UNIVERSITY OF CALGARY

Using Computer Vision to predict Optimal Vitamin D Dosage

by

Sergiu Cociuba

A THESIS

SUBMITTED TO THE CUMMING SCHOOL OF MEDICINE

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF BACHELOR OF HEALTH SCIENCES HONOURS

Bachelor of Health Sciences
Cumming School of Medicine
University of Calgary
Calgary, AB

March 2024

Table of Contents

Contents

Abstract

Machine learning combined with high-resolution peripheral quantitative computed tomography (HR-pQCT) was used to predict the 36-month failure load of the tibia and radius from patients undergoing different vitamin D dosages to identify optimal dosage. Analyzing data from a 311-participant clinical trial with models such as linear regression, support vector regression, random forest, neural network, and a novel dynamic affine feature map transform (DAFT) neural network, this study trains predictive models on biomarker data, bone measurements, and HR-pQCT images. Explainable artificial intelligence (AI) like shapely additive explanations and med cam were employed to explain model predictions. Results show the random forest outperformed all models for predicting failure load of the tibia ,while the support vector regression outperformed all models in predicting failure load of the radius. The DAFT model demonstrated promising potential for multimodal data integration and highlighted the need for larger datasets to train more complex models that have the potential to have better performance.

Acknowledgments

I would like to thank my supervisor Dr. Steven Boyd for giving me the opportunity to conduct research within his lab and for his invaluable patience, guidance, and expertise that helped me successfully complete this project. A special thanks to the members of the Boyd lab for their advice and support that helped me complete this project.

Reflection

I was faced with the difficult task of not just understanding but mastering several machine learning models to optimize and train them effectively. Each model was different in their approach to solving a task, each with its unique mathematical concepts, parameters, and optimization techniques, many of which were complex and foreign to me. I developed a strategy of breaking down complex sections into digestible pieces which enabled a gradual but solid understanding of the mechanisms driving each model. I learned to appreciate the nuances of how each model optimized its loss, an insight crucial for their effective training and optimization.

Visual aids played an important role in my learning process. Explanations were significantly easier to understand by visual representations, especially through YouTube tutorials that portrayed concepts like hyperplanes in support vector regression and convolutional layers in neural networks. These visual insights were important in making learning the models less abstract, making it easier to understand.

The best part of this project was when I finally managed to figure out how to get med cam to work. I spent several weeks figuring out how to properly set it up into my model, obtain the attention maps, and then upscaled them. Even after that, I had to figure out how to use ITK snap to overlay the images, then write code to group the pixel intensities so I could make my own heatmap. The documentation for med cam only provided the generation of the attention map, but not how to properly visualize this. This was both the most challenging part of the project, along with the best part because once this difficult task was completed, the sense of fulfilment was nice.

The discovery of the DAFT model was massive in this project, a find that aligned perfectly with my project's needs. This experience of relentless pursuit, learning, and eventual success has imbued me with valuable lessons and knowledge that I will carry forward into my master’s program. Here, I will apply these insights to create machine learning models aimed at optimizing hospital wait times, among other future endeavors in the field. This project, marked by many challenges that I was able to overcome, has equipped me with technical skills but has also greatly improved my critical problem-solving skills and enhanced my ability to work independently.

List of Tables

List of Figures

Introduction

Osteoporosis is characterized by a decrease in bone mineral density (BMD), having a large impact on global health. The World Health Organization identifies osteoporosis as individuals whose BMD falls under 2.5 standard deviations (SD) or more below the average BMD of a healthy female adult's hips[1]. This measurement, referred to as the T-score, serves as a crucial marker for diagnosing osteoporosis[1]. 2.7 million hip fractures occurred worldwide in 2010 alone, half of which were linked to osteoporosis, highlighting the global impact of osteoporosis[2]. In Canada, recent studies indicate that the prevalence of osteoporosis is around 14.2% to 18.8%, calling for further action to improve detection, treatment, and management for osteoporosis[3]. The clinical consequences of osteoporosis greatly decrease the quality of life of individuals suffering from osteoporosis. They experience financial burdens, disability, and premature mortality. Various risk factors, including sex (with women at a higher risk compared to men), age, lifestyle factors like inactivity, smoking, and excessive alcohol consumption, contribute to the development of osteoporosis[4,5].

Bone is classified into two types: cortical bone, the hard outer layer that provides structural support, and trabecular bone, a porous inner structure involved in metabolic processes such as calcium homeostasis and bone marrow production[1]. Although less dense, trabecular bone contributes to bone strength, flexibility, and shock absorption due to its lattice-like structure[1]. Bone integrity is sustained through a continuous cycle of bone remodeling, which involves the removal of old bone by cells called osteoclasts and the formation of new bone by cells called osteoblasts[1]. After reaching peak bone mass in early adulthood, a balance is typically maintained between bone formation and resorption[1]. However, risk factors for osteoporosis can disrupt this balance, leading to cortical bone thinning and trabecular bone shrinkage, thereby increasing the risk of fractures[1].

Measurements of BMD by dual x-ray absorptiometry remains the gold standard for obtaining a patient's T-score to diagnose osteoporosis[6]. Despite being the gold standard, it cannot assess fracture risk or the efficacy of pharmaceutical interventions because it cannot distinguish between cortical and trabecular bone, assess bone microarchitecture, or assess bone strength[6]. Recent studies have highlighted that bone microarchitecture and mechanical properties like failure load are associated with fracture risk[7,8], indicating a need for better technology over DXA to be able to capture bone microarchitecture and mechanical properties.

High-resolution peripheral quantitative computed tomography images (HR-pQCT) has emerged as a superior technology, offering high resolution 3D images that distinguish the cortical and trabecular bone and measure mechanical properties like failure load using finite element analysis (FEA)[7]. FEA, a computational technique that predicts bone mechanical properties by simulating the bone physical properties and geometry, allowing the calculation of failure load, important for assessing bone quality[7]. Furthermore, studies have shown that HR-PQCT's results are correlated with DXA, indicating HR-pQCT is on par with DXA[8].

Optimal dosage for vitamin D supplementation remains a controversial topic in literature. There are several studies which have attempted to clarify which dosage results in significant improvements in bone health, but many come to contradicting conclusions. For example, some trials have supplemented with 560 and 500 IU of vitamin D reported significant improvements in BMD[9,10], while other studies have concluded that 500 IU of vitamin D results in no significant change in BMD[11]. Furthermore, much of the literature in terms of higher doses of vitamin D indicate no significant change in BMD between higher ($\geq$ 4000 IU) and lower ($\leq$1000) dosages of vitamin D supplementation[11,12,13,14].

Machine learning has revolutionized the prediction and identification of diseases, with algorithms varying in their capacity to understand complex relationships between inputs (features) and outputs (targets)[15]. While linear regressions are more straightforward and less computationally intensive, models like convolutional neural networks, which are more complex and computationally demanding, can discern more complex relationships[15]. Supervised learning models optimize a loss function to adjust model weights, aiming for convergence. Typically, a dataset is divided into training (70%), validation (20%), and testing (10%) portions to refine and test the model, respectively. This process mitigates overfitting, a common challenge where a model memorizes rather than learns from data[15].

Machine learning models like linear regressions, support vector regressions, and random forest excel with tabular datasets, while convolutional neural networks excel on image, text, and audio datasets[16]. The development of multimodal models, or models that combine images and tabular data, represents a promising direction for improving machine learning accuracy and has seen success in recent literature[17].

However, the transparency of machine learning decision-making processes creates a challenge for clinical implementation, with models gaining the term “black box”. Explainable artificial intelligence (EAI) software, like shapely additive explanations (SHAP) and med cam, address this by clarifying how model features impact predictions and visualizing the importance of image pixels, respectively. These tools pave the way for more transparent and justifiable use of machine learning in clinical settings.

Rational

These challenges in determining optimal vitamin D dosage to improve bone health showcase the need for new approaches. A bioinformatic perspective could be taken, and a predictive model capable of predicting bone failure from various bone parameters, biomarkers, and medical images could be created to address this issue. Such a model would lay the groundwork to hopefully create a solution to enable healthcare providers to identify optimal vitamin D dosages for each patient, potentially mitigating the risk of osteoporosis-related fractures and their devastating consequences. This approach not only attempts to address the gaps identified using bioinformatics, but also introduces a personalized medicine dimension, potentially impacting public health by reducing the incidence and severity of osteoporosis.

Consideration of Sex and Gender

This study considers sex, as male and female are used as parameters for the machine learning models, and the impact sex has on the models is captured. Unfortunately, due to lack of data and knowledge in literature, gender cannot be considered for this project.

Objectives

Research Question:

Can machine learning be used to predict failure load of the tibia and radius at 36 months on patients undergoing vitamin D supplementation to identify optimal dosage of vitamin D?

Objective 1:

Train a linear regression, support vector regression, random forest, neural network, and a multimodal neural network to predict failure load of the tibia and radius at 36 months.

Objective 2:

Assess if deep learning using both tabular data and medical images leads to better performance in predicting bone failure of the radius and tibia at 36 months, which will be used to identify optimal vitamin D dosage.

Objective 3:

Use explainable AI software to identify which features and area of the medical images are most important for predicting failure load of the tibia and radius at 36 months.

Methods:

All code can be viewed from this repository: https://github.com/CerJeo-C/MDSC508_Thesis

Dataset used in project:

The dataset used came from a single-center, 3-year, double-blind, randomized clinical trial that took place from August 2013 to December 2017[13]. Three parallel groups received daily doses of vitamin D from 400 IU, 4000 IU, and 10 000 IU[13]. Participants were recruited from the general population using letters, posters, and public media. The age range were healthy men and woman aged 55 to 70 years. Exclusion criteria included a serum Vitamin D level of 30 nmol/L or greater than 125 nmol/l or vitamin D supplementation of 2000 IU or greater in the past 6 months.[13] The inclusion criteria were a DXA scan at the lumbar spine that obtained a total hip areal bone mineral density T score of -2.5 SD, a vitamin D serum level between 30 and 125 nmol/L, and normal serum calcium levels[13]. There were 311 participants in the study, and HR-pQCT images were obtained at baseline and 36 months for tibia and radius[13]. Biomarker and bone measurements taken at baseline and 36 months for both radius and tibia[13]. Bone failure was estimated using FEA[13]. 27 features were chosen based on an initial literature search on bone biomarkers and bone measurements that were correlated with failure load of both the tibia and radius.

Preprocessing:

The first step in preprocessing the tabular data was separating the data from the baseline timepoint and all other timepoints, as the models only require the baseline timepoint feature values. The baseline feature values were extracted into their own comma separated value file format (.csv) and then the 36-month time point for failure load of the tibia and radius was appended to the dataset, serving as the target variable for the models.

A python script was created for the linear regression, support vector regression, random forest, and Dynamic affine feature map transform (DAFT) neural network, that would normalize all tabular features. The data was split into a train, validate, test set with a 70/20/10 split.

