SUPPLEMENTARY MATERIALS OF "BONEMET: AN OPEN LARGE-SCALE MULTI-MODAL MURINE DATASET FOR BREAST TUMOR BONE METASTASIS DIAGNOSIS AND PROGNOSIS"

Anonymous authors
Paper under double-blind review

OUTLINE

This document provided supplementary materials to support our main paper. Section A provides details of breast tumor bone metastasis model creations and data collection processes. Section B presents additional experimental settings and results. Section C lists three examples of API for CT Image Segmentation, Registration and RoI-based CT Image Cropping, and section D talks about future potential application of our BoneMet dataset.

A   DETAILS OF DATASET AND DATA COLLECTION

In order to build such large-scale BoneMet dataset, the team has spent more than five years with three to four graduate students assisted with over ten undergraduates and summer students. The experimental expenses include mice purchase, shipping, per diem, surgical tools, cell culture, and user fee to access micro-CT scanners, mechanical testers, microscopy. Usually, it takes approximately one year's training to become an expert in inoculating cancer to mice and 2 to 3 month each batch of the mice (around 20 mice) for data acquisition, image processing including CT reconstruction, segmentation, registration, and image analysis.

Image Acquisition: For each scan, mice were anesthetized with 3% (v/v) isoflurane and held in the built-in holder, which was rotated 208-degree with a step of 0.8 degrees and there are 260 X-ray images taken in total. One frame was taken per step with the following settings: 900 ms exposure time, X-ray of 200 mA current and 50 kVp, and a 0.5 mm Al filter, cone-beam angle Horizontal of 25.888290 (deg) and Vertical 17.423092 (deg). The in vivo scanning of the hindlimbs lasted around 4 min per animal and the accumulated radiation exposure (< 600 mGy) was low.

3D Reconstruction: The 3D volume or 2D cross-sectional images of micro-CT were reconstructed from 260 X-ray projections using the NRecon® software (Bruker) with a voxel size of 7 to 10.6 µm. The 3D reconstructed CT is acquired based on the traditional filtered backprojection (FBP),which uses a 1D filter on the projection data before backprojecting (2D or 3D) the data onto the image space. Then, the 3D bone reconstructed CT were optimized with the several steps such as center of rotation adjustment,beam hardening correction,ring artifact reduction and reconstruction filtering in order to get the reconstructed micro-CT images with high quality. Among these steps, center of rotation adjustment is used to ensure that the reconstruction is symmetrical, beam hardening correction is to compensate artifacts in the scan. Since the 260 X-ray images were taken by the rotation of cameras, the ring artifacts generated during the reconstruction process is minimized by ring artifact reduction. The median filter is employed to remove the salt-and-pepper noise in the reconstructed images, which is the common noise happening in the medical images.

Segmentation:The the left and right tibiae were segmented from reconstructed CT images by our python segmentation package. A global threshold value (75/255) was applied and the processed images were found to agree well with the gray-scaled images. Two limbs of each mice were separated by their momentum intensity in their relative positions and the tibia and femur of each limb were segmented by identify the specific structure of the knee, where the position of the knee is found by comparing the cross-sectional areas of each 2D slices of the bone and the minimal is where the knee located.

