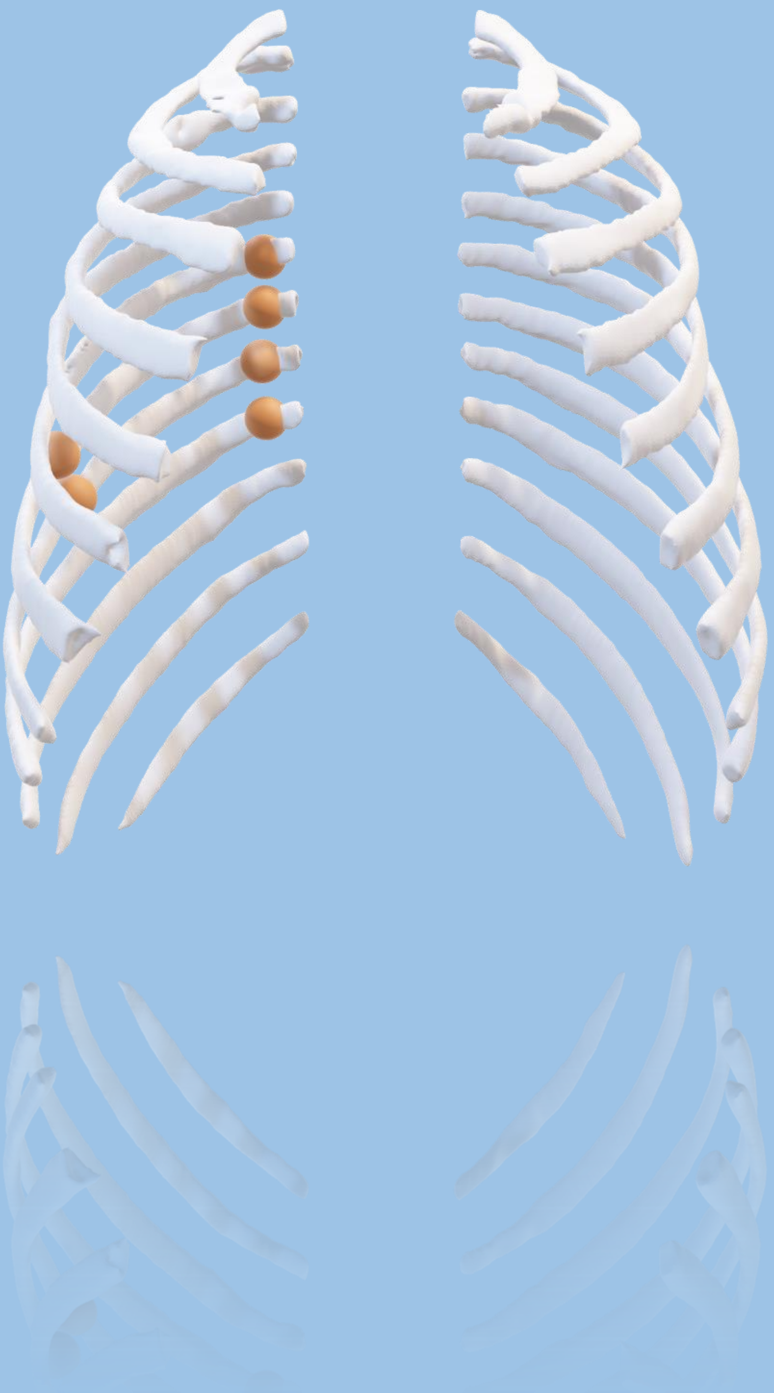
Deep Learning-Based Automatic Detection and Classification of Rib Fractures from CT scans



MSc Technical Medicine thesis

Noor Borren

Universiteit Leiden

TU Delft Delft University of Technology

ERASMUS UNIVERSITEIT ROTTERDAM

Deep Learning-Based Automatic Detection and Classification of Rib Fractures from CT scans

by

Noor Borren

Student number : 4450450

26th of September, 2023

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Department of Trauma Surgery and Radiology & Nuclear Medicine, Erasmus MC
January 2023 – September 2023

Supervisor(s):

M.M.E. (Mathieu) Wijffels, MD, PhD, Erasmus MC
T. (Theo) van Walsum, Eng, PhD, Erasmus MC

Thesis committee members:

T. (Theo) van Walsum, Eng, PhD, Erasmus MC (chair)
M.M.E. (Mathieu) Wijffels, MD, PhD, Erasmus MC
M. (Maarten) van der Elst, MD, PhD, TU Delft

Table of Contents

Preface

Before you lays not only the culmination of my 9.5-month thesis journey but also the product of seven eventful years of studying the interesting and intriguing field of Clinical Technology. Reflecting on this period fills me with happiness as it has been a profound learning experience in the clinical, technical, and personal dimensions of my life.

During my bachelor's thesis, I became increasingly intrigued by Machine Learning in the medical field. This is a topic I could further explore in the master Technical Medicine. Although COVID-19 made parts of studying harder, it did give me the possibility to combine my study with a side-job for a medical AI startup. Then, the medical-technical internships in the second year of our master's program allowed me to both deepen and broaden this knowledge. These internships also offered the unique opportunity to experience the medical world up close, in all its diversity. For my thesis project, I wanted to find a place where I could use and improve my Machine Learning skills within an engaging department. Therefore, the combination between the Trauma Surgery and the Biomedical Imaging Group was perfect!

I want to extend my heartfelt gratitude to my supervisors Mathieu and Theo for all their guidance along the way. Mathieu, from the first moment you called me after I send you an inquiry email, your enthusiasm has been contagious. You gave me opportunities throughout my thesis, such as presenting for different groups and guiding another student, which I really appreciated and enjoyed. During the surgeries and outpatient clinic moment(s), your feedback was consistently insightful which helped me in becoming a better Clinical Technologist. Theo, thank you for always taking extra time in your busy schedule whenever I needed it. Your expertise, valuable insights and kindness always made our discussions enjoyable and really helpful for my project. I also want to express my gratitude to all the researchers in the Trauma Surgery and Biomedical Imaging Group who contributed along the way. More specifically, I want to thank Alex. You sat literally next to me the entire length of the journey and was always open for discussions about my topic or simply for a cup of coffee.

Lastly, I want to express my appreciation to my family, friends and roommates. Thank you for being there and your support!

Now, let's see what the future holds.

Noor Borren
September 2023, Rotterdam

List of Abbreviations

AI	Artificial Intelligence
Bbox	Bounding box
CoM	Centre-of-Mass
CSV	Comma Separated Values
CT	Computed Tomography
CWIS	Chest Wall Injury Society
DCRibFrac	Detection and Classification of Rib Fractures
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
GUI	Graphical User Interface
IoU	Intersection-over-Union
JSON	JavaScript Object Notation
ML	Machine Learning
NIfTI	Neuroimaging Informatics Technology Initiative
SSRF	Surgical Stabilization of Rib Fractures

Abstract

Introduction: Trauma-induced rib fractures are a common injury, affecting millions of individuals globally each year. Although anteroposterior thoracic radiographs are part of the standard posttraumatic screening, the most sensitive modality, and therefore golden standard for diagnosing rib fractures, is computed tomography (CT). Still, between 19.2% and 26.8% of rib fractures are missed. Another problem encountered in rib fracture treatment management is the large interobserver variability on their taxonomy. This thesis aims to automate rib fracture detection and improve consistency in their classification by developing a Deep Learning (DL) model, using CT data.
Methods: The rib fractures were classified according to the Chest Wall Injury Society (CWIS) taxonomy, evaluating rib fracture's type, displacement and location. Furthermore, the ribs were numbered from 1 up to and including 12 from cranio-caudal direction. For the detection and three CWIS labels, three classification models of the nnDetection framework were trained. The rib numbering consisted of a trained nnU-Net segmentation model. The four models were combined to obtain the proposed DCRibFrac model.
Experiments and results: The dataset is composed of retrospectively collected and anonymized CT scans of 100 randomly selected patients (1010 rib fractures) who were admitted to the Erasmus MC following blunt chest trauma. On the internal test set, DCRibFrac achieved a detection sensitivity of 77%, precision of 79%, and F1-score of 78%, with a mean false-positives per scan of 2.26. The type labels had the lowest scores, with sensitivities between 17% and 90%. The displacement labels had sensitivities between 43% and 91%. The location labels had the highest scores, with sensitivities between 88% and 96%. The rib number was correct in 72% of the rib fractures when wrong segmentations were excluded.
Conclusion: The proposed DL model automates acute rib fracture detection and reaches a sensitivity that is on par with clinicians. This model is the first, to the authors' knowledge, to incorporate the CWIS taxonomy and shows its potential for achieving a consistent classification.

1. Introduction

1.1 Clinical motivation

The Dutch healthcare system is facing challenging times, with a predicted personnel shortage of 135.000 by 2031, primarily affecting hospitals and elderly care facilities [1]. Maintaining the status quo is no longer feasible and change is needed. In the meantime, medical technology is starting to play a bigger role and holds promise to alleviate part of this pressure [2].

Amongst others, an area with great potential for (technological) improvement is rib fracture treatment. Trauma-induced rib fractures is a common injury, affecting millions of individuals globally each year, with a prevalence of 10-40% in trauma patients [3-6]. Common causes are high-energy trauma (i.e. falling from height or car accidents) as well as lower-energy traumas in older patients (i.e. fall from standing height) [7]. Rib fractures, in general, lead to high morbidity and causes mortality in combination with other conditions, such as hemothorax, pneumothorax, extremity fractures and injuries to soft tissue [3]. Moreover, inadequate pain control can lead to respiratory complications such as pneumonia, due to impaired coughing and insufficient breathing [8].

While most rib fractures are managed conservatively, there is growing interest in surgical stabilization of rib fractures (SSRF) as an alternative approach [9]. For patients with a flail chest, characterised by three or more consecutively fractured ribs at multiple locations per rib, SSRF is increasingly recognised, based on evidence, as the preferable treatment option [10, 11]. For these patients, SSRF can result in benefits such as a shorter hospital stay, shorter duration of mechanical ventilation, decreased pneumonia risk and improved cost-effectiveness compared to conservative treatment [12-15]. Another group that can benefit from surgical fixation, are the patients with a non-union (non-healing) of their rib fracture a few months after the trauma [16]. However, there are no internationally acknowledge guidelines for this patient group yet [17], and there may be other types of rib fractures that could benefit from SSRF.

One factor for the limited implementation of SSRF guidelines could be the inconsistency in rib fracture classification, which makes communication in clinical practice and scientific research difficult. The ideal classification system should predict outcomes, has an optimal inter- and intra-observer agreement and covers all fracture entities, enabling a consistent and non-confusing discussion. Up to 2020, there was no standardised and widely used rib fracture classification system yet. The Chest Wall Injury Society (CWIS) addressed this problem by a Delphi consensus study [18]. However, literature shows a significant interobserver variability among clinicians using this system, emphasizing the need for a more robust and reliable computed tomography- (CT-)scoring approach [19]. Another limiting factor is the number of missed fractures. In the primary survey of trauma patients, an anteroposterior thoracic radiograph is the standard of care. However, the sensitivity is low for rib fractures and 50-80% remain undetected [20-22]. Therefore, the golden standard for diagnosing rib fractures is computed tomography (CT). This does improve the sensitivity but missed rib fractures are still common, ranging from 19.2% to 26.8% [19, 23-25].

Artificial Intelligence (AI) has the potential to address these challenges by improving the detection sensitivity and classification consistency of rib fractures, offering valuable insights and guidance for optimal treatment strategies. This benefits the healthcare system by improving patient outcomes and cost-effectiveness while alleviating the burden on the healthcare system.

1.2 Related work

In recent years, an uptake of interest for Deep learning (DL) models, a subset of AI, in rib fracture detection can be seen, with one study published on this topic in 2018 [26], and in the year of 2022, already eight studies were published [27-34]. (For a brief technical introduction, please refer to Appendix A.) Not surprisingly, DL models hold promise in enhancing rib fracture detection and establishing a more consistent classification compared to clinicians. (Table 1)

In general, studies prioritised a high sensitivity over a low false-positive rate. This emphasis on sensitivity may be attributed to the clinicians' capability to be more specific than sensitive in their assessments. Consequently, these models are often designed to work collaboratively with clinicians allowing them to complement each other. In clinical practice, the achievable sensitivity of a model will therefore partly depend on the time needed for a clinician to review the positively labelled spots and exclude false-positives.

Although achieving high sensitivities, many studies failed to report on key dataset characteristics (i.e., data selection methods, pixel sizes, patient demographics), the definition of false-positives and testing on an external test set. This makes it difficult to determine how robust these high sensitivities are in context of real-world datasets. Furthermore, existing classification systems are inconsistent and not up-to-date with the CWIS taxonomy standards. (Table 1) These reported classifications offer limited clinical value as they have no treatment consequences.