For the ResNet model and DAFT model, all image data underwent normalization of all voxels, padding to a size of 800,700,168, and down sampled by a factor of 2 due to computational limits. The tibia and radius image data were then converted into tensor format and pickled (compressed) to save storage space. The image data for both the tibia and radius were separated, each being split into train, validate, and test set with a 70/20/10 split.

All scripts were run on the University of Calgary's high performance computing cluster ARC with a Nvidia A100 tensor core GPU.

Performance Metrics

1) Mean Squared Error (MSE) $= \frac{1}{n}\sum_{i=0}^{n-1}(\mathrm{yi} - \hat{\mathrm{y}}\mathrm{i})^2$

MSE measures the average squared difference between actual and predicted values. It's used in regression problems. The formula involves squaring the difference between each actual value and the prediction produced by the model, then averaging these squares across all observations. Lower MSE values mean better model performance. However, MSE penalizes larger errors more and doesn't indicate direction of the error, only the magnitude of the error is provided.

2) Mean Absolute Error (MAE) $= \frac{1}{n}\sum_{i=0}^{n-1} | \mathrm{yi} - \hat{\mathrm{y}}\mathrm{i} |$

MAE calculates the average absolute difference between actual and predicted values, also for regression. It adds up the absolute differences between each actual and predicted value, then averages these. Lower MAE indicates a better model. Unlike MSE, MAE treats all errors equally, but also doesn't show if the model is over or under predicting.

3) R-squared (R^2): $R^2(y, \hat{y}) = \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}$

R^2, or the coefficient of determination, measures how well the independent variables explain the variance in the dependent variable. It's a ratio that compares the model's performance to a baseline average model. Values range from 0 to 1, where 1 means perfect prediction. Higher R^2 values indicate better performance.

Using all three-performance metrics provides more information about the model, as MSE highlights if the model is weak to outliers, MAE is an indication of robustness, and R^2 provides explanatory power of the model.

SHAP (Shapely additive explanations):

This library will identify how much each feature contributes to each prediction of the target by removing some features to identify its impact[19]. This is done for all combinations of features while training. This provides a score that can be compared to other features to identify which features contribute to the performance of the model the most. This score was obtained from the linear regression, support vector regression, and the random forest. Each feature had 27 SHAP scores for each combination. The absolute means was taken to identify the general impact each feature had on the model.

Training the models:

Five models were trained, with the linear regression, support vector regression (SVR), and random forest being trained on the tabular data, while the ResNet was trained on the HR-pQCT images, and the DAFT model was trained on both tabular data and HR-pQCT images.

For the Linear regression model, the linear regression was trained in a k-fold cross validation loop 5 times. K-fold Cross validation is a way to subset the training data and validation data, to train on k-1 subsamples. This allows the model to be trained with 5 different datasets to assess the performance and generalizability of the model. Then, the entire training dataset was used to train a new model. The performance metrics and SHAP scores were recorded and stored in .csv files when the test dataset was passed into the model.

For SVR and random forest model, the bayes search CV library was used from sci kit learn to optimize the parameters of both these models[18]. Only these two models were chosen as linear regression does not have any parameters to optimize, and neural networks cannot be optimized using this method due to computational restraints and were optimized through trial and error. A similar method to linear regression was employed, with the only difference is that during each fold of the k-fold cross validation, the parameters of the SVR and random forest were being optimized by the bayes search CV method, generating 50 different sets of parameters for each fold, and choosing the best performing set of parameters for that fold. Like the linear regression, once the optimized model was trained, the best fold was taken, and a new model was trained on the entire training dataset with the optimized parameters. The performance metrics and SHAP scores were recorded and stored in .csv files when the test dataset was passed into the model.

For the ResNet model, the images were loaded in using a custom dataset class. This class would unpickle the tensor from a given directory, extract the image and its labels from the pickled file, down sample the image by a factor of 2, and add the batch dimension which is an extra dimension in the tensor that indicates to the model how many images are processed at one time. The model is then loaded, with MSE as the loss function used to optimize the model. Adaptive moment estimation, or Adam, was used to optimize the learning rate as training progressed. The starting learning rate was 0.001. The batch size used was 4. Additionally, early stopping was included, and after 20 iterations, or epochs without improvement, the model will automatically stop training. Furthermore, with each epoch, if a MSE was obtained, the early stopping count would reset back to 0, and the model would be saved. Since k-fold cross validation was used to train the model, each time a better performance was recorded, it was recorded for that fold. Saving the model means that if any issues arose during training, it could be reloaded from its last checkpoint, and training could resume. Additionally, this allowed for loading the best model found during training and passing the test dataset through. During testing, the best model was loaded, and the test dataset was passed through. The performance for MSE,

R^2, and MAE was recorded, along with the med cam activation maps produced for each image in the test dataset.

The DAFT model was trained and tested in a similar way to the ResNet. The difference was that the tabular data was extracted from the pickled tensors, whereas in the ResNet, the tabular data is completely ignored. The other major difference is that there are no med cam results for the DAFT model, as the model combines the image and tabular data within its layers, which results in med cam being unable to produce meaningful results.

For both the ResNet and Daft model, the parameters used to generate the model were optimized manually. Several iterations with different combinations of parameters were used to find the best performing parameters.

Med Cam:

To gain insight into what the ResNet model was learning from the HR-pQCT images, the med cam library was used. Med cam is a library that visualizes the weights associated with pixels found deep in the neural networks[20]. Once the ResNet was fully trained, the test images were passed through, and med cam would generate an attention map for each image or assign pixel intensities based on the weight of the pixels. The attention map was upscaled to the original image size and overlayed on the original image using ITK snap, a visualization software. Using ITK snap, the attention map was treated as a segmentation mask, and the intensities of the pixels were divided into 10 groups and colored, giving a heatmap of important pixels. Most important pixels are red, and yellow, while least important pixels are green, blue, and black.

Results:

Model Performance Results:

Performance metrics of each model for predicting failure load of the tibia are shown in Table 1A. The random forest model outperformed all models for predicting failure load of the tibia at 36 months with an MSE of 848558.80, a R^2 score of 0.82, and an MAE of 517.37. The ResNet model showed the worst performance out of all the models, with an MSE of 1243463.60, a R^2 score of 0.73, and a MAE of 765.43. The linear regression, however, had the lowest R^2 of 0.64. The combination of tabular data and neural network resulted in the DAFT model outperforming most models with a MSE of 897060.00, a R^2 score of 0.82, and a MAE of 544.65. The support vector regression outperforms the DAFT model with a slightly better R^2 score of 0.81, and MAE of 520.74.

Table 1B highlights the performance metrics for each model for predicting failure load of the radius after 36 months. The SVR demonstrated the best performance with an MSE of 106544.18, an R^2 score of 0.87, and a MAE of 192.01. The ResNet model was outperformed by all other models, with an MSE of 161205.94, a R^2 of 0.80, and an MAE of 311.55. The linear regression's performance was almost as good as the SVR, with a MSE of 107575.19, R^2 of 0.87, and a MAE of 193.73. All models trained on the tabular dataset outperformed all neural networks.

Table 2A highlights the difference between testing and training R^2 results of models trained to predict failure load of the tibia at 36 months, indicating which models are overfitting. The ResNet model suffered the most from overfitting, with a difference between training and testing R^2 of 0.26, more than double all other models. The linear regression had the least amount of overfitting, with a difference between training and testing R^2 of 0.09.

Table 1: Performance metrics of the linear regression (LR), support vector regression (SVR), random forest (RF), convolutional neural network (ResNet), and the multimodal convolutional neural network (DAFT) for predicting failure load of the tibia (A) and radius (B) at 36 months. MSE is mean squared error, R2 is R^2 score, and MAE is mean absolute error. Bold indicates the best performance in that metric.

A	MSE	R2	MAE	B	MSE	R2	MAE
LR	915579.40	0.80	586.39	LR	107575.19	**0.87**	193.73
SVR	924506.51	0.80	520.74	SVR	**106544.18**	**0.87**	**192.01**
RF	**848558.80**	**0.82**	**517.37**	RF	115954.55	0.86	223.49
ResNet	1243463.60	0.73	765.43	ResNet	161205.94	0.80	311.55
Daft	897060.00	0.81	544.65	Daft	130516.20	0.84	253.70

The difference between testing and training R^2 results of models trained to predict failure load of the radius at 36 months is found in figure 2B. The ResNet model had the highest degree of overfitting, with a difference between training and testing R^2 of 0.19. The linear regression and random forest model had the least degree of overfitting, with a difference between training and testing R^2 of 0.09.

Table 2: Training and testing R^2 (R2) score for linear regression (LR), support vector regression (SVR), random forest (RF), ResNet, and DAFT model for predicting failure load of tibia (A) and radius (B) at 36 months. Bold indicates least overfitting.

A	Train R2	Test R2	Difference	B	Train R2	Test R2	Difference
LR	0.89	0.80	0.09	LR	0.96	0.87	0.09
SVR	0.96	0.80	0.16	SVR	0.95	0.87	0.08
RF	0.94	0.82	0.12	RF	0.94	0.86	0.09
ResNet	0.99	0.73	0.26	ResNet	0.99	0.80	0.19
Daft	0.94	0.81	0.14	Daft	0.97	0.84	0.13

SHAP Analysis Results:

Figure 1 highlights the SHAP analysis for the linear regression trained to predict failure load of the tibia at 36 months. The SHAP analysis identified baseline failure load of the tibia as the most impactful feature with a score of 1929. Trabecular BMD of the tibia and total BMD of the radius were the next important features, with scores of 711 and 550, respectively. Among the vitamin D dosage groups, 400 IU had a SHAP score of 53, 4000 IU had 38, and 10 000 IU had 16. Sex showed an equal influence on the model with both male and female features scoring 71. The least impactful features were baseline vitamin D serum levels and cortical porosity of the radius with scores of 5 and 2, respectively. Overall, the model relies heavily on baseline failure load of the tibia and minimally on dosage groups.

The SHAP analysis results for the SVR model trained to predict bone failure load of the tibia are shown in figure 2. It highlights baseline failure load of the tibia as the most impactful feature with a score of 1492. Female sex and C-terminal telopeptide levels were the next most impactful features for this model, with a score of 179 and 133 respectively. The least impactful features were identified as cortical thickness of the radius with a score of 10, cortical porosity of the radius with a score of 11, and baseline vitamin D serum levels with a score of 12. The vitamin D dosage groups had a score of 46, 59, and 40 for 400 IU, 4000 IU, and 10 000 IU respectively.

The random forest model trained to predict bone failure load of the tibia showed baseline failure load of the tibia was the only important feature with a score of 1675, as shown in Figure 3. All other features had a score of ≤18, indicating little impact on the model.

