Cervical Spine Fracture Detection

Team presentation 3

W210 Capstone
Fall 2022 - Section 3
Fengyao Luo, Weijia Li, Jane Hung, Minjie Xu

Agenda

Background

Product Demo

Dataset

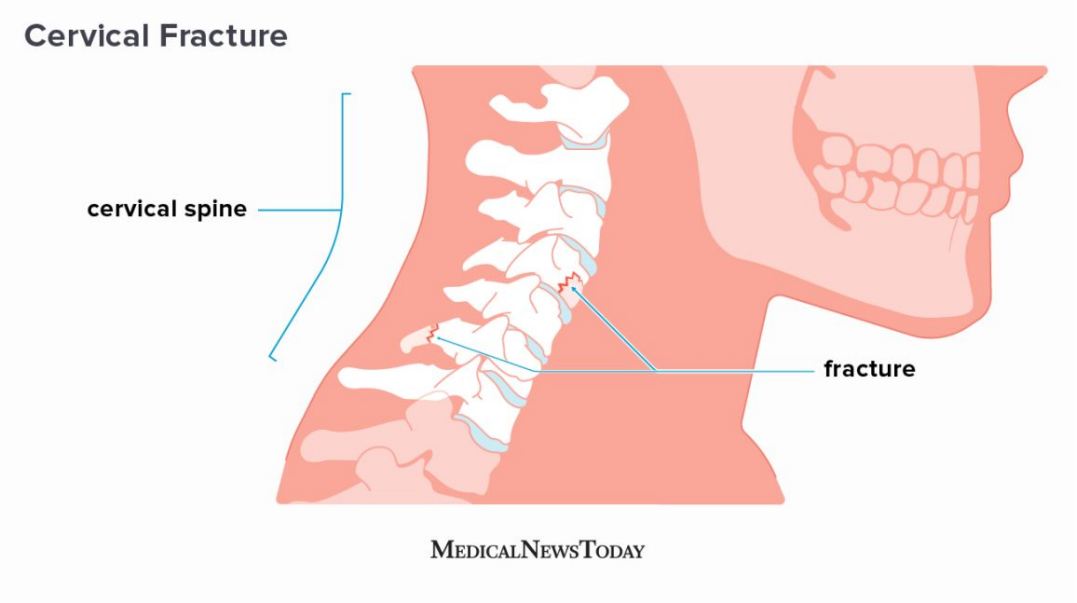
Pipeline

Models

Model Performance

Future Work

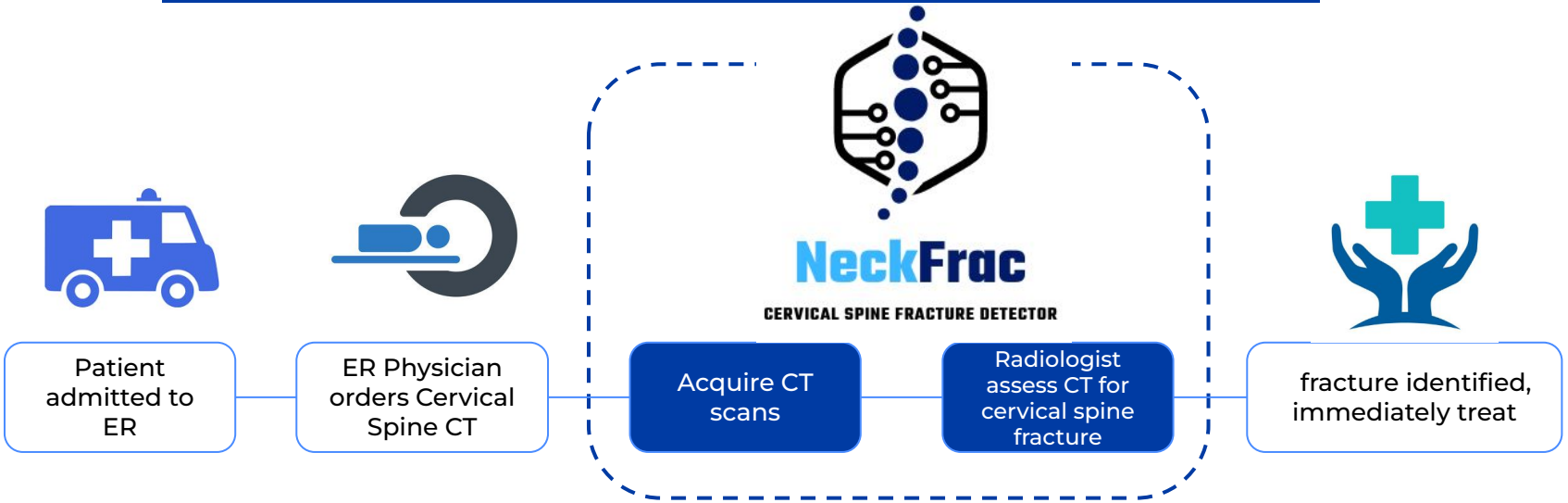
Background



- Cervical Spine Fractures (broken neck)
- **1.5 million** vertebral compression fractures occur
- 3 millions patients per year
- only **25–33%** of incident radiographically identified vertebral fractures are clinically diagnosed
- Quickly detecting and determining the location of vertebral fractures is essential for prevent paralysis after trauma.
- RSNA Cervical Spine Fracture AI Challenge

Mission

Quicker, better, more accurate diagnosis to save lives.

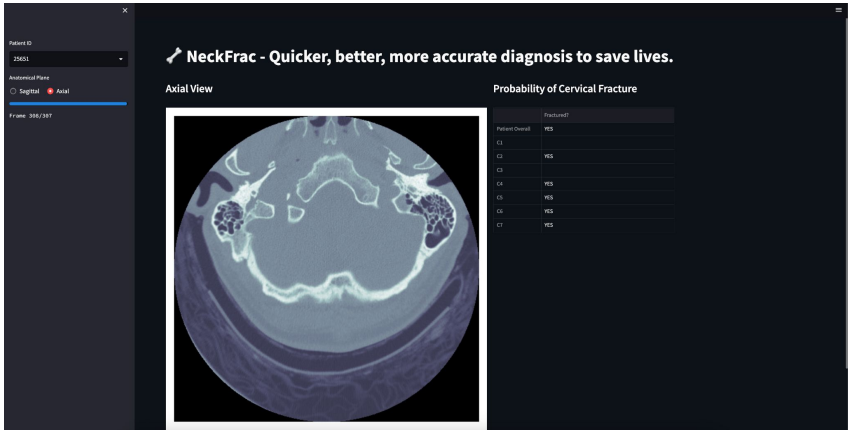


- Traditionally, this process takes 30 mins on average with the first priority.
- With NeckFrac, this process will take less than 10 mins.

Product Demo

Introduce Tool ▶️

Fracture Bounding Boxes ▶️



Datasets

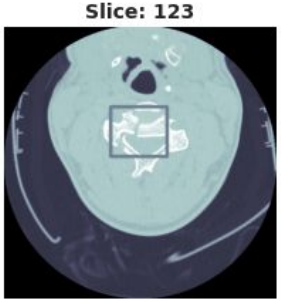
Train.csv
2019 patients in train set (balanced at patient_overall level)

	StudyInstanceUID	patient_overall	C1	C2	C3	C4	C5	C6	C7
0	1.2.826.0.1.3680043.6200	1	1	1	0	0	0	0	0
1	1.2.826.0.1.3680043.27262	1	0	1	0	0	0	0	0
2	1.2.826.0.1.3680043.21561	1	0	1	0	0	0	0	0
3	1.2.826.0.1.3680043.12351	0	0	0	0	0	0	0	0
4	1.2.826.0.1.3680043.1363	1	0	0	0	0	1	0	0

Train Image / Metadata
No missing data, Image size varies, need to resize to 512x512

- Patient ID
- Slice Number
- Image Size
- Slice Thickness
- Image Position Patient
- Image Orientation Patient

