ENGINEER - Vol. LVIII, No. 04, pp. [23-36], 2025

DOI: https://doi.org/10.4038/engineer.v58i4.7714

Enhanced Prediction of Osteoporosis with Improved Contextual Accuracy Utilizing Expanded Classifications and Explainable AI

H. M. Gammulle and C. K. Walgampaya

Abstract: Many classification models shown in medical research achieve high statistical accuracy, yet are not examined for their contextual accuracy, lack of which can lead to misleading predictions that do not align with human reasoning or the science involved. This will cause distrust and alienate those unfamiliar with deep learning and training models, even if they understood the medical condition being predicted. To address this, explainable AI can evaluate a model's contextual accuracy by assessing how reasonable its decision-making process appears to be. In this paper utilization of class augmentation as a method to improve contextual accuracy in classification models is explored. This forces the model to rely on finer details within the image, ultimately enhancing its reasoning ability. This is explored within the context of using AI models to diagnose patients with Osteoporosis using X-ray images instead of DXA. Osteoporosis is a silent, bone density disorder that needs to be diagnosed as early as possible to avoid many complications, since in serious cases, even a simple sneeze can fracture a bone. Understanding why a model would classify a patient as healthy or having osteoporosis would help physicians judge this decision better. Explainable AI is used to evaluate this decision making, providing insights into how well the model's decision making aligns with human logic. To validate the findings, two different x-ray datasets were used, where the images within each dataset appeared visually similar to one another within the dataset itself. Results show that utilizing class augmentation when training a model before later training it for a specific use case with a smaller number of classes can give better explainability to the resulting model. In a medical setting, models can be forced to learn better clinical markers during training. Additionally, this had consistently lowered the number of false positives.

Keywords: Bone mass density, Contextual accuracy, Deep learning, Explainable AI, Feature recognition, Osteoporosis, X-ray images

1. Introduction

Osteoporosis is a bone density disorder where the body fails to maintain the bone condition properly. It is considered to be a silent disease that shows little to no symptoms till the last minute. This leads to weak bones with low mineral density that may break from a simple fall or, in more serious cases, a sneeze. A patient can break a bone while doing very normal day to day activities if they have osteoporosis. Breaking a bone itself is a serious complication that often leads to permanent pain and can even contribute towards patients losing height. It can also cause some patients to develop a stooped or hunched posture. Osteopenia is the state of having less bone mineral density (BMD) than normal, yet is not as severe as Osteoporosis is [1]. Bone density disorders are identified in practice using a special method called Dual Energy X-Ray Absorptiometry (DEXA or DXA) where two different x-rays are used in small doses with different energy peaks. The two different x-rays are absorbed differently by both bone and soft tissue. Using the image captured this way, two proxy values for bone density measurements are obtained, named T-score and Z-score [2], [3]. Though this is the definitive way of quantitatively measuring bone mineral density and osteopenia/osteoporosis, there have been many efforts to use normal x-ray images to detect and/or predict osteoporosis in a patient, where much of these efforts involve utilizing deep learning for this task. Some of them will be briefly explained within this paper.

Mr. H. M. Gammulle, Mechanical Engineering Undergraduate, Faculty of Engineering,University of Peradeniya, Kandy. Sri Lanka
Email:harindugml@gmail.com
https://orcid.org/0009-0005-8044-1360

Eng. (Prof.) C. K. Walgampaya, AMIE(SL), B.Sc. Eng. (Hons)(Peradeniya), MSc. (Louisville, USA), PhD (Louisville, USA),
Professor, Faculty of Engineering,
University of Peradeniya.
Email:ckw@pdn.ac.lk
https://orcid.org/0000-0003-2192-474X

Deep learning has many applications that makes diagnosis and treatment of diseases more convenient and sometimes, actually possible compared to how it used to be [4]. The field of AI is a rapidly evolving field where Convolutional Neural Networks (CNNs) hold a significant place as powerful tools that can extract data in the form of a simple .csv file or a database of thousands of images. CNNs are a type of artificial neural networks used in deep learning that are composed of an input layer, several different types of hidden layers including convolution layers and an output layer. The convolution layers store weights and filters that allow the network to extract features from the data given as input [5]. Though CNNs can be made from scratch, better results can be obtained by using an already existing model that is provided by various groups. These models can be altered to fit the users' requirements. This type of model is referred to as a transfer learning model. In this paper, a transfer learning model is made using the ResNet50 model provided by keras, a residual deep learning model [6] and implemented to detect osteoporosis.

Although the training process is seemingly simple, the outcome of it, the model that is obtained, is commonly referred to as a black box that magically does what it was trained to do. The user is given no reason as to why the model does what it does, even though it does so with good accuracy. Contextual accuracy is an important factor for determining how a model understands the data it processes to give an output. This is a relative metric which is used to compare a model's ability to gather information from data, compared to how an average human being can do the same. Contextual accuracy can help determine if a classification model has learned patterns that align with its training data properly, so as to enable it to perform well when it is given data that it has not seen before, which is the end use case for such models if they get implemented in the real world.

This is the reason why explainable AI (XAI) has been an emerging field as means to explain how a model functions and how it understands data. In this research paper, Shapley Additive exPlanations (SHAP) are used to describe the ability to interpret information from data of different models. It is based on the Shapley Value, which is a concept used in game theory. When used in machine learning context, it helps recognize the contribution of features of data that led to the model's prediction [7]. Using explainable AI tools, the contextual accuracy can be evaluated through random samples from the dataset. This becomes increasingly important as the problem definition becomes simpler. For example, "what is a car?". A model trained to detect cars might only focus on wheels for detection. For such a model, an image of a garage with 2 wheels showing would also classify as a car. This is an error an average human being would not make. As such, evaluating the contextual accuracy would enable AI systems with high accuracy as well as high contextual accuracy to be more acceptable and believable compared to the current state where they are simply a black box that magically predicts an outcome.

While there is a possibility that research in this field would help uncover a completely unrelated way to definitely predict if a patient has osteoporosis or not using x-ray imagery instead of DXA, the currently accepted methods mentioned before involve calculating a proxy value for bone mineral density, where the area of the bone within the x-ray image is used. As such, the same logic is applied to a model trained on these images. Current studies in detecting osteoporosis using deep learning techniques lack depth considering the explainability of the models trained. Purpose of this research is to propose a model that has high accuracy as well as proper explainability for detecting osteoporosis using x-ray imagery compared to the unexplained models proposed so far. After all, telling a patient that they are osteoporotic purely because they are left-handed would seem unscientific, unreasonable as well as unbelievable for the patient.

