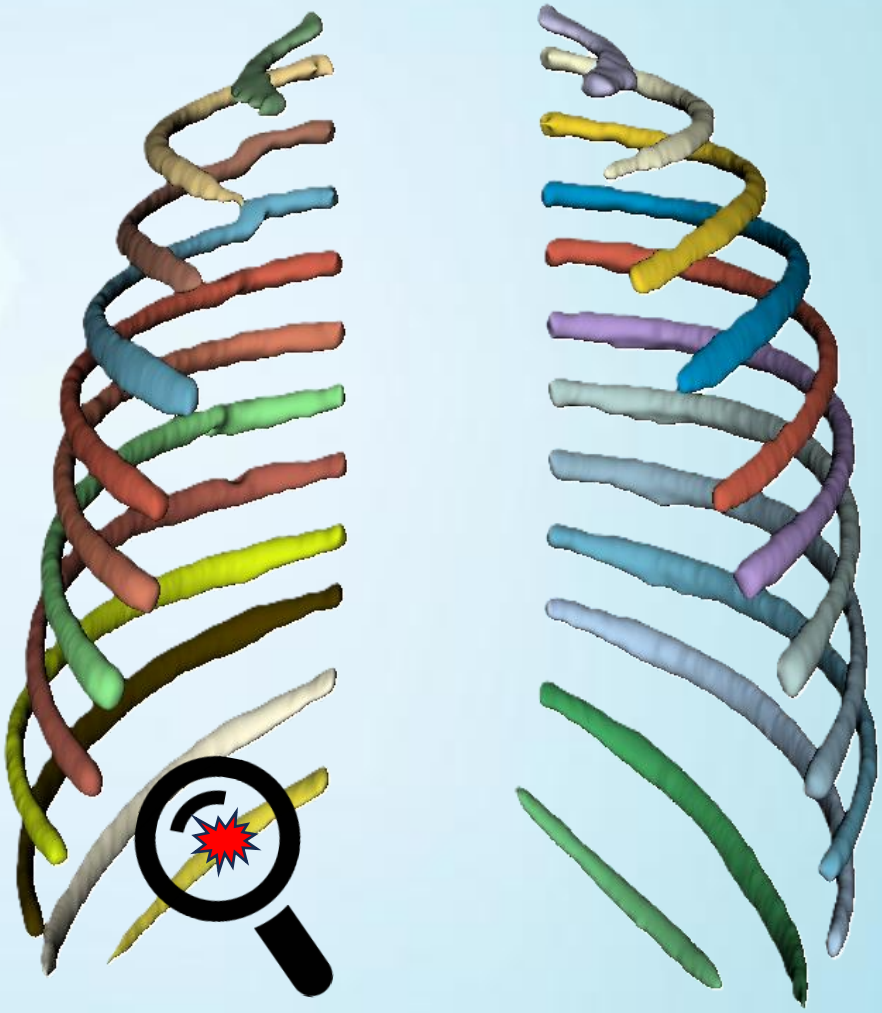
Optimisation of Deep Learning Based Automatic Rib Fracture Detection and Classification



MSc Thesis Technical Medicine

Victoria Marting

Universiteit Leiden
The Netherlands

TU Delft Delft University of Technology

Erasmus University Rotterdam

OPTIMISATION OF DEEP LEARNING BASED AUTOMATIC RIB FRACTURE DETECTION AND CLASSIFICATION

Victoria Marting

Student number: 4884493

24-09-2024

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 & Biomedical Imaging Group Rotterdam
Rotterdam Erasmus Medical Centre

January 2024 – September 2024

Supervisors:

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

Thesis committee members:

M.M.E. (Mathieu) Wijffels, MD, PhD, Erasmus MC
T. (Theo) van Walsum, Eng, PhD, Erasmus MC
D.K. (Dorottya) de Vries, MD, PhD, Leiden Universitair Medisch Centrum

Table of Contents

Preface

I am proud to present this thesis, the result of nine months of dedicated work. It marks the end of my journey as a student. These six years have flown by and have enriched me personally and professionally. I am deeply grateful for all the people I have met and the opportunities that have come my way during this period.

The successful completion of this project would not have been possible without the invaluable guidance of my supervisors Theo and Mathieu. I have learned so much from both of you throughout this process. Your knowledge and enthusiasm during our weekly meetings have been a constant source of motivation, making this thesis journey a rewarding experience. Your accessibility and support made it a true pleasure to work with you. It was an honour to build upon the work of Noor for my thesis. Beyond the thesis, I thoroughly enjoyed the clinical aspects of my graduation internship at the Department of Trauma Surgery, where I developed my skills as a medical professional. Also, the weekly IGIT meetings with Theo were both educational and enjoyable.

I would like to give a special thanks to Cile and Alex for always being by my side during my time here and helping me where needed. Your support, combined with our countless coffee breaks and lunch conversations, made this experience not only productive but also genuinely fun. I'll miss those moments! Lastly, I would like to thank my family, friends, roommates and boyfriend, for their encouragement along this journey. From now on I will go through life as a Technical Physician! I am excited to see what the future holds!

Victoria Marting *September 2024*

List of Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
CoM	Centre of Mass
CT	Computed Tomography
CWIS	Chest Wall Injury Society
DCRibFrac	Detection and Classification of Rib Fractures
DL	Deep Learning
EV	External Validation
FPPS	False Positives Per Scan
FixCon	Fixation versus Conservative
HET	High Energy Trauma
IoU	Intersection over Union
LET	Low Energy Trauma
NRS	Numeric Rating Scale
PR-curve	Precision-Recall Curve
RCT	Randomized Controlled Trial
SSRF	Surgical Stabilization of Rib Fractures
TS	TotalSegmentator
VAS	Visual Analogue Scale

Abstract

Introduction: Trauma-induced rib fractures are a common injury, affecting millions of individuals globally each year. The number and characteristics of these fractures influence whether a patient is treated conservatively or surgically. Rib fractures are typically diagnosed using CT scans, yet 19.2% to 26.8% of fractures are still missed during assessment. Another challenge in managing rib fractures is the interobserver variability in their classification. In 2023, a deep learning-based algorithm for the automatic detection and classification of rib fractures (DCRibFrac v1.0) was developed based on the Chest Wall Injury Society (CWIS) taxonomy. Although DCRibFrac v1.0 demonstrated promising results, there remained room for improvement, particularly in the accuracy of rib number labelling. This project aims to develop and assess DCRibFrac v2.0, an enhanced version of the original algorithm.

Methods: Two novel approaches for automatic rib number labelling were proposed and evaluated: a pre-trained deep learning model named TotalSegmentator and a custom-developed nnUNet. Additionally, three nnDetection models were developed based on multi-centre data for the automatic detection of fractures and the classification of their type, displacement, and location. The performance of DCRibFrac v2.0 was evaluated. Finally, an external validation was conducted to assess the generalizability and robustness of DCRibFrac v2.0.

Results: For the development and evaluation of DCRibFrac v2.0 a total of 170 patients were included. The custom-developed nnUNet was the best-performing method for rib number labelling, correctly labelling 95.5% of all ribs and 98.4% of fractured ribs in 30 patients. On the internal test set, DCRibFrac v2.0 achieved a detection sensitivity of 80%, a precision of 87%, and an F1 score of 83%, with a mean FPPS (false positives per scan) rate of 1.11. Classification sensitivity varied across fracture types, with the lowest being 25% for complex fractures and the highest being 97% for posterior fractures. The correct rib number was assigned to 94% of the detected fractures. The detection and classification performance on the external validation dataset was slightly better, with a fracture detection sensitivity of 84%, precision of 85%, F1 score of 84%, FPPS of 0.96 and 95% of the fractures assigned the correct rib number.

Conclusion: The developments resulting in DCRibFrac v2.0 have improved the automatic detection and classification of rib fractures, with particular advancements in rib number labelling performance and detection precision. These improvements are important steps towards establishing a more accurate and standardised method for rib fracture assessment, which could enhance clinical decision-making and improve patient outcomes in trauma care.

1. Introduction

1.1 Traumatic Rib Fractures

Traumatic rib fractures are a common injury following thoracic trauma and are often caused by high force to the chest wall (1). Rib fractures account for 10% of all trauma admissions with a prevalence of 10-40% among trauma patients (1–3). These injuries result from high-energy trauma (HET) in younger patients, such as falls from heights or car accidents, frequently accompanied by other injuries. In elderly patients, they result from low-energy trauma (LET) (4–6). In general, rib fractures lead to high morbidity and can cause mortality when combined with other conditions such as haemothorax, pneumothorax, and soft tissue injuries (3,7). The thoracic pain caused by rib fractures limits patients to cough and breathe deeply, which can result in atelectasis and pneumonia (8). An increased number of fractures and older age are associated with increased rates of morbidity and mortality (3,5).

1.2 Rib Fracture Management

Rib fractures can be managed through surgical intervention or conservative treatment. A combination of optimal pain management, pulmonary physical therapy, oxygen suppletion and mechanical ventilation is considered essential for the conservative management of patients with rib fractures (2,9,10). Despite this treatment strategy, mortality and complications such as pulmonary contusion and pneumonia still occur often (1,4,10,11). Traditionally, patients were treated conservatively, nowadays more data is supporting the positive effects of early surgical stabilisation of rib fractures (SSRF) (12).

SSRF aims to improve respiratory mechanics, reduce pain and prevent pulmonary complications by inserting rib fracture stabilising systems. However, operative treatment increases the risk of surgical site infections with or without implant infections, potentially necessitating multiple additional surgical procedures.

In the case of a flail chest, defined as at least three consecutive fractured ribs with two or more fracture lines per rib, SSRF is preferred. This condition leads to a segment of the chest wall that moves paradoxically compared to the rest of the chest, resulting in insufficient breathing physiology (Figure 1). For patients with a flail chest, studies have shown that SSRF reduces the risk of pneumonia, shortens hospital length of stay and decreases the duration of mechanical ventilation (13–16). Therefore, SSRF is considered the best treatment option for flail pattern patients.

For patients with multiple simple rib fractures, it remains unclear whether a conservative or surgical approach is more beneficial in terms of patient outcomes and cost-effectiveness, as well as which factors should influence this decision (11,17–19). It frequently occurs that these patients, as well as flail patients, present with non-union(s) of their rib fracture(s) a few months after the trauma. It is still a matter of debate if this patient group, with multiple simple (non-flail) fractures, could benefit from SSRF (20).

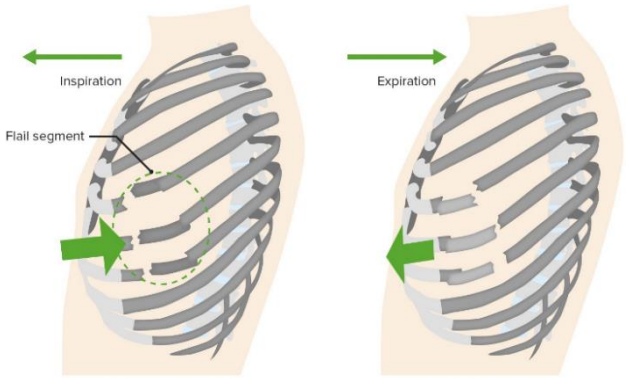


Figure 1: Illustration showing the characteristic segmental rib fractures in flail chest, where a segment of the rib cage moves independently from the rest of the chest wall.

A multi-centre randomised clinical trial (RCT) is currently conducted by Erasmus MC, comparing the surgical and non-surgical treatment strategies for patients with multiple simple (non-flail) rib fractures, to evaluate which treatment yields the best outcomes (FixCon trial) (11). It is expected that SSRF for these patients will reduce the incidence of pneumonia and shorten the hospital length of stay compared to non-surgical treatment. The findings of this trial will provide valuable insights into the clinical outcomes associated with each treatment strategy, offering evidence that can inform future decision-making processes. One of the key factors influencing treatment decisions is the fracture characteristics, such as the location, type and number of rib fractures (2).

1.3 Rib Fracture Classification

A reason for the limited implementation of SSRF guidelines is the inconsistency in rib fracture classification. This inconsistency complicates communication in both clinical practice and scientific research. Given that rib fracture classification plays a critical role in decision-making, improving the reliability and accuracy of classification, would be highly beneficial for optimizing treatment strategies. This enhancement would ensure that patients receive the most appropriate treatment through a standardised evaluation of their rib fractures.

In 2020, the Chest Wall Injury Society (CWIS) published SSRF guidelines, including a rib fracture classification system. This classification system was conducted through a Delphi consensus study (21). The proposed taxonomy is based on three characteristics: the fracture type, fracture displacement and fracture location on the rib bow. The CWIS taxonomy distinguishes the fracture line type as *simple*, *wedge* or *complex*, the displacement as *undisplaced*, *offset* or *displaced*, and the location as *anterior*, *posterior*, or *lateral* (see Figure 2–4 for the definition of each class). However, significant interobserver variability remains among clinicians in rib fracture classification using the CWIS taxonomy, particularly concerning the type and displacement classification (22).

Computed Tomography (CT) is the most effective imaging modality for diagnosing rib fractures resulting from trauma and is the golden standard. Nevertheless, literature shows that 19.2% to 26.8% of rib fractures are still missed during the diagnostic process, which is done manually by radiologists and other healthcare professionals (22–24). Additionally, the manual classification of these fractures is time-consuming. To address these issues, a robust, (semi)automatic and reliable CT-based classification scoring approach is necessary.

In the past years, several deep learning (DL) methods for rib fracture detection and classification have been published (25–32). DL is a subset of machine learning (ML), which is a subset of artificial intelligence (AI). DL is based on neural networks, which are designed to automatically learn and extract features from example data without explicit programming. These features can be used for the classification of new data. The studies show that DL models hold promise in enhancing rib fracture detection and establishing a more consistent classification compared to clinicians. However, these models are not up to date with the CWIS taxonomy standards, and it is unclear how robust these models are. The reported classification systems offer limited clinical value as they have no treatment consequences.

Figure 2: Schematic representation of fracture type according to the CWIS rib fracture taxonomy (21).
A) simple: single fracture line,
B) wedge: single fracture line with an additional line that does not run through the entire width of the rib. This creates a chipped-off fragment,
C) complex: two or more fracture lines that extend across the entire rib width, resulting in the presence of one or more fragments.

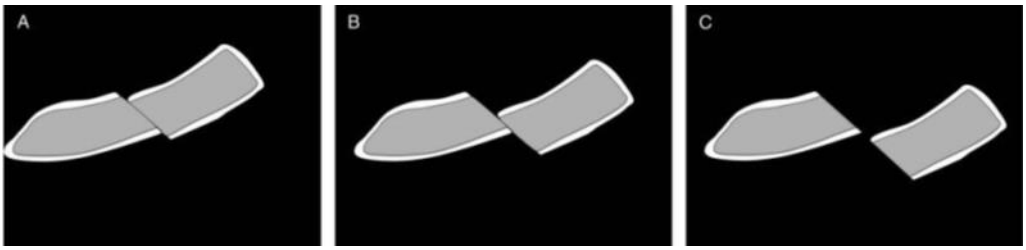


Figure 3: Schematic representation of fracture displacement according to the CWIS rib fracture taxonomy (21).
A) undisplaced: more than 90% cortical contact,
B) wedge: between no cortical contact and 90% cortical contact,
C) displaced: no cortical contact.

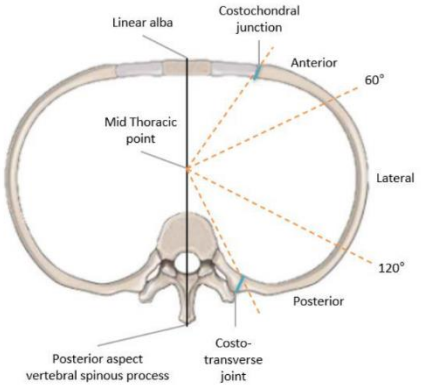


Figure 4: Schematic representation of fracture location according to the CWIS rib fracture taxonomy. The mid-thoracic point is defined as the middle of the line drawn between the linear alba and the posterior aspect of the vertebral spinous process (21).
Anterior: the angle between 0° and 60° relative to the mid thoracic point,
Lateral: the angle between 60° and 120° relative to the mid thoracic point,
Posterior: the angle between 120° and 180° relative to the mid thoracic point.

1.4 Automated CWIS Taxonomy

In 2023, NB proposed a DL-based classification algorithm based on the CWIS taxonomy, named DCRibFrac v1.0 (33). The detection sensitivity was 77%, which is on par with that of clinicians but did not surpass them. In general, DCRibFrac v1.0 shows potential for improved detection sensitivity and consistent classification of acute rib fractures from CT scans, but further refinements are needed. The results indicate that certain classes are more difficult to classify than others (Table 1). In particular, DCRibFrac v1.0 did not yield satisfactory results for the rib fracture type classification. The lack of performance could be attributed to the size of the dataset, as overfitting was still evident with a training dataset of n=81. Therefore, a larger dataset is needed to decrease overfitting and improve detection and classification. Additionally, the large label imbalance might have made learning the less frequent labels' characteristics more challenging (complex, wedge and displaced rib fractures). The significant interobserver variability, demonstrated by interobserver agreement studies, might also be an explanation for these results (22,33).