Bounding Box
235 patients (12% of train set) have the bounding box



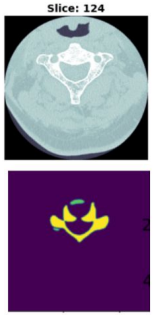
- Most of patient have 15-25 bounding boxes
- Patients rarely have 100 bounding boxes, distribution right skewed

Segmentation
87 patients (4% of train set) have segmentation labelled

- Image has overlap of adjacent vertebrae
- Need to predict C1-C7 for rest of train set

	C1	C2	C3	C4	C5	C6	C7
150	1	1	0	0	0	0	0

Pipeline (1/4)



1) Read Scans

```
array([[25, 22, 24, ..., 22, 33, 26],
       [31, 26, 23, ..., 29, 36, 13],
       [28, 28, 23, ..., 26, 27, 12],
       ...,
       [10, 27, 37, ..., 56, 68, 34],
       [25, 14,  4, ..., 50, 75, 48],
       [41, 19, 13, ..., 30, 58, 47]], dtype=uint16)
```

2) Feature Engineering

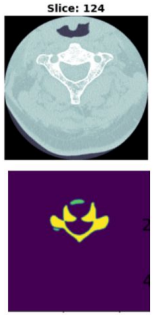
3) Model

4) Evaluation

5) Inference

Fracture ?

Pipeline (2/4)



1) Read Scans

```
array([[25, 22, 24, ..., 22, 33, 26],
       [31, 26, 23, ..., 29, 36, 13],
       [28, 28, 23, ..., 26, 27, 12],
       ...,
       [10, 27, 37, ..., 56, 68, 34],
       [25, 14,  4, ..., 50, 75, 48],
       [41, 19, 13, ..., 30, 58, 47]], dtype=uint16)
```

2) Feature Engineering

3) Model

4) Evaluation

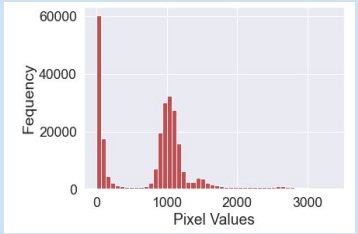
5) Inference

Fracture ?

HU Rescaling

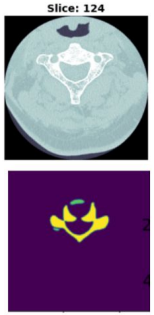
Normalization

Min: HU = -1000 (air)
Max: HU = 1900 (bone)



- **Random Anisotropy** — Make images look anisotropic 25% of times
- **Random Affine** — Apply a random affine transformation
- **Random Noise** — Apply Gaussian noise 25% of times
- **Random Flip**

Pipeline (3/4)



1) Read Scans

```
array([[25, 22, 24, ..., 22, 33, 26],
       [31, 26, 23, ..., 29, 36, 13],
       [28, 28, 23, ..., 26, 27, 12],
       ...,
       [10, 27, 37, ..., 56, 68, 34],
       [25, 14,  4, ..., 50, 75, 48],
       [41, 19, 13, ..., 30, 58, 47]], dtype=uint16)
```

2) Feature Engineering

3) Model

4) Evaluation

5) Inference

Fracture ?

EfficientNet

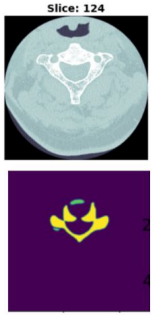
Treat each vertebrate as its own standalone EffNet Model

2.5D Model

Stage 1:
Predict Segmentation with 2.5D UNet

Stage 2:
Predict Fracture with Resnet 50 spatial dropout bi-GRU

Pipeline (4/4)



1) Read Scans

```
array([[25, 22, 24, ..., 22, 33, 26],
       [31, 26, 23, ..., 29, 36, 13],
       [28, 28, 23, ..., 26, 27, 12],
       ...,
       [10, 27, 37, ..., 56, 68, 34],
       [25, 14,  4, ..., 50, 75, 48],
       [41, 19, 13, ..., 30, 58, 47]], dtype=uint16)
```

2) Feature Engineering

3) Model

4) Evaluation

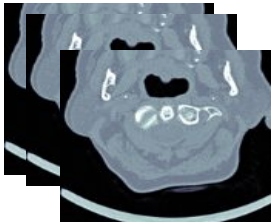
5) Inference

Fracture ?

4) Evaluation

Train Val Loss	Recall	Precision
Accuracy	F2 Score	FP FN Rate

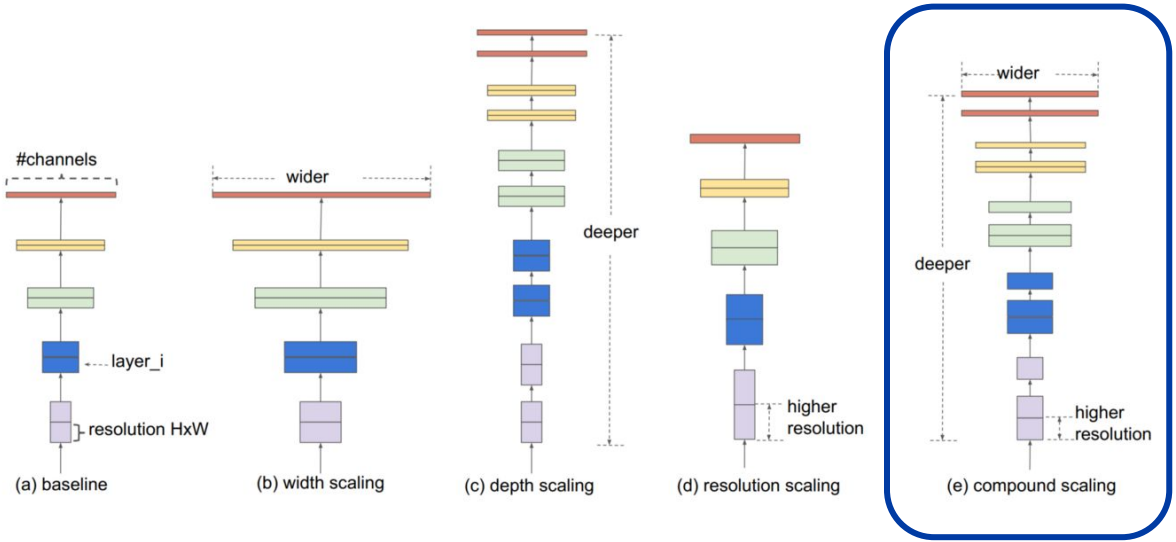
5) Inference



StudyInstanceUID	prediction_type	fractured	pred_flag
1.2.826.0.1.3680043.14833	patient_overall	0.599315107	1
1.2.826.0.1.3680043.14833	C1	0.68948257	1
1.2.826.0.1.3680043.14833	C2	0.420390278	1
1.2.826.0.1.3680043.14833	C3	0.389513612	1
1.2.826.0.1.3680043.14833	C4	0.362404943	1
1.2.826.0.1.3680043.14833	C5	0.733516157	1
1.2.826.0.1.3680043.14833	C6	0.156630933	0
1.2.826.0.1.3680043.14833	C7	0.836084545	1

Model I: EfficientNet (1/2)

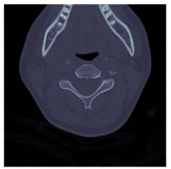
Why EfficientNet?

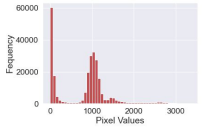


It's all about Scale!