Registration (alignment): After the segmentation, the week 0 scans of each mice tibia were aligned to a registration reference tibia CT, where the long axis positioned vertically and the anterior-posterior and medial-lateral axes arranged orthogonally, and the transformation and rotation were manually adjust by Dataviwer®. Then the registered tibiae CT at week 0 scans served as baseline scan for their subsequent week scans. The mutual information maximizing between those of tibiae CT from different mice at week 0 and the registration reference tibia CT, and between those of the sequential scans and themselves week 0 baselines. The mutual information is a measure of the statistical dependence or information shared between the image intensities of the two images. Before transformation (rotation, translation, scaling, *etc.*), initial coarse alignment is performed to reduces the search space for the registration algorithm, then transformation applied to the moving image, maximizing the mutual information between the fixed image and the transformed moving image by gradient descent optimization. Once the optimal parameters are found, the final transformation will be applied to the moving image to align it with the fixed image.
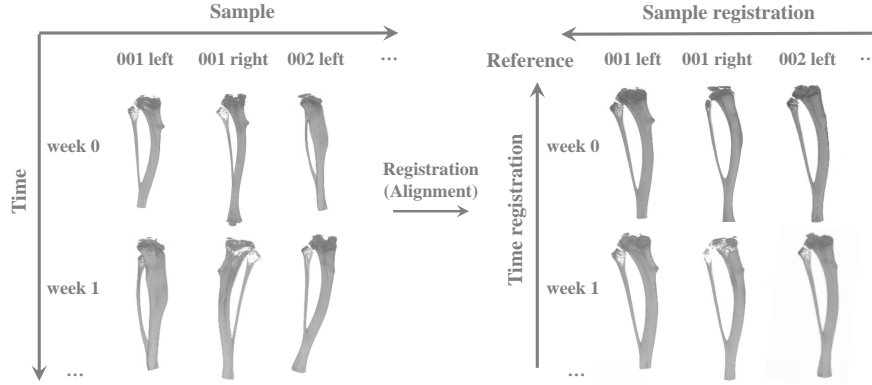

Figure 1: Examples of 3D CT scans from the Seg-CT and the Regist-CT imagery. **Left:** 3D CT scans from the Seg-CT imagery (*i.e*, w/o registration); and **Right:** 3D CT scans from the Regist-CT imagery, which are registered relative to their vertically aligned reference at various time points and across different animals.

Selection of Region of Interest: The overlapped composite RoI-CT images were generated with different pixel values assigned in each pixel based on various criterion: the pixel with bone (above threshold value of 75) in baseline and sequential scans were assigned to the value of 180 (light gray), the non-bone pixel (below threshold) at week 0 later became bone (above threshold) is assigned to 240 (white),the bone pixels (above threshold) at week 0 later became non-bone (below threshold) is set the value to 60 (dark gray), the pixel without bone in both time points were gave the value to 0 (black). After the generation of CT composites, the RoI section was selected automatically with the proximal tibia-fibula junction as the landmark, where the number of contoured circles of bone tissues in each 2D slice changes. Then, the fibula in the 2D slices was manually cropped by CTAn® with interpolated mask.

Analysis and Quantification: The desired analysis such as bone volume changes was conducted on the overlaid tibiae ROI or registered CT to access the bone structure morphology changes, metastatic breast tumor disease progression and mechanical properties impairment over time with and without treatment effects. Cortical polar moment inertia (Ct.pMOI), bone mineral density (Ct.BMD), and tissue mineral density (Ct.TMD) as well as trabecular bone volume fraction (Tb.BV/TV), thickness (Tb.Th), separation (Tb.Sp), bone mineral density (Tb.BMD), and tissue mineral density (Tb.TMD) were calculated using CTAn® 3D analysis software. The overlaid subsequent scans with the baseline scan (Week 0) of the same tibia make it possible to quantify the changes (Delta values) of each tibia at subsequent weeks relative to week 0.

Table 1: Overview of the six components in our BoneMet dataset

Components	Advantages	Applications
Rotation-X-Ray	Low cost, wide availability, and minimal radiation exposure	Pre-training and fine-tuning deep learning models for bone metastasis early detection and monitoring
Recon-CT	Reveal the precise location and size of bone lesions with detailed 3D structure and 2D cross-sections	Training deep learning models for bone metastasis localization, segmentation, and sparse angle reconstruction
Seg-CT	Enhance the diagnosis efficiency by segmenting limited regions	Training deep learning models for bone segmentation and region-specific analysis of metastatic lesions
Regist-CT	Increase sensitivity, making subtle changes in the bone structures detectable	Training deep learning models for longitudinal analysis, early diagnosis, prognosis of metastatic changes, disease progression and treatment effects analysis
RoI-CT	Enable the precise diagnosis and quantitative studies	Training deep learning models for quantitative analysis and prediction of metastatic bone structural and mechanical properties
MiceMediRec	Detailed demographic information about the animals and the disease	Enabling multimodal deep learning models for comprehensive disease diagnosis, prognosis, and generalist biomedical AI diagnosis

