. The ratio of training, validation, and test was set to be 8:1:1. The main goal of this experiment was to compare the performance differences among each network and to evaluate the impact of prostate and breast cancer WBBS images on network training. The hyperparameters used in the experiment are in Table 1, and the results are depicted in Tables 2 and 3, compared in Table 4. The qualitative results of bone segmentation are shown in Figures 7 and 8.

**Table 1.** Hyperparameters were used for the 10-fold cross-validation experiments with each neural network.

Hyperparameters	Mask R-CNN	Double U-Net	DeeplabV3 Plus
Learning Rate	0.005	0.0005	0.0005
Batch Size	4	4	4
Epochs	100	200	200

**Table 2.** The comparing results of 10-fold cross-validation on prostate cancer WBBS image dataset.

Category	Mask R-CNN		Double U-Net		DeeplabV3 Plus	
	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity
Skull	97.22	94.43	96.05	96.13	95.34	95.91
Spine	93.90	88.62	91.16	91.30	89.94	89.79
Chest	95.33	93.58	94.83	94.52	93.61	93.87
AR_humerus	91.82	84.80	89.65	90.18	87.42	87.88
AL_humerus	92.46	85.30	89.76	90.12	87.94	89.02



Table 2. Cont.

Category	Mask R-CNN		Double U-Net		DeeplabV3 Plus	
	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity
PR_humerus	91.72	84.41	88.68	89.55	85.77	87.25
PL_humerus	89.94	82.01	87.89	88.78	87.50	83.64
Pelvis	92.32	88.26	90.76	90.83	90.99	87.84
Femurs	88.40	81.75	86.08	84.85	85.59	82.60
Kidney	86.13	79.23	82.45	82.73	80.15	81.87
Average	91.93	86.24	89.73	89.90	88.43	87.97
Average (w/o kidney)	92.57	87.02	90.54	90.70	89.34	88.64

The F1 scores are 89.71, 90.62, and 88.99 for Mask R-CNN, Double U-Net, and DeeplabV3, respectively. The Double U-Net has the best F1-score.

Table 3. The comparing results of 10-fold cross-validation on the breast cancer WBBS image dataset.

Category	Mask_R-CNN		Double U-Net		DeeplabV3 Plus	
	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity
Skull	97.24	94.23	96.18	95.88	95.91	93.24
Spine	93.20	88.61	91.15	90.68	90.56	87.76
Chest	95.17	93.48	94.10	94.32	92.78	93.40
AR_humerus	89.67	80.23	87.21	86.01	85.88	81.90
AL_humerus	89.07	81.20	86.44	84.97	87.15	80.26
PR_humerus	89.65	82.10	87.58	86.46	85.41	83.66
PL_humerus	88.28	80.08	87.34	86.39	86.62	81.92
Pelvis	92.22	88.27	90.86	90.24	91.34	87.54
Femurs	89.95	81.39	87.05	84.83	88.06	81.71
Kidney	87.21	80.71	84.37	83.74	83.91	77.59
Average	91.17	85.03	89.23	88.35	88.76	84.90
Average (w/o Kidney)	91.61	85.51	89.77	88.86	89.30	85.71

The F1-scores are 88.45, 89.31, and 87.47 for Mask R-CNN, Double U-Net, and DeeplabV3, respectively. The Double U-Net has the best F1-score.