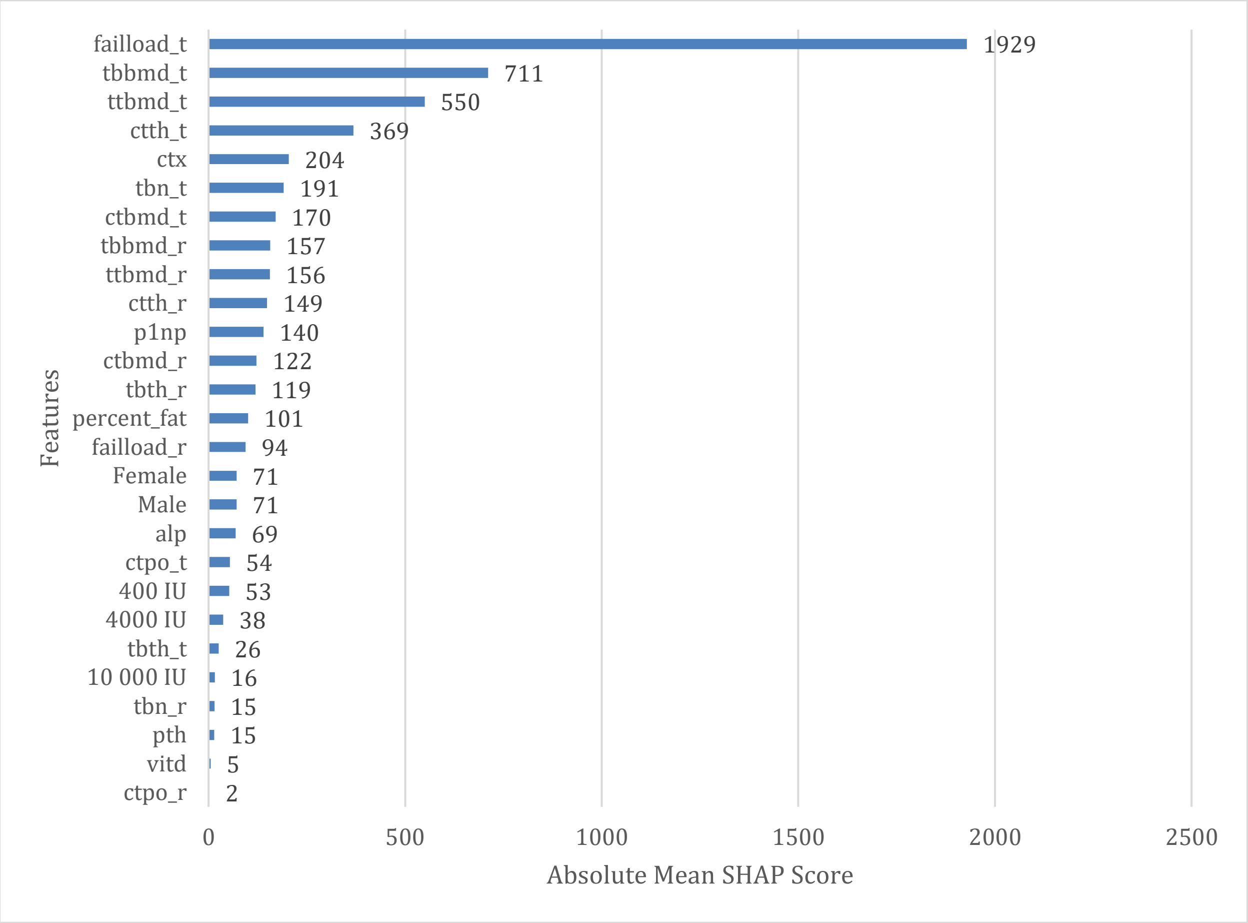

Figure 1: SHAP values obtained from the linear regression predicting failure load of tibia at 36 months. Feature abbreviations are as follows: **Vitd** : Vitamin D serum level. **pth**: Parathyroid. **ctx**: C-terminal telopeptide. **alp**: Alkaline phosphatase. **p1np**: Procollagen Type 1 N-Terminal Propeptide**. percent_fat**: percentage of body fat. **ttbmd_r**: Total body bone mineral density of radius. **ctbmd_r**: Cortical bone mineral density of radius. **tbbmd_r**: Trabecular bone mineral density of radius. **tbth_r**: Trabecular bone thickness of radius. **tbn_r**: Trabecular bone of radius. **ctth_r**: Cortical bone thickness of radius. **ctpo_r**: Cortical porosity of radius. **ttbmd_t**: Total body bone mineral density of tibia. **ctbmd_t**: Cortical bone mineral density measurement of tibia. **tbbmd_t**: total trabecular bone mineral density of tibia. **tbth_t**: Trabecular bone thickness of tibia **tbn_t**: Trabecular bone number total of tibia. **ctth_t**: Cortical bone thickness total measurement of tibia. **ctpo_t**: Cortical porosity total measurement of tibia. **failload_r**: failure load of radius. **failload_t**: failure load of tibia.

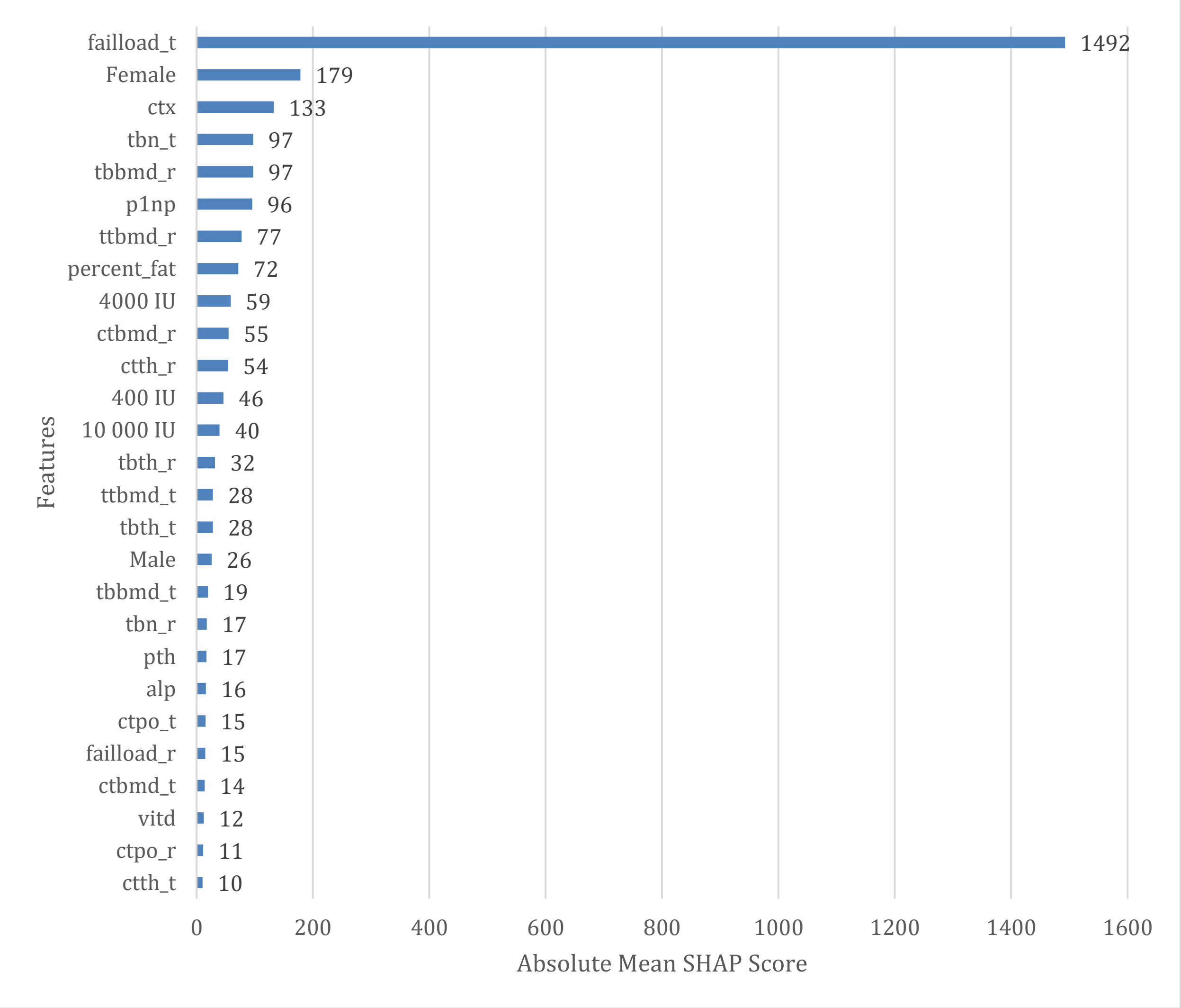

Figure 2: SHAP values obtained from the support vector regression predicting failure load of tibia at 36 months. Feature abbreviations are as follows: **Vitd** : Vitamin D serum level. **pth**: Parathyroid. **ctx**: C-terminal telopeptide. **alp**: Alkaline phosphatase. **p1np**: Procollagen Type 1 N-Terminal Propeptide**. percent_fat**: percentage of body fat. **ttbmd_r**: Total body bone mineral density of radius. **ctbmd_r**: Cortical bone mineral density of radius. **tbbmd_r**: Trabecular bone mineral density of radius. **tbth_r**: Trabecular bone thickness of radius. **tbn_r**: Trabecular bone of radius. **ctth_r**: Cortical bone thickness of radius. **ctpo_r**: Cortical porosity of radius. **ttbmd_t**: Total body bone mineral density of tibia. **ctbmd_t**: Cortical bone mineral density measurement of tibia. **tbbmd_t**: total trabecular bone mineral density of tibia. **tbth_t**: Trabecular bone thickness of tibia **tbn_t**: Trabecular bone number total of tibia. **ctth_t**: Cortical bone thickness total measurement of tibia. **ctpo_t**: Cortical porosity total measurement of tibia. **failload_r**: failure load of radius. **failload_t**: failure load of tibia.