EfficientNet paper — Tan, M., & Le, Q. v. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learning, ICML 2019*, *2019-June*. Link: https://arxiv.org/abs/1905.11946

Model I: EfficientNet (2/2)

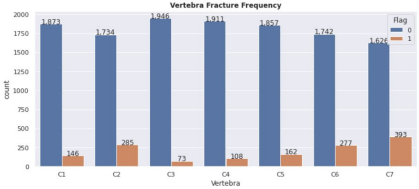




(512 , 512)

(128 , 128)

Training Image → HU Rescaled → Normalized → Rescale Image



Class Weight → Model

Minimum validation loss: 0.14071543514728546

EfficientNet for C1 | EfficientNet for C2 | • • • | EfficientNet for C7

Cross Validation

Inference on Validation Images

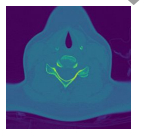
	row_id	fractured
0	1.2.826.0.1.3680043.22327_C1	0.596053
1	1.2.826.0.1.3680043.22327_C2	0.616532
2	1.2.826.0.1.3680043.22327_C3	0.329862
3	1.2.826.0.1.3680043.22327_C4	0.577175
4	1.2.826.0.1.3680043.22327_C5	0.434725
5	1.2.826.0.1.3680043.22327_C6	0.453186
6	1.2.826.0.1.3680043.22327_C7	0.507186
7	1.2.826.0.1.3680043.22327_patient_overall	1.000000

Model II: 2.5D UNet + biGRU

Stage 1: Predict Segmentation

Load Data
dicom/nii files

Normalize & Quantization



Save to Local

2.5D

Cross Validation

Train dataset
(87 examples)

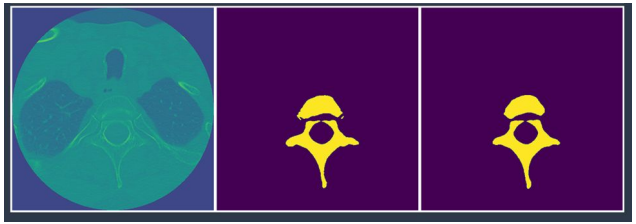
Augmentation

UNet℠
(pretrained: Imagenet)

Save Models

Stage 2: Predict Fracture

Infer Segmentation



Crop Voxel

Cross Validation

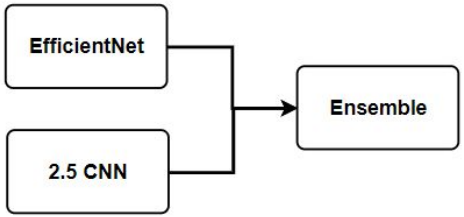
Train dataset
(2017 examples)

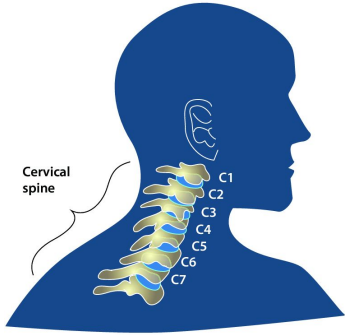
Augmentation

Resnet 50
spatial dropout
bi-GRU

Save Models

Ensemble Model Performance



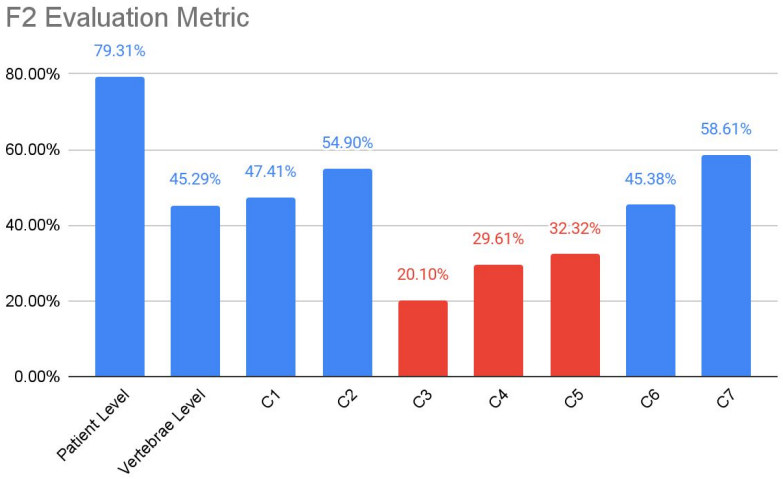


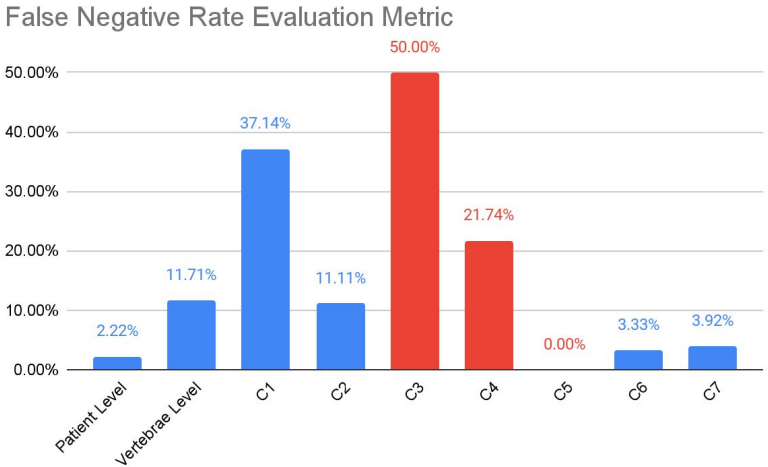
* Assign variable classification threshold for each vertebrae based on F2

* Model does well at detecting overall patient level fractures according to F2

* C3-C5 are the worst performing vertebrae in the model

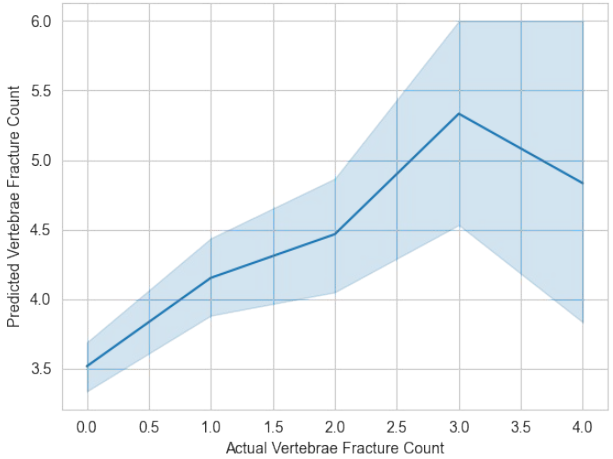
Best Classification Thresholds	
Vertebrae	**Classification Threshold**
C1	0.42
C2	0.36
C3	0.19
C4	0.19
C5	0.13
C6	0.22
C7	0.36

