Student name - Ishan Jakkulwar

School name - Queen Elizabeth’s School Barnet

Teacher mentor name - Michael Noonan

Osteoporosis Multi-Regional Classification

*A Novel Multifocal Transfer Learning Neural Network Model For Proactive, Specific, Cheaper, And More Efficient Detection Of Osteoporosis Via Radiographs*

**Table of Contents**  Page Number

Introduction *4*

*(hypothesis, variables, background on topic, rationale,*

*what I hope to learn from this research)*

Materials and Methods *6*

*(what I did and what I used)*

Results *10*

*(data and findings)*

Discussion *11*

*(a discussion of my results, what they mean*

*and how the data relates to my hypothesis)*

Conclusions *12*

(*relate the discussion to the hypothesis, and check*

*whether or not it was supported, mention future work)*

References *14*

*(All sources used)*

Introduction

This is a novel, multifocal transfer learning neural network model for proactive, specific, cheaper, and more efficient detection of osteoporosis via radiographs. I started the project with the focus of developing a short and long-term solution to the late, costly detection of osteoporosis, derived from issues seen in the rationale(project concept)

It involves creating four machine learning models on Google Colab, which are trained on a custom dataset composed of knee and spine X-rays. This allows the models to predict whether or not the uploaded X-ray has osteoporosis or a prominent case of osteopenia.

My fourth and best model, which used the efficient net b-zero transfer learning model to optimise accuracy, precision, recall, and f1 score, achieved a test accuracy-87.5%, Recall-80%, and precision around the 70-75% mark, with the potential of being integrated into hospitals and reducing socio-economic burdens of osteoporosis-related fractures worldwide.

Background

Osteoporosis is a systemic skeletal disorder characterised by low bone mass and deterioration of bone tissue, which increases susceptibility to fractures. Millions worldwide have been diagnosed with osteoporosis, which leads to a lack of mobility, chronic pain, and diminished quality of life.

I was inspired to create this project when a relative of mine who was living in India recently got diagnosed with osteoporosis. He stressed how the lateness of the diagnosis paired with the cost caused him to experience a severe fracture, which would be permanently damaged if only slightly.

Bone health assessments are vital for diagnosing osteoporosis. The gold standard, DEXA (dual-energy x-ray absorptiometry) scans, is currently being used to identify and diagnose bone disease, and it does this pretty effectively.

However, DEXA machinery is very expensive and specialised and thus generally unavailable in low-resource settings(in terms of the United Kingdom it will cost the NHS more than double the price of an X-ray), preventing the diagnosis from being timely. DEXA is also only used after a fall or serious injury has occurred which causes bone to shatter.

Therefore, early intervention and detection are critical for the prevention of osteoporosis. In addition, globally, the cost of living crisis has already driven many into poverty, so finding a cheaper alternative could alleviate the stress put on them. One of the options I came to was using X-rays, which are already common practise, a lot cheaper, and much more known about. I concluded that picking up patterns using machine learning was the best option, as it was faster, cheaper, and more efficient.

By using a machine learning algorithm to identify x-rays with osteoporosis in different regions(the spine and the knee have the most risk of osteoporotic lethal fractures, and are also the most common places where osteoporosis occurs from lifestyle factors and ageing), there will be a significant improvement in detection of osteoporosis and early prevention.

Bone Density Calculations + Osteoporosis Classification Parameters

It does so by analysing the bone density of the patient and comparing it to an average bone density for their age and sex, which it then uses to generate a T-score value. Below are equation(s):

T score- How an individual’s Bone density(density of bones) differs from that of their age and sex



Normal: T-score ≥ -1

Osteopenia (Low Bone Mass): T-score between -1 and -2.5

Osteoporosis: T-score ≤ -2.5

Hypotheses

Primary Endpoint: The primary endpoint will be the improvement in classification accuracy of osteoporosis stages compared to traditional diagnostic methods, measured through sensitivity and specificity rates.

Secondary Endpoint: The secondary endpoint will evaluate the tool's effectiveness in reducing time to diagnosis and treatment initiation, along with assessing patient outcomes related to fracture risk and overall health improvement over time.