Table 1: Results of 16 studies describing the detection of rib fractures and reporting sensitivities. Nine studies also classified rib fractures for which only the studies with the lowest and highest scores are presented. All studies were published before February of 2023.

	Sensitivity	**F1-score**	**False-positives per scan**
Detection	0.645-0.971 [25, 26, 28, 29, 31-42]	0.652-0.940 [25, 28, 32, 37, 39, 41, 42]	0.14-2.71 [29, 31, 33-35]
Classification			
• Acute	0.68-0.92 [33, 40]	0.85, 0.87^{+} [33, 42]	
• Healing	0.86-1 [33] [41, 50]	0.82, 0.86^{+} [33, 42]	
• Old	0.59-0.97 [33, 40]	0.77-0.94 [33, 42]	
• Displaced	0.92-1 [34, 35]	0.89 [32]	
• Non-displaced	0.74-0.85 [31, 34]	0.78 [32]	
• Buckle	0.58-0.83 [35, 37]	Not reported	

$^{+}$*Only two articles reported F1-scores.*

Considering the challenges and opportunities presented, the primary objective of this thesis is to automate and improve rib fracture detection and set a standard for the taxonomy of rib fractures from CT to promote clear and standardised communication, aid in predicting clinical outcomes and provide support for scientific research.

1.3 Contributions

The key contributions of this thesis are:

- A new open-source labelling software for the CWIS classification;
- An ensemble model where four DL models are combined to detect and classify rib fractures from CT scans. The model's detection sensitivity aligns with those achieved by clinicians and other DL models;
- Assessment of using non-standardised CT scans derived from routine clinical practice.

This thesis is divided into six chapters. *Chapter 1* gives a general introduction to this thesis and explains the relevance in broader (social) context. A brief summary of literature is given on the topic of DL-based detection and classification of rib fractures. *Chapter 2* covers the methods developed in this project to obtain the proposed DCRibFrac model. *Chapter 3* presents the experiments and results where the ground truth establishment is evaluated and results of DCRibFrac, on the internal test set, are presented. *Chapter 4* discusses and interprets the findings, presents limitations and gives recommendations for future research. Lastly, *Chapter 5* summarizes the main findings.

2. Methods

This project introduces a model designed for the automatic detection and classification of acute rib fractures (DCRibFrac) from CT scans. Section 2.1 introduces the three CWIS classifications used in this project. This classification system is extended by incorporating the rib number, on which a fracture is located, as a fourth label. Section 2.2 describes the method's technical details with three more detailed sections. Section 2.2.1 covers the implementation of the nnU-Net framework. Similarly, Section 2.2.2 provides insights into the nnDetection framework. Lastly, section 2.2.3 presents the integration of these models to form DCRibFrac. The code can be accessed at https://gitlab.com/radiology/igit/msc-projects/noor-borren/dcrib.

2.1 CWIS taxonomy of rib fractures

The CWIS set a standard for the taxonomy of rib fractures by conducting a Delphi consensus study. This classification is based on three characteristics; the fracture line's type, fracture's displacement and the fracture's location on the rib bow. The fracture line's type could be described with the labels (Figure 1) [18]:

- Simple: characterised by a single fracture line that runs through the rib;
- Wedge: characterised by a single fracture line with an additional line that does not run through the entire width of the rib. This creates a chipped-off fragment;
- Complex: characterised by two or more fracture lines that extend across the entire rib width, resulting in the presence of one or more fragments.

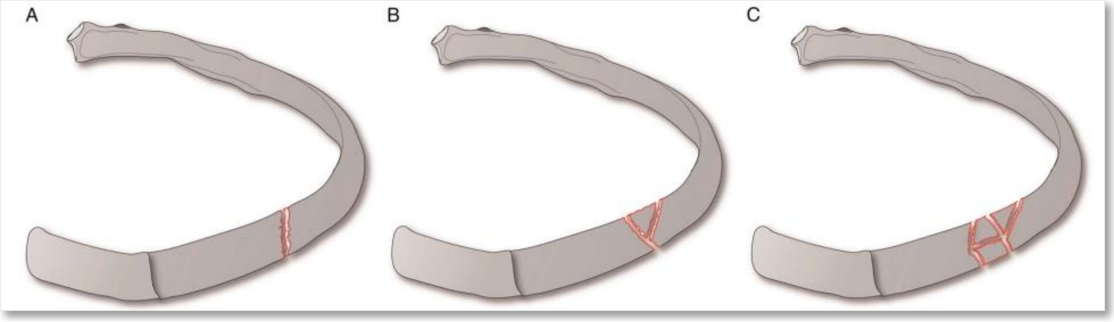


Figure 1: Schematic representation of a A) simple; B) wedge and C) complex fracture [18].

The fracture displacement could be described as (Figure 2) [18]:

- Undisplaced: more than 90% cortical contact;
- Offset: between no cortical contact and 90% contact;
- Displaced: no cortical contact.

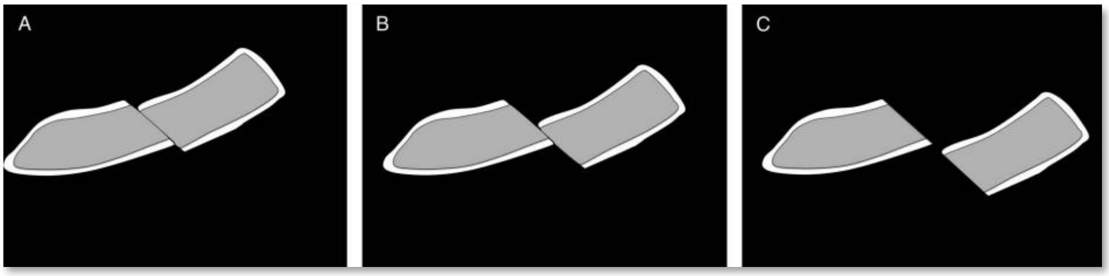


Figure 2: Schematic representation of A) an undisplaced; B) an offset and C) a displaced fracture [18].

In the CWIS study, there was consensus reached on defining three anatomical sectors for the fracture's location on the rib bow. This encompasses only the bony part of the rib, thus excluding the chondral area, medial to the costochondral joint, and excludes the paravertebral section, medial to the costotransverse joint. (Figure 3) However, there was no consensus on where the boundaries of these three sectors should be. One of the proposals was to define the sectors with angles from the mid-thoracic point [18]. This point is defined as the middle of the line drawn between the linear alba and the posterior aspect of the vertebral spinous process. This method is the most objective and describes the fracture location as (Figure 3):

- Anterior: defined as the angle between 0° and 60°;
- Lateral: defined as the angle between 60° and 120°;
- Posterior: defined as the angle between 120° and 180°.

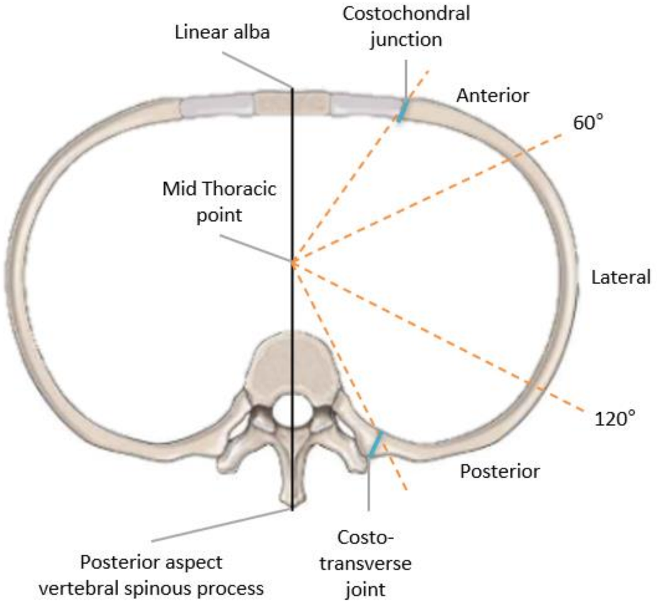


Figure 3: Schematic overview of the third proposal of the CWIS standard for defining the anatomical sectors of the rib.

In this project, a fourth classification was added for denoting the rib number on which the fracture is located.

2.2 Methods overview

This project proposes a model designed for the detection and classification of acute rib fractures (DCRibFrac). Figure 4 provides an overview of DCRibFrac's pipeline, which consists of three main components. First, to detect the rib fractures and assign the three CWIS labels, the nnDetection framework is employed [43]. Second, to determine the rib on which the fracture is located, the nnU-Net framework is utilised [44]. Lastly, the models are combined in post-processing steps to generate the final output, the detection of a rib fracture with four labels. To provide a comprehensive understanding, the following paragraph will explain nnU-Net first, as nnDetection builds upon the concept introduced by nnU-Net.

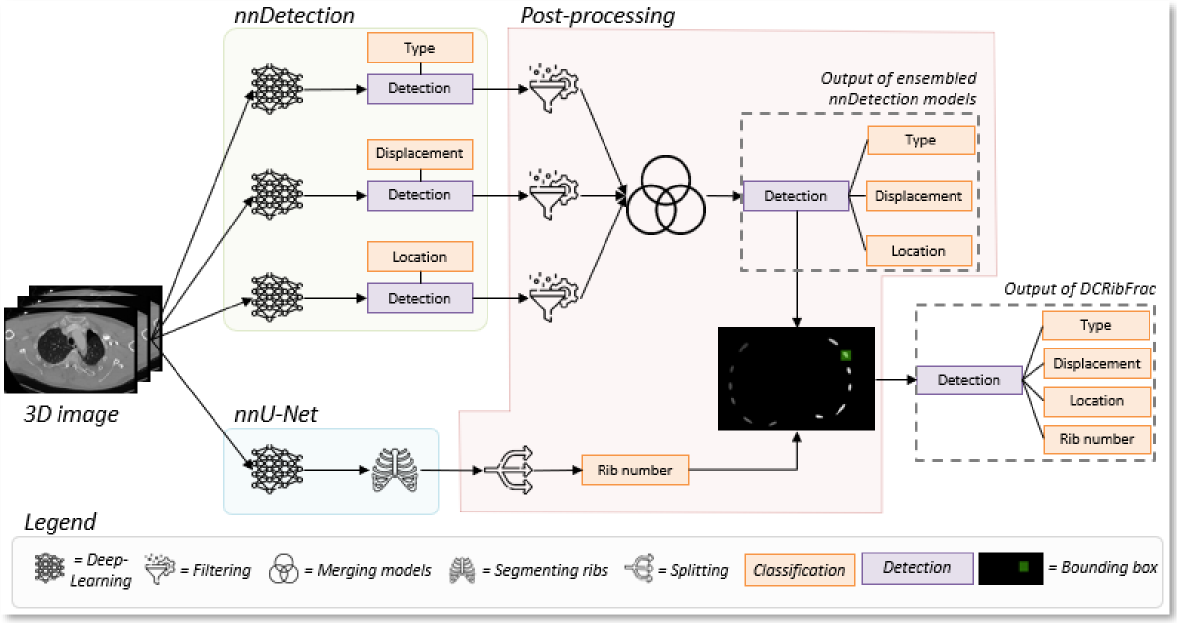


Figure 4: Overview of DCRibFrac's pipeline for the automatic detection and classification of rib fractures. The CT scans of the patient are inputs for four DL-models, each responsible for one of the classifications. The results are post-processed for the final result; a detected rib fracture with four labels.

2.2.1 nnU-Net

In this project, rib segmentations are needed to determine the rib on which the fracture is located. To accomplish this, the state-of-the-art medical segmentation framework nnU-Net is utilised. This framework was chosen because it handles the entire pipeline, including preprocessing, training, and post-processing, autonomously across a wide variety of segmentation tasks. This simplifies the implementation, ensures reproducibility and eliminates laborious hyperparameter tuning. nnU-Net can train three networks, all of which are based on the original U-Net architecture [44]. In this project, the 3D low resolution U-net was trained, as the segmentation task does not require a high level of precision. For a more in-depth understanding of this framework's training process, refer to Appendix B.