2. Related Works

There have been many attempts at accurately predicting osteoporosis using x-ray imagery using models with different architectures. While transfer learning had been commonly utilized in this aspect in various studies, the explainability of these methods hasn't been explored. As explained earlier, such a black box model has unfavourable implications in the medical field. The paper by Wani et al. from 2022 [8] discusses using several transfer learning models based on AlexNet, VggNet-16, VggNet-19, and ResNet where AlexNet was able to achieve the highest accuracy of 91.1%. This study had focused on the comparison of statistical performance of different transfer learning models, with their explainability not being within their scope. In a different paper by Derkatch et al. an ensemble method was used to predict nonvertebral and hip fractures. This study found that their convolutional neural

network (CNN) had correctly identified 87.4% of vertebral fractures as well as 88.4% of non-vertebral fractures [9]. Another paper by Krishnaraj et al. discusses simulating DXA in CT using deep learning and segmentation where they managed to achieve an 82% overall accuracy [10]. Lee et al. in 2020 explored feature extraction and machine learning in the context of using it to predict bone density from spine x-ray images where they utilized various classification algorithms and feature extraction models. The best result they got was 0.66 precision with a F1 score of 0.73 [11]. While these studies had focused on various methodologies to yield these results, they had not explored the explainability of their models. To our knowledge, no prior study had integrated post-hoc interpretability methods to assess the contextual accuracy of a model, a gap our work addresses.
In 2023, Zhu et al. have discussed the effect of training using k+m classes for a model that needs to classify for k classes and how it affects the model's generalization and reliability performance. What they found was that the added m classes served as regularization [12]. Here, they dive deep into class augmentation and come to the conclusion that class augmentation can improve confidence reliability for misclassification detection, confidence calibration and out-of-distribution detection. The aforementioned study is directly related to the gap observed in the previous work mentioned earlier.
Within the field of explainable AI, SHapley Additive exPlanations (SHAP) is a commonly used tool to help interpret model predictions. It is commonly used in the medical field to analyse data and discover new insights about different medical conditions. In reverse, if we know what to look for, we can evaluate the contextual accuracy of a model as well. This ensures that the model would be reliable and would provide meaningful predictions. A study by Hu et al. in 2023, used this method to predict bone cement leakage in patients who underwent percutaneous vertebral augmentation (PVA) which is done for osteoporotic vertebral compression fracture (OVCF). They used SHAP to identify several important factors related to cement leakage through this [13]. Another paper done by Khanna et al. used SHAP and other similar tools such as LIME (Local Interpretable Model Explainer), ELI5, Qlattice, etc. to assess the influence various factors posed towards risk of getting osteoporosis [14]. Another study by Liu et al. explores accuracy of models made to identify patients with high-risk for fracture and implementing early interventions by utilizing SHAP in their study [15].

3. Methodology

3.1 Datasets

This study utilized two different datasets. The main dataset used was for the classification of osteoporosis and comprised x-ray images of the knee area. It was divided into three classes: normal, osteopenia and osteoporosis [16]. The second dataset used was utilized as a control experiment to evaluate the results obtained by using the main dataset. This dataset comprised chest X-ray images and was classified as normal, pneumonia and covid [17].
A conditional preprocessing algorithm was followed for the main dataset, named 'Dataset 1' shown in Figure 1. This was done to address various issues in the dataset that led it to be inconsistent and had an imbalance of classes. The way these two datasets were processed and used is shown below.

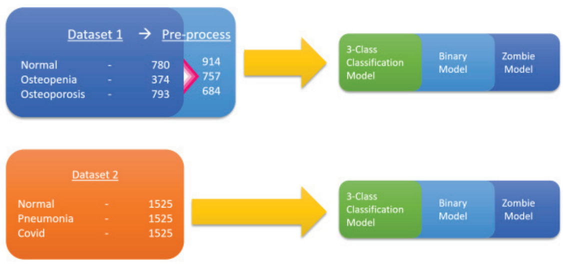

Figure 1 - Utilization of the two datasets

The resulting pre-processed dataset was then pre-processed by adding noise and decreasing the brightness of the image. The second dataset used, named 'Dataset 2' above, was only pre-processed to add noise and change brightness. Both datasets were then used to train three different types of models for each dataset, which are explained further in this paper.

3.2 Preprocessing

Many steps were followed for preprocessing the knee x-ray images in the main dataset due to many reasons. These steps were taken separately to address irregularities within the dataset, non-uniformity of the images due to random padding as well as improving model performance by removing dataset imbalance. Some of the images had both legs in the x-ray while most of the images had only one. Almost every image observed had random padding on at least one side which made the percentage area of the bones in the image vary significantly, which proved to be suboptimal in previous trials. The preprocessing steps that were taken can be summarized as follows.

i. Crop the black padding pixels from the top of the image
ii. Crop the black padding pixels from the bottom of the image
iii. Crop the black padding pixels from the left of the image
iv. Crop the black padding pixels from the right of the image
v. Check the middle vertical third of the image to see if it is dark (less than 50, 50, 50 in RGB value)
vi. If v. is true, split the image vertically in half
vii. Add 40% random noise to the image
viii. Clip the values of the image so they are within range
ix. Decrease the brightness of the image to 80%

These preprocessing steps were taken to mitigate the aforementioned issues of the dataset. They were essential to make sure that the dataset was as uniform as possible and to make sure the model does not train on irrelevant image features. These steps are shown visually in Figure 2. The green area shown in each step is the active area of processing that is done. The grey region is the inactive area of processing.

These images were pre-processed so as to split images into two different images if they contained two knees instead of one. This was achieved by evaluating if the average value of pixels in the middle third of the image when divided vertically is light or dark. Since these were x-ray images, there were black padding pixels in some images. This was removed by using an algorithm that would keep slicing one line of pixel from the image from each direction till the pixels in the next line is not entirely black.

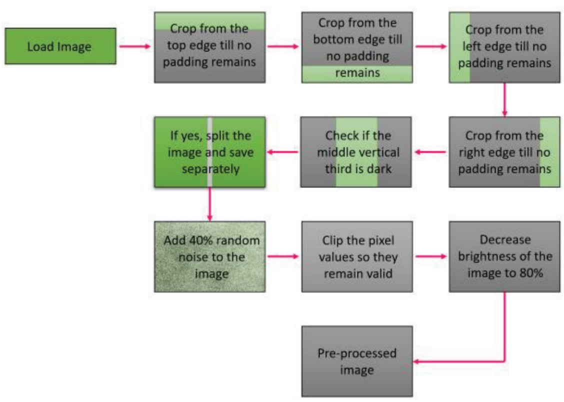

Figure 2 - Preprocessing steps used to process the dataset

The second dataset was not specifically pre-processed to do any augmentations as the dataset was of acceptable quality. Instead, the second dataset and the pre-processed version of the first dataset underwent a general preprocessing step where random noise was added while reducing the brightness of the image to 80%

3.3 Training Models

Using these datasets, three different versions of a transfer learning model were trained for each dataset. All models shared the same ResNet50 architecture as a common base, followed by a flatten layer, a dropout layer, a dense layer with 128 neurons with relu activation, another dropout layer, followed by the final output layer. This is visually represented in Figure 3 below. The base model shown is the ResNet50 model by Keras with its ImageNet weights loaded, but its layers were not frozen.

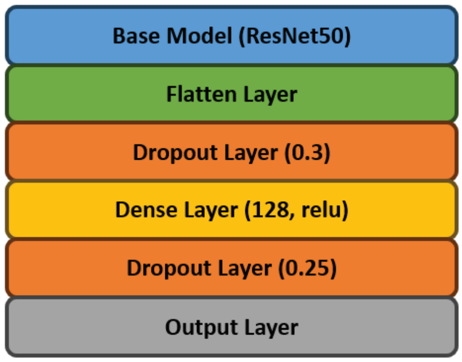

Figure 3 - Basic structure of the models used

The three different versions of models will be referred to as 'multiclass', 'binary' and 'zombie' models within the context of this paper. The multiclass model was made by using the dataset as is and training the above structured model using them. The classes were left as they were and since both datasets had three classes, the output layer also was given three classes. The binary model was made by grouping two of the classes of the respective dataset into one, so that the classes were of normal and abnormal. Appropriately, the output layer of the model was changed to manage two classes instead of three.

The zombie model was made by using the multiclass model itself as a base and creating what would essentially be a different kind of transfer learning model. Here, the original output layer of the trained multiclass model was removed and the rest of the layers were frozen. A new output layer of just two classes was added instead and the model was retrained. No other layers were added. This meant that the only weights being trained were the ones created by adding the new output layer.

