Ostfalia
Hochschule für angewandte Wissenschaften

Faculty of Computer Science

On Deep-learning-based osteoporotic vertebral fracture prediction and risk assessment in CT Images

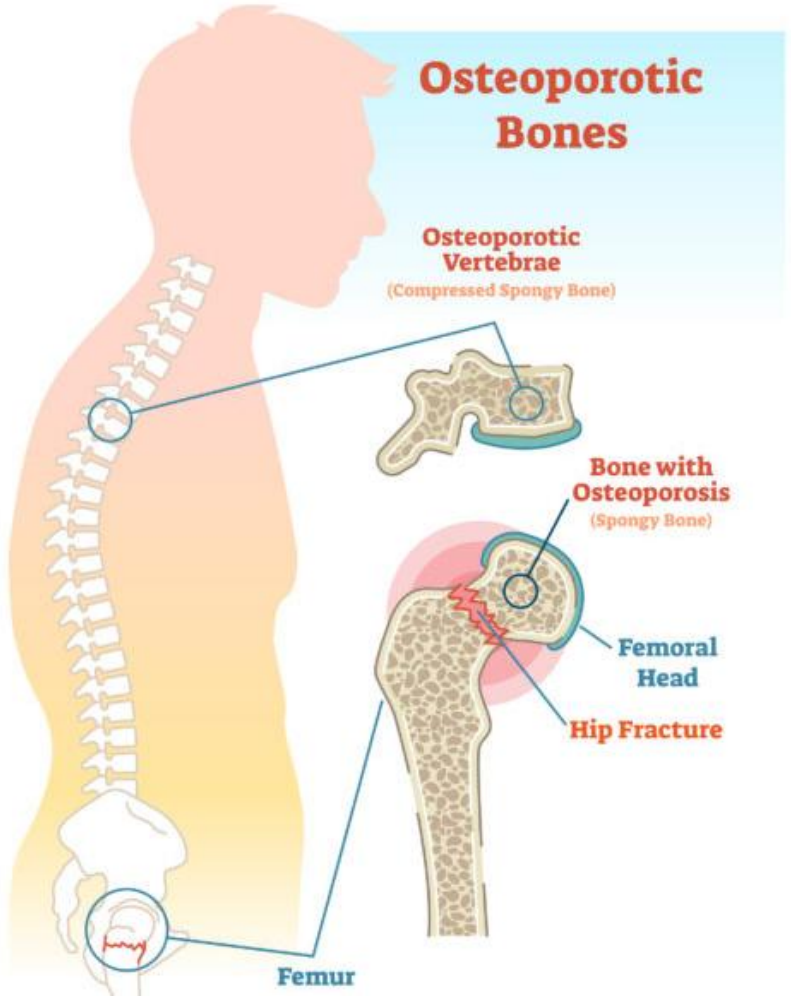
Shaikh Mohd Faraz, Prof. Carsten Meyer

Contents

- Motivation
- Methodology
- Inference pipeline
- Model Training & Results
- Self supervised learning (SSL) based 3D pretraining
- Finetuning 2D pretrained models
- Finetuning a CT based large vision language model
- Summary
- Discussions

Motivation

- Osteoporosis related fractures impact quality of life
  - Effects bone especially of spine and hip leading to fracture
  - 1 in 5 deaths in men within 6 months of fracture .. [1]
- Men are severely underdiagnosed prior to the fracture
  - Spine fractures are common in men with above 65 age ..[1, 2]
- Gold standard methods
  - DXA: measures BMD, but needs dedicated scan equipment which are not widely available in many clinics
  - FRAX: High specificity, but missed substantial patients who developed major fractures in 10 year follow up.. [3]
- CT scans of vertebra could detect high risk people in opportunistic setting



References:
[1] Prasad, D., Nguyen, M.H.: Chronic hepatitis, osteoporosis, and men: underrecognised and underdiagnosed. The Lancet Diabetes & Endocrinology 9(3), 14
[2] Robert A. Adler, Update on osteoporosis in men. Best Practice & Research Clinical Endocrinology & Metabolism. 2(5), (2018)
[3] Jiang et. al., Diagnostic accuracy of FRAX in predicting the 10-year risk of osteoporotic fractures using the USA treatment thresholds

Image reference:
[1] Compo ortho, https://comportho.com/anti-aging/health-tip-risk-factors-for-male-osteoporosis