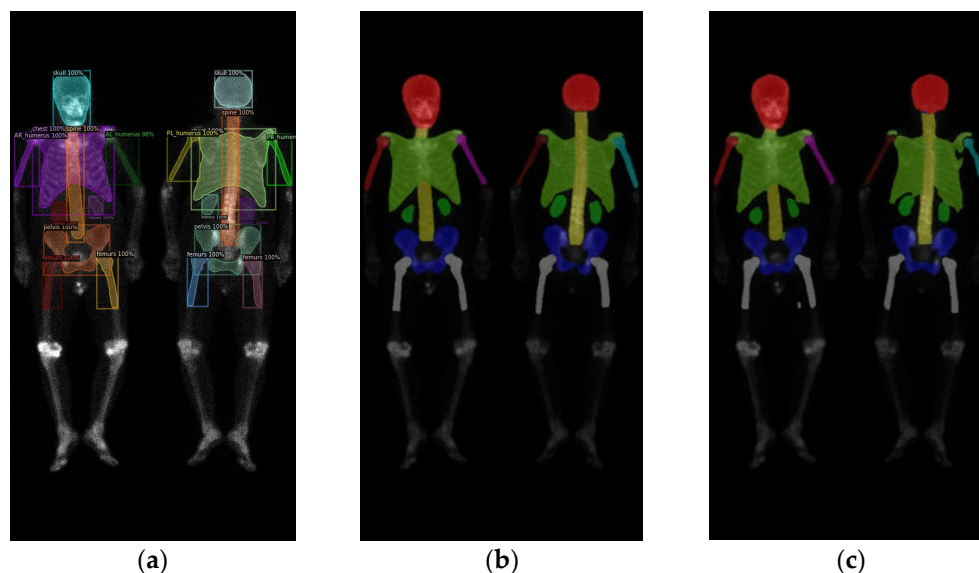
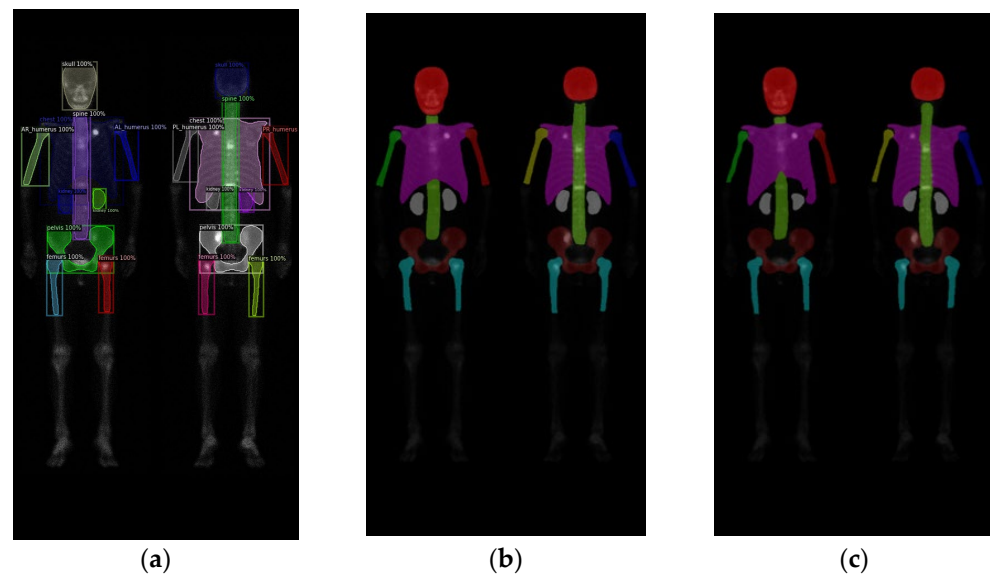


Figure 7. The qualitative results of three models on prostate cancer WBBS images: (a) Mask R-CNN, (b) Double U-Net, and (c) DeeplabV3 plus.

Table 4. The comparing results of 10-fold cross-validation on the two image datasets.

Database	Mask_R-CNN			Double U-Net			DeeplabV3 Plus		
	Pre.	Sen.	F1-Score	Pre.	Sen.	F1-Score	Pre.	Sen.	F1-Score
Prostate cancer	92.57	87.02	89.71	90.54	90.70	90.62	89.34	88.64	88.99
Breast cancer	91.61	85.51	88.45	89.77	88.86	89.31	89.30	85.71	87.47

Pre. = Precision, Sen. = Sensitivity.



**Figure 8.** The qualitative results of three models on breast cancer WBBS images: (a) Mask R-CNN, (b) Double U-Net, and (c) DeeplabV3 plus.

### 3.2. 10-Fold Cross-Validation with Data Augmentation

After the above experiments, we chose Double U-Net for investigation because it slightly outperformed on F1-score. Following, we fine-tuned the epoch to trade-off the training time and the performance to see what best performance we could reach. The images of training for prostate cancer and breast cancer were augmented 63 times by using rotation, scaling, and brightness adjustment described in the methods. Again, the hyperparameters are in Table 5, and the quantitative results of the 10-fold cross-validation are in Table 6.

