Deep learning approach to assess damage mechanics of bone tissue

Sabrina Chin-yun Shen[1,2], Marta Peña Fernández[3], Gianluca Tozzi[4], and Markus J. Buehler[2,5*]

[1] Department of Materials Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, Massachusetts 02139, United States of America

[2] Laboratory for Atomistic and Molecular Mechanics (LAMM), Massachusetts Institute of Technology, 77 Massachusetts Ave. 1-165, Cambridge, Massachusetts 02139, United States of America

[3] School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK

[4] Zeiss Global Centre, School of Mechanical and Design Engineering, University of Portsmouth, PO1 3DJ, UK

[5] Center for Computational Science and Engineering, Schwarzman College of Computing, 77 Massachusetts Ave., Cambridge, Massachusetts 02139, United States of America

*Corresponding author, mbuehler@mit.edu

Abstract

Machine learning methods have the potential to transform imaging techniques and analysis for healthcare applications with automation, making diagnostics and treatment more accurate and efficient, as well as to provide mechanistic insights into tissue deformation and fracture in physiological and pathological conditions. Here we report an exploratory investigation for the classification and prediction of mechanical states of cortical and trabecular bone tissue using convolutional neural networks (CNNs), residual neural networks (ResNet), and transfer learning applied to a novel dataset derived from high-resolution synchrotron-radiation micro-computed tomography (SR-microCT) images acquired in uniaxial continuous compression *in situ*. We present the systematic optimization of CNN architectures for classification of this dataset, visualization of class-defining features detected by the CNNs using gradient class activation maps (Grad-CAMs), comparison of CNN performance with ResNet and transfer learning models, and perhaps most critically, the challenges that arose from applying machine learning methods to an experimentally-derived dataset for the first time. With optimized CNN architectures, we obtained trained models that classified novel images between failed and pristine classes with over 98% accuracy for cortical bone and over 90% accuracy for trabecular bone. Harnessing a pre-trained ResNet with transfer learning, we further achieved over 98% accuracy on the cortical dataset, and 99% on the trabecular dataset. This demonstrates that powerful classifiers for high-resolution SR-microCT images can be developed even with few unique training samples and invites further development through the inclusion of more data and training methods to move towards novel, fundamental, and machine learning-driven insights into microstructural states and properties of bone.

Keywords: Bone; AI; ML; convolutional; neural network; images; microCT; modeling

1. Introduction

Medical imaging and understanding of bone fracture mechanics are critical for detection, diagnoses, and treatment of bone injuries. However, X-ray computed-tomography (CT) and other imaging methods are often difficult to manually interpret as a result of image ambiguity, noise, and other limitations. Diagnostics are even more prone to error due to their reliance on human judgement, which is imperfect and often limited by situational effects[1–3]. In fact, there is a 3-5% estimated error rate in diagnostics involving imaging, resulting in 40 million annual misdiagnoses worldwide with subsequent consequences in patient outcomes and economic fallout[1]. Here, deep learning methods have a clear advantage.

As automated and algorithmic entities, machine learning models can take an objective and comprehensive approach to medical image analysis [4]. Numerous studies have investigated machine learning applications for various disease states, demonstrating their growing potential in detection, classification, segmentation, and other problems [5–7]. With specific regards to bone fracture, Yahalomi et.al. successfully trained a model that detected wrist fractures in X-ray images with an accuracy of 96%, while up to 30% of wrist X-ray images are incorrectly diagnosed in clinical settings [8]. Tanzi et. al. developed a multistage convolutional neural network that classified proximal femur X-ray images into 5 fracture types with average accuracy of 81% and demonstrated that their tool helped specialists to classify fracture images with a 14% increase in accuracy [9]. These are just two of numerous similar studies—nevertheless, there remain large hurdles between state-of-the-art machine learning methods and translation into clinical settings, including lack of data and lack of algorithm robustness [10]. In bone tissue mechanics specifically, there is a lack of fundamental understanding of how, and what, deep neural networks can be designed to decipher the complex minutiae that characterize the mechanical behavior of bone at different dimensional scales, and particularly, damage initiation and progression at the microscale[11]. Key to overcoming these hurdles is continual expansion of datasets, including development of novel datasets, and investigation of various machine learning architectures, including development of strategies that can maximally harness limited datasets.

In this work, we utilize a novel dataset of high-resolution synchrotron-radiation micro-computed tomography (SR-microCT) images of cortical bone and trabecular bone to develop tools that can label mechanical states of bone undergoing uniaxial continuous compression *in situ*[12] (details see Methods). We derive a large dataset by augmenting two-dimensional cross-sections of a limited set of 3-dimensional SR-microCT scans. There are several previous studies that investigate applied machine learning in bone tissue mechanics[6–9], however to our knowledge, no previous work specifically examines such exploitation of high-resolution SR-microCT images derived from controlled experiments, which contain abundant information of bone microstructure but are each individually expensive to obtain. This contrasts with previously investigated datasets that comprise of clinical images, which are not uniformly collected and have limited voxel resolution due to technical and radiation dosage limitations[13]. As such, these microCT scans present an opportunity to gain fundamental understanding of the microstructural properties that characterize bone mechanics. Furthermore, because the SR-microCT scans were collected *in situ* during continuous compression up to failure, they uniquely contain data on various strain states of the tissue that ultimately allow study of bone damage initiation and progression, as well as detection limits of trained classification models.

With this novel dataset, we explore classification efficacy of convolutional neural networks (CNNs) and transfer learning strategies with a Residual Neural Network (ResNet). CNNs are widely regarded as a

break-through technology for computer vision, and therefore are a popular choice for image recognition and classification in machine learning [14]. CNNs function by sequentially sliding sets of small filters over input images to detect specific features in a hierarchical manner, producing activation maps of where features occur. Despite the success of deep CNNs, it has been empirically shown that performance of traditional CNN models is limited by a maximum threshold for depth [15]. This limitation can be alleviated with the use of ResNets, which incorporate skip connections that add outputs from previous layers to outputs of later, deeper layers. Such skip connections allow models to learn identity functions that guarantee each layer performs at least as well as the previous one, easing the training of substantially deeper networks[15].

As deep neural nets have become increasingly advanced and able to solve increasingly complex problems, a new strategy for harnessing their power has emerged: transfer learning. Transfer learning is a method that makes use of knowledge gained while training a model for one problem by applying it to a different, usually related, problem [21], such as applications to predict mechanical properties of biomimetic and biological materials[16–20]. Oftentimes this involves fixing the layers of a pre-trained network such that its weights cannot be updated and replacing only the final layer of the network with a new layer to be tuned to data for the new problem. In this way, previously learned feature extractors can be applied to the new data, reducing the quantity of training data needed for the new problem. Alternatively, the entire network can be set as trainable, simply using the pre-trained model to initialize weights. This is known as fine-tuning and is especially applicable for transfer learning problems where the new dataset is very different from the original one.

Overall, this study aims to systematically investigate application of machine learning strategies in a novel instance of high-resolution image analysis. Such investigation contributes a first step toward bridging the gap between deep learning and bone tissue mechanics with this data, which has the potential to enhance accuracy and efficiency in the characterization of bone deformation and fracture, and ultimately to allow better understanding of the influence of pathological conditions on the biomechanical response of bone tissue.

2. Materials and methods

2.1 Dataset Preparation

The datasets comprise of SR-microCT scans of 5 unique samples each of bovine cortical bone (CB) and trabecular bone (TB) at various strain states, ranging from totally pristine (i.e. unloaded) to totally failed (i.e. maximum compression). The dataset was collected as part of a previous study where continuous SR-microCT images were acquired during uniaxial compression of the bone samples at the Diamond-Manchester Imagining Branchline I13-2 at Diamond Light Source (UK)[22]. The dataset is limited to 5 samples each of CB and TB as a result of experimental costs and beamtime limitations.

