AAAS
BME Frontiers
Volume 2022, Article ID 9763284, 10 pages
https://doi.org/10.34133/2022/9763284

BMEF
A SCIENCE PARTNER JOURNAL

Research Article

Breast Cancer Induced Bone Osteolysis Prediction Using Temporal Variational Autoencoders

Wei Xiong[1], **Neil Yeung**[1], **Shubo Wang**[2], **Haofu Liao**[3], **Liyun Wang**[2] and **Jiebo Luo**[1]

[1]*Department of Computer Science, University of Rochester, Rochester, USA*
[2]*Department of Mechanical Engineering, University of Delaware, USA*
[3]*Amazon Web Services, USA*

Correspondence should be addressed to Wei Xiong; wei.xiong@rochester.edu

Received 1 November 2021; Accepted 14 March 2022; Published 7 April 2022

Objective and Impact Statement. We adopt a deep learning model for bone osteolysis prediction on computed tomography (CT) images of murine breast cancer bone metastases. Given the bone CT scans at previous time steps, the model incorporates the bone-cancer interactions learned from the sequential images and generates future CT images. Its ability of predicting the development of bone lesions in cancer-invading bones can assist in assessing the risk of impending fractures and choosing proper treatments in breast cancer bone metastasis. *Introduction*. Breast cancer often metastasizes to bone, causes osteolytic lesions, and results in skeletal-related events (SREs) including severe pain and even fatal fractures. Although current imaging techniques can detect macroscopic bone lesions, predicting the occurrence and progression of bone lesions remains a challenge. *Methods*. We adopt a temporal variational autoencoder (T-VAE) model that utilizes a combination of variational autoencoders and long short-term memory networks to predict bone lesion emergence on our micro-CT dataset containing sequential images of murine tibiae. Given the CT scans of murine tibiae at early weeks, our model can learn the distribution of their future states from data. *Results*. We test our model against other deep learning-based prediction models on the bone lesion progression prediction task. Our model produces much more accurate predictions than existing models under various evaluation metrics. *Conclusion*. We develop a deep learning framework that can accurately predict and visualize the progression of osteolytic bone lesions. It will assist in planning and evaluating treatment strategies to prevent SREs in breast cancer patients.

1. Introduction

Bone is among the most common sites of cancer metastasis, whereby primary cancers originating from other places such as breast, prostate, colon, and kidney spread to various bones including the spine, hip, and skull. Osteolytic bone lesions, which result from pathological bone loss due to tumor invasion, are developed in around 75% patients with stage IV breast cancer, the most common nonskin cancer among women in the United States [1]. The destruction of bone is driven by the "vicious" cycle between breast cancer cells and bone-reabsorbing osteoclasts, in which one would reinforce the activity of the other [2, 3]. As the result, cancer patients with bone metastasis can suffer from more than two adverse skeletal-related events (SREs) such as severe bone pain, spinal cord compression, and bone fractures during the course of the disease. It is difficult to treat SREs [3].

In current practice, computed tomography (CT) is routinely employed to evaluate bone structure, which shows 95% specificity in detecting bone metastasis with relatively low sensitivity (73%) [4]. To improve the accuracy, CT has been combined with other imaging modalities such as positron emission tomography (PET) or magnetic resonance imaging (MRI) to detect cancer bone metastasis [4]. Despite the advance in detecting cancer bone metastasis using these hybrid imaging approaches, it is still challenging to predict fracture risk, which is crucial for planning treatments and improving clinical outcomes. While prophylactic stabilization has been suggested to increase overall and immediate postoperative survival, the well-accepted Mirels' scoring scheme used in clinical practice lacks reproducibility and

specificity (13%-50%) for long bones with metastatic lesions [5]. Several CT-based fracture risk assessment studies have attempted to address this challenge, for example, by conducting rigidity analysis using 2D transaxial CT images [6] or finite element analysis using 3D volumetric images [7–9]. These approaches, tested on CT images at single time points, could potentially be more useful if future bone structures were accurately predicted. Prior studies have also investigated prediction of SREs using machine learning methods, but limited efforts on early assessment of bone fracture in metastatic bone disease have been made.

Although sequential clinical images may be captured for breast cancer patients, there have been few attempts to utilize prior scans in assessing bone lesions and the associated fracture risk. In a preclinical study, the accuracy of bone metastasis prediction based on one prior dual-modality image was found to be approximately 85% in a rat breast cancer metastasis model [10]. Since bone metastasis lesions often progress with time, a temporal sequence of images may offer more cues leading to better detection of bone lesions. To this end, we have performed a preclinical imaging study in mice to test if a sequence of CT images could be used to predict breast cancer-induced bone destruction utilizing deep learning. The intratibial cancer inoculation model showed the full progression of bone destruction, from trabecular bone loss within the bone marrow to full-thickness perforation of bone cortex, similar to clinical observations in breast cancer patients [11].

Based on the preclinical study and the recent progress on medical image analysis with deep learning methods [12–15], we showed that deep learning methods can be an effective approach to predicting bone metastasis. To achieve this, in this study, we collected a dataset composed of microcomputed tomography (micro-CT) scans of murine skeleton, which were taken at 3-6 successive time points while breast cancer metastasis and bone lesions progressed. Specifically, we aimed to predict and visualize the progression process of bone lesions using a deep learning framework. We benchmarked our model against state-of-the-art future prediction deep models, the 3D Generative Adversarial Networks (3D-GAN) [13, 16], the Convolutional Long Short-Term Memory Network (C-LSTM) [17], and the PredRNN [18].

The main novelty of our approach is that we modeled the bone lesion prediction problem as a video prediction problem [16, 19]. Inspired by the success of recent natural video prediction methods [19], we adopted a temporal variational autoencoder (T-VAE) [20, 21] model that captures both the spatial and temporal patterns of cancer-induced bone osteolysis. We used variational autoencoders to model the distribution of the future states. Moreover, we proposed an edge-aware loss that encourages our model to pay more attention to the valid pixels in the sparse CT slices, as an effective way to address the challenging class imbalance between the foreground class pixels (e.g., the bone) and the background class pixels within the CT image data. Given the bone CT images taken at the first three weeks, our model was shown to predict the progression of bone lesions at the fourth week with an accuracy significantly higher than the benchmark models.