**Table 5.** Hyperparameters for training Double U-Net.

Hyperparameters	Double U-Net
Learning Rate	0.0005
Batch Size	4
Epochs	20

**Table 6.** 10-fold cross-validation on Double U-Net, used augmentation.

Fold Number	Prostate		Breast	
	Precision	Sensitivity	Precision	Sensitivity
1	86.67	96.05	83.95	94.84
2	87.01	94.92	86.18	95.26
3	91.22	91.33	81.14	96.05
4	93.01	91.37	81.87	96.32
5	85.69	94.85	84.35	96.18
6	94.18	89.28	96.23	76.73
7	96.10	86.81	95.64	85.26
8	93.43	88.31	95.37	84.49
9	92.99	87.74	95.57	85.51
10	93.89	88.12	94.97	89.19
Average	91.42	90.88	89.53	89.98

The F1 scores are 91.15 and 89.75, respectively.

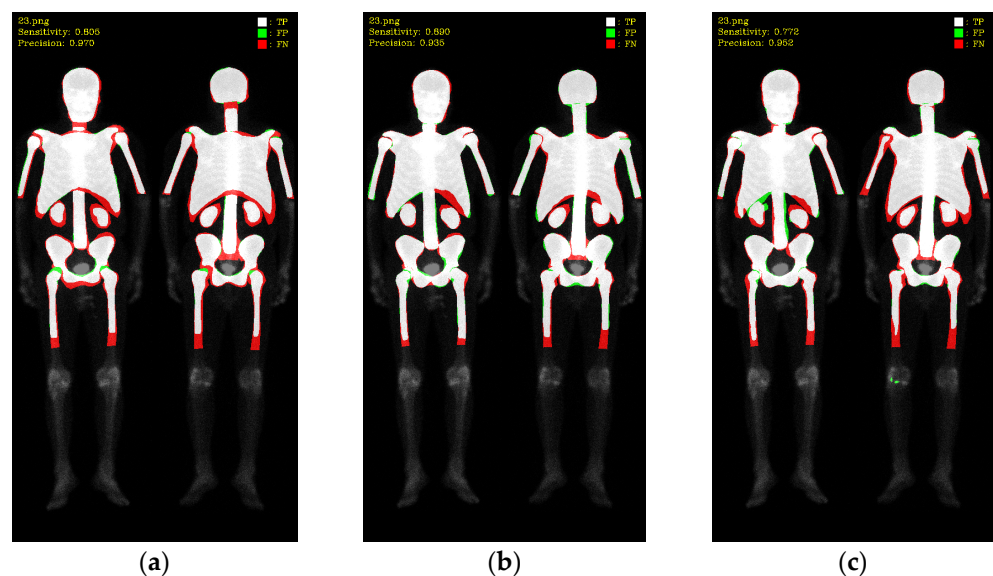
## 4. Discussion

This study utilised Mask R-CNN, Double U-Net, and DeeplabV3 plus for skeleton segmentation comparison on prostate cancer and breast cancer WBBS images. The quantitative results were investigated via 10-fold cross-validation. Based on the quantitative findings,