Table 2: Details of MiceMediRec Dataset

Source	Parameters	Description
Medical record	Age	Unit: week
	Sex	
	Date	The date of tumor inoculation
	Body Weight	Unit: g
	Tibiae tumor inoculation	Records of each tibiae with tumor or without tumor
	Treatment	Treatment to metastatic breast tumor, like chemotherapy. Dose Unit: μL/g
	Diagnosis	Diagnosis to bone metastasis, like bone lesion
Quantitative analysis	Average Radiant Efficiency	IVIS signal to breast tumor in the bone. Unit: μW/cm^2
	Bone volume	Bone total volumes. Unit: mm^3
	Cort. pMOI	Cortical bone polar moment of inertia. Unit: mm^4
	Cort. Ar	Cortical bone area. Unit: mm^2
	Cort. Th	Cortical bone thickness. Unit: mm
	Trab. BV/TV	Trabecular bone volume relative to the marrow volume. Unit: %
	Trab. Th	The average thickness of the individual trabeculae. Unit: mm
	Trab. N	The number of trabeculae per unit length. Unit: mm^{-1}
	Trab. Sp	The average distance between trabeculae. Unit: mm
	Displacement	the deformation of the bone in response to applied load. Unit: μm
	Reaction force	The force exerted by a constraint on the bone in response to an applied load. Unit: N
	Stiffness	Bone resists deformation in response to an applied force. Unit: N/m
	Yield load	The force where the bone begins to deform permanently. Unit: N
	Maximum load	The highest force the bone can safely withstand before failure. Unit: N
	Work to fracture	The energy bone absorbed before fracture. Unit: N·m

B SUPPORTING EXPERIMENTAL SETTINGS AND RESULTS

B.1 TIBIAE AUTO-SEGMENTATION

Conventional nnU-net as well as emerging foundation models such as MedSAM have demonstrated the improved accuracy and robustness for universal medical image segmentation (4). TotalSegmentator, which is pretrained on the nnU-Net segmentation algorithm with the dataset of 1204 CT examinations, presents a deep learning model to automatically and robustly segment all major anatomic structures on body CT images (10). In this experiment, we use the TotalSegmentator to segment the right tibia from hindlimb in our Recon-CT dataset with variable resolutions by downsizing, and employ our paired Seg-CT dataset as the ground truth. As shown in Figure 2, the left side is the depiction of the 3D render of hindlimb CT from the Recon-CT dataset. After inference, the nnU-Net model successfully identified the left tibia from other part of the hindlimb, including the femur, spine, and hips in all CTs with variable resolutions and most of the tibiae are segmented accurately by the TotalSegmentator, with better segmentation in higher resolution CT when compared with the ground truth tibia from our Seg-CT dataset. This experiment underscores the importance of high resolution dataset for effective bone segmentation. The segmentation of the tibia using the nnU-Net deep learning model hold significant potential for applications in disease characterization, surgical and radiation therapy planning.
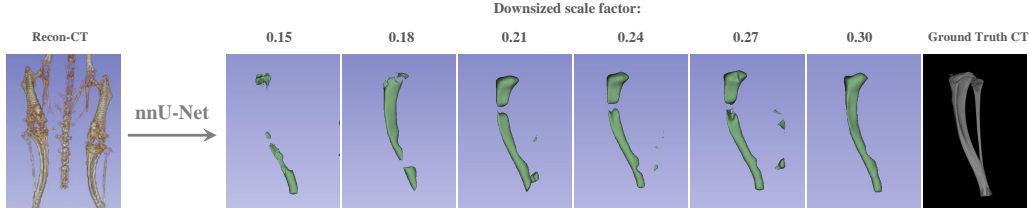
Figure 2: Illustration of tibiae 3D CT auto-segmentation on the Recon-CT dataset from low (left) to high (right) resolutions by adjusting the downsizing scale factor. The right tibia is segmented by nnU-Net. Compared with the ground truth, the TotalSegmentator performs better on high resolution CTs.
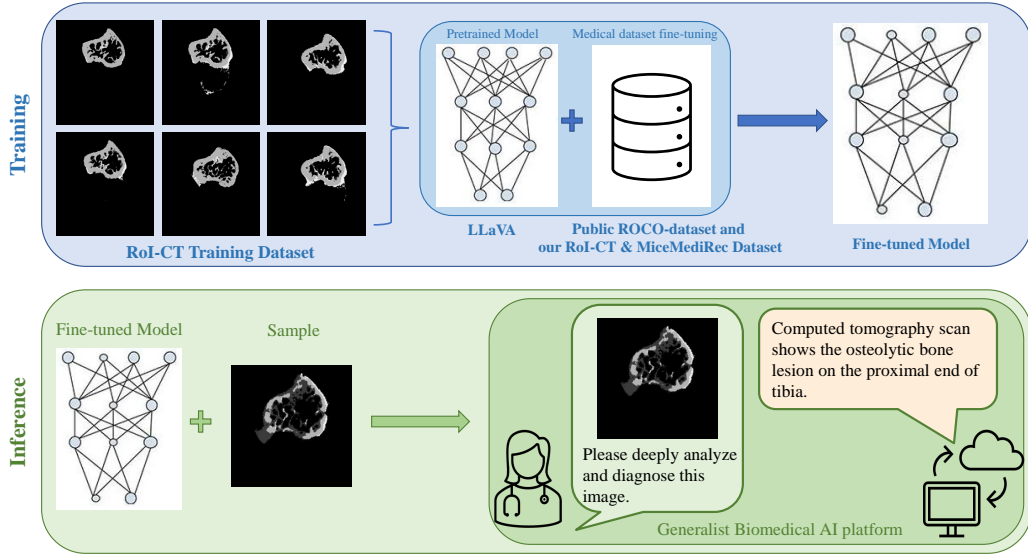