3.4 Evaluating the Models Using SHAP

Using randomly selected images as samples, SHAP values were generated for each model to see how the image was used to arrive at a conclusion. Outputs of the final output layer were also compared to get a better idea about how confident the model was at classifying each of those images. The term osteo is used in this paper to refer to cases which belongs to either osteoporosis or osteopenia in instances where only normal and osteo class are present.

4. Results and Discussion

4.1 Training Models

The three models trained using these different ways showed varying accuracies, albeit in a narrow band about 95%. The accuracies were as expected due to the distribution of images within the dataset. Table 1 as well as Figures 4 and 5 show the classification report, model accuracy curve and model loss curve binary model.

Table 1 – Classification report of the binary model

	Precision	Recall	F1-score	Support
Normal	0.96	0.93	0.95	148
Osteo	0.95	0.97	0.96	185
Accuracy			0.95	333
Macro Avg	0.95	0.95	0.95	333
Weighted Avg	0.95	0.95	0.95	333

Accuracy of the model: 95.2%

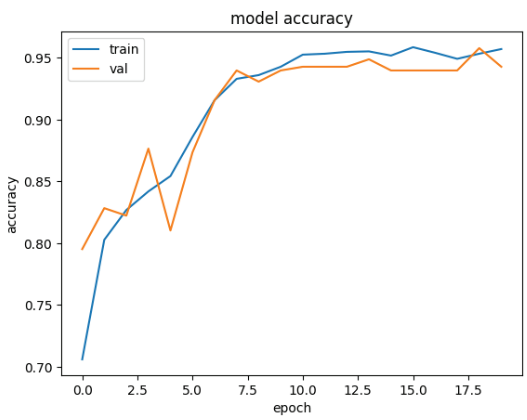

Figure 4 - Accuracy curve of the binary model

The three models trained using these different ways showed varying accuracies, albeit in a narrow region between 93% and 96.1%. According to the classification report of the binary model shown in Table 1, the binary model had achieved an accuracy of 95.2%.

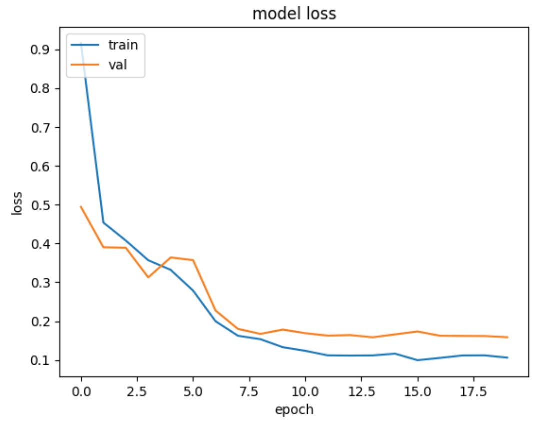

Figure 5 - Loss curve of the binary model

It had also achieved a precision of 95% for identifying Osteo cases and 96% for normal cases.

Table 2 – Classification report of the multiclass model

	Precision	Recall	F1-score	Support
Normal	0.95	0.93	0.95	409
Osteopenia	0.94	0.76	0.84	79
Osteoporosis	0.92	0.96	0.94	349
Accuracy			0.93	837
Macro Avg	0.93	0.89	0.91	837
Weighted Avg	0.93	0.93	0.93	837

Accuracy of the model: 93.4%

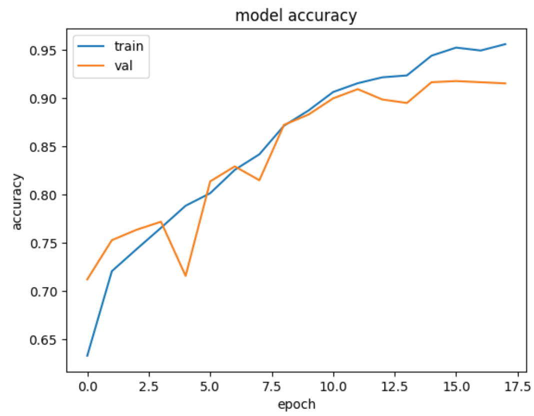

Figure 6 - Accuracy curve of the multiclass model

Table 2 as well as Figures 6 and 7 show the same for the multiclass model that was trained using the same dataset. As seen in the classification report shown in Table 2, it was observed that the multiclass model had achieved an overall accuracy of 93.4%.

It was able to differentiate among normal, osteopenia and osteoporosis cases with 95%, 94% and 92% accuracy.

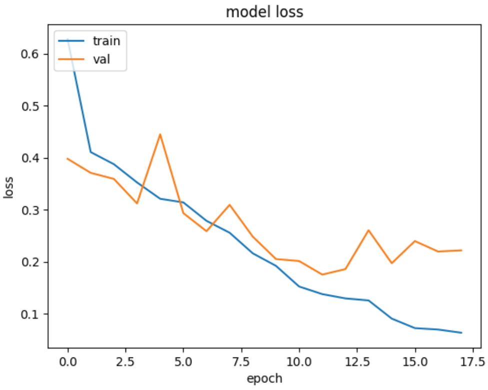

Figure 7 - Loss curve of the multiclass model

This model has achieved a higher accuracy than many of the other models in several other studies which also used the same dataset as the main dataset used in this paper, details of which are shown in Table 3 below.

Table 3 - Comparison of models discussed in previous work using knee x-ray imagery, trained for multiclass classification

Author and Year of Publication	*Accuracy in multiclass classification*	
	Classifier	*Accuracy*
Wani et al. [7]	AlexNet	91.0%
Kumar et al. [18]	ResNet18	91.4%
Abubakar et al. [19]	GoogleNet	90.0%
Yang [20]	Late-Fusion	82.0%
Sarhan et al. [21]	Vgg-19	92.0%
This study	ResNet50	93.4%

Table 4 as well as Figures 10 through 12 shows details related to the zombie model made by using the multiclass model as a base.

Table 4 – Classification report of the zombie model

	Precision	Recall	F1-score	Support
Normal	0.93	0.98	0.96	141
Osteo	0.98	0.95	0.97	192
Accuracy			0.96	333
Macro Avg	0.96	0.96	0.96	333
Weighted Avg	0.96	0.96	0.96	333

Accuracy of the model: 96.1%

Here, the model only had to train its binary classification output layer. Meaning these accuracy and loss curves can be assumed to be less significant compared to Figures 4 to 7 for the binary and multiclass models, since they were trained from the ground up on the dataset used, whereas this was used only to finetune the output layer.

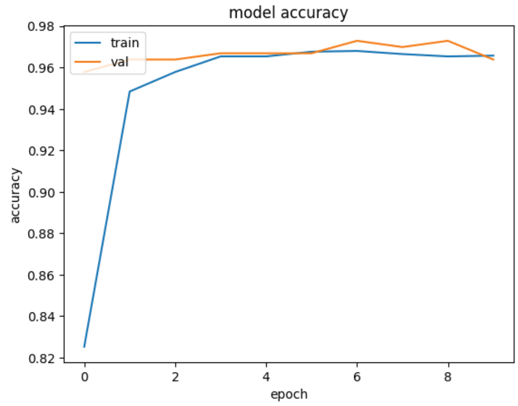

Figure 8 - Accuracy curve of the zombie model

The percentages shown in Table 5 are incorrect predictions done by the models, expressed as a fraction of the actual number of images in their respective classes in the dataset. According to the table, it can be seen that the Binary model had wrongly classified 6.76% of its validation data as negative, while the same for the zombie model is almost one third of this percentage, at 2.17%. Meanwhile, the percentages for false negatives had shown a different behaviour. The binary model had managed to give less false positives than the zombie model, at 3.24% compared to 5.50% of the zombie model. This is a 2.26% increase in false negatives, compared to a decrease in false positives by 4.49%.