Briefly, cortical and trabecular bone were cored from fresh-frozen bovine mid-femoral diaphysis and tibial condyle, respectively, and cylindrical samples (4mm diameter for CB and 6 mm diameter for TB) were extracted in the proximal-distal direction. Prior to the experiment, the ends of the samples were embedded into brass endcaps, achieving a nominal free length of 8 mm and 12 mm for CB and TB, respectively[22]. Samples were mounted in a loading stage (CT5000, Deben, UK) and continuous loading was performed at 0.1 mm/min up to 6-7% apparent strain with SR-microCT images acquired simultaneously. A total of 1201 projection images per scan with an exposure time of 10 ms and 15 ms

for cortical and trabecular bone, respectively, were acquired under load uninterruptedly and resulted in an effective voxel size of 6.5 µm. The projection images were reconstructed into 3D datasets and rigidly registered to the first scan for each sample. 3D reconstructions of the bone samples in their initial (unloaded) states are shown in **Figure 1**, along with their corresponding stress-strain curves under compression. Further details on sample preparation, experimental testing, and image acquisition are reported elsewhere[12,22].

Each microCT scan was labeled with its bone sample name (CB5, CB6, CB8, CB12, CB22; TB5, TB6, TB9, TB11, TB12) and loading state. The registered 3D microCT scans were center-cropped to remove lens framing, histogram matched to normalize brightness, and median-filtered (3x3x3 kernel window) to remove noise, then sliced into 300 2D images from each direction (XY, XZ, YZ). Sample images are shown in **Figure 2**. While total fracture is apparent in some planes, it is not necessarily easy to distinguish in others, such as the image from the XZ plane in the CB examples shown. This ultimately yielded a large dataset of 2-dimensional bone images ranging between totally pristine, partially failed, and totally failed. For the initial classification problem, pristine vs. failed bone, only images from the completely unloaded state ('pristine' bone) and the highest loading state ('failed' bone) for each bone sample were used. Classifiers for cortical and trabecular bone were trained separately in parallel due to their largely differing microstructure[23]. For example, the bone volume fraction ranged from 12.8% to 30.9% in trabecular bone samples and from 96.5% to 97.2% in cortical bone samples.

2.2 Convolutional Neural Network: Structure and Training

Taking the well-known AlexNet architecture as inspiration[14], CNNs were constructed using TensorFlow and Keras, open-source software libraries for machine learning[24,25]. The AlexNet CNN structure consists of alternating convolution+ReLU and MaxPooling layers, then two dense layers with a final softmax activation function for classification[14]. To optimize the CNN architecture for our use case, parameters including number of layers, learning rate, kernel size, and batch size were varied systematically, and the effects on classification of each bone type observed.

For model training, 'failed' and 'pristine' images as previously described were divided into training, validation, and testing datasets. For each bone type, images were divided simply by bone sample; that is, for each experiment, images from 3 biological samples were placed in the training set (5400 images), images from 1 biological sample were placed in the validation set (1800 images), and images from the 1 remaining sample were placed in the testing set (1800 images). This division was chosen because images from the same bone sample display similar bone and crack features, especially when taken from the same plane, so dividing based on bone sample maximally ensures that the model encounters completely new data during validation and testing stages. Furthermore, reserving all images from one bone sample for the testing set enables deeper investigation of model performance, such as classification ability on images where that same bone sample is only partially failed. Images were randomly augmented with horizontal flips or shearing before being fed into the learning algorithm to avoid overfitting and increase variability in training. Models were trained on 20 epochs, enough for training accuracy to plateau.

To evaluate performance of each CNN structure, k-fold cross validation with a validation and test set was utilized, which increases confidence in algorithm performance and eliminates selection bias from random data assignment into training, validation, and testing datasets[26]. As described above, the total datasets for both CB and TB were split based on bone sample, thus k=5 for each (**Table 1**). In k-fold cross validation for a particular CNN structure, one by one, each bone sample was used as the test set. For

each bone sample test set, the remaining bone samples were each sequentially used as the validation set with images from the remaining three (*k-2)* bone samples combined as the training set. In this way, each CNN structure was trained and tested 20 times considering every possible permutation of the 5 bone samples as training, validation, and testing sets, thus accounting for the variations in both microstructure and failure mechanism within the same bone sample type. Then, for a holistic representation of performance, each structure was evaluated based on average accuracy over all 20 trials in the cross validation.

Using model performance results from k-fold cross validation for each variation of the CNN structure, 'optimized' CNN architectures were determined for classification of CB and for classification of TB.

2.3 CNN Visualization & Detection Limits

To gain insight into the optimized CNNs after training, and to ensure that they detect features related to bone mechanics, Gradient Class Activation Mapping (Grad-CAM) was used to visualize how the CNNs determine classifications. The Grad-CAM method improves interpretability of CNNs by generating activation maps that highlight the discriminative regions in an image that a CNN uses to identify a particular class [27]. For each trained model, this involves computing the gradient between the class output and the last convolutional layer in the CNN, then multiplying this gradient with the final convolutional layer to produce a heatmap showing CNN activations on the input image. Because the cross-validation procedures described above produced 20 trained models each, for both CB and TB, the models with the best classification performance from cross-validation of their respective optimized structures were selected for visualization.

To further verify and investigate the CNNs, using the same two models that were visualized with Grad-CAM, we considered fracture detection limits by measuring classification performance on the spectrum of images between the 'totally failed' and 'totally pristine' mechanical states. These images, taken at various degrees of compressive loading, were not used in model training. For each loading state, the trained models were tasked to classify between the 'totally pristine' state, labeled 'pristine', and that specific loading state, labeled 'failed'.

2.5 Mixed Dataset

As described, for the initial classification task, bone images were divided into cross-validation groups based on the bone samples they were derived from. Because this presents a useful but simplistic investigation, the optimized CNN architectures were also trained with the data more thoroughly mixed; i.e. randomly split by both biological sample and direction of slicing. This increases variability in training data, and therefore increases generalizability[28]. Although all images from the same bone sample are related, images from each slicing plane depicts bone and crack features differently, as shown in **Figure 2**, thus vary enough that they can be considered different pieces of data. As a result, a model trained on images from one plane of a sample, and tested on images from another plane of the same sample, is not considered to be tested on the same data it is trained on. Separating by bone sample and slicing plane yielded 15 groups of image data (5 bone samples by 3 directions each) for the CB and TB datasets each, which were randomly divided into 5 sets (3 groups each; **Table 1**). These 5 sets of image data were used in 5-fold cross validation as described above to evaluate performance of the optimized CNNs on these more complex 'mixed' datasets.

2.6 ResNet Transfer Learning

Taking our trained-from-scratch CNNs as baselines, we then investigated the efficacy of transfer learning strategies. Our transfer learning study utilized ResNet-50, a 50-layer deep residual neural network with 5 blocks, each of which contain both convolutional layers and identity layers[15]. ResNet-50 was trained on over a million images from the ImageNet database, and can classify images into 1000 object categories including everyday objects such as keyboard, pencil, and various animals [15]. Using transfer learning with this network allows us to harness the knowledge it possesses while saving computational expense[29]. Constructing the transfer learning model involved loading the ResNet-50 architecture and its corresponding pre-trained weights without its final layer, setting the pre-trained weights to be untrainable, and adding a new dense layer to perform classification for our new problem[30]. Hyperparameters were maintained from the optimized CNNs for consistency. 5-fold cross validation was performed on both the bone-split datasets and the mixed-split datasets for CB and TB described above. For each of these, the transfer learning method was repeated with fine-tuning enabled by setting pre-trained weights from ResNet-50 as trainable. Because transfer learning is expected to decrease the training data and time necessary, transfer learning models were trained on only 5 epochs, enough for training accuracy curves to plateau. The ResNet-50 structure was also trained from scratch (i.e. with randomly initialized weights) over 20 epochs for comparison with transfer learning results and the optimized CNNs.

3. Results and Discussion

3.1 Dataset Preparation

Image post-processing, particularly histogram matching and median filtering, was found to be critical to the development of successful deep learning classifiers. In preliminary trials where the SR-microCT scans were directly cropped and sliced into 2D images without these processing steps, classifiers that were able to distinguish perfectly between images from the pristine state and images from any loaded state were easily trained. Upon visualization with the Grad-CAM method[31], however, it was found that the models were not classifying based on fracture, but simply based on noise that was present in the pristine images and not in the others—likely an artifact of the experimental methods involved in collection of the SR-microCT scans (Appendix A). The post-processing steps taken mitigated this effect, similarly to other image analysis techniques such as digital volume correlation, where image post-processing plays a fundamental role on the optimization of the method[32].

