# OXRAY: Database to Diagnose Osteoporosis Condition and Classify using Transformer

**Pooja S Dodamani[1], Dr Ajit Danti[2], Dr. Shivanand Dodamani[3] , Dr. Vivek Patil[4]**

**Abstract**

Advances in the medical Image processing field anticipate the need to research and evaluate various applications in medical field. Biomedical is one such field that is considered most prominent in progress of diseases. Osteoporosis is the disease which inhibit the risk of fractures in human body due to weakening of bone. The bone becomes porous and weak which results in wear and tear of tissues resulting in fracture risk. In the present scenario Bone mass density (BMD) is well-known and acceptable standard by WHO to diagnose osteoporosis disease. In BMD scan the energy emitted by the x-ray beams is being passed through bone region and is absorbed and other part of the bone is not absorbed. So denser the bones mean it has good mineral content which absorbs more energy and less dense which absorbs less energy. The energy absorbed per pixel is measured in g/cm by converting. The density of bone is calculated with each pixel is based on the number of pixels in particular area. But still, it is questionable by many researchers. So motivated by this we have come up with custom x-ray images database for researchers to carry rigorous analysis in medical domain and forecast the findings by examining the accuracy of medical x-ray images results. our custom database of x-ray images has groups of Spine, Knee, Hand, Femur, Leg, Shoulder bones with details of Indian patients with unique patient ID, age, gender, diagnosis, image type. This Data can be used by young research scholars to work extensively in the osteoporosis bone disease and forecast promising results for betterment of medical society. In this study, we propose a Vision Transformer (ViT) model for classification of osteoporosis using X-ray images. The ViT model utilizes a self-attention mechanism to capture larger dependencies between image patches and has shown good results for various computer vision tasks. We trained and evaluated the model on a dataset of X-ray images, with images for each class normal and osteoporosis. The optimized hyperparameters were the RectifiedAdam optimizer with a learning rate of 0.0001, 500 epochs, and a mini-batch size of 16. The model achieved an overall accuracy of 88%. The proposed ViT classification model shows potential as a reliable and efficient tool for osteoporosis diagnosis, which could aid in early detection and treatment of the disease.

**Keywords:** *X-rays, DEXA, BMD, ViT*

[1] *Department of Computer Science, Christ University, Bangalore, India*

[2] *Department of Computer Science, Faculty of Engineering, Christ University, Bangalore, India*

[3]*Department of Orthopedics, Zydus Hospital, Dahod, India*

[4]*Department of Orthopedics, JNMC Hospital, Belgaum, India*

* Corresponding author's Email: pooja.dodamani@res.christuniversity.in

## 1. Introduction

Osteoporosis is a common skeletal disorder that affects millions of people worldwide, especially the elderly. Early detection and diagnosis of osteoporosis can help prevent complications such as fractures and improve treatment outcomes. X-ray imaging is a widely used diagnostic tool for osteoporosis, as it can reveal changes in bone density and structure. However, accurately diagnosing osteoporosis from X-ray images can be challenging due to the subtle differences in bone density and structure between normal and osteoporotic bones.

So motivated by this we have come up with custom x-ray scan images database for researchers to carry out rigorous analysis in medical domain and forecasts the findings by examining the accuracy of medical x-ray images but still it is questionable by many researchers Based on the number of pixels in the particular area into aerial density. Osteoporosis is a disease which included the advances in medical field its treatment and diagnosis is related to the development of Engineering method in Medical technology to develop various Algorithms, Applications and Devices to solve health care related problem by new Innovation's main aim of Biomedical Engineering is to improve the medical field with technical advancement and Improve Human Health.[26] Biomedical engineering in the past have created many medical devices to save life so, in future also we need to innovate to make our life better and save humans from life-threatening diseases. Hence every technical research or innovation is useful tool for human health and progress. X-ray was considered to be one of the greatest innovations which change the history in medical field .so with advancement in the x-ray field initially single photon absorptiometry SPA found in year 1963 by Cameron and Sorenson was used to measure the bone density in the areas such as femoral neck and lumber spine. But the correlation between the peripheral skeleton lumbar spine bone density is moderate which results in error for clinical evaluation. Also due to its limitations that limit body parts measurement. Dual photon absorptiometry (DPA) uses too distant beams radiation energy to evaluate and measure the bone density with mathematical calculations. In 1980s the invention of single photon absorptiometry (SEXA) and Dual photon absorptiometry (DEXA). SEXA had lower error rate in measuring bone density compared to SPA. DEXA use two energy sources to distinguish between the soft tissue and heart issue that uses higher energy and low energy to measure bone density and currently is well-known gold standard excepted by WHO to measure bone density but DEXA also have certain limitation major in lack of standardization in Bone tissue measurement. The regional assessment is inferior in DEXA, body part considered for measurement are very few it's costly and not affordable by economically weaker society. so, there is a lot of scope to research and come up with a tool to predict the bone density using x-rays due to its affordability and traditionally accepted way to diagnosis osteoporosis condition. It is very much observed in people of age group 40-80years men and women especially post-menopausal woman are at high-risk condition is left untreated it will result in permanent disability or cost one's life. It's important to address this condition and make affordable device so that each and every country can afford it. Osteoporosis is condition which makes bone more fragile it's because our body tends to reabsorb the old one and create new bone in the body, when the absorption is more than the creation there is a loss in the bone tissue which leads to bone mineral loss. This loss results in fracture occurrence in hip, vertebrae, spine etc. And people who have had fracture have the risk of having the second fracture occurrence and various bone break. So, bone mineral density test is the well-known method currently and based on all above literature survey and deficiency in the current state of art there is a need to research in the bio medical field and come up with affordable tools to solve the problems of human kind.[1]