2. Results

2.1. Task Definition. The specific task in this work is that given the first three weeks of murine bone CT scans, we predict and generate CT scan images at the fourth week. Figure 1 shows the visualization of the input transverse/axial CT slices, the ground-truth CT slice at the fourth week, and the generated CT slice by our temporal VAE (T-VAE) model. Comparing our generated images with the ground-truth data, the visual results demonstrate that our model can predict the bone lesions accurately.

2.2. Evaluation Metrics. We evaluate the quality of the generated predictions utilizing the quantitative metrics of peak signal to noise ratio (PSNR), structural similarity index measurement (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [22] score of the generated predicted fourth frame against the ground-truth fourth frame.

2.3. Comparison with Other Models. We compare our T-VAE model with several state-of-the art future frame prediction models: the 3D Generative Adversarial Networks (3D-GAN) [16], the Convolutional LSTM (C-LSTM) [23], and the PredRNN [18] for the slice prediction task. 3D-GAN learns a distribution of future frames with a pair of generator and discriminator composed of 3D convolutions [24–26]. C-LSTM first projects images into feature maps with convolution operations and then learns the temporal information with an LSTM model. PredRNN uses spatio-temporal memory flows to build the architecture for accurate future prediction.

2.3.1. Qualitative Comparison. Figure 2 shows the visual results generated by different prediction models. In Figure 2, there is no lesion in the slices of the first three weeks but lesion may occur in the fourth week. This makes the prediction very challenging. On such cases, all the compared models fail to make accurate bone lesion predictions. In contrast, our model is able to capture subtle temporal progression patterns in the first three weeks and predict the state of the bone image in the fourth week accurately. Notably, 3D-GAN makes a few correct lesion predictions. However, it tends to overpredict the lesion; i.e., the lesion region is much larger than the ground-truth lesion or lesion is predicted but there is no lesion in the ground-truth.

2.3.2. Quantitative Comparison. We also provide quantitative results in Table 1 as a complementary to the visual results. From Table 1, our T-VAE model achieves higher PSNR and SSIM scores and lower LPIPS scores, demonstrating the superiority of our model against the existing models. It is worth noting that both our T-VAE model and the C-LSTM model use LSTMs to describe the temporal information in a sequence. However, our T-VAE model can encode the distribution of the generated data, which is better at dealing with major changes in the temporal progression of bone. Furthermore, our proposed edge-aware loss effectively allows the model to focus on the valid pixels in the images, thus providing more accurate predictions.

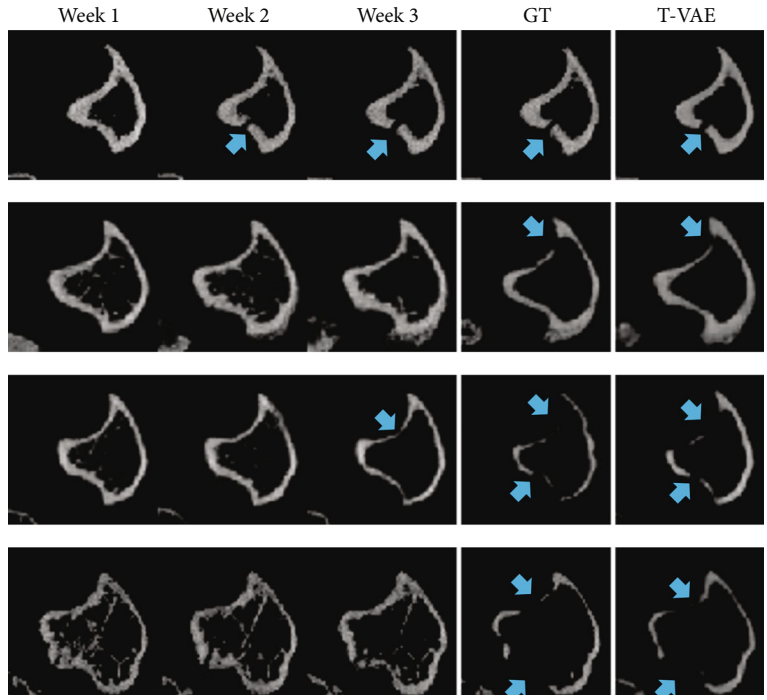

Figure 1: Each row displays the first three weeks of bone CT slices (input), the ground-truth image in week 4, and *our* predicted result in week 4. Blue arrows indicate osteolytic lesions.