Table 1: Detection and Classification Performance DCRibFrac v1.0

	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

The currently used method for rib number labelling utilizes a nnUNet for the initial segmentation of the ribs. Segmentation is a technique used to divide data sets into multiple image segments to change the representation of the image in a more understandable way. In this process, 2D pixels and 3D voxels are assigned labels that share the same properties. In this case, one segment was created specifically to identify the ribcage. During post-processing, this segment is first improved by morphological operations (34). Additionally, the segmentation is split into multiple segments, which are then counted. Lastly, the centres of mass (CoMs) are calculated to order the segments (see Figure 5).

Using this method, the rib number labelling yielded favourable results for patients with minor displaced rib fractures but was ineffective for severely dislocated ribs. In the internal test dataset, the rib numbers were labelled correctly in only 8 out of 19 patients (42%). Figure 6 illustrates two cases where the rib number labelling was incorrect. Figures 6A and 6B represent the same patient. The fracture encircled in Figure 6B has caused the two fragments to be labelled as separate ribs (Figure 6A), whereas they should be labelled as multiple fragments of the same rib. Figure 6C illustrates an example where multiple ribs were incorrectly merged due to morphological operations, resulting in a single segment containing multiple ribs. To improve the performance of DCRibFrac v1.0, the accuracy of rib number labelling should be enhanced.

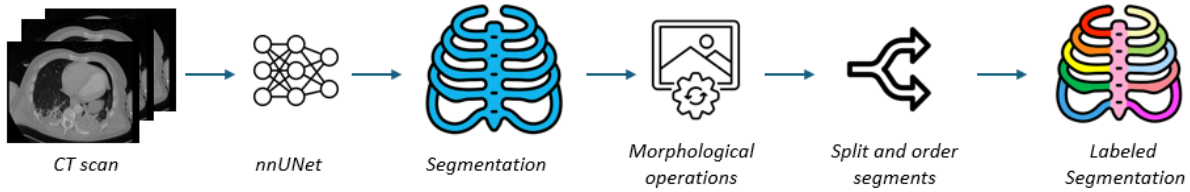


Figure 5: Pipeline for automatic rib number labelling in DCRibFrac v1.0.

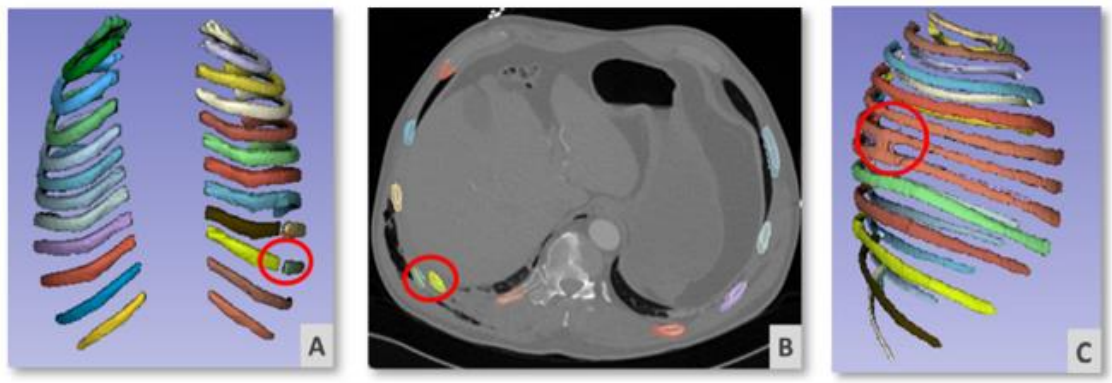


Figure 6: Examples of wrongly segmented and labelled ribs.
A) Posterior view of a 3D segmented model where the fracture splits the segmentation into two separate fragments,
B) Axial slice of the fracture encircled in A,
C) Lateral view of a 3D segmented model where ribs are merged due to dilation.

1.5 Objective and Contributions

The primary objective of this project is to develop and evaluate DCRibFrac v2.0, an improved version of DCRibFrac v1.0. The contributions are as follows:

- Implement and evaluate an improved rib number labelling method,
- Enhance the classification performance by expanding the dataset with multi-centre FixCon data,
- Assess interobserver agreement of CWIS classification to evaluate the consistency and agreement among different observers using the CWIS taxonomy,
- Perform an external validation within the Netherlands to evaluate the generalizability and reliability of DCRibFrac v2.0.

2. Materials and Methods

This section outlines the development of DCRibFrac 2.0 and details its advancements from DCRibFrac v1.0. Section 2.1 describes the framework of DCRibFrac v1.0. Sections 2.2 and 2.3 detail the steps taken to enhance the classification and the rib number labelling performance, respectively. Finally, the setup for external validation is explained in Section 2.4.

2.1 DCRibFrac v1.0

The algorithm development of DCRibFrac v1.0 involved creating four DL models, three nnDetection models and one nnUNet, each targeting a specific classification category: type, displacement, location or rib number (35,36). DCRibFrac v1.0 was developed and evaluated using 100 thoracic CT scans from trauma patients exclusively from EMC.

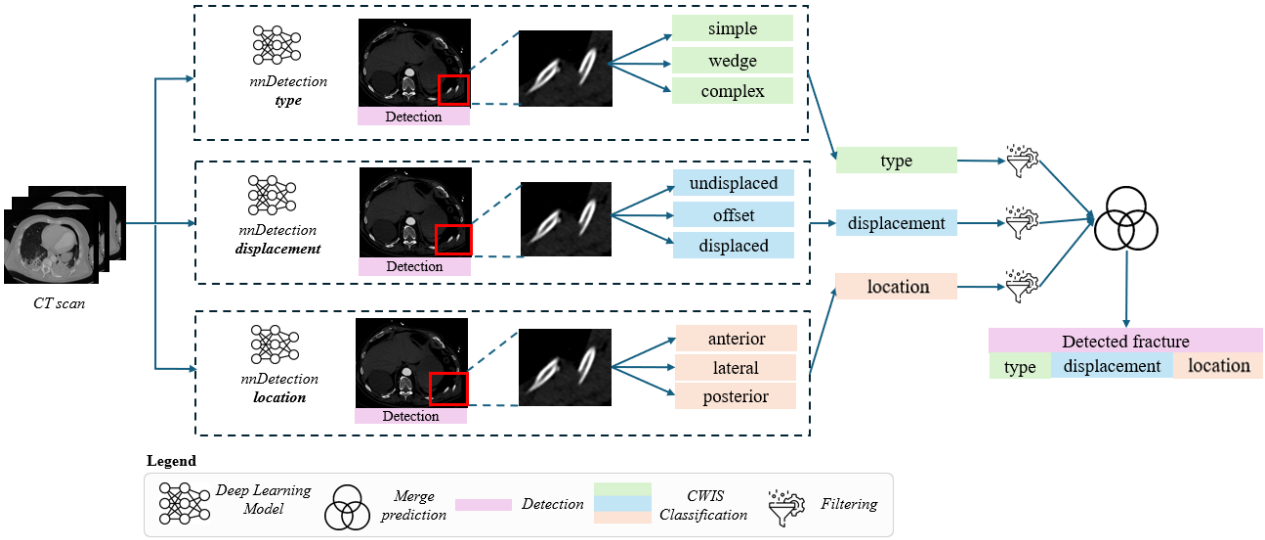


Figure 7: Overview of the pipeline for automatic detection and CWIS classification of rib fractures. The dotted boxes indicate the steps which are taken within the DL-models. During post-processing the predictions are combined.

2.1.1 CWIS Classification

Figure 7 illustrates the pipeline to obtain the predicted CWIS classification. Three separate nnDetection models were developed for the classification of the type, displacement and location of the fracture, as the nnDetection framework processes only one classification at a time. Each nnDetection model produces bounding box coordinates. Each bounding box indicates a detected fracture and is accompanied by a probability score and the corresponding labels for each detection. Consequently, one detected rib fracture with the complete CWIS classification has three different bounding boxes. For further details on nnDetection, refer to Appendix A. During post-processing, the results from the three nnDetection models are ensembled by combining their bounding boxes. Initially, each nnDetection model's output is filtered to eliminate overlapping bounding boxes, retaining only those with the highest probability scores while discarding those with lower scores. Two bounding boxes overlap if their Intersection-over-Union (IoU) score exceeds 50%. The IoU score is calculated using the following formula:

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

where bbox1 and bbox2 are bounding boxes, ∩ indicating the intersection and ∪ the union of the boxes. Additionally, the models are merged with the requirement 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 a union. In cases where only two out of three models overlap, one of the labels cannot be assigned. In such instances, a label 'unknown' is assigned.

2.1.2 Rib number labelling

To determine the rib number on which a detected fracture is located, rib segmentations are needed. The state-of-the-art medical segmentation framework nnU-Net was utilised for this task. This framework was chosen because it handles the entire pipeline, including preprocessing, training, and post-processing, autonomously across a wide variety of segmentation tasks. In DCRibFrac v1.0 a nnU-net was trained to create one single segment representing the ribcage. For further details on nnUNet, refer to Appendix B. As explained in the introduction, the segmentation of the ribcage by nnUNet is post-processed to obtain the number of fractured ribs. The nnUNet segmentation is first improved by morphological operations. Additionally, the single segment is split into multiple segments, which are then counted. Lastly, the centres of mass (CoMs) are calculated to order the segments, see Figure 5.

2.1.2 Merging predictions

The labelled segmentations are merged with the results of nnDetection. This is done by converting the bounding box coordinates of nnDetection into a binary label map. Subsequently, the binary label map and the numbered rib segmentations are merged. Bounding boxes that do not overlap with the numbered rib segmentations are discarded and in case of overlap, the fourth label, indicating the rib number, is assigned. This concludes the DCRibFrac model version 1.0. For a more in-depth explanation of DCRibFrac v1.0, please refer to (33).

2.2 CWIS Classification Improvement

The effectiveness and reliability of DL models are influenced by both the quantity and quality of the data used for training. Large and diverse datasets enable models to learn more effectively, generalise better to new data and provide accurate predictions (37,38). Therefore, to enhance the performance of the CWIS classification, the current dataset will be expanded.

2.2.1 Data Acquisition

For the development of DCRibFrac v2.0, available data from the multi-centre FixCon trial is included to expand the dataset. The most important inclusion and exclusion criteria for the FixCon trial are as follows (11):

Table 2: Inclusion and exclusion criteria of the FixCon study

Inclusion Criteria	**Exclusion criteria**
Age >18 years	Neurotraumatic changes leading to mechanical ventilation
Of ribs 4 t/m 10, either: - Simple (non-flail) fractures of 3 or more ribs, with at least one rib displaced by the width of the shaft - Simple (non-flail) fractures of 3 or more ribs accompanied by severe pain (VAS or NRS > 6)	Rib fractures due to cardiopulmonary resuscitation.
	SSRF is not possible due to additional traumatic injuries
	Flail chest, based on radiological or clinical findings
Blunt chest trauma	Decreased sensory or motor function due to (previous) cervical or thoracic spine failure.
Hospital presentation within 72 hours of trauma	Previous rib fractures or pulmonary problems

Figure 8: Map of hospital locations: blue markers indicate hospitals contributing data for the development of DCRibFrac v2.0. The red marker represents the hospital used for external validation.

Initially, all available FixCon patients were included in the dataset for the development of DCRibFrac v2.0. Patients were excluded if their post-trauma thoracic CT scans did not cover all ribs. The FixCon dataset consists of data from thirteen Dutch hospitals, including ErasmusMC, Amphia Hospital, Bravis Hospital, Catharina Hospital, Deventer Hospital, Haga Hospital, Ikazia Hospital, Isala Hospital, Maasstad Hospital, Maastricht University Medical Centre, Rijnstate Hospital, Spaarne Hospital, and University Medical Centre Groningen. These hospitals, marked by the blue markers in Figure 8, cover a large area of the Netherlands. Incorporating data from multiple centres provides a more extensive dataset, which can enhance the model's performance. Multi-centre data encompasses a wider variety of patient demographics and clinical conditions. This diversity helps ensure the model is not overfitted to a specific patient population and is more likely to perform well across different groups.

2.2.2 Ground Truth

To obtain the ground truth of the internal test and training set, all rib fractures were labelled manually according to the CWIS taxonomy. For the development of DCRibFrac v1.0, the dataset from EMC has already been labelled by a single researcher. The additionally included FixCon data was manually labelled using the MeVisLab tool developed for DCRibFrac v1.0 (33). This semi-automatic tool stores the coordinates and manually classified labels for each fracture in the required format. Specifically, it records each fracture's coordinates, type, displacement, location, and rib number. A comprehensive overview of the labelling software and the application in this project is explained in Appendix C.

Labelling is done independently by two researchers (VM and MvD). In cases where there was disagreement on the type or displacement class label, an experienced trauma surgeon (MW) solved the disagreement to obtain an accurate ground truth dataset. When one of the two labellers missed a fracture, the trauma surgeon made the final determination on whether a fracture was present. The classification of the fracture location (anterior/lateral/posterior) by VM was based on a measurement, as explained in Appendix C, while MvD's classification was done subjectively. Therefore, VM's classification is considered the truth for this category and disagreements were not solved by the experienced trauma surgeon.

The inter-observer agreement between the labellers will be evaluated using Krippendorff's Alpha and Cohen's Kappa. Krippendorff's Alpha handles categorical variables as well as missing data. Cohen's Kappa is designed to measure the agreement between two raters. Together, these metrics provide a comprehensive evaluation of the consistency and reliability of the labelling process (39,40).

2.2.3 Training and Evaluation

The training dataset, which includes the CT scans and their corresponding ground truth, was used to train three new nnDetection models. To train the three nnDetection models, a five-fold cross-validation strategy was implemented for each model. The post-processing steps for nnDetection remained the same as in DCRibFrac v1.0. The model's ability to automatically detect and classify fractures will be evaluated

by comparing its predictions on the internal test dataset against its ground truth. The detection performance will be assessed on fracture level using sensitivity, precision, F1-score, and false positives per scan (FPPS) as evaluation metrics. A false-positive was defined as a 3D bounding box that did not overlap with the midpoint of a blob in the ground truth. For the classification performance, sensitivity, precision, F1 scores and confusion matrices were utilised to present the results.

2.3 Rib number labelling

In DCRibFrac v1.0, the rib number labelling task had favourable results for patients with minor displaced rib fractures, but it was ineffective for those with severely dislocated ribs. In eleven out of nineteen patients in the internal test set, the labelling of the segmented ribs was inaccurate for at least one rib. In this report, the rib number labelling approach proposed in DCRibFrac v1.0 will be referred to as nnUNet-PP (nnUNet, followed by post-processing), see Figure 5. Two novel approaches for rib number labelling (TotalSegmentator and nnUNetOnly) were developed and evaluated, explained in the following sections.

2.3.2 TotalSegmentator

Recently, the TotalSegmentator tool was released, a pre-trained DL segmentation model based on nnUNet (41). This tool automatically segments and labels all major anatomical structures in the body on CT scans, including the left and right ribs 1 to 12. TotalSegmentator has demonstrated a mean Dice score of 0.94 on the test set for many anatomical structures, ranging from the skeleton to the cardiovascular system to the gastrointestinal tract. It was developed using a large and diverse dataset of over 1,200 CT scans, which included pathological cases such as fractured bones. Detailed information on fracture characteristics and the causes of the fractures (e.g., whether they were traumatic or pathological) was not available. The performance suggests that the model is promising and worth considering. However, the study presenting the tool notes some limitations, such as occasional confusion between neighbouring ribs. Despite these shortcomings, TotalSegmentator is expected to provide improved accuracy compared to nnUNet-PP and its potential should be further evaluated. This model is the basis for the two new approaches developed and assessed in this study. The first novel approach is based solely on TotalSegmentator. The pipeline for obtaining a labelled segmentation using TotalSegmentator is shown in Figure 9. CT scans are fed to the TotalSegmentator model, which results in 24 segments, one for each rib. These labelled segments can be combined with the results of the nnDetection models during post-processing, as done in DCRibFrac v1.0.

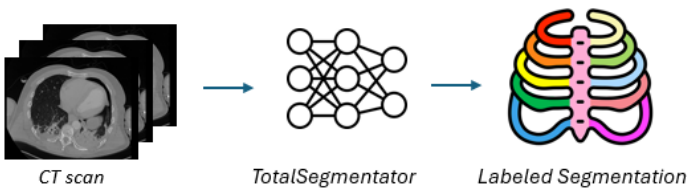


Figure 9: Pipeline to obtain labelled rib segmentation using TotalSegmentator.

2.3.3 nnUNetOnly

The second novel approach involves the development of a new nnUNet based on the results from TotalSegmentator and is named nnUNetOnly. In this process, the results from TotalSegmentator are manually refined using 3D Slicer to create a correct training set. This refined dataset is utilized to develop and train a new nnUNet model. The training dataset used to develop the three nnDetection models is also employed for both the development and evaluation of nnUNetOnly. This dataset is divided into a training and validation dataset. The training set was used to train the nnUNet, while the validation set was used to assess the performance of nnUNetOnly by comparing its performance with TotalSegmentator, see Section 2.3.5. Figure 10 shows the workflow for the development and evaluation of this model. The aim of nnUNetOnly is to create a model which directly and automatically segments and labels each rib separately, see Figure 11.