3.2 CNN Optimization and Performance

During CNN optimization, when varying one parameter, all others were held constant at “baseline” values selected from preliminary trials of model training: layers=3, learning rate = 0.0001, kernel size=(5,5), batch size=20. For each trial, the CNN was trained using minibatch gradient descent, Adam optimization, and categorical cross-entropy loss over 20 epochs. **Figure 3** summarizes how model performance, defined by classification accuracy, varied based on several different parameters for the CB and TB datasets. The CB and TB datasets largely displayed similar trends. When varying number of hidden layers in the CNN architecture, classification accuracy increased with increasing number of layers, then decreased slightly after 3 layers for CB and 5 layers for TB. As more layers increase the complexity of the CNN and the features it can detect, this eventual decrease in performance may be attributable to overfitting to the training set[33]. Classification accuracy similarly increased, then

decreased as learning rate increased, with maximums at 0.0001 for CB and 0.001 for TB. This trend is expected, as large learning rates can result in divergence, while learning rates too small can result in getting stuck at local minima or simply taking too long to train[34]. CNN performance varied minimally as batch size was varied, peaking at batch size 10 before decreasing and increasing slowly for CB, and approximately plateauing after peaking at batch size 20 for TB. This contrasts from the typical effect observed, where small batch sizes induce a regularizing effect and better overall performance[35]. Finally, classification accuracy decreased with increasing kernel size, which is reasonable because large kernel sizes introduce more parameters and can make the receptive field too large relative to the original image[36].

From this systematic assessment of CNN architecture, we conclude that 4 hidden layers, learning rate = 0.0001, batch size = 10, and kernel size = (5,5) yields the optimal CNN for the classification of CB fracture SR-microCT scans (Figure 3). These values for learning rate, batch size, and kernel size each produced the best performance when varying the respective parameters. Three hidden layers yielded the best performance while varying layers with the baseline parameters, but produced a model that performed relatively poorly with the other optimized parameters. The 4-layer model performed only negligibly worse during optimization, well within the first standard deviation for overall accuracy in the 3-layer model. Thus the 'optimized' model was changed to have 4 layers, and this structure yielded an impressive average classification accuracy of 0.983 over its cross-validation. This was higher than all but one of the models investigated during cross-validation. The model with baseline parameters and batch size 10 had average classification accuracy of 0.988, however with standard deviation of .017, this difference in performance is negligible.

For TB, the optimal CNN structure was determined to have 5 layers, learning rate = 0.001, batch size 20, and kernel size = (3,3). These parameter values each yielded the highest performing models in their respective optimizations, and together formed a model that achieved average classification accuracy of 0.901 over its cross-validation, higher than any of the models investigated during optimization. The optimized CNN architectures for classification of CB and TB are depicted in **Figure 4a** and **4b**.

These results highlight the intricacies involved in determining the structure of an appropriate deep learning model for different applications. From the investigation of CB, we observed that simple optimization is not always sufficient, and the interplay of different parameters have significant effects on model performance. Furthermore, we observed that different datasets, even when related like the two bone datasets investigated here, can require different machine learning structures for optimal classification performance. This makes sense as different datasets may have differing critical features, such as largely different porosities, which require varying levels of hierarchical feature extraction and different learning parameters to sufficiently capture. Throughout optimization of the CNNs, it was evident that models generally performed better on the CB dataset than the TB dataset. Even after separate optimization for each dataset, classification accuracy was nearly 10% lower for the TB dataset. This further indicates the uniqueness of different datasets and may be attributable to the larger variations in microstructure of TB compared to CB samples. As previously reported, while TB11 and TB12 showed a bone volume fraction above 25% and a plate-like structure, TB6 and TB5 displayed a considerably lower bone volume fraction (i.e. below 17%) and a rod-like structure, which yielded to different failure mechanisms. Trabecular rods experienced significant bending prior to fracture, whereas plate-like structures failed at lower strain levels by buckling. Conversely, the microstructure of the

unique CB samples was fairly consistent (i.e. 3.3 ± 0.3 % porosity, 0.76 ± 0.03 degree of anisotropy), and all samples displayed a longitudinal crack, piercing the volume, after failure[22].

While we have selected the described optimized CNNs, we note that several of the CNN structures investigated produced models with similarly good performances. Further, there are more complex CNN architectures that we did not investigate[37–40]. This means that our selected structures are not the only ones satisfactory for our current study, and it is possible that different structures could produce even better results. However, we emphasize that these results are valuable as proof-of-concept for classification of bone fracture states in SR-microCT scans even with just a few original samples and invite further study of this dataset.

3.3 CNN Visualization & Detection Limits

Figure 3c shows the test set classification results for the best performing trained models from cross-validation of the optimized CNNs for CB and TB. These models were used for visualization and investigation of detection limits. The model trained on CB5, CB8, and CB22, validated on CB6, and tested on CB12 achieved perfect classification accuracy on the test set. The average softmax outputs for both the 'failed' and 'pristine' images were very close to 1, indicating that the model was not only correct, but also confident in its classifications. For TB, the model trained on TB6, TB11, and TB12, validated on TB9, and tested on TB5 achieved test set classification accuracy of 0.986, with softmax outputs close to 1 but lower than those of the CB model, especially for pristine images. To identify the features that contributed most to a specific classification, **Figure 5a** depicts the Grad-CAM heatmaps for a 'pristine' image and a 'failed' image from the CB test set, as well as the heatmaps overlaid on the original images. The trained model appears to universally detect native features of the cortical bone microstructure, such as Haversian and Volkmann canals. In the failed image, the model also detects and is highly activated at the visible crack features. Interestingly, the model is able to distinguish between 'pristine' and 'failed' images very well even though it is activated by bone features present in both image groups, allowing us to speculate that the model may detect differences in bone microstructure under loaded conditions. Detection of crack features is further verified by the detection limits of the model (**Figure 5b**), which shows how its overall classification accuracy is perfect at high loading states, then decreases sharply at Load 11, and hovers at approximately 0.6 before dropping to 0.5 at Load 1 (the unloaded state), where all images the model is tasked to classify are actually 'pristine'. Note that classification accuracy of pristine images at each loading stage is constant because the 'pristine' category images are always the same. As evident in **Figure 4c**, cracks are easily distinguishable at loading states 13 and above, but difficult to identify at loading state 12, and not visible at loading states 11 and below. This corresponds with the drop in classification accuracy and further verifies that the model distinguishes between 'failed' and 'pristine' images based on crack features. Interestingly, because the model is able to distinguish between loaded and 'pristine' images (albeit minimally) even when there are no distinct crack features, it may be able to detect indicators or predictors of failure during the linear-elastic mechanical response of the tissues that are indistinguishable to the human eye.

The Grad-CAM visualizations and detection limits for the TB model are shown in **Figure 6**. Here we see surprising results—similarly to the CB analysis, classification accuracy of the model is very good at higher loading states, then dips when fractures are no longer visible to the human eye. Fascinatingly, however, classification accuracy remains relatively high at lower loading states, even returning to near perfect accuracy at loading states 3 and 4. This is unexpected because the images at lower loading states

resemble the pristine images more and more closely, so it was anticipated that classification accuracy would decrease. Grad-CAM visualization for the 'failed' image reveals that the CNN was activated at some, but not all sites of bone fracture. To further investigate, Grad-CAM was used to visualize model activations throughout the whole spectrum of loading states, shown in **Figure 7a.** The particular slice shown was classified correctly as 'failed' at loading states 3 and 4, but incorrectly as 'pristine' (which was expected) at all other loading states until visible fractures appeared. Strangely, there are no apparent differences in the activations detected in loading states 3 and 4, and the other loading states where the images were classified as 'pristine'. Furthermore, the activated regions do not appear to correspond with bone or fracture features. This contrasts with the Grad-CAM activations for CB, shown in **Figure 7b**. This indicates that while the model has been trained to detect some fractures, evidenced by the dropoff in classification accuracy once fractures are no longer apparent, it also detects and determines classification based on something not obviously visible and potentially not actually associated with bone mechanics or fracture—for example, image artifacts such as the ring artifacts clearly visible in the images[41]. This suggests the need for an even more robust image pre- and post-processing pipeline prior to modelling with machine learning. For TB, the high activation on the left image borders and low activation on the right image borders also suggest some boundary effect as an artifact of imaging or model training, however its origin is unclear as the images were cropped from the center of larger SR-microCT scans prior to training.