After the ribs are segmented, they are split into individual ribs to facilitate the determination of the rib number. This is achieved by the following steps:

1. Segmentation improvement – the segmented ribs are refined using the morphological operation *opening*, to remove small artifacts, and *closing,* to fill holes. The kernel size of 2x2x2 and the number of iterations, one and three respectively, were empirically determined on a subset of the training set.
2. Rib splitting – the segmented ribs are split into individual ribs using the *SciPy* library's *label* function. Only regions larger than 500 voxels are retained using the *remove_small_objects* function. If the region count is less than 22 (for patients without a 12th rib), an error message is outputted. In cases of severe rib displacement due to a fracture, which leads to a single rib consisting of two regions, the rib numbering is influenced. Therefore, if the number of regions exceeds 24, three additional iterations of the morphological operation *dilation* are performed with the same kernel size in an attempt to merge these regions. If the region count is still higher than 24, an error message is outputted. The error messages indicate that the numbered rib segmentations cannot be utilised.
3. Laterality definition - the centre-of-mass (CoM) is calculated for each region with the function *regionprops* of the *skimage* library. By computing the mean of all CoMs, the regions are split in left and right by having their CoM x-coordinate higher or lower than that mean, respectively.

4. Order definition - two lines are fitted through the left and right CoM points with the function *Line.best_fit*. Each CoM point is then projected onto the line to determine the order in which the CoMs should be labelled. The rib numbers are assigned to the regions associated with the CoMs, starting with one for the first right rib and ending with 24 for the left 12th rib.

The coupling of rib fractures with these numbered rib segmentations is addressed in section 2.2.3.

2.2.2 nnDetection

nnDetection is a framework for semantic segmentations, which can also be utilised as an object detector, and follows the same self-configuring strategy as nnU-Net [43, 44]. Part of the framework consists of Retina U-Net. This merges the object detector RetinaNet, also used by one of the top five models in literature [33], with a U-Net architecture. Retina U-Net uses the Feature Pyramid Network, which extracts features at different scales, enabling analysis of objects with varying sizes. To enhance the classification without introducing unnecessary complexity, Retina U-Net uses two additional layers dedicated to the classification task solely (P1 and P0 in Figure 5). This optimizes the model's classification without making the model inefficient [45].

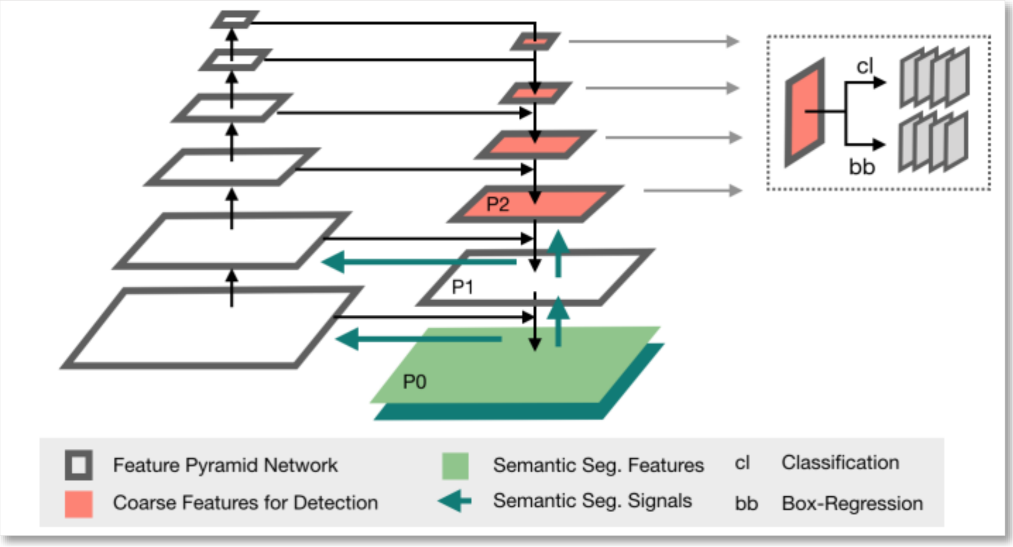


Figure 5: A schematic 2D representation of Retina U-Net [45].

The nnDetection framework handles only one classification at a time, for example the fracture's type. Therefore, the framework is used for training three separate models, each dedicated to one of the CWIS classifications. The detection of rib fractures are indicated by spherical objects (referred to as "blobs" throughout this project), with their midpoint corresponding to the middle of the rib fracture. In separate files, information regarding the label information are given. Additional information on how these files are constructed and an example are given in Appendix C.

A trained nnDetection model produces bounding box coordinates, accompanied by probability scores, and one of the three labels for each detection. (Figure 4) Consequently, one rib fracture detection with the complete CWIS classification can have three different bounding boxes. To address this, an ensemble approach is used by combining the results from the three models. First, each individual nnDetection model is filtered to remove overlapping bounding boxes where the ones with the lowest probability scores are discarded. Two bounding boxes are considered overlapping if their Intersection-over-Union (IoU) score exceeds 50%. IoU is calculated by:

$$IoU = \frac{bbox_1 \cap bbox_2}{bbox_1 \cup bbox_2}$$

where $bbox_1$ and $bbox_2$ are bounding boxes, ∩ indicates the intersection and ∪ the union of the boxes. Secondly, the models are merged with the requirement that there is an overlap of bounding boxes from at least two models, indicating the detection of a potential rib fracture by at least two models. These bounding boxes are combined through union. In cases where only two of the three models overlap, one of the labels cannot be assigned. In such instances, a label 'unknown' is assigned.

2.2.3 Combining classification results

The last step combines the merged nnDetection models with the numbered rib segmentations. This is done by converting the bounding box coordinates into a binary label map. Subsequently, the binary label map and the numbered rib segmentations are compared. Bounding boxes that do not overlap with the numbered rib segmentations are discarded and for the others, the fourth label, indicating the rib number, is assigned. This concludes the DCRibFrac model. (Figure 4)

3. Experiments and Results

3.1 Data

The data were retrospectively collected and anonymized CT scans of 100 randomly selected patients (1010 rib fractures) who were admitted to the Erasmus MC following blunt chest trauma. An ethical approval waiver was granted by the Institutional Review Board under the reference number MEC-2023-0039. Further details about the data selection process can be found in Appendix D.

In 98 of the 100 patients, the scans were acquired with the Siemens SOMATOM Definition Edge scanner. These scans had an in-plane image size of 512x512 pixels, with variations in the Z-direction between 216 and 1513 slices. Additional patient and image characteristics are provided in Table 2.

The labelling process was conducted in software specifically developed for this project within the Free MeVisLab SDK and performed by a single researcher (NB) [46]. For additional information regarding this software and the labelling procedure, refer to Appendix E. Ground truth label distribution in the dataset is presented in Table 2 and visualised in Appendix F.1, showing a similar distribution between the left and right sides.

Table 2: Dataset characteristics of patients and CT images

Variables	Internal training set	Internal test set
No. patients	81	19
No. fractures (%)	803 (80)	207 (20)
Mean age (range)	55 (20-86)	58 (37-86)
Gender, female:male (%)	19:62 (23:77)	3:16 (16:84)
No. CT slice thickness = 2 mm (%)	51 (63)	12 (63)
No. CT slice thickness = 1 mm (%)	8 (10)	0
No. CT slices thickness < 1 mm (%)	22 (27)	7 (37)
No. CT pixel spacing > 0.9 mm (%)	16 (20)	5 (26)
No. CT pixel spacing 0.7< x <0.9 mm (%)	49 (60)	10 (53)
No. CT pixel spacing < 0.7 mm (%)	16 (20)	4 (21)
No. fractures for Type (%)		
- Simple	597 (74)	145 (70)
- Wedge	138 (17)	40 (19)
- Complex	68 (9)	22 (11)
No. fractures for Displacement (%)		
- Undisplaced	506 (63)	103 (50)
- Offset	195 (24)	79 (38)
- Displaced	102 (13)	25 (12)
No. fractures for Location (%)		
- Anterior	159 (20)	21 (10)
- Lateral	364 (45)	116 (56)
- Posterior	280 (35)	70 (34)

3.2 Implementation

Training of the nnDetection and nnU-Net models was performed on the GPU cluster of the Erasmus MC with the 2090 Ti 11GB and Nvidia A40 48GB GPU’s. A random 80-20 train-test split was used on the number of fractures with the corresponding patients. To ensure class balance for the minority class, stratified sampling was performed. Furthermore, during training, a five-fold cross-validation strategy was implemented for each model. The post-processing of DCRibFrac was developed in Python version 3.7.

3.3 Evaluation metrics

The Krippendorff's Alpha and Fleiss' Kappa were selected as tests to assess interobserver agreement on a subset of the training set. Krippendorff's Alpha is suitable for multiple observers. It has a scale from 0 to 1, accommodates categories, handles missing data and takes both the level of agreement and disagreement into account [47, 48]. Fleiss' Kappa is an extension of Cohen's Kappa and can be used for multiple observers. This metric assesses the level of agreement but does not account for missing data. It has a scale from -1 to 1, where -1 shows a lower agreement than would be expected by chance and positive values indicate better agreement than expected by chance [49]. Cohen's kappa was used to assess the intra-observer variability which also scales from -1 to 1 and can be interpreted as the Fleiss' kappa [50]. (Table 3)

Table 3: Interpretation of interobserver agreement tests [19].

Krippendorff's Alpha and Kappa values	Interpretation
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-0.90	Strong
> 0.91	Almost perfect

Evaluation of DCRibFrac was performed at the per-fracture level. The quantitative evaluation for the rib fracture detection consisted of the sensitivity, precision, F1-score, number of false positives per scan (FPPS) and precision-recall curve. A false-positive was defined as a 3D bounding box that did not overlap with the midpoint of a blob in the ground truth, where each blob could only be coupled to one bounding box. For the classification, sensitivity, precision and confusion matrices were utilised to present the results.

The qualitative evaluation, important because there is a high chance of missed rib fractures in the ground truth, was performed in 3D Slicer [51].

3.4 Results

In this section, the experimental results are presented. In section 4.4.1, the inter- and intra-observer agreement on a subset of the test set is assessed. In section 4.4.2, two dataset sizes are compared to evaluate the influence on training of a nnDetection model. In section 4.4.3, the ensembled DCRibFrac models' detection performance with associated thresholds are shown to gain a better understanding of the model's behaviour. In section 4.4.4, DCRibFrac is utilised to detect and classify rib fractures in the test set to assess the overall performance. Additionally, the results of the qualitative assessment are presented.

3.4.1 Inter- and intra-observer agreement

To assess the ground truth in the current dataset, an inter- and intra-observer agreement study was performed on a subset of 50 rib fractures from the test set. Two researchers (NB, MvD) and a Trauma surgeon (MW) labelled the rib fractures according to the CWIS taxonomy. For the intra-observer agreement test, one researcher (NB) labelled the rib fractures twice with a one-month interval.

In total, there were 58 rib fractures noted by the three observers of which 44 rib fractures were seen by all observers. The results of both studies are seen in Table 4 and 5 where a slight discrepancy between the Fleiss' Kappa and Krippendorff's Alpha measures for the interobserver agreement can be observed. Notably, the location classification achieved the highest agreement among all observers. Furthermore, the intra-observer study yielded to the higher agreements than the inter-observer study.

Table 4: Interobserver agreement for the CWIS classification of rib fractures

Label	Fleiss' Kappa (95% CI)	Interpretation	Krippendorff's Alpha (95% CI)	Interpretation
Type	0.23 (0.11-0.38)	Fair	0.34 (0.20-0.47)	Fair
Displacement	0.02 (-0.01-0.11)	Slight	0.46 (0.32-0.58)	Moderate
Location	0.14 (0.07-0.27)	Slight	0.63 (0.48-0.76)	Substantial

Table 5: Intra-observer agreement for the CWIS classification of rib fractures

Label	Cohen's kappa (95% CI)	Interpretation
Type	0.71 (0.52-0.89)	Substantial
Displacement	0.80 (0.65-0.95)	Substantial
Location	0.84 (0.69-0.99)	Strong

3.4.2 Influence of dataset size

To assess the impact of training with more data on the nnDetection models, an experiment with two dataset sizes was performed. The first training existed of 35 patients with 293 rib fractures. Then, training was done with the entire training dataset of 81 patients with 803 rib fractures.