Table 5 - Comparison of incorrect predictions of the models (Here FP – False positives and FN – False negatives)

Model	**Incorrect predictions by the model**			
	False positives	***False negatives***	***FP as %***	***FN as %***
Binary	10	6	6.76%	3.24%
Zombie	3	10	2.17%	5.50%

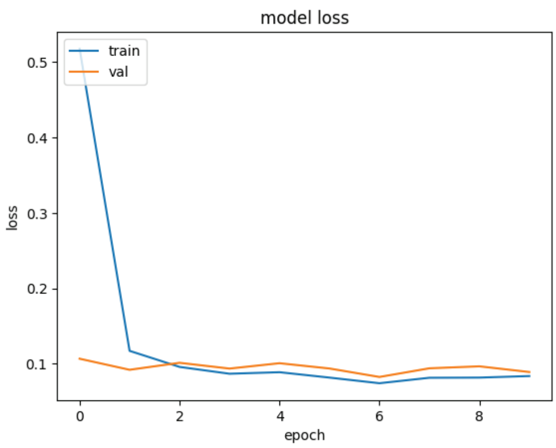

Figure 9 - Loss curve of the zombie model

Overall, from the above data, it can be seen that the Binary model performed worse compared to the zombie model, since the zombie model's accuracy (96.1%) was higher than that of the binary model (95.2%). Though the zombie model's precision towards normal images did decrease (93% compared to binary model's 96%), it had managed to avoid false positives more than the binary model did.

4.2 Training Models – Control Experiments

Since the models needed to be tested outside the main dataset, the same models were made for the second dataset as well and compared. Similar to the main dataset, three models were made for this dataset as well. Figures 10 and 11 as well as Table 6 show statistics of multiclass model trained using this dataset.

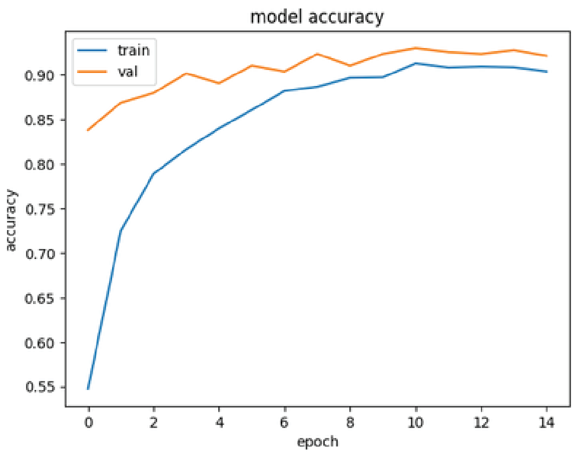

Figure 10 - Accuracy curve of the multiclass model

It can be seen here that the model performed to similar levels as the previous set in the multiclass case. Despite the normal class having less class accuracy at 85%, the overall accuracy of the model reached 93%. The curves are rather smooth and the loss curve never takes a drastic turn. Performance wise, for the purpose of this study, it can be said that the two multiclass models do perform relatively similar to each other in most cases.

Figures 12 and 13, and Table 7 show statistics of the binary model trained using the second dataset.

Table 6 – Classification report of the multiclass model

	Precision	Recall	F1-score	Support
Covid	0.97	0.93	0.95	139
Normal	0.85	0.99	0.91	79
Pneumonia	0.99	0.88	0.93	168
Accuracy			0.93	458
Macro Avg	0.94	0.93	0.93	458
Weighted Avg	0.94	0.93	0.93	358

Accuracy of the model: 93.0%

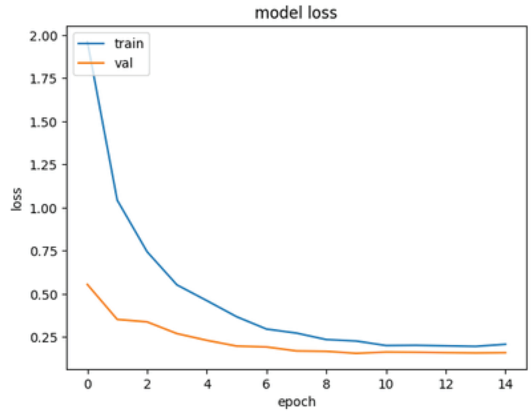

Figure 11 - Loss curve of the multiclass model

Table 7 – Classification report of the binary model

	Precision	Recall	F1-score	Support
Abnormal	0.97	0.90	0.93	307
Normal	0.82	0.95	0.88	151
Accuracy			0.91	458
Macro Avg	0.90	0.92	0.91	458
Weighted Avg	0.92	0.91	0.92	458

Accuracy of the model: 91.5%

As seen from the above figures, the model showed an average accuracy of 91.5%. In this instance as well, the normal class did show a much lower class-accuracy at 82% which is proving to be a repeating pattern within this specific dataset.

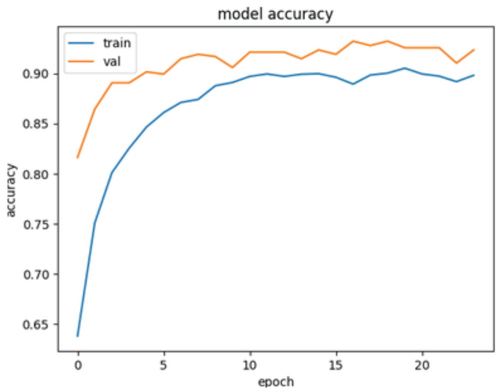

Figure 12- Accuracy curve of the binary model

While this has the potential to be of concern if this was a classification performance comparison, the objective of training these models is for them to be used later on as a comparative tool to evaluate the behaviours and performance of the models by utilizing explainable AI. This will be further discussed later.

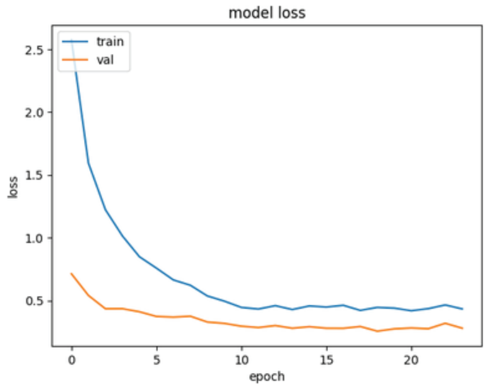

Figure 13 - Loss curve of the binary model

Figures 14 and 15 as well as Table 8 show statistics of the zombie model made using the same data as well as the multiclass model as base model.