3.4 Mixed Dataset

After training with the new mixed datasets, the optimized CNN models for CB and TB achieved average classification accuracies of 0.967 and 0.926, respectively (**Figure 8**). This represents a small decrease in classification accuracy for CB, and a slight increase for TB. The disparity is likely attributable to the different characteristics of CB and TB. Because CB is less porous and did not often show fracture in all three directions, rather demonstrating a predominance of longitudinal cracks[12] , it is possible that some partitions of the data hindered the model's ability in learning to detect all indicators of failure. TB, alternatively, did not exhibit a dominant direction for failure as it resulted from an overall structural collapse following bending of transversally oriented trabecular rods and buckling of longitudinally oriented trabecular plates (with respect to the applied load)[42] , as reported in Fernández et. al.[22] Thus, mixing the data likely exposed the model to a better representation of the entire distribution space of bone images and fracture features.

3.5 Transfer Learning

The transfer learning method involved pre-loading ResNet-50 weights, then training a final dense layer with our bone fracture data. For reference, this network structure has 23,561,152 parameters, only 4,098 of which are trainable. The optimized CB CNN discussed above has 19,265,858 trainable parameters, and the TB CNN has 20,446,018. With both the CB and TB datasets, the transfer learning network was trained with batch size 20, learning rate = 0.0001, minibatch gradient descent, Adam optimizer, and categorical cross-entropy loss. With data split by bone, also shown in **Figure 8**, the transfer learning network achieved average classification accuracy of 0.959 on the CB dataset and 0.951 on the TB dataset over 5-fold cross validation. When the network was trained with fine-tuning enabled, the model achieved classification accuracy of 0.988 for CB and 0.996 for TB. This means that for CB, transfer learning without fine tuning performed worse than the optimized CNN, and transfer learning with fine tuning performed just scarcely better. For TB, transfer learning performed better than the

optimized CNN, and transfer learning with fine tuning performed best, achieving near perfect accuracy. Improvement with fine-tuning enabled makes sense, as fine-tuning allows the models to adapt their feature detectors for attributes specific to the bone fracture problem[43]. Similar patterns emerged with the mixed split datasets for both CB and TB; the transfer learning models performed better than their respective optimized CNNs, and transfer learning with fine tuning performed best of all. Results from these and from training the ResNet structure from scratch are also shown in **Figure 8** for comparison; the ResNet performed very poorly with the CB dataset, and approximately as well as the optimized CNN for the TB dataset. In every case, transfer learning with the pre-trained ResNet and fine-tuning enabled yielded the highest classification accuracy despite having fewer training epochs and similar number of training parameters as compared to the optimized CNNs, demonstrating the power of the transfer learning strategy.

Finally, investigation of detection limits of models trained with transfer learning and fine tuning revealed that the transfer learning models for CB and TB displayed the same patterns as their respective optimized CNNs (**Figure 9**). That is, their classification accuracies are perfect or near perfect until they drop notably at loading states where fractures are no longer clearly visible, then nearly perfectly trace the shapes of the CNN curves as they fluctuate at lower loading states. These similar curves indicate that the unexpected trends of very high classification accuracy at very low loading states are attributable to features in the images, rather than anomalies resulting from model training or overfitting. Interestingly, whether these features were intentional or not (i.e. strain indicators or artifacts from experimental methodology), they are largely indistinguishable from looking at the images, but each of the deep learning models were able to detect and exploit them. This is exciting, as it indicates that deep learning has the potential to harness information on bone tissue local mechanics captured by SR-microCT scans that is not readily available from visual observation, or even to inform development of better image processing methods, though this will indubitably require further investigation into elucidating internal mechanisms of deep learning models.

4. Conclusion

The optimized CNN models and ResNet transfer learning models were all able to achieve extremely high average classification accuracy on both cortical bone and trabecular bone fracture datasets. The transfer learning models with fine tuning enabled uniformly performed the best, achieving average classification accuracy of 0.988 and 0.996 for CB and TB images split by bone sample, respectively, and 0.997 and 0.990 for CB and TB images split by both bone and SR-microCT scan slicing direction. This demonstrates that transfer learning strategies are highly advantageous for saving computing power and making use of even very limited datasets like the one described here. Our results highlight not only the importance of selecting proper deep learning strategies and architectures for a particular dataset, but also emphasize how crucial it is to properly investigate and pre-process the dataset, even when (or perhaps especially when) it is experimentally-derived under apparently well-controlled conditions. Similarly, it is clearly imperative to have accountability mechanisms for explicating how a particular model is working.

These results serve as a first foray into applied machine learning with high-resolution microCT images for greater understanding of bone microstructure and local mechanics. Looking forward, we anticipate making further strides to harness the microstructural and mechanical information contained in this dataset by applying different deep learning methods, such as developing regression models to predict strain levels or utilizing generative adversarial networks (GANs) to predict full-field strain from microCT

images. This, in conjunction with further investigation of how to best process the SR-microCT scans from raw experimental data into appropriate deep learning datasets, will generate more advanced models for understanding and prediction of bone tissue mechanics.

Acknowledgements

We acknowledge funding from NSF GRFP, as well as the IBM-MIT Watson AI Lab, Army Research Office (W911NF1920098), the Office of Naval Research (N000141612333, N00014171-2320 and N000142012189), and AFOSR-MURI (FA9550-15-1-0514). Additional support from NIH is acknowledged (U01 EB014976). We acknowledge Diamond Light Source (DLS) for time in the Diamond-Manchester Imaging Branchline I13-2 under proposal MG22575.

References

1. Itri, J. N., Tappouni, R. R., McEachern, R. O., Pesch, A. J. & Patel, S. H. Fundamentals of diagnostic error in imaging. *Radiographics* **38**, (2018).

2. Hallas, P. & Ellingsen, T. Errors in fracture diagnoses in the emergency deparment - Characteristics of patients and diurnal variation. *BMC Emergency Medicine* **6**, (2006).

3. Waite, S. *et al.* Interpretive Error in Radiology. *American Journal of Roentgenology* **208**, 739–749 (2017).

4. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Medical Image Analysis* vol. 42 (2017).

5. Payan, A. & Montana, G. Predicting Alzheimer's disease a neuroimaging study with 3D convolutional neural networks. in *ICPRAM 2015 - 4th International Conference on Pattern Recognition Applications and Methods, Proceedings* vol. 2 (2015).

6. Pranata, Y. D. *et al.* Deep learning and SURF for automated classification and detection of calcaneus fractures in CT images. *Computer Methods and Programs in Biomedicine* **171**, (2019).

7. Antony, J., McGuinness, K., O'Connor, N. E. & Moran, K. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. in *Proceedings - International Conference on Pattern Recognition* vol. 0 (2016).

8. Yahalomi, E., Chernofsky, M. & Werman, M. Detection of Distal Radius Fractures Trained by a Small Set of X-Ray Images and Faster R-CNN. in *Advances in Intelligent Systems and Computing* vol. 997 (2019).

9. Tanzi, L. *et al.* Hierarchical fracture classification of proximal femur X-Ray images using a multistage Deep Learning approach. *European Journal of Radiology* **133**, (2020).

10. Currie, G., Hawk, K. E., Rohren, E., Vial, A. & Klein, R. Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging. *Journal of Medical Imaging and Radiation Sciences* vol. 50 (2019).

11. Rho, J. Y., Kuhn-Spearing, L. & Zioupos, P. Mechanical properties and the hierarchical structure of bone. *Medical Engineering and Physics* **20**, 92–102 (1998).

12. Peña Fernández, M., Kao, A. P., Witte, F., Arora, H. & Tozzi, G. Low-cycle full-field residual strains in cortical bone and their influence on tissue fracture evaluated via in situ stepwise and continuous X-ray computed tomography. *Journal of Biomechanics* **113**, 110105 (2020).

13. Burghardt, A. J., Link, T. M. & Majumdar, S. High-resolution computed tomography for clinical imaging of bone microarchitecture. in *Clinical Orthopaedics and Related Research* vol. 469 2179–2193 (Springer New York LLC, 2011).

14. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* **60**, (2017).

15. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* vols. 2016-December (2016).

16. Gu, G. X., Chen, C. T. & Buehler, M. J. De novo composite design based on machine learning algorithm. *Extreme Mechanics Letters* **18**, 19–28 (2018).