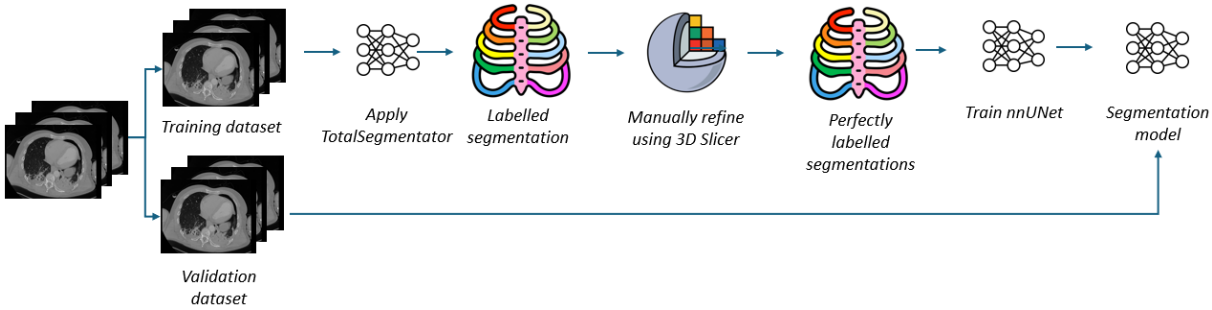


Figure 10: Workflow for the development of nnUNetOnly

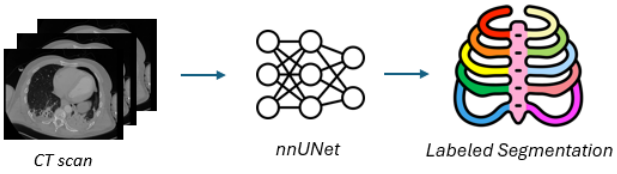


Figure 11: Pipeline to obtain labelled rib segmentation using nnUNetOnly.

2.3.4 Experiment: nnUNet-PP versus TotalSegmentator

To evaluate the performance of TotalSegmentator in rib number labelling, the segmentations will be compared to the results of nnUNet-PP. The evaluation is based on 30 EMC CT scans. The selected CT scans were not utilised for the development of the nnUNet in nnUNet-PP. The ground truth for rib number labelling was established by one researcher (VM), who created a spline by manually tracing and labelling each rib in the axial slices of the CT scans using 3D Slicer software 5.6.2 (42). The splines are created with the *Curve* function under the *Markups* tab. Each spline was created by following the rib from cranial to caudal, starting at the first rib. Figure 12A illustrates an axial slice with these splines overlaid, while Figure 12B presents a 3D model of the splines. To assess the accuracy of rib number labelling generated by nnUNet-PP and TotalSegmentator, the labelled splines (GT) were projected onto the same 3D image space as the segmentations. A qualitative assessment was then conducted through visual inspection, see Figure 12C. Evaluation is conducted for each rib. During qualitative analysis in 3D Slicer, the following aspects are assessed:

- Does the label of the spline match the label of the segment it intersects?
- Does the spline intersect with multiple segments or no segments at all?

If the ground truth spline intersects with a segmentation that carries the same label, the rib number label for that rib is categorised as correct. Conversely, if a spline intersects with a segment where the labels do not match or does not intersect with a segment at all, that rib is categorized as incorrect.

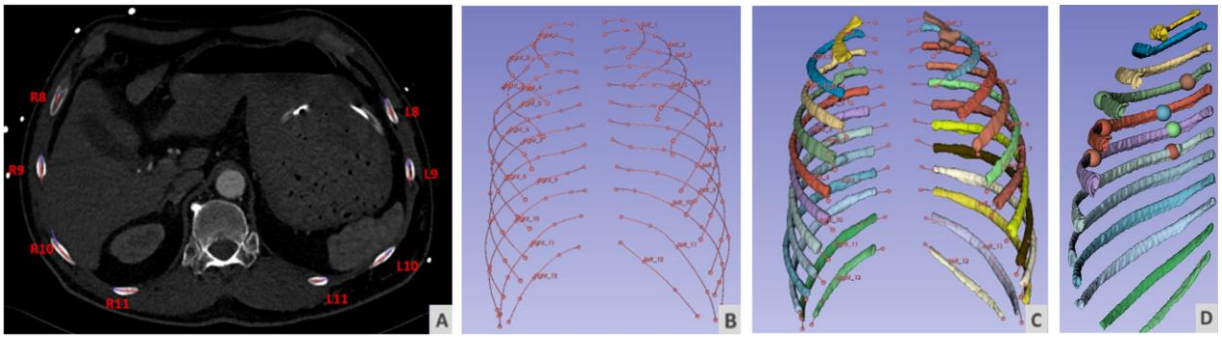


Figure 12: Example of how the ground truth was created and used for evaluation. A) axial slice with labelled pink splines. B) display of 3D model of all splines in one patient. C) segmentation overlaid with the splines. D) Segmentation overlayed with blobs which represent fractures.

This evaluation process is applied to the results of both nnUNet-PP and TotalSegmentator, allowing for a comparative analysis to determine the superior approach. However, the clinical relevance lies more in the accurate labelling of specifically fractured ribs. Label maps containing the location of fractures are available, as they were created for the development of the nnDetection models. These label maps contain spheres which represent the location of a fracture. By visualising these spheres in 3D Slicer, the performance of the approaches specifically for fractured ribs will be evaluated (Figure 12D).

2.3.5 Experiment: TotalSegmentator versus nnUNetOnly

This experiment aimed to determine which approach is superior and should be implemented in DCRibFrac 2.0, TotalSegmentator or nnUNetOnly. The ground truth is established in a similar way as done in the previously explained experiment, where nnUNet-PP and TotalSegmentator were compared. Instead of comparing the segmentations of TotalSegmentator and nnUNetOnly with splines, they are compared with labelled control points for qualitative evaluation, since creating splines is time-consuming. This comparison is performed using 3D Slicer as well, through the *Point List* function under the *Markups* tab. The control points are placed in the middle of the rib bow, with one control point assigned per rib. During the qualitative analysis in 3D Slicer, the following aspects are evaluated by visual inspection:

- Does the label of the control point match the label of the segment it overlaps?
- Is it feasible that the form of the segment represents a single rib?
- In the case of severely dislocated fractures: Do the two fragments have the same label?

TotalSegmentator and nnUNetOnly will be evaluated by comparing their results to this ground truth using the validation dataset, as explained in Section 2.3.3. The evaluation will be done on a per-rib level, focusing on fractured ribs. Based on the results of this experiment, the best method will be implemented in DCRibFrac v2.0. Figure 13 provides an overview of the data utilisation, and the experiments performed in this project. The dotted line encircles the contributions of this project.

2.4 External validation

Apart from the internal validation using an internal test set, the final algorithm is evaluated by performing an external validation. External validation is needed to assess a model's reproducibility and generalizability, which is necessary in a clinical setting. ML models must be robust, meaning they should reliably perform even in contexts that may differ subtly from those represented in the training data (43). Performing an external validation will give insight into the robustness and reliability of DCRibFrac v2.0. Frequently, the performance observed on external datasets is poorer than the performance appraised on original datasets (44). The external data, in this case, refers to a set of new data from another hospital which is not used for the development of DCRibFrac v2.0. The CT scans in the FixCon RCT originating

from Zuyderland Hospital are used for external validation, see the red pinpoint in Figure 8. This dataset was chosen due to its size, making it the second-largest dataset available, aside from the EMC dataset.

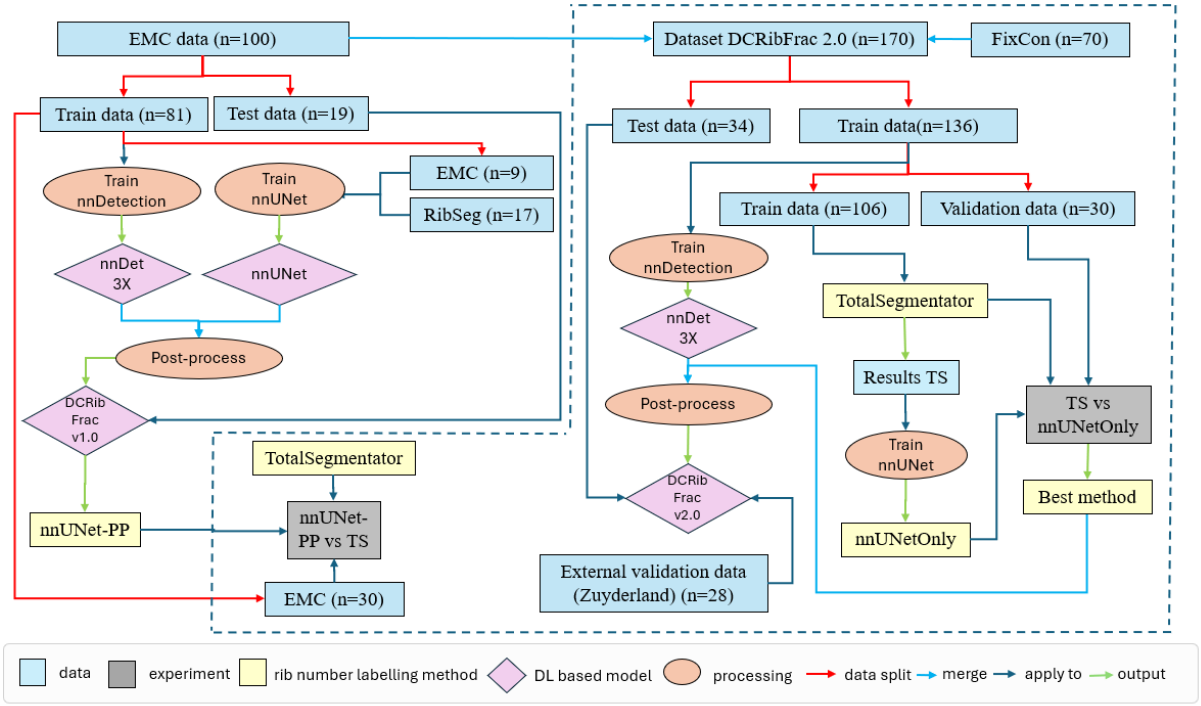


Figure 13: Flowchart illustrating the data utilization and model development process for automatic rib fracture detection and classification. The dotted line encircles the contributions of this project.

3. Results

This section presents the outcomes of the experiments detailed in Section 2. Section 3.1 provides an overview of the dataset used to enhance fracture detection, CWIS classification and rib number labelling. Section 3.2 discusses the findings from the interobserver agreement study. Section 3.3 outlines the results of the experiments where the three different rib number labelling approaches, nnUNet-PP, TotalSegmentator, and nnUNetOnly, were evaluated. Section 3.4 introduces the final pipeline and the performance of DCRibFrac v2.0. The results from the external validation are covered in Section 3.5. Finally, Section 3.6 summarises the CWIS classification performance across various DCRibFrac models and test sets.

3.1 Dataset

An additional 70 anonymised FixCon CT scans have been added to the ErasmusMC dataset, bringing the total to 170 CT scans, including 1509 fractures. The dataset characteristics of the included CT scans for the development of DCRibFrac v2.0 are presented in Table 3. Refer to Appendix D for details on the data characteristics of the EMC dataset, used to develop the initial model, DCRibFrac v1.0. Additionally, Appendix D visualises the distribution of fracture characteristics per rib for both the EMC and the added FixCon datasets separately, providing insights into how the dataset has expanded. The label distribution of the added FixCon data is similar to the EMC data. The fraction of complex and anterior fractures is slightly smaller in the added FixCon dataset. All scans have an in-plane image size of 512x512, and the number of slices varies between 195 and 1680. For the development of DCRibFrac v2.0, the dataset is split into an internal training set and an internal test set using an 80/20 split based on the number of fractures, respectively. To ensure class balance, stratified sampling is performed, focusing on the minority classes.

Table 3: Data characteristics internal train and test set

Variables	**Internal training set**	**Internal test set**
No. patients (%)	136 (80)	34 (20)
Amphia Hospital Breda	12 (9)	1 (3)
Bravis Hospital Bergen op Zoom	3 (2)	-
Catharina Hospital Eindhoven	1 (1)	-
Deventer Hospital	1 (1)	-
Erasmus Medical Centre Rotterdam	77 (57)	23 (68)
Haga Hospital The Hague	9 (6)	2 (6)
Ikazia Hospital Rotterdam	2 (1)	-
Isala Hospital Zwolle	3 (2)	3 (9)
Maasstad Hospital Rotterdam	20 (15)	1 (3)
Maastricht University Medical Centre	3 (2)	2 (6)
Rijnstate Hospital Arnhem	-	1 (3)
Spaarne Hospital Haarlem	4 (3)	1 (3)
University Medical Centre Groningen	1 (1)	-
No. fractures (%)	1203 (80)	306 (20)
No. fractures for Type (%)		
- Simple	900 (75)	235 (77)
- Wedge	203 (17)	53 (17)
- Complex	100 (8)	18 (6)
No. fractures for Displacement (%)		
- Undisplaced	696 (57)	178 (58)
- Offset	348 (29)	79 (26)
- Displaced	159 (13)	49 (16)

No. fractures for Location (%)		
- Anterior	171 (15)	49 (16)
- Lateral	606 (50)	167 (55)
- Posterior	426 (35)	90 (29)
No. CTs with slice thickness (%)		
- > 2 mm	47 (35)	17 (50)
- = 1 mm	33 (24)	4 (12)
- < 1 mm	55 (41)	13 (38)
No. CTs with pixel spacing (%)		
- > 0.9 mm	23 (17)	4 (12)
- 0.7< x <0.9 mm	83 (61)	20 (58)
- < 0.7 mm	30 (22)	10 (30)

3.2 Interobserver Agreement & Ground Truth

3.2.1 Interobserver Agreement

The interobserver agreement study was conducted based on the FixCon dataset, excluding the EMC data. Two observers independently detected and classified rib fractures according to the CWIS taxonomy. In total, there were 519 rib fractures noted by the two observers, of which 467 rib fractures were seen by both observers. Out of the 52 fractures detected by only one observer, 16 were identified by observer 1, while the remaining 36 were detected by observer 2. To evaluate interobserver agreement on fracture type, displacement, and location, Cohen's Kappa statistic was calculated based on the 467 fractures (Table 4). Additionally, Krippendorff's Alpha was calculated to account for missing data. Disagreements were observed in 144 fractures for 162 classification tasks, with 40 disagreements in type classification, 52 in displacement classification, and 70 in location classification.

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

Label	Cohens Kappa (95%CI)	Krippendorff's Alpha (95% CI)	Interpretation
Type	0.74 (0.67 - 0.82)	0.76 (0.69 - 0.82)	Substantial
Displacement	0.82 (0.78 - 0.87)	0.82 (0.78 - 0.87)	Strong
Location	0.74 (0.69 - 0.80)	0.73 (0.68 - 0.79)	Substantial

3.2.2 Detection and Classification Accuracy

Observer 1 (VM) identified 483 fractures, while observer 2 (MvD) identified 503 fractures. For the 52 fractures detected by only one observer, observer 3 made the final determination on whether a fracture was present. Out of the 16 fractures detected solely by observer 1, 12 were confirmed as actual fractures, resulting in 4 false positives by observer 1. For the 36 fractures detected by observer 2, 16 were confirmed as actual fractures by observer 3, resulting in 20 false positives. Assuming that all fractures are detected by either observer 1 or observer 2, the overall detection sensitivity is 96.7% for observer 1 and 97.4% for observer 2, with a precision of 99.3% for observer 1 and 96% for observer 2.

Table 5 shows the percentage of agreements between observer 3 and either observer 1 or observer 2 regarding classification. For simple fractures, observer 3 agreed with observer 1 in 83.3% of cases, which is higher than the 16.7% agreement with observer 2. In contrast, for wedge fractures, observer 3 agreed more often with observer 2 (64.7%) compared to observer 1 (35.3%). Since both observers have different strengths, combining their decisions will lead to higher overall accuracy of the ground truth.

Table 5: Agreement in CWIS classification between observer 3 and either observer 1 or observer 2

Label	Observer 1 (%)	Observer 2 (%)
Simple	83,3	16,7
Wedge	35,3	64,7
Complex	100,0	0,0
Undisplaced	90,0	10,0
Offset	67,9	31,3
Displaced	21,4	78,6

3.3 Results Rib Number Labelling Experiments

3.3.1 Qualitative Evaluation nnUNet-PP versus TotalSegmentator

A total of 716 ribs were labelled and evaluated in 30 patients, accounting for the fact that two patients had only 11 pairs of ribs. Among these, 316 fractures were identified in 233 ribs with the majority of the fractures occurring in the 3rd to 7th ribs. TotalSegmentator labelled 94% of all ribs correctly, which was 54% for nnUNet-PP. Focusing specifically on the fractured ribs, nnUNet-PP correctly labelled 50% of fractured ribs, while TotalSegmentator labelled 94% of fractured ribs correctly. Figure 14 shows the percentage of correctly labelled ribs on the y-axis, with the corresponding number of ribs displayed within the bars.