Deep learning techniques have shown promising results in image classification tasks, including medical image analysis.[4] In particular, the Vision Transformer (ViT) architecture has gained attention as a powerful tool for image classification, even surpassing traditional convolutional neural networks (CNNs) in certain cases.[18] ViT replaces the patch-based approach of CNNs with a permutation-based approach, treating each pixel in the image as a separate entity. This results in improved performance on certain image classification tasks, making it a promising approach for osteoporosis diagnosis using X-ray images.[19]

In this paper, we present custom x-ray database can be used to carry out research related to bone as there is no much research done and data available on imaging modalities due to ethical restrictions.[24] most of them have discussed various methods related to imaging modalities but no such data is made available for research Publicly, motivated by this we are offering our custom database and a ViT classification model for osteoporosis diagnosis using X-ray images. We evaluate the model on a dataset of X-ray images from both normal and osteoporotic patients, and compare its performance to traditional CNN-based models. Our results demonstrate the potential of ViT for accurately diagnosing osteoporosis from X-ray images, which could ultimately improve patient outcomes and quality of life.

Our custom database consists of Spine, Knee, Hand, Femur, Leg, Shoulder x-ray scan images with her patient ID, gender, position, age, diagnosis.

Osteoporosis is a common skeletal disorder that affects millions of people worldwide, especially the elderly. Early detection and diagnosis of osteoporosis can help prevent complications such as fractures and improve treatment outcomes. X-ray imaging is a widely used diagnostic tool for osteoporosis, as it can reveal changes in bone density and structure. However, accurately diagnosing osteoporosis from X-ray images can be challenging due to the subtle differences in bone density and structure between normal and osteoporotic bones.

So motivated by this we have come up with custom x-ray scan images database for researchers to carry out rigorous analysis in medical domain and forecasts the findings by examining the accuracy of medical x-ray images but still it is questionable by many researchers Based on the number of pixels in the particular area into aerial density. Osteoporosis is a disease which included the advances in medical field its treatment and diagnosis is related to the development of Engineering method in Medical technology to develop various Algorithms, Applications and Devices to solve health care related problem by new Innovation's main aim of Biomedical Engineering is to improve the medical field with technical advancement and Improve Human Health.[26] Biomedical engineering in the past have created many medical devices to save life so, in future also we need to innovate to make our life better and save humans from life-threatening diseases. Hence every technical research or innovation is useful tool for human health and progress. X-ray was considered to be one of the greatest innovations which change the history in medical field .so with advancement in the x-ray field initially single photon absorptiometry SPA found in year 1963 by Cameron and Sorenson was used to measure the bone density in the areas such as femoral neck and lumber spine. But the correlation between the peripheral skeleton lumbar spine bone density is moderate which results in error for clinical evaluation. Also due to its limitations that limit body parts measurement. Dual photon absorptiometry (DPA) uses too distant beams radiation energy to evaluate and measure the bone density with mathematical calculations. In 1980s the invention of single photon absorptiometry (SEXA) and Dual photon absorptiometry (DEXA). SEXA had lower error rate in measuring bone density compared to SPA. DEXA use two energy sources to distinguish between the soft tissue and heart issue that uses higher energy and low energy to measure bone density and currently is well-known gold standard excepted by WHO to measure bone density but DEXA also have certain limitation major in lack of standardization in Bone tissue measurement. The regional assessment is inferior in DEXA, body part considered for measurement are very few it's costly and not affordable by economically weaker society. so, there is a lot of scope to research and come up with a tool to predict the bone density using x-rays due to its affordability and traditionally accepted way to diagnosis osteoporosis condition. It is very much observed in people of age group 40-80years men and women especially post-menopausal woman are at high-risk condition is left untreated it will result in permanent disability or cost one's life. It's important to address this condition and make affordable device so that each and every country can afford it. Osteoporosis is condition which makes bone more fragile it's because our body tends to reabsorb the old one and create new bone in the body, when the absorption is more than the creation there is a loss in the bone tissue which leads to bone mineral loss. This loss results in fracture occurrence in hip, vertebrae, spine etc. And people who have had fracture have the risk of having the second fracture occurrence and various bone break. So, bone mineral density test is the well-known method currently and based on all above literature survey and deficiency in the

current state of art there is a need to research in the bio medical field and come up with affordable tools to solve the problems of human kind.[1]

Deep learning techniques have shown promising results in image classification tasks, including medical image analysis.[4] In particular, the Vision Transformer (ViT) architecture has gained attention as a powerful tool for image classification, even surpassing traditional convolutional neural networks (CNNs) in certain cases.[18] ViT replaces the patch-based approach of CNNs with a permutation-based approach, treating each pixel in the image as a separate entity. This results in improved performance on certain image classification tasks, making it a promising approach for osteoporosis diagnosis using X-ray images.[19]

In this paper, we present custom x-ray database can be used to carry out research related to bone as there is no much research done and data available on imaging modalities due to ethical restrictions.[24] most of them have discussed various methods related to imaging modalities but no such data is made available for research Publicly, motivated by this we are offering our custom database and a ViT classification model for osteoporosis diagnosis using X-ray images. We evaluate the model on a dataset of X-ray images from both normal and osteoporotic patients, and compare its performance to traditional CNN-based models. Our results demonstrate the potential of ViT for accurately diagnosing osteoporosis from X-ray images, which could ultimately improve patient outcomes and quality of life.