17. Gu, G. X., Chen, C.-T., Richmond, D. J. & Buehler, M. J. Bioinspired hierarchical composite design using machine learning: simulation, additive manufacturing, and experiment. *Materials Horizons* **5**, (2018).

18. Guo, K., Yang, Z., Yu, C.-H. & Buehler, M. J. Artificial intelligence and machine learning in design of mechanical materials. *Materials Horizons* **17**, 2021 (2021).

19. Yu, C. H., Qin, Z. & Buehler, M. J. Artificial intelligence design algorithm for nanocomposites optimized for shear crack resistance. *Nano Futures* **3**, 35001 (2019).

20. Buehler, M. J. Liquified protein vibrations, classification and cross-paradigm de novo image generation using deep neural networks. *Nano Futures* **4**, 1–12 (2020).

21. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* vol. 22 (2010).

22. Fernández, M. P. *et al.* Time-resolved in situ synchrotron-microCT: 4D deformation of bone and bone analogues using digital volume correlation. *Acta Biomaterialia* (2021) doi:10.1016/j.actbio.2021.06.014.

23. Ott, S. M. Cortical or Trabecular Bone: What's the Difference? *American Journal of Nephrology* vol. 47 373–375 (2018).

24. TensorFlow. https://www.tensorflow.org/.

25. Keras: the Python deep learning API. https://keras.io/.

26. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, (2006).

27. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision* **128**, 336–359 (2016).

28. Therrien, R. & Doyle, S. Role of training data variability on classifier performance and generalizability. in *Medical Imaging 2018: Digital Pathology* (eds. Gurcan, M. N. & Tomaszewski, J. E.) vol. 1058109 5 (SPIE-Intl Soc Optical Eng, 2018).

29. Weiss, K., Khoshgoftaar, T. M. & Wang, D. D. A survey of transfer learning. *Journal of Big Data* **3**, 9 (2016).

30. Transfer learning with TensorFlow Hub | TensorFlow Core. https://www.tensorflow.org/tutorials/images/transfer_learning_with_hub.

31. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision* **128**, 336–359 (2016).

32. Peña Fernández, M., Barber, A. H., Blunn, G. W. & Tozzi, G. Optimization of digital volume correlation computation in SR-microCT images of trabecular bone and bone-biomaterial systems. *Journal of Microscopy* **272**, 213–228 (2018).

33. Deep Learning #3: More on CNNs & Handling Overfitting | by Rutger Ruizendaal | Towards Data Science. https://towardsdatascience.com/deep-learning-3-more-on-cnns-handling-overfitting-2bd5d99abe5d.

34. Smith, L. N. A DISCIPLINED APPROACH TO NEURAL NETWORK HYPER-PARAMETERS: PART 1 – LEARNING RATE, BATCH SIZE, MOMENTUM, AND WEIGHT DECAY. *arXiv* (2018).

35. Kandel, I. & Castelli, M. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express* **6**, 312–315 (2020).

36. Deciding optimal kernel size for CNN | by Sabyasachi Sahoo | Towards Data Science. https://towardsdatascience.com/deciding-optimal-filter-size-for-cnns-d6f7b56f9363.

37. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (International Conference on Learning Representations, ICLR, 2015).

38. Szegedy, C. *et al.* Going deeper with convolutions. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* vols. 07-12-June-2015 1–9 (IEEE Computer Society, 2015).

39. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* vols. 2016-December 2818–2826 (IEEE Computer Society, 2016).

40. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2323 (1998).

41. Vo, N. T., Atwood, R. C. & Drakopoulos, M. Superior techniques for eliminating ring artifacts in X-ray micro-tomography. *Optics Express* **26**, 28396 (2018).

42. Peña Fernández, M. *et al.* Full-Field Strain Analysis of Bone-Biomaterial Systems Produced by the Implantation of Osteoregenerative Biomaterials in an Ovine Model. *ACS Biomaterials Science and Engineering* **5**, 2543–2554 (2019).

43. Guo, Y. *et al. SpotTune: Transfer Learning through Adaptive Fine-tuning*. (2019).

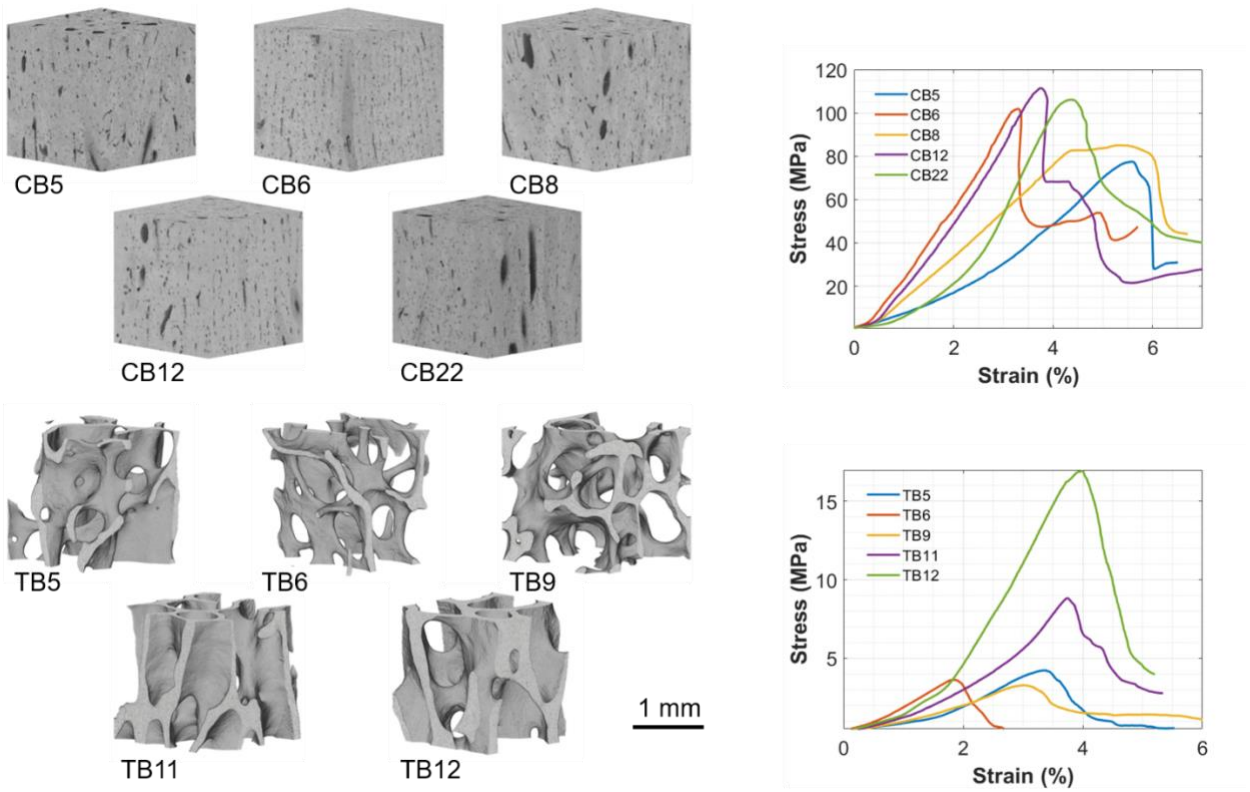


Figure 1: 3D-reconstructed SR-microCT scans of 5 cortical bone and 5 trabecular bone samples with corresponding stress-strain curves from compression testing. Note the varied microstructure between cortical bone and trabecular bone, including bone porosity differences.