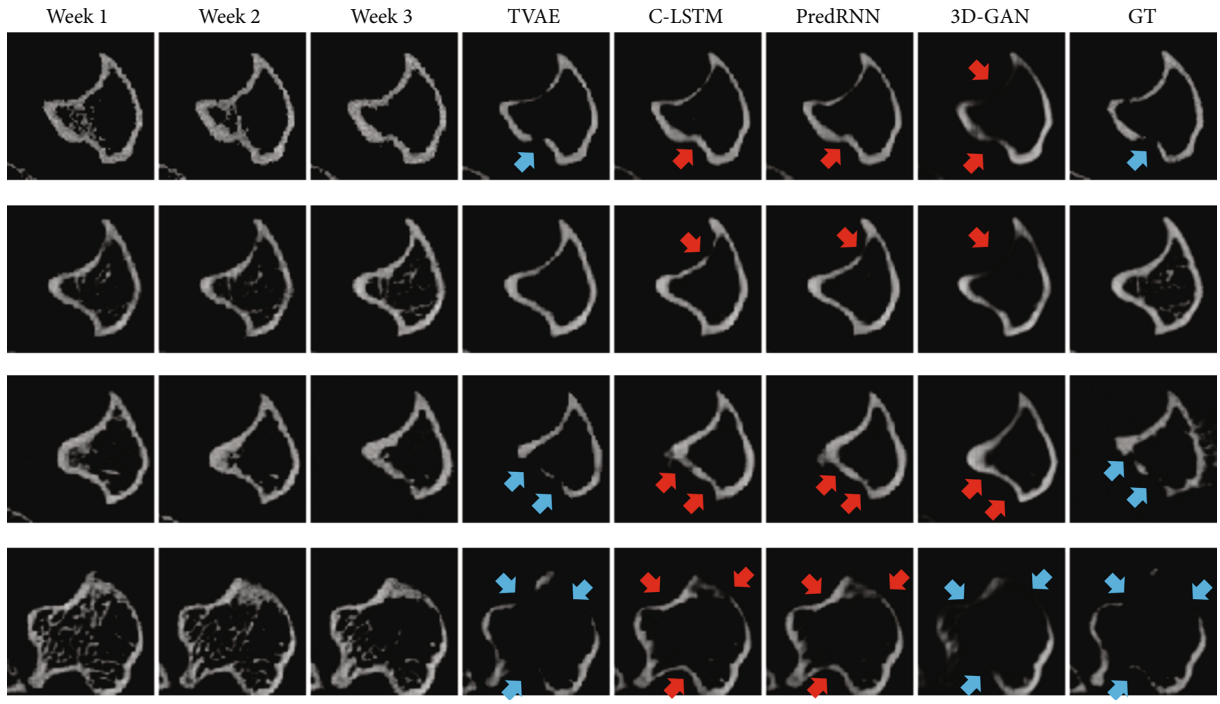

Figure 2: Each row displays the first three weeks of bone CT slices (input), followed with the *predicted* results from T-VAE (2D), C-LSTM, PredRNN, 3D-GAN, and ground-truth image in week 4. Blue arrows indicate lesions in the ground-truth image or correct osteolytic lesion predictions. Red arrows indicate wrong lesion predictions. Please pay special attention to the boundary regions of the bone (i.e., the tibial cortical bone), which are the primary regions to determine whether there is lesion or not. Qualitatively, *our* model T-VAE provides the best results among all the models.

Table 1: Average PSNR, SSIM, and LPIPS score of predictions from different models. ↑ means higher is better. ↓ means lower is better.

Model	PSNR ↑	SSIM ↑	LPIPS ↓
C-LSTM	22.81	0.786	0.087
3D-GAN	21.98	0.760	0.095
PredRNN	22.52	0.798	0.089
T-VAE (2D)	23.44	0.816	0.079

2.4. Ablation Study. We conduct an ablation study to evaluate the effects of different components and loss functions. Specifically, we compare our model using the proposed *edge-aware* loss with the model using a plain *mean square error* (MSE) loss. Note that our full model takes 2D slices as the input and uses 2D convolutions in the encoders and decoders to extract features. As an alternative, we can instead adopt a volume composed of 48 2D slices as the input and use 3D convolutions in the encoders and decoders to project the volume. We denote this variant of our model as T-VAE (3D) and denote our default version (using 2D convolution) as T-VAE (2D). The quantitative results of different versions of our model are shown in Table 2.

2.4.1. MSE Loss versus Edge-Aware Loss. Concluded from Table 2, utilizing the edge-aware loss during training, the T-VAE (2D) model achieves better scores than using the plain MSE loss. By factoring in the Gaussian mask of the first week's scan using the edge-aware loss, the T-VAE (2D) model is subsequently able to produce sharper and higher quality generations evidenced by the gains in the image reconstruction metrics PSNR and SSIM.

2.4.2. T-VAE (2D) versus T-VAE (3D). We also compared our T-VAE (2D) version with T-VAE (3D) version. The only difference is that for the 2D version, we use 2D convolutions to extract features from each 2D image slice, while for the 3D version, we use 3D convolutions to extract features from each 3D volume (each volume of data contains 48 2D image slices). The encoder and decoder of the T-VAE (3D) model have similar architectures as the generator and discriminator of DCGAN [25]. Theoretically, the 3D model should be better at capturing the volumetric nature of the data than the 2D model. However, in practice, the T-VAE (2D) model outperforms the T-VAE (3D) model in terms of both quantitative performance shown in Table 2 and qualitative performance shown in Figure 3. In Figure 3, T-VAE (3D) overpredicts the lesion as indicated by the red arrows. In contrast, our T-VAE (2D) model predicts the lesion much more accurately. Moreover, volumes predicted by our T-VAE (2D) model contain more content details. These results indicate a sign of overfitting of the T-VAE (3D) model. We suspect the reason could be that the dataset does not have enough 3D volumes as training data in comparison to the training set of 2D slices, even when data augmentation is used. We plan to obtain more 3D volumes to address this issue.

Table 2: Quantitative results of ablation study. ↑ means higher is better. ↓ means lower is better.

Models	PSNR ↑	SSIM ↑	LPIPS ↓
T-VAE (2D)+MSE loss	22.57	0.785	0.083
T-VAE (2D)+edge-aware loss	23.44	0.816	0.079
T-VAE (3D-DCGAN)+MSE loss	22.21	0.790	0.084
T-VAE (3D-DCGAN)+edge-aware loss	22.77	0.782	0.087
T-VAE (3D-UNet)+edge-aware loss	22.87	0.774	0.085

To make the ablation study for the model architecture more solid, we include the result of another 3D version of our model. Two recent works [13, 27] have used powerful 3D-GAN models for 3D data generation and achieved promising performance. Both of them have used a 3D U-Net as the generator. We hypothesize that a more effective architecture could compensate for the lack of data. The 3D U-Net is such a design that uses full skip connections ("full" means each layer in the encoder is connected to the corresponding layer in the decoder) to transfer knowledge from the encoder to the decoder, leading to a more effective usage of data. Therefore, we follow their architecture and use a 3D U-Net for volume modeling instead of the plain 3D encoder-decoder architecture. The results are shown in Table 2. Using a 3D U-Net instead of the vanilla DCGAN architecture can improve the PSNR and LPIPS scores of our T-VAE 3D version. The reason could be that more content features are transferred from the previous frames to the future frames.

2.5. Diversity of Prediction. Our model contains a stochastic component that samples the latent vector z_t from an encoded latent distribution. Such a design is based on the stochastic nature of the progression of bone lesion. Given the bone slices of the previous weeks, there is still reasonable uncertainty on the bone state of the next week. The distribution intuitively reflects the range of plausible outcomes of the bone progression based on the previous weeks' scans x_1, x_2, x_3. Depending on the choice of z_t, the constitution of the resulting prediction x'_4 can vary as shown in Figure 4. Interestingly, given different z_t, the boundary regions (the tibial cortical bone) of our predicted bone remain largely unchanged. Such regions are the key factors for a clinician to determine if there is lesion in the bone or not. In the meanwhile, different contents are generated in the nonboundary areas (the bone marrow regions), showing the diversity of our results.

