

From the Department of Clinical Sciences, Danderyd
Hospital
Karolinska Institutet, Stockholm, Sweden

FRACTURE EVALUATION AND PREDICTION OF OUTCOME AFTER A FRACTURE USING ARTIFICIAL NEURAL NETWORKS

Jakub Olczak



**Karolinska
Institutet**

Stockholm 2024

All previously published papers were reproduced with permission from the publisher.

Published by Karolinska Institutet.

Printed by Universitetsservice US-AB, 2024

© Jakub Olczak <https://orcid.org/0000-0002-7706-6951>

The comprehensive summary chapter of this thesis is licensed under CC BY 4.0. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/> Other licences or copyright may apply to illustrations and attached articles.

ISBN 978-91-8017-824-2

DOI <https://doi.org/10.69622/27194766>

Cover illustration: Watercolor artwork by Märta Nummelin.

Fracture evaluation and prediction of outcome after a fracture using artificial neural networks

Thesis for Doctoral Degree (Ph.D)

By

Jakub Olczak, MD

The thesis will be defended in public at Danderyds sjukhus, Föreläsningssal Aulan, Entrévägen 2, målpunkt K, plan 3, 2024-12-04, 09:00

Principal Supervisor:

Docent. Max Gordon, MD

Karolinska Institutet

Department of Clinical Sciences, Danderyd Hospital

Division of Orthopedics

Opponent:

Docent Hans Berg, MD

Karolinska Institutet

Department of Physiology and Pharmacology

Division of Environmental physiology

Co-supervisor(s):

Professor Olof Sköldenberg, MD

Karolinska Institutet

Department of Clinical Sciences, Danderyd Hospital

Division of Orthopedics

Examination Board:

Professor Li Felländer-Tsai, MD

Karolinska Institutet

Department of Department of Clinical Science,

Intervention and Technology

Division of Orthopedics

Ali Sharif Razavian, PhD

Kungliga Tekniska Högskolan (previously)

Docent Maria Cöster, MD

Uppsala University

Department of Department of Surgical Sciences

Division of Orthopedics and Hand surgery

Docent Mattias Rantalainen

Karolinska Institutet

Department of Department of Medical Epidemiology
and Biostatistics

Division of Orthopedics

To my family

Popular science summary of the thesis

Young men and postmenopausal women are most at risk of a fracture. We know which treatments work when treating fractures, on average. However, every patient is unique and has unique circumstances and individual resources. For example, we know that some patient groups have a 50% risk of dying within one year, e.g., some elderly and frail patients. We just do not know which individuals will suffer.

A fracture can have a considerable impact on your quality of life. Constant pain, the inability to take care of yourself, depression, and disability are common lifelong outcomes. However, you are at risk of even worse outcomes as there are strong links between fractures and death.

Using artificial intelligence (AI) and machine learning (ML), we can now analyze massive amounts of data in ways that have never been possible. We will use this approach to create a new form of individualized predictions that will predict how your life is expected to change due to the fracture. This concept is called personalized medicine.

First, we were able to analyze fractures in X-ray images using artificial intelligence. Not all fractures are the same, and we needed to determine the type of fracture, as the type of fracture matters in the choice of treatment.

However, not all patients are the same. We need to map out the characteristics of the patient. Patients will fill out a form about their health status before the injury while waiting to see the doctor or on the X-ray results. Doctors will use that information with the AI model to guide treatment toward the best outcome by tailoring the surgery and aftercare to the problematic areas.

Using healthcare data from thousands of patients, we teach artificial intelligence to predict the outcome after a fracture. We hope to focus on those areas where the patient will suffer the most and complications will follow. By focusing on preventing specific outcomes, healthcare professionals will minimize the negative effects of having a fracture on the patient and society.