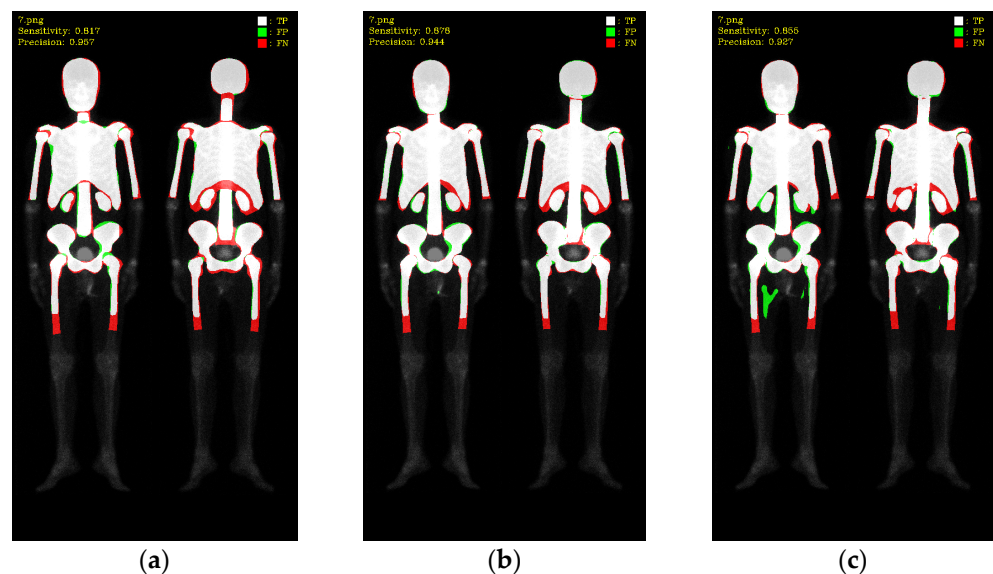


Mask R-CNN exhibited higher precision than Double U-Net by 2.03% in the prostate cancer dataset and 1.84% in the breast cancer dataset. Mask R-CNN also exhibited higher precision than DeeplabV3 by 3.23% in the prostate dataset and 2.31% in the breast dataset. On the other hand, Double U-Net (90.70% & 88.86%) demonstrated higher sensitivity than Mask R-CNN (87.02% & 85.51%) and DeeplabV3 plus (88.64% & 85.71%). This indicated that Mask R-CNN had lower false positives (FP) during prediction, while Double U-Net had lower false negatives (FN).

To better understand these results, we visualized the predictions, where white color represented TP, green color represented FP and red color represented FN (as shown in Figures 9 and 10). Mask R-CNN's predictions shifted inward slightly compared to the ground truth (GT), resulting in more FN in the edge regions and only a few FP. Double U-Net's predictions aligned well with the GT along the edges, resulting in slightly fewer FN but more FP. DeeplabV3 plus exhibited irregularities along the edges compared to the other two models, leading to noticeable erroneous FP and an overall increase in FP.



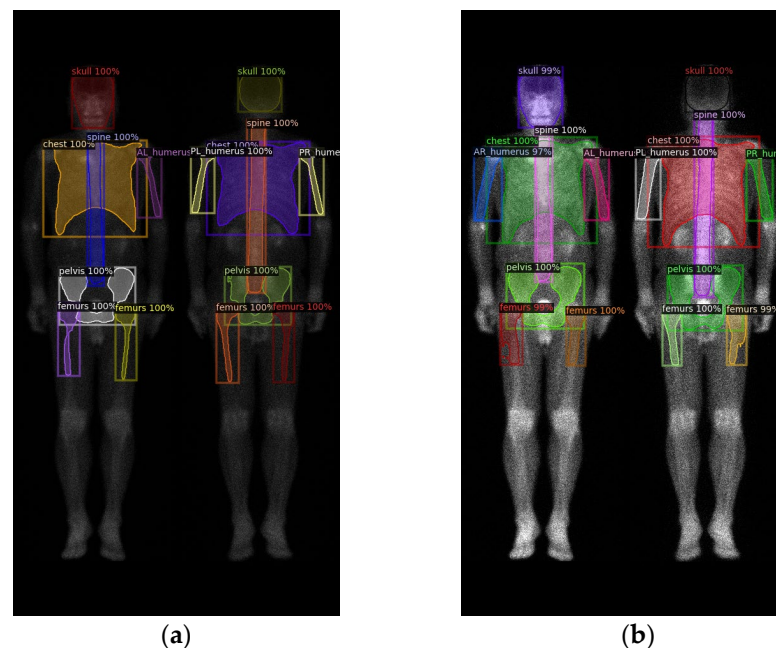
**Figure 9.** Qualitative comparisons on three models: (a) Mask R-CNN, (b) Double U-Net, (c) DeeplabV3 plus. White: TP, red: FP, and green: FN.



**Figure 10.** Qualitative comparisons on three models: (a) Mask R-CNN, (b) Double U-Net, (c) DeeplabV3 plus.

These findings shed light on the performance of different models for skeleton segmentation, emphasizing the trade-off between FP and FN. Further improvements can be explored to address the limitations observed, particularly in the case of DeeplabV3 plus, to enhance its stability and accuracy.

Further investigation of Mask R-CNN results revealed an increase in false negatives (FN) when predicting smaller categories, such as the humerus and kidneys, as shown in Figure 11a. This result could be caused by the following reasons:



**Figure 11.** (a) Original test segmentation result with missing right humerus in the frontal view. (b) Segmentation result after adjusting the brightness to 2.5 times and retest.