Figure 4 also shows the uncertainty map of the diverse predictions. The intensity of each pixel in the uncertainty map indicates the probability of the predicted pixel in that location to be nonzero. From the uncertainty map, we observe that pixels in the boundary regions show a very high probability, while pixels in the nonboundary regions show a lower probability; i.e., our model can capture the core temporal patterns and generate pixels with a very high confidence and low uncertainty in the important regions (boundary regions). Our model also enables the diversity by generating pixels with

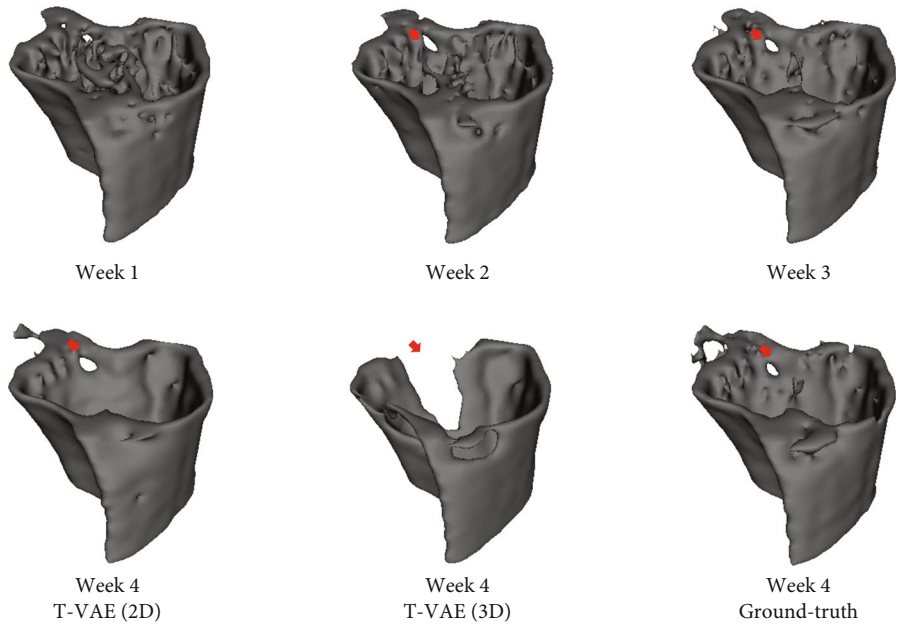

Figure 3: Visualization of the predicted volume of T-VAE (2D) and T-VAE (3D). T-VAE (2D) predicts the accurate locations and size of the lesion, while T-VAE (3D) overpredicts the lesion region; i.e., the lesion region is significantly larger than the lesion in the ground-truth. Red arrows indicate the lesion regions.

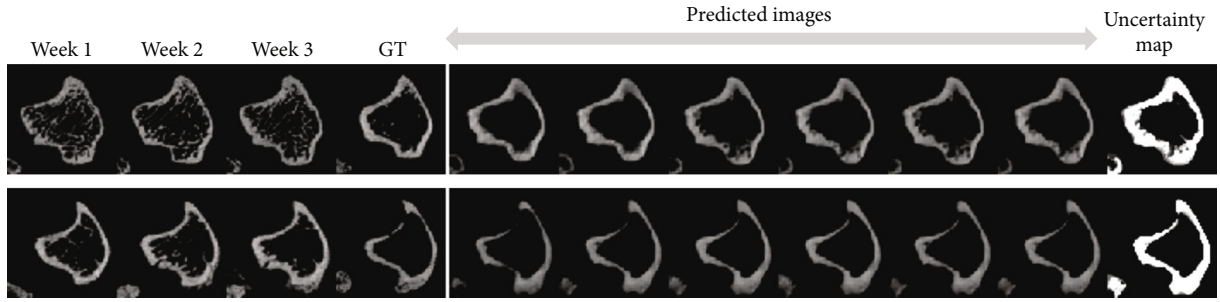

Figure 4: In each row, we show the slices of the first three weeks, the ground-truth slice in the fourth week, six predicted slices by sampling various z_t of our model, and the uncertainty map of the predicted slices. We observe the primary part of the image; i.e., the boundary of the bone (the tibial cortical bone) remains unchanged, while the other parts (the bone marrow regions) are diverse. Please zoom in to see the details of the predictions and the uncertainty map.

a relatively lower confidence (higher uncertainty) in the non-boundary regions.

2.6. Failure Cases. Figure 5 shows two cases where our model does not make very accurate predictions. In the first row, our model is able to predict two of the lesions but fails to predict all the lesions. In the second row, our model fails to predict the lesion. In summary, our model sometimes cannot predict the lesion in a very accurate way when the lesion occurs in the fourth week, but there is no lesion in the first three weeks. We will address these challenging cases in our future work.

3. Discussion

In this study, we showed that bone lesions can be reasonably predicted and visualized by utilizing deep learning models. Given the scans from the previous three time points, our T-VAE model generated diverse and plausible images for the fourth time point. Our model outperformed various future prediction models including the 3D-GAN, the Convolutional LSTM, and the PredRNN, as measured in various reconstruction similarity metrics such as PSNR and SSIM, as well as perceptual metrics such as LPIPS.

During the prediction process of the model, a latent vector, z_t, is sampled from a learned distribution. The sampling of this vector is stochastic. The choice of z_t affects the appearance of the final prediction. On a high level, the latent vector is supposed to represent the stochastic changes within the bone cortex between the discrete CT images of each week. Although the diversity of the latent vector can result in a distribution of plausible outcomes of plausible predictions as evidenced by Figure 4, the most important regions such as the boundary regions that can indicate the lesions usually remain unchanged or only slightly changed.