Claim: A deep learning-based tool will be created that successfully improves multiclass osteoporosis classification, facilitates early identification, streamlines treatment procedures, and enhances the management of osteoporosis-related diseases, thereby positively impacting the overall medical field, so that patients diagnosed with osteoporosis can take bisphosphonates as soon as possible

Method

In developing the osteoporosis detection model, I faced several challenges, primarily due to the limited dataset. Through testing with different Neural Networks, even creating my own CNN(which had a low accuracy compared to the final model), I managed to address this problem by utilising transfer learning with EfficientNet-B0, which leveraged pre-trained knowledge to improve performance despite the small data size. In addition, I applied data augmentation techniques such as rotation, scaling, and brightness adjustments which I used to artificially expand the dataset to ensure better generalisation. Data preparation and cleaning were crucial in eliminating inconsistencies as I found throughout my iterative process, and I balanced the dataset through oversampling and random sampling to prevent class imbalances that occurred.

I spent around one month on each stage and spent six months on data preparation.

The FFive-StepProcess for Machine Learning Projects(created by me through iterative trial and error to find the most efficient path for machine learning projects):

Data Preparation-

Data acquisition involved gathering and cleaning X-ray images from multiple sources to ensure diversity in image quality, as well as variation in patient demographics (age, sex). Data cleaning included correcting image orientation, resolution normalisation, and quality cheques to standardise images across sources.​

Feature Extraction-

Annotated images of osteoporotic fractures were referenced during feature exploration to identify key diagnostic markers, such as bone density and structural integrity variations, which informed me of the selection of features the model would focus on during training, and further my knowledge in the field.​

Data Augmentation-

Data augmentation was performed to simulate natural variations in medical imaging. Techniques included random rotations, flips, scaling, and brightness adjustments. This helps my model generalise better to real-world X-rays by the variations in image perspective and intensity that might occur in clinical practise and widens the limited dataset.​

Model Training-

Model training involved producing a convolutional neural network (CNN) architecture, optimised through transfer learning techniques(in which efficient net b0 was decided on, due to its balance between speed and accuracy). Training hyperparameters (e.g., epochs/learning rate, batch size) were used to reduce over and under-fitting.​

Model Testing + Tuning

The trained model was tested on a held-out dataset of X-ray images to measure its effectiveness in identifying osteoporotic changes. Metrics such as accuracy, recall, precision, and F1-score were recorded to assess diagnostic reliability. The model’s predictions were compared against true labels to gauge sensitivity and specificity, ensuring clinical relevance.​

**Materials**

Type of data: Qualitative

Type of data: Secondary research

Language: Python

Types of sampling:

Stratification: Allows my training, validation, and test sets to correctly reflect each class and region. It helps in keeping the datasets balanced so that the model learns from all different categories without overfitting into one particular class.

Random Sampling: Randomised sampling can be applied to randomly select images in each class and region for training, validation, and testing splits. Randomised sampling has the least selection bias, ensuring that no subset of data keeps showing up in one part of the training pipeline.