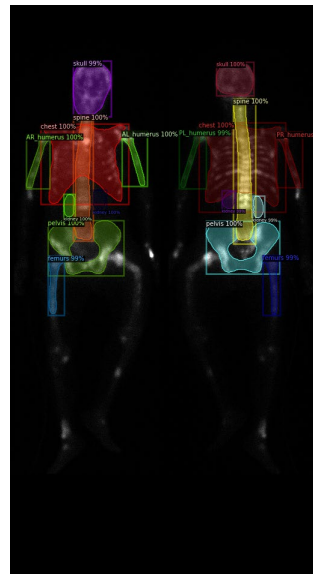
First, the insufficient brightness in the WBBS image may hinder feature detection. The brightness of WBBS images depends on the counts collected by the scintillation crystal, which can be influenced by factors such as patient thickness and radiopharmaceutical activity. In cases where the received counts are insufficient, resulting in inadequate image brightness, deep neural network models may struggle to make accurate judgments or even make errors. Adjusting the image brightness and conducting further tests can help alleviate this situation, as shown in Figure 11b.

Second, abnormal patient positioning in the WBBS image could cause another issue. In a few instances, patient positioning in the WBBS image deviates to some extent from standard clinical positions. This deviation made challenges for CNN prediction, as shown in Figure 12. The degree of deviation is closely related to the patient's clinical condition and is difficult to entirely avoid in clinical practice. While other previous studies might manually exclude misleading images to prevent such occurrences, this study aimed to maintain a dataset that reflects real clinical scenarios, thereby we did not exclude any case. To enhance the network's ability to predict WBBS images with unusual positioning, future considerations include employing hard negative mining techniques to improve the model's generalization capabilities.

Third, the model's insensitivity to features of small objects in WBBS images could also decrease performance. Quantitative results indicated relatively low precision for categories such as upper limbs, femurs, and kidneys, which correspond to smaller objects. This suggested Mask R-CNN facing certain difficulties in segmenting smaller regions.

These findings highlighted specific challenges encountered during the skeleton segmentation process, particularly related to image brightness, abnormal patient positioning,

and the segmentation of smaller objects. Addressing these challenges could improve the performance of the Mask R-CNN model.



**Figure 12.** A segmentation result showed the absence of the frontal and dorsal left femur due to abnormal patient position, while this abnormal position was rare and did not exist in the training dataset.

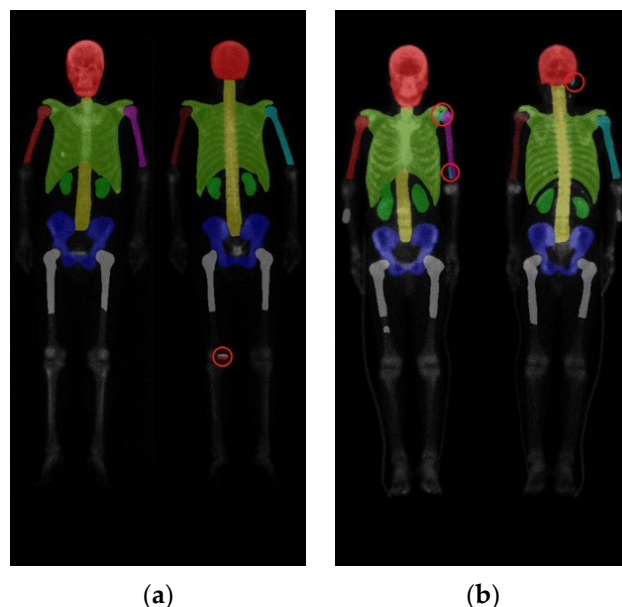
On the other hand, we observed that DeeplabV3 plus and Double U-Net tended to mix categories, resulting in unstable performance. Double U-Net and DeeplabV3 plus did not exhibit the category missing issue observed in Mask R-CNN, but they experienced problems such as category confusion and masks appearing in unintended areas, with DeeplabV3 plus being particularly affected. The issue of category confusion during prediction in semantic segmentation network architectures was not explicitly mentioned in [20,25]. However, in our experiments, we did observe this problem. Figure 13a showed an incorrect segmentation in the knee area in a Double U-Net skeleton segmentation result, while Figure 13b depicted category confusion in the upper limbs and head in a DeeplabV3 plus skeleton segmentation result.