Our study explored the potential of using deep generative models to predict the occurrence and progression of

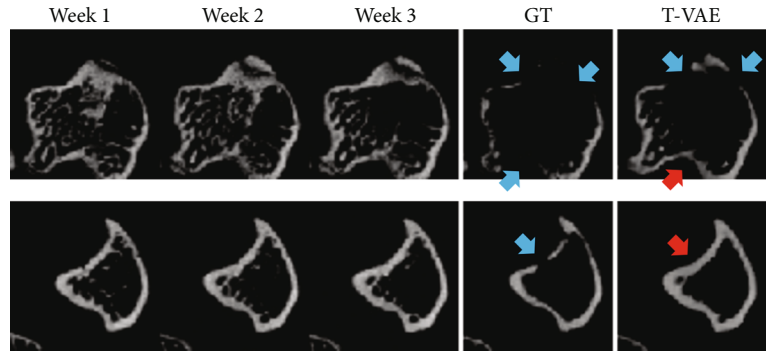

Figure 5: Failure samples generated by our method. In each row, we show the slices of the first three weeks, the ground-truth slice in week 4, and our predicted slice. Red arrows indicate wrong lesion predictions. Blue arrows indicate lesions in the ground-truth image or correct lesion predictions.

bone lesions with a reasonable accuracy using sequential scans in a preclinical CT dataset. Future studies should be performed using datasets containing images from different breast cancer subtypes or other cancer types. The breast cancer used in the current dataset was the aggressive triple-negative breast cancer. The approach will need to be tested on clinical sequential datasets, which are expected to present several technical challenges. Due to the health risk associated with radiation, the scan frequency and subsequently the intervals between scans may vary from patients to patients. How to model temporal patterns with varying time intervals remains a challenge.

Although the resolution of clinical CT scans is typically lower than that of preclinical CT scans, the thicker human bone cortex could compensate this limitation during the detection of bone lesions and their effects on bone structural compromise.

Despite these challenges, the present study provides the foundation of a deep learning framework, which could lead to early detection of osteolytic bone lesions and the ability of predict fracture risk associated with metastatic breast cancer. The approach can be applied to other cancer bone metastasis. Our long-term goal is to assist planning and evaluating treatment strategies to prevent painful or even fatal adverse skeletal-related events in cancer patients.

4. Materials and Method

4.1. Data Acquisition and Preprocessing. The dataset consists of CT image scans of the metaphysis of tibiae of adult female mice, which received either breast cancer cells or phosphate-buffered saline (PBS), followed by weekly scans using a CT scanner with a voxel size of 7 micron. Each bone volume is a data point $x \in \mathbb{R}^{100\times100\times48}$ with height and width 100 and number of slices 48. The data were processed using SITK [28] to register the CT scans. Each resulting 2D slice is a gray-scale, single-channel image. The slice was finally resized to be $64 \times 64 \times 1$ and then normalized. There were 251 mice and approximately 100 of them had no tumor. The ratio of slices with lesion emergence and lack of lesion emergence in nontumor mice was explicitly adjusted in the experiments. Each mouse had 3-6 weekly 3D scans. In the current study, we only considered samples with 4 weeks of data points to ensure temporal consistency.

Our preprocessing protocols are as following. We utilize an 80 : 20 split of the mice that contain only four weeks of data points, resulting in 88 CT volume sequences (each sequence contains four volume data points) as the training set and 23 CT volume sequences as the test set. Basic data augmentations are applied to the training set including random horizontal flips on the training set. For more details on the training parameters, refer to Subsection 4.2.8.

It is worth noting that there is currently no other publicly available datasets that contain CT scans of bone with bone lesions and have temporal sequences of CT scans rather than single discrete scans without temporal progression. Therefore, we do not benchmark our method on other datasets.

4.2. Method

4.2.1. Overview of Temporal Variational Autoencoders (T-VAE). Our model is derived from the recent model [19]. Given data at the previous time steps $x_1, ..x_{t-1}$, the goal of our T-VAE model is to predict data at the t-th step x_t. Our full model is composed of three modules: a prior estimator E_α, a future predictor P_β, and a latent inference model I_γ, where α, β, and γ are learnable parameters of these models, respectively. The prior estimator learns a prior representation that describes the distribution of the data in the current time step. It encodes the uncertainty of the temporal progression of data sequences. The predictor takes the prior and data from the previous time steps $x_1, ..x_{t-1}$ to generate data in the current time step x_t. To facilitate the training, the inference model is used to map the sequence of data to latent representations. An illustration of our framework is shown in Figure 6. We describe the modules in detail as below.

4.2.2. Prior Estimator. The vanilla variational autoencoder adopts the standard Gaussian $\mathcal{N}(0, 1)$ distribution as the prior to model the variance in the data. For temporal data, however, this is not the optimal encoding scheme. One primary reason is that the uncertainty at the current step is

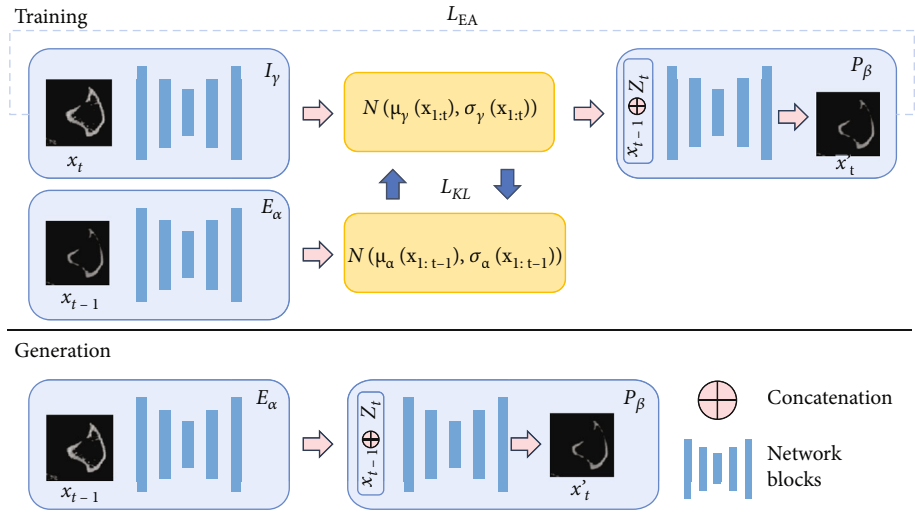

FIGURE 6: The training and generation process of T-VAE.