Training with a small dataset led to more overfitting, seen by the validation curve starting to increase after just a few epochs and the larger gap between the validation and training curves. The addition of more data reduced overfitting, seen by the plateauing validation curve with a smaller gap between the training and validation curve. (Figure 6)

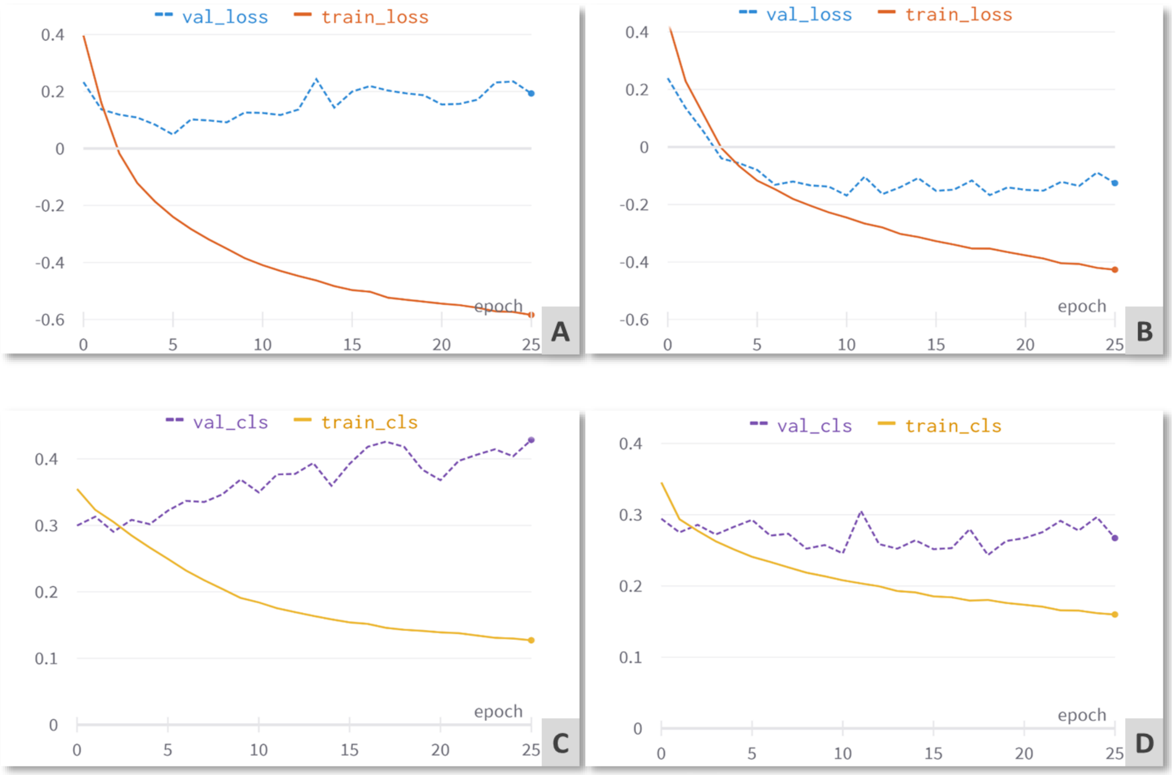


Figure 6: A,B) The train and validation loss of training with 35 and 81 patients, respectively; C,D) The classification loss for training with 35 and 81 patients, respectively.

3.4.3 Performance of ensembled nnDetection models on the validation set

To gain further insight into the detection performance of the ensembled nnDetection models on the validation set, following cross-validation, a precision-recall curve was generated with corresponding thresholds. (Figure 7) Also, the probability score threshold could be chosen to achieve a detection sensitivity of 82%. This sensitivity surpasses the range typically observed among clinicians, which falls between 73.2% and 80.8% [19, 23-25].

The precision-recall curve illustrates that DCRibFrac cannot reach a sensitivity of one, as the highest sensitivity, at a threshold of zero, is 0.95. A threshold of 0.448 was chosen to achieve the aimed sensitivity.

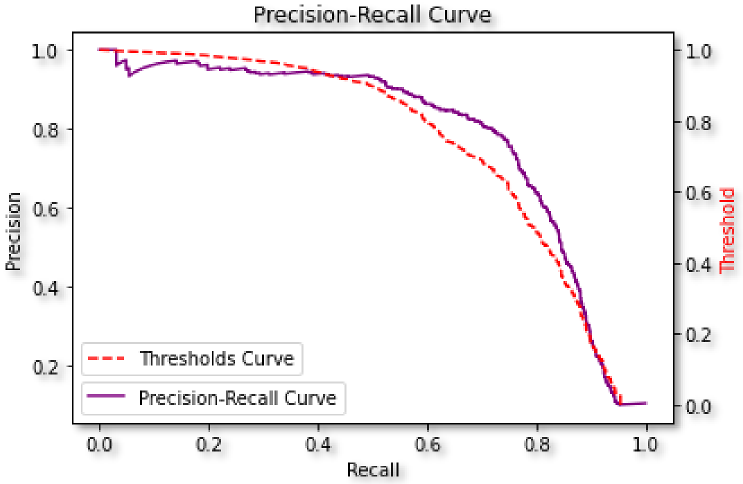


Figure 7: Precision-Recall curve with thresholds of DCRibFrac's detection performance on the validation set.

3.4.4 Performance of DCRibFrac on the test set

To evaluate the performance of DCRibFrac on unseen data, the detection and classification of rib fractures was performed on the test set. The total time for running the DCRibFrac pipeline for a patient was between 30-90 minutes depending on the image size.

DCRibFrac achieved a detection sensitivity of 77%, precision of 79% and F1-score of 78%, with a mean FFPS of 2.26 on the test set. All classification labels were assigned, as there were no cases with only two overlapping nnDetection models. An example of a correctly detected rib fracture can be seen in Figure 8A.

The qualitative evaluation for the detection of rib fractures revealed two additional true-positives that were not labelled in the ground truth. (Figure 8B) False-positives mostly resulted from old fractures, indicated by callus (Figure 8C), paravertebral fractures and cysts. The missed fractures were mostly simple fractures with a small interruption of the cortical bone and severely dislocated fractures with a translation in the z-direction. (Figure 9) More examples of false-positives and -negatives can be seen in Appendix F.2.

In terms of the type classification, complex fractures had a sensitivity and precision of 17% and 30%, wedge fractures 30% and 42%, and simple fractures 90% and 75%, respectively. For the displacement classification, displaced fractures had a sensitivity and precision of 43% and 75%, offset fractures 78% and 79%, and undisplaced fractures 91% and 83%, respectively. The location classification showed the highest performance, where posterior fractures had a sensitivity and precision of 96% and 84%, lateral fractures 88% and 95%, and anterior fractures 88% and 88%, respectively. (Appendix F.3, Table 7) These were compared with the results of the interobserver agreement study and are presented in Appendix F.3 (Table 8,9,10). Interestingly, for the two labels that were most distinct in the location and displacement classifications, namely the anterior-posterior and undisplaced-displaced labels, there were no wrong classifications. (Figure 11) The numbered rib segmentations were incorrect for six patients, indicated by the error messages. (Figure 10A,B) This resulted in 72% correctly classified rib number labels when these six patients were excluded.

The qualitative evaluation of the numbered rib segmentations revealed the presence of eleven wrong segmentations instead of six. There was no error message for these five patients because the merging and splitting of ribs were in equilibrium, with 11 or 12 regions observed on each side. (Figure 10C,D) In the correct segmentations, the bounding box sometimes overlapped with a more cranial rib segmentation first, resulting in a wrong rib number.

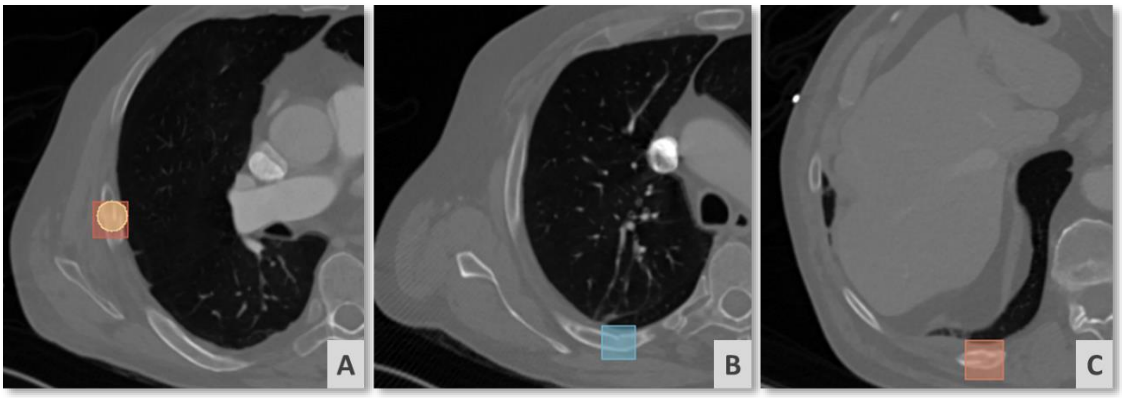


Figure 8: Three DCRibFrac detections in one patient at different levels. Squares indicate detections and circles represent the ground truth; A) A true-positive as the circle overlaps with the square; B) A true-positive which was not present in the ground truth; C) A false-positive because of an old, thus not an acute, fracture as indicated by the callus around the fracture.

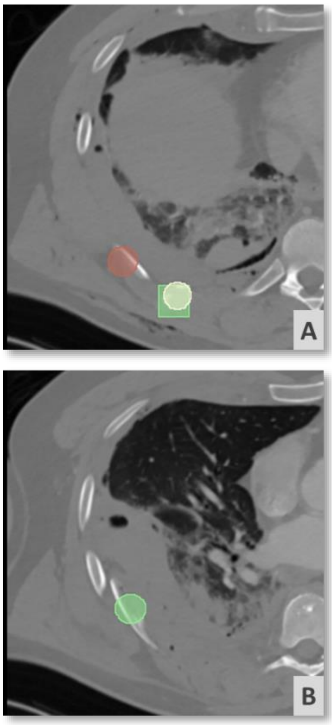


Figure 9: Example of two blobs for indicating one severely dislocated rib fracture. A) Red, left blob on slice 238; B) Green blob on slice 255.

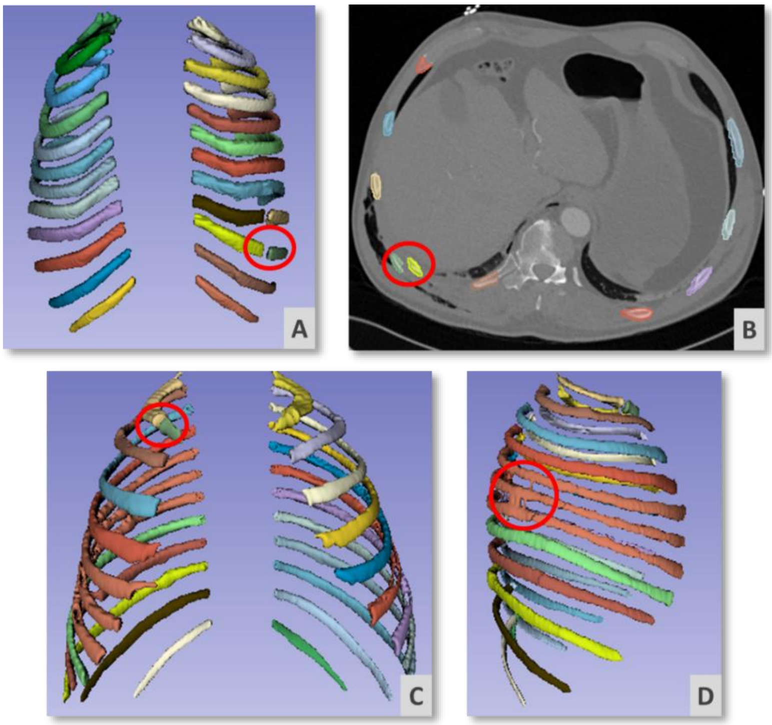


Figure 10: Examples of wrongly segmented ribs with the regions of interest indicated by the red circles. A) Posterior view of a 3D segmented model where the fracture splits the segmentation in two; B) Axial slice of the fracture encircled in A. C, D) Anterior and lateral view of a fracture causing the rib to be split in two and dilation that caused the merging of three ribs, respectively. The counting did not result in errors as the total count for the right side was still 11.