This problem stemmed from different network architectures. Mask R-CNN utilizes parallel branch networks to independently determine categories and select the appropriate masks based on individual region-of-interest (ROI). Consequently, different ROIs could be distinguished independently, and masks could be treated as separate entities. In contrast, traditional fully convolutional network (FCN) architectures performed category and mask predictions simultaneously, leading to competition between different categories and masks. Additionally, due to the design of having one category per mask, FCN-based methods could not treat ROIs independently. Another critical factor was using the Sigmoid activation function and average binary cross-entropy loss in the branch networks, which mitigated the adverse effects of cross-category competition encountered in traditional FCN methods. This design yielded excellent instance segmentation results and avoided category overlap or confusion. From the experiments, Mask R-CNN demonstrated itself more suitable for skeleton segmentation in WBBS images than the other two network architectures.

From experiments shown in Tables 2–4, one might think that the models' performance is close to each other, and there might not be a statistically significant difference. It is crucial to consider the context of image segmentation in deep learning. In this task, precision and sensitivity are calculated pixel-wise. Therefore, even a small difference in percentage points can have a significant impact.

In the experiments involving data augmentation, it was observed that it contributed to a slight performance improvement. As the model already performed reasonably well without data augmentation, the addition of data augmentation only led to marginal per-

formance gains. According to related literature [29], incorporating data augmentation helped reduce overfitting at higher learning rates, allowing the model to be trained for more epochs without sacrificing accuracy. Further experiments and investigations were warranted to explore the impact of data augmentation in more depth.



**Figure 13.** (a) Segmentation result of Double U-Net with a segmentation error in the distal part of the leg. (b) Segmentation result of Deeplabv3 plus showing category confusion in the upper limbs and head region.

The limitation of this study is the scarcity of original data and the homogeneity of its source. In the future, it is desirable to establish collaborations with other medical centers to acquire cross-centre data, thereby improving the performance and generalization ability of the models. Additionally, we only investigated three relatively common network architectures, and it would be an attractive research direction to explore newer architectures, such as transformer-based networks. Different nuclear medicine imaging modalities, such as planar and SPECT, differ in the resulting images. It would be worth investigating whether these differences lead to heterogeneity in model predictions. This is an area for further exploration in the future.

## 5. Conclusions

In this study, we investigated three CNN models on bone segmentation of the WBBS images. We found that only one model was suitable for this goal, Mask R-CNN. The Double U-Net and Deeplabv3 + had a problem with ‘category confusion’, which humans would never have. We used a pixelwise scale to examine the model performance. The best performance we had ever made for Mask R-CNN was the precision, sensitivity, and F1-score: 0.93, 0.87, 0.90 for the prostate cancer dataset and 0.92, 0.86, 0.88 for the breast cancer dataset, which was the average of 10-fold cross-validation.

**Author Contributions:** Conceptualization, D.-C.C.; methodology, D.-C.C.; software, P.-N.Y. and Y.-Y.C.; validation, P.-N.Y. and Y.-Y.C.; formal analysis D.-C.C.; investigation, Y.-C.L.; resources, Y.-C.L. and D.-C.C.; data curation, P.-N.Y. and Y.-Y.C.; writing—original draft preparation, P.-N.Y.; writing—review and editing, D.-C.C.; visualization, D.-C.C.; supervision, D.-C.C.; project administration, D.-C.C.; funding acquisition, D.-C.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Science and Technology Council (NSTC), Taiwan, grant number MOST 111-2314-B-039-040.

**Institutional Review Board Statement:** The study was approved by the Institutional Review Board (IRB) and the Hospital Research Ethics Committee (CMUH106-REC2-130, approved on 27 September 2017) of China Medical University.

**Informed Consent Statement:** Patient consent was waived by IRB due to this is a retrospective study, and only images were used without the patient's identification.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** We thank National Center for High-performance Computing (NCHC) for providing computational and storage resources.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

According to the previous research results, we established a skeleton segmentation website equipped with a deep learning framework, enabling clinical physicians to utilize website functions online to assist in calculating BSI and conducting clinical diagnoses, thereby achieving the purpose of the previous research. This website has multiple functions that allow clinical physicians to upload images and perform simple post-processing on the images within the website. Finally, execute the skeleton segmentation deep learning model for skeleton segmentation. The public IP address of the website is 140.128.65.129, and login credentials are required (username: wbbsweb, password: wbbswebpass).