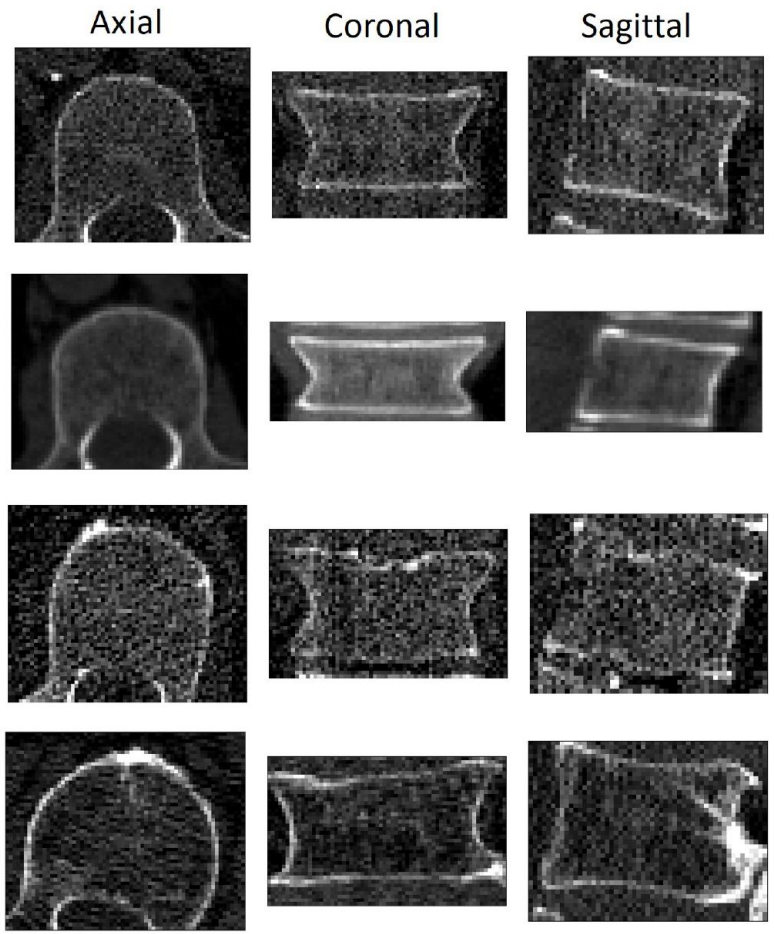
Methodology

Dataset

- MrOS dataset, CT scans of L1-L2 vertebrae
  - 2549 male subjects
  - 92 cases of incident fractures within 10 years
- Follow up fracture information, Age, BMI
- Labels: one or more incident vertebral fracture in 10-year follow up
- Stratified into 4 folds