Our custom database consists of Spine, Knee, Hand, Femur, Leg, Shoulder x-ray scan images with her patient ID, gender, position, age, diagnosis.

## 1. Related work

Dai et,al. The transformer has received less attention in medical field of image analysis. TransMed, a novel image classification design based on transformers, is presented in this paper. TransMed employs a hybrid model that combines CNN and transformers, with CNN acting as a low feature extractor to generate feature sequences of images and transformers extracting high dependencies among sequences for improved performance. TransMed outperformed previous state-of-the-art models in two multi-modal medical image classification datasets, improving accuracy by 10.1% and 1.9%, respectively. These findings show the utility of deep network transformers for image analysis, especially in multi-modal systems. Thus, the transformer structure shows promise in numerous medical image analysis tasks.

Iqbal et,al. One of the most often found malignancies in women is breast cancer. Early breast cancer detection is crucial, and various imaging modalities such as As initial screening procedures, ultrasound, mammography, and MRI are employed. However, random variation, amorphous shapes, and hazy surroundings of tumor area, accurately segmenting breast tumours is a difficult task. The existing CNN-based methods have limitations in extracting information about the overall context, with less effective outcomes. The BTS-ST network, which combines Swin-Transformer into conventional CNN-based U-Net to enhance global modelling capabilities, is suggested as a solution to this problem. The proposed network also includes Spatial Interaction, Feature Compression, and Relationship Aggregation blocks to Increase the segmentation accuracy of small tumour regions, combine global dependencies from Swin-Transformer and CNN hierarchically, and feature representation capabilities of irregularly shaped tumours. The results of the suggested method's evaluation on multimodality datasets based on ultrasound, mammography, and MRI revealed that it performed better than other cutting-edge approaches.

Xin et, al. Skin cancer is a growing threat to human health, and the application of artificial intelligence to identify dermoscopic images has resulted in significant advances in disease detection and treatment. However, existing approaches based on CNNs only extract features of small objects and fail to locate important parts. To address this issue, the researchers propose SkinTrans, an improved transformer network, which incorporates vision transformers (VIT) that have shown impressive results in conventional categorization challenges. The SkinTrans model features overlapping, multi-scale sliding windows, multi-scale patch embedding, and label shuffling for a balanced sampling of skin cancer datasets. The model is used on two skin cancer datasets achieving promising results, demonstrating the potential of using VIT-based models in skin cancer classification.

Le Dinh et, al. The author proposes the use of chest X-ray images as a fast-screening method for

early COVID-19 detection. The authors created X-ray datasets of chest images by collecting images from available datasets. They conducted experiments on different deep learning models using both transformer- and convolution-based methods, to classify COVID-19, pneumonia, and normal cases, as well as assess COVID-19 severity. Results indicate that using chest images for COVID-19 detection with promising result and reliable performance. Transformer-based models outperformed on all metrics.

Wang et,al. The challenge of accurately detecting and adapting to new strains of coronaviruses in respiratory diseases such as SARS-CoV-2 is significant. In order to solve this, researchers suggest using chest radiography pictures to diagnose and prevent a variety of coronavirus infections using a Continuous Learning method called CoroTrans-CL. To lessen the issue of catastrophic forgetting, this method combines the Elastic Weight Consolidation and Herding Selection Replay techniques using the Swin Transformer architecture. The researchers created a useful benchmark dataset with diverse coronavirus strains, and they explain the suggested method in five iterative learning stages that depict the timeline of various coronavirus epidemics. This method offers a practical remedy for continuous diagnosis of infections with mutant SARS-CoV-2 viruses.

Okolo et,al. This study examines using a deep learning network based on Transformers to classify chest X-ray images automatically to address the bottleneck of requiring expert radiologists in remote areas. The study evaluates the performance using four chest X-ray image datasets encompassing varied diseases, the state-of-the-art model, ViT, was compared to an unique enhanced Vision Transformer model called IEViT. Findings revealed that IEViT outperformed ViT across all datasets analysed in terms of F1-score, sensitivity, and precision, reaching an F1-score between 96.39% and 100% and outperforming ViT by up to +5.82%. The proposed model demonstrates promising potential for accurately identifying the many diseases visible in chest X-ray images.

Manzari et, al. The effectiveness of deep medical diagnosis systems in preventing adversarial attacks is a concern, as inaccurate diagnosis could have disastrous consequences. To deal with this, we suggest a hybrid CNN-Transformer model that combines the minimizing the high quadratic complexity of self-attention while leveraging the localization of CNNs and the global connection of Transformers with an efficient convolution operation. We also aim to learn smoother decision boundaries by augmenting shape information through feature permutation in mini-batches, improving generalization ability. Our model outperforms state-of-the-art studies on standardized MedMNIST-2D datasets with less computational complexity.

## 2. Data

In the medical field there is no specific database explicitly made to evaluate BMD and made publicly available for researchers to carry out research.[5] Considering the current scenario, we have created custom database which can be used to measure bone mineral density from the bone x-ray scan images.