## References

1. Coleman, R. Metastatic bone disease: Clinical features, pathophysiology, and treatment strategies. *Cancer Treat. Rev.* **2001**, *27*, 165–176. [CrossRef] [PubMed]
2. National Health Insurance Research Database. Available online: <https://www.mohw.gov.tw/cp-16-70314-1.html> (accessed on 12 May 2022).
3. O'Sullivan, G.J.; Carty, F.L.; Cronin, C.G. Imaging of bone metastasis: An update. *World J. Radiol.* **2015**, *7*, 202–211. [CrossRef]
4. Coleman, R.E. Clinical features of metastatic bone disease and risk of skeletal morbidity. *Clin. Cancer Res.* **2006**, *12*, 6243–6249. [CrossRef] [PubMed]
5. Brenner, A.I.; Koshy, J.; Morey, J.; Lin, C.; DiPoce, J. The bone scan. *Semin. Nucl. Med.* **2012**, *42*, 11–26. [CrossRef]
6. Imbriaco, M.; Larson, S.M.; Yeung, H.W.; Mawlawi, O.R.; Erdi, Y.; Venkatraman, E.S.; Scher, H.I. A new parameter for measuring metastatic bone involvement by prostate cancer: The Bone Scan Index. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **1998**, *4*, 1765–1772.
7. Dennis, E.R.; Jia, X.; Mezheritskiy, I.S.; Stephenson, R.D.; Schoder, H.; Fox, J.J.; Heller, G.; Scher, H.I.; Larson, S.M.; Morris, M.J. Bone scan index: A quantitative treatment response biomarker for castration-resistant metastatic prostate cancer. *J. Clin. Oncol.* **2012**, *30*, 519–524. [CrossRef] [PubMed]
8. Anand, A.; Morris, M.J.; Kaboteh, R.; Båth, L.; Sadik, M.; Gjertsson, P.; Lomsky, M.; Edenbrandt, L.; Minarik, D.; Bjartell, A. Analytic validation of the automated bone scan index as an imaging biomarker to standardize quantitative changes in bone scans of patients with metastatic prostate cancer. *J. Nucl. Med.* **2016**, *57*, 41–45. [CrossRef]
9. Nakajima, K.; Edenbrandt, L.; Mizokami, A. Bone scan index: A new biomarker of bone metastasis in patients with prostate cancer. *Int. J. Urol.* **2017**, *24*, 668–673. [CrossRef]
10. Armstrong, A.J.; Nordle, O.; Morris, M. Assessing the Prognostic Value of the Automated Bone Scan Index for Prostate Cancer—Reply. *JAMA Oncol.* **2019**, *5*, 270–271. [CrossRef]
11. Ulmert, D.; Kaboteh, R.; Fox, J.J.; Savage, C.; Evans, M.J.; Lilja, H.; Abrahamsson, P.A.; Björk, T.; Gerdtsen, A.; Bjartell, A.; et al. A novel automated platform for quantifying the extent of skeletal tumour involvement in prostate cancer patients using the Bone Scan Index. *Eur. Urol.* **2012**, *62*, 78–84. [CrossRef]
12. Reza, M.; Kaboteh, R.; Sadik, M.; Bjartell, A.; Wollmer, P.; Trägårdh, E. A prospective study to evaluate the intra-individual reproducibility of bone scans for quantitative assessment in patients with metastatic prostate cancer. *BMC Med. Imaging* **2018**, *18*, 8. [CrossRef] [PubMed]
13. Armstrong, A.J.; Anand, A.; Edenbrandt, L.; Bondesson, E.; Bjartell, A.; Widmark, A.; Sternberg, C.N.; Pili, R.; Tuveson, H.; Nordle, O. Phase 3 assessment of the automated bone scan index as a prognostic imaging biomarker of overall survival in men with metastatic castration-resistant prostate cancer: A secondary analysis of a randomized clinical trial. *JAMA Oncol.* **2018**, *4*, 944–951. [CrossRef] [PubMed]
14. Anand, A.; Morris, M.J.; Kaboteh, R.; Reza, M.; Trägårdh, E.; Matsunaga, N.; Edenbrandt, L.; Bjartell, A.; Larson, S.M.; Minarik, D. A preanalytic validation study of automated bone scan index: Effect on accuracy and reproducibility due to the procedural variabilities in bone scan image acquisition. *J. Nucl. Med.* **2016**, *57*, 1865–1871. [CrossRef] [PubMed]