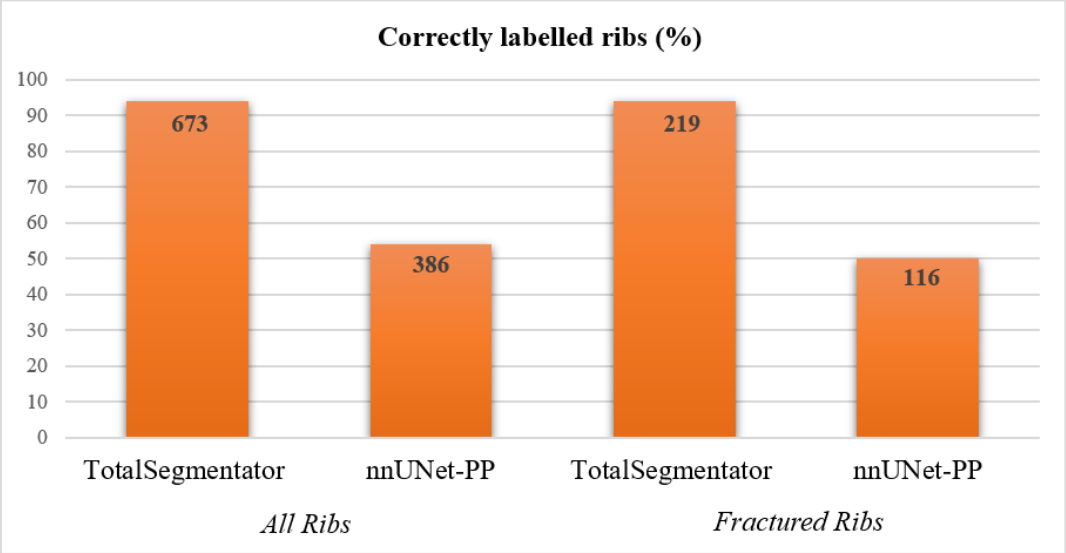


Figure 14: Fraction of correctly labelled (fractured) ribs using nnUNet-PP and TotalSegmentator

Relabelling of the rib segments provided by TotalSegmentator was necessary since the initial labelling of TotalSegmentator is consistently incorrect. For instance, the first left rib (L1) was always mislabelled as the third right rib (R3), without exception. Consequently, all labels were automatically corrected and renamed before analysis.

In two out of the six patients with incorrectly labelled ribs using TotalSegmentator, the mislabelling was due to the patients having 11 pairs of ribs rather than the standard 12 pairs, which was something TotalSegmentator apparently could not process well. This issue accounts for 17 of the 43 inaccurately labelled ribs. nnUNet-PP effectively addresses this problem, as it does not require 24 classes and classifies all ribs in these patients correctly. Therefore, in cases with 11 pairs of ribs, nnUNet-PP appears to be the superior method. However, the prevalence of having 11 pairs of ribs is only 3.4% (45). Figure 15A illustrates an example of a patient with 11 pairs of ribs and the corresponding results from TotalSegmentator. Misclassification of a single rib often led to additional inaccuracies in the labelling of other ribs for the same patient using TotalSegmentator.

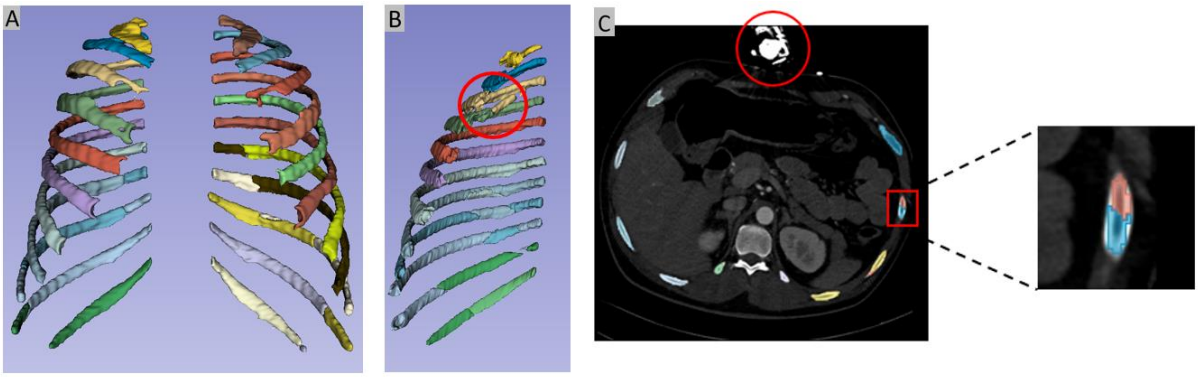


Figure 15: Examples of wrongly labelled ribs using TotalSegmentator. A) Patient with 11 pairs of ribs, B) Part of a drain misclassified as a rib, C) Misclassification due to scattering.

Among the remaining 26 misclassified ribs by TotalSegmentator, 8 were from two patients who had thorax drains at the time of the CT scan. Parts of these drains were mistakenly classified as ribs, resulting in multiple errors (Figure 15B). In another case, a metal object resting on the patient caused scattering, resulting in the misclassification of 12 ribs (Figure 15C). For the remaining patient, the misclassifications of 6 ribs could not be explained other than by severe displacement of fractures.

Figure 16A shows an example of a correct segmentation and labelling result by TotalSegmentator. In the case of dislocated ribs, nnUNet-PP is not able to accurately label the ribs, which explains its low performance. In Figure 16B, displaced fragments of the fractured 8th and 9th left ribs, encircled in red, are classified incorrectly as separate fragments by nnUNet-PP and correctly as a single fragment by TotalSegmentator (Figure 16C). Using nnUNet-PP often leads to merged rib segments (Figure 16D) and excessive thickness due to the morphological operations applied during postprocessing (Figure 16E). In contrast, TotalSegmentator does not exhibit this issue (Figure 16F). Overall, TotalSegmentator demonstrates a clear advantage over nnUNet-PP in terms of rib number labelling accuracy. It is robust and performs well, particularly in standard cases, with 12 pairs of ribs.

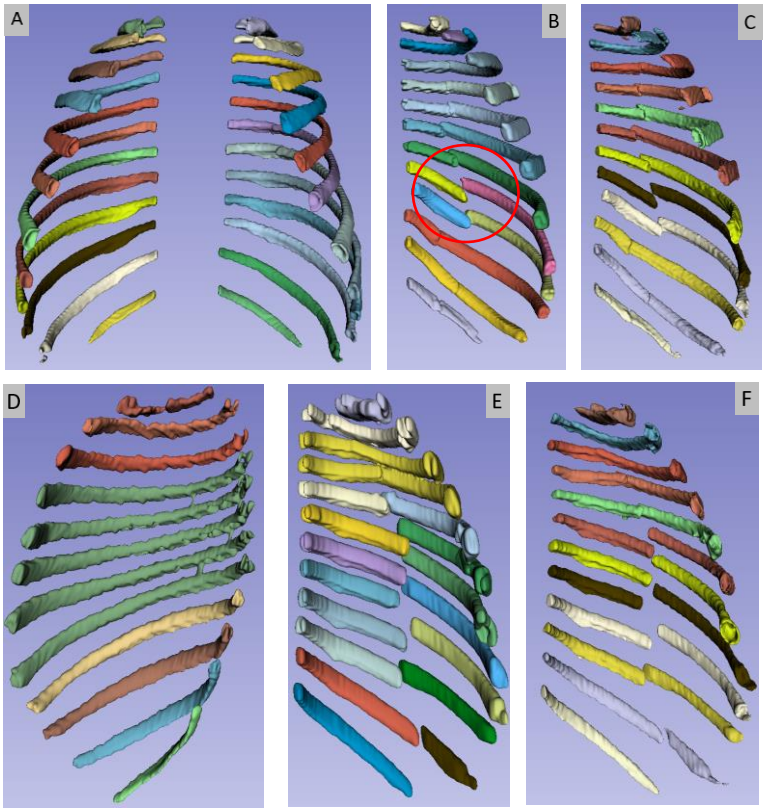


Figure 16: Example of results using nnUNet-PP (B, D, E), and TotalSegmentator (A, C, F).

3.3.2 Qualitative Evaluation TotalSegmentator versus nnUNetOnly

The developed nnUNet was designed to handle a variable number of ribs, unlike TotalSegmentator, which required assigning 24 classes (12 pairs of ribs). This flexibility may allow the nnUNet to improve rib number labelling in patients with only 11 pairs of ribs. A total of 716 ribs (30 patients) were evaluated, including two cases where only 11 pairs of ribs were present. Among these 716 ribs, 246 ribs were fractured. TotalSegmentator correctly labelled 92,5% of all ribs and 95,5% of the fractured ribs. In comparison, nnUNetOnly labelled 95,5% of all ribs and 98,4% of the fractured ribs correctly. Figure 17 shows the percentage of correctly labelled ribs on the y-axis, with the corresponding number of ribs displayed within the bars.

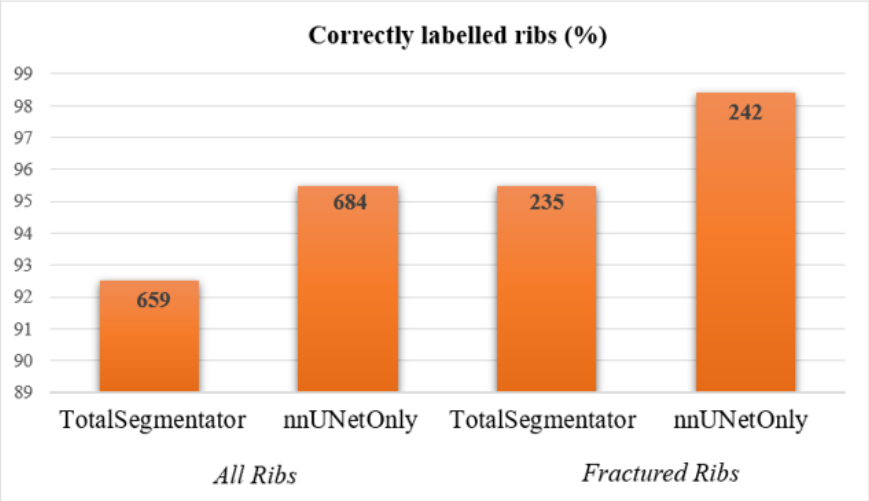


Figure 17: Number of incorrectly labelled (fractured) ribs using TotalSegmentator and nnUNetOnly.

There was at least one incorrectly labelled rib in 7 patients when using nnUNetOnly, of which 5 were labelled incorrectly by TotalSegmentator as well. TotalSegmentator had at least one error in 11 patients. Figure 18(A-D) illustrates two cases where nnUNetOnly outperformed TotalSegmentator. In Figure 18(A, B), the results for a patient with only 11 pairs of ribs are shown. nnUNetOnly correctly assigned 22 labels, while TotalSegmentator assigned 24 labels, leading to inaccurate rib number labels. However, for another patient with 11 pairs of ribs, nnUNetOnly was not error-free, since 6 ribs were mislabelled (Figure 18E). TotalSegmentator labelled 8 ribs incorrectly for this same patient (Figure 18F). Figure 18C presents an example of incorrect labelling of severely displaced fractures by TotalSegmentator. This highlights the challenges in accurate rib labelling under complex conditions, which nnUNetOnly managed better (Figure 18D). There were two cases where TotalSegmentator outperformed nnUNetOnly. Figure 18H illustrates one of these cases, where the 9th right rib is labelled incorrectly by nnUNetOnly but correctly by TotalSegmentator. Additionally, the rib segments are notably thin at certain points, compared to the segmentation of TotalSegmentator.

Even though the cause of mislabelling cannot always be determined, these results indicate that nnUNetOnly performs better than TotalSegmentator. McNemar's test statistics show a significant difference with a p-value = 0.023 based on the rib number labelling results of fractured ribs. Consequently, nnUNetOnly has been incorporated into the pipeline of DCRibFrac v2.0.

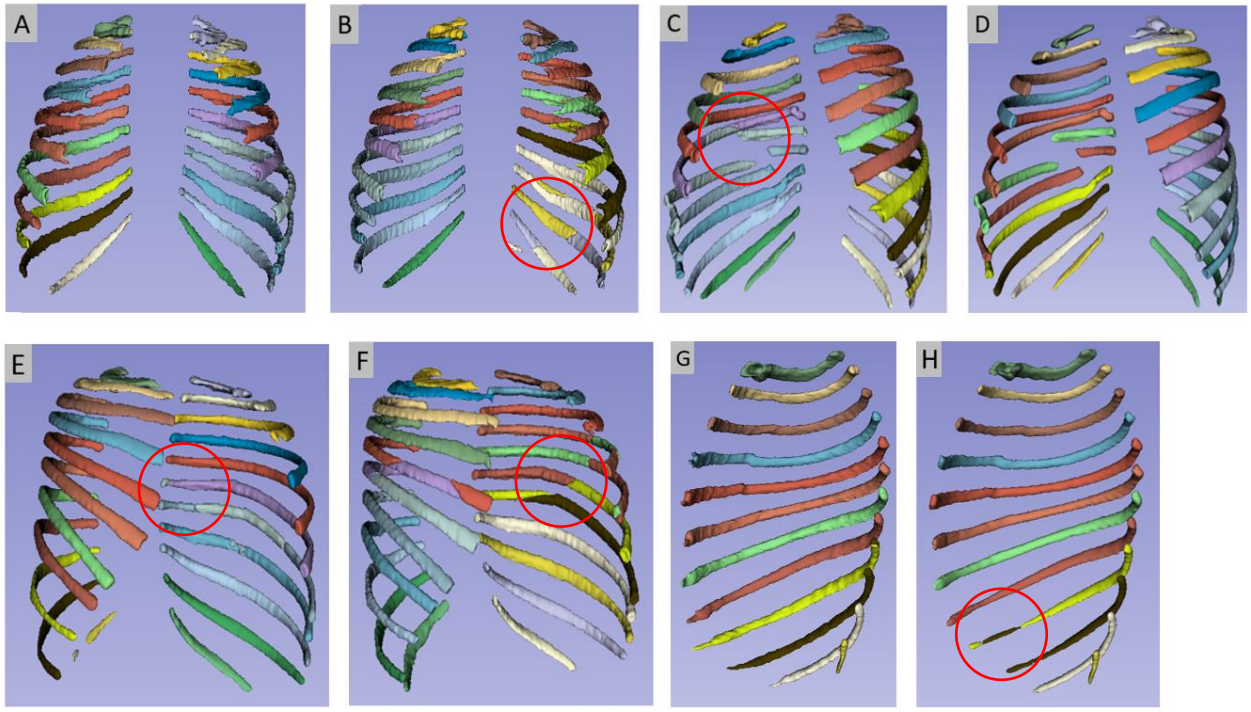


Figure 18: Examples of results using TotalSegmentator (B, C, F, G) and nnUNetOnly (A,D, E, H) of four patients.

3.4 Final Pipeline and Performance

Figure 19 represents the final pipeline of DCRibFrac v2.0. The newly developed nnUNet (nnUNetOnly) has replaced the initial nnUNet and postprocessing steps (nnUNet-PP). Additionally, the nnDetection models have been retrained with more data to enhance performance. The ensembling methods, however, remain unchanged from those used in DCRibFrac v1.0.

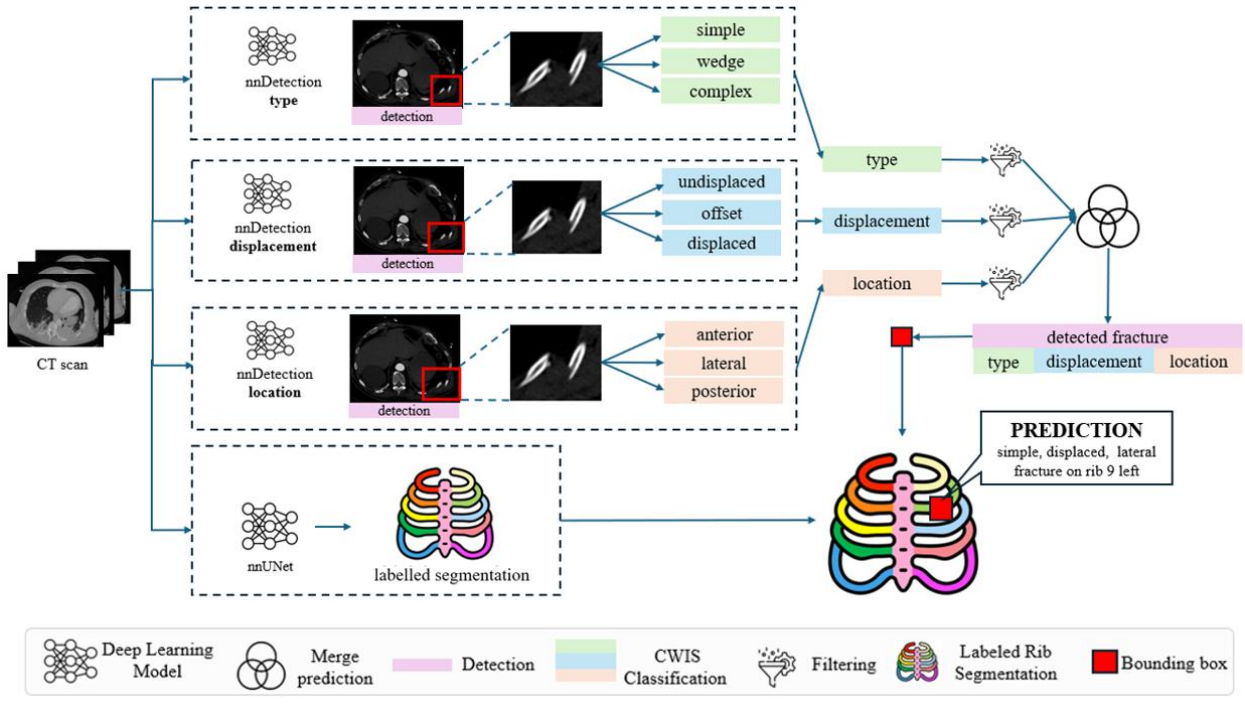


Figure 19: Pipeline of DCRibFrac v2.0

3.4.1 Validation Results

The precision-recall (PR) curve of the validation set for DCRibFrac v2.0, along with the corresponding thresholds, was analysed to gain insight into the performance of the ensembled nnDetection models (Figure 20). The probability score threshold required to achieve a target sensitivity of 82% (dotted vertical line) was determined based on this curve. This sensitivity surpasses the range typically observed among clinicians, which is between 73.2% and 80.8% (22–24). Figure 20 compares the PR curves of DCRibFrac v1.0 and DCRibFrac v2.0. The increased area under the curve (AUC) for DCRibFrac v2.0 reflects an overall improvement in model performance, indicating better precision across various recall levels. At the desired sensitivity of 82%, the threshold for DCRibFrac v2.0 is higher than it was for v1.0. This higher threshold translates to a more selective model, resulting in improved precision compared to DCRibFrac v1.0, reducing the likelihood of false positives while maintaining a similar recall.