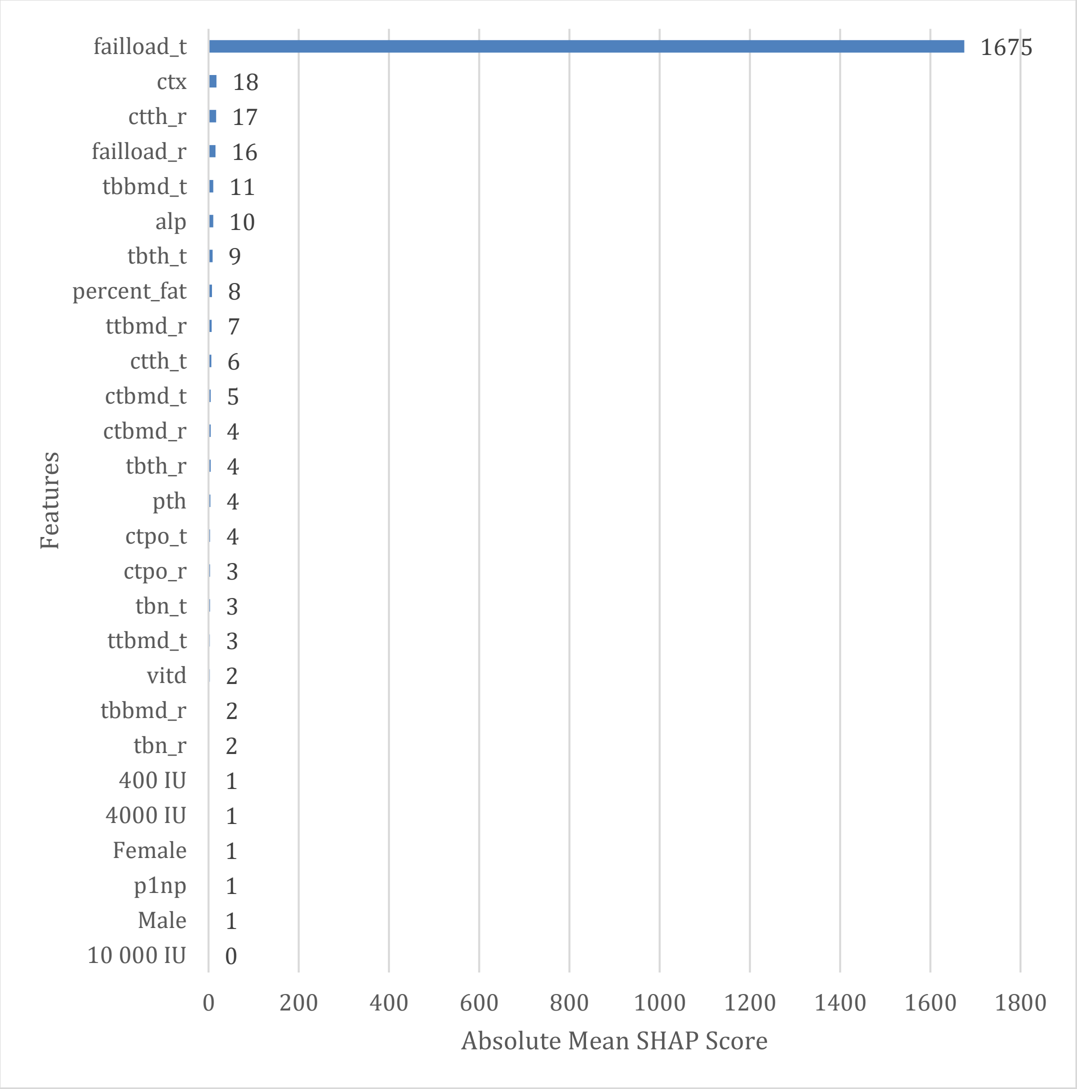

Figure 3: SHAP values obtained from the random forest predicting failure load of tibia at 36 months. Feature abbreviations are as follows: **Vitd** : Vitamin D serum level. **pth**: Parathyroid. **ctx**: C-terminal telopeptide. **alp**: Alkaline phosphatase. **p1np**: Procollagen Type 1 N-Terminal Propeptide**. percent_fat**: percentage of body fat. **ttbmd_r**: Total body bone mineral density of radius. **ctbmd_r**: Cortical bone mineral density of radius. **tbbmd_r**: Trabecular bone mineral density of radius. **tbth_r**: Trabecular bone thickness of radius. **tbn_r**: Trabecular bone of radius. **ctth_r**: Cortical bone thickness of radius. **ctpo_r**: Cortical porosity of radius. **ttbmd_t**: Total body bone mineral density of tibia. **ctbmd_t**: Cortical bone mineral density measurement of tibia. **tbbmd_t**: total trabecular bone mineral density of tibia. **tbth_t**: Trabecular bone thickness of tibia **tbn_t**: Trabecular bone number total of tibia. **ctth_t**: Cortical bone thickness total measurement of tibia. **ctpo_t**: Cortical porosity total measurement of tibia. **failload_r**: failure load of radius. **failload_t**: failure load of tibia.

Baseline failure load of the radius was seen as the most important feature for the linear regression trained to predict failure load of the radius with a score of 658 according to figure 4. Features that will minimally impact the model were identified as total BMD of the radius, trabecular number of the radius, and baseline failure load of the tibia with scores of 116, 88, and 81 respectively. Sex showed an equal impact on the model, both male and female having a score of 2, while both being the lowest impactful features along with cortical BMD of the radius, total BMD of the radius, parathyroid hormone levels (PTH), and trabecular number of the radius, with scores of 1, 1, 2, and 2 respectively. The dosage groups had minimal impact on the model, with a score of 8, 11, 3 for 400 IU, 4000 IU, and 10 000 IU respectively.

Figure 5 highlights the SHAP analysis for the SVR trained to predict failure load of the radius at 36 months. Baseline failure load of the radius was identified as the most impactful feature with a score of 494. Baseline failure load of the tibia, cortical thickness of the radius, and total BMD of the radius were the next most impactful features with scores of 112, 74, and 71 respectively. Cortical porosity of the tibia, PTH, and cortical porosity of the radius were the least impactful features with scores of 1, 3, and 3 respectively. Female had a SHAP score of 28, while male had a SHAP score of 14, half of the impact than female. 400 IU treatment group had a score of 5, while 4000 IU had 5, and 10 000 IU had a score of 17.

The random forest model trained to predict failure load of the radius had baseline failure load of the radius as the most impactful feature to the model with a score of 610 according to figure 6. Additionally, Baseline failure load of the tibia, C-terminal telopeptide, and trabecular number of the radius were minimally impactful to the model with a score of 36, 27, and 20 respectively. The dosage groups of 400 IU, 4000 IU, and 10 000 IU had a score of 0, indicating no impact on the model. Sex had small impact on the model, with male having a score of 2, and female having a score of 3.

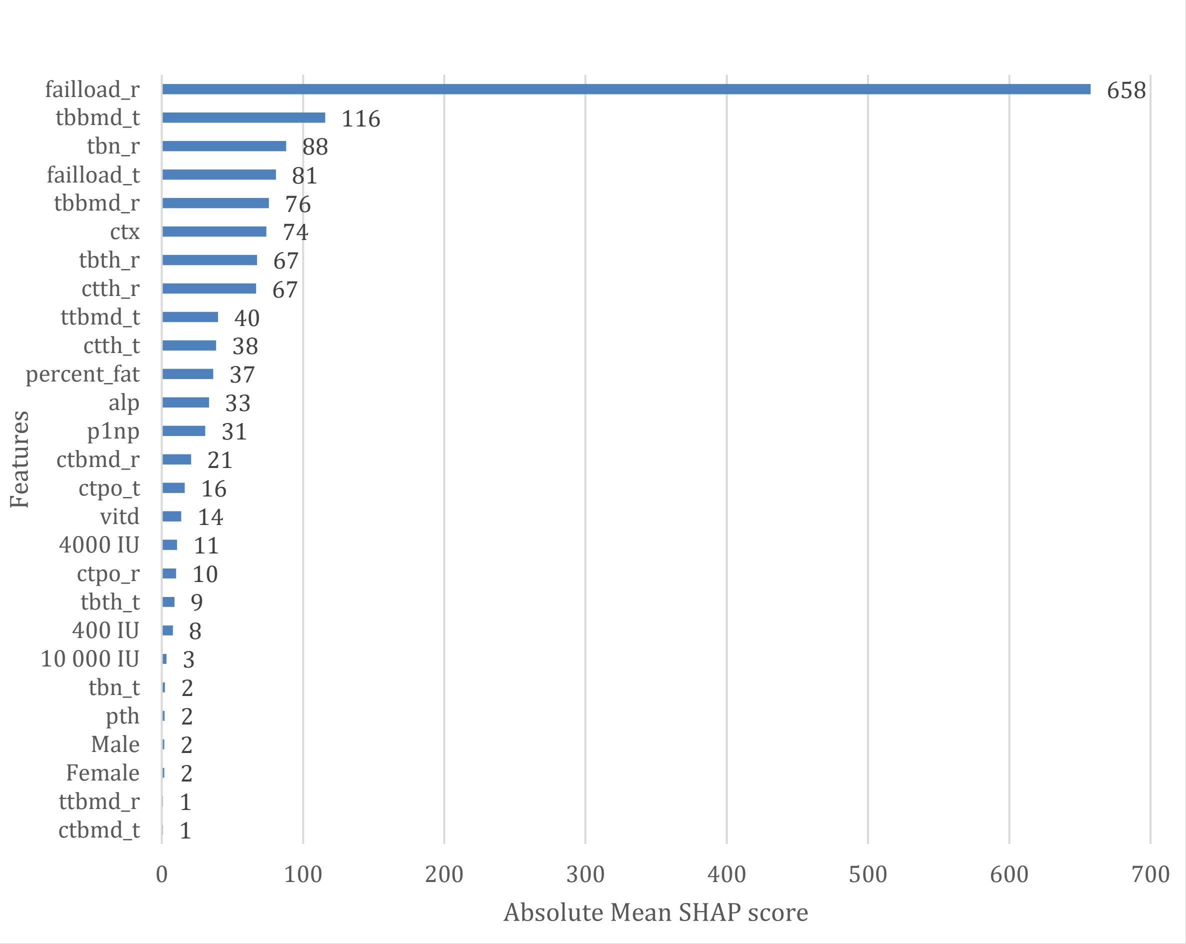

Figure 4: SHAP values obtained from the linear regression predicting failure load of radius at 36 months. Feature abbreviations are as follows: **Vitd** : Vitamin D serum level. **pth**: Parathyroid. **ctx**: C-terminal telopeptide. **alp**: Alkaline phosphatase. **p1np**: Procollagen Type 1 N-Terminal Propeptide**. percent_fat**: percentage of body fat. **ttbmd_r**: Total body bone mineral density of radius. **ctbmd_r**: Cortical bone mineral density of radius. **tbbmd_r**: Trabecular bone mineral density of radius. **tbth_r**: Trabecular bone thickness of radius. **tbn_r**: Trabecular bone of radius. **ctth_r**: Cortical bone thickness of radius. **ctpo_r**: Cortical porosity of radius. **ttbmd_t**: Total body bone mineral density of tibia. **ctbmd_t**: Cortical bone mineral density measurement of tibia. **tbbmd_t**: total trabecular bone mineral density of tibia. **tbth_t**: Trabecular bone thickness of tibia **tbn_t**: Trabecular bone number total of tibia. **ctth_t**: Cortical bone thickness total measurement of tibia. **ctpo_t**: Cortical porosity total measurement of tibia. **failload_r**: failure load of radius. **failload_t**: failure load of tibia.