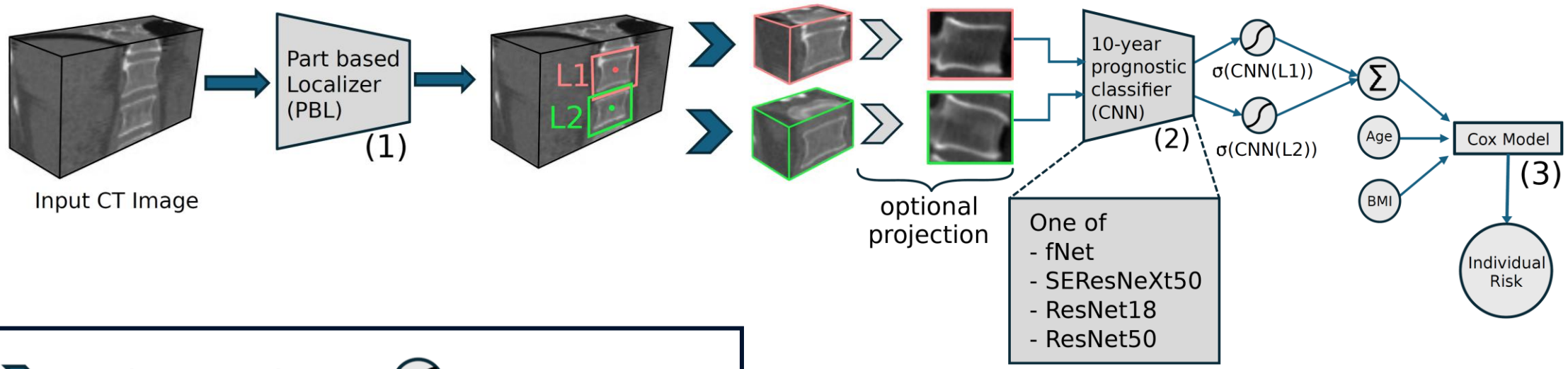
Image + Clinical data

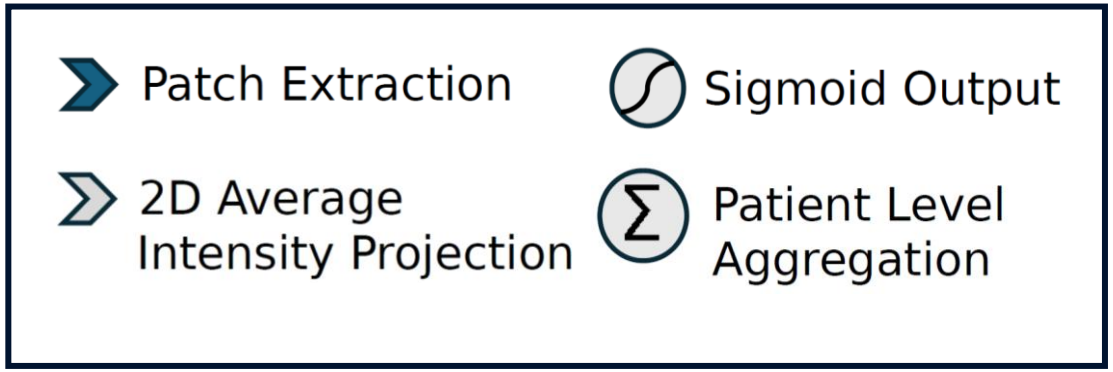
- CNN model
  - Input: 2D or 3D image patches of vertebra
  - Output: Fracture in 10 years: yes/no
  - Evaluation metrics: AUROC, AUPRC
- Cox model
  - Input: Sigmoid values from CNN model, age, BMI
  - Output: Time to fracture/ follow up time, censoring info.
  - Evaluation metrics: Hazard ratio, C-index



Examples of central axial, coronal and sagittal slices (width 1mm each) of CT images from different image acquisition sites of the MrOS dataset

Inference pipeline





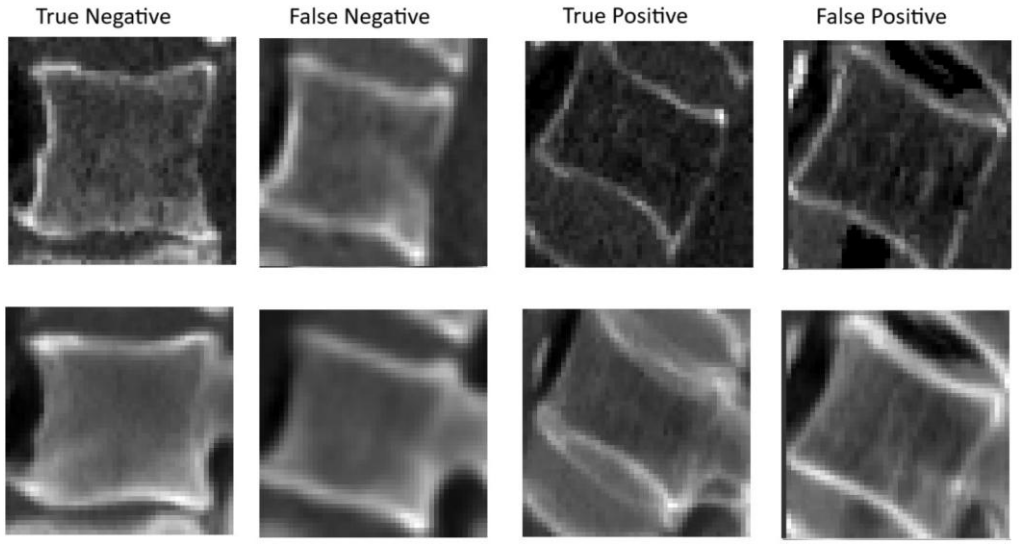
Model Training and Results

Training details

- 4 CNN based model architectures (No pretraining): fNet, ResNet18, ResNet50 & SEResNeXt50
- Model were modified based on input dimension: 2D or 3D
- 4-fold nested cross validation using early stopping