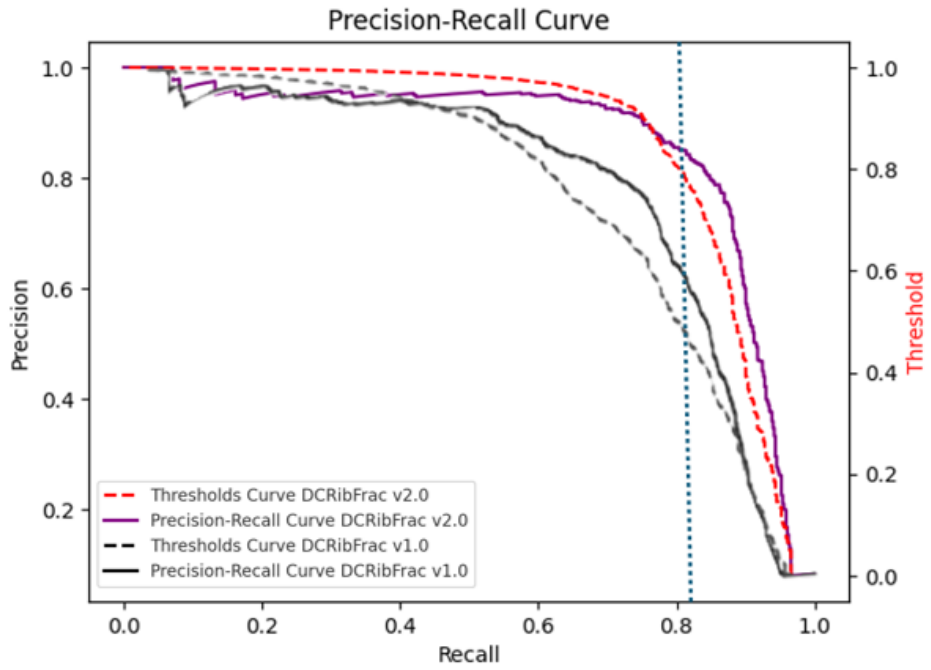


Figure 20: Precision-Recall curve CWIS classification of DCRibFrac v1.0 and DCRibFrac v2.0

3.4.2 Performance DCRibFrac v2.0 on Internal Test Set

To assess the performance of DCRibFrac v2.0 on unseen data, the rib fracture detection, CWIS classification and rib number labelling results on the internal test set were evaluated. DCRibFrac v2.0 achieved a detection sensitivity of 80%, a precision of 87%, and an F1 score of 83% on the internal test set, with a mean FPPS of 1.11. All classification labels were successfully assigned, as no cases involved only two overlapping nnDetection models. False positives resulted from old fractures, indicated by callus formation around the fracture (Figure 21D). Misclassifications also occurred in the region where the anterior rib transitions to the costal cartilage, due to irregularities in the cortical bone in this region.

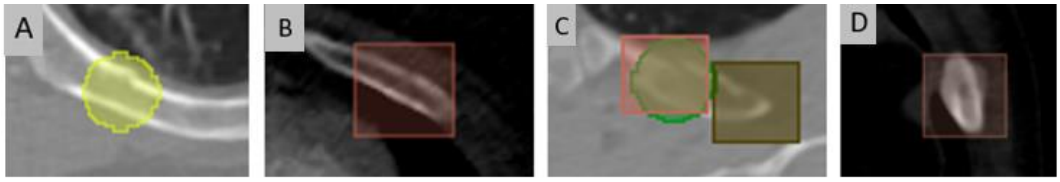


Figure 21: Example of a missed fractures (A), a true positive (B), a fracture classified as two fractures resulting in a false positive (C), a false positive due to callus formation (D). Squares indicate the detected fracture by DCRibFrac v2.0 and circles indicate fractures labelled as the ground truth.

Additionally, large fractures (often with displacement) were assigned two fracture labels instead of one, leading to false positives (Figure 21C). Missed fractures are frequently undisplaced fractures with small interruption of the cortical bone (Figure 21A). The qualitative evaluation of the detection of rib fractures revealed three additional true positives that were not labelled in the ground truth (Figure 21B). Table 8 presents the characteristics of the missed fractures and the fraction per class that was missed. Figure 22 illustrates the classification performance per class of all detected fractures in a confusion matrix. For type classification of the detected fractures, simple fractures had a sensitivity of 95% and a precision of 87%. Wedge fractures had a sensitivity of 36% and a precision of 47%, while complex fractures had a sensitivity of 25% and a precision of 40%. Regarding displacement, the sensitivity and precision were 90% and 89% for undisplaced fractures, 80% and 69% for offset fractures, and 61% and 95% for displaced fractures. For fracture location, the sensitivity and precision were 79% and 93% for anterior fractures, 94% and 93% for lateral fractures, and 97% and 93% for posterior fractures.

Table 8: Characteristics of missed fractures in the internal test set

	Simple	Wedge	Complex	Undisplaced	Offset	Displaced	Anterior	Lateral	Posterior
Missed fractures (%)	82	15	3	66	8	26	24	58	18
Missed fractures per class (%)	22	17	11	23	6	33	30	22	12

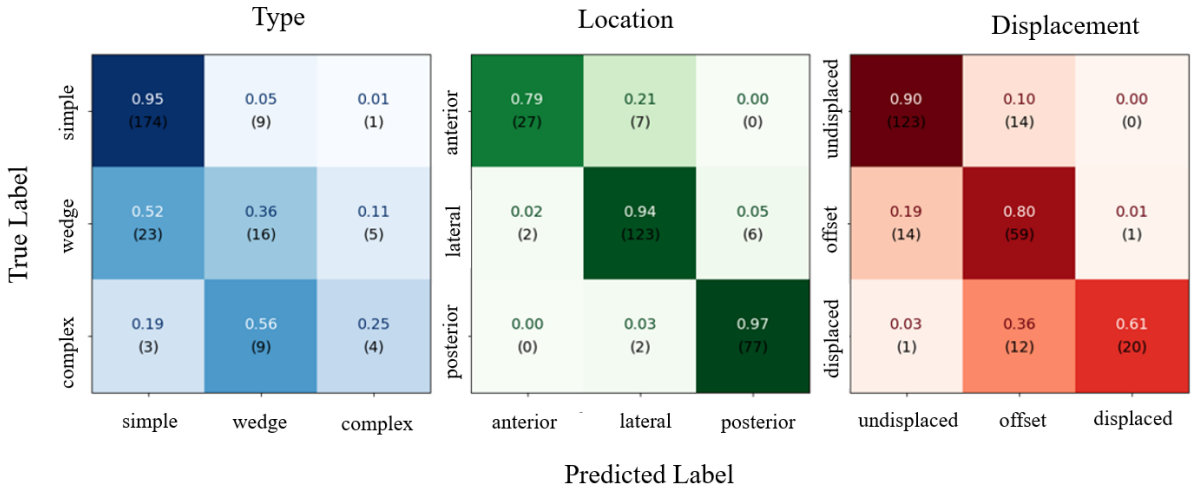


Figure 22: Confusion matrix of CWIS classification on the internal test set

Of the 244 detected fractures by nnDetection, 94% were assigned the correct rib number. Of the 15 misclassified ribs, 7 were due to a postprocessing step. During postprocessing, the rib segments are overlaid with the bounding box of each detected fracture. If there is an overlap between a rib segment and a bounding box, the bounding box is assigned to that rib. However, when neighbouring ribs are close to each other because of a displaced fracture or anatomical variation, the bounding box may intersect with multiple ribs. In this case, the more cranial rib will get assigned the bounding box, resulting in a misclassification. Figure 23 illustrates two examples of misclassifications due to this issue.

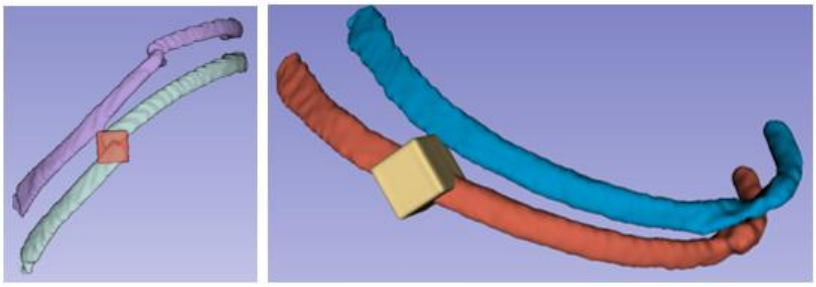


Figure 23: Two examples of predicted bounding boxes, which overlap slightly with a more cranial rib, resulting in incorrect rib number labelling.

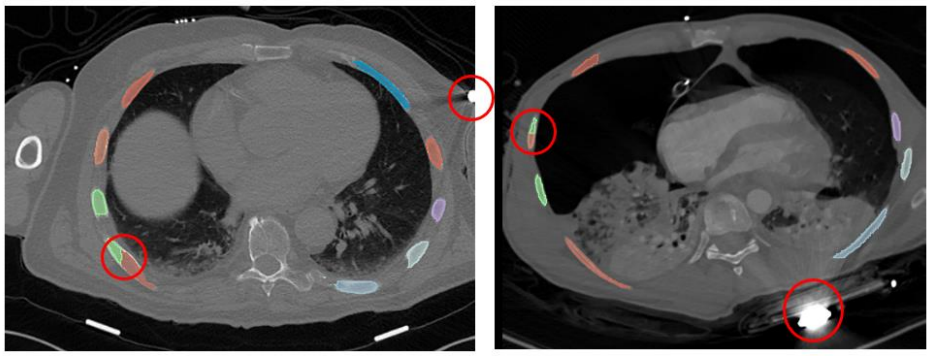


Figure 24: Two examples of incorrectly labelled segmentations by nnUNet due to an object causing scattering.

The remaining six misclassified ribs were the result of inaccurate labelling by the developed nnUNet in four patients. One of these patients had 11 pairs of ribs. Although 22 segments were created for this case, some ribs were still mislabelled, leading to one fracture being assigned to the wrong rib number. In two other patients, an object resting on the thorax caused scattering, resulting in four misclassifications (Figure 24). For the remaining patient, no clear reason for the misclassifications could be identified.

The runtime of DCRibFrac v2.0 for a single patient ranged from 20 to 90 minutes on the GPU cluster at Erasmus MC, utilising the 2090 Ti 11GB and Nvidia A40 48GB GPUs when the nnDetection models and the nnUNet model were run parallel. Detection and classification with the nnDetection models took between 7 and 80 minutes per scan for a single classification task (type, displacement, location), depending on the number of slices, with an average runtime of 23 minutes per scan. Segmentation using the nnUNet model took approximately 2 minutes per scan, while postprocessing required between 5 and 15 minutes.

3.4.3 DCRibFrac v1.0 versus DCRibFrac v2.0

To ensure a fair one-on-one comparison of detection and classification performance between DCRibFrac v1.0 and DCRibFrac v2.0, both models were evaluated on the internal test set of DCRibFrac v1.0, which consists of 19 EMC patients with 207 fractures. This comparison gives insight into the impact of incorporating additional and multi-centre datasets on the performance of DCRibFrac (Table 6). However, since DCRibFrac v2.0 was trained using a subset of this internal test set, the nnDetection models were retrained with a slightly adjusted training and testing split, excluding the images that were part of the test set.

The detection sensitivity, precision and F1 score on the EMC test set were 79%, 88% and 83%, respectively, with a FPPS rate of 1.16. The performance of DCRibFrac v1.0 on this same internal test set was 77%, 79%, 78% and 2.26 respectively. Table 6 presents the percentage of fractures per class which were not detected. The last column represents the difference in detection performance per class, with positive values indicating improvement of detection sensitivity using DCRibFrac v2.0 for that specific class, compared to v1.0. The confusion matrix in Figure 25 visualizes the results of the classification performance per class of DCRibFrac v1.0 and v2.0 on the internal test set of DCRibFrac v1.0.

Table 6: Percentage of missed fractures per class on the internal test set of DCRibFrac v1.0

	DCRibFrac v1.0	DCRibFrac v2.0	Difference (%)
Simple	28	28	0
Wedge	8	0	8
Complex	18	14	4
Undisplaced	28	26	2
Offset	9	6	3
Displaced	44	48	-4
Anterior	24	5	19
Lateral	19	20	-1
Posterior	29	29	0

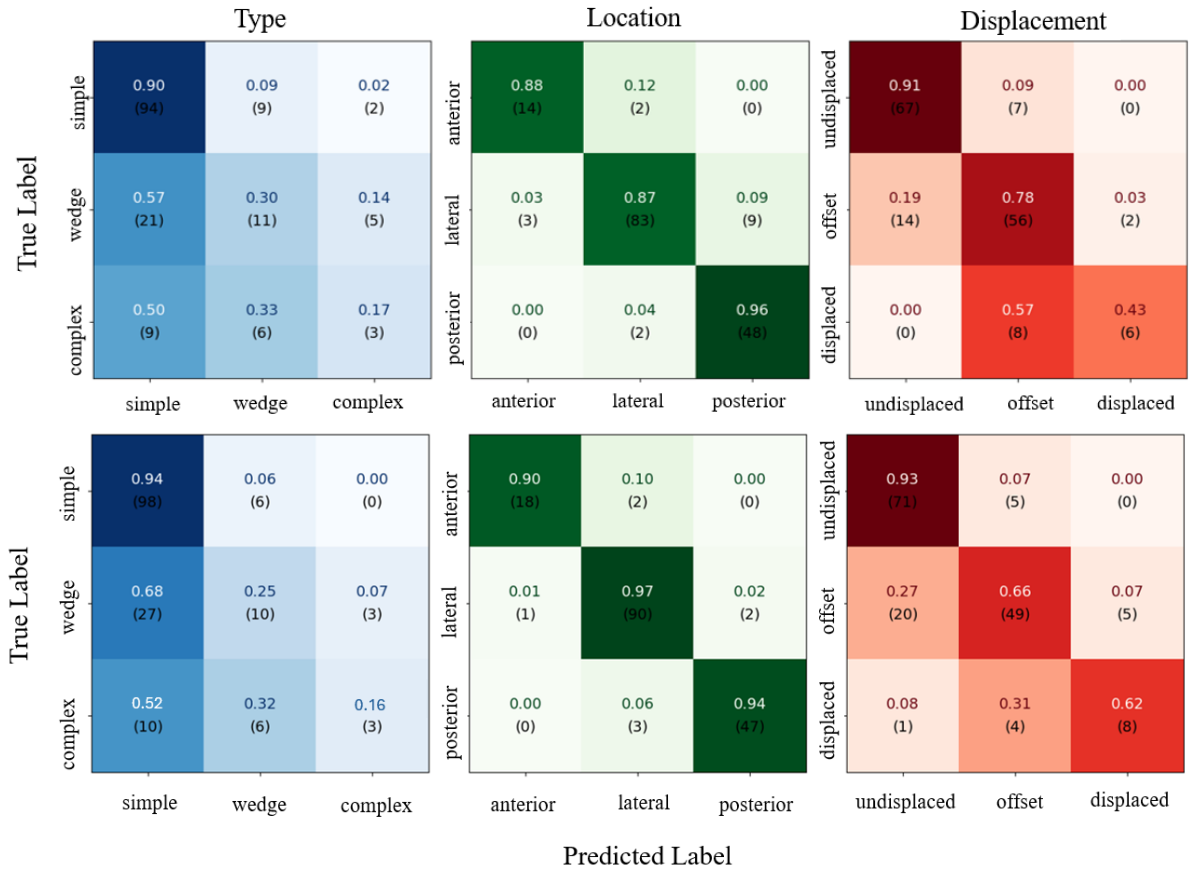


Figure 25: Confusion matrix showing the performance of DCRibFrac v1.0 (top) and DCRibFrac v2.0 (bottom) on the internal test set of DCRibFrac v1.0 (19 EMC patients)

3.5 External Validation

The pipeline of DCRibFrac v2.0 was applied to the dataset from Zuyderland Hospital, to assess its performance on external data. As done for the internal test and training set of DCRibFrac v2.0, the external validation dataset was labelled by two observers (VM and MvD). If there was a conflict in classification, an experienced trauma surgeon (MW) solved the disagreement to obtain an accurate ground truth dataset. The image size of the external validation dataset was 512x512 for every patient. The number of slices differed between 251 and 1414 slices. Table 7 shows the data characteristics of the Zuyderland dataset and the data of the internal test set for comparison. For more details on the external validation dataset, please refer to Appendix E. Interestingly, the distribution of fractures within this dataset shows a higher prevalence of rib fractures on the left side.