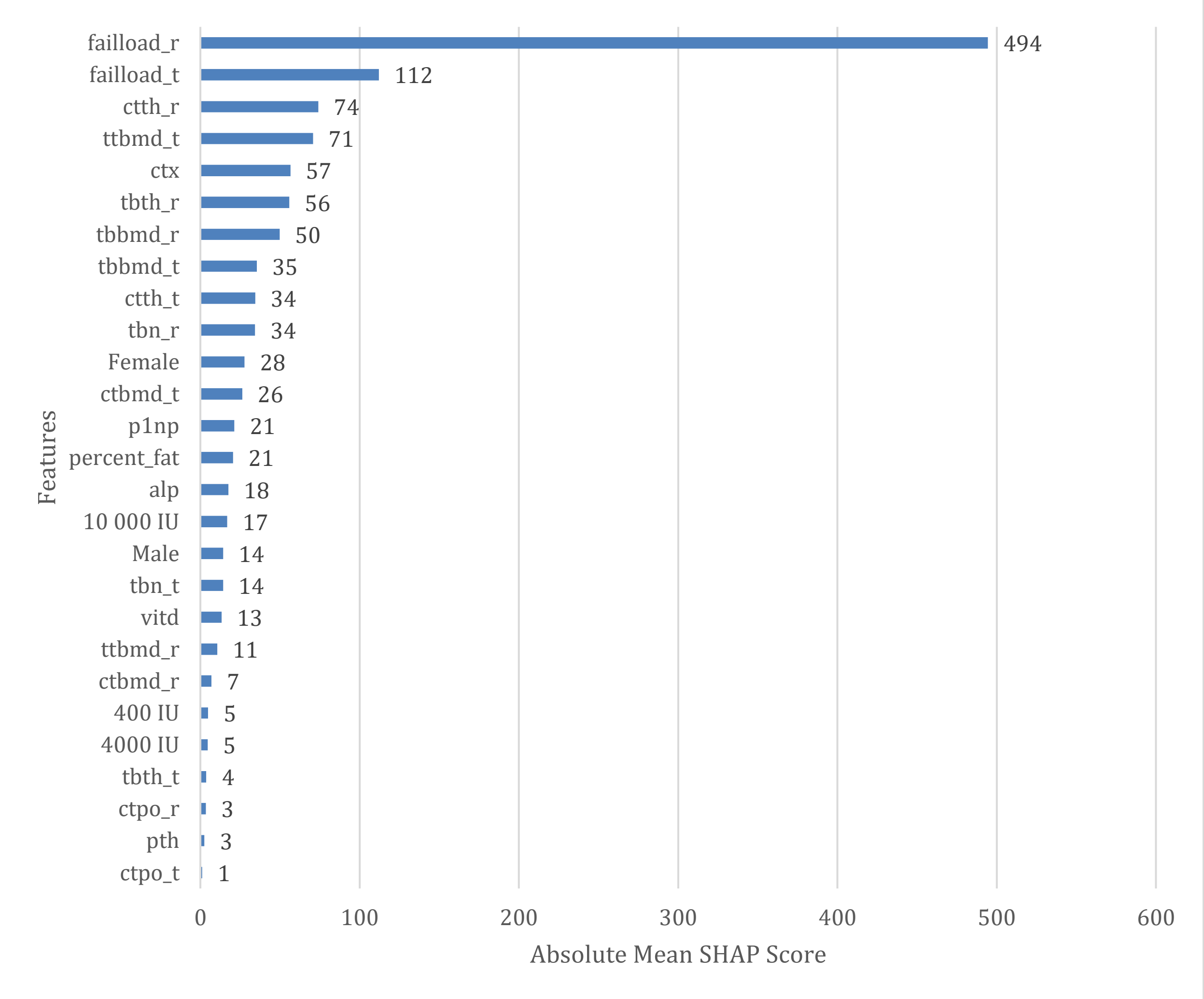

Figure 5: SHAP values obtained from the support vector regression predicting failure load of radius at 36 months. Feature abbreviations are as follows: **Vitd** : Vitamin D serum level. **pth**: Parathyroid. **ctx**: C-terminal telopeptide. **alp**: Alkaline phosphatase. **p1np**: Procollagen Type 1 N-Terminal Propeptide**. percent_fat**: percentage of body fat. **ttbmd_r**: Total body bone mineral density of radius. **ctbmd_r**: Cortical bone mineral density of radius. **tbbmd_r**: Trabecular bone mineral density of radius. **tbth_r**: Trabecular bone thickness of radius. **tbn_r**: Trabecular bone of radius. **ctth_r**: Cortical bone thickness of radius. **ctpo_r**: Cortical porosity of radius. **ttbmd_t**: Total body bone mineral density of tibia. **ctbmd_t**: Cortical bone mineral density measurement of tibia. **tbbmd_t**: total trabecular bone mineral density of tibia. **tbth_t**: Trabecular bone thickness of tibia **tbn_t**: Trabecular bone number total of tibia. **ctth_t**: Cortical bone thickness total measurement of tibia. **ctpo_t**: Cortical porosity total measurement of tibia. **failload_r**: failure load of radius. **failload_t**: failure load of tibia.

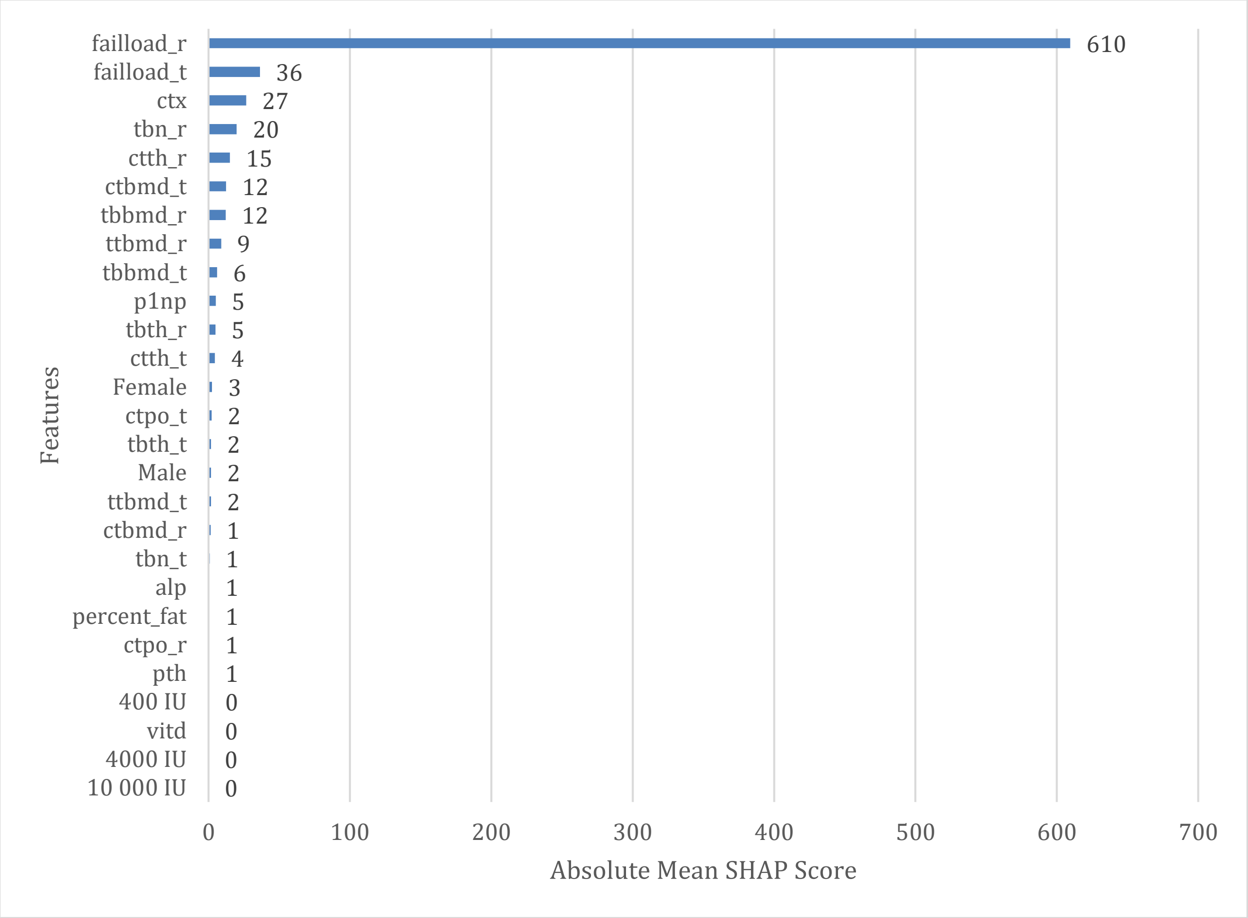

Figure 6: SHAP values obtained from the random forest predicting failure load of radius at 36 months. Feature abbreviations are as follows: **Vitd** : Vitamin D serum level. **pth**: Parathyroid. **ctx**: C-terminal telopeptide. **alp**: Alkaline phosphatase. **p1np**: Procollagen Type 1 N-Terminal Propeptide**. percent_fat**: percentage of body fat. **ttbmd_r**: Total body bone mineral density of radius. **ctbmd_r**: Cortical bone mineral density of radius. **tbbmd_r**: Trabecular bone mineral density of radius. **tbth_r**: Trabecular bone thickness of radius. **tbn_r**: Trabecular bone of radius. **ctth_r**: Cortical bone thickness of radius. **ctpo_r**: Cortical porosity of radius. **ttbmd_t**: Total body bone mineral density of tibia. **ctbmd_t**: Cortical bone mineral density measurement of tibia. **tbbmd_t**: total trabecular bone mineral density of tibia. **tbth_t**: Trabecular bone thickness of tibia **tbn_t**: Trabecular bone number total of tibia. **ctth_t**: Cortical bone thickness total measurement of tibia. **ctpo_t**: Cortical porosity total measurement of tibia. **failload_r**: failure load of radius. **failload_t**: failure load of tibia.

Med Cam Heatmap Results:

In comparison with Figure 8 and 9, Figure 7 indicates that the ResNet model for tibia failure load is putting more weight for slices found in the middle of the HR-pQCT images. The surface slices and bottom slices are not weighted heavily, indicated majority of the image being labelled blue or black. Furthermore, Figure 5 shows that the ResNet model is focusing on parts of the cortical bone on the right side of the tibia, indicated by the red pixels. Figure 5 also shows a space in the middle of the of the tibia, the trabecular bone, that is not weighted heavily, indicated by the black and blue labeling. Additionally, there are many weighted pixels on the boundary between the padding (appears grey) and the actual tibia, as demonstrated by figures 7, 8, and 9.

Unfortunately, generating med cam results from the attention maps were unsuccessful as when the image was upscaled, the attention map was too far off the radius, preventing accurate and reliable interpretation.