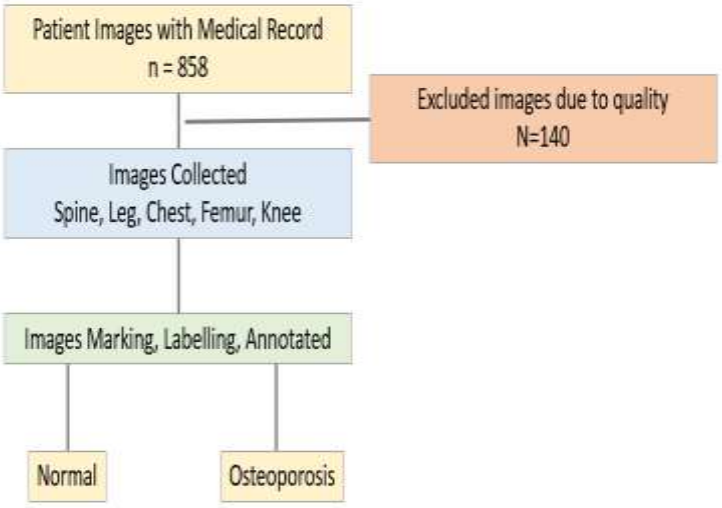



Figure 1. Shows Database Structure

The custom database has 858 x-ray images from research hospital with various scan Images. It has 488 female and 370 male images that include Spine, Knee, Hand, Femur, Leg, Shoulder x-ray bone scan images. The scanned images are categorized as follows

- Image details
- Labelling Images
- Annotating images

The details are explained below

### 3.1. Image details

The custom databases are created with respect to Indian context. Among 858 x-ray images totally there are 206 Spine images, 350 Knee images, 52

Hand images, 128 Femur images, 80 Leg images, 42 Shoulder images of x-ray bone.

The images are collected from the hospital in person from the period of 2018 to 2022 physically in Diacom format. The details of the image collected or presented in the table format. The custom dataset is being grouped into seven groups such as a SP, KN, H, FE, L, SH where are 206 Spine images, 350 Knee images, 52 Hand images, 128 Femur images, 80 Leg images, 42 Shoulder images of x-ray bone. This grouping of particular images in categories makes researchers work with each category and understand efficiently.[6]

**3.2. Labelling images**

The custom data set is being annotated carefully so that it is easy to experiment and understand. The images are converted from Diacom to PNG portable network graphics format and labelled. The labelling consists of Patient ID, age, gender, body part and diagnosis details for example the labelling of the images KN_001.PNG which refers to x-ray image where KN represent Knee image with patient ID 001. Similarly, SP_001.PNG which refers to x-ray image where SP represents spine image with patient ID 001. FE_001.PNG refers to x-ray images where FE represents femur image with patient ID 001. L_001.PNG refers to x-ray image L represents leg image with patient ID 001.H_001.PNG refers to x-ray means where H represents Hand image with patient ID 001. SH_001.PNG refers to Shoulder image with patient ID 001.

| Knee | Femur | Shoulder | Hand | Leg |
|---|---|---|---|---|
| Subject Id: KN_001 | Subject Id: FE_001 | Subject Id: CT_001 | Subject Id: H_001 | Subject Id: L_001 |
| Age:43 | Age:68 | Age:35 | Age:78 | Age:54 |
| Gender: Male | Gender: Female | Gender: Male | Gender: Female | Gender: Female |
| Diagnosis: Normal | Diagnosis: Osteoporosis | Diagnosis: Normal | Diagnosis: Osteoporosis | Diagnosis: Osteopenia |

Table 1 Shows Image Labels in the Database

**3.3. Annotating images**

The custom databases being labelled with the supervision of medical practitioner for all images. They are all carefully annotated manually as given below

- patient ID
- age
- gender
- diagnosis
- part of the image
- anterior or posterior

The custom data set is created to solve various problems associated in orthopaedics domain related to Bone and make young research to motivate and solve the issues associated by making data available publicly. The labelling will make the researcher work easier to carry out the analysis with proper Report related to spine, femur, hand leg, elbow, Femur, chest annotating of x-ray scan images.

Figure 2 Shows Sample Database Images

**3.4. Data Analysis**

Osteoporosis is the disease that is related to human bone mineral density and the diagnosis is

by analysing BMD values in the patient based on the trabecular bone

pattern that results in wear and tear of Bone tissues.[3] One major problem in this osteoporosis condition is increasing day by day and greater attention should be taken to solve this medical problem from childhood to adult age. so that people don't suffer from severe disability due to occurrence of fractures and breakage of Bone because the healing may or may not be possible at later age. Currently the well-known a DEXA defines BMD with T-score and Z-score values. T-score ranges -1 standard deviation is normal, -1 to -2.5 standard deviation is Osteopenia and below -2.5 standard deviation is for osteoporosis. The interpretations of data annotation are shown in the Table 1. Looking at the obtained BMD values and evaluation of these T-score, Z-Score values proper medication can be given to patients and treated such as Calcitonin, Hormone replacement therapy, bisphosphonates & SERMS as prescribed by medical practitioner orthopaedist.[2] By properly analysing and treating patients with their medical history report obtained one can be treated for osteoporosis disease. The validation of accuracy can be realized by verifying their models and reports obtained from medical practitioner with appropriate analysis.