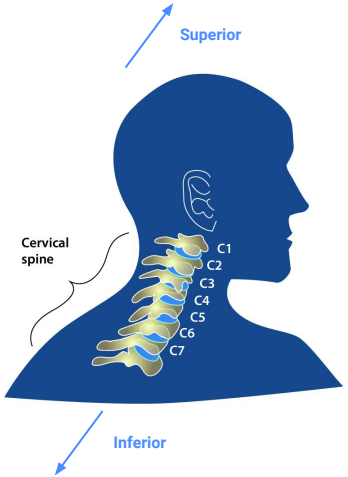


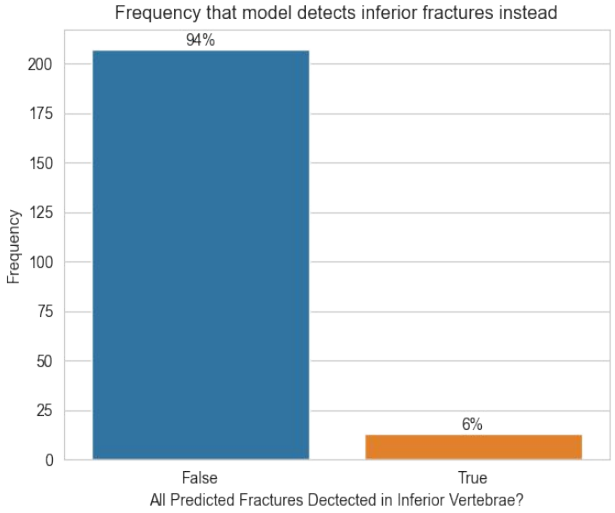


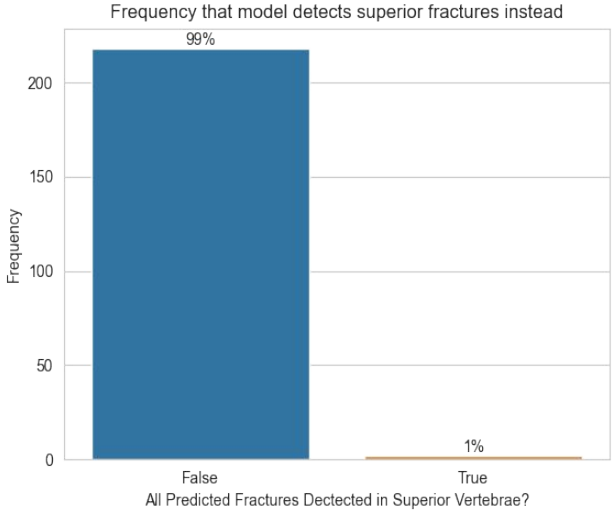
Ensemble Model Error Analysis

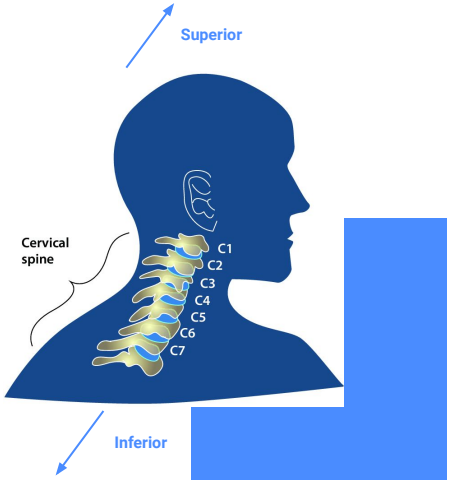
Model predicts mostly 3.5 - 5.5 fractures on average regardless of actual vertebrae fracture count

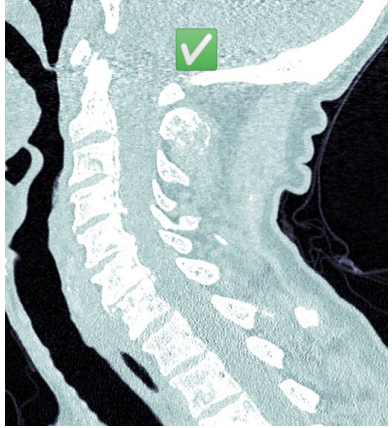












* Common for model to overpredict the number of fractures

* Better orientation leads to better predictions

Patient
1.2.826.0.1.3680043.21982

Vertebrae Label	Actual Vertebrae Fracture	Predicted Vertebrae Fracture
C1	-	-
C2	-	-
C3	-	-
C4	**-**	**Yes**
C5	**Yes**	**Yes**
C6	**Yes**	**Yes**
C7	**-**	**Yes**

Patient
1.2.826.0.1.3680043.17370

Vertebrae Label	Actual Vertebrae Fracture	Predicted Vertebrae Fracture
C1	-	-
C2	**Yes**	**Yes**
C3	-	-
C4	-	-
C5	-	-
C6	-	-
C7	**Yes**	**Yes**

Business Impact

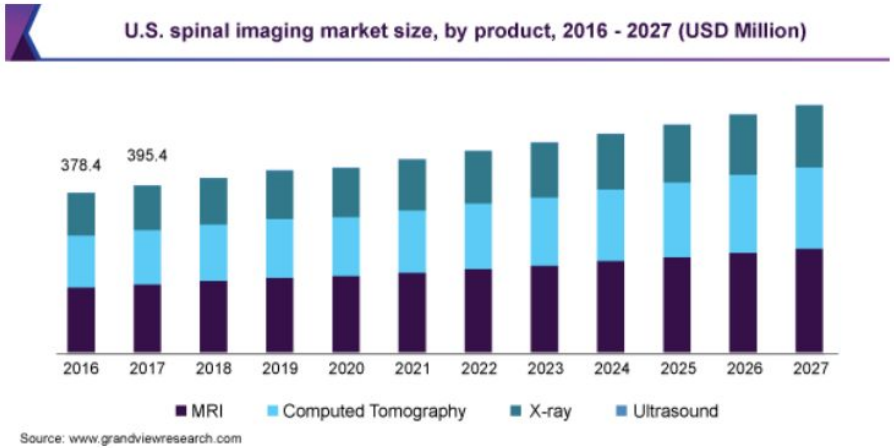
	Recall / Sensitivity	Inference Time
Radiologist	93%	33-43 min
AIDOC Model	76%	3-8 min
NeckFrac	**98%**	**2min**

Future Work

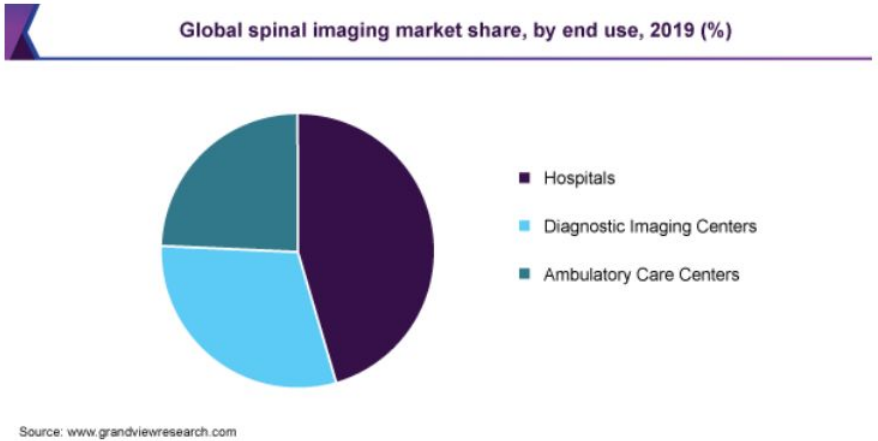
- Labeling stretched images for better error analysis
- Incorporating soft tissue into model to identify fracture
- Combine three anatomical planes - axial, sagittal, and coronal - for modeling
- Designing a model API for use in CT imaging devices
- Rank the fracture severity (low, medium, high) based on probability

Appendix

Market Analysis & Problem Statement



- The global spinal imaging market size was valued at **$1.6 billion** in 2019 with expect growth rate of **5.2%** from 2020 to 2027
- Estimated market value of CT in U.S. is over **$150 million** in 2022
- The demand is continuing growing rapidly



- MVP and product version
- Mission statement:
  Quicker, better, more accurate diagnosis to save lives.

Competitors' Products

Company/Author	Product Name	Performance
aidoc ALWAYS ON AI	C-spine	A FDA-approved CNN in cervical spine fracture detection with overall accuracy 92%, sensitivity 76%, and specificity 97%
香港中文大學 The Chinese University of Hong Kong	Ofeye 1.0	Achieve accuracy 93.9%, sensitivity of 86%, and specificity 97.1%
Joseph Redmon University of Washington	YOLOv3 software	CNN (darknet-53) models with accuracy 93%, sensitivity 91%, and specificity 93%.
NIH Clinical Center America's Research Hospital	Not disclosure	The computer system has a accuracy 95%, sensitivity 95.7%, and specificity 77.3%. Localization of fractures: a false-positive rate of 0.29 per patient.

EDA - Part I

Patient Overall Fracture:

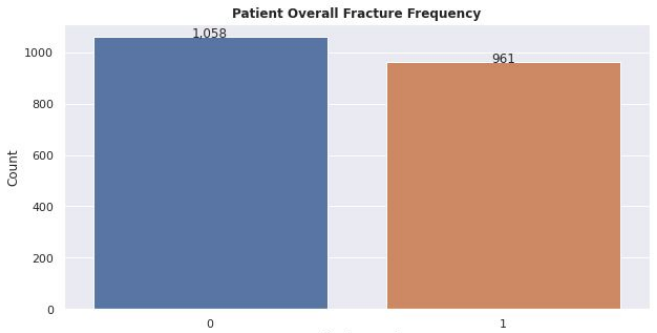
- 48% of patients in the train set have broken bones. The training data is well-balanced.