Table 7. Data characteristics Zuyderland Hospital

Variables	**External test set**	**Internal test set**
No. patients	28	36
No. fractures	193	309
No. fractures for Type (%)		
- Simple	134 (70)	235 (77)
- Wedge	33 (17)	53 (17)
- Complex	26 (13)	18 (6)
No. fractures for Displacement (%)		
- Undisplaced	99 (51)	178 (58)
- Offset	56 (29)	79 (26)
- Displaced	38 (20)	49 (16)

No. fractures for Location (%)		
- Anterior	5 (3)	49 (16)
- Lateral	102 (53)	167 (55)
- Posterior	86 (44)	90 (29)
No CTs with slice thickness (%)		
- >2.0 mm	1 (4)	17 (50)
- =1 mm	-	4 (12)
- <1 mm	27 (96)	13 (38)
No CTs with pixel spacing (%)		
- Pixel spacing > 0.9 mm	5 (18)	4 (12)
- Pixel spacing 0.7< x <0.9 mm	23 (82)	20 (58)
- Pixel spacing < 0.7 mm	-	10 (30)

The detection sensitivity on the external validation dataset was 84%, with a precision of 85%, an F1-score of 84% with a FPPS of 0.96. Table 8 presents the characteristics of the missed fractures and the percentage of missed fractures per class.

Concerning the classification performance of all detected fractures, simple fractures had a sensitivity of 93% and a precision of 83%, see Figure 26. Wedge fractures had a sensitivity of 47% and a precision of 43%, while complex fractures had a sensitivity of 22% and a precision of 83%. Regarding displacement, the sensitivity and precision were 88% and 81% for undisplaced fractures, 65% and 63% for offset fractures, and 67% and 86% for displaced fractures. For fracture location, the sensitivity and precision were 86% and 95% for lateral fractures, and 95% and 86% for posterior fractures.

Of the detected fractures, 8 were assigned an incorrect rib number, resulting in a classification accuracy of 95%. 5 out of these 8 can be explained by the bounding boxes of the predictions overlapping with the cranially neighbouring rib, as explained in Section 3.4.2. The remaining 3 cases are misclassified due to incorrect rib number labelling by nnUNet.

Table 8: Characteristics of missed fractures from the external validation dataset

	Simple	Wedge	Complex	Undisplaced	Offset	Displaced	Anterior	Lateral	Posterior
Missed fractures (%)	87	3	10	88	6	6	16	45	39
Missed fractures per class (%)	20	3	12	27	5	4	100	14	14

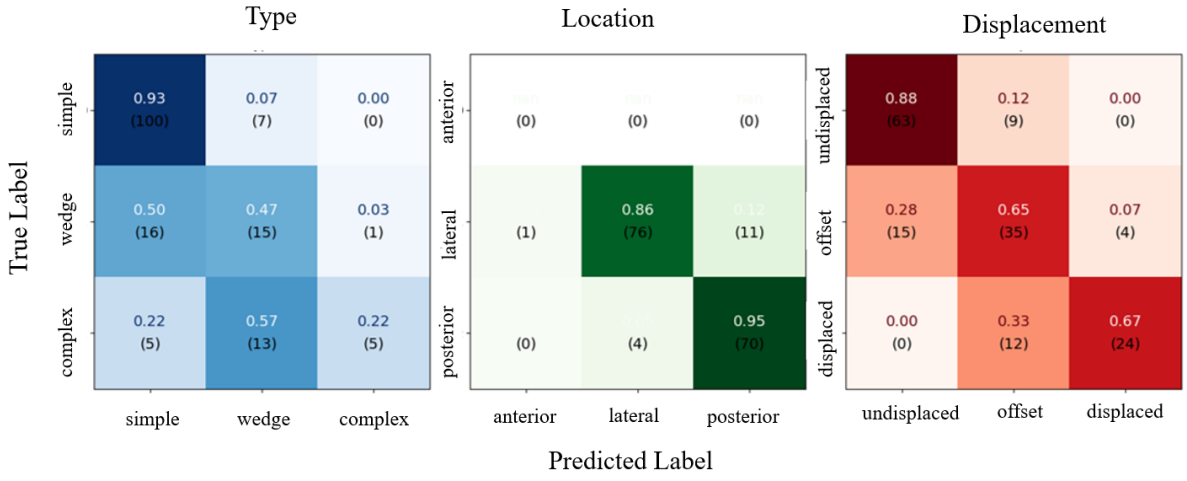


Figure 26: Confusion matrix of CWIS classification on external validation dataset

3.6 Overview Classification Performance

Table 9 gives an overview of all the datasets that have been used for testing and training each version of DCRibFrac. Table 10 presents the results obtained from the tests conducted on these datasets.

Table 9: DCRibFrac versions and their datasets used for training and testing

DCRibFrac version	Training	Testing
v1.0	EMC (n=81)	EMC (n=19)
v2.0 (EMC)	Multi-Centre (n=137)	EMC (n=19)
v2.0	Multi-Centre (n=136)	Multi-Centre (n=34)
EV v2.0	-	Zuyderland (n=28)

EV= External validation

Table 10: Performance of each DCRibFrac version on several datasets

	Test data	Detection	Type			Displacement			Location		
			Simple	Wedge	Complex	Undisplaced	Offset	Displaced	Anterior	Lateral	Posterior
Sensitivity	v1.0	0.77	0.90	0.30	0.17	0.91	0.78	0.43	0.88	0.88	0.96
	v2.0 (EMC)	0.79	0.94	0.25	0.16	0.93	0.66	0.62	0.90	0.97	0.94
	v2.0	0.80	0.95	0.37	0.25	0.90	0.80	0.61	0.79	0.94	0.97
	EV v2.0	0.84	0.93	0.47	0.22	0.88	0.65	0.67	—	0.86	0.95
Precision	v1.0	0.79	0.75	0.42	0.30	0.83	0.79	0.75	0.88	0.95	0.84
	v2.0 (EMC)	0.88	0.72	0.45	0.50	0.77	0.84	0.61	0.95	0.95	0.96
	v2.0	0.87	0.87	0.47	0.40	0.89	0.69	0.95	0.93	0.93	0.93
	EV v2.0	0.85	0.83	0.43	0.83	0.81	0.63	0.86	—	0.95	0.86
F1-score	v1.0	0.78	0.82	0.35	0.21	0.86	0.78	0.55	0.88	0.92	0.90
	v2.0 (EMC)	0.83	0.82	0.32	0.24	0.84	0.74	0.61	0.92	0.96	0.95
	v2.0	0.83	0.91	0.41	0.31	0.89	0.74	0.74	0.85	0.93	0.95
	EV v2.0	0.84	0.87	0.45	0.35	0.84	0.64	0.75	—	0.90	0.90
FPPS	v1.0	2.26									
	v2.0 (EMC)	1.16									
	v2.0	1.11									
	EV v2.0	0.96									

EV= External validation, EMC= Erasmus Medical Centre

4. Discussion

4.1 Performance of DCRibFrac v2.0

This project aimed to enhance the performance of DCRibFrac v1.0 by improving its automatic detection, CWIS classification, and rib number labelling of rib fractures. To accomplish this, a new rib number labelling method was developed, evaluated, and implemented. Additionally, three nnDetection DL models, responsible for automatic fracture detection and CWIS classification, were retrained with an expanded multi-centre dataset. These advancements resulted in DCRibFrac v2.0, an improved version of DCRibFrac v1.0.

The detection sensitivity of DCRibFrac v2.0 achieved 80% on its internal test set, with a precision of 87%, an F1-score of 83%, and a mean FPPS of 1.11. This detection sensitivity is at the upper limit of the range observed for clinicians, which spans from 73.2% to 80.8% (22–24,46). Compared to DCRibFrac v1.0, v2.0 demonstrates improved performance across all detection metrics on the same internal test set: sensitivity increased from 77% to 79%, precision from 79% to 88%, F1-score from 78% to 83%, and FPPS decreased from 2.26 to 1.11. Missed fractures are most often simple, undisplaced fractures. Besides the fact that these fracture classes occur most often, simple undisplaced fractures can be subtle and may not show major changes in bone structure, and therefore do not create distinctive features that are easily detectable.

Classification performance remains challenging, particularly for accurately identifying the type of fractures. Label imbalance may cause the model to become biased toward the more prevalent classes (simple fractures), leading to suboptimal classification for less frequent fracture types (complex and wedge). A significant proportion of the underrepresented complex and wedge fractures are misclassified as simple, leading to low sensitivity for wedge and complex fractures and reduced precision for simple fractures. Overall, the location classification has improved, especially for lateral fractures, and performs well across all classes. DCRibFrac v2.0 demonstrates an increase in classification sensitivity for displaced fractures, improving from 43% to 62% on the internal test set of v1.0, and achieving 61% sensitivity on the internal test set of v2.0. However, the sensitivity for offset fractures decreased from 78% to 66% on the v1.0 internal test set. The addition of more multi-centre data to the training set could have resulted in the model extracting other features and patterns, which results in improved classification of displaced fractures and decreased classification of offset fractures on this test set. Nevertheless, the performance on the internal test set of DCRibFrac v2.0 shows an improved sensitivity of 81% for offset fractures.

The automatic rib number labelling performance has significantly improved, with 94% and 95% of rib fractures being assigned the correct rib number in the internal and external test sets, respectively. The results indicate that the presence of fractures does not necessarily lead to a higher rate of mislabelling, suggesting robust performance even in cases with fractures.

The external validation results confirm the strong potential of DCRibFrac v2.0, with performance metrics slightly better than the internal test set. The external test set demonstrated a detection sensitivity of 84%, a precision of 85%, an F1-score of 84%, and a FPPS rate of 0.96. The label distribution of the external validation set is similar to the label distribution of the internal test set of DCRibFrac v1.0, see Table 7. However, the fraction of complex and posterior fractures is larger in the external validation dataset and lower for anterior fractures, compared to the internal test set. The results of DCRibFrac v2.0 on the internal test set show that only 11% of complex and 12% of posterior fractures are missed, in contrast to 22%, 17%, 30%, and 22% for simple, wedge, anterior, and lateral fractures, respectively. Therefore, the increase in detection performance on the external validation set can be attributed to the fact that complex and posterior fractures are detected more often, which results in higher detection sensitivity for the external validation dataset, since it contains relatively more complex and posterior fractures, compared to the internal test set.

Another factor potentially contributing to the improved classification performance on the external validation dataset could be the difference in CT scan slice thickness. In the external validation

set, 89% of the CT scans had a slice thickness of 0.5 mm, compared to only 30% in the internal test set, where the majority of scans had a larger slice thickness. Since no resampling is performed during preprocessing, and the images retain their original slice thickness and pixel spacing, the smaller slice thickness in the external dataset likely provides more detailed imaging, leading to better detection and classification of fractures.

Despite the relatively small dataset, the detection sensitivity and FPPS rate are comparable with other DL-based methods for automatic rib fracture detection (Table 11). Including only 70 additional patients has enhanced sensitivity and lowered the FPPS. The DL methods achieving the highest sensitivities (>90%) are generally trained on larger datasets, indicating that further dataset expansion will improve detection performance.

Table 11: Comparison with other DL methods that noted both the sensitivity and false positives per scan (FPPS) for detecting rib fractures. '' Denotes a missing value*

	Year	**Patients, fractures**	**Sensitivity**	**FPPS**
Zhou et al. (26)	2022	640, 2853	95%	0.17
Niiya et al. (25)	2022	918, *	93%	1.9
Li et al. (47)	2023	18172,*	93%	0.5
Meng et al. (48)	2021	8829, 34699	92%	0.14
Wang et al. (27)	2022	9265, 43803	85%	0.35
Wu et al. (49)	2021	10943, 9590	85%	0.764
Azuma et al. (32)	2022	539, 4906	84%	2.71
Zhou et al. (50)	2020	1049, 25054	83%	1.1
DCRibFrac v2.0	**2024**	**170, 1509**	**80%**	**1.11**
Zhang et al. (51)	2021	3580, 15947	80%	0.43
Weikert et al. (52)	2020	159, 991	66%	0.16
Kaiume et al. (53)	2021	39, 256	65%	1.1

4.2 Strengths and Limitations

The study design presents several strengths and limitations. Firstly, a strength of DCRibFrac v2.0 is the establishment of the ground truth for the additional FixCon data by multiple observers, in contrast to the single-observer approach employed in DCRibFrac v1.0. This approach reduces bias and enhances the accuracy of the ground truth, as classification interpretations can vary between observers, as demonstrated by the interobserver agreement study. The interobserver agreement study underscores the complexity of establishing a uniform ground truth among observers and highlights the necessity for a standardised classification method. By involving multiple observers, the performance metrics more accurately reflect the true sensitivity of the model. However, a limitation is the absence of multiple observers for the previously annotated EMC patients, which is a large fraction of the dataset used for the development and evaluation of DCRibFrac v2.0. This inconsistency in labelling may affect the overall reliability and robustness of the model.

Another strength is the inclusion of multi-centre data, which enhances the heterogeneity of the dataset. This increased diversity makes the findings more representative of a broader population and thereby improves the generalizability of the results. The variability introduced by different patient demographics, clinical practices, and equipment across centres contributes to the robustness of the predictive model. This variability helps the model become more resilient to overfitting, as it is trained on a wider range of scenarios. The improved performance, as reflected in the results, underscores the benefits of incorporating multi-centre data in developing a more reliable and generalizable model.

A limitation of the study is the lack of direct comparison between the classifications made by DCRibFrac v2.0 and those of a clinical expert. This comparison is essential for validating the model's clinical relevance and accuracy.

4.2 Future research

The current results indicate that expanding the dataset has led to increased sensitivity and precision in detection and classification. This suggests that further expansion of the dataset size could continue to enhance performance. To improve CWIS classification, future efforts should focus on augmenting the dataset with underrepresented fracture classes. This would help refine the model's ability to accurately classify these types of fractures as well.

As mentioned in the previous section, the ground truth for the EMC dataset was established by a single observer. Integrating labels from an additional observer could potentially enhance the accuracy of the ground truth and improve the model's accuracy.

Future research should focus on modifying the method for combining the rib segmentation and CWIS classification results to prevent the fractures from being assigned an incorrect rib label due to the overlap of the bounding box with more cranially laying ribs. For example, this issue can be solved by assigning the bounding box to the rib it overlaps the most. Based on the results of the internal test set, solving this issue will expectedly result in ~50% reduction of incorrectly labelled fractures, which translates to a correct rib number label for 97% of the fractured ribs. The developed nnUNet could be further improved by training a nnUNet with a dataset that includes more challenging cases, such as CT scans with artefacts and those featuring 11 rib pairs.

The time required to process a single CT scan through the entire pipeline lies between 20 and 90 minutes per patient, depending on the number of slices. This processing time limits the model's applicability in acute care settings. However, in non-acute settings, the model could improve detection sensitivity and support clearer communication. Trauma-related CT scans often cover a larger area than just the thorax, which is the region of interest for DCRibFrac v1.0. To reduce processing time, an additional preprocessing step, such as cropping the image to the region of interest, could be implemented.

Conducting an international external validation will provide valuable insights into the model’s generalizability and robustness across different populations and medical imaging protocols. This evaluation helps ensure that the model remains accurate, reliable, and applicable in various clinical settings, not only within The Netherlands. Testing the model on a diverse international dataset confirms its effectiveness in detecting and classifying rib fractures under different conditions, thereby supporting its broader applicability and potential for global use.

Future research could investigate the impact of automated rib fracture detection and classification systems on the clinical decision-making processes of radiologists and other medical specialists, with a focus on changes in decision-making patterns. Additionally, it would be valuable to evaluate how the integration of these systems influences clinical outcomes.

Finally, the DCRibFrac model could be enhanced by integrating additional clinical data to provide more comprehensive information for the decision-making process between surgical and conservative treatment options. Including risk factors that predict mortality in patients with blunt chest wall trauma, such as age over 65 years and pre-existing conditions like cardiopulmonary disease, could result in a more clinically relevant prediction model (54,55).