## 3. Research Methodology

The ViT model can also be used for osteoporosis diagnosis. The input X-ray image is divided into N patches, and linear embeddings are computed for each patch. To preserve the patch's positional information, position embeddings are added to these embeddings. For classification, learnable patch is added using a multilayer perceptron (MLP) head. After that, position embeddings and combined patch are sent into the transformer encoder model. The pretrained ViT is used in this application Models are fine-tuned on X-ray images to improve their performance on the osteoporosis diagnosis task. The X-ray images are scaled to the necessary resolution, and because the pretrained ViT models need three input channels, the same grayscale X-ray image is transferred into the remaining two channels.

The ViT model works by dividing N patches of size PPC from image I, which has dimensions RHWC, where N = HW/P2 (H: height, W: width, C: number of channels). Each patch is then processed using a linear projection to generate patch embeddings x1pE, x2pE, ..., xNpE. To retain the positional information of each patch, in position embeddings Epos are added to the patch embeddings. In addition, learnable embedding I class is appended with patch embedding series, similar to [class] in BERT.[25]

The transformer encoder model, combination of multi-layer perceptron (MLP) blocks and alternating layers of multiheaded self-attention (MSA), is then fed the patch and position embeddings. The input is subjected to layer normalization (LN) in first block of transformer layer, followed by multiheaded self-attention and residual connection. Next, second block applies LN to the output of the first block, by a multi-layer perceptron and another residual connection. The multi-layer perceptron consists of two FC layers with Gaussian error linear unit nonlinearity.

After passing through the transformer encoder layers, the output of the last layer is further normalized using LN to generate a final latent representation z0L. This latent representation is used to perform classification via an attached MLP head. The final output y is obtained by applying LN to z0L.[7]

For the specific task of osteoporosis diagnosis using X-ray images, the model is trained to classify the X-ray image as either normal or osteoporosis. ViT model can be a promising approach for osteoporosis diagnosis using X-ray images by leveraging its ability to extract larger dependencies and position information.

### 4.1. Data Pre-processing

The input images are being resized to a fixed size to ensure that all images have the same dimensions. The size of the images can be set to 224 x 224 pixels or higher depending on the requirements of the ViT model.

The pixel values of the images are normalized to ensure that they lie in the range of -1 to 1, which is required by ViT models. Normalization is performed after resizing the images to ensure that all images have the same size and pixel value range. Normalization can be performed by dividing the pixel values by 255 (maximum pixel value) and then subtracting 0.5, so that the pixel values lie in the range [-0.5, 0.5]. Finally, the pixel values are multiplied by 2 to obtain the desired range of [-1, 1].

Both resizing and normalization are crucial steps in data pre-processing for ViT models, as they ensure that input are of similar size and have same pixel value range, which is necessary for effective training and accurate predictions.[23]

### 4.2. Data Augmentation

Data augmentation is a well-known method in deep

learning to artificially increase the volume of train data creating new, modified versions of the original data. The goal is to improve the system ability to generalize to unseen data.

In the case of the ViT model, data augmentation can be used to create new variations of the input images by applying a series of random transformations. These transformations can include cropping, flipping, rotation, and color adjustments, among others.[22] To apply data augmentation to the training data, we used the PyTorch transforms module. This module provides a set of pre-defined image transforms that can be combined into a pipeline using the transforms.Compose() function. This pipeline can then be applied to the input data using a PyTorch Dataset and DataLoader. It's important to note that data augmentation should only be applied to the training data, and not to the validation or test data. This is because we want to evaluate the model's performance on the original, unmodified data. Additionally, it's important to choose the appropriate techniques depending on the nature of data.[7]

Overall, data augmentation is a powerful technique that can help improve the performance of the ViT model, especially when the size of the training dataset is limited.

**4.3. ViT Architecture for Classification**

ViT, or Vision Transformer, is a deep learning architecture that uses the principles of the Transformer model to analyze images by Dosovitskiy et al. in 2020. ViT is a purely attention-based architecture that eliminates the need for hand-designed convolutional neural network (CNN) layers and achieves superior performance on various image classification benchmarks.[9]

The ViT architecture consists of three main components: patch embedding, multi-head self-attention, and a fully connected MLP head.[8]

1. Patch Embedding: In this step, the input image is cut into set of patches, each of size P x P pixels, and flattened into a 1D vector. The patch size can be set to 16x16 or 32x32 pixels depending on the image size and complexity. The flattened patches are then projected into a high-dimensional vector space using a learnable linear projection. The resulting vectors are referred to as patch embeddings and are the input to the multi-head self-attention layer.
2. Multi-Head Self-Attention: In this layer, the patch embeddings are fed into multiple attention heads that compute attention weights for each patch based on its relationships with other patches in the image. The attention weights are computed using dot product attention, where each patch is treated as a query and the other patches are treated as keys and values. The attention weights are then used to compute a weighted sum of the patch embeddings, which represents a global representation of the image. This process is repeated for each attention head, resulting in a set of attention-based features for the image.
3. MLP Head: In this final step, the attention-based features are passed to sequence of fully connected layers (MLP) to produce the final output. The MLP head can have one or more layers and can be followed by a SoftMax layer for classification. The model is optimized by standard backpropagation and stochastic gradient descent during training.

$SoftMax(i) = e^{(i)} / sum(e^{(j)})$, for j=1 to N

where i is a particular value, e is the exponential function, and j iterates over all values in the set. The softmax function normalizes a set of values into a probability distribution, where each value is transformed into a probability value between 0 and 1, and the sum of all probabilities equals 1.