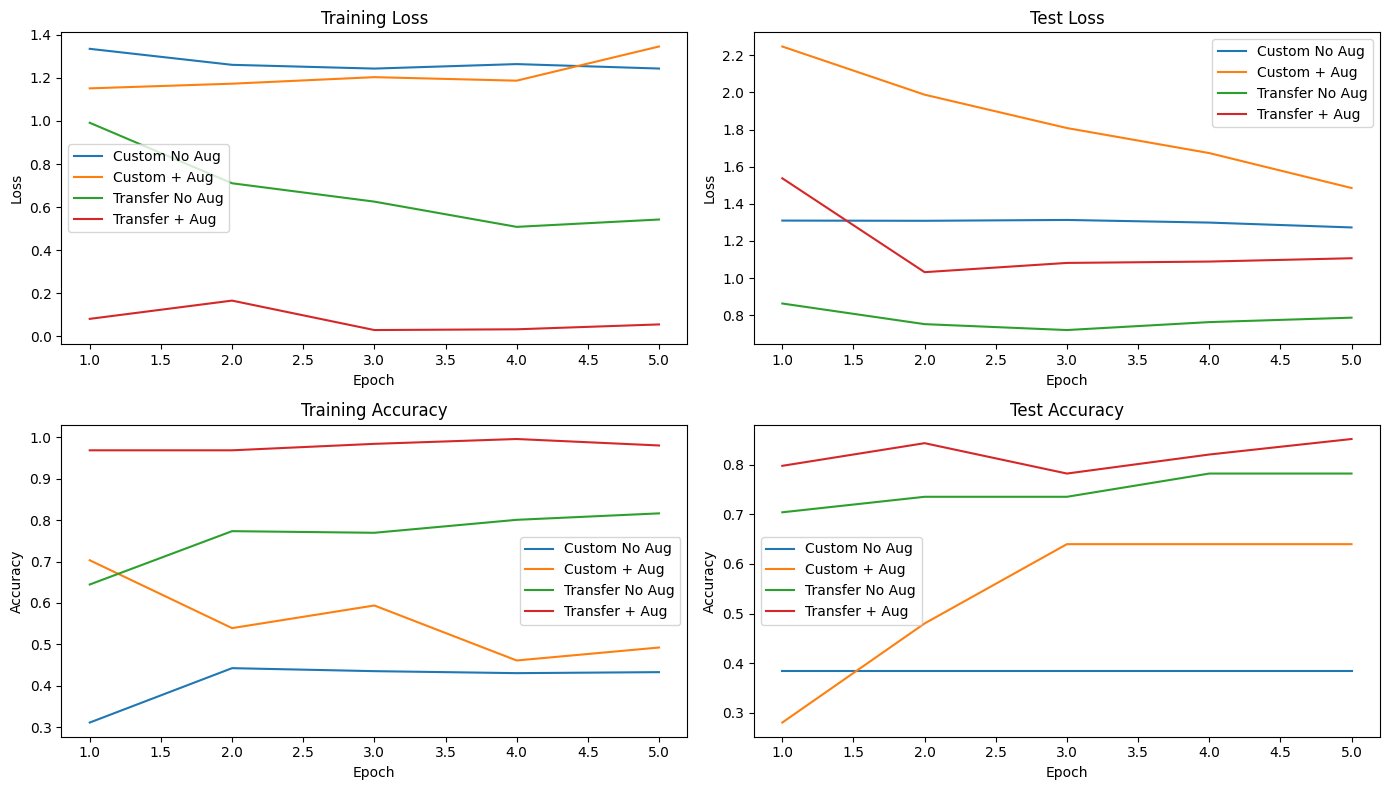
Oversampling for data augmentation: If any class imbalance exists-for example, more normal images than osteoporotic ones oversampling, using various data augmentation techniques, synthesises more variations to present the model with underrepresented classes. This might be useful to balance the data and improve the feature learning capability of the model from its minority classes.

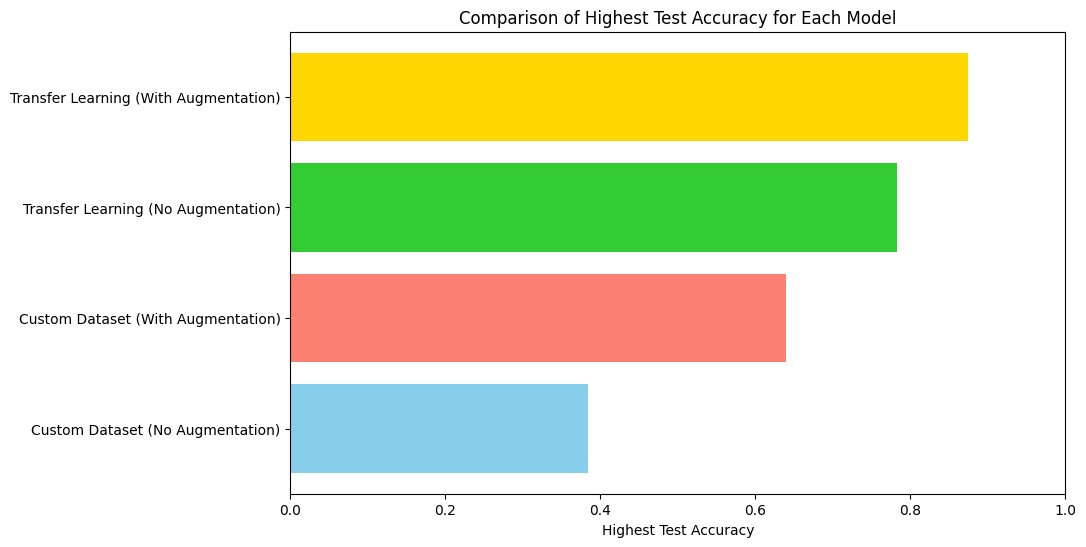
Cross-Validation Sampling: Though cross-validation is used less in deep learning due to computational costs, for smaller datasets, this technique-for example, k-fold cross-validation-provides assurance that the model generalises well. A common process is to divide the data into several "folds" and rotate through the various folds, using different folds as validation sets while the remainder is used for training.