Conclusion

In conclusion, this project introduces DCRibFrac v2.0, an enhanced DL-based algorithm for automatic rib fracture detection and classification. Through the development and implementation of a novel rib number labelling method and the retraining of DL models using a more comprehensive, multi-centre dataset, improvements have been achieved in automatic rib fracture detection, CWIS classification, and rib number labelling. Specifically, the detection sensitivity and precision were 80% and 87%, respectively, with an FPPS rate of 1.11 and an accuracy of 94% in rib number labelling. Key enhancements include improvements in rib number labelling and a reduction in false positives, advancing towards a more reliable and standardised approach for rib fracture detection and CWIS classification. The findings of this thesis contribute to the development of a predictive model that supports the automation of the decision-making process for patients with blunt chest trauma, potentially improving clinical outcomes.

Bibliography

1. *Peek J. et al., Traumatic rib fractures: A marker of severe injury. A nationwide study using the National Trauma Data Bank. Trauma Surg Acute Care Open. 2020;5(1).*

2. *He Z. et al., The ideal methods for the management of rib fractures. J Thorac Dis, 2019. 11(Suppl 8): p. S1078-s1089.*

3. *Ziegler, D.W. et al., The morbidity and mortality of rib fractures. Journal of Trauma and Acute Care Surgery, 1994. 37(6).*

4. *Van Vledder M.G. et al., Patterns of injury and outcomes in the elderly patient with rib fractures: a multi-center observational study. Eur J Trauma Emerg Surg. 2019 Aug 1;45(4):575–83.*

5. *Bergeron E. et al., Elderly trauma patients with rib fractures are at greater risk of death and pneumonia. J Trauma, 2003 Mar;54(3): p. 478-85.*

6. *Barnea Y. et al., Isolated rib fractures in elderly patients: mortality and morbidity. Can J Surg. 2002 Feb;45(1):43-6.*

7. *Flagel B.T. et al., Half-a-dozen ribs: The breakpoint for mortality. Surgery. 2005 Oct;138(4):717–25.*

8. *Peek J. et al., Comparison of analgesic interventions for traumatic rib fractures: a systematic review and meta-analysis. Eur J Trauma Emerg Surg. 2019 Aug 1;45(4):597–622.*

9. *May L. et al., Rib fracture management. BJA Educ. 2016 Jan 1;16(1):26–32.*

10. *Mukherjee K. et al., Non-surgical management and analgesia strategies for older adults with multiple rib fractures: A systematic review, meta-analysis, and practice management guideline from the Eastern Association for the Surgery of Trauma. J Trauma Acute Care Surg. 2023 Mar 1;94(3):398-407.*

11. *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. 2019 Jul 30; 14: p. 38-38.*

12. *de Moya M. et al., Rib fixation: Who, What, When? Trauma Surg Acute Care Open. 2017 Apr 27;2(1)*

13. *Otaka S. et al., Effectiveness of surgical fixation for rib fractures in relation to its timing: a retrospective Japanese nationwide study. Eur J Trauma Emerg Surg. 2022 Apr 1;48(2):1501–8.*

14. *Bhatnagar A. et al., Rib fracture fixation for flail chest: What is the benefit? J Am Coll Surg. 2012 Aug;215(2):201–5.*

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 Dec;258(6):914–21.*

16. *Marasco S.F. et al., Prospective randomized controlled trial of operative rib fixation in traumatic flail chest. J Am Coll Surg. 2013 May;216(5):924–32.*

17. *Pieracci F.M. et al., Indications for surgical stabilization of rib fractures in patients without flail chest: surveyed opinions of members of the Chest Wall Injury Society. Int Orthop. 2018 Feb 1;42(2):401–8.*

18. *Kasotakis G. et al., Operative fixation of rib fractures after blunt trauma: A practice management guideline from the Eastern Association for the Surgery of Trauma J Trauma Acute Care Surg. 2017 Mar;82(3):618-626.*

19. *Qiu M. et al., Potential Benefits of Rib Fracture Fixation in Patients with Flail Chest and Multiple Non-flail Rib Fractures. Indian Journal of Surgery. 2016 Dec 1;78(6):458–63.*

20. *de Jong M.B. et al., Surgical treatment of rib fracture nonunion: A single center experience. Injury. 2018 Mar 1;49(3):599–603.*

21. *Edwards J.G. et al., Taxonomy of multiple rib fractures: Results of the chest wall injury society international consensus survey. Journal of Trauma and Acute Care Surgery. 2020 Feb 1;88(2):E40–5.*

22. *Van Wijck S.F.M. et al., Interobserver agreement for the Chest Wall Injury Society taxonomy of rib fractures using computed tomography images. Journal of Trauma and Acute Care Surgery. 2022 Dec 1;93(6):736–42.*

23. *Cho S.H. et al., 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 Oct;85(1018).*

24. *Yao L. et al., Rib fracture detection system based on deep learning. Sci Rep. 2021 Dec 1;11(1).*

25. *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 Dec 1;12(1).*

26. *Zhou Q.Q. et al., Precise anatomical localization and classification of rib fractures on CT using a convolutional neural network. Clin Imaging. 2022 Jan 1;81:24–32.*

27. *Wang S. et al., Assessment of automatic rib fracture detection on chest CT using a deep learning algorithm. Eur Radiol. 2023 Mar;33(3):1824-1834.*

28. *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 Sep 1;154.*

29. *Chai Z. et al. ORF-Net: Deep Omni-supervised Rib Fracture Detection from Chest CT Scans. IEEE Trans Med Imaging. 2024 May;43(5):1972-1982.*

30. *Inoue T. et al., Automated fracture screening using an object detection algorithm on whole-body trauma computed tomography. Sci Rep. 2022 Oct 3;12(1):16549).*

31. *Su, Y. et al., Rib fracture detection in chest CT image based on a centernet network with heatmap pyramid structure. SIViP 2023 ,17, 2343–2350.*

32. *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 Apr 1;29(2):317–28*

33. *Borren N. Deep Learning-Based Automatic Detection and Classification of Rib Fractures from CT scans MSc Technical Medicine thesis. TU Delft Repository. 2023 Sept. Available from: http://repository.tudelft.nl*

34. *Sunil Bhutada et al., Opening and closing in morphological image processing. World Journal of Advanced Research and Reviews. 2022 Jun 30;14(3):687–95.*

35. *Baumgartner M. et al., nnDetection: A Self-configuring Method for Medical Object Detection. MICCAI. 2021 Jun 1, vol 12905*

36 . *Isensee F. et al., nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021 Feb 1;18(2):203–11.*

37. *Munappy A.R. et al., Data management for production quality deep learning models: Challenges and solutions. Journal of Systems and Software. 2022 Sep 1;191.*

38. *Bailly A. et al., Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. Comput Methods Programs Biomed. 2022 Jan 1;213.*

39. *Cohen, J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 1960, 20(1), 37–46.*

40. *Hayes, A.F. Answering the Call for a Standard Reliability Measure for Coding Data. Communication Methods and Measures, 2007. **1**(1): p. 77-89.*

41. *Wasserthal J. et al., TotalSegmentator: Robust Segmentation of 104 Anatomical Structures in CT images. Radiol Artif Intell. 2023 Jul 5;5(5):e230024.*

42. *Fedorov A. et al., 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging. 2012 Nov;30(9):1323–41.*

43. *Cabitza F. et al., The importance of being external. methodological insights for the external validation of machine learning models in medicine. Comput Methods Programs Biomed. 2021 Sep 1;208.*

44. *Yu A.C. et al., External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. Radiol Artif Intell. 2022 May 1;4(3).*

45. *Gonzales-Portillo G.S. et al., The prevalence of 11 ribs and its potential implications in spine surgery. Clin Neurol Neurosurg. 2021 Apr 1;203.*

46. *Banaste N. et al., Whole-body CT in patients with multiple traumas: Factors leading to missed injury. Radiology. 2018 Nov 1;289(2):374–83.*

47. *Li N. et al., An automatic fresh rib fracture detection and positioning system using deep learning. British Journal of Radiology. 2023 Jun 1;96(1146).*

48. *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.*

49. *Wu M. et al., Development and evaluation of a deep learning algorithm for rib segmentation and fracture detection from multicenter chest CT images. Radiol Artif Intell. 2021 Sep 1;3(5).*

50. *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 Jul 1;21(7):869–79.*

51. *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 Feb 1;94(1118)*

52. *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 Jul 1;21(7):891–9.*

53. *Kaiume M et al., Rib fracture detection in computed tomography images using deep convolutional neural networks. Medicine (United States). 2021 May 21;100(20):E26024.*

54. *Battle C. et al, Risk factors that predict mortality in patients with blunt chest wall trauma: An updated systematic review and meta-analysis. Emerg Med J. 2023 May;40(5):369-378.*

55. *Gupta A.K. et al. Evaluation of risk factors for prognosticating blunt trauma chest. Polish Journal of Surgery. 2021 Jul 19;94(1):12–9.*

56. *Jaeger P.F. et al., Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection. 2018 Nov 21.*

Supplementary Materials

Appendix A nnDetection

nnDetection is a framework for semantic segmentations, which can also be used as an object detector and follows the same self-configuring strategy as nnU-Net (see Appendix B). The framework consists of a Retina U-Net. The Retina U-Net is specifically designed to combine the strengths of both object detection and semantic segmentation in a unified framework.

Retina U-Net uses the Feature Pyramid Network, which extracts features at different scales, enabling analysis of objects with varying sizes. The layers in Figure 1 represent different levels of the feature pyramid. The pyramid structure is used to capture multi-scale features, which are essential for detecting objects of various sizes. Each layer of the pyramid corresponds to a different resolution or scale of the image features.

The red layers are the coarse feature maps used for object detection on different scales. These layers aggregate information from the corresponding levels of the feature pyramid, allowing the network to detect objects of different sizes. The green layer represents the semantic segmentation features, which are used for classifying pixels into different categories and distinguishing between different objects. The network produces two outputs: the classification of the detected objects and the coordinates of their bounding boxes.

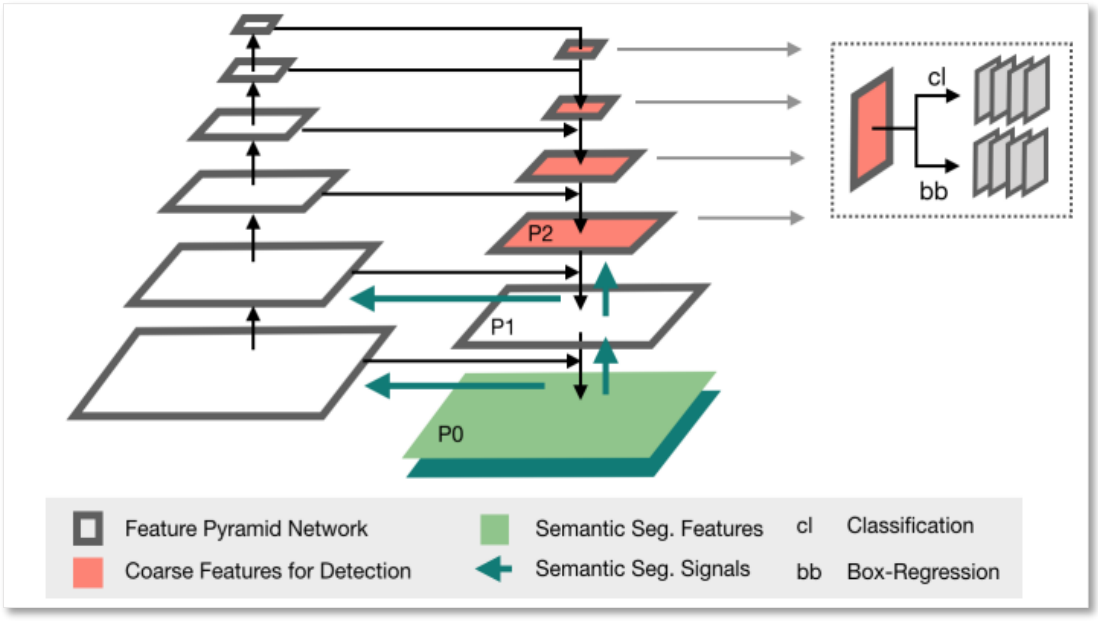


Figure 1: A schematic 2D representation of Retina U-Net (56).

Appendix B nnUNet

The nnU-Net framework is an advanced, fully automated method for medical image segmentation, capable of handling a wide variety of segmentation tasks. It functions end-to-end, covering all steps from preprocessing to training and post-processing, while adapting to different challenges with minimal manual intervention. nnU-Net is based on the U-Net architecture, shown in Figure 1.

The process begins with the model analysing an input image. As the image passes through multiple layers, the model learns important features while the image size gradually decreases (downsampling). At the network's deepest point, the image becomes small, but the model has captured a lot of detailed features. The model then increases the size of the image back to the original size, while combining the learned features (upsampling). The final output is a segmented version of the image, where each part is labelled according to what it represents, for example, a ribcage.

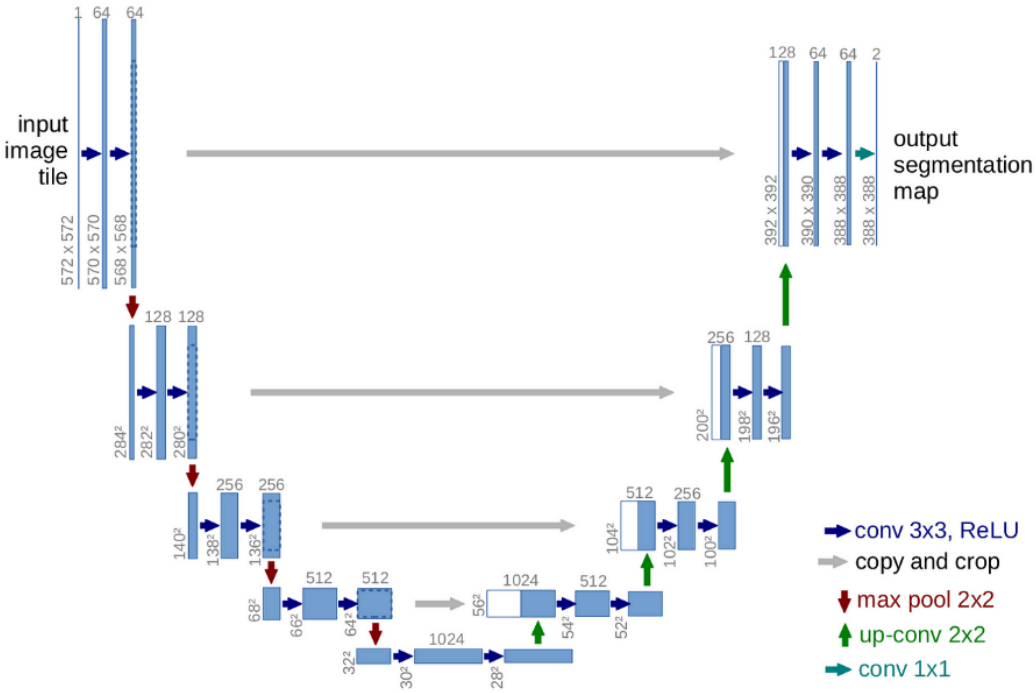


Figure 1: U-Net Architecture for Image Segmentation

Appendix C: Labelling Software

This appendix was written by N. Borren for the development of DCRibFrac v1.0 (33)

A comprehensive overview of the labelling software is given. First, a description of the module network is given. Then, a manual belonging to the GUI will be described.

C.1 Module Network

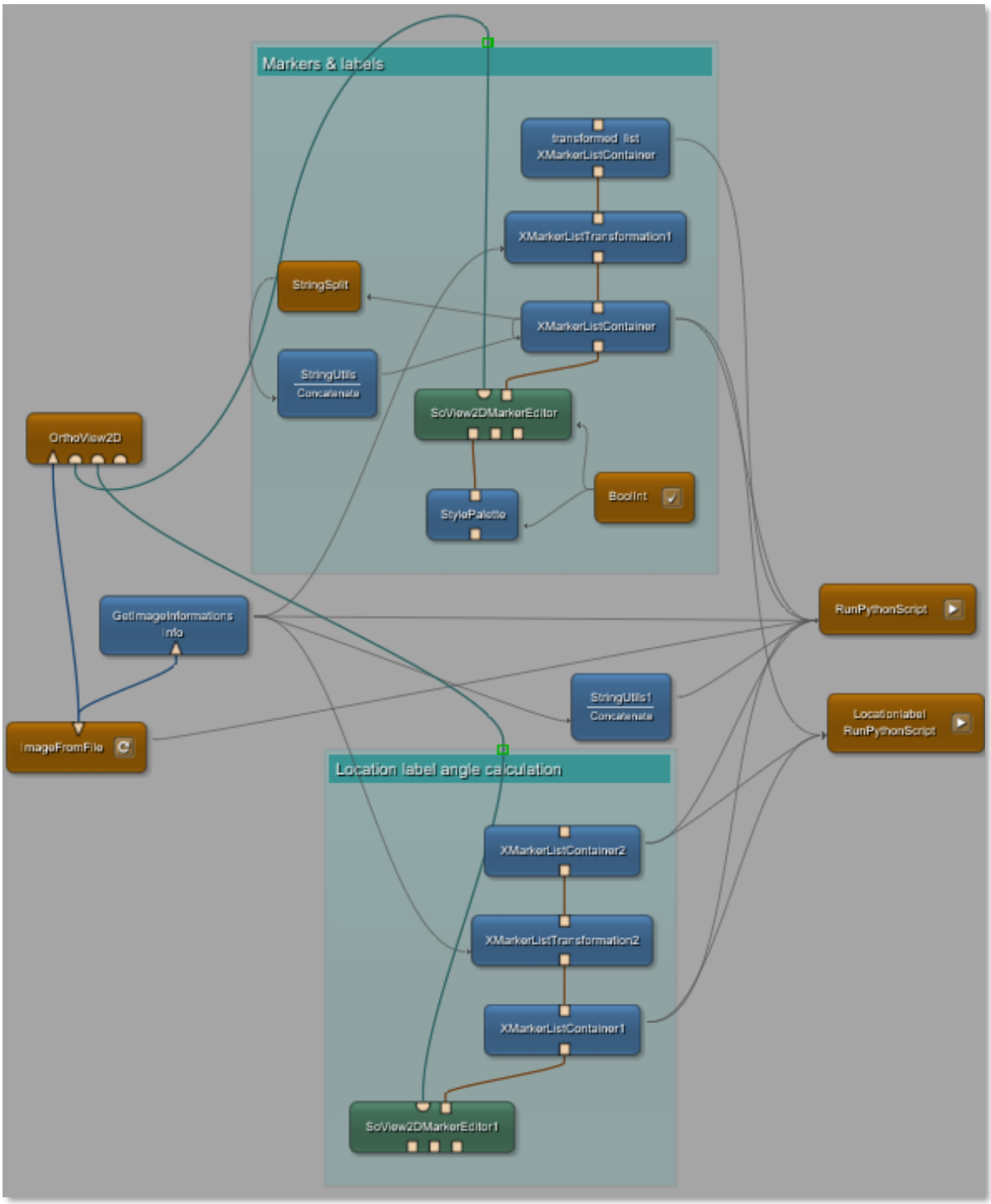


Figure 2: Module network of MeVisLab for the labelling of rib fractures.