Populärvetenskaplig sammanfattning

Yngre män och postmenopausala kvinnor löper störst risk för att råka ut för en fraktur. Vi vet vilka behandlingar som fungerar vid olika frakturer, i genomsnitt. Men varje patient är unik, har unika förutsättningar och individuella resurser. Till exempel så vet vi att vissa patientgrupper har en 50% risk att dö inom ett år, till exempel vissa äldre och sköra patienter. Vi vet bara inte vilka dessa individer är.

En fraktur kan ha en betydande inverkan på din livskvalitet. Konstant smärta, oförmåga att ta hand om dig själv, depression och funktionshinder är vanliga och ofta livslånga konsekvenser. Dessutom är du, efter en fraktur, i riskzonen för ännu värre utfall då det finns starka kopplingar mellan frakturer och död.

Med hjälp av artificiell intelligens (ofta kallat för "AI") och maskininlärning ("ML") kan vi idag analysera enorma mängder data på sätt som tidigare inte varit möjliga. Vårt mål är att använda dessa metoder för att skapa en ny form av individualiserade prognoser som förutspår hur just ditt liv förväntas förändras på grund av frakturen. Detta koncept kallas för precisionsmedicin.

Först lyckades vi att hitta frakturer i röntgenbilder med hjälp av artificiell intelligens. Men alla frakturer är inte likadana, och vi fokuserade då på att fastställa typen av fraktur, eftersom typen av fraktur spelar roll vid valet av behandling. Men inte alla patienter är likadana. Vi måste kartlägga patientens förutsättningar.

Målet är ett system där patienterna först fyller i ett formulär om sitt hälsotillstånd före skadan, till exempel medan de väntar på att träffa läkaren eller på röntgenresultaten. Läkare, sjuksköterskor, fysioterapeuter och annan sjukvårdspersonal kommer sedan att använda den informationen tillsammans med en AI-modell för att försöka förstå vilka patientens förväntade problemområden kommer att vara. Genom att välja operation och fokusera eftervården till de förväntade problemområdena, är målet att uppnå bästa möjliga resultat – i termer av förtida död och livskvalitet – för patienten.

Genom att använda data från tusentals patienter lär vi den artificiella intelligensen att förutspå utfallen efter en fraktur. Vi hoppas kunna fokusera på de områden där patienten kommer att lida mest och där komplikationer kommer att uppstå. Genom att fokusera på att förebygga specifika utfall kommer vårdpersonal att minimera de negativa effekterna av en fraktur på både patienten och samhället.

Abstract

Background: Improved interpretation of orthopedic trauma could improve patient outcomes. The radiograph is the predominant tool in orthopedic emergency decision-making. Machine learning-guided radiographic interpretation could help improve patient outcomes.

Aims: 1) Explore convolutional neural networks (CNN) for orthopedic trauma imaging and fracture and classification in medical imaging. 2) Study CNNs on combined imaging and registry data to predict patient outcomes after trauma. 3) Evaluate the generalizability of this approach through external validation.

Methods: Study I used CNNs and transfer learning to detect fractures in auto-labeled wrist, hand, ankle, and foot radiographs. Study II and Study III doubled down on ankle fractures using the AO Foundation-/Orthopedic Trauma Association (AO) 2018 standard. We manually labeled thousands of ankle exams and trained a CNN to classify fractures. In Study III, we externally validated a CNN model against a different site and implemented active learning to improve the model. Study IV linked fractures in the Swedish Fracture Registry (SFR) to the trauma radiographs and developed models that, based on the initial radiograph, predicted patient-reported outcome measures (PROM) or death after one year.

Results

Study I: Deeper CNN architectures outperformed, with the best correctly classifying 83% of cases, compared to 82% for the human reviewers. For secondary outcomes, the CNN performed near-perfectly for body parts and excellently in exam view. A manual review of 400 random training cases found that the auto-generated labels were the problem.

Study II: The CNN performed well on the primary task. However, several outcomes were too rare to be included in the training, testing, or error bounding. For example, type A fractures were challenging to train, and there were many AO subgroups.