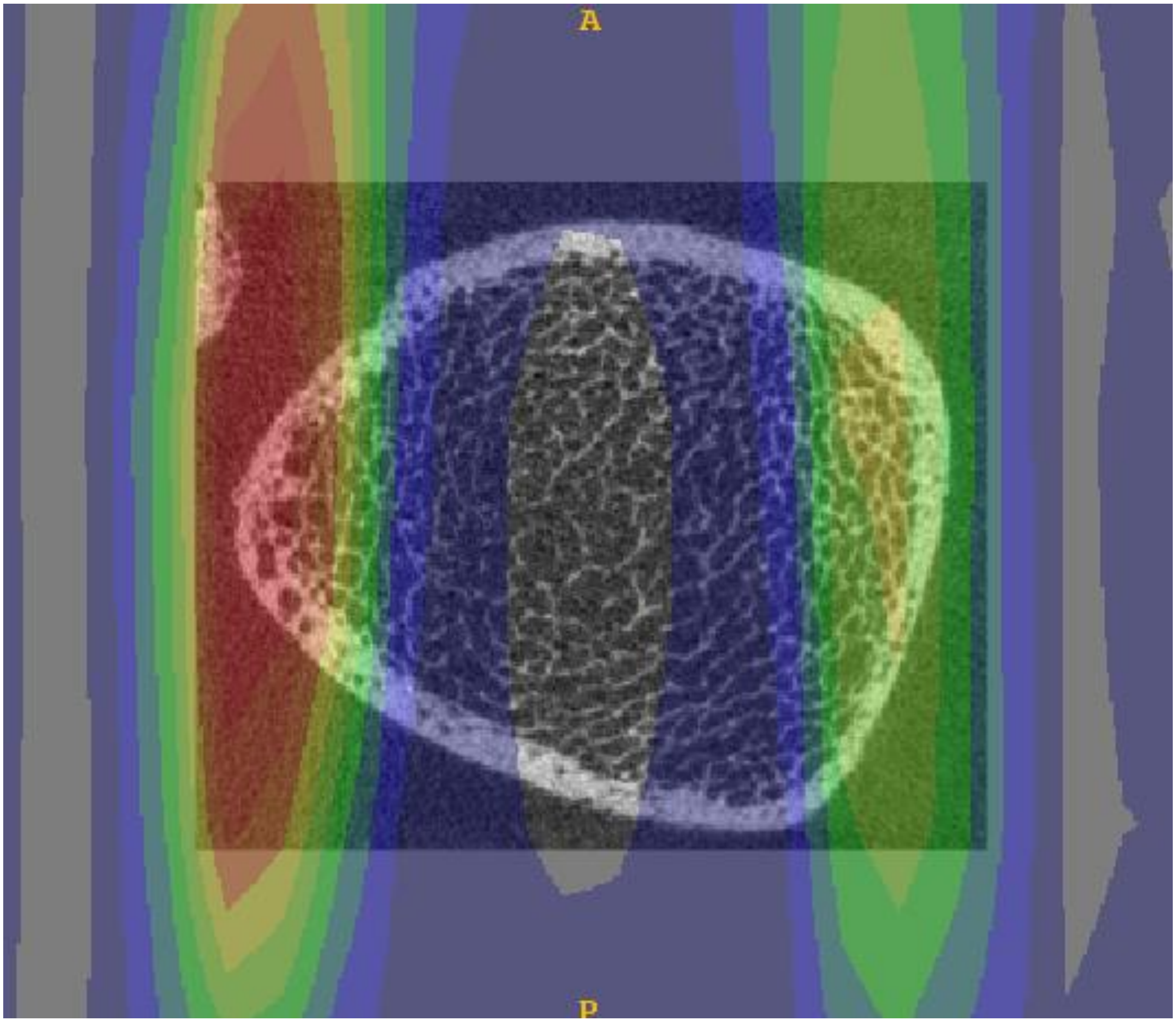

Figure 7. Image corresponds to slice 45 of a HR-pQCTimage of the tibia. Heatmap indicates pixels weighted the highest in the ResNet model. Pixels weighted highest to lowest appear as red, orange, yellow, green, blue, and black.

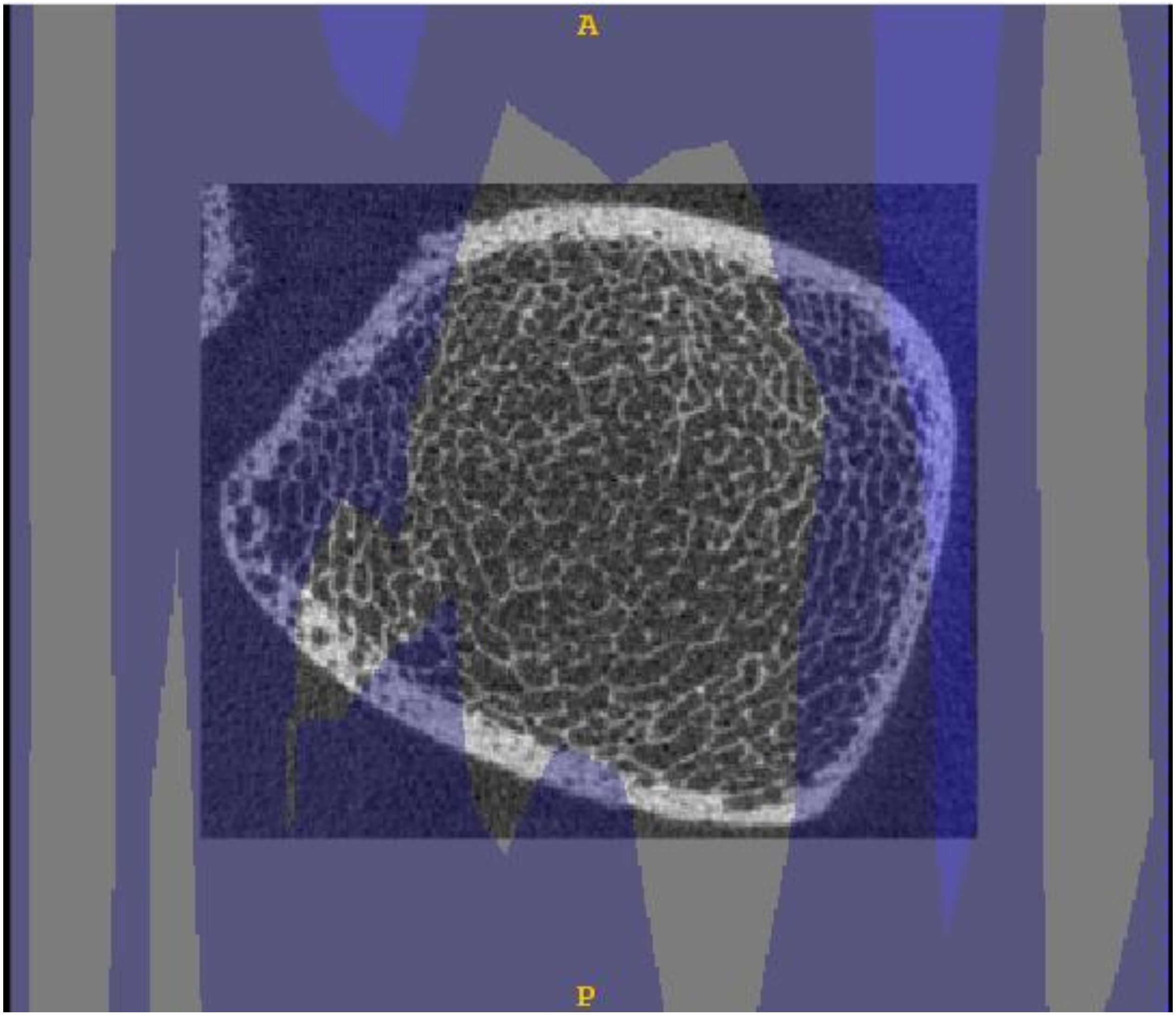

Figure 8. Image corresponds to slice 18 of a HR-pQCR image of the tibia. Heatmap indicates pixels weighted the highest in the ResNet model. Pixels weighted highest to lowest appear as red, orange, yellow, green, blue, and black.

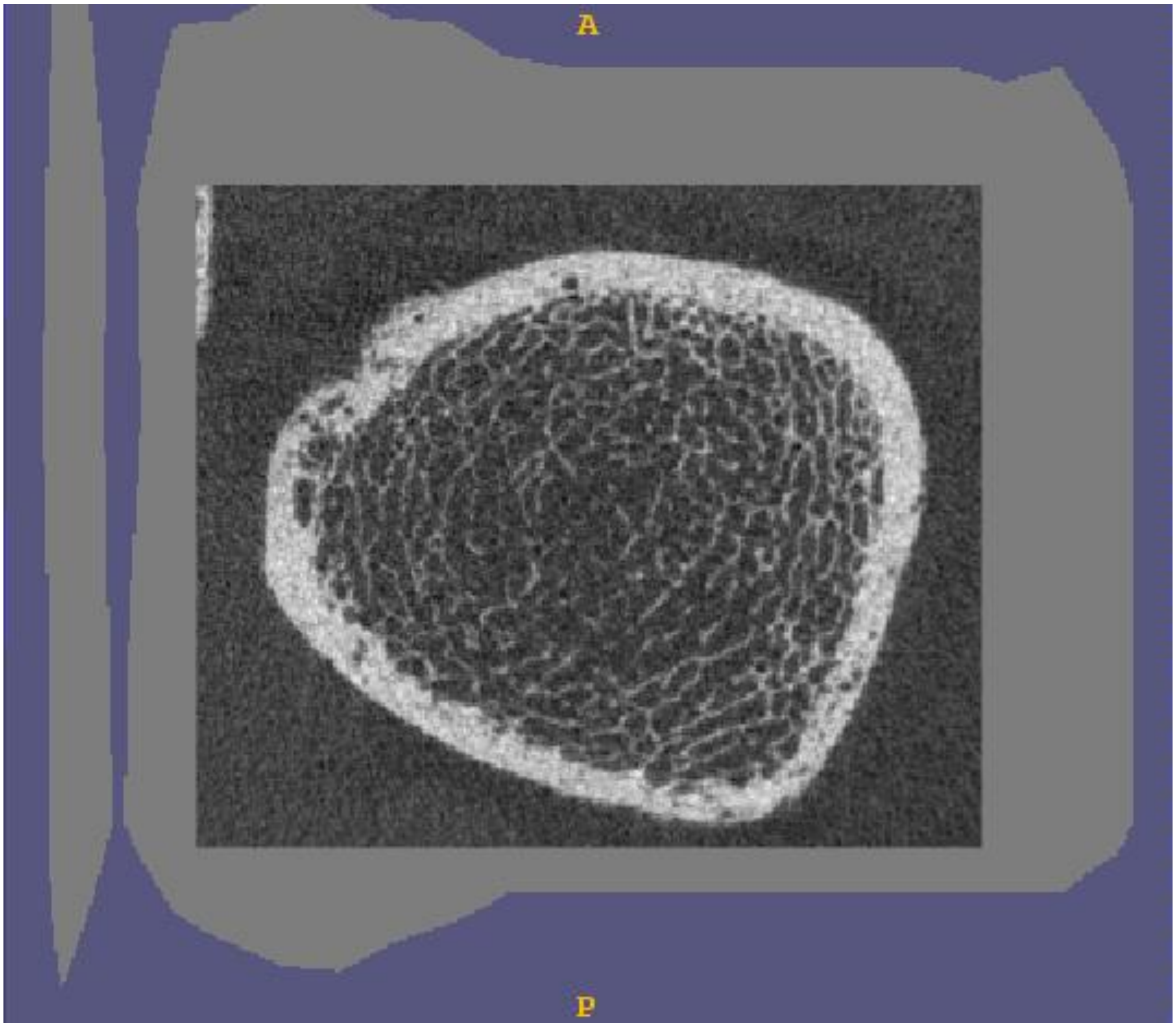

Figure 9. Image corresponds to slice 84 of a HR-pQCR image of the tibia. Heatmap indicates pixels weighted the highest in the ResNet model. Pixels weighted highest to lowest appear as red, orange, yellow, green, blue, and black.