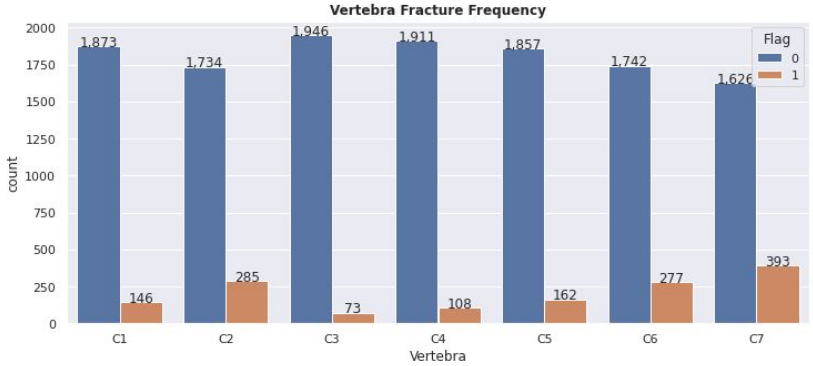
Vertebrae Fracture Frequency:

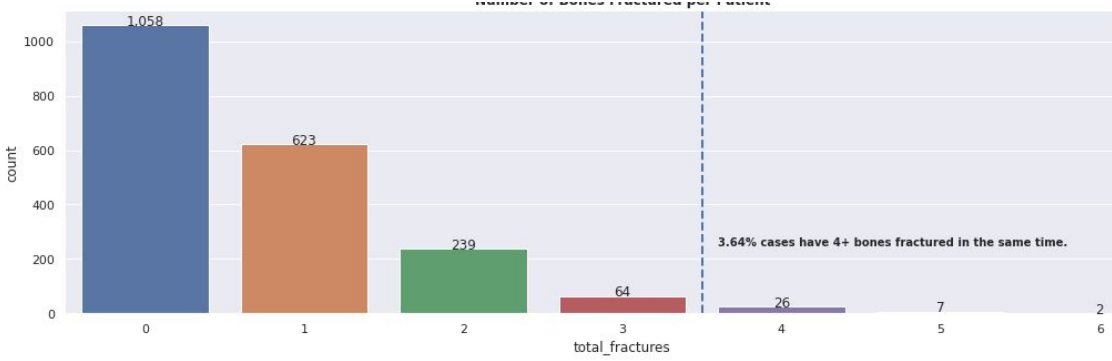
- C7, C2 bones have most frequent injuries. C3, C4 bones have the least frequent injuries.

N of Bones Fractured:

- Majority (65%) of patients broke only 1 bone, and very few (3%) patients broke over 4 bones.
- Patients, who have over 1 broken bone, usually broke the bones in the nearby area, such as broke C6, C7 or C1, C2 at the same time.







EDA - Part II

Patient slice counts:

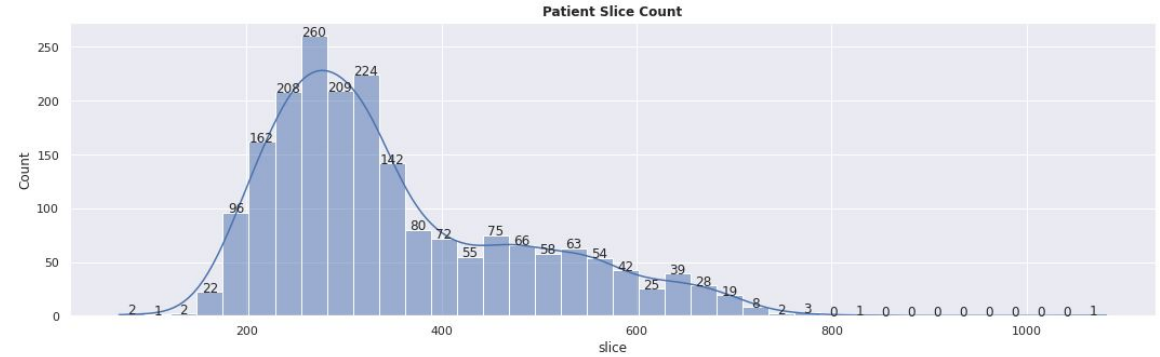
- Majority of patients have 200-400 CT scans in the train set. Distribution right skewed.

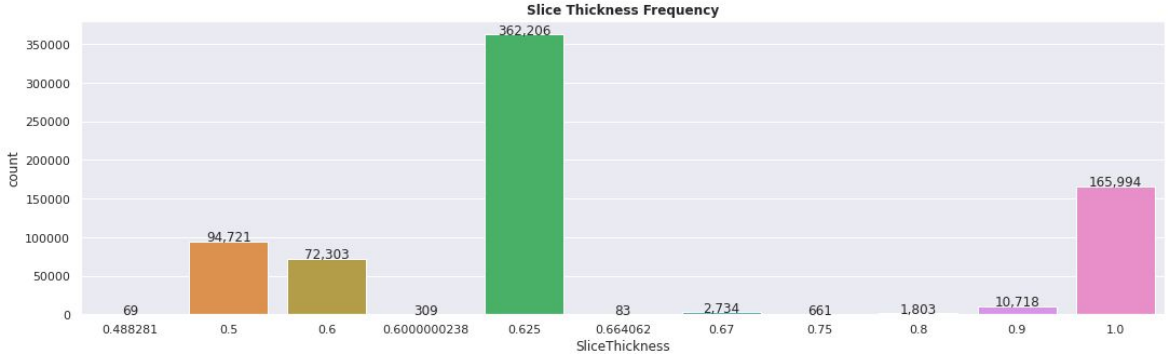
Slice Thickness Frequency:

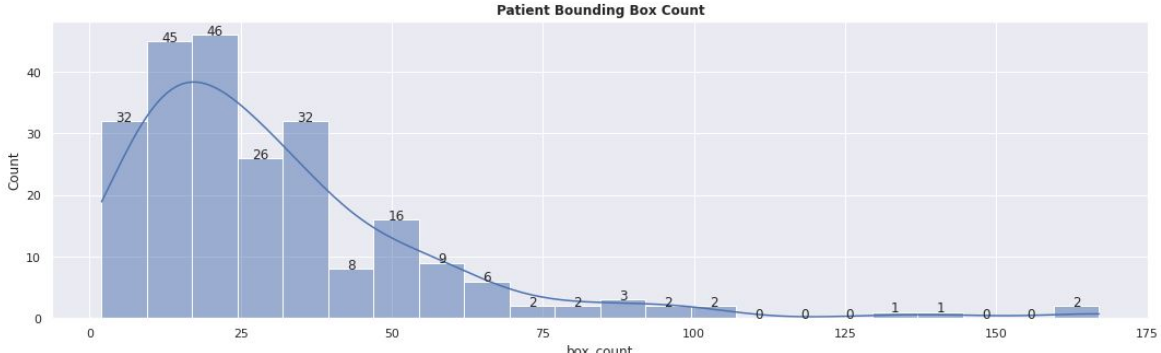
- Most of patients' CT scan slice thickness is 0.62
- Many variation

Bounding box counts:

- Most of patient have 15-25 bounding boxes
- Patients rarely have 100 bounding boxes, distribution right skewed