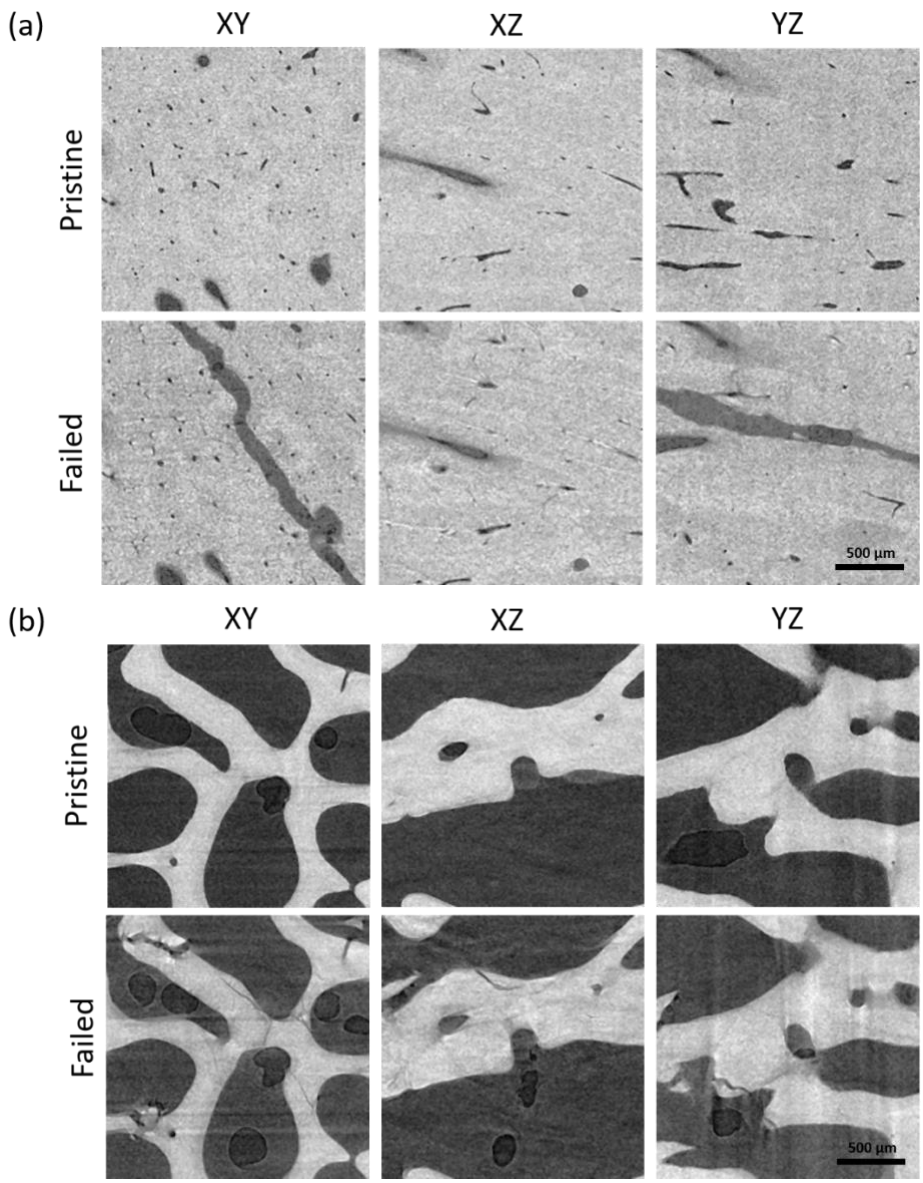


Figure 2: Sample images from one (a) cortical bone sample and one (b) trabecular bone in the datasets. From left to right, images are XY, YZ, and XZ slices from the respective SR-microCT scans for pristine and failed states.

Table 1: Bone image data were divided into 5 groups for k-fold cross validation of model classification accuracy based on (a, b) bone sample and based on (c, d) bone sample and image slicing plane.

(a)

0	1	2	3	4
CB5 XY	CB6 XY	CB8 XY	CB12 XY	CB22 XY
CB5 XZ	CB6 XZ	CB8 XZ	CB12 XZ	CB22 XZ
CB5 YZ	CB6 YZ	CB8 YZ	CB12 YZ	CB22 YZ

(b)

0	1	2	3	4
TB5 XY	TB6 XY	TB9 XY	TB11 XY	TB12 XY
TB5 XZ	TB6 XZ	TB9 XZ	TB11 XZ	TB12 XZ
TB5 YZ	TB6 YZ	TB9 YZ	TB11 YZ	TB12 YZ

(c)

0	1	2	3	4
CB8 YZ	CB5 XY	CB6 YZ	CB5 XZ	CB6 XY
CB12 XZ	CB6 XZ	CB8 XY	CB5 YZ	CB12 XY
CB22 YZ	CB8 XZ	CB12 YZ	CB22 XZ	CB22 XY

(d)

0	1	2	3	4
TB5 XZ	TB5 XY	TB6 XZ	TB9 XZ	TB5 YZ
TB6 XY	TB6 YZ	TB11 XZ	TB11 XY	TB9 YZ
TB9 XY	TB12 YZ	TB12 XY	TB11 YZ	TB12 XZ

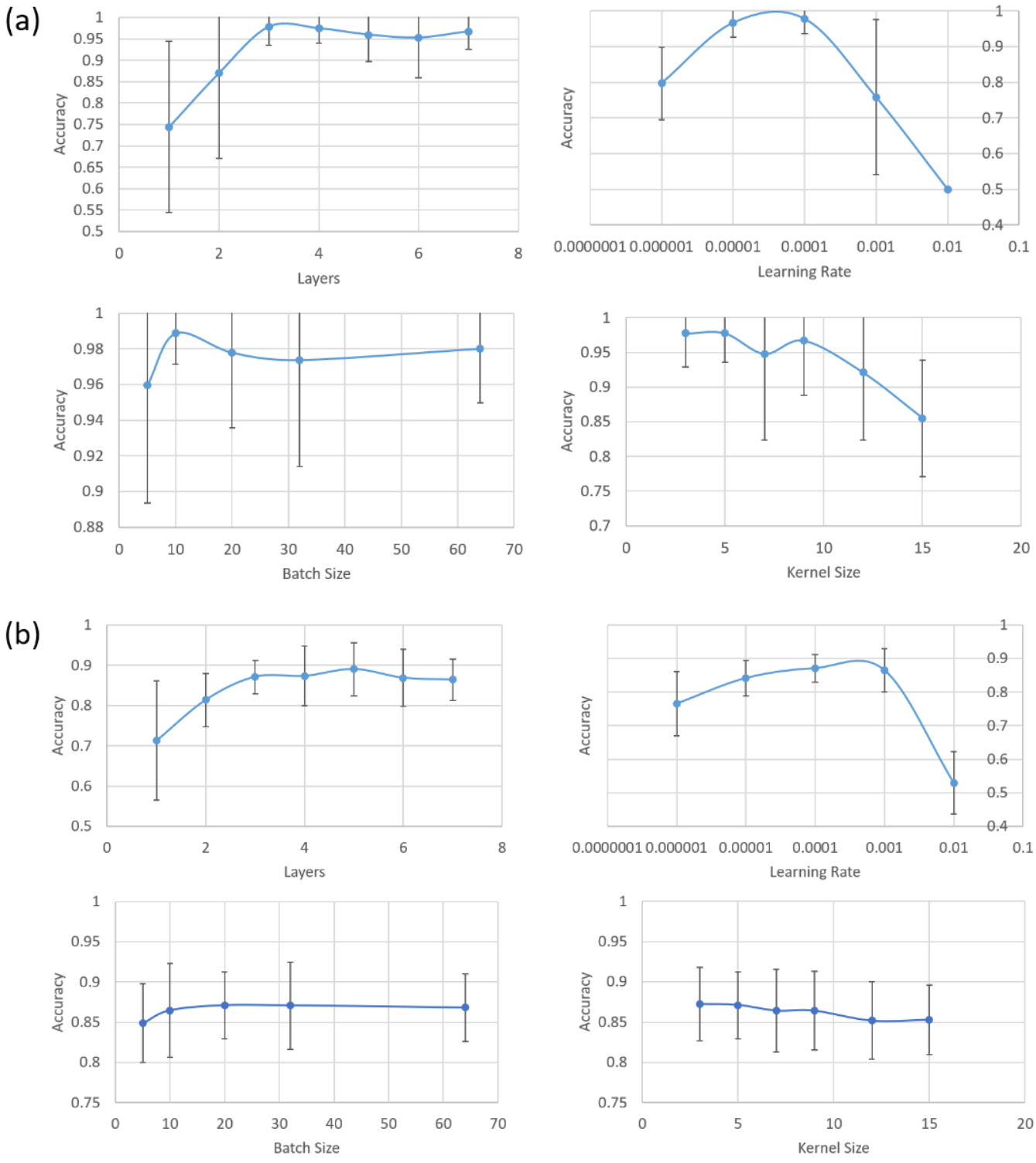
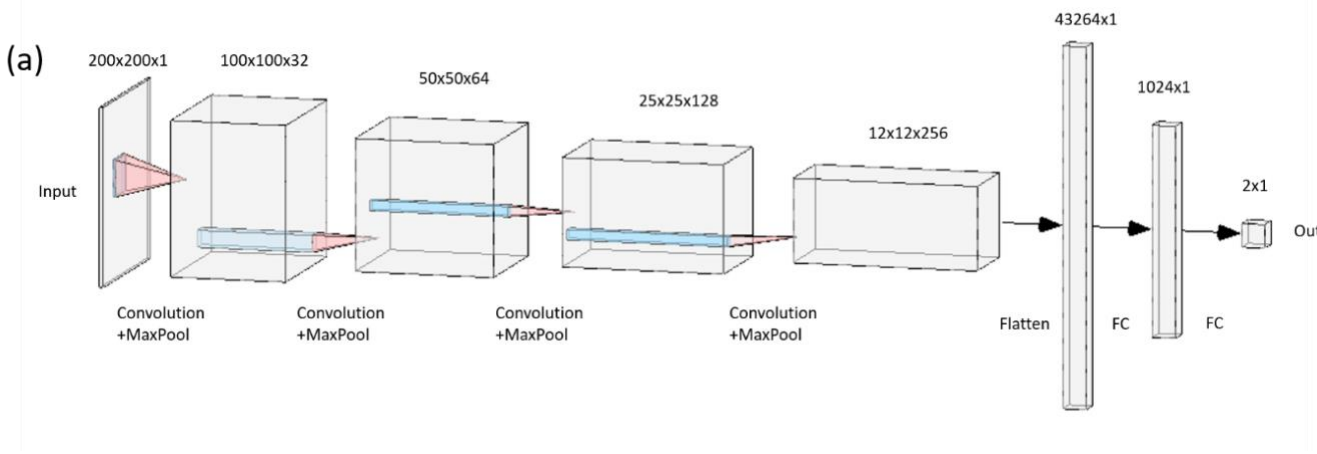
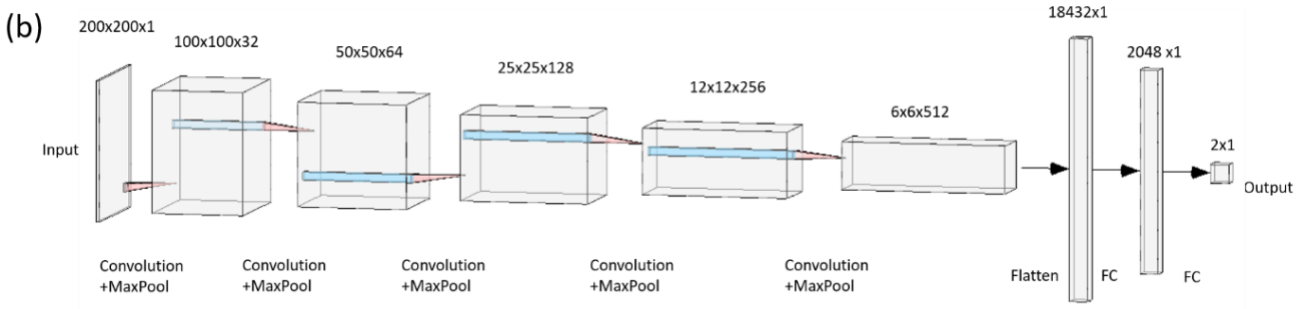