not fully stochastic but influenced by the states of the history data. Therefore, we introduce the prior estimator model E_α. It is a recurrent model that calculates the prior of each time step $z_\alpha(t)$ as a Gaussian distribution conditioned on data at the previous steps. At each time step t, we have

$$\mathbf{z}_\alpha(t) \sim \mathcal{N}(\mu_\alpha(x_{1:t-1}), \sigma_\alpha(x_{1:t-1})). \tag{1}$$

In the following, we use $\mu_\alpha(t)$ and $\sigma_\alpha(t)$ to represent $\mu_\alpha(x_{1:t-1})$ and $\sigma_\alpha(x_{1:t-1})$ for simplicity, respectively. $\mu_\alpha(t)$ and $\sigma_\alpha(t)$ are calculated using our prior estimator E_α. We have

$$\mu_\alpha(t), \sigma_\alpha(t) = E_\alpha(x_{t-1}, h_\alpha(t-1)), \tag{2}$$

where $h_\alpha(t-1)$ is the hidden state in the prior estimator at the $t-1$ step.

4.2.3. Future Predictor. Our future predictor P_β is a recurrent model based on LSTM networks and the variational autoencoder (VAE). Specifically, it is composed of an encoder to extract features from each frame, an LSTM model to capture the temporal information in a sequence, and a decoder that maps the features back to image frames.

At each time step t, our predictor P takes the hidden state at the previous time step $h_\beta(t-1)$, and the latent representation at the current time step $\mathbf{z}_\gamma(t)$, to generate the current frame x'_t. The hidden state $h_\beta(t-1)$ encodes the temporal pattern among the data at the previous time steps $x_1, ..x_{t-1}$. During training, the latent representation $\mathbf{z}_\gamma(t)$ is computed by the latent inference model I_γ, which we will formulate later. We have

$$x'_t = P_\beta(h_\beta(t-1), \mathbf{z}_\gamma(t)). \tag{3}$$

4.2.4. Latent Inference Model. Our latent inference model I_γ is only used during training. It adopts another LSTM model to learn the temporal patterns in the sequence. The latent inference model is used to calculate the posterior $p(z_t \mid x_{1:t})$ used in the future predictor. We have

$$\mathbf{z}_\gamma(t) = I_\gamma(h_\gamma(t-1), x_t), \tag{4}$$

where $h_\gamma(t-1)$ is the hidden state of the inference model at time step t, representing the history temporal information of the data sequence $x_1, ..x_{t-1}$. x_t is the ground-truth frame at the t-th time step. Similar to the prior in the prior estimator, the posterior representation $\mathbf{z}_\gamma(t)$ also follows a conditional Gaussian distribution. We have

$$\mathbf{z}_\gamma(t) \sim \mathcal{N}(\mu_\gamma(x_{1:t}), \sigma_\gamma(x_{1:t})), \tag{5}$$

where $\mu_\gamma(x_{1:t})$ and $\sigma_\gamma(x_{1:t})$ are the outputs of the LSTM in our latent inference model at time step t.

4.2.5. Learning Objectives. During training, we learn the modules by minimizing the reconstruction loss (implemented with mean square error) between the predicted frame and the ground-truth frame at time step t. We have

$$\mathcal{L}_{\text{recon}} = \sum_{t=1}^{T} (x'_t - x_t)^2, \tag{6}$$

where T is the total number of time steps in each training sequence. Note that using only the $\mathcal{L}_{\text{recon}}$ will result in a deterministic model that cannot provide any stochasticity. The inference model may also degrade to memorize the ground-truth data x_t at the t-th step. To address these issues, we follow the idea of the variational autoencoder and enforce the posterior representation of the inference model $\mathbf{z}_\gamma(t)$ to be close to the prior representation of the prior estimator $\mathbf{z}_\alpha$

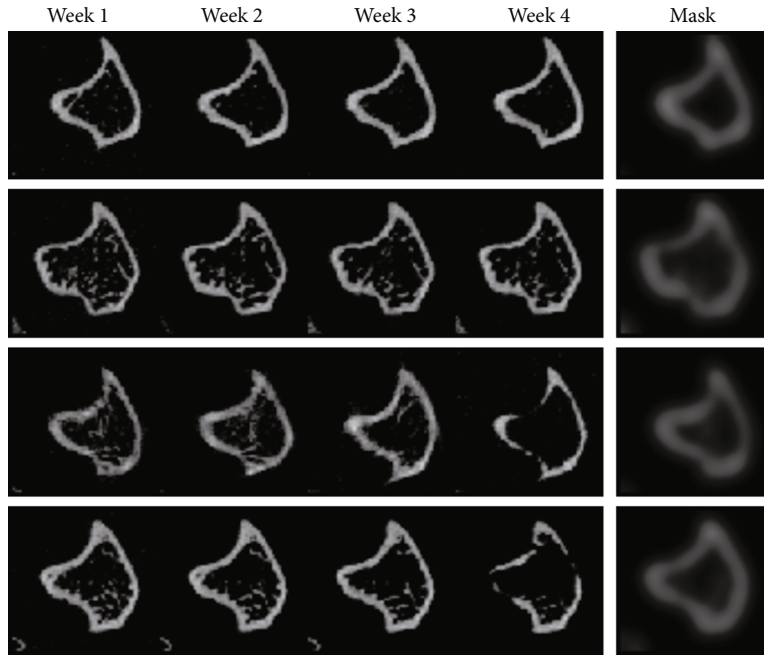

Figure 7: Visualization of the Gaussian masks G that are used in the edge-aware loss.

(t) at each time step t. We use the KL divergence loss to achieve our goal. We have

$$\mathscr{L}_{KL} = \sum_{t=1}^{T} D_{\mathrm{KL}}\left(\mathbf{z}_{\gamma}(t) \| \mathbf{z}_{\alpha}(t)\right), \tag{7}$$

where $D_{\mathrm{KL}}(p\|q)$ denotes the Kullback-Leibler divergence (KL divergence), formulated as

$$D_{\mathrm{KL}}(p\|q) = \sum_{i}^{N} p_i \log_2 \frac{p_i}{q_i}, \tag{8}$$

where p and q both denote probability distribution and q denotes a "target" probability distribution [29].