Result

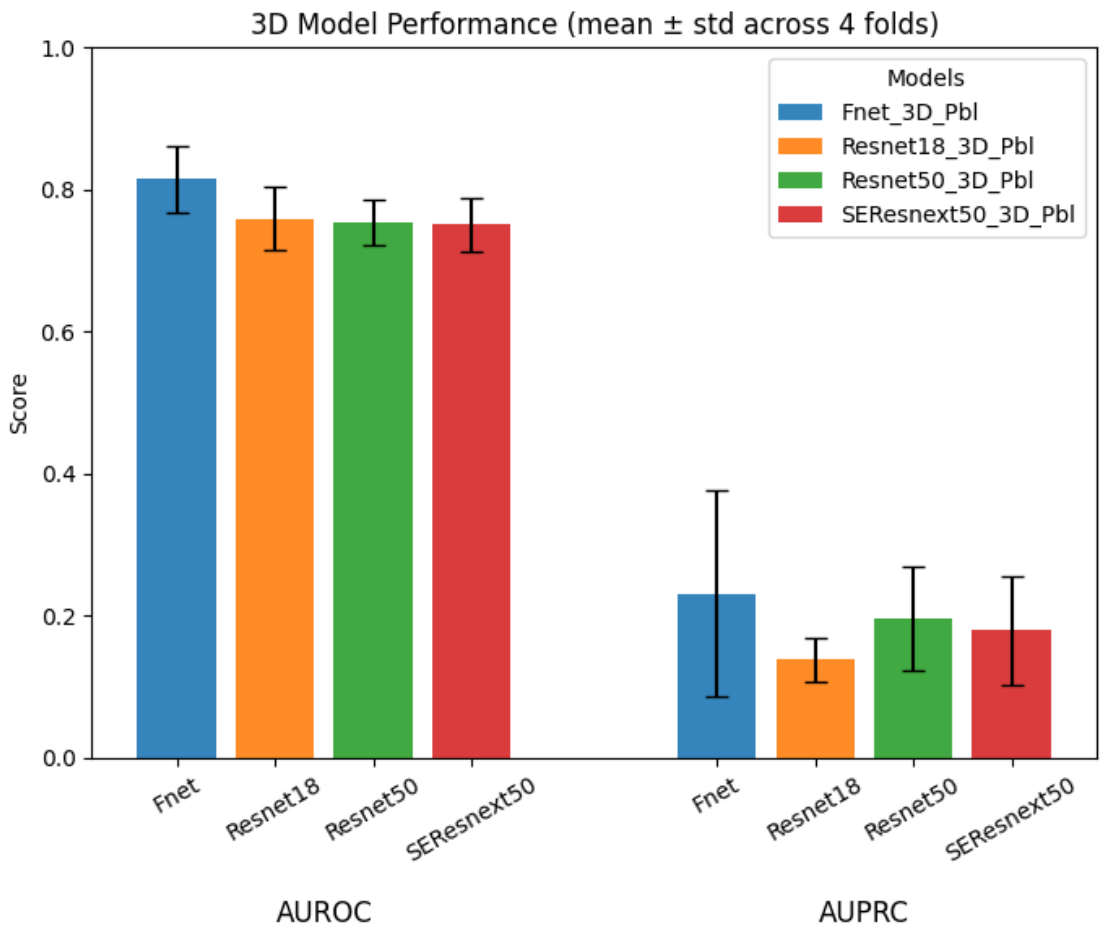
- 3D Model: fNet with 1.17 M params
  AUROC : 81.5 %, AUPRC : 23.1 %
- 2D Model: ResNet18 with 11.17 M params
  AUROC : 80.7 %, AUPRC : 25.0 %
- Improvement of C-index from baseline of 63 (Age + BMI) to 78 (Image + Age + BMI)
- standardized Hazard Ratio: 2.5