Figure 3: Average classification accuracy of CNNs over 5-fold cross-validation with validation and test set as it varied based on number of hidden layers, learning rate, batch size, and kernel size for the (a) CB dataset and (b) TB dataset. Error bars denote standard deviation.





(c)

	Pristine Acc.	Failed Acc.	Overall Acc.	Failed Softmax	Pristine Softmax
CB	1.000	1.000	1.000	0.999	0.997
TB	0.971	1.000	0.986	0.997	0.959

Figure 4: Optimized CNN structures for classification of (a) cortical bone (CB) images and (b) trabecular bone (TB) images. These AlexNet inspired architectures consist of alternating convolution and MaxPooling layers, then two fully connected dense layers. (c) Classification performance on the test sets for models with the best performance from cross-validation of the optimized structures, including classification accuracy of failed images, pristine images, and overall accuracy, as well as the average softmax outputs for the failed and pristine images which corresponds to prediction confidence.

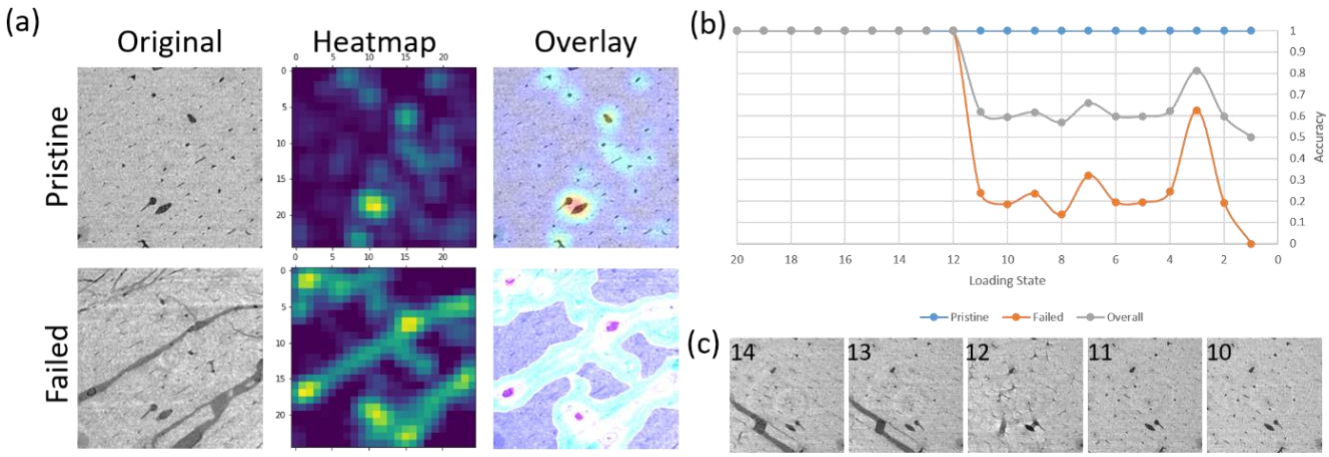


Figure 5: (a) Grad-CAM activation heatmaps for the optimized CB CNN on pristine and failed CB images, and the heatmaps overlaid over the original images. (b) Classification accuracy of pristine images, failed (loaded) images, and overall classification accuracy at all loading states ranging from totally pristine (loading state 1) to totally failed (loading state 20) for CB. (c) Slices of the CB test set bone sample at loading states 10-14; fracture is only easily visible above loading state 12.

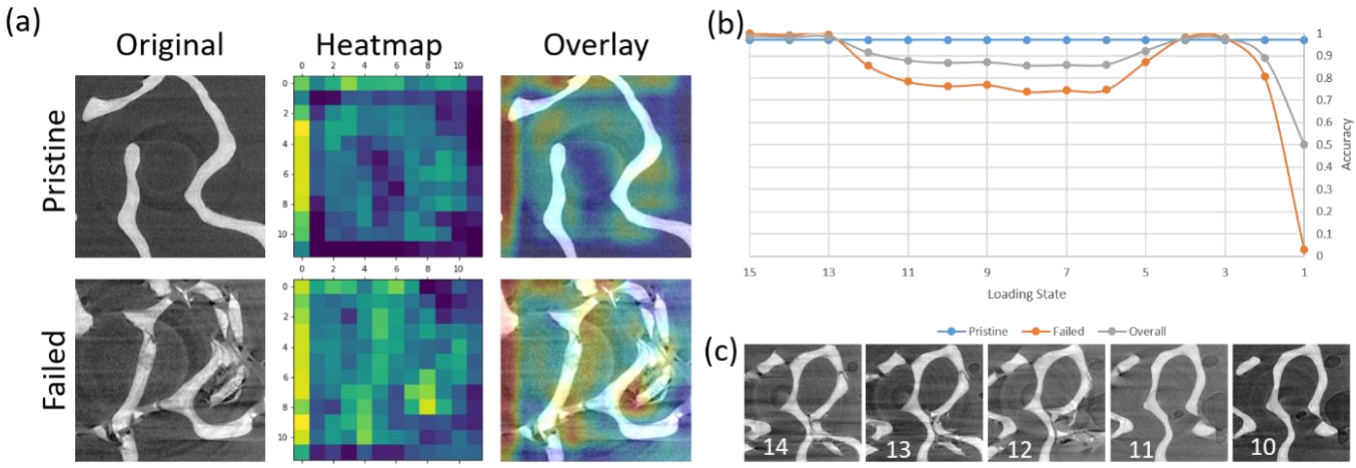


Figure 6: (a) Grad-CAM activation heatmaps for the optimized TB CNN on pristine and failed TB images, and the heatmaps overlaid over the original images. (b) Classification accuracy of pristine images, failed (loaded) images, and overall classification accuracy at all loading states ranging from totally pristine (loading state 1) to totally failed (loading state 15) for TB. (c) Slices of the TB test set bone sample at loading states 10-14; fracture is only easily visible above loading state 11.