At each time step, the KL divergence loss forces the inference model to match prior distributions rather than memorizing history data, so that the predictor that is conditioned on the posterior representation can learn new patterns which does not exist in the previous data.

4.2.6. Edge-Aware Learning. Since the pixels of a CT imaging data usually exhibit sparseness, i.e., after normalization, most of the pixels are close to 0 and only a small portion of the pixels are valid. Directly adopting the reconstruction loss in equation (6) can lead to inaccurate prediction or blurry results. To address this issue, we encourage the model to pay more attention to the valid pixels in the CT slices/frames. These pixels form the edges and boundary of the bone structure. A typical solution is to use the focal loss [30]. However, the conventional focal loss is usually used with the cross-entropy loss or other losses with similar forms. It is not proper enough to use it on the reconstruction loss. We propose an edge-aware loss, borrowing the idea from focal loss, but with a rather simple yet effective form. Notice that in our task the significant edges or boundary are usually already provided by data at the first week, as in the first week, the cancer is just injected to the bones, and it takes time for the bone lesion induced by cancer. Therefore, data in the first week usually preserves all the edges and boundary. It is indeed a good mask that can guide the model where to pay attention to. One remaining issue is that data in the first week is too sharp to serve as a mask, which can harm the optimization of the model. To solve this issue, we blur the data using Gaussian kernels and then use it as the mask indicating important regions. We modify the reconstruction loss in equation (6) as follows:

$$\mathscr{L}_{EA} = \sum_{t=1}^{T} (1 + \lambda G(x_1)) \left(x'_t - x_t\right), \tag{9}$$

where $G(x_1)$ denotes to the Gaussian blurred edge mask, λ that ranges from 0 to 1 is the weight for the edge mask, and σ is the standard deviation for the Gaussian kernel. Figure 7 shows examples of our generated masks.

The total loss is a combination of the edge-aware reconstruction loss and the KL divergence loss. We have

$$\mathscr{L} = \mathscr{L}_{EA} + 0.0001 * \mathscr{L}_{KL}. \tag{10}$$

4.2.7. Testing Phase. Notably, the models adopted in the testing phase differ from those adopted in the training phase. The primary difference is that the latent inference model cannot be used, as it uses ground-truth data x_t to calculate the posterior latent representation $\mathbf{z}_{\gamma}(t)$. In the future predictor, instead of using the posterior representation to calculate the final predicted frame at time step t, we sample latent

representations from the prior learned by the prior estimator. We modify equation (3) to the following form:

$$x'_t = P_\beta\left(h_\beta(t-1), \mathbf{z}_\alpha(t)\right). \quad (11)$$

In this way, we can generate a new frame at time step t using data at the previous time steps $x_1, ..x_{t-1}$ and the sampled latents from the learned prior.

4.2.8. Architecture and Training Parameters. Our T-VAE model follows the encoder-decoder architecture of the vanilla VAE. The encoder and decoders of T-VAE use a DCGAN [25] discriminator and generator architecture, respectively [25]. To enrich the content information in the output slices, we connect a few layers in the encoder to the corresponding layers in the decoder (full skip-connections for each layer are not applied). The default version of our model uses 2D convolutions in the encoders and decoders. The input and output of the model at each time step is a single 2D slice. The decoder has a sigmoid output layer. The dimension of the latent vectors is $|z_t| = 10$, and the dimension of the output vector of the encoder is given by $g_{\text{dim}} = 128$. We train the T-VAE model with a batch size of 48, for a max of 200 epochs and early stopping is applied. We use similar training parameters for 3D-GAN, C-LSTM, and PredRNN. During training, we apply a random crop of 96×96 on the slices and then resize the slices to 64×64. During testing, we directly resize each slice to 64×64 and input to the model 2D slices of the first three weeks (all the three slices are from the same location of the given 3D volume), and then, our model generates the slice in the fourth week of that location. The resulting slices in all locations form a predicted 3D volume.

As a variant of our default model, our model can take a sequence of 3D volumes instead of 2D slices as the inputs. We denote this variant as T-VAE (3D). The inputs to this model are the first three weeks of 3D volumes (each composed of 48 2D slices). We replace the 2D convolutions of the T-VAE model with 3D convolutions in the encoders and decoders. The other parts of the model remain unchanged. The output of our model is the predicted 3D volume containing 48 2D slices. We use a smaller batch size of 4 but keep the same number of max epochs.

We use the Adam optimizer with momentum $\beta_1 = 0.9$ and a learning rate of 0.002. We conduct a hyperparameter search on λ and σ values of the Gaussian mask G and find that $\lambda = 1.0$ and $\sigma = 5.0$ yield the best results. The experiments are implemented with PyTorch and run on a NVIDIA 1080Ti GPU.

4.3. Animal and Human Studies. Animal experiments were approved by the Institutional Animal Care and Use Committee (IACUC) and conducted in an accredited animal facility at the University of Delaware. In brief, female C57BL/6J mice (Jackson Laboratory, Bar Harbor, ME, USA) were inoculated with *Mus musculus* mesenchymal-like Py8119 breast cancer cells (ATCC, Manassas, VA, USA, CRL-3278™) through intratibial injection at the age of 14 weeks old. Separate animals injected with PBS served as nontumor control group. Tumor-inoculated animals started developing osteolytic bone lesions 2 weeks after injection. Animals were also subjected to physical activities such as local cyclic compression of the tibia and treadmill running, which regulated tumor growth and the progression of bone lesions. The details of the experimental studies have been published [31]. *In vivo* micro-CT (μCT) scans were conducted weekly for 3-6 weeks using SKYSCAN® 1276 (Bruker, Kontich, Belgium). The imaging settings included 900 ms exposure time, 200 mA current and 50kVp, a 0.5 mm Al filter, and a 7 μm voxel size. Images were reconstructed using NRecon® software (Bruker). For all the images, the volume of interest was the proximal tibial metaphysis of 2.1 mm thick (300 slices) below the growth plate. To ensure consistency of bone orientation, the weekly scans for each bone were registered using its first scan, which was registered using a common reference tibia for all the animals.

Data Availability

The micro-CT data are available on a data server maintained at the University of Delaware. Free download is available upon request.

Conflicts of Interest

The authors have declared no conflicts of interest.