One of the main advantages of ViT is that it can be trained end-to-end on large-scale datasets without the need for hand-designed CNN layers. This makes ViT a versatile and effective architecture for various image classification tasks.[21]

In summary, the ViT architecture uses the Transformer model to perform global context reasoning on a sequence of image patches, which allows it to achieve best performance on image classification models. Visual transformers are a type of deep learning model which exhibits amazing performance in computer vision problems, including medical field.[20] They can be used to extract meaningful features using medical images and classify them based on various criteria, including the presence of osteoporosis. However, it is important to note that using visual transformers for osteoporosis classification requires a large amount of high-quality data, which can be challenging to obtain. In addition, developing an accurate visual transformer model needs evaluation of numerous parameters, with architecture of the model, the amount of dataset, and the training and validation procedures.

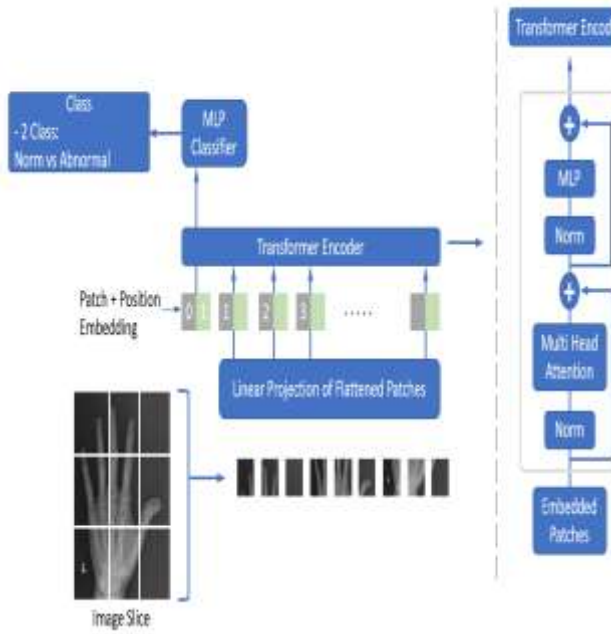



Fig 3. shows Architecture Diagram of ViT

Algorithm Vision Transformer

1. Initialize the model parameters including the weights and biases of the attention layers and feedforward layers.
2. Load the training data and labels.
3. Set the learning rate, LR, to 0.0001 and the batch size, BS, to 16.
4. Divide training data into batches of size BS.
5. For each batch, perform the following operations:
    - Initialize the gradients of the model parameters to zero: grad_W = 0, grad_b = 0, grad_H = 0.
    - Perform data augmentation on the images in the batch.
    - Reshape the images to the required size of 224x224: X = reshape (X, (BS, 224, 224, 3)).
    - Compute the forward pass of the ViT model on the batch of images: Z = ViT_forward(X, W, b, H).
    - Compute the loss using the cross-entropy loss function: L = cross_entropy_loss(Z, Y).
    - Compute the gradients of loss with the model parameters using backpropagation:

      grad_Z = derivative_of_cross_entropy_loss(Z , Y)

      grad_X, grad_W, grad_b, grad_H = ViT_backward(grad_Z, X, W, b, H)
    - Update the model parameters using stochastic gradient descent (SGD) with the learning rate of LR:

      W = W - LR * grad_W

      b = b - LR * grad_b

      H = H - LR * grad_H
6. Repeat steps 5 for all batches in the training data.
7. Calculate the accuracy on the validation set using the forward pass of the ViT model on the validation data: val_accuracy = ViT_accuracy(val_data, val_labels, W, b, H).
8. If the validation accuracy has improved since the previous epoch, save the model parameters: if val_accuracy > best_val_accuracy: best_val_accuracy = val_accuracy;

best_W = W;

best_b = b;

best_H = H.

9. Repeat steps 5 to 8 for a total of 500 epochs.

## 4. Result and Discusion

The input image intensities were initially rescaled to a range of -1 to 1 as per the requirement for ViT models. The incorporation of Techniques for enhancing the data, such as random cropping and flipping, increased the model's precision. The ViT architecture was fine-tuned during the training process with all its parameters. For an Using a 224 x 224 resolution of image, the hyperparameters that were optimized based on validation accuracy were the RectifiedAdam optimizer with a learning rate of 0.0001, number of epochs as 500, and a mini-batch size of 16. Among all the models, The best performance was achieved using B/16 model, with 88.71% validation accuracy.

During the classification task, the loss metric used cross-entropy as evaluation metric for both training and validation. The confusion matrix and loss metrics to evaluate the model during testing. Additionally, the performance of the models is analyzed using the Precision, Recall, F1-score respectively. The hyperparameters are optimized using the validation set. The best hyperparameters, which resulted in the highest accuracy during 5-fold cross-validation, are selected to evaluate the metrics on test set.

Performance metrics measured

1. Accuracy: The proportion of correct predictions to the total number of predictions. Accuracy = (TP + TN) / (TP + TN + FP + FN)

where TP is true positives, TN is true negatives, FP is false positives, and FN is

Δ

false negatives.