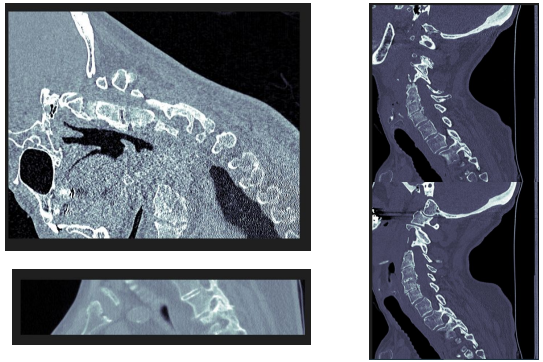




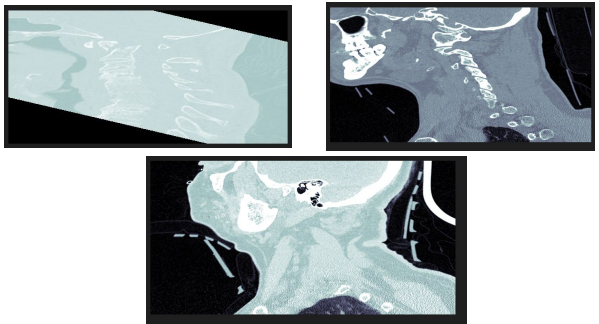


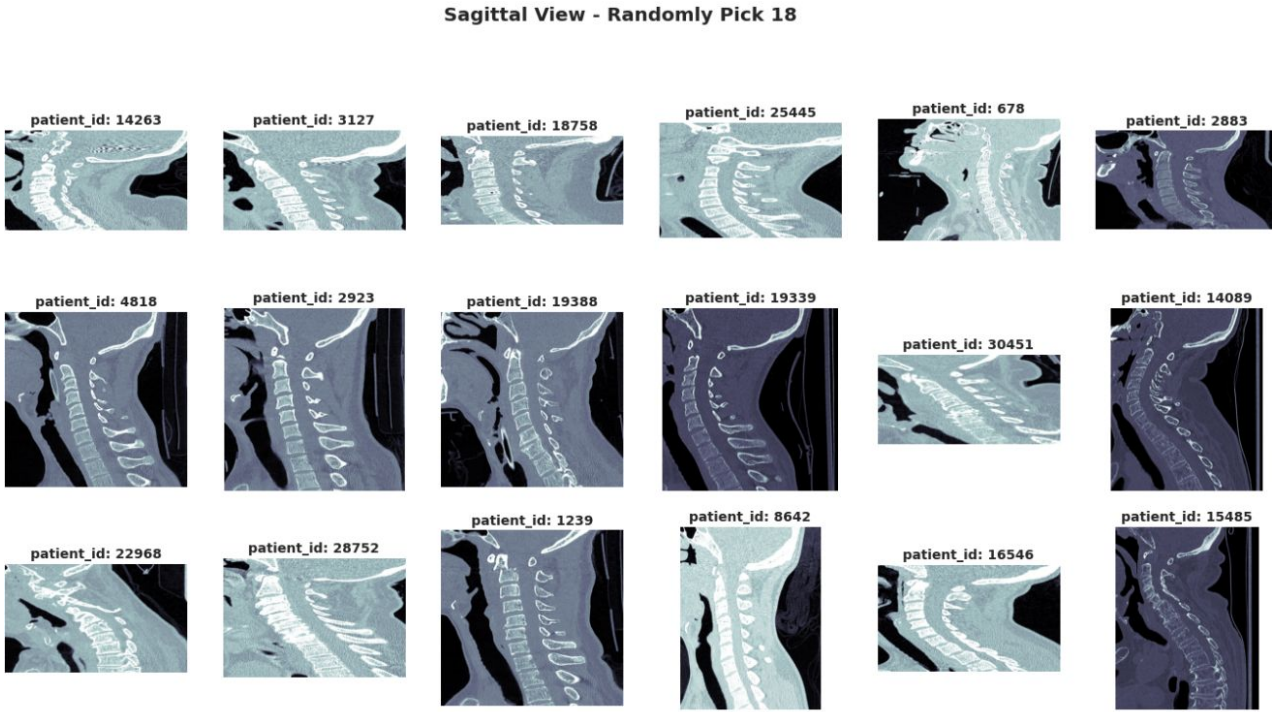
Sagittal Views

Scans to ignore:
Due to funky image



Scans to ignore:
Due to image size



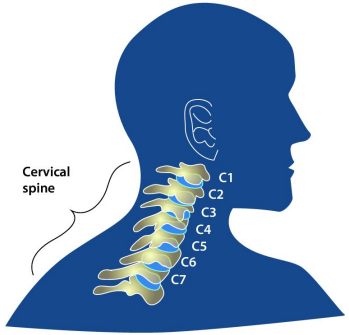


EfficientNet Model Performance

* Assign variable classification threshold for each vertebrae based on F2 and false positive count

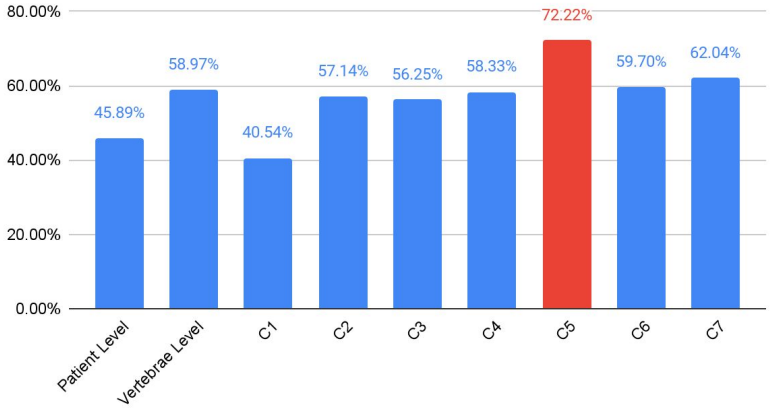
* With only 500 patients used in validation, the model struggles with C3-C5 fracture detection according to F2

* We miss detection of 72.22% of C5 fractures

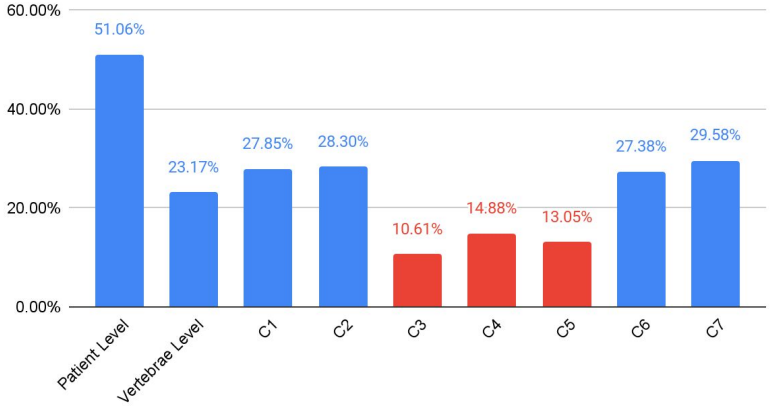


Best Classification Thresholds	
Vertebrae	**Classification Threshold**
C1	0.67
C2	0.42
C3	0.31
C4	0.34
C5	0.73
C6	0.21
C7	0.81