Study III: The external validation data differed from the training site in important ways. It included weight-bearing studies, mostly type A fractures, with fewer views per study. The CNN external validation performance improved with active learning on type A fractures but decreased somewhat for other types.

Study IV: We tried a range of network configurations and found that the CNN's ability to predict PROM after one year (PROM1) or death was variable. At best, the root mean squared errors (RMSE) and mean average errors (MAE) were on par with the standard deviation.

Conclusions

Study I: We succeeded in predicting fractures in radiographs at the level of human reviewers. The CNN performance for individual radiographs was better than indicated by the automatic fracture labels generated for the study.

Study II: We successfully implemented a CNN for ankle fracture classification using the AO 2018 standard, looking at the complete exam rather than individual images.

Study III: The initial external validation dataset performance was acceptable but not good enough. We successfully improved external validity using internal training data and active learning. External validation is essential when reporting CNN model performance.

Study IV: We performed a series of experiments to train a CNN to predict PROM after one year and got our models to learn the most common value or the mean for the PROMs, i.e., overfits. We explore different ways to improve performance.

List of scientific papers

- I. Olczak J, Fahlberg N, Maki A, Razavian A S, Jilert A, Stark A, Sköldenberg O, Gordon M. Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthopaedica*, 2017, 88:6, 581–586, DOI: 10.1080/17453674.2017.1344459
- II. Olczak J, Emilson F, Razavian A, Antonsson T, Stark A, Gordon M. Ankle fracture classification using deep learning: automating detailed AO Foundation/Orthopedic Trauma Association (AO/OTA) 2018 malleolar fracture identification reaches a high degree of correct classification. *Acta Orthopaedica*. 2021a Jan 2;92(1):102–8. DOI: 10.1080/17453674.2020.1837420
- III. Olczak, J., Prijs, J., Ijpma, F. *et al.* External validation of an artificial intelligence multi-label deep learning model capable of ankle fracture classification. *BMC Musculoskelet Disord* 25, 788 (2024). <https://doi.org/10.1186/s12891-024-07884-2>
- IV. Olczak J and Gordon M. Artificial intelligence for predicting patient-reported outcome measures (PROM) from the Swedish Fracture Registry, based on trauma radiographs (manuscript).

Scientific papers not included in the thesis

- V. Olczak J, Kiani NA, Zenil H, Tegner J. Topological Evaluation of Methods for Reconstruction of Genetic Regulatory Networks. In: 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) [Internet]. Bangkok, Thailand: IEEE; 2015 [cited 2024 Mar 29]. p. 468–73. Available from: <http://ieeexplore.ieee.org/document/7400604/>
- VI. Kiani NA, Zenil H, Olczak J, Tegnér J. Evaluating network inference methods in terms of their ability to preserve the topology and complexity of genetic networks. *Seminars in Cell & Developmental Biology*. 2016 Mar 1;51:44–52.
- VII. Olczak J, Pavlopoulos J, Prijs J, Ijpmma FFA, Doornberg JN, Lundström C, et al. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. *Acta Orthop*. 2021 May 14;1–13.
- VIII. Oliveira e Carmo L, van den Merkhof A, Olczak J, Gordon M, Jutte PC, Jaarsma RL, et al. An increasing number of convolutional neural networks for fracture recognition and classification in orthopaedics. *Bone Jt Open*. 2021 Oct 20;2(10):879–85.
- IX. Prijs J, Liao Z, Ashkani-Esfahani S, Olczak J, Gordon M, Jayakumar P, et al. Artificial intelligence and computer vision in orthopaedic trauma: the why, how, and what. *The Bone & Joint Journal*. 2022 Aug 1;104-B(8):911–4.
- X. Prijs J, Liao Z, To MS, Verjans J, Jutte PC, Stirler V, et al. Development and external validation of automated detection, classification, and localization of ankle fractures: inside the black box of a convolutional neural network (CNN). *Eur J Trauma Emerg Surg*. 2023 Apr;49(2):1057–69.
- XI. Dankelman LHM, Schilstra S, Ijpmma FFA, Doornberg JN, Colaris JW, Verhofstad MHJ, et al. Artificial intelligence fracture recognition on computed tomography: review of literature and recommendations. *Eur J Trauma Emerg Surg*. 2023 Apr 1;49(2):681–91.
- XII. Olczak J and Gordon M. From Radiologist Report to Image Label: Assessing Latent Dirichlet Allocation in Training Neural Networks for Orthopedic Radiograph Classification. [Internet]. arXiv; 2024. Available from: <https://doi.org/10.48550/arXiv.2408.13284>