Criteria for sampling: Data will be sampled from other people’s research articles and public datasets such as Kaggle

| Data Sources | [Normal Spine RadioGraph](https://radiopaedia.org/search?scope=all&commit=Search&q=normal+spine+radiograph)  [Normal, Osteopenic, Osteoporotic Knee Xray](https://link.springer.com/article/10.1007/s44196-024-00615-4#:~:text=In%20the%20context%20of%20knee,of%20knee%2Drelated%20bone%20disorders)  [Overview of Osteoporosis](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5335887/)  [Previous research papers on spine](https://www.researchgate.net/publication/343179468_Deep_learning_of_lumbar_spine_X-ray_for_osteopenia_and_osteoporosis_screening#:~:text=The%20results%20showed%20that%20in,95%25%20CI%3A%2067.3%2D91.8)  [Imaging of osteoporotic fractures](https://link.springer.com/article/10.1007/s40134-013-0032-x)  [Lit Reviews](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11044779/)  [Spine Osteoporotic X-ray](https://huggingface.co/datasets/TrainingDataPro/spine-x-ray?row=35) |
| --- | --- |
| Software and Environment | [Google Collab](https://colab.research.google.com/)  Python |
| Python Libraries and Frameworks | [Pytorch](https://pytorch.org)-  [OpenCV](https://opencv.org/)-For additional image processing and transformations (e.g., resizing, cropping).  Argumentation [collab](https://colab.research.google.com/github/fastai/fastai/blob/master/nbs/10b_tutorial.albumentations.ipynb#scrollTo=_I4Gqd-QHKv_)- For data augmentation techniques to improve model generalization  [NumPY](https://numpy.org/)- For numerical operations and matrix manipulation.  [Matplotlib](https://matplotlib.org/)- For visualizing images, data distribution, and model performance. |
| Evaluation Metrics and Tools | Accuracy, Precision, Recall, F1-score- Basic classification metrics to evaluate model performance. Accuracy may not be used as much to reduce overfitting.  Confusion Matrix-To understand misclassifications among classes.  ROC-AUC (Receiver Operating Characteristic - Area Under Curve): Useful for multiclass classification problems to evaluate model discrimination capabilities. |
| Data Augmentation Techniques | Rotation, Scaling, Flipping, and Brightness Adjustments: To enhance data variability.  Synthetic Data Generation: To create new samples |
| Documentation and Reporting Tools | Google Docs: For writing project reports, abstracts, and tracking findings. |
| Machine Learning Models | Custom CNN Architectures: to explore a customized CNN structure tailored to my X-ray data. |

**Results**

****

****

*Highest Recall - 0.800*

*Highest F1 Score - 0.719*

*Highest Precision - 0.700*

**Discussions**

We employed the following metrics to assess model performance, because of their significance within the medical imaging industry already:​ ​

*Accuracy*: Measures the proportion of correct predictions over the total predictions.​ ​

*Precision*: Defined as the ratio of true positive predictions to the total positive predictions (true positives + false positives). ​ ​

*Recall*: Measures the model's ability to correctly identify all true positives.

*F1-Score*: The harmonic mean of precision and recall. ​

To evaluate the performance of my osteoporosis detection model, statistical analyses were

conducted to assess its predictive capabilities. These analyses included accuracy, precision, recall, F1-score, and overall training trends. Each metric provides insights into the model's effectiveness and robustness, particularly in a medical application where both sensitivity and specificity are critical.​

Accuracy - achieved a maximum test accuracy of 87.15%, indicating reliable overall classification performance.​

Precision - A precision of 0.70 suggests that approximately 66.4% of the positive predictions were correctly classified as osteoporotic.​

Recall -. Recall of 0.800 highlights the model’s success in identifying patients with osteoporosis, aligning with the priority of minimising false negatives in medical diagnostics.​

F1 Score - With a value of 0.70, the F1 score reflects the balance between precision and recall, validating the model’s effectiveness in a binary classification setting.​

Training Loss and Accuracy-Training loss and accuracy trends were analysed across 5 epochs( each epoch split in half so technically 10). The results showed a steady improvement:​ Training Loss decreased consistently, reflecting the model's ability to minimise errors during optimisation. Training Accuracy increased from 95+% in Epoch 1 to around 100 % in Epoch 5, showing the model’s ability to generalise within the training dataset.​

Test Loss and Accuracy - Test loss exhibited a gradual decline, while test accuracy improved:​ From epochs 1-5 the test accuracy shifted between ~78 % to ~87% in the final epochs. This suggests the model avoided significant overfitting, although there may still be room for optimisation.​