Figure 3: Illustration of the training and inference process, and Generalist Biomedical AI platform for diagnosis of breast tumor bone metastasis given a 2D cross-section image from RoI-CT dataset.

B.2 GENERALIST BIOMEDICAL AI DIAGNOSIS OF METASTATIC BREAST TUMOR ON BONE

Generalist medical AI (GMAI) models can flexibly interpret different combinations of medical modalities, including data from imaging, electronic health records and laboratory results, and in turn produce expressive outputs for disease diagnosis assistance with advanced medical reasoning abilities (6). In this experiment, we used a LLaVA (Large Language and Vision Assistant), which is a combination of the cutting-edge LLaMa 2 text generator and OpenAI's CLiP for image embedding, and fine-tuned with ROCO-dataset (the medical text-image pairs) and our RoI-CT imagery with MiceMediRec text dataset pairs, to generate CT analysis to assist breast tumor bone metastasis diagnosis (2) (5). As shown in Figure 3, the Web interface demonstrates the ability of fine-tuned LLaVA model to generate the diagnosis given a 2D cross-sectional image with breast tumor bone metastasis from RoI-CT dataset.

We have justified the selection of benchmark methods based on their relevance and performance in similar tasks before, as well as their compatibility with the BoneMet dataset. Additionally, we plan to include more comparative analyses with alternative methods in future work to further evaluate the dataset and benchmarks.

4

C   THE DETAILS OF THREE APIS AND THEIR USAGE EXAMPLES

C.1   CT IMAGE SEGMENTATION.

This API provides a simple interface to segment the 3D Reconstructed CT (Recon-CT) images into separate CT scans for the spine, left tibia, left femur, right tibia, and right femur. It can handle individual or batched segmentation of the Recon-CT scans. The API reads the 3D CT scans, identifies the appropriate indices to split the images, and saves the segmented scans to the specified output paths. Given the time point (e.g., the week after tumor inoculation), the input folder path, and the output folder path, Figure 4 exhibits how to utilize the CT Image Segmentation API to automatically segment the tibiae.

```python
config = {
    "week": " week 0",
    "masterfolder": r"F:\Recon-CT\week 0",
    "masterout": r"F:\Seg-CT\week 0"
}

splitter = ReconCTSegmentation(config)

# Split a single image
input_folder = r"F:\Recon-CT\week 0\871"
image_title = "871"
splitter.split_image(input_folder, image_title, config["masterout"])

# Split multiple images
for folder in os.listdir(config["masterfolder"]):
    if folder[0:10] in [871, 872, 873, ...]:
        input_folder = os.path.join(config["masterfolder"], folder)
        image_title = os.path.basename(folder)[0:12]
        splitter.split_image(input_folder, image_title, config["masterout
    "])
```

Figure 4: Example of our CT Image Segmentation API.