There are two main parts in the module network; the marker and label assignments, and the calculation of the angle from which the location label can be determined. Each module will briefly be described. 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. Here, error-messages are also defined for when not all labels are assigned to a marker.

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 result of this script is an angle in degrees which can help in deciding which location label should be assigned to the marker.
The module network comes together in the GUI that is explained in the next chapter.

C.2 GUI Manual

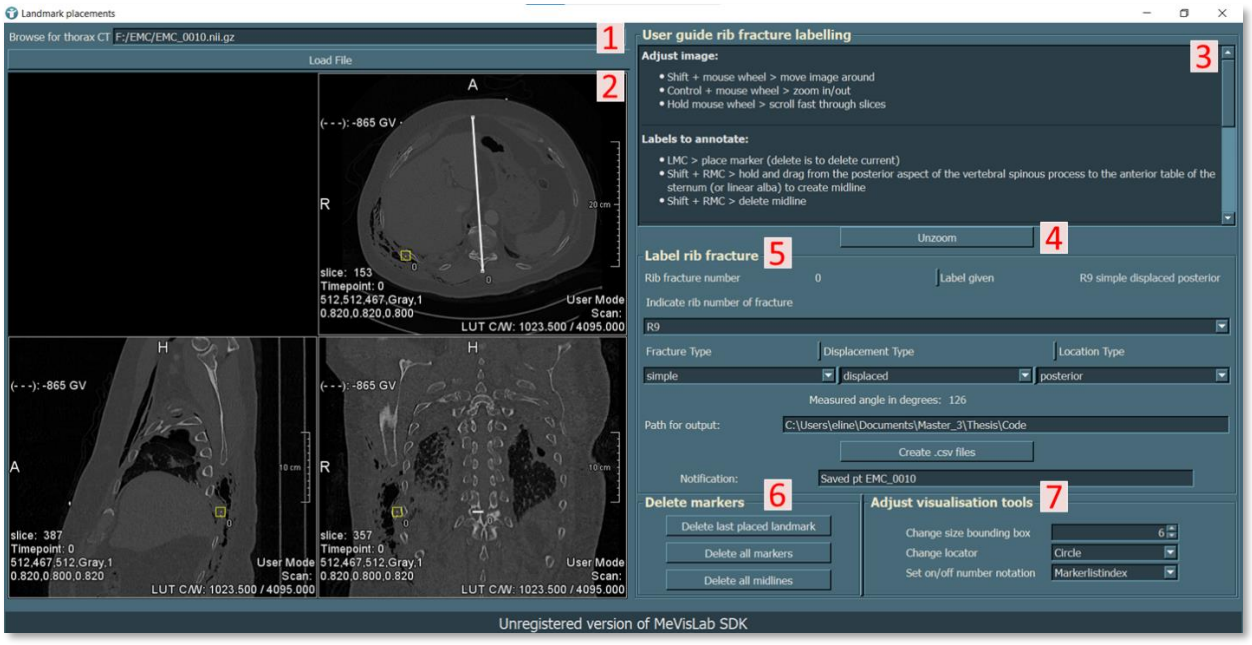


Figure 3: The GUI of the MeVisLab labelling software with numbers indicating the different sections.

The GUI will be explained according to the different sections in the interface, corresponding to the numbers in Figure 3.

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 anterior table of the sternum to create the midline. The angle will be given in section 5.

A marker can be deleted by clicking on it and using *delete* from 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 definition of labels 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. Label the rib fracture. In the first line, the rib fracture number that is currently selected and the given labels are shown. In the drop-down menus the four labels can be assigned. 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]*. When 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 D: Label Distribution Internal Dataset

Table 1: Dataset Characteristics Internal Test and Training set DCRibFrac v1.0

Variables	Internal training set	Internal test set
No. patients	81	19
No. fractures (%)	803 (80)	207 (20)
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)

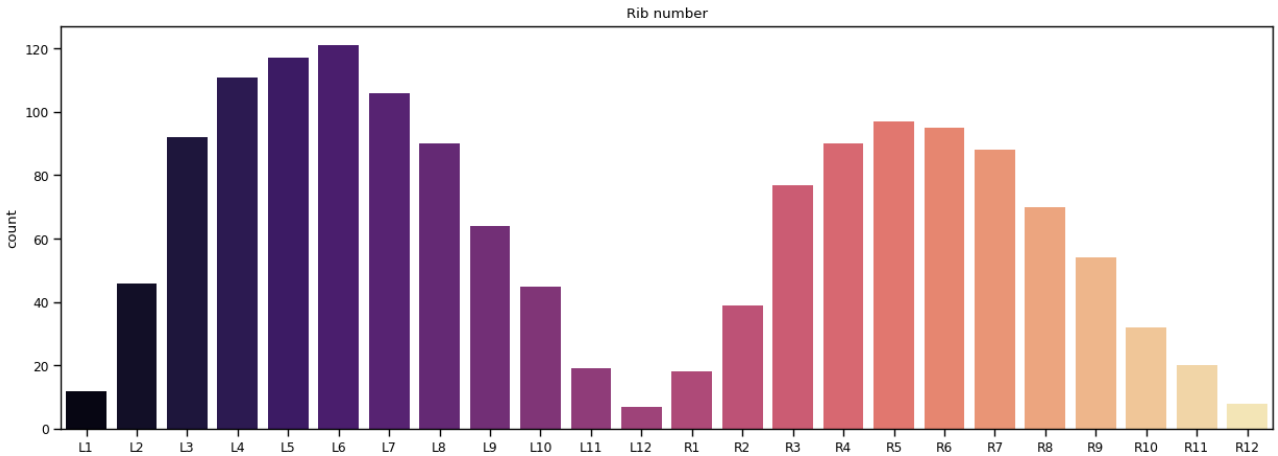


Figure 1: Number of fractures per rib in the internal test and training set of DCRibFrac v2.0

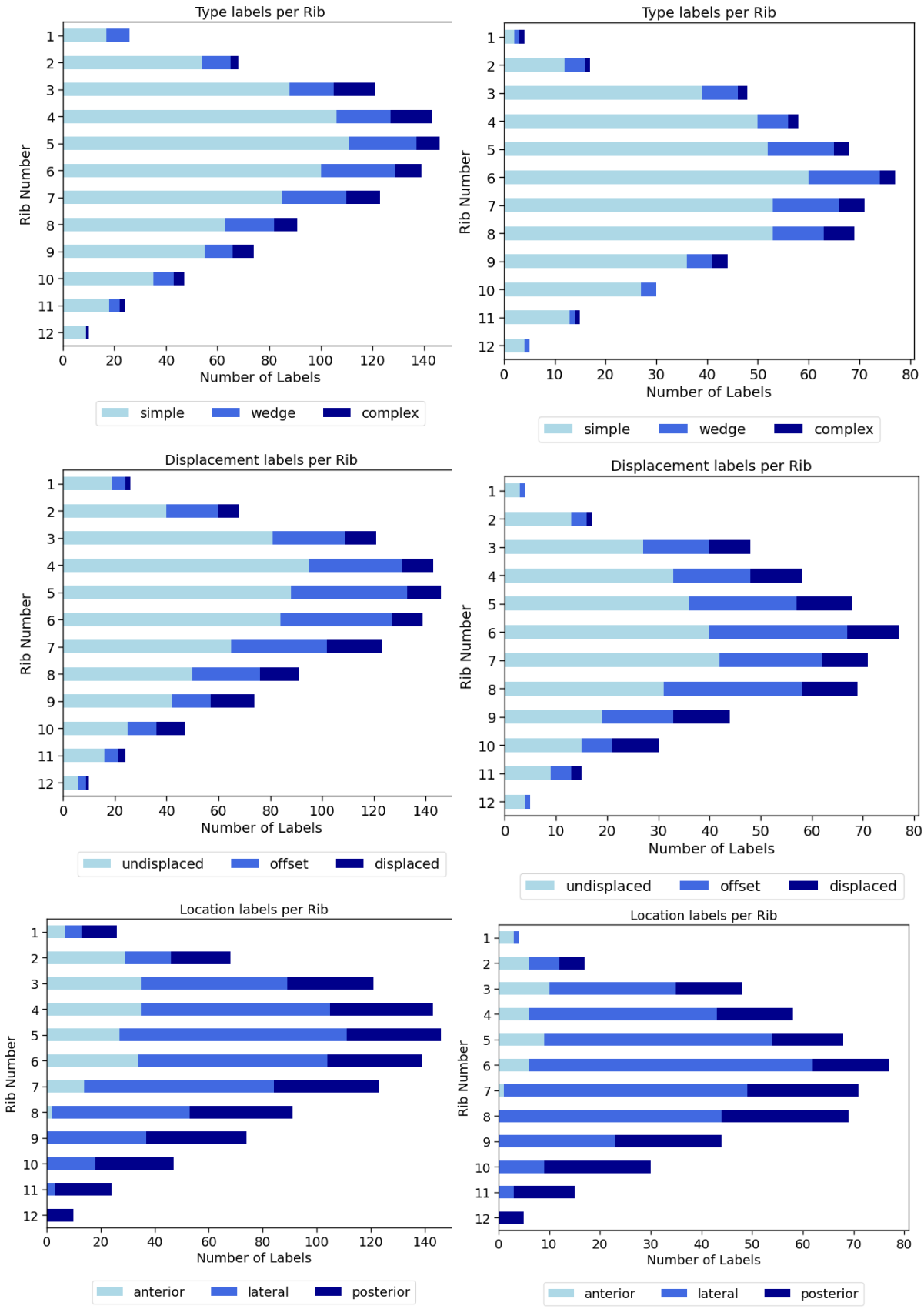


Figure 2: Label distribution of the EMC dataset (left) and the additional FixCon data (right). Distribution per rib number for the labels type (top), displacement (middle), and location (bottom).

Appendix E: Label Distribution External Validation Dataset

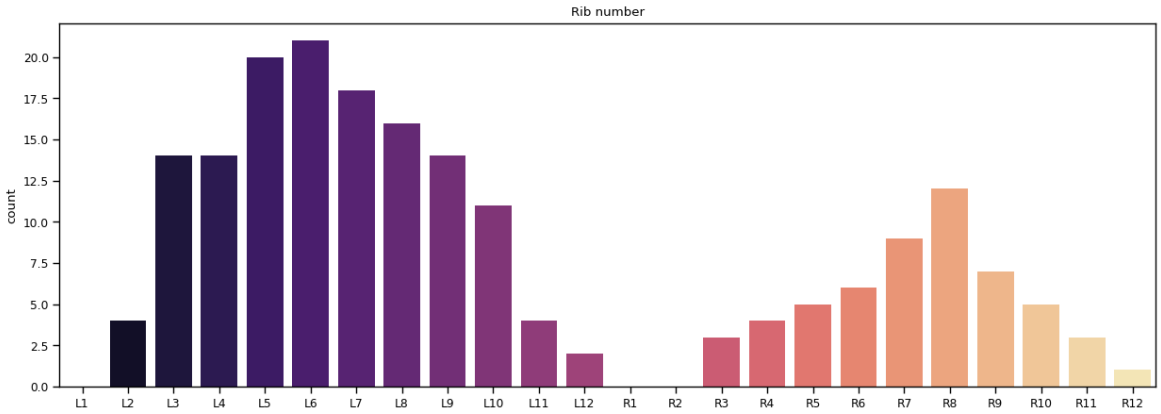


Figure 1: Number of fractures per rib in the external test set

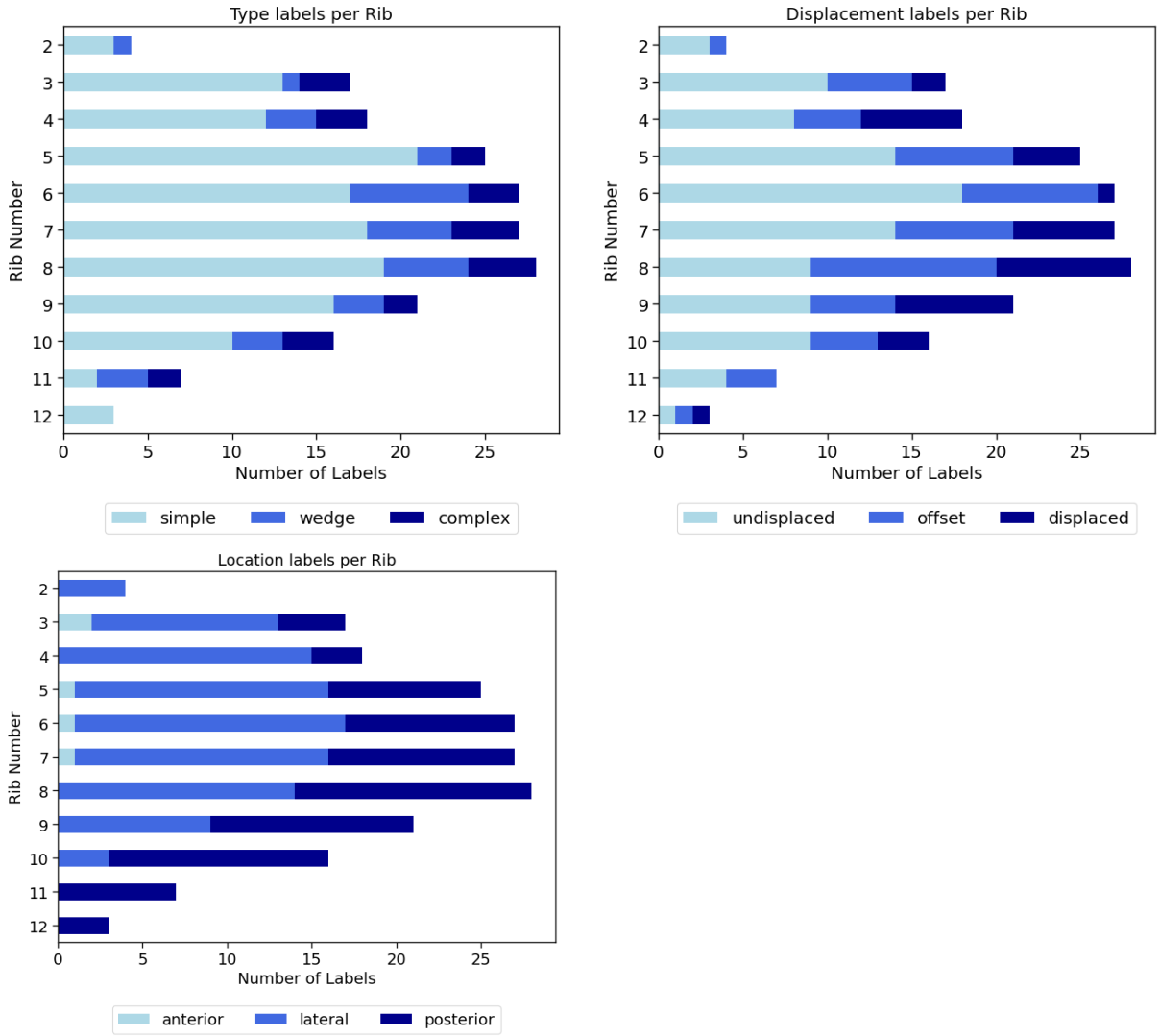


Figure 2: Fracture characteristics of external validation dataset. Distribution per rib number for the labels type (top left), displacement (top right), and location (bottom).