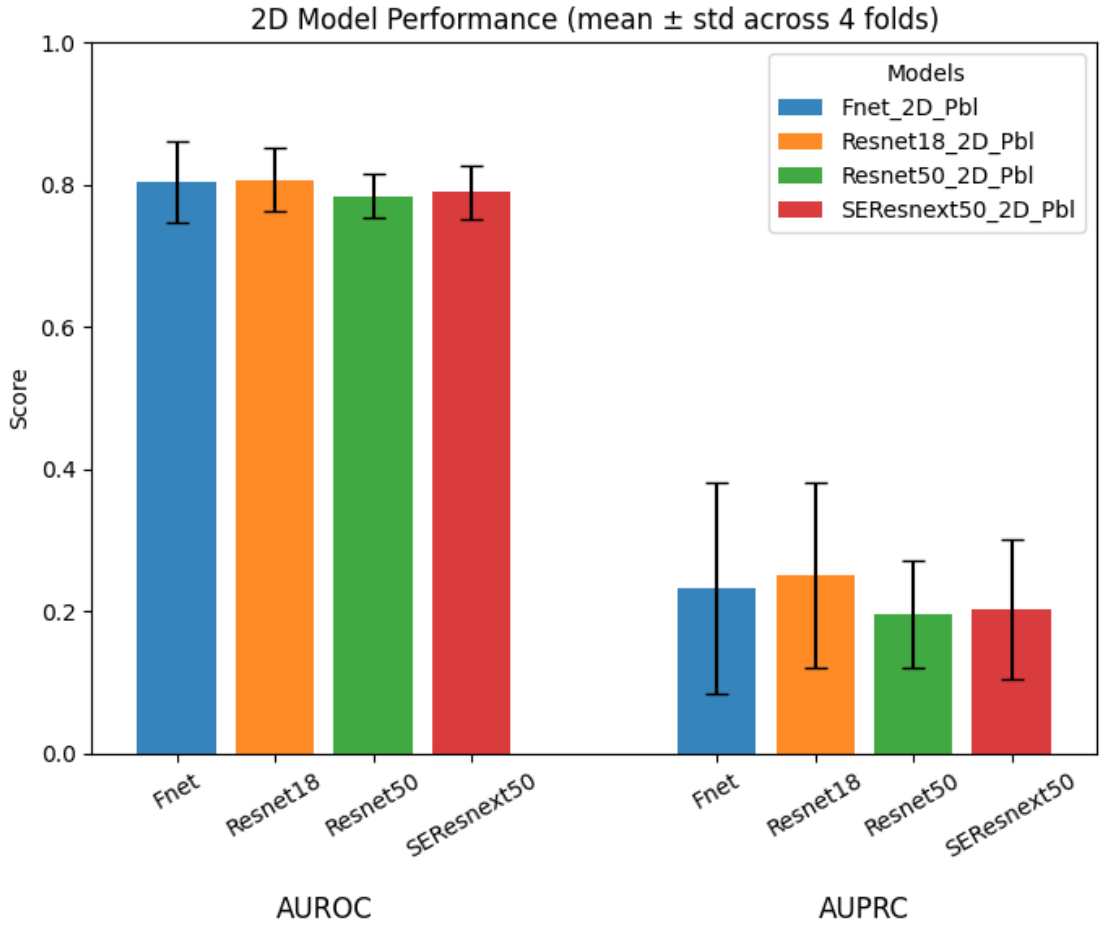


Central sagittal slices of width 1 mm (top) and 30 mm (bottom).
Labels on top represents model output for those patches

Model Training and Results

- 2D model perform better than 3D variants except for the fNet
- High standard deviation across folds especially for AUPRC





Self supervised learning (SSL) based 3D pretraining

- SE-Resnet50 architecture containing squeeze & excitation blocks
- Pretrained using BYOL on about 40K 3D vertebral patches from public datasets
- Outperformed the model trained from scratch on vertebral fracture diagnostic task
- Could finetuning SSL pretrained models on MrOS dataset improve vertebral fracture prognosis?