2. Precision: The proportion of true positives to the total number of positive predictions.

Precision = TP / (TP + FP)

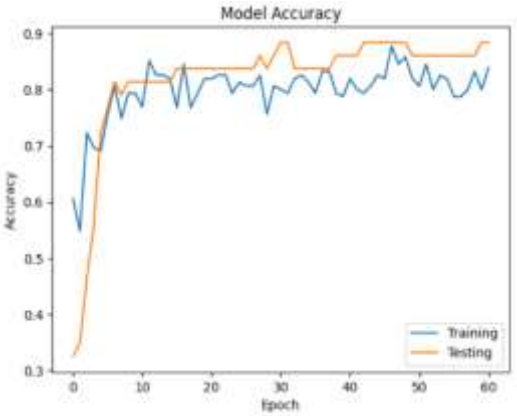
3. Recall (sometimes termed sensitivity): The proportion of genuine positives to the total number of actual positives.

Recall = TP / (TP + FN)

4. F1-score: A precision and recall weighted average that accounts for both metrics.

F1-score = 2 * (precision * recall) / (precision + recall)

5. Confusion matrix: A table that summarizes the performance of a classifier by showing the number of true positives, true negatives, false positives, and false negatives.




The confusion matrix analysis revealed that the model achieved a high sensitivity and specificity for both the normal and osteoporotic classes, which is crucial for accurate diagnosis. The high-performance metrics demonstrate that the ViT model is a promising approach for osteoporosis diagnosis using X-ray images.

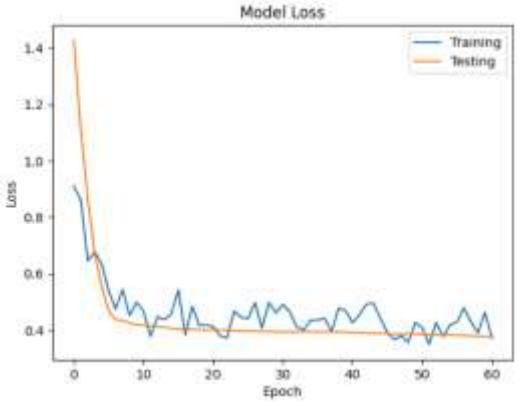



Figure 4. shows Training Curves of Model Accuracy and Model Loss

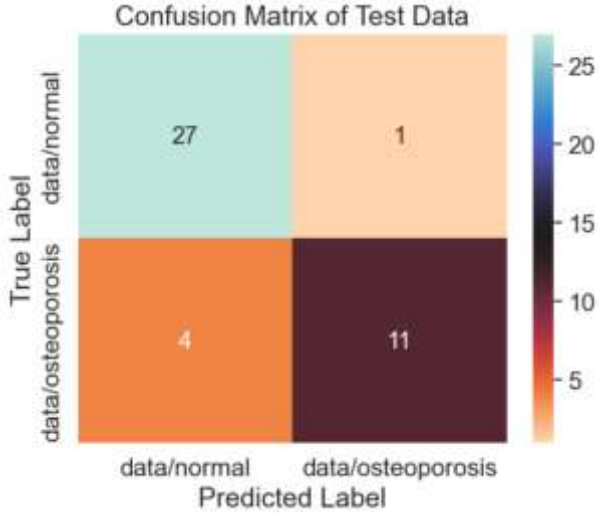



Figure 5. shows Confusion Matrix

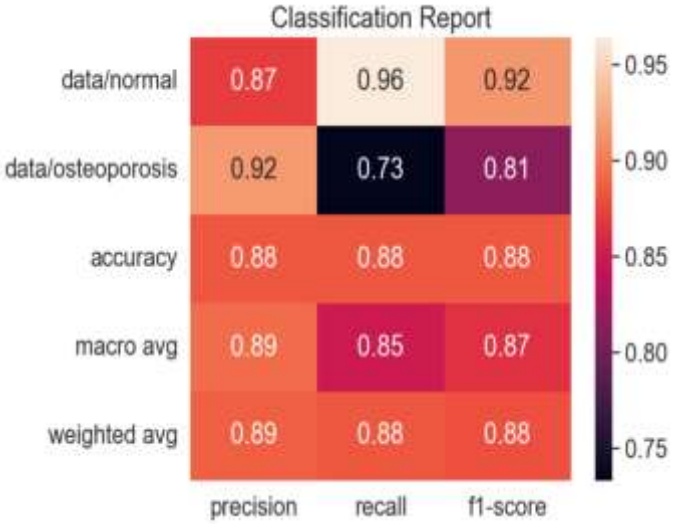



Figure 6. Shows Classification Report

## 5. Conclusion

In this paper we present custom database related to bone x-ray scan images for Orthopedic and biomedical research that represent necessary data and details required to perform Bone related research with proper

image data structure details, labelling, annotation annotating details of Femur, Spine, Hand, Chest, Leg, elbow, Knee x-ray bone images by making them available to research community educational purposes. To identify and solve the problem with respect to Bone Disease. The custom databases is acquired and made available by following all the necessary formalities mentioned in the medical field is being obtained from Hospital. The ViT classification model shows promise for osteoporosis diagnosis using X-ray images. Our ViT model achieved a high accuracy of 88% on the validation set, and a similar performance was achieved on the test set. However, more extensive evaluations on larger datasets and compared with other state-of-the-art models are necessary to establish the full potential of the ViT model for osteoporosis diagnosis. Overall, our results suggest that the ViT model has the potential to be an effective tool for osteoporosis diagnosis, with the advantage of being computationally efficient and easily interpretable. Fine-tuning on a larger X-ray image dataset can further improve its performance on this task. This database will be made available to researchers based on the request and agreeing certain ethical formalities.