C.2   CT IMAGE REGISTRATION.

This API helps researchers with the tibia registration on Seg-CT dataset. It can handle individual or batched registration of the segmented tibiae CTs. The API loads the reference and target CT scans, performs initial transformation, and registers the target CT scan to the reference CT scan. Then the registered CT scan and the transformation are saved to the specific output folder. Given the time point (e.g., the week after tumor inoculation), the slices range of reference and target subjects, the input folder path, the reference folder path, and the output folder path, Figure 5 illustrates how to utilize the CT Image Registration API to automatically align the segmented tibiae.

C.3   RoI-BASED CT IMAGE CROPPING.

This API provides a simple interface to crop the region of interest (tibia proximal end) on Regist-CT dataset. It can handle batched cropping of the Regist-CT dataset. The API reads the overlapped 3D Regist-CT composite processed by our python package, identifies the proximal tibia-fibular junction, selects appropriate indices to split the images, and saves the cropped to the specified output paths. Given the input folder path, the output folder path, and index of the first selected slice below the tibia-femoral junction, Figure 6 demonstrates how to utilize the RoI-based CT Image Cropping API to automatically crop the proximal end of tibiae.

D   BAD-CASE ANALYSIS

Regarding to the bad case analysis, we conducted the following analyses: label noise and data imbalance. Label noise is not a concern in our primary dataset, as positive and negative labels are

```
config = {
    "workspace": r"F:\Seg-CT\week 0",
    "outputdir": r"F:\Regist-CT\week 0",
    "refdir": r"F:\reference",
    "img_z_range": [None, None],
    "ref_z_range": [None, None],
    "initial_transform_angles": [np.pi * i / 16 for i in range(-16, 10)],
    "BASELINE_REG": True, # week 0 (True) or sequential scans (False)
}

# Initialize the registration instance
registration = CTRegistration(config)

# Register a single CT scan
input_folder = r"F:\Seg-CT\week 0"
ct_id = "871 week 0 left tibia"
week = 0
output_folder = config["outputdir"]
registration.register_ct(input_folder, ct_id, week, output_folder)

# Register a batch of CT scans
input_folder = r"F:\Seg-CT\week 0"
ct_ids = ["871 week 0 left tibia", "871 week 0 right tibia", "872 week 11
    left tibia", ...]
week = 0
output_folder = config["outputdir"]
registration.batch_register(input_folder, ct_ids, week, output_folder)
```

Figure 5: Example of our CT Image Registration API.

```
# Configuration
config = {
    "foldername": "selected 300 slices below proximal Tibia-fibular
    junction",
    "first_slice_selected": "first slice selected",
    "last_slice_selected": "last slice selected",
    "first_slice_selected_below_t-f_junction": 0  # Index of the first
    selected slice below the tibia-fibular junction
}

# Initialize the RoICropper
cropper = RoICompositeCropper(config)

# Crop the RoI from CT images
input_folder = r"F:\Regist-CT\Tibia w0w5composite"
output_folder = os.path.join(input_folder, config["foldername"])
first_slice_selected = config["first_slice_selected"]
last_slice_selected = config["last_slice_selected"]
first_slice_below_tf_junction = config["first_slice_selected_above_t-
    f_junction"]

cropper.crop_roi(input_folder, output_folder, first_slice_selected,
    last_slice_selected, first_slice_below_tf_junction)
```

Figure 6: Example of our RoI-based CT Image Cropping API.

assigned at the animal level. However, we manually created a separate dataset with noisy labels. To address label noise in this separate dataset, we applied Label Smoothing (8) with a smoothing factor of 0.1, following the DeiT approach (9). The model's accuracy improved from 71.43% (without Label Smoothing) to 95.86% (with Label Smoothing), demonstrating the technique's effectiveness in reducing errors associated with noisy labels.