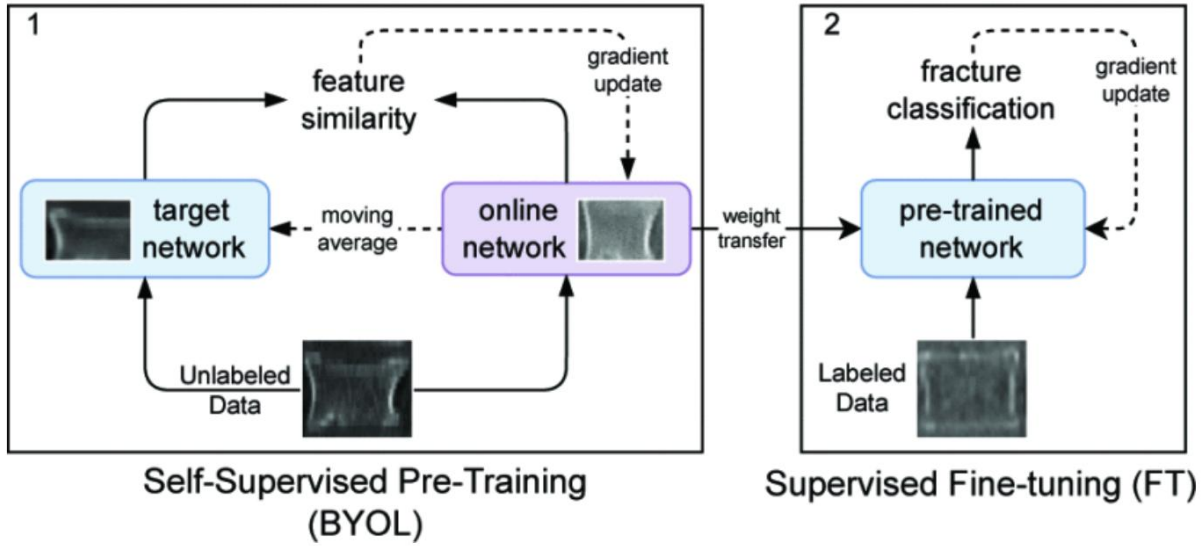


Image ref: Laue et. el. Using unlabeled data in self-supervised training improves automatic vertebral body fracture assessment

Finetuning 3D pretrained SSL model

Methodology: 4 –fold cross validation on MrOS

- used recommended preprocessing
- different learning rates and optimizer settings

Result: Small improvement over AUROC and AUPRC compared to non-pretrained variant

Model	Pretraining	Layerwise weight decay	AUROC	AUPRC
SEResnet50_3D	X	X	80.67 ± 2.43	15.49 ± 3.17
SEResnet50_3D	✓	X	**81.72 ± 1.58**	**20.13 ± 9.18**
SEResnet50_3D	✓	✓	80.93 ± 3.66	18.32 ± 7.18
fNet_3D	X	X	80.39 ± 5.07	21.61 ± 12.05