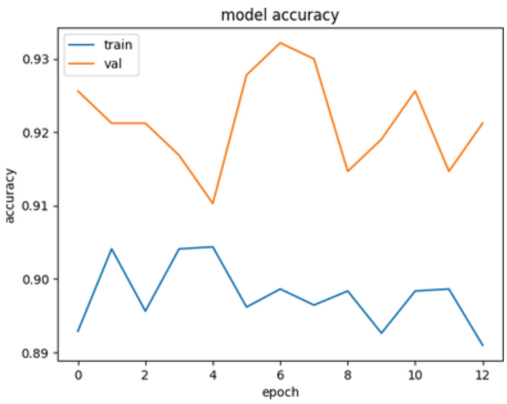

Figure 14 - Accuracy curve of the zombie model3

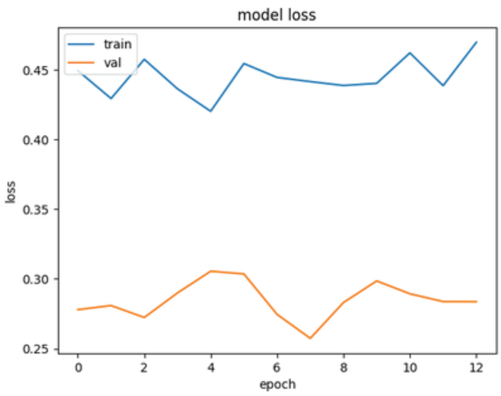

Figure 15 - Loss curve of the zombie model

According to the above figures as well as Table 9 below, it can be seen that the model had achieved an average accuracy of 92.6% despite the repeated pattern of the normal class having less class accuracy. While the variation of accuracies and losses in the zombie model can be seen as hectic, it should be noted that the variation is much exaggerated on the plot as the scale is more sensitive than in previous cases.

Table 8 – Classification report of the zombie model

	Precision	Recall	F1-score	Support
Abnormal	0.98	0.91	0.94	307
Normal	0.83	0.97	0.90	151
Accuracy			0.93	458
Macro Avg	0.91	0.94	0.92	458
Weighted Avg	0.93	0.93	0.93	458

Accuracy of the model: 92.6%

This is to be expected when the only weights that were trained in this step were the ones that were tied to the output layer, and that there were over 50 layers that were frozen that were already trained on the same dataset it is being trained on here.

Using the above data, as well as data obtained from confusion matrices of the models, the FP and FN of the binary and zombie models can be compared. This can help further solidify or disprove any assumptions made using the previous set of results shown in Table 5.

Table 9 - Comparison of incorrect predictions of the models for the second dataset

Model	**Incorrect predictions by the model**			
	False positives	***False negatives***	***FP as %***	***FN as %***
Binary	31	8	10.10%	5.30%
Zombie	29	5	9.45%	3.31%

The behaviour of the 2 models, as shown in Table 9, is not the exact same behaviour shown by the two models trained on the previous dataset. Instead of FN% increasing for the zombie model like in Table 5, it can be seen that there is a sharp decrease in this value. Hence it is safe to say that the findings from Table 5 related to FN are not a recurring phenomenon in the zombie models made.

4.3 Utilization of SHAP for Explainability

Integrating SHAP into the models gave better meaning to the results obtained before. Even though the models showed similar results as seen before, the shapley values for the models were vastly different. The shapley plot contains the Shapley values for each section of the image. The positive values shown in red/pink show sectors that contributed positively towards the prediction while the negative values shown in blue show sectors that contributed negatively towards the prediction.

The results obtained as shown in Figures 16 through 18 have a few key takeaways. The more obvious of which being that the zombie model had retained its ability to recognize features the same way the multiclass model had after being trained. This is proven by how identical the shapley plots for the multiclass model and the zombie model is. In essence, it can be said that the child model (the binary model) inherited this trait from its parent model (multiclass model). The sectors of the image that contributed towards the prediction show a drastic difference between the binary and the zombie model. The interpretation here is akin to the zombie model properly recognizing the bone and predicting whether the patient was healthy or if he had osteoporosis while the binary model focused more on artifacts in the image rather than the bone itself. This overall behaviour of the three models was a common feature observed throughout all the random samples chosen for generating shapley plots.

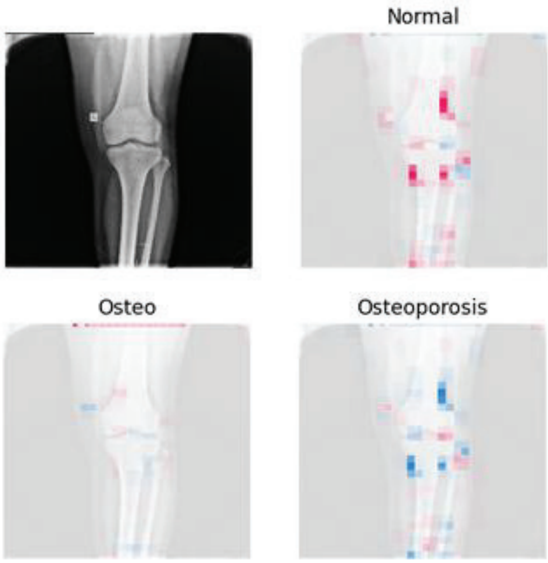

Figure 16 - Shapley plots obtained by implementing SHAP to the multiclass model using a normal image

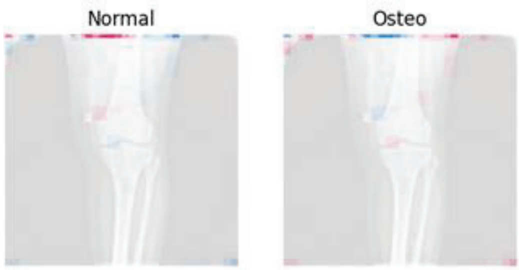

Figure 17 - Shapley plots obtained by implementing SHAP to the binary model using the same normal image

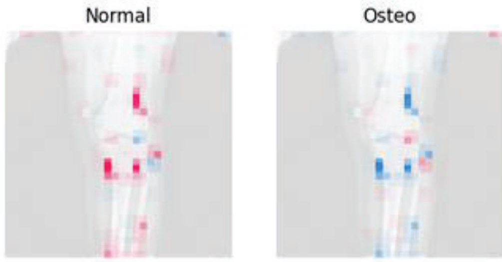

Figure 18 - Shapley plots obtained by implementing SHAP to the zombie model using the same normal image

The less obvious result here is that in the multiclass model, it can be seen that the model was able to classify if the image was normal or not initially. This can be seen by how the shapley plots for normal were the opposite of the addition of shapley plots for osteopenia and osteoporosis. The model was clearly able to predict that the patient was not suffering from osteoporosis as well. Yet it did struggle with deciding if the patient suffered from osteopenia or not which can be seen by the lack of any bright coloured areas in the shapley plot. This however was not the same result observed for osteopenia images. The multiclass only predicted this example to be an osteopenia image with an 88.3% confidence. However, this accuracy improves when the same image was given to the zombie model, which predicted this to be an osteo image with 98.4% confidence. Yet in both cases, the prediction does not agree with the shapley plots acquired for each case. Meanwhile, the binary model showed better explainability for the decision while showing a confidence of 97.5% for the image being an osteopenia image.