We also examined the effects of data imbalance in our dataset, which had a negative-to-positive ratio of 5:1. To mitigate the challenges posed by this imbalance, we implemented Focal Loss (3). The model trained with Cross-Entropy Loss achieved an accuracy of 66.67%, whereas the model trained with Focal Loss improved the accuracy to 69.36%.

E   FUTURE POTENTIAL APPLICATION OF BONEMET DATASET

The BoneMet dataset also holds other significant promise for advancing various applications. For example, it can be leveraged to develop foundation models and self-supervised contrastive learning techniques (11), which will enhance the model's ability to learn robust and generalizable features from the dataset without extensive labeled data. Additionally, the dataset can facilitate the prediction of multi-angle X-ray images, providing a comprehensive view of bone metastases from different perspectives with reduced radiation exposure and improving diagnostic accuracy and aid in better visualization of complex anatomical structures. Moreover, the BoneMet dataset can be utilized for finite element analysis (FEA) prediction of metastatic bone mechanical properties (7). By integrating FEA with deep learning models, researchers and clinic doctors can easily predict how metastatic lesions affect the mechanical integrity of bones without running the complicated and time-consuming finite element simulation, which is crucial for the broad application of finite element analysis in assessing fracture risk and planning appropriate treatments. Lastly, the dataset can be used to train 3D CT registration models, such as the Convolutional Neural Networks (ConvNets) and Deep Learning Image Registration (DLIR) framework (1). These models can accurately align 3D CT scans over time or across different imaging modalities, enabling precise monitoring of disease progression and the effectiveness of treatments.

F   ETHICAL STATEMENT

All the animal procedures including cancer implantation and micro CT scans have been approved by the authors' Institutional Animal Care and Use Committee (IACUC). We made every effort to minimize animal suffering throughout the research process. In brief, the pain and stress associated with cancer implantation and cancer growth were carefully monitored (via signs of inflammation, loss of body weight, changes of normal behaviors, and tumor burden) and managed (via administration of painkillers and humane sacrifice). Mice were anesthetized using 3% isoflurane gas, ensuring a deep, stable sleep state that prevented movement during surgery and the imaging process with minimized stress or discomfort. Additionally, to reduce radiation exposure, each mouse received 5 weekly micro CT scans of lower limbs using a Bruker/Skyscan in vivo scanner (Bruker 1276). The advantage of this scanner vs. other commercially available in vivo scanners is the low radiation exposure during each session (< 500 mGy) due to the faster scan speed (900 ms) and fewer projections (260). The radiation dose (< 0.5 Gy) is much lower than the lethal radiation dose for mice (10.5 Gy). Our BoneMet dataset adheres to all relevant regulations, including those concerning animal welfare and data protection. Furthermore, we are committed to following guidelines for ethical AI development, including ensuring transparency in AI model development, avoiding bias, and maintaining data privacy, ensuring that the use of this dataset aligns with best practices in responsible AI research.

G   LIMITATIONS

Although animal models of breast cancers provide valuable insights into disease development and allow testing the accuracy and sensitivity of various diagnosis methods, we acknowledge the limitations of using animal models in general and the specific model adopted in our study. One potential limitation is the difference in physiology and immune response between mice and humans, leading to the dramatic divergence of the disease progression speed (weeks in mice vs. years in humans). The other limitation is that the mice used in the study were inbred mice with highly

homogeneous gene backgrounds, which were maintained under well-controlled living environments. Thus, the animal subjects do not account for the large variability of human diversity regarding gene background, lifestyles (such as diet and physical activities), and underlying health conditions (like Parkinson's disease and obesity). Despite these limitations, the large quantities with detailed labels of animal datasets like ours provide sequential images covering the entire disease development and progression time course, which can be useful to test the feasibility and performance of new diagnosis and prognosis tools prior to human trials.

REFERENCES

[1] Bob D De Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019.

[2] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

[4] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[5] Vir Chau Markus Zhang. Babydoctor. https://github.com/photomz/BabyDoctor, 2023. GitHub.

[6] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[7] Renzo Phellan, Bahe Hachem, Julien Clin, Jean-Marc Mac-Thiong, and Luc Duong. Real-time biomechanics using the finite element method and machine learning: Review and perspective. *Medical Physics*, 48(1):7–18, 2021.

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[9] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[10] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023.

[11] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.