## References

1. Fathima S. M. N, Tamilselvi R, Beham M. P. XSITRAY: A Database for the Detection of Osteoporosis Condition. Biomed Pharmacol J 2019;12(1).
2. Subasinghe HWAS, Lekamwasam S, Ball P, Morrissey H, Waidyaratne E. Estimating regional bone mineral density-based T-scores using clinical information; tools validated for postmenopausal women in Sri Lanka. Osteoporos Sarcopenia. 2020 Sep;6(3):122-128. doi: 10.1016/j.afos.2020.08.004. Epub 2020 Sep 16. PMID: 33102805; PMCID: PMC7573505.
3. ESTIMATION OF BONE MINERAL DENSITY TO IDENTIFY OSTEOPOROSIS, Journal of Xi'an University of Architecture & Technology ISSN No : 1006-7930, Volume XIII, Issue 6, 2021
4. Jang, Miso, et al. "Opportunistic osteoporosis screening using chest radiographs with deep learning: Development and external validation with a cohort dataset." *Journal of Bone and Mineral Research* 37.2 (2022): 369-377.
5. MS, Lin MH, Lee CP, Yang YH, Chen WC, Chang GH, Tsai YT, Chen PC, Tsai YH. Chang Gung Research Database: A multi-institutional database consisting of original medical records. Biomed J. 2017 Oct;40(5):263-269. doi: 10.1016/j.bj.2017.08.002. Epub 2017 Nov 10. PMID: 29179881; PMCID: PMC6138604.
6. Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, Maria de la Iglesia-Vayá, PadChest: A large chest x-ray image dataset with multi-label annotated reports, Medical Image Analysis, Volume 66, 2020,101797, ISSN 1361-8415, https://doi.org/10.1016/j.media.2020.101797.
7. Tummala, Sudhakar, et al. "Classification of Brain Tumor from Magnetic Resonance Imaging Using Vision Transformers Ensembling." *Current Oncology* 29.10 (2022): 7498-7511.
8. Shin, Hyunji, et al. "Vision Transformer Approach for Classification of Alzheimer's Disease Using 18F-Florbetaben Brain Images." *Applied Sciences* 13.6 (2023): 3453.
9. Qi, Zheng, et al. "Privacy-preserving image classification using vision transformer." *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022.
10. Dai, Yin, Yifan Gao, and Fayu Liu. "Transmed: Transformers advance multi-modal medical image classification." *Diagnostics* 11.8 (2021): 1384.
11. Iqbal, Ahmed, and Muhammad Sharif. "BTS-ST: Swin transformer network for segmentation and classification of multimodality breast cancer images." *Knowledge-Based Systems* (2023): 110393.
12. Xin, Chao, et al. "An improved transformer network for skin cancer classification." *Computers in Biology and Medicine* 149 (2022): 105939.

13. Le Dinh, Tuan, et al. "Covid-19 chest x-ray classification and severity assessment using convolutional and transformer neural networks." *Applied Sciences* 12.10 (2022): 4861.
14. Wang, Boyuan, Du Zhang, and Zonggui Tian. "CoroTrans-CL: A Novel Transformer-Based Continual Deep Learning Model for Image Recognition of Coronavirus Infections." *Electronics* 12.4 (2023): 866.
15. Okolo, Gabriel Iluebe, Stamos Katsigiannis, and Naeem Ramzan. "IEViT: An enhanced vision transformer architecture for chest X-ray image classification." *Computer Methods and Programs in Biomedicine* 226 (2022): 107141.
16. Manzari, Omid Nejati, et al. "MedViT: A robust vision transformer for generalized medical image classification." *Computers in Biology and Medicine* (2023): 106791.
17. Ma, DongAo, et al. "Benchmarking and Boosting Transformers for Medical Image Classification." *Domain Adaptation and Representation Transfer: 4th MICCAI Workshop, DART 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*. Cham: Springer Nature Switzerland, 2022.
18. Sheng, Yiwei, and Sihan Ren. "Medical image classification based on enhanced Vision Transformer." *International Conference on Electronic Information Engineering, Big Data, and Computer Technology (EIBDCT 2022)*. Vol. 12256. SPIE, 2022.
19. Xue, Linyan, et al. "Osteoporosis Prediction in Lumbar Spine X-Ray Images Using the Multi-Scale Weighted Fusion Contextual Transformer Network."
20. Xiong, Yuxuan, Bo Du, and Pingkun Yan. "Reinforced transformer for medical image captioning." *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*. Springer International Publishing, 2019.
21. Tanzi, Leonardo, et al. "Vision transformer for femur fracture classification." *Injury* 53.7 (2022): 2625-2634.
22. Steiner, Andreas, et al. "How to train your vit? data, augmentation, and regularization in vision transformers." *arXiv preprint arXiv:2106.10270* (2021).
23. Gao, Xiaohong, Yu Qian, and Alice Gao. "COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models." *arXiv preprint arXiv:2107.01682* (2021).
24. Duong, Linh T., et al. "Detection of tuberculosis from chest X-ray images: Boosting the performance with vision transformer and transfer learning." *Expert Systems with Applications* 184 (2021): 115519.
25. J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL HLT 2019–2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf. 1 (2018) 4171–4186. https://doi.org/10.48550/arxiv.1810.04805 .
26. Sukegawa, Shintaro, et al. "Identification of osteoporosis using ensemble deep learning model with panoramic radiographs and clinical covariates." *Scientific reports* 12.1 (2022): 1-10.