Confusion Matrix Analysis​ A confusion matrix was used to evaluate prediction performance:​ True Positives (TP): Cases correctly identified as osteoporotic.​ True Negatives (TN): Normal cases are accurately classified.​ False Positives (FP): Normal cases misclassified as osteoporotic.​

False Negatives (FN): Osteoporotic cases missed by the model. The relatively low false positive (FP) rate, combined with a high true positive (TP) rate, aligns with the model’s high recall, confirming its effectiveness in identifying at-risk individuals.​ ​

**Conclusions**

The solution I have found to the problem of late, costly, and inefficient detection of osteoporosis offers a clinically viable alternative to conventional diagnostic methods and it addresses key challenges such as early detection, accuracy, and accessibility. The findings highlight several critical contributions:

Revisiting Engineering Criteria​ Model Accuracy:

The proposed transfer learning-based model achieved a high classification accuracy of ~87%, accompanied by a precision of 0.700, recall of 0.800, and an F1 score of 0.719. These metrics indicate how the model was quite efficient, and one that with tuning and added data could be integrated into hospitals.

Generalisation-The machine learning model displayed impressive results on the limited data and showed minimal over and underfitting when run.

Efficiency- With a total training time of 511 seconds, the approach is time efficient with fast results, although it can be improved with a higher or premium CPU on Google Colab.

Bench Impacts​-This study serves as a proof-of-concept for leveraging CNNs to analyse medical images for osteoporosis diagnosis.​ The architecture can be extended to other applications, such as identifying fractures, sub-classifying bone diseases, or improving anomaly detection in X-ray images.​ Incorporating transfer learning minimises the need for large annotated datasets, making the solution adaptable to underrepresented populations or rare diseases.​

Bedside Impacts-This diagnostic tool provides clinicians with an accessible and scalable decision-support system, enabling early intervention and treatment planning.​ The solution's rapid processing speed facilitates point-of-care use, empowering healthcare providers to screen patients quickly and accurately.​ By promoting early detection, this method could significantly reduce the socio-economic burden of osteoporosis-related fractures.​

Future Work

The concept uses for the technology created:

​ -The future use of the deep learning-driven X-ray analysis system with CNNs to detect and classify osteoporosis in knee and spine X-rays, will be far superior compared with the conventional diagnosis methods. ​

-It will achieve more homogenous and sensitive bone density loss, improving fracture risk prediction, enabling timely interventions, and lessening the growing healthcare costs for the management of osteoporosis.

-It will become a standardised tool in hospitals​ and other medical facilities​

Future work for technology​

Model Refinement: Investigate advanced architectures or techniques (like ensemble learning or higher-level transfer learning models) to enhance model performance.​

Data Diversity: Collect a more diverse(and larger) data set for increased evaluation metrics(which I want to maximise). For example, the range of ages and sexes

Integration with Clinical Workflows: Develop strategies for integrating the tool into existing healthcare systems to streamline workflows.​ ​

**References**

**Data, T. (2024). spine-x-ray. [online] Huggingface.co. Available at:**

[**https://huggingface.co/datasets/TrainingDataPro/spine-x-ray?row=35​**](https://huggingface.co/datasets/TrainingDataPro/spine-x-ray?row=35%E2%80%8B)

**‌Radiopaedia (2016). normal spine radiograph | Search | Radiopaedia.org. [online] Available at:** [**https://radiopaedia.org/search?scope=all&commit=Search&q=normal+spine+radiograph**](https://radiopaedia.org/search?scope=all&commit=Search&q=normal+spine+radiograph)**.**

**​ Gobara, M. (2024). Knee Osteoporosis Dataset multiclass. Kaggle.com. [online] doi:​** [**https://doi.org/10.1007/s44196-024-00615-4**](https://doi.org/10.1007/s44196-024-00615-4)**)\*\*.​**

**Sozen, T., Ozisik, L. and Calik Basaran, N. (2019). An overview and management of osteoporosis. European Journal of Rheumatology, [online] 4(1), pp.46–56. doi:​ https://doi.org/10.5152/eurjrheum.2016.048​** [**www.radiologymasterclass.co.uk**](http://www.radiologymasterclass.co.uk)**. (n.d.).**

**Imaging of Musculoskeletal Disorders - Osteoporosis. [online] Available at:** [**https://www.radiologymasterclass.co.uk/tutorials/musculoskeletal/imaging-joints-bones/osteoporosis\_x-ray#top\_3rd\_img**](https://www.radiologymasterclass.co.uk/tutorials/musculoskeletal/imaging-joints-bones/osteoporosis_x-ray#top_3rd_img)