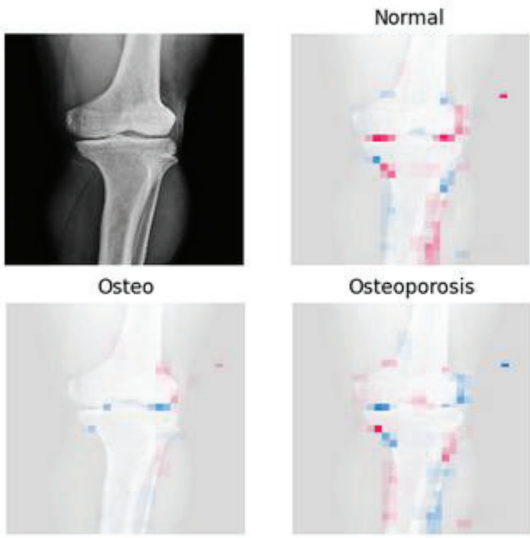

Figure 19 - Shapley plots obtained by implementing SHAP to the multiclass model using an osteopenia image

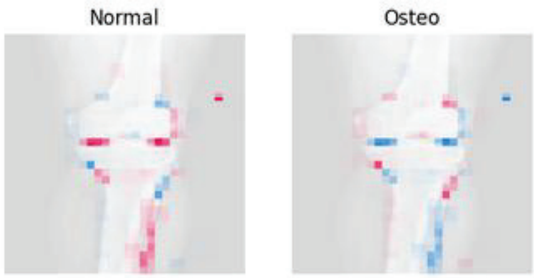

Figure 20 - Shapley plots obtained by implementing SHAP to the zombie model

Unlike the previous example, here it can be seen that the model is utilizing finer details of the bone to predict the class. This may be due to the slight imbalance of training data in each class in the dataset compared to the other two classes. This became more plausible when the same trend observed for normal cases returned for the osteoporosis class as well.

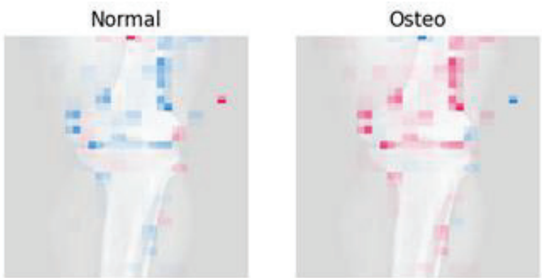

Figure 21 - Shapley values obtained by implementing SHAP to the binary model using the same osteopenia image

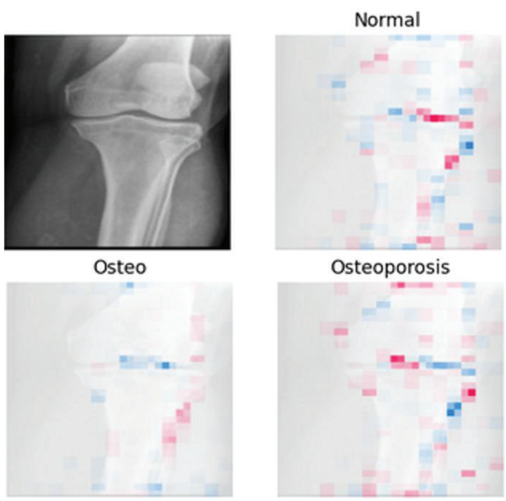

Figure 22 - Shapley plots obtained by implementing SHAP to the multiclass model using an osteoporosis image

Considering the presence of primarily same coloured regions within the SHAP plots in Figure 21, the model seems to be able to get stronger evidence for an image being as well as not being an osteopenia image in this case compared to both cases discussed before. The zombie model as usual behaves similar to the multiclass despite only classifying to classes instead of three.

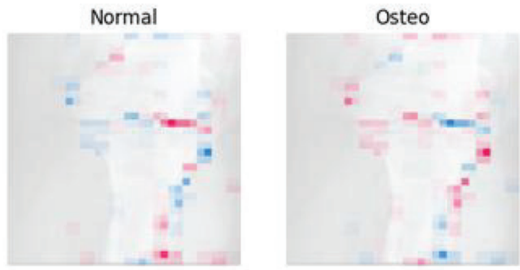

Figure 23 - Shapley plots obtained by implementing SHAP to the zombie model using an osteoporosis image

The trend of the multiclass model and the zombie model being more explainable was evident throughout all samples observed whereas the binary model lacked any explainability despite still holding a high accuracy.

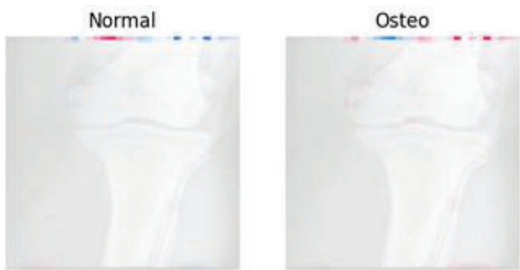

Figure 24 - Shapley plots obtained by implementing SHAP to the binary model using an Osteoporosis image

One possible reason why the shapley values for the binary model for osteoporosis and normal cases existing primarily as a line in the very top of the image could be the existence of a narrow region of high contrast. An example is shown in Figure 25 below.

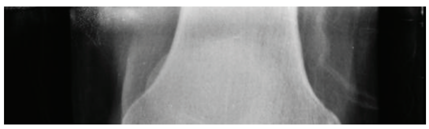

Figure 25 - High contrast narrow band at the top of the image

Some of the samples observed were bad due to the existence of labelling within the image which labelled if the leg was the left leg or the right leg.

Through Figures 26 to 28, it can be clearly seen how an artifact had affected the prediction irrespective of which model used. Yet it can also be seen that the prediction did not entirely depend on the visible artifact for the multiclass as well as zombie models.

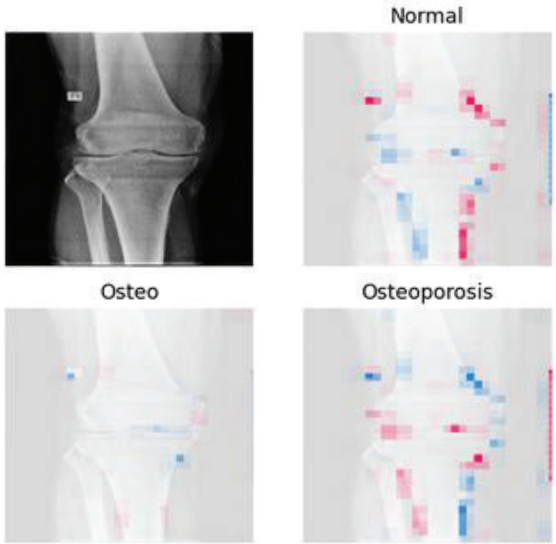

Figure 26 - A multiclass example for anomalies within the image due to visible and normally invisible artifacts

Despite the artifact affecting the prediction, there are more pixels that affected the decision that provided more contextual accuracy for the decision. It can also be seen that the zombie model showed better contextual accuracy, considering how the zombie model as well as the multiclass model had been able to avoid depending on the background pixels for its prediction.

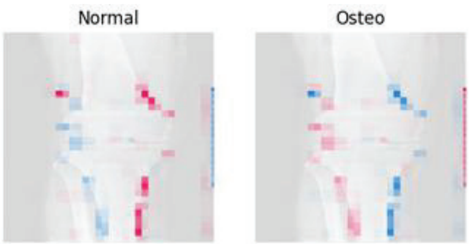

Figure 27 - Zombie model shapley plot for the same example used in Figure 26

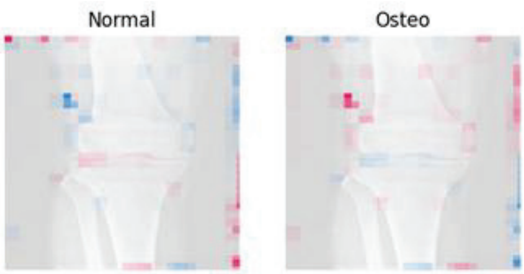

Figure 28 - Binary model shapley plot for the same example used in Figure 26

In this specific example, it can also be seen that there is another anomaly along the right edge of the image which cannot be normally seen, but had heavily affected the decision of every model. The reason for this seems unclear. It is possible that this is a result of scan lines appearing over the image. Yet, in this case as well, it is evident that the zombie model did not solely rely on it for a prediction whereas the binary model had been significantly influenced by it.