False Negative Rate Evaluation Metric

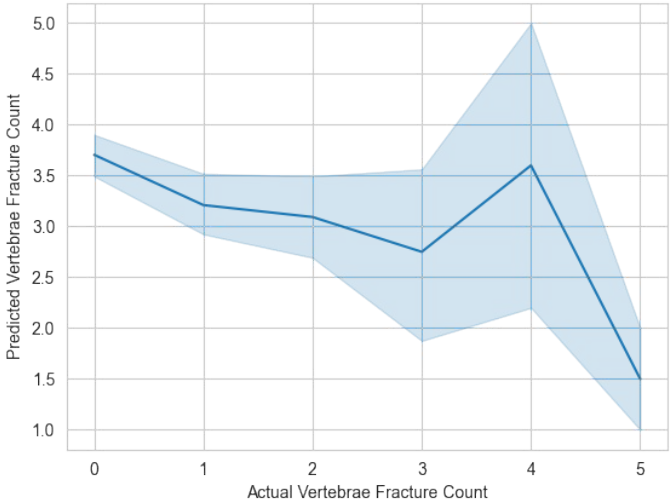


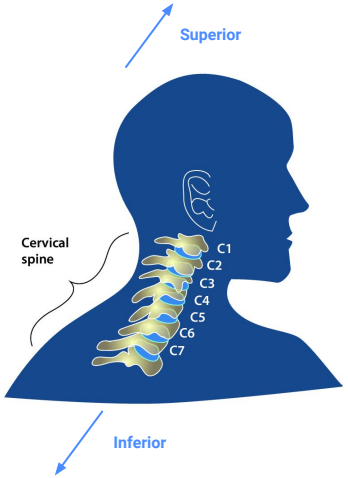
F2 Evaluation Metric

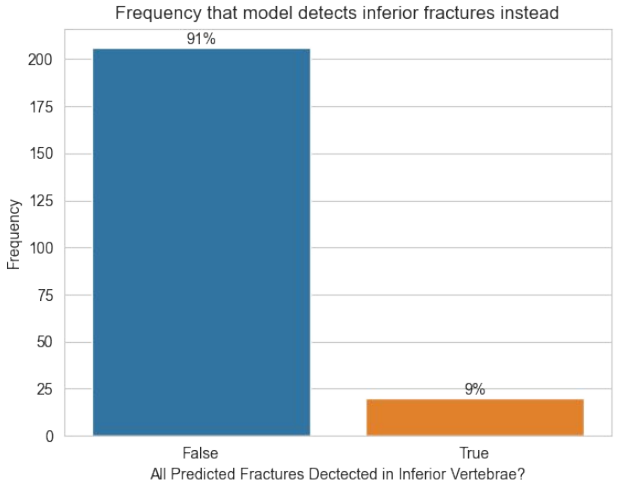


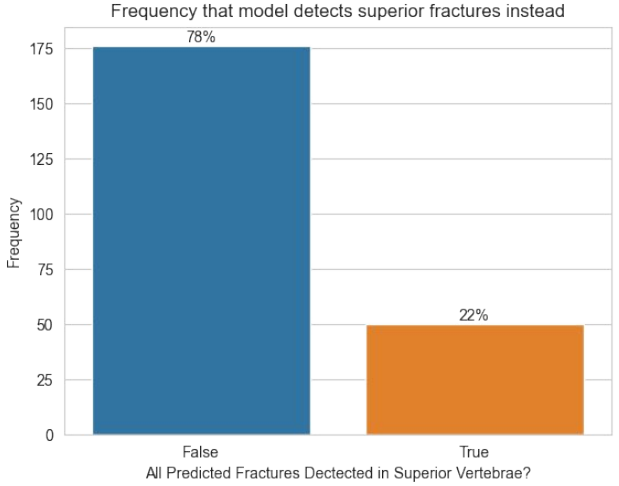
EfficientNet Error Analysis

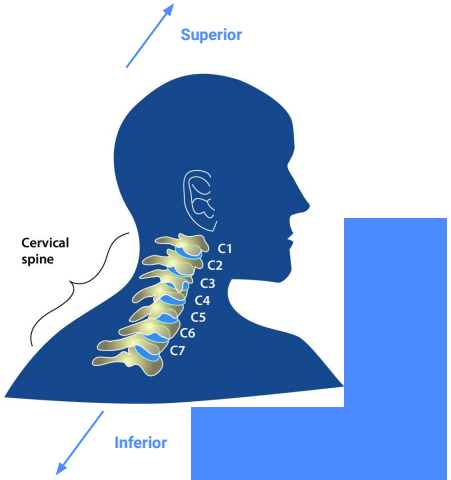
Model predicts mostly 2 - 4 fractures on average regardless of actual vertebrae fracture count









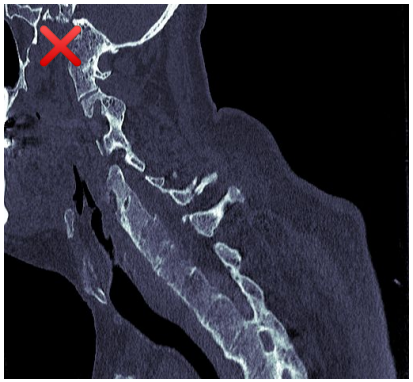


* Predict all inferior vertebrae fractures 9% of the time

* Predict all superior vertebrae fractures 22% of the time

1.2.826.0.1.3680043.21561

Actual Vertebrae Fracture	Predicted Vertebrae Fracture
None	None
C2	**None**
None	None
None	**C4**
None	None
None	**C6**
None	None



1.2.826.0.1.3680043.22438

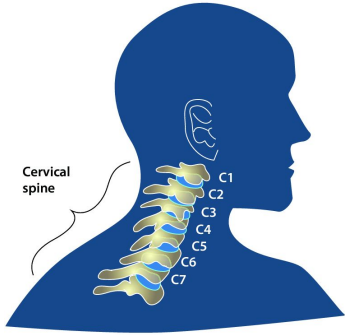
Actual Vertebrae Fracture	Predicted Vertebrae Fracture
None	None
None	**C2**
None	None
None	None
None	None
C6	**None**
C7	**None**

2.5D UNet + biGRU Model Performance

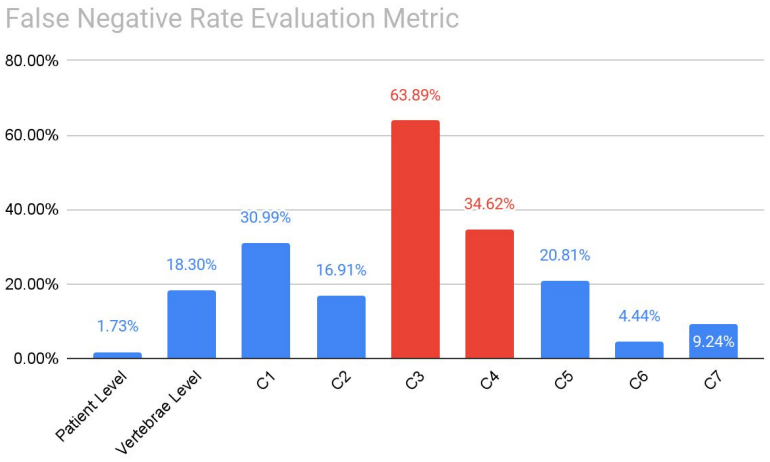
* Assign variable classification threshold for each vertebrae based on F2

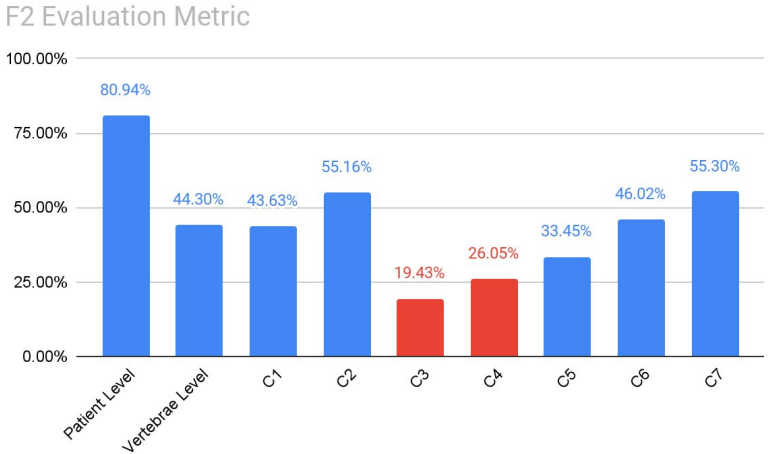
* Model does well at detecting overall patient level fractures according to F2