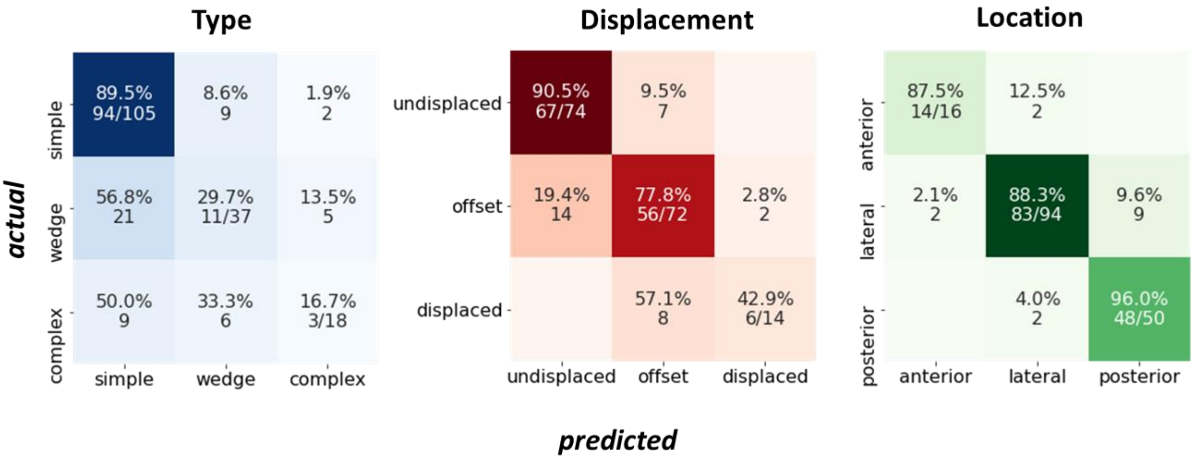


Figure 11: Confusion matrices of the labels in the test set. The diagonal represents the correctly predicted labels.

4. Discussion

The aim of this project was to automate and improve the acute rib fracture detection and set a standard for the taxonomy of rib fractures. The proposed model, DCRibFrac, was implemented by combining three nnDetection models and one nnU-Net model for the detection of rib fractures and classification, according to the CWIS taxonomy. Although the rib fracture detection was automated, the sensitivity yielded to 77%, which is on par with that of clinicians but did not surpass them. The classification demonstrated commendable sensitivity and precision scores for the displacement, except for the displacement label, and location classifications. However, DCRibFrac could not set a proper standard for the rib fracture type classification. The last label, the rib number, had favourable results for patients with minor displaced rib fractures but did not work on severely dislocated ribs. In general, DCRibFrac shows potential for improved detection sensitivity and a consistent classification of acute rib fractures from CT scans but further refinements are needed.

The achieved detection sensitivity aligns with the findings reported in the literature, despite using a comparatively smaller dataset. (Table 6) From the loss graphs of the two dataset sizes, overfitting is still evident in the larger dataset size, used for training DCRibFrac. As the parameters of nnDetection are difficult to manually adjust, addition of a larger dataset is needed to decrease overfitting and improve both the detection and classification. More specifically, adding data of simple fractures with a minimal cortical interruption and severely dislocated fractures with a translation in the z-direction should be incorporated. Moreover, to decrease the FPPS, training with data of old fractures could lead to improved performance as the model will learn to not detect these types of rib fractures. The publicly available RibFrac dataset, primarily existing of old fractures, is suited to incorporate in the training dataset to decrease the FPPS [52].

Table 6: Comparison with other DL methods that noted both the sensitivity and false-positives per scan (FPPS) for detecting rib fractures. *Denotes a missing value

	Patients, fractures	Sensitivity	FPPS
Zhou et al. [33]	640, 2853	94.9%	0.17
Niiya et al. [29]	918, *	93.5%	1.9
Meng et al. [37]	8829, 34699	92.2%	0.14
Wang et al. [31]	9265, 43803	85%	0.35
Wu et al. [41]	10943, 9590	84.9%	0.764
Azuma et al. [34]	539, 4906	83.7%	2.71
Zhou et al. [42]	1049, 25054	83.2%	1.1
Zhang et al. [35]	3580, 15947	79.4%	0.43
DCRibFrac	**100, 1010**	**77%**	**2.26**
Weikert et al. [40]	159, 991	65.7%	0.16
Kaiume et al. [39]	39, 256	64.5%	1.1

The approach for determining the rib number was suboptimal. The choice of using CoMs for the rib order is only suited for ribs that are not segmented in multiple regions. Still, to improve this approach for non-severely displaced fractures, the morphological operation erosion should be applied to the bounding box label map. This makes sure that the bounding box has no overlap with more cranial rib segmentations. However, (severely) displaced ribs are common. Particularly, the algorithm’s output of splitting the ribs’ segmentation into single ribs was poor in an additional five patients despite the output remaining error-free. As a consequence, this model cannot be used unsupervised.

The Krippendorff's Alpha results of the interobserver agreement study closely resemble the results obtained in a large interobserver agreement by Van Wijck et al [19]. This underscores the complexity of establishing a uniform ground truth among observers. Notably, Fleiss' Kappa indicated an even weaker agreement than the Krippendorff's Alpha, as it did not consider the level of disagreement and label order (e.g., the difference between the anterior and posterior label is the largest) [48]. In this project, a single researcher established the ground truth. This potentially introduced bias as some classifications could be interpreted differently by other observers. Nevertheless, the substantial and strong consistencies observed in the intra-observer agreement suggest that random errors may be minimal.

This study has some limitations. First, the establishment of the ground truth was done by one observer. Therefore, the model was not only impacted by the data that it trained on but the calculation of the performance metrics is potentially not an accurate representation of the real sensitivity. This is also indicated by the additional true-positives encountered in the qualitative assessment. Therefore, all sensitivities, precisions and F1-scores should be interpreted cautiously. Secondly, the large label imbalance might have made the learning of the less frequent labels' characteristics more challenging. This is exemplified by the sensitivities observed for the categories with the least amount of rib fractures in the dataset: wedge, complex and displaced rib fractures. This shortcoming might be overcome by increasing the number of these rib fractures in the training dataset. However, the results of the interobserver studies might be an explanation for the poor results too. Potentially, the CWIS classification system is not suited for a clear definition of rib fractures as the distinction between the different groups may not be clear enough. Thirdly, there was no direct comparison with clinicians on the same dataset that the model was trained on. As a result, the generic percentages of missed fractures by clinicians might not accurately represent the performance on the current dataset. Still, the threshold for the minimal sensitivity was chosen based on the percentages reported in literature. Another limitation of this study was the absence of an external test set to evaluate the generalisability of DCRibFrac.

DL models are highly dependent on the data that they are trained on and tend to perform better with homogeneous datasets [53, 54]. However, real-world datasets are heterogeneous due to variations in CT scanners, slice thickness, imaging protocols and populations. In this project, the dataset remained as close to the real-world as possible by not excluding scans based on the available kernels, slice thickness, pixels spacing, imaging protocol or minor motion artefacts. However, the scans of the Erasmus MC were primarily acquired by one CT scanner and predominantly represented a male population. This could limit the clinical relevance in other populations or hospitals. Moreover, for a more comprehensive evaluation, paravertebral and costo-chondral fractures should be included too.

Another point for the clinical relevance is that DCRibFrac is not expected to yield to direct speed advantages in the radiologists' workflow of the acute setting. The current total reading time for a CT scan already falls within the time needed to run DCRibFrac, with rib fracture detection alone taking approximately 3 to 7 minutes [32, 41, 42]. Therefore, fastening improvements to DCRibFrac should be made to have a running time that falls within the total reading time. However, in the non-acute setting, DCRibFrac holds potential benefits. When optimised, it could improve the detection sensitivity and establish a consistent standard for rib fracture classification. This does not only enable clear and uniform communication but also has implications for predicting clinical outcomes and support further scientific research.

Future research should aim to improve DCRibFrac by exploring the previously mentioned approaches. Additionally, improvements for the numbered rib segmentations are needed. First, the problem with severely displaced rib parts should be resolved when making use of the CoMs. A potential approach is to use label information. If a displaced label is assigned to a fracture, local morphological operations such as dilation could be applied to merge the fragmented regions together. As an alternative to using CoMs for determining the order of the ribs, Lessmann et al. [66] presented a method to count the vertebrae, which were later used as inputs for counting ribs [55, 56].

Furthermore, it would be interesting to uncouple the detection and classification task. This could entail using nnDetection models for the detection of rib fractures and then experimenting with alternative methods for the classification. For instance, recent work done by Edamadaka et al. introduced a deterministic formula for calculating the percentage of displacement [57]. Similarly, the location on the rib could be mathematically calculated in degrees. Then, the classification could be made more objectively by setting boundary conditions for each label based on the numerical values.

A benefit of this approach is that the numerical values would be well suited as inputs for a prediction model too. For example, a model could predict whether rib fractures will naturally heal or remain fractured. In these prediction models, the input of clinical data, as described by Zhou et al. [69], could improve results even further [58]. This information could aid clinicians in making informed decisions about surgically fixating ribs that are unlikely to heal naturally. Preferably, this decision has to be made within the first 3-7 days after trauma which is too soon for basing this decision on radiological signs in the CT scans [59]. By integrating these approaches, future research could lead to clinicians delivering optimised care and improve patient outcomes in the management of rib fractures.

5. Conclusion

In conclusion, this thesis presents DCRibFrac, a DL model that aims to automate and improve acute rib fracture detection, and set a standard for their taxonomy from the golden standard CT. While the current detection sensitivity is comparable to that of clinicians, further refinement holds promise of surpassing the sensitivity reported for clinicians (73.2% - 80.8%) [19, 23-25]. Notably, this project is the first, to the authors' knowledge, to incorporate the CWIS taxonomy into a DL classification model and shows its potential for achieving a consistent classification. Future research should focus on improving and fastening the different components of DCRibFrac and advancing towards prediction models to increase the clinical added value for acute rib fracture patient management and treatment planning.

6. Bibliography

Reference List

1. Helder, C., *Kamerbrief: Nieuwe prognose verwachte personeelstekort*, M.v.L.Z.e. Sport, Editor. 2022.
2. GuptaStrategists, *Uitweg uit de schaarste*. 2022.
3. Ziegler, D.W. and N.N. Agarwal, *The morbidity and mortality of rib fractures.* Journal of Trauma and Acute Care Surgery, 1994. **37**(6).
4. Flagel, B.T., et al., *Half-a-dozen ribs: The breakpoint for mortality.* Surgery, 2005. **138**(4): p. 717-725.
5. Wijffels, M.M.E., et al., *Early fixation versus conservative therapy of multiple, simple rib fractures (FixCon): protocol for a multicenter randomized controlled trial.* World journal of emergency surgery : WJES, 2019. **14**: p. 38-38.
6. He, Z., et al., *The ideal methods for the management of rib fractures.* J Thorac Dis, 2019. **11**(Suppl 8): p. S1078-s1089.
7. Bergeron, E., et al., *Elderly trauma patients with rib fractures are at greater risk of death and pneumonia.* J Trauma, 2003. **54**(3): p. 478-85.
8. Beard, L., et al., *Analgesia of Patients with Multiple Rib Fractures in Critical Care: A Survey of Healthcare Professionals in the UK.* Indian journal of critical care medicine : peer-reviewed, official publication of Indian Society of Critical Care Medicine, 2020. **24**(3): p. 184-189.
9. Marc de, M., N. Ram, and B. Walter, *Rib fixation: Who, What, When?* Trauma Surgery & Acute Care Open, 2017. **2**(1): p. e000059.
10. Pieracci, F.M., et al., *Consensus statement: Surgical stabilization of rib fractures rib fracture colloquium clinical practice guidelines.* Injury, 2017. **48**(2): p. 307-321.
11. Dehghan, N., et al., *Flail chest injuries: A review of outcomes and treatment practices from the National Trauma Data Bank.* Journal of Trauma and Acute Care Surgery, 2014. **76**(2).
12. Otaka, S., et al., *Effectiveness of surgical fixation for rib fractures in relation to its timing: a retrospective Japanese nationwide study.* European Journal of Trauma and Emergency Surgery, 2022. **48**(2): p. 1501-1508.
13. Bhatnagar, A., J. Mayberry, and R. Nirula, *Rib fracture fixation for flail chest: what is the benefit?* J Am Coll Surg, 2012. **215**(2): p. 201-5.
14. Marasco, S.F., et al., *Prospective randomized controlled trial of operative rib fixation in traumatic flail chest.* J Am Coll Surg, 2013. **216**(5): p. 924-32.
15. Leinicke, J.A., et al., *Operative management of rib fractures in the setting of flail chest: a systematic review and meta-analysis.* Ann Surg, 2013. **258**(6): p. 914-21.
16. de Jong, M.B., et al., *Surgical treatment of rib fracture nonunion: A single center experience.* (1879-0267 (Electronic)).
17. Minervini, F., et al., *Nonunion of traumatic rib fractures: a suitable indication for surgery?* European Journal of Trauma and Emergency Surgery, 2022. **48**(4): p. 3165-3169.
18. Edwards, J.G., et al., *Taxonomy of multiple rib fractures: Results of the chest wall injury society international consensus survey.* J Trauma Acute Care Surg, 2020. **88**(2): p. e40-e45.
19. Van Wijck, S.F.M., et al., *Interobserver agreement for the Chest Wall Injury Society taxonomy of rib fractures using computed tomography images.* J Trauma Acute Care Surg, 2022. **93**(6): p. 736-742.
20. Kara, M., et al., *Disclosure of unnoticed rib fractures with the use of ultrasonography in minor blunt chest trauma.* European Journal of Cardio-Thoracic Surgery, 2003. **24**(4): p. 608-613.
21. Griffith, J.F., et al., *Sonography compared with radiography in revealing acute rib fracture.* AJR Am J Roentgenol, 1999. **173**(6): p. 1603-9.