15. Reza, M.; Wirth, M.; Tammela, T.; Cicalese, V.; Veiga, F.G.; Mulders, P.; Miller, K.; Tubaro, A.; Debruyne, F.; Patel, A.; et al. Automated bone scan index as an imaging biomarker to predict overall survival in the Zometa European Study/SPCG11. *Eur. Urol. Oncol.* **2021**, *4*, 49–55. [[CrossRef](#)]
16. Wuestemann, J.; Hupfeld, S.; Kupitz, D.; Genseke, P.; Schenke, S.; Pech, M.; Kreissl, M.C.; Grosser, O.S. Analysis of bone scans in various tumor entities using a deep-learning-based artificial neural network algorithm—Evaluation of diagnostic performance. *Cancers* **2020**, *12*, 2654. [[CrossRef](#)]
17. Yoshida, A.; Higashiyama, S.; Kawabe, J. Assessment of software for semi-automatically calculating the bone scan index on bone scintigraphy scans. *Clin. Imaging* **2021**, *78*, 14–18. [[CrossRef](#)]
18. Koizumi, M.; Wagatsuma, K.; Miyaji, N.; Murata, T.; Miwa, K.; Takiguchi, T.; Makino, T.; Koyama, M. Evaluation of a computer-assisted diagnosis system, BONENAVI version 2, for bone scintigraphy in cancer patients in a routine clinical setting. *Ann. Nucl. Med.* **2015**, *29*, 138–148. [[CrossRef](#)]
19. Koizumi, M.; Miyaji, N.; Murata, T.; Motegi, K.; Miwa, K.; Koyama, M.; Terauchi, T.; Wagatsuma, K.; Kawakami, K.; Richter, J. Evaluation of a revised version of the computer-assisted diagnosis system, BONENAVI version 2.1.7, for bone scintigraphy in cancer patients. *Ann. Nucl. Med.* **2015**, *29*, 659–665. [[CrossRef](#)]
20. Shimizu, A.; Wakabayashi, H.; Kanamori, T.; Saito, A.; Nishikawa, K.; Daisaki, H.; Higashiyama, S.; Kawabe, J. Correction to: Automated measurement of bone scan index from a whole-body bone scintigram. *Int. J. Comput.-Assist. Radiol. Surg.* **2020**, *15*, 401. [[CrossRef](#)]
21. Cheng, D.C.; Liu, C.C.; Hsieh, T.C.; Yen, K.Y.; Kao, C.H. Bone metastasis detection in the chest and pelvis from a whole-body bone scan using deep learning and a small dataset. *Electronics* **2021**, *10*, 1201. [[CrossRef](#)]
22. Cheng, D.C.; Hsieh, T.C.; Yen, K.Y.; Kao, C.H. Lesion-based bone metastasis detection in chest bone scintigraphy images of prostate cancer patients using pre-train, negative mining, and deep learning. *Diagnostics* **2021**, *11*, 518. [[CrossRef](#)] [[PubMed](#)]
23. Cheng, D.C.; Liu, C.C.; Kao, C.H.; Hsieh, T.C. System of Deep Learning Neural Network in Prostate Cancer Bone Metastasis Identification Based on Whole Body Bone Scan Images. U.S. Patent US11488303B2, 1 November 2022.
24. Brown, M.S. Computer-Aided Bone Scan Assessment with Automated Lesion Detection and Quantitative Assessment of Bone Disease Burden Changes. U.S. Patent US20140105471, 7 April 2015.
25. Huang, K.B.; Huang, S.G.; Chen, G.J.; Li, X.; Li, S.; Liang, Y.; Gao, Y. An end-to-end multi-task system of automatic lesion detection and anatomical localization in whole-body bone scintigraphy by deep learning. *Bioinformatics* **2023**, *39*, btac753. [[CrossRef](#)] [[PubMed](#)]
26. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *arXiv* **2018**, arXiv:1703.06870v3.
27. Jha, D.; Riegler, M.A.; Johansen, D.; Halvorsen, P.; Johansen, H.D. Double U-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. *arXiv* **2020**, arXiv:2006.04868.
28. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *arXiv* **2018**, arXiv:1802.02611v3.
29. Bhuse, P.; Singh, B.; Raut, P. Effect of data augmentation on the accuracy of convolutional neural networks. In *Information and Communication Technology for Competitive Strategies (ICTCS 2020) ICT: Applications and Social Interfaces*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 337–348.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.