* C3 and C4 are the worst performing vertebrae in the model

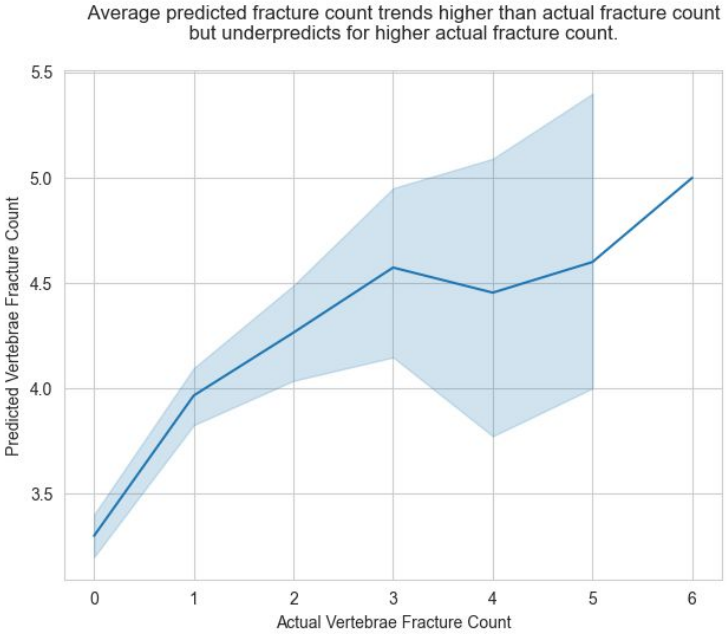


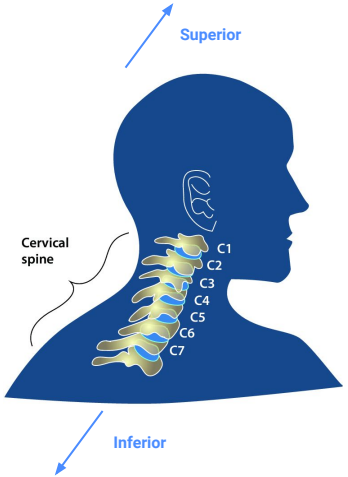
Best Classification Thresholds	
Vertebrae	**Classification Threshold**
C1	0.23
C2	0.34
C3	0.23
C4	0.15
C5	0.19
C6	0.23
C7	0.42

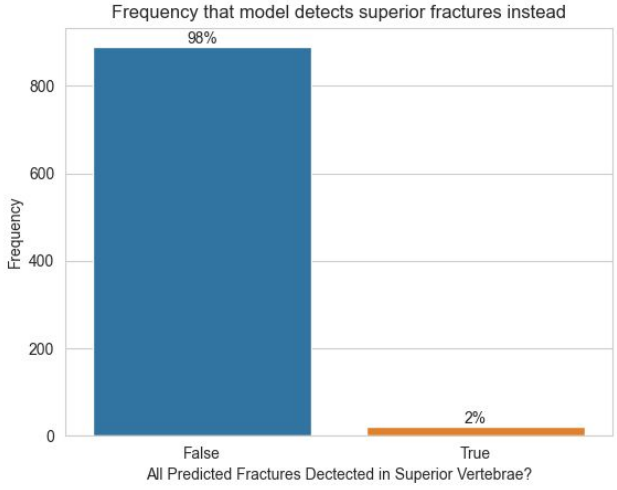


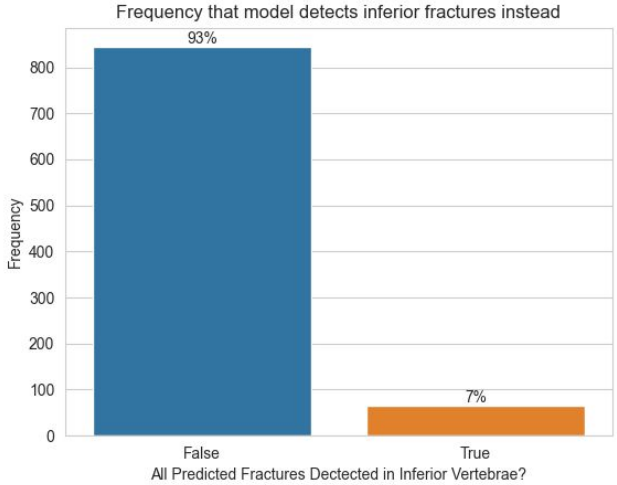


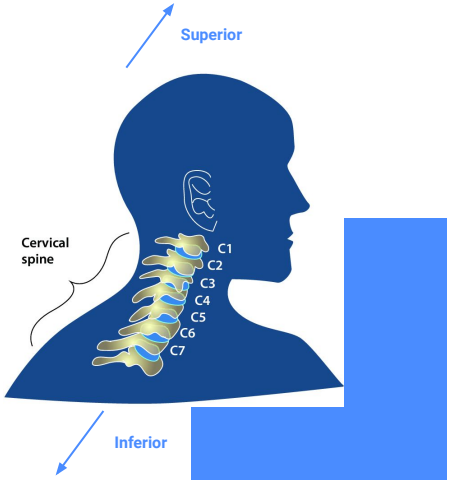
2.5D UNet + biGRU Error Analysis





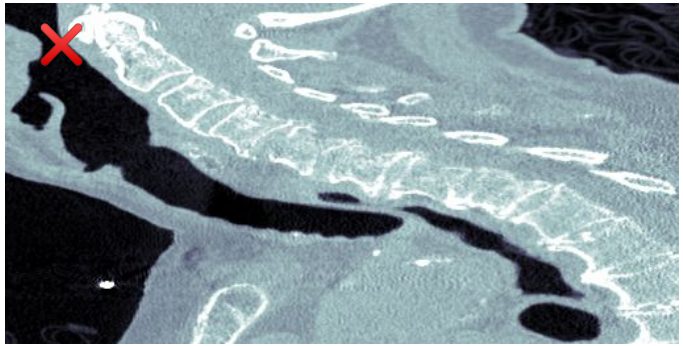






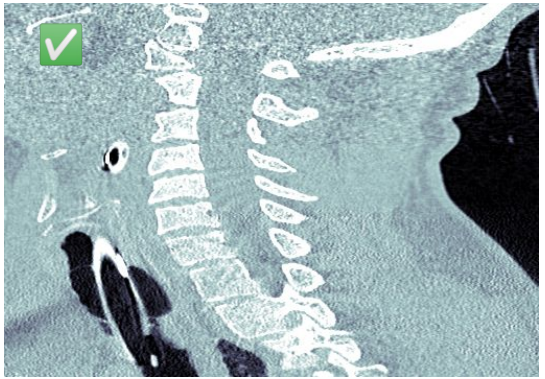
* Common for model to overpredict the number of fractures

* Better orientation leads to better predictions



1.2.826.0.1.3680043.10005

Actual Vertebrae Fracture	Predicted Vertebrae Fracture
None	None
None	**C2**
None	**C3**
None	**C4**
None	**C5**
None	**C6**
None	**C7**



1.2.826.0.1.3680043.10515

Actual Vertebrae Fracture	Predicted Vertebrae Fracture
None	None
None	**C2**
None	None
None	None
None	None
None	None
C7	**C7**

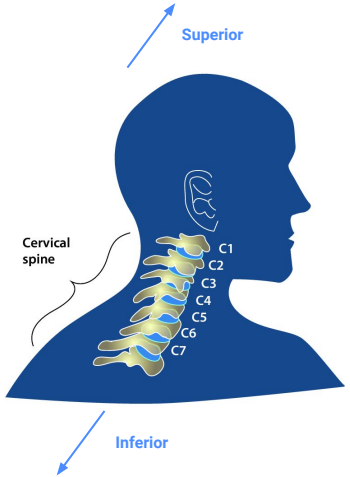
Error Analysis

* The ensemble model typically overpredicts the number of fractures since we prioritize not missing positive fractures (F2)

* **Inferior fracture detection**:
Actual fracture is in C1, but we predict fractures in C2 and C3 instead

* **Superior fracture detection**:
Actual fracture is in C7, but we predict fractures in C1 and C2 instead

* The model preferentially predicts all inferior vertebrae 6% of the time and predicts all superior vertebrae 1% of the time



Model predicts mostly 3.5 - 5.5 fractures on average regardless of actual vertebrae fracture count