22. Bansidhar, B.J., J.A. Lagares-Garcia, and S.L. Miller, *Clinical rib fractures: are follow-up chest X-rays a waste of resources?* Am Surg, 2002. **68**(5): p. 449-53.
23. Banaste, N., et al., *Whole-Body CT in Patients with Multiple Traumas: Factors Leading to Missed Injury.* Radiology, 2018. **289**(2): p. 374-383.
24. Cho, S.H., Y.M. Sung, and M.S. Kim, *Missed rib fractures on evaluation of initial chest CT for trauma patients: pattern analysis and diagnostic value of coronal multiplanar reconstruction images with multidetector row CT.* Br J Radiol, 2012. **85**(1018): p. e845-50.
25. Yao, L., et al., *Rib fracture detection system based on deep learning.* Sci Rep, 2021. **11**(1): p. 23513.
26. Chen, M., P. Du, and J.Y. Zhao, *SCRFD: Spatial Coherence Based Rib Fracture Detection.* 2018: p. 105-109.
27. Chai, Z.Z., et al., *ORF-Net: Deep Omni-Supervised Rib Fracture Detection from Chest CT Scans.* 2022. **13433**: p. 238-248.
28. Inoue, T., et al., *Automated fracture screening using an object detection algorithm on whole-body trauma computed tomography.* Sci Rep, 2022. **12**(1): p. 16549.
29. Niiya, A., et al., *Development of an artificial intelligence-assisted computed tomography diagnosis technology for rib fracture and evaluation of its clinical usefulness.* Sci Rep, 2022. **12**(1): p. 8363.
30. Su, Y.P., et al., *Rib fracture detection in chest CT image based on a centernet network with heatmap pyramid structure.* Signal Image Video Process., 2022.
31. Wang, S., et al., *Assessment of automatic rib fracture detection on chest CT using a deep learning algorithm.* Eur Radiol, 2022.
32. Yang, C., et al., *Development and assessment of deep learning system for the location and classification of rib fractures via computed tomography.* Eur J Radiol, 2022. **154**.
33. Zhou, Q.Q., et al., *Precise anatomical localization and classification of rib fractures on CT using a convolutional neural network.* Clin Imaging, 2022. **81**: p. 24-32.
34. Azuma, M., et al., *Detection of acute rib fractures on CT images with convolutional neural networks: effect of location and type of fracture and reader's experience.* Emerg Radiol, 2022. **29**(2): p. 317-328.
35. Zhang, B., et al., *Improving rib fracture detection accuracy and reading efficiency with deep learning-based detection software: A clinical evaluation.* Br J Radiol, 2021. **94**(1118).
36. Castro-Zunti, R., et al., *Assessing the speed-accuracy trade-offs of popular convolutional neural networks for single-crop rib fracture classification.* Comput Med Imaging Graph, 2021. **91**.
37. Meng, X.H., et al., *A fully automated rib fracture detection system on chest CT images and its impact on radiologist performance.* Skelet Radiol, 2021. **50**(9): p. 1821-1828.
38. Hu, Y., et al., *Slice grouping and aggregation network for auxiliary diagnosis of rib fractures.* Biomed Signal Process Control, 2021. **67**.
39. Kaiume, M., et al., *Rib fracture detection in computed tomography images using deep convolutional neural networks.* Medicine, 2021. **100**(20): p. E26024.
40. Weikert, T., et al., *Assessment of a deep learning algorithm for the detection of rib fractures on whole-body trauma computed tomography.* Korean J Radiol, 2020. **21**(7): p. 891-899.
41. Wu, M., et al., *Development and evaluation of a deep learning algorithm for rib segmentation and fracture detection from multicenter chest ct images.* Radiology: Art Int, 2021. **3**(5).
42. Zhou, Q.Q., et al., *Automatic detection and classification of rib fractures on thoracic ct using convolutional neural network: Accuracy and feasibility.* Korean J Radiol, 2020. **21**(7): p. 869-879.
43. Baumgartner, M., et al. *nnDetection: A Self-configuring Method for Medical Object Detection*. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. 2021. Cham: Springer International Publishing.
44. Isensee, F., et al., *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.* Nature Methods, 2021. **18**(2): p. 203-211.

45. Jaeger, P., et al., *Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection*. 2018.
46. *MeVisLab: Download* 2023 Jan 15]; Available from: https://www.mevislab.de/download.
47. Hayes, A.F. and K. Krippendorff, *Answering the Call for a Standard Reliability Measure for Coding Data.* Communication Methods and Measures, 2007. **1**(1): p. 77-89.
48. Gwet, K., *On Krippendorff's Alpha Coefficient.* 2015.
49. Fleiss, J.L. and J. Cohen, *The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability.* Educational and Psychological Measurement, 1973. **33**(3): p. 613-619.
50. McHugh, M.L., *Interrater reliability: the kappa statistic.* Biochem Med (Zagreb), 2012. **22**(3): p. 276-82.
51. Fedorov A., et al., *3D Slicer as an Image Computing Platform for the Quantitative Imaging Network.* Magnetic Resonance Imaging 2012 Nov: p. 1323-41.
52. Jin, L., et al., *Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet.* EBioMedicine, 2020. **62**.
53. Thambawita, V., et al., *Impact of Image Resolution on Deep Learning Performance in Endoscopy Image Classification: An Experimental Study Using a Large Dataset of Endoscopic Images.* Diagnostics (Basel), 2021. **11**(12).
54. Vali-Betts, E., et al., *Effects of Image Quantity and Image Source Variation on Machine Learning Histology Differential Diagnosis Models.* J Pathol Inform, 2021. **12**: p. 5.
55. Lessmann, N. *Rib segmentation*. 2022 [cited 2023 7th of August]; Available from: https://grand-challenge.org/algorithms/rib-segmentation/#information.
56. Lessmann, N., et al., *Iterative fully convolutional neural networks for automatic vertebra segmentation and identification.* Medical Image Analysis, 2019. **53**: p. 142-155.
57. Edamadaka, S., et al., *FasterRib: A deep learning algorithm to automate identification and characterization of rib fractures on chest computed tomography scans.* Journal of Trauma and Acute Care Surgery, 2023. **95**(2).
58. Zhou, Q.Q., et al., *Automatic detection and classification of rib fractures based on patients' CT images and clinical information via convolutional neural network.* Eur Radiol, 2021. **31**(6): p. 3815-3825.
59. Fokin, A.A., et al., *Surgical Stabilization of Rib Fractures: Indications, Techniques, and Pitfalls.* JBJS Essent Surg Tech, 2020. **10**(2): p. e0032.
60. Chan, H.-P., et al., *Deep Learning in Medical Image Analysis.* Advances in experimental medicine and biology, 2020. **1213**: p. 3-21.
61. Maes, F., et al., *The Role of Medical Image Computing and Machine Learning in Healthcare*, in *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks*, E.R. Ranschaert, S. Morozov, and P.R. Algra, Editors. 2019, Springer International Publishing: Cham. p. 9-10.
62. Li, F.F., J. Johnson, and S. Yeung, *Lecture 11: Detection and Segmentation*. 2018, Stanford.
63. Mujahid, M., et al., *Pneumonia Classification from X-ray Images with Inception-V3 and Convolutional Neural Network.* Diagnostics (Basel), 2022. **12**(5).
64. Soffer, S., et al., *Deep learning for pulmonary embolism detection on computed tomography pulmonary angiogram: a systematic review and meta-analysis.* Scientific Reports, 2021. **11**(1): p. 15814.
65. Yang, J., et al. *RibSeg Dataset and Strong Point Cloud Baselines for Rib Segmentation from CT Scans*. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. 2021. Cham: Springer International Publishing.

7. Supplementary material

Appendix A: Short introduction to technical aspect of the thesis

AI entails various techniques, including Machine Learning (ML), which is widely used in the medical imaging field. Traditional ML methods rely on the developers' expertise and knowledge to quantify features that are used for a classifier. This classifier can then make predictions on new data by the adjusted weights of these features. In contrast, Deep learning (DL), a subset of ML, automates the feature extraction (on raw data) [60]. This can result in superior performance in comparison with ML, when the dataset is sufficiently large, especially in medical image analysis [60, 61].

One of the areas where DL is widely used, is in Computer Vision Tasks. Within this broad term there are a few tasks specifically relevant for medical image analysis; image classification, object localisation, object detection and semantic segmentation [62]. First, image classification is concerned with classifying one image with one label at a time. For instance, classification of pneumonic or healthy lungs on a chest X-rays [63]. In object localisation, the objective is to locate (referred to as "detect" throughout this project) one or more objects in an image. This is usually done by defining a bounding box that captures the object(s) of interest. An example of this is the detection of pulmonary embolisms [64]. Object detection combines the two previous tasks. Thus, in an image, an object is simultaneously detected and classified. This is applicable to this project where rib fractures are detected and classified. Taking the object detection a step further is to delineate, or segment, an object and classify it. This is referred to as semantic segmentation [62]. For instance, segmenting each rib and stating which rib number it is.

Appendix B: Training process of nnU-Net

Data used for training nnU-Net consisted of the RibSeg opensource dataset [65], which contains 490 segmented ribs, and a part of the Erasmus MC dataset. The RibSeg dataset consists of mainly healing and old fractures, thus does not represent the patients in this project with acute fractures. However, it could still be used as a basis, because creating a segmented dataset from scratch was not possible due to time limits in this project. Therefore, an iterative approach was taken by gradually adding Erasmus MC data in two training loops. The first training was done on ten RibSeg patients. The trained model then made a prediction on nine randomly chosen Erasmus MC patients, containing contrast-enhanced images as these are not present in the RibSeg dataset. These predicted segmentations were evaluated and improved in 3D Slicer [51]. The corrected segmentations were all input for the second training which consisted of 17 RibSeg patients and the nine corrected Erasmus MC patients. A schematic overview is process can be seen in Figure 12.

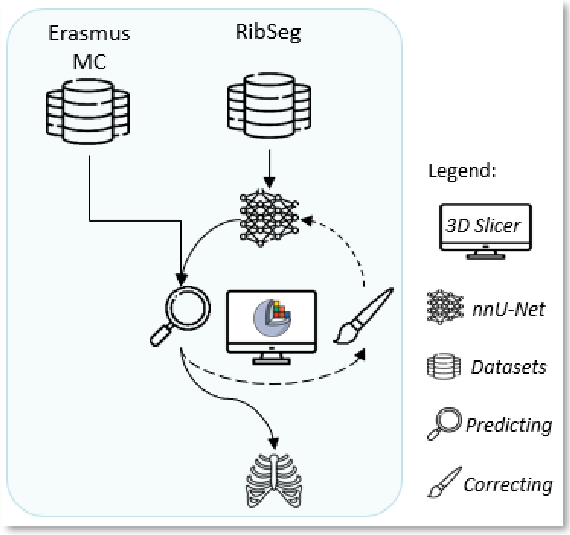


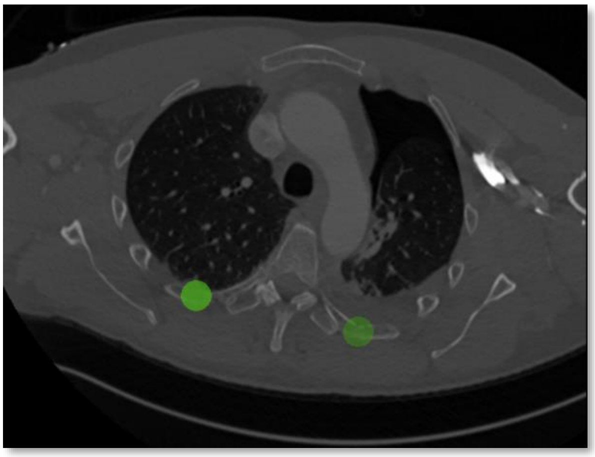
Figure 12: Schematic overview of the process to train the nnU-Net model.

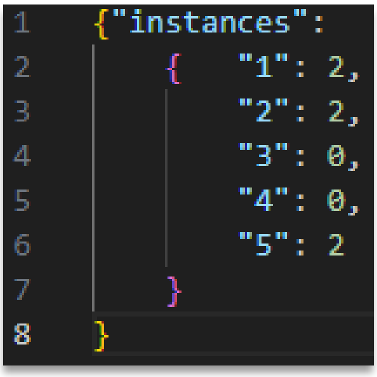
Appendix C: Label pre-processing; from CSV-file to NIfTI and JSON files

The information of the rib fractures needs to be converted into the specific label files used in the nnDetection framework. This entailed one NIfTI file containing the label map for the location of the rib fractures and three corresponding JSON files with CWIS classifications for the fracture. The fourth label, representing the rib number, did not require additional formatting.