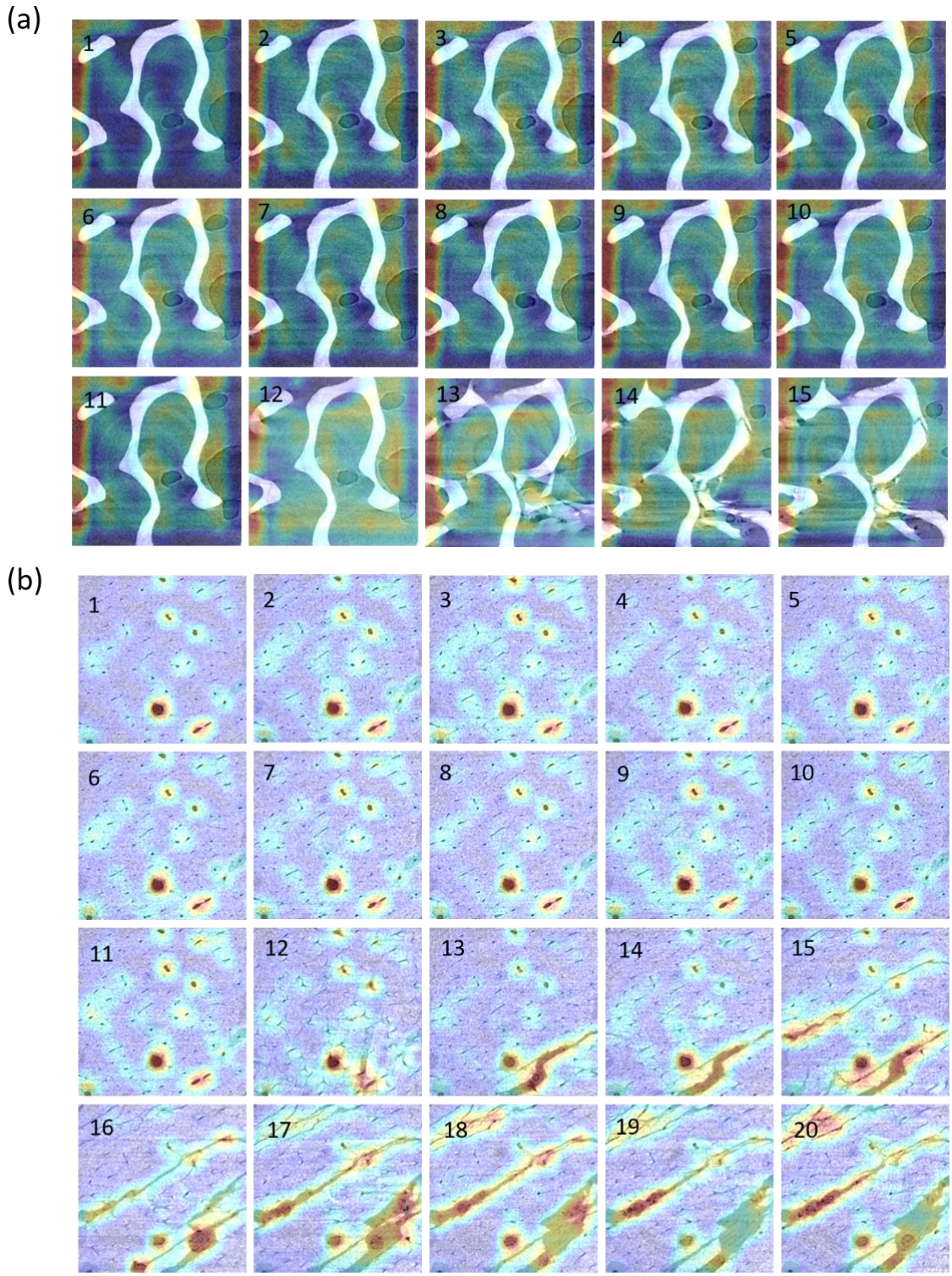


Figure 7: Visualization of Grad-CAM heatmap activations for slices of all loading conditions in (a) trabecular bone and (b) cortical bone.

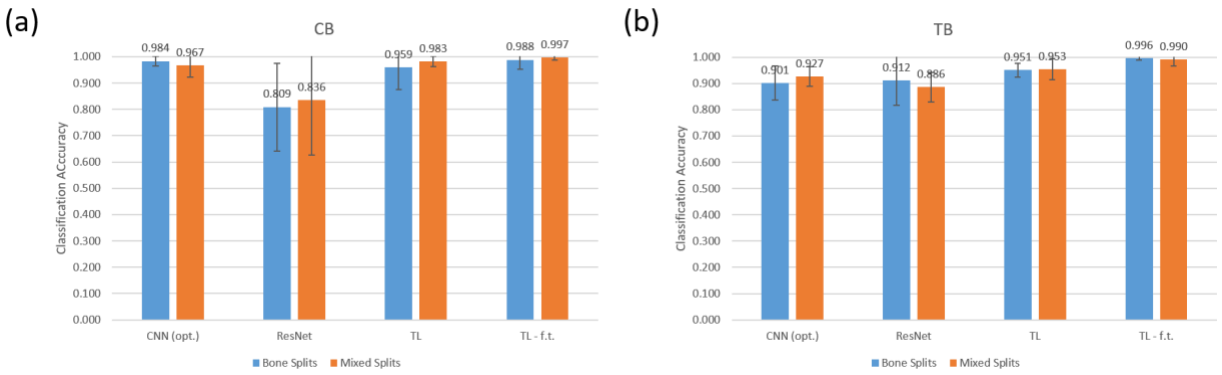


Figure 8: Average classification accuracy over 5-fold cross validation for the optimized CNNs, ResNet trained from scratch, ResNet transfer learning, and ResNet transfer learning with fine tuning for (a) cortical bone and (b) trabecular bone. Error bars denote standard deviation.

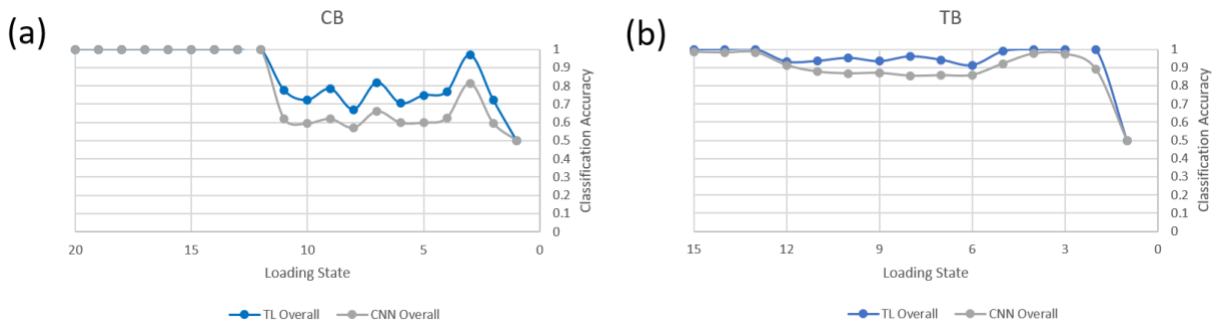


Figure 9: Detection limits of the transfer learning models with fine tuning for (a) CB and (b) TB as determined by overall classification accuracy between pristine (loading state 1) and failed (loaded) images at different loading states. Note the similar patterns between classification accuracy of the transfer learning models and the optimized CNNs across the spectrum of loading states.