An interesting conundrum can be found here. ResNet50 and other pretrained models offered are often trained on large datasets with a large class size such as ImageNet. As shown in a paper by Dewt et al. [22] it can be seen that a model that can classify into many such classes has high contextual accuracy considering how the shapley values for each prediction look. Yet the binary model used in this paper was based off ResNet50, a similar model that has the ability to classify into many classes provided by the ImageNet dataset. It can be seen that this ability to accurately extract features from the images was lost during the process of training the model after adding extra layers. This can be explained by the base model not being frozen during training. The base model ResNet50 does possess the contextual accuracy for classifying images accurately. This ability is lost when it is trained for two classes. Yet it is retained to a certain degree when trained to differentiate among three classes of similar looking images instead. And this ability is carried over to the zombie model as well.

4.4 Utilization of SHAP for Explainability in the Control Experiment

The second set of models were trained specifically to be utilized here. This would be able to give a better insight as to why models behave the way they do. The following shows the SHAP plots obtained for the same image using the three different models made.

The behaviour observed in the below figures can be considered similar to the behaviour of the multiclass model of the osteoporosis dataset. Both, the multiclass model as well as the zombie model, have shown similar behaviour based on their shapley plots while the binary model has behaved drastically different to them. And similar to the previous case, it can be seen that the model underutilized available data to come to a conclusion. The accuracies of the zombie model (92.6%) is higher than the accuracy of the binary model (91.5%) and hence it can be safely assumed that the reason why the binary model only picked a part of the image for its decision was not due to contextual reasons. Underutilizing information found in an image to come to a worse prediction would not be indicative of anything else.

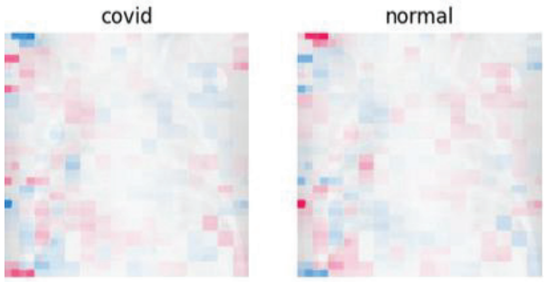

Figure 29 - Shapley plots obtained for a pneumonia image using the zombie model

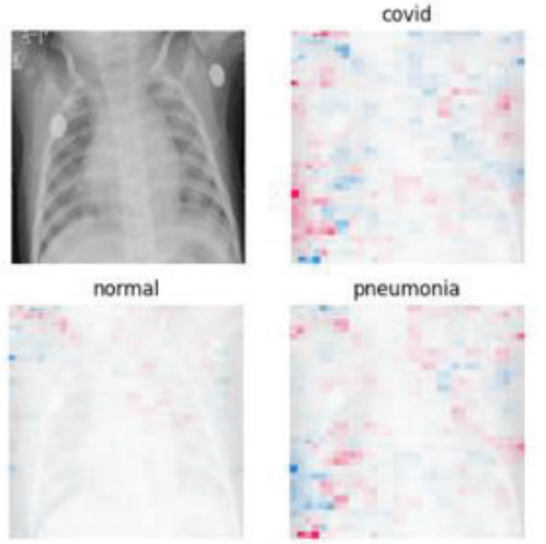

Figure 30 - Shapley plots obtained for a pneumonia image using the multiclass model

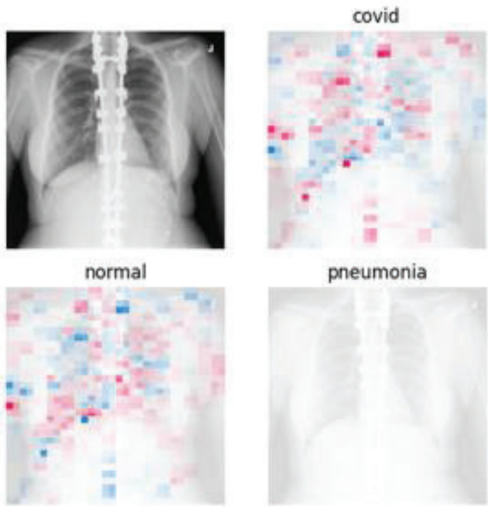

Figure 31 - Shapley plots obtained for a normal image using the multiclass model

Compared to the behaviour shown previously, it can be seen in Figure 31 through to 34 that, the binary model did utilize more of the data in the entire image instead. This in comparison is more normal behaviour of the model.

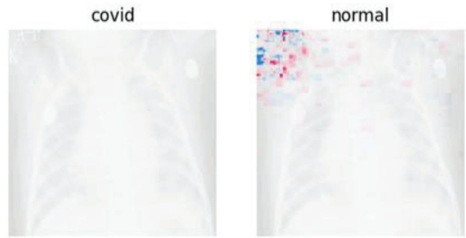

Figure 32 - Shapley plots obtained for a pneumonia image using the binary model

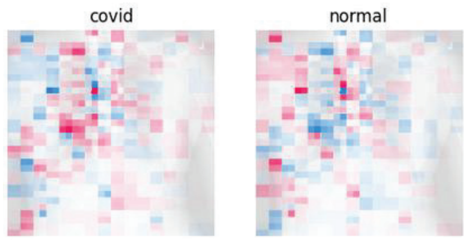

Figure 33 - Shapley plots obtained for a normal image using the zombie model

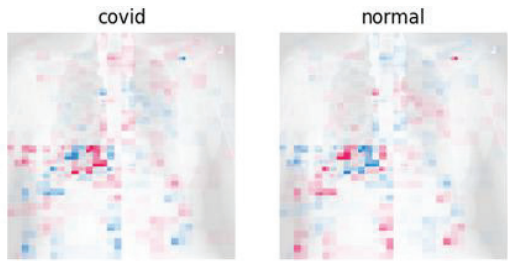

Figure 34 - Shapley plots obtained for a normal image using the binary model

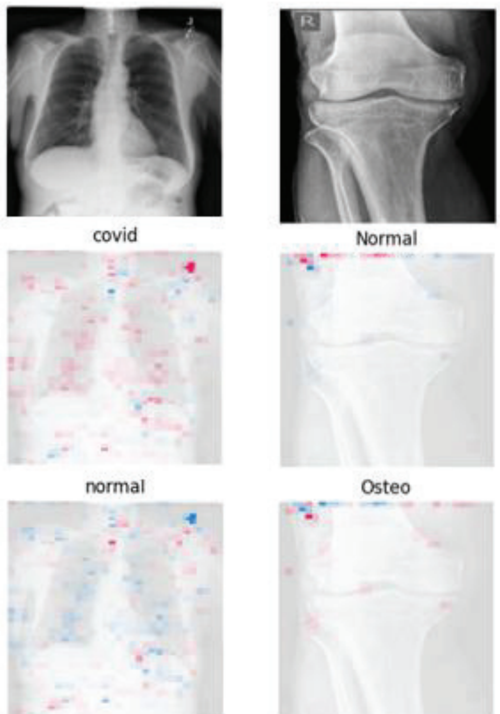

Figure 35 - Results obtained from the binary models using the 2 datasets

For comparison, consider the results obtained by the two zombie models using SHAP, shown in Figure 35. The behaviour of the model can be safely regarded as being heavily dependent on the dataset, since the behaviour is different despite being the same model structure, and the only difference being the dataset used to train it. The binary model underutilized image data, relying more on noise and unrelated artifacts for predictions. It also had lower accuracy than the zombie model, which performed better both statistically (as shown in classification reports) and contextually (as seen in SHAP plots).