mean ± std of AUROC and AUPRC over 4 CV folds

Finetuning 2D pretrained model

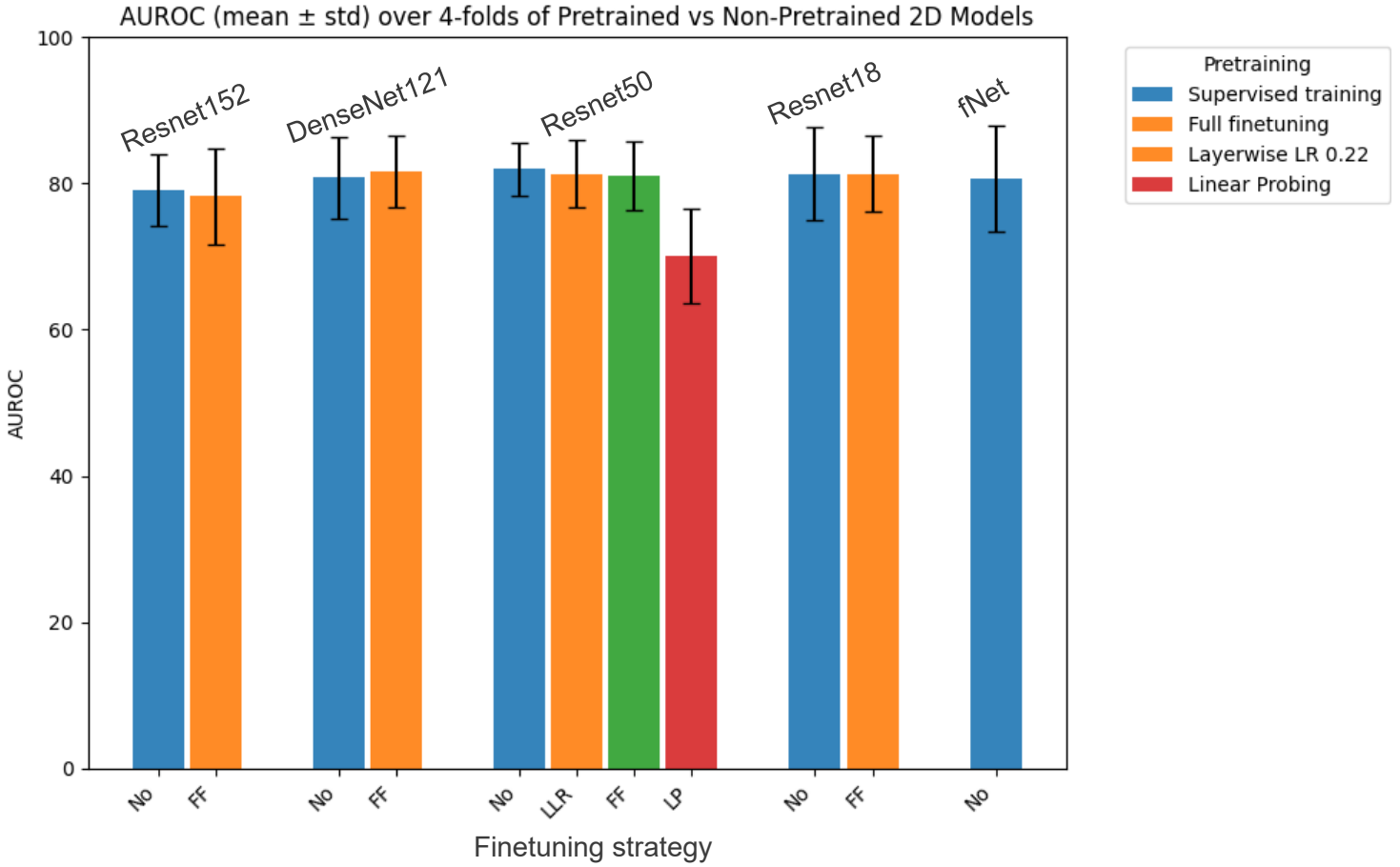
Could finetuning Imagenet pretrained 2D models outperform non-pretrained variant?

- **Models**
  - ResNet18, ResNet50, ResNet152, DenseNet121, fNet
- **Hyperparameters**
  - learning rate, batch size, patch size
  - For ResNet50: test layer wise finetuning & linear probing
- **Metrics**
  - AUROC, AUPRC (mean ± std)

Finetuning 2D pretrained model

General Impressions

- Learning rate has the highest impact on model performance, followed by batch size and input size.
- No pretraining > Layer wise finetuning > Full finetuning > Linear probing
- Pretrained ResNet18/50/152 ~ Non-pretrained versions
- Pretrained DenseNet121 > Non-pretrained, small improvement and may not be significant



CT based large vision language models

Explored CT based 3 large pretrained vision language models: CT-CLIP, VISTA 3D, MERLIN

CT- CLIP [1]

- Pretrained on about 25k Chest CTs
- Multiple abnormality detection based on soft tissues ex. Bronchiectasis, arterial wall calcification

VISTA 3D [2]

- Pretrained on about 11k CT images
- Developed primarily for Segmentation task, Image encoder can be used to extract features

MERLIN [3]

- discussed in the next slide…

[1] Hamamci et el. Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography
[2] He et el. VISTA3D: A Unified Segmentation Foundation Model For 3D Medical Imaging
[3] Blankemeier et. el. Merlin: A vision language foundation model for 3d computed tomography

Merlin a Vision-Language Model, Blankemeier et. el (2024)

- Merlin: Vision-Language model pretrained on 3D CT datasets
- Structured (EHR) and unstructured data (Reports, abdominal CT scans): ~6.4M images from 15k CTs
- Text Encoder: Longformer pretrained model
- Image Encoder: Resnet152 (pretrained 2D, inflated to 3D)
- Merlin paper presents results related to vertebral fracture
  - Vertebral fracture diagnosis (full spine) on VerSe-2019, F1-score of 0.767 using zero-shot learning
  - 5-year vertebral fracture prediction (in house dataset): AUROC of 0.8 using finetuning