The decision to use blobs instead of the more conventional bounding boxes was made because bounding boxes have the limitation that they are axis-aligned. However, blobs are isotropic, therefore invariant to rotation and easily constructed from the midpoints. The radius of the blobs was 10 voxels in the voxel-coordinate system of each NIfTI file. This was a balance between overlapping the majority of the rib fracture and avoiding too much overlap with other structures. To create these spherical blobs, the pixel spacing of the images needed to be accounted for. It is common that the in-plane pixel spacing is different from the pixel-spacing in the z-direction. Therefore, to construct a spherical blob instead of an ellipsoid, this is corrected for. Lastly, the pixel values assigned to the blobs are unique. This ensures that each blob is linked to their corresponding label in the JSON files. Each label map retains the same image shape, direction, origin and spacing as the original image.

The JSON files contain the label information in numerical format. For instance, anterior fractures are assigned a value of 0, lateral fractures a value of 1, and posterior fractures a value of 2. In Figure 13, an example of an axial label map overlaid on the original image with the corresponding JSON file for the label 'location' is seen.





```
{"instances":
    {   "1": 2,
        "2": 2,
        "3": 0,
        "4": 0,
        "5": 2
    }
}
```

Figure 13: Left) An example of the original image with the green spheres indicating the overlayed label map; Right) Corresponding JSON file for the location on the rib bow, where instances 1 and 2 correspond to the blobs seen in the left image and having assigned the posterior label.

Appendix D: Data selection

Chest CT images of patients with rib fractures were retrieved and anonymised through the Trialbureau of the Erasmus MC and accessible through XNAT. CT scans that did not include all ribs, only had fractures with callus formation, indicating older fractures, or had significant motion artifacts around a rib fracture were excluded from the dataset. In case of a patient with multiple CT reconstructions, the scan with the smallest slice thickness was chosen, and if there was more than one image with the smallest slice thickness, the one reconstructed with the lowest kernel number was chosen, irrespective of given contrast. The DICOM files were converted to NIfTI files.

Appendix E: Labelling software

A comprehensive overview of the labelling software and the application in this project is explained in this section. First, the module network is explained. Then, a manual belonging to the graphical user interface (GUI) will be given.

E.1 Module Network

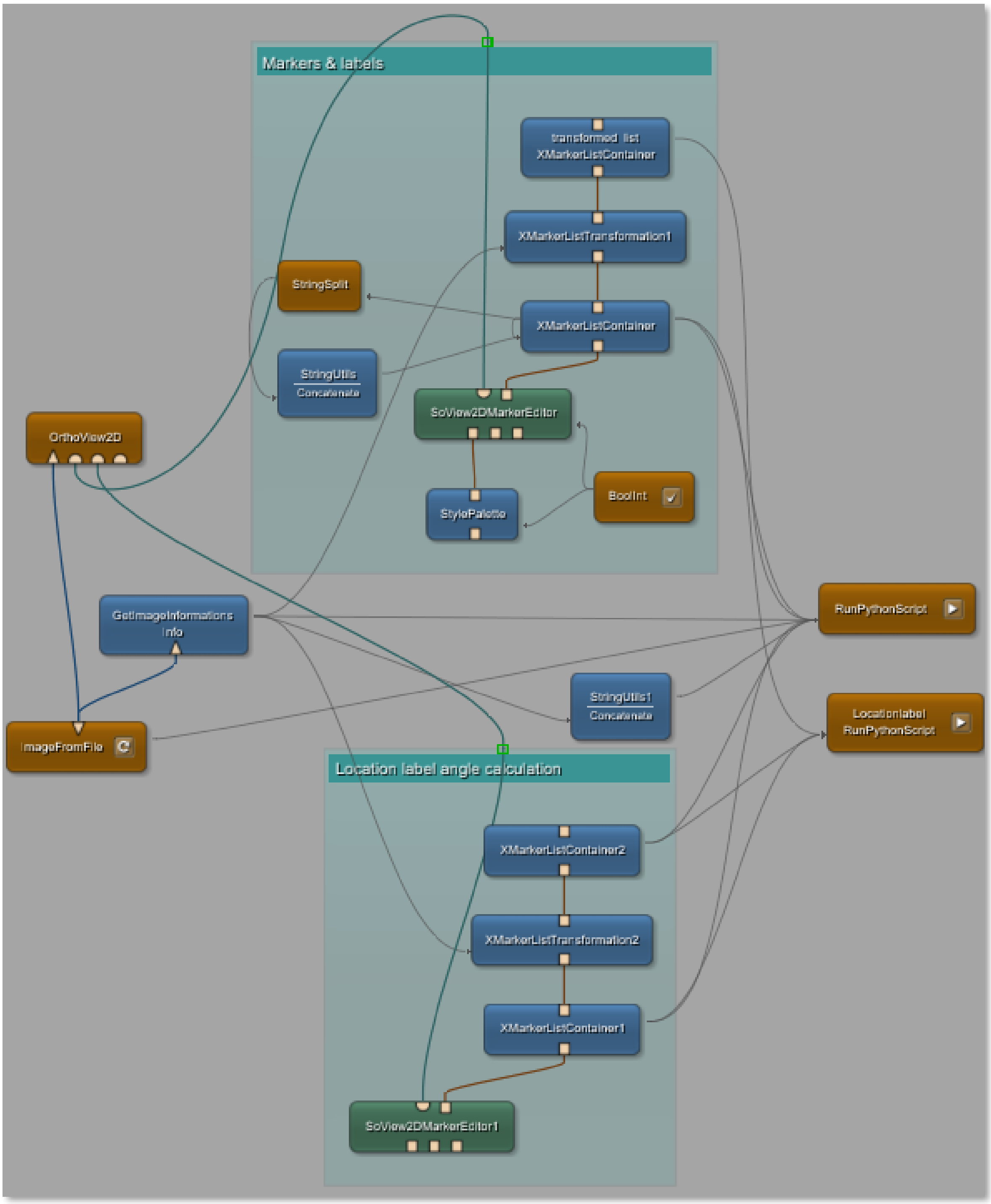
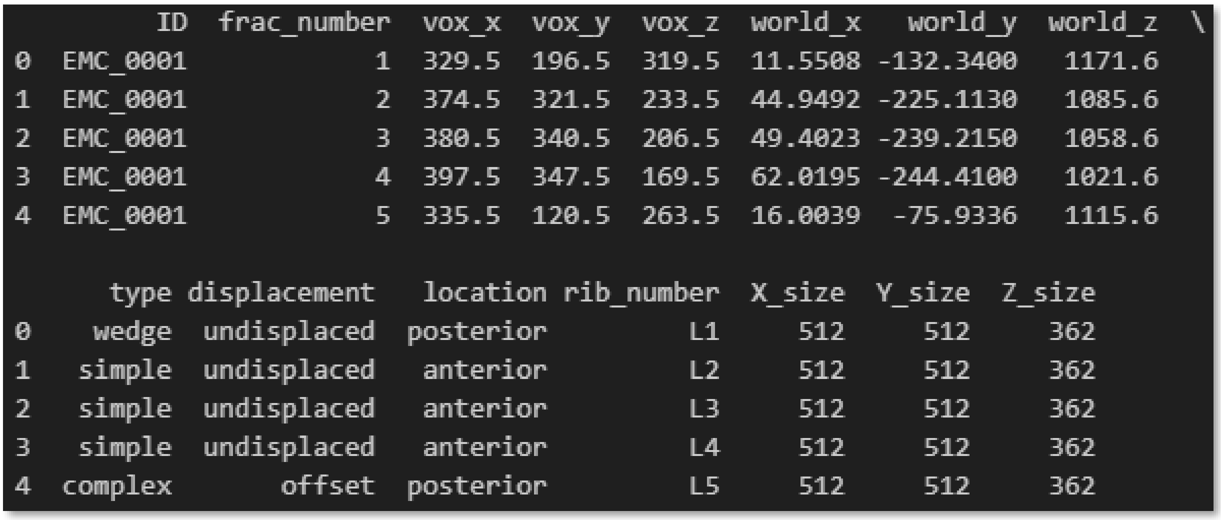


Figure 14: Module network of MeVisLab for the labelling of rib fractures.

There are two main parts in the module network; the marker & labels assignment, and the calculation of the angle from which the location label can be determined. Each module will briefly be described.

(Figure 14) For both parts, the same input is needed from the *OrthoView2D,* which is used to visualise the image that is loaded through *ImageFromFile*. Then, for the midpoints and label assignments:

SoView2dMarkerEditor – used for setting markers in the middle of the rib fractures;
StylePalette – used for setting different colours per marker to be able to distinguish them;
BoolInt – used to set the height and width of the visualisation bounding box around the marker;
XMarkerListContainer – used to merge all markers with their labels to a string;
StringSplit & StringUtils – used to concatenate all four labels belonging to one marker;
XMarkerListTransform – used to transform the world coordinates of the marker points to voxel coordinates. The voxel to world transformation matrix is given by the *GetImageInformationsInfo*;
StringsUtils1 – used to obtain the image size information;
RunPythonScript – used to combine all information, reformat it and save it as a CSV-file. For an example of the output, refer to Figure 15. If not all labels are assigned, instead of saving the CSV-file, an error-messages is outputted in the GUI.



```
         ID  frac_number  vox_x  vox_y  vox_z  world_x   world_y  world_z  \
0  EMC_0001            1  329.5  196.5  319.5  11.5508 -132.3400   1171.6
1  EMC_0001            2  374.5  321.5  233.5  44.9492 -225.1130   1085.6
2  EMC_0001            3  380.5  340.5  206.5  49.4023 -239.2150   1058.6
3  EMC_0001            4  397.5  347.5  169.5  62.0195 -244.4100   1021.6
4  EMC_0001            5  335.5  120.5  263.5  16.0039  -75.9336   1115.6

      type displacement   location rib_number  X_size  Y_size  Z_size
0    wedge  undisplaced  posterior         L1     512     512     362
1   simple  undisplaced   anterior         L2     512     512     362
2   simple  undisplaced   anterior         L3     512     512     362
3   simple  undisplaced   anterior         L4     512     512     362
4  complex       offset  posterior         L5     512     512     362
```

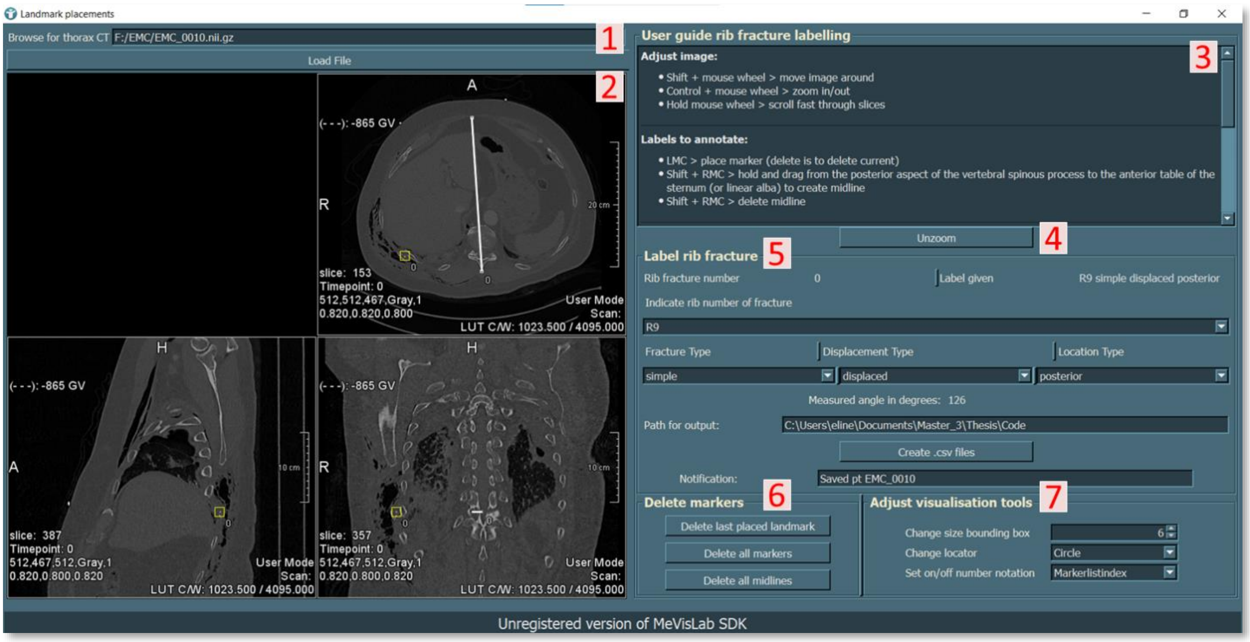
Figure 15: Example of the CSV output file for the first five fractures of a patient is shown. The voxel and world coordinates of the fractures are given, the four labels and the image shape of the NIfTI file.