5. Conclusion

Convolutional Neural Networks hold significant promise in medical applications, and has a high usage in early diagnosis and population level analysis of medical conditions. The trust towards these models and their predictions from the general public depends more on the trustworthiness of the public towards their healthcare systems as a whole. Ensuring the barrier between a model and the physician is as transparent as possible as well as interpretable by layman can help foster this trust between the two parties to a certain extent. Lack of such transparency would risk worsening existing scepticism towards the health sector as a whole.

For a model that is to be trained to detect a specific medical condition, it was shown throughout this study that utilizing class augmentation initially before training for the smaller class problem can be beneficial. This was shown by the comparison of shapley plots of the multiclass, binary and zombie models. It can be seen that, by doing so, the model was forced to learn clinically meaningful features. It was also seen that, doing so, significantly reduced the number of false positives of given by the zombie model, even though the effect on this on false negatives was not as clear. Yet the % of false positives did reduce in both cases and can be considered a perk of utilizing this method.

In the work done in this paper, one limitation observed was the lack of augmentable classes. Image augmentation involving changing pixel values was not used here since it can create a circumstance where there exist images in all classes that looked similar even to the model. Apart from the normal, osteopenia and osteoporosis classes, there were no other classes that can be created for this issue. In the context of Osteoporosis, this study would benefit highly from the existence of a dataset that ideally had classes based anatomical orientation as well. Instead of there being normal, Osteopenia and Osteoporosis, there would be normal-front/side, Osteopenia-front/side, and Osteoporosis-front/side as well. From the results obtained, it can be expected that if such a dataset was used to create a zombie model of two classes in the same way, that would have much higher contextual accuracy, and most probably, statistical accuracy as well.

References

1. Bone Health & Osteoporosis Foundation, “What is osteoporosis and what causes it?”, January 24, 2024. [Online]. Available: https://www.bonehealthandosteoporosis.org/patients/what-is-osteoporosis/

2. Dimai, H. P., “Use of dual-energy X-ray absorptiometry (DXA) for diagnosis and fracture risk assessment; WHO-criteria, T-and Z-score, and reference databases”, *Bone*, Vol. 104, 2017, pp. 39-43.

3. Radiological Society of North America (RSNA) and American College of Radiology (ACR), “Bone Densitometry (Dexa, DXA)”, [Online]. Available: https://www.radiologyinfo.org/en/info/dexa

4. GeeksforGeeks, “Applications of deep learning in Healthcare”, June 6, 2024. [Online]. Available: https://www.geeksforgeeks.org/applications-of-deep-learning-in-healthcare/

5. Dreessen, O., “Introduction to Convolutional Neural Networks: What is Machine Learning? - Part 1”, February 2023. [Online]. Available: https://www.analog.com/en/resources/analog-dialogue/articles/max78000-article-series-part-1.html

6. He, K., Zhang, X., Ren, S., & Sun, J., “Deep Residual Learning For Image Recognition”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

7. Lundberg, S., “A Unified Approach To Interpreting Model Predictions”, arXiv preprint arXiv:1705.07874, 2017.

8. Wani, I. M., & Arora, S., “Osteoporosis Diagnosis In Knee X-Rays By Transfer Learning Based On Convolution Neural Network”, *Multimedia Tools and Applications*, Vol. 82, No. 9, 2023, pp. 14193-14217.

9. Derkatch, S, Kirby, C., Kimelman, D., Jozani, M.J., Davidson, M., and Leslie, W.D., "Identification Of Vertebral Fractures By Convolutional Neural Networks To Predict Nonvertebral And Hip Fractures: A Registry-Based Cohort Study Of Dual X-Ray Absorptiometry." *Radiology 293*, no. 2 (2019): 405-411.

10. Krishnaraj, A., Barrett, S., Bregman-Amitai, O., Cohen-Sfady, M., Bar, A., Chettrit, D., Orlovsky, M. and Elnekave, E., 2019. "Simulating Dual-Energy X-Ray Absorptiometry In CT Using Deep-Learning Segmentation Cascade"., *Journal of the American College of Radiology*, 16(10), pp.1473-1479.

11. Lee, S., Choe, E.K., Kang, H.Y., Yoon, J.W. and Kim, H.S., 2020. "The Exploration Of Feature Extraction And Machine Learning For Predicting Bone Density From Simple Spine X-Ray Images In A Korean Population". *Skeletal radiology*, 49, pp.613-618.

12. Zhu, F., Zhang, X.Y., Wang, R.Q. and Liu, C.L., 2022. "Learning By Seeing More Classes". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6), pp.7477-7493.

13. Hu, Y.L., Wang, P.Y., Xie, Z.Y., Ren, G.R., Zhang, C., Ji, H.Y., Xie, X.H., Zhuang, S.Y. and Wu, X.T., 2023. "Interpretable Machine Learning Model To Predict Bone Cement Leakage In Percutaneous Vertebral Augmentation For Osteoporotic Vertebral Compression Fracture Based On Shapley Additive Explanations" *Global Spine Journal,* p.21925682231204159.

14. Khanna, V.V., Chadaga, K., Sampathila, N., Chadaga, R., Prabhu, S., Swathi, K.S., Jagdale, A.S. and Bhat, D., 2023. "A Decision Support System For Osteoporosis Risk Prediction Using Machine Learning And Explainable Artificial Intelligence". *Heliyon, 9* (12).

15. Liu, M., Wei, X., Xing, X., Cheng, Y., Ma, Z., Ren, J., Gao, X. and Xu, A., 2024. "Predicting Fracture Risk For Elderly Osteoporosis Patients By Hybrid Machine Learning Model". *Digital Health*, 10, p.20552076241257456.

16. Gobara, M., "Multi-Class Knee Osteoporosis X-Ray Dataset", *Kaggle*, September 14, 2024. [Online]. Available: https://www.kaggle.com/datasets/mohamedgobara/multi-class-knee-osteoporosis-x-ray-dataset

17. Asraf, A., "Covid19, Pneumonia And Normal Chest X-Ray PA Dataset", *Mendeley Data*, April 9, 2021.

18. Kumar, S., Goswami, P., & Batra, S., "Fuzzy Rank-Based Ensemble Model for Accurate Diagnosis of Osteoporosis in Knee Radiographs", *International Journal of Advanced Computer Science and Applications*, Vol. 14, No. 4, 2023.

19. Abubakar, U. B., Boukar, M. M., Adeshina, S., & Dane, S., "Transfer Learning Model Training Time Comparison for Osteoporosis Classification on Knee Radiograph of RGB and Grayscale Images", *WSEAS Transactions on Electronics*, Vol. 13, 2022, pp. 45-51.

20. Yang, T. S., "Recognition And Classification Of Knee Osteoporosis And Osteoarthritis Severity Using Deep Learning Techniques", PhD dissertation, National College of Ireland, Dublin, 2022.

21. Sarhan, A. M., Gobara, M., Yasser, S., Elsayed, Z., Sherif, G., Moataz, N., Yasir, Y., Moustafa, E., Ibrahim, S., & Ali, H. A., "Knee Osteoporosis Diagnosis Based On Deep Learning", *International Journal of Computational Intelligence Systems,* Vol. 17, No. 1, 2024, p. 241.

22. Dewt, C., Chen, R. C., Yu, H., & Jiang, X., "XAI For Image Captioning Using SHAP", *Journal of Information Science & Engineering,* Vol. 39, No. 4, 2023.