Discussion:

The Random Forest model outperformed all other models in predicting the tibia's failure load at 36 months, whereas the ResNet model performed the worst. Additionally, for predicting radius failure load at 36 months, the SVR model outperformed all models, with ResNet again showing the poorest performance. The most significant feature for predicting tibia failure load across the linear regression, SVR, and Random Forest models was the baseline failure load of the tibia. Similarly, the baseline failure load of the radius was key for predicting the radius failure load at 36 months. Additionally, ResNet demonstrated the greatest degree overfitting to the training dataset in both the radius and tibia prediction models, in contrast to linear regression, which showed minimal overfitting. Med cam images indicate that ResNet focuses on cortical bone in the inner slices of HR-pQCT images.

The random forest model outperformed all other models across all performance metrics for predicting failure load of the tibia at 36 months. It is possible that linear regression and support vector regression fail to capture as complex relationships as the random forest model, leading to its superior performance. Literature demonstrates that multimodal models like DAFT can outperform other neural networks, gradient boosters, linear regression models, and random forest models[17,21,22]. The results of the DAFT model in both failure load of radius and tibia could be contributed to the limited dataset used to train the model. 311 samples may not be enough for the DAFT model to fully learn the complex relationship between the tabular data, HR-pQCT images and failure load of the tibia and radius. Increasing the dataset would most likely improve the DAFT model, as literature that utilized the DAFT model had a dataset size of approximately 600 – 1300 samples[17]. Furthermore, these results highlight that the more complex the model, the more predictive power the model has, but also the need for more samples increases as well, as the neural network-based models were outperformed by the tabular data-based models due to lack of data[15].

In contrast, the SVR model outperformed all other models across all performance metrics for predicting failure load of the radius. This is contrary to the previous conclusion where the SVR was not performing as well as the random forest because of not being a complex enough model to capture the relationships between the features and failure load. The results comparing test and train R^2 scores showed that the random forest had little overfitting occurring, indicating that the random forest was most likely not outperformed by the SVR due to dataset size. The SVR outperformed the random forest by only a small margin, so subtle differences in the variability of the data between the baseline tibia failure loads and baseline radius failure loads could contribute to the reduced performance of the random forest.

The most impactful feature according to the SHAP scores for each model in both tibia and radius models was their corresponding baseline failure load. This is expected as the target variable is to predict the 36-month failure load of tibia or radius. Additionally, in both the radius and tibia models, only the baseline failure load of the tibia and radius was important to the random forest models. The random forest models outperformed most of the models in both bone models, and this implies that many of the features provided to the model are not needed to make accurate predictions of failure load. This is contradictory to both the models, and existing literature. Firstly, Female, C-terminal telopeptide, total BMD of tibia, trabecular BMD of tibia, and cortical thickness of the radius have been shown to have a degree of impact in the SVR and linear regression models. It is well known that there are differences between men and women's tibia and radius, with women having less dense bones [29]. Additionally, BMD is known to be correlated with failure load, making the SHAP results found in the random forest contradictory to existing literature and thus the results seen in the linear regression and SVR is supported by the existing literature [29]. Furthermore, C-terminal telopeptide (CTX), a marker of bone turnover, impacts bone failure. This is supported in literature, as an increased rate of bone turnover in comparison to bone formation can either result in bone loss or gain and thus impacting failure load, depending which process is occurring at a higher rate[1,26].

Regarding the impact that the different vitamin D dosage groups had on the models, the results show that they were not impactful to the decision making of the models. This is shown in Figures 1-6, where each treatment group had very low SHAP scores, ranging from 0-56 across all the models. This is supported by the existing literature that found no significant impact on bone given various doses of vitamin D[12,13,14].

The results shown in Figures 7, 8, and 9 highlight why the ResNet model performed poorly, as many of the red labelled pixels are found in the padding of the image. This indicates that the model is learning the boundary between the tibia and padding of the image, measuring the size of the tibia, which is an indication of the size which is correlated to failure load. While this is not wrong, failure load is not only correlated to bone size, but also bone microarchitecture like BMD[1] and should be focused on both bone size and microarchitecture.

Table 7 also explains the poor performance of the ResNet, demonstrating that it is overfitting the most, being the most successful in the training dataset, while performing the worst in the test dataset. The same limitation that the DAFT model faces applies to the ResNet model, as a larger sample size is required for the model to begin outperforming the other models. Additionally, models like ResNet are complex and prone to overfitting, which is supported by Figure 4 and existing literature[15].

Looking at Figure 7, the ResNet model focuses on the cortical bone found in the middle slice the most, as demonstrated by the highest weighted pixels only showing up in the center slices of the image. These results support existing literature, as cortical bone contributes more significantly to bone strength and its failure than trabecular bone due to its load-bearing capacity and structural properties, such as cortical thickness, cross-sectional area, and area moment of inertia[25].

Limitations:

Many bone measurements potentially provide redundant information, negatively impacting the performance of the models[27]. This is seen by Figures 1-6, with many of the features having low SHAP scores. A potential solution is principal component analysis (PCA) to reduce feature dimensionality, which was not pursued to maintain model interpretability through explainable AI techniques. Studies have shown PCA to greatly improve the performance of machine learning models and provides a future direction for improving model performance[28].

Furthermore, the study was limited to healthy participants aged 55-77 years, which constrains the model's relevance to different age groups or individuals with health issues. In the future, a wider age range should be considered to increase the scope of what the models can predict.

Additionally, the research utilized a singular architecture for each model. Exploring a broader range of architectures, such as Dense Net or Vision Transformers for neural networks, or investigating various libraries offering random forest, linear regression, and support vector regression models, could be valuable future directions to pursue to improve our model's performance in predicting bone failure.

Med Cam results have limitations, as output of the attention maps are much lower resolution in comparison with the original input image. This forces the attention map to be upscaled to the original image. This results in the loss of information, resulting in the interpretation of the weights to have a degree of error. Studies that have utilized Med Cam do not describe this limitation, possibly due to the design of Med Cam meant to work with 3D segmentation and classification problems, not regression tasks[23,24]. Potential adjustments could be done on med am to optimize and adapt the software to be more compatible with regression tasks.

Future Directions:

In the future, increasing the size of the dataset used to train the models will help resolve the overfitting issue found in our more complex models. Additionally, a more diverse age group should be included to increase the scope of data the model can predict on.

Additionally, PCA will be explored to reduce the number of redundant features in the tabular dataset, which will improve performance. Furthermore, methods to extract which features belong in the PCA's chosen as inputs to the model will also be investigated, to combine both performance improvements and retain interpretability of the model.

Several improvements to resolving the issues with upscaling low resolution attention maps will be explored. For example, adjusting in which layers attention maps are produced, to obtain them in earlier layers that have not been down sampled and adjusting existing architectures to minimize down sampling of lower layers. Additionally, more sophisticated upscaling methods could be explored to upscale attention maps more effectively.

Broadening the scope of architectures could also be explored. For example, there are many different libraries in different coding languages that could provide better performance, such as TensorFlow and Keras.

Conclusion:

This project combined machine learning and HR-pQCT imaging to predict optimal vitamin D dosages by assessing the failure load of the radius and tibia at 36 months in a clinical trial with 311 participants. We trained models including linear regression, support vector regression, random forest, neural networks, and the DAFT neural network. Findings showed the random forest model was most effective for the tibia, and support vector regression was best for the radius. The DAFT model showed potential for integrating multimodal data but needs larger datasets for full effectiveness. Using SHAP values and Med-CAM, we gained insights into model decisions and confirmed the minimal impact of vitamin D dosage on predictions. This research lays groundwork for future studies to enhance model complexity and incorporate more features, advancing machine learning in personalized medicine for optimizing osteoporosis treatment strategies.

References:

1. Poduval, Murali, and Karthik Vishwanathan. 2023. "Definition and Evolution of the Term Osteoporosis." *Indian Journal of Orthopaedics*, October. https://doi.org/10.1007/s43465-023-01013-2.
2. Odén, Anders, Eugene V. McCloskey, Helena Johansson, and John A. Kanis. "Assessing the Impact of Osteoporosis on the Burden of Hip Fractures." Calcified tissue international 92, no. 1 (2013): 42–49.
3. Wade, S. W., C. Strader, L. A. Fitzpatrick, M. S. Anthony, and C. D. O'Malley. 2014. "Estimating Prevalence of Osteoporosis: Examples from Industrialized Countries." *Archives of Osteoporosis* 9 (1): 182. https://doi.org/10.1007/s11657-014-0182-3.
4. Xiao, P.-L., A.-Y. Cui, C.-J. Hsu, R. Peng, N. Jiang, X.-H. Xu, Y.-G. Ma, D. Liu, and H.-D. Lu. 2022. "Global, Regional Prevalence, and Risk Factors of Osteoporosis According to the World Health Organization Diagnostic Criteria: A Systematic Review and Meta-Analysis." *Osteoporosis International* 33 (10): 2137–53. https://doi.org/10.1007/s00198-022-06454-3.
5. Johnston, Catherine Bree, and Meenakshi Dagar. "Osteoporosis in Older Adults." The Medical clinics of North America 104, no. 5 (2020): 873–884.
6. Zhou, Bin, Ji Wang, Y. Eric Yu, Zhendong Zhang, Shashank Nawathe, Kyle K. Nishiyama, Fernando Rey Rosete, Tony M. Keaveny, Elizabeth Shane, and X. Edward Guo. "High-Resolution Peripheral Quantitative Computed Tomography (HR-pQCT) Can Assess Microstructural and Biomechanical Properties of Both Human Distal Radius and Tibia: Ex Vivo Computational and Experimental Validations." Bone (New York, N.Y.) 86 (2016): 58–67.
7. Boutroy, Stephanie, Bert Van Rietbergen, Elisabeth Sornay-Rendu, Francoise Munoz, Mary L Bouxsein, and Pierre D Delmas. 2008. "Finite Element Analysis Based on In Vivo HR-pQCT Images of the Distal Radius Is Associated With Wrist Fracture in Postmenopausal Women." *Journal of Bone and Mineral Research* 23 (3): 392–99. https://doi.org/10.1359/jbmr.071108.
8. MacNeil, Joshua A., and Steven K. Boyd. 2007. "Accuracy of High-Resolution Peripheral Quantitative Computed Tomography for Measurement of Bone Quality." *Medical Engineering & Physics* 29 (10): 1096–1105. https://doi.org/10.1016/j.medengphy.2006.11.002.
9. Bæksgaard, L., K. P. Andersen, and L. Hyldstrup. 1998. "Calcium and Vitamin D Supplementation Increases Spinal BMD in Healthy, Postmenopausal Women." Osteoporosis International 8 (3): 255–60. https://doi.org/10.1007/s001980050062.
10. Buckley, Lenore M. 1996. "Calcium and Vitamin D 3 Supplementation Prevents Bone Loss in the Spine Secondary to Low-Dose Corticosteroids in Patients with Rheumatoid Arthritis: A Randomized, Double-Blind, Placebo-Controlled Trial." Annals of Internal Medicine 125 (12): 961. https://doi.org/10.7326/0003-4819-125-12-199612150-00004.
11. Laird, Eamon, Mary Ward, Emeir McSorley, J. J. Strain, and Julie Wallace. 2010. "Vitamin D and Bone Health: Potential Mechanisms." *Nutrients* 2 (7): 693–724. https://doi.org/10.3390/nu2070693.
12. Grimnes, G., R. Joakimsen, Y. Figenschau, P. A. Torjesen, B. Almås, and R. Jorde. 2012. "The Effect of High-Dose Vitamin D on Bone Mineral Density and Bone Turnover Markers in Postmenopausal Women with Low Bone Mass—a Randomized Controlled 1-Year Trial." *Osteoporosis International* 23 (1): 201–11. https://doi.org/10.1007/s00198-011-1752-5.
13. Burt, Lauren A., Emma O. Billington, Marianne S. Rose, Duncan A. Raymond, David A. Hanley, and Steven K. Boyd. 2019. "Effect of High-Dose Vitamin D Supplementation on Volumetric

Bone Density and Bone Strength: A Randomized Clinical Trial.” *JAMA* 322 (8): 736–45. https://doi.org/10.1001/jama.2019.11889.

14. Cooper, Lucy, Phillip B Clifton-Bligh, M Liza Nery, Gemma Figtree, Stephen Twigg, Emily Hibbert, and Bruce G Robinson. 2003. “Vitamin D Supplementation and Bone Mineral Density in Early Postmenopausal Women.” *The American Journal of Clinical Nutrition* 77 (5): 1324–29. https://doi.org/10.1093/ajcn/77.5.1324.
15. Deo, Rahul C. 2015. “Machine Learning in Medicine.” *Circulation* 132 (20): 1920–30. https://doi.org/10.1161/CIRCULATIONAHA.115.001593.
16. Borisov, Vadim, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2024. “Deep Neural Networks and Tabular Data: A Survey.” *IEEE Transactions on Neural Networks and Learning Systems*, 1–21. https://doi.org/10.1109/TNNLS.2022.3229161.
17. Wolf, Tom Nuno, Sebastian Pölsterl, and Christian Wachinger. 2022. “DAFT: A Universal Module to Interweave Tabular Data and 3D Images in CNNs.” *NeuroImage* 260 (October): 119505. https://doi.org/10.1016/j.neuroimage.2022.119505.
18. scikit-optimize developers. "skopt/searchcv.py." GitHub, last modified March 2, 2024. https://github.com/scikit-optimize/scikit-optimize/blob/de32b5f/skopt/searchcv.py#L30.
19. Lundberg, Scott, and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” https://doi.org/10.48550/ARXIV.1705.07874.
20. Otkowski, Karol, Camila Gonzalez, Andreas Bucher, and Anirban Mukhopadhyay. 2020. “M3d-CAM: A PyTorch Library to Generate 3D Data Attention Maps for Medical Deep Learning.” https://doi.org/10.48550/ARXIV.2007.00453.
21. Płotka, Szymon, Michal K. Grzeszczyk, Robert Brawura-Biskupski-Samaha, Paweł Gutaj, Michał Lipa, Tomasz Trzciński, Ivana Išgum, Clara I. Sánchez, and Arkadiusz Sitek. 2023. “BabyNet＋＋: Fetal Birth Weight Prediction Using Biometry Multimodal Data Acquired Less than 24 Hours before Delivery.” Computers in Biology and Medicine 167 (December): 107602. https://doi.org/10.1016/j.compbiomed.2023.107602.
22. Borsos, Balázs, Corinne G. Allaart, and Aart Van Halteren. 2024. “Predicting Stroke Outcome: A Case for Multimodal Deep Learning Methods with Tabular and CT Perfusion Data.” Artificial Intelligence in Medicine 147 (January): 102719. https://doi.org/10.1016/j.artmed.2023.102719.
23. Gotkowski, Karol, Camila Gonzalez, Andreas Bucher, and Anirban Mukhopadhyay. 2020. “M3d-CAM: A PyTorch Library to Generate 3D Data Attention Maps for Medical Deep Learning.” https://doi.org/10.48550/ARXIV.2007.00453.
24. Solano-Rojas, Braulio, and Ricardo Villalón-Fonseca. 2021. “A Low-Cost Three-Dimensional DenseNet Neural Network for Alzheimer’s Disease Early Discovery.” Sensors 21 (4): 1302. https://doi.org/10.3390/s21041302.
25. Osterhoff, Georg, Elise F. Morgan, Sandra J. Shefelbine, Lamya Karim, Laoise M. McNamara, and Peter Augat. 2016. “Bone Mechanical Properties and Changes with Osteoporosis.” Injury 47 (June): S11–20. https://doi.org/10.1016/S0020-1383(16)47003-8.
26. Greenblatt, Matthew B., Joy N. Tsai, and Marc N. Wein. 2017. “Bone Turnover Markers in the Diagnosis and Monitoring of Metabolic Bone Disease.” Clinical Chemistry 63 (2): 464–74. https://doi.org/10.1373/clinchem.2016.259085.
27. Shimizu, Ayame, and Kei Wakabayashi. 2021. “Examining Effect of Label Redundancy for Machine Learning Using Crowdsourcing.” In The 23rd International Conference on Information Integration and Web Intelligence, 87–94. Linz Austria: ACM. https://doi.org/10.1145/3487664.3487677.

28. Howley, Tom, Michael G. Madden, Marie-Louise O’Connell, and Alan G. Ryder. "The effect of principal component analysis on machine learning accuracy with high dimensional spectral data." In International Conference on Innovative Techniques and Applications of Artificial Intelligence, pp. 209-222. London: Springer London, 2005.
29. Burghardt, Andrew J, Galateia J Kazakia, Sweta Ramachandran, Thomas M Link, and Sharmila Majumdar. 2010. “Age- and Gender-Related Differences in the Geometric Properties and Biomechanical Significance of Intracortical Porosity in the Distal Radius and Tibia.” *Journal of Bone and Mineral Research* 25 (5): 983–93. https://doi.org/10.1359/jbmr.091104.

Appendix

Appendix 1: Linear regression model parameters for failure load of tibia (A) and radius (B)

A			B		
Parameter	Value	Optimized	Parameter	Value	Optimized
fit_intercept	TRUE	No	fit_intercept	TRUE	No
copy_X	TRUE	No	copy_X	TRUE	No
n_jobs	0	No	n_jobs	0	No
positive	FALSE	No	positive	FALSE	No

Appendix 2: Support vector regression model parameters for failure load of tibia (A) and radius (B)

A			B		
Parameter	Value	Optimized	Parameter	Value	Optimized
C	5.45	Yes	C	1.84	Yes
coef0	6.88	Yes	coef0	7.78	Yes
degree	5.00	Yes	degree	5.00	Yes
epsilon	0.37	Yes	epsilon	0.45	Yes
gamma	0.03	Yes	gamma	0.03	Yes
kernel	poly	Yes	kernel	poly	Yes
max_iter	2616.00	Yes	max_iter	9483.00	Yes
shrinking	TRUE	Yes	shrinking	TRUE	Yes
tol	0.00	Yes	tol	0.00	Yes

Appendix 3: Random Forest model parameters for failure load of tibia (A) and radius (B)

A			B		
Parameter	Value	Optimized	Parameter	Value	Optimized
bootstrap	TRUE	Yes	bootstrap	TRUE	Yes
ccp_alpha	0.09	Yes	ccp_alpha	0.08	Yes
criterion	friedman_mse	Yes	criterion	poisson	Yes
max_depth	50.00	Yes	max_depth	46.00	Yes
max_features	None	Yes	max_features	None	Yes
max_leaf_nodes	40.00	Yes	max_leaf_nodes	540.00	Yes
max_samples	0.50	Yes	max_samples	1.00	Yes
min_impurity_decrease	0.64	Yes	min_impurity_decrease	0.45	Yes
min_samples_leaf	3.00	Yes	min_samples_leaf	9.00	Yes
min_samples_split	10.00	Yes	min_samples_split	2.00	Yes
min_weight_fraction_leaf	0.00	Yes	min_weight_fraction_leaf	0.00	Yes
n_estimators	526.00	Yes	n_estimators	1999.00	Yes

Appendix 4: ResNet parameters for failure load of tibia (A) and radius (B)

A			B		
Parameter	Value	Optimized	Parameter	Value	Optimized
block	basic	Yes	block	basic	Yes
layers	[2, 2, 2, 2]	Yes	layers	[2, 2, 2, 2]	Yes
block_inplanes	[64, 128, 256, 512]	Yes	block_inplanes	[64, 128, 256, 512]	Yes
spatial_dims	3.00	No	spatial_dims	3.00	No
n_input_channels	1	No	n_input_channels	1	No
conv1_t_size	(7, 7, 7)	Yes	conv1_t_size	(7, 7, 7)	Yes
conv1_t_stride	(2, 2, 2)	Yes	conv1_t_stride	(2, 2, 2)	Yes
no_max_pool	TRUE	Yes	no_max_pool	TRUE	Yes
shortcut_type	B	No	shortcut_type	B	No
widen_factor	1	No	widen_factor	1	No
num_classes	1	No	num_classes	1	No
feed_forward	TRUE	No	feed_forward	TRUE	No
bias_downsample	TRUE	No	bias_downsample	TRUE	No

Appendix 5: DAFT model parameters for failure load of tibia (A) and radius (B)

A			B		
Parameter	Value	Optimized	Parameter	Value	Optimized
in_channels	1	No	in_channels	1	No
n_outputs	1	No	n_outputs	1	No
bn_momentum	0.10	Yes	bn_momentum	0.10	Yes
n_basefilters	4	Yes	n_basefilters	4	Yes