Authors' Contributions

W. Xiong and N. Yeung designed and implemented the deep learning framework and conducted the experiments. W. Xiong, H. Liao, and J. Luo conceived the idea of the edge-aware loss and other implementation ideas. S. Wang imaged and collected the original dataset and helped with visualization of the 3D volumes. L. Wang conceived and designed the animal study, from which the dataset was obtained. J. Luo and L. Wang conceived this study and provided guidance and feedback. All authors contributed to the writing of the manuscript. Wei Xiong and Neil Yeung contributed equally to this work.

Acknowledgments

The animal work and micro-CT scanning was partially supported by the National Institutes of Health (R01AR054385 to L. Wang). The image prediction work was partially supported by the National Science Foundation (1704337 to J. Luo).

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," *CA: a Cancer Journal for Clinicians*, vol. 70, no. 1, pp. 7–30, 2020.

[2] G. R. Mundy, "Metastasis to bone: causes, consequences and therapeutic opportunities," *Nature Reviews. Cancer*, vol. 2, no. 8, pp. 584–593, 2002.

[3] M. Clemons, K. A. Gelmon, K. I. Pritchard, and A. H. Paterson, "Bone-targeted agents and skeletal-related events in breast cancer patients with bone metastases: the state of the art," *Current Oncology*, vol. 19, no. 5, pp. 259–268, 2012.

[4] F. Pesapane, K. Downey, A. Rotili, E. Cassano, and D.-M. Koh, "Imaging diagnosis of metastatic breast cancer," *Insights Into Imaging*, vol. 11, no. 1, pp. 1–14, 2020.

[5] T. A. Damron and K. A. Mann, "Fracture risk assessment and clinical decision making for patients with metastatic bone disease," *Journal of Orthopaedic Research®*, vol. 38, no. 6, pp. 1175–1190, 2020.

[6] A. Nazarian, V. Entezari, D. Zurakowski et al., "Treatment planning and fracture prediction in patients with skeletal metastasis with CT-based rigidity analysis," *Clinical Cancer Research*, vol. 21, no. 11, pp. 2514–2519, 2015.

[7] L. C. Derikx, J. B. van Aken, D. Janssen et al., "The assessment of the risk of fracture in femora with metastatic lesions: comparing case-specific finite element analyses with predictions by clinical experts," *The Journal of Bone and Joint Surgery. British volume*, vol. 94-B, no. 8, pp. 1135–1142, 2012.

[8] A. Sternheim, O. Giladi, Y. Gortzak et al., "Pathological fracture risk assessment in patients with femoral metastases using CT-based finite element methods. A retrospective clinical study," *Bone*, vol. 110, pp. 215–220, 2018.

[9] F. Eggermont, L. C. Derikx, N. Verdonschot et al., "Can patient-specific finite element models better predict fractures in metastatic bone disease than experienced clinicians? Towards computational modelling in daily clinical practice," *Bone & joint research*, vol. 7, no. 6, pp. 430–439, 2018.

[10] S. Ellmann, L. Seyler, J. Evers et al., "Prediction of early metastatic disease in experimental breast cancer bone metastasis by combining PET/CT and MRI parameters to a model-averaged neural network," *Bone*, vol. 120, pp. 254–261, 2019.

[11] L. E. Wright, P. D. Ottewell, N. Rucci et al., "Murine models of breast cancer bone metastasis," *BoneKEy reports*, vol. 5, p. 804, 2016.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science*, pp. 234–241, Springer, 2015.

[13] A. Elazab, C. Wang, S. J. S. Gardezi et al., "GP-GAN: brain tumor growth prediction using stacked 3D generative adversarial networks from longitudinal MR images," *Neural Networks*, vol. 132, pp. 321–332, 2020.

[14] W. Li, V.-D. Nguyen, H. Liao, M. Wilder, K. Cheng, and J. Luo, "Patch transformer for multitagging whole slide histopathology images," in *Lecture Notes in Computer Science*, pp. 532–540, Springer, 2019.

[15] W. Li, Y. Lu, K. Zheng et al., "Structured landmark detection via topology-adapting deep graph learning," in *Computer Vision – ECCV 2020*, pp. 266–283, Springer, 2020.

[16] W. Xiong, W. Luo, L. Ma, W. Liu, and J. Luo, "Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks," in *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2364–2373, Salt Lake City, UT, USA, 2018.

[17] L. Zhang, "Spatio-temporal convolutional LSTMs for tumor growth prediction by learning 4D longitudinal patient data," 2019, https://arxiv.org/abs/1902.08716.

[18] Y. Wang, "PredRNN: A Recurrent Neural Network for Spatio-temporal Predictive Learning," 2021, https://arxiv.org/abs/2103.09504.

[19] E. Denton and R. Fergus, "Stochastic video generation with a learned prior," *CoRR*, 2018, http://arxiv.org/abs/1802.07687.

[20] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013, https://arxiv.org/abs/1312.6114.

[21] D. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[22] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *in Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, Salt Lake City, UT, USA, 2018.

[23] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional LSTM network: a machine learning approach for precipitation nowcasting," *CoRR*, 2015, https://arxiv.org/abs/1506.04214.

[24] I. Goodfellow, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[25] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016, https://arxiv.org/abs/1511.06434.

[26] W. Xiong, Y. He, Y. Zhang, W. Luo, L. Ma, and J. Luo, "Fine-grained image-to-image transformation towards visual recognition," in *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5840–5849, Seattle, WA, USA, 2020.

[27] Y. Chen, A. Jakary, S. Avadiappan, C. P. Hess, and J. M. Lupo, "QSMGAN: improved quantitative susceptibility mapping using 3D generative adversarial networks with increased receptive field," *NeuroImage*, vol. 207, p. 116389, 2020.

[28] Y Z B H R, *SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research*, 2019.

[29] J. Shlens, "Notes on Kullback-Leibler Divergence and Likelihood," 2014, https://arxiv.org/abs/1404.2000.

[30] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *CoRR*, 2017, https://arxiv.org/abs/1708.02002.

[31] S. Wang, S. Pei, M. Wasi et al., "Moderate tibial loading and treadmill running, but not overloading, protect adult murine bone from destruction by metastasized breast cancer," *Bone*, vol. 153, p. 116100, 2021.