Similarly, the location label angle calculation is set up. However, instead of using markers to which the labels are assigned, it uses midlines from which angles can be calculated. The *SoView2DMarkerEditor* is now used in vector mode. Then, the *RunPythonScript* is used to calculate the angle between the marker and the drawn midline. The calculation of this angle was based on the following formula:

$$\cos\theta = \frac{\mathbf{a} * \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$$

where **a** represents the vector from the mid-thoracic point to the linear alba and **b** represents the vector from the mid-thoracic point to the landmark. The dot product between these two vectors was divided by the product of their lengths to obtain the cosine of the angle. The angle was calculated on the slice where the landmark was placed, in the centre of the rib fracture. The result of this script is an angle in degrees which can help in deciding which location label should be assigned. The module network comes together in the GUI that is explained in the next chapter.

E.2 GUI manual



Browse for thorax CT F:/EMC/EMC_0010.nii.gz

(- - -): -865 GV

slice: 153
Timepoint: 0
512,512,467,Gray,1
0.820,0.820,0.800
User Mode
Scan:
LUT C/W: 1023.500 / 4095.000

slice: 387
Timepoint: 0
512,467,512,Gray,1
0.820,0.800,0.820
User Mode
Scan:
LUT C/W: 1023.500 / 4095.000

slice: 357
Timepoint: 0
512,467,512,Gray,1
0.820,0.800,0.820
User Mode
Scan:
LUT C/W: 1023.500 / 4095.000

User guide rib fracture labelling

Adjust image:

- Shift + mouse wheel > move image around
- Control + mouse wheel > zoom in/out
- Hold mouse wheel > scroll fast through slices

Labels to annotate:

- LMC > place marker (delete is to delete current)
- Shift + RMC > hold and drag from the posterior aspect of the vertebral spinous process to the anterior table of the sternum (or linear alba) to create midline
- Shift + RMC > delete midline

Label rib fracture

Rib fracture number 0
Label given R9 simple displaced posterior
Indicate rib number of fracture R9
Fracture Type simple
Displacement Type displaced
Location Type posterior
Measured angle in degrees: 126
Path for output: C:\Users\eline\Documents\Master_3\Thesis\Code
Notification: Saved pt EMC_0010

Delete markers

Adjust visualisation tools

Change size bounding box 6
Change locator Circle
Set on/off number notation Markerlistindex

Figure 16: The GUI of the MeVisLab labelling software with numbers indicating the different sections.

The GUI will be explained according to the different sections in the interface, corresponding to the numbers in Figure 16.

1. Give the path to the NIfTI image and click on ‘Load File’.
2. The images are shown in axial, sagittal and coronal slices. After identifying the rib fracture, left mouse click in the middle of the fracture to set a marker. Here, shown as the yellow square. If you want to calculate the angle for the location label, hold shift + right mouse click from the posterior aspect of the vertebral spinous process to the linear alba to create the midline (direction of vector is important). The angle will be outputted in section 5.
   A marker can be deleted by clicking on it and using *delete* on your keyboard. Shift + right mouse click on the midline begin or endpoint to delete the midline.
3. A quick user guide on how to use the software. Additional tips are given for adjusting the images in section 2. Moreover, a short description of the classification system is given.
4. When going through the slices of different patients, the field of view is sometimes not correct and it seems like there is no image showing. Press *Unzoom* to set the field of view to the current patient.
5. In the first line, the rib fracture number that is currently selected and the assigned labels are shown. In the drop-down menus, the four labels can be allocated. Then, the output path needs to be defined. Once all rib fractures are marked and given their labels, the *Create .csv files* button can be clicked. If all rib fractures have all four labels and there is at least one midline drawn, the notification will output *Saved pt [name patient]*. If labels are missing, the notification will output which marker’s label is missing. When no midline is defined, the output is *Draw at least one midline*.
6. If the patient data is saved and a new patient is loaded, all markers and midlines of the former patient should be deleted. To do this, click on *Delete all markers* & *midlines*.
7. The visualisation of section 2 can be changed a little. The visualisation bounding box can be changed, which is purely for visualisation purposes as it does not influence the marker coordinates. Lastly, the locator and the number notation next to the yellow box can be changed.

Appendix F: Supplementary graphs

F.1 Dataset

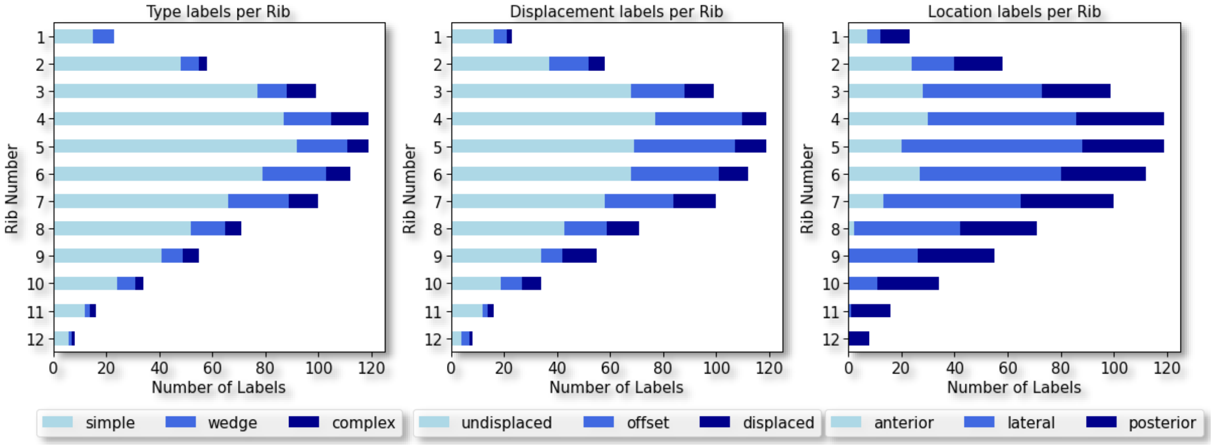


Figure 17: Schematic representation of the labels per rib in the training dataset. Left, Middle, Right) Distribution per rib number for the labels type, displacement and location, respectively.

F.2 Detection

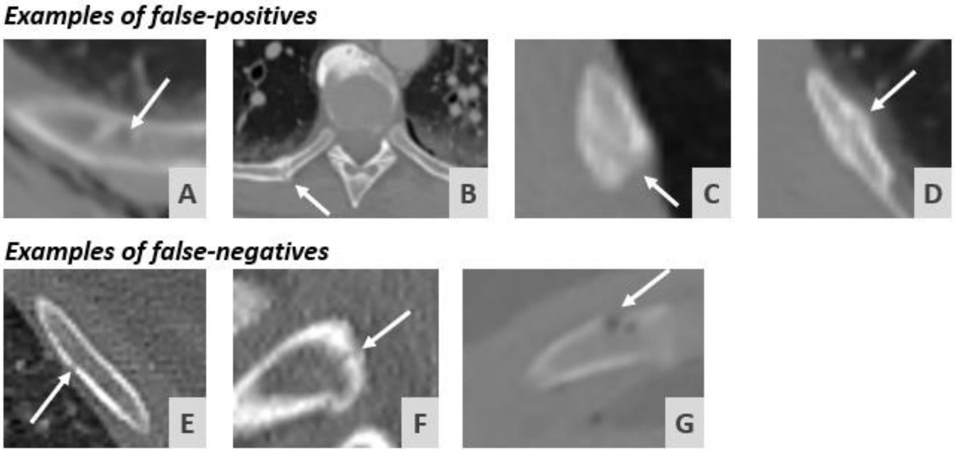


Figure 18: Examples of wrong detection outputs indicated with the white arrow. A) Cyste; B) Paravertebral fracture; C, D) Old fractures with callus; E, F, G) Small fractures that remained unnoticed by DCRibFrac.

F.3 Classification

Table 7: Results of the proposed model on the internal test set.

	Sensitivity	Precision	F1-score	FPPS
Detection	0.77	0.79	0.78	2.26
Type				
• Simple	0.90	0.75	0.82	
• Wedge	0.30	0.42	0.35	
• Complex	0.17	0.30	0.21	
Displacement				
• Undisplaced	0.91	0.83	0.86	
• Offset	0.78	0.79	0.78	
• Displaced	0.43	0.75	0.55	
Location				
• Anterior	0.88	0.88	0.88	
• Lateral	0.88	0.95	0.92	
• Posterior	0.96	0.84	0.90	

Table 8: Cohen's kappa scores with 95% confidence interval for type label comparison between observers and DCRibFrac on 34 rib fractures from the test set

	Observer A	Observer B	Observer C	DCRibFrac
Observer A	x	0.53 (0.26-0.79)	0.37 (0.06-0.68)	-0.02 (-0.39-0.34)
Observer B	0.53 (0.26-0.79)	x	0.37 (0.06-0.67)	-0.18 (-0.54-0.17)
Observer C	0.37 (0.06-0.68)	0.37 (0.06-0.67)	x	0.15 (-0.27-0.56)
DCRibFrac	-0.02 (-0.39-0.34)	-0.18 (-0.54-0.17)	0.15 (-0.27-0.56)	x

Table 9: Cohen's kappa scores with 95% confidence interval for displacement label comparison between observers and DCRibFrac on 34 rib fractures from the test set

	Observer A	Observer B	Observer C	DCRibFrac
Observer A	x	0.73 (0.54-0.93)	0.54 (0.30-0.78)	0.28 (0.01-0.56)
Observer B	0.73 (0.54-0.93)	x	0.38 (0.14-0.63)	0.20 (-0.05-0.50)
Observer C	0.54 (0.30-0.78)	0.38 (0.14-0.63)	x	0.13 (-0.15-0.41)
DCRibFrac	0.28 (0.01-0.56)	0.20 (-0.05-0.50)	0.13 (-0.15-0.41)	x

Table 10: Cohen's kappa scores with 95% confidence interval for location label comparison between observers and DCRibFrac on 34 rib fractures from the test set

	Observer A	Observer B	Observer C	DCRibFrac
Observer A	x	0.63 (0.39-0.87)	0.74 (0.51-0.98)	0.45 (0.16-0.76)
Observer B	0.63 (0.39-0.87)	x	0.54 (0.28-0.80)	0.58 (0.33-0.84)
Observer C	0.74 (0.51-0.98)	0.54 (0.28-0.80)	x	0.36 (0.05-0.67)
DCRibFrac	0.45 (0.16-0.76)	0.58 (0.33-0.84)	0.36 (0.05-0.67)	x

Appendix G: Implementation details nnDetection

The nnDetection framework was slightly changed for this project. First, the *nevergrad* library used within nnDetection was not up-to-date with the newer *numpy* functions. Therefore, in the following scripts, changes needed to be made:

- Nevergrad/parametrization/data.py at line 16: *np.int* changed to *int* and *np.float* changed to *float*;
- Nevergrad/optimization/base.py at line 95: *np.int* changed to *int*;
- Nevergrad/optimization/utils.py at line 149: *np.float* changed to *float.*

Secondly, the blobs used within this project all had exactly the same voxel size. This consistency caused issues with a particular definition within nnDetection, where bounding box sizes falling within the 0.005 upper and lower percentile were filtered out. Consequently, in this project, the filtering resulted in the removal of all blobs. Adjustments were made as follows:

- nnDetection/nndet/planning/architecture/boxes/base.py line 398-425: at the return statement *boxes_np[mask.astype(bool)]* changed to *boxes_np*.

Lastly, a recommended modification, though not necessary, involved changing the training logger to an online version. By implementing this change, it becomes possible to track the models in real-time without the need to log in to the GPU cluster. To utilise this feature, an account needs to be made on the website *wandb.ai*. Then, make the following adjustments in the code:

- nnDetection/scripts/train.py line 28: add to the imports *from pytorch_lightning.loggers import WandbLogger* and at line 189-190: pl_logger = WandbLogger(project=cfg[“task”], etc.
- In the slurm script, an account-specific key should be exported. An example: *export WANDB_API_KEY=c34f87beebb843a3ce5d9293c0c5bffc45905a3*