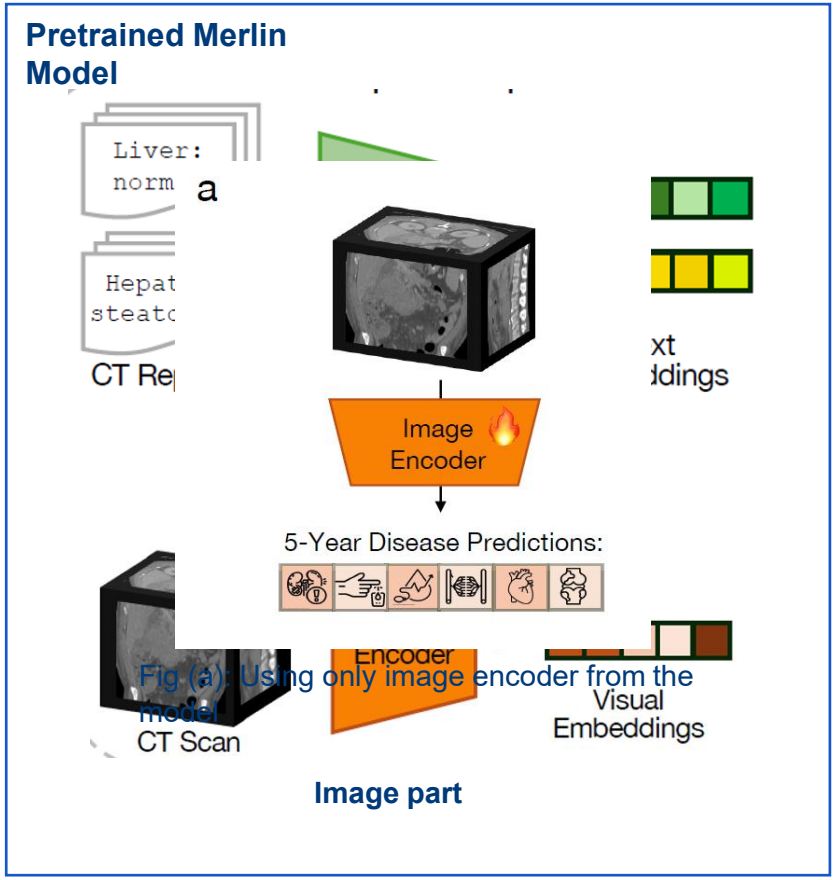


Fig (a): Using only image encoder from the

Image ref: Blankemeier et. el. Merlin: A vision language foundation model for 3d computed tomography

Finetuning Merlin for vertebral fracture prognosis on MrOS dataset

Methodology

- Finetuned only the Image encoder on MrOS dataset by replacing the final layer with fully connected layers
- Preprocessing: changed HU Clipping range, data augmentation to make it suitable for MrOS dataset
- Hyperparameters: different learning rates, finetuning BN layer only, layerwise learning rate, optimizers, input patch sizes

Finetuning Merlin for vertebral fracture prognosis on MrOS dataset

Results

All Models were trained on MrOS dataset i.e. either finetuned or supervised trained from scratch

Model	Pretraining	AUROC	AUPRC
Merlin_3D	No	81.71 ± 4.85	20.97 ± 12.26
Merlin_3D	**Imagenet**	**83.46 ± 2.07**	**24.51 ± 14.76**
Merlin_3D	CT + Reports	81.43 ± 5.86	21.91 ± 13.2
fNet_3D	No	80.39 ± 5.07	21.61 ± 12.05
SEResnet50_3D	SSL	81.72 ± 1.58	20.13 ± 9.18

mean ± std of AUROC and AUPRC over 4 CV folds

Conclusions

- Slightly better performance for the Imagenet pretrained model, but the CT pretrained model performs worse than the Imagenet model
- Repeat experiments shows higher std over 4-folds for non-pretrained model as compared to pretrained

Summary

- CT based vertebral fracture prediction shows
  - Improved prediction compared to Age and BMI only model
  - Moderate impact of CNN architectures on classification performance
  - 2D models performs slightly better than corresponding 3D (except fNet)
  - AUROC and C-index are comparable to other state of the art methods in literature
- Several attempts to improve performance using pretrained model shows minor improvement
  - 2D: Imagenet pretraining
  - 3D: Supervised pretraining on diagnostic vertebral fracture, SSL pretrained
- Large vision language model when finetuned has higher AUROC but it may not be statistically significant due to high variability across 4 folds

Thank you