Contents

1	Literature review.....	1
1.1	Data.....	1
1.1.1	Imaging data.....	2
1.1.2	Registries – The Swedish Fracture Registry.....	3
1.2	Fracture classification.....	5
1.2.1	Classification systems.....	6
1.2.2	Utility and problems of fracture classification from radiographs.....	7
1.2.3	Ankle fracture classification – Lauge–Hansen vs. Danis–Weber vs. AO ankle.....	8
1.3	Artificial intelligence, deep learning, and machine learning modeling.....	11
1.3.1	The case for ANNs – The Universal approximation theorem.....	12
1.3.2	Training ANNs.....	12
1.3.3	Recent advances.....	14
1.3.4	AI in medicine.....	16
1.3.5	ML and ANNs in orthopedics.....	17
1.3.6	ML and ANN for outcome prediction in Orthopedics.....	17
1.4	Ethical considerations and methodological biases.....	18
1.5	Discussion and conclusion.....	22
2	Research aims.....	23
3	Materials and methods.....	25
3.1	Study design.....	25
3.2	Data sources.....	25
3.2.1	Danderyd Hospital, Stockholm, Sweden (DS).....	25
3.2.2	Flinders Medical Centre, Adelaide, Australia (FMC).....	25
3.2.3	The Swedish Fracture Registry (SFR).....	25
3.2.4	Region Stockholm (RS).....	26
3.2.5	Region Gotland (RG).....	26
3.3	Neural networks.....	26
3.3.1	Images and imaging.....	26
3.3.2	Image transformations.....	28
3.3.3	Convolutional neural networks.....	29
3.4	Statistics.....	30

3.4.1	Balanced vs. imbalanced problems	31
3.4.2	Dataset size selection	31
3.4.3	Accuracy	32
3.4.4	Precision and recall	32
3.4.5	Area under the curve (AUC)	32
3.4.6	Area under the receiver operating characteristic curve (AUROC)	33
3.4.7	The area under the precision-recall curve (AUPR)	33
3.4.8	Bootstrapping confidence intervals (CI)	34
3.4.9	Top-N performer	34
3.4.10	Weighted average	34
3.5	Data and population	35
3.5.1	Study I	35
3.5.2	Study II	36
3.5.3	Study III	37
3.5.4	Study IV	38
3.6	Modelling	40
3.6.1	Study I	40
3.6.2	Study II	41
3.6.3	Study III	43
3.6.4	Study IV	43
4	Results	47
4.1	Study I	47
4.1.1	Primary outcome – Fracture detection	47
4.1.2	Secondary outcomes	49
4.2	Study II	49
4.3	Study III	54
4.3.1	Flinders data (EVD)	55
4.3.2	Danderyd (IVD)	56
4.4	Study IV	60
5	Discussion	71
5.1	Fracture detection using CNNs	71
5.2	Imaging-based patient outcome prediction	73
5.3	Study I	74
5.3.1	Discussion of results	74
5.3.2	Strengths	75
5.3.3	Limitations	75

5.4	Study II.....	76
5.4.1	Discussion of results.....	76
5.4.2	Strengths.....	77
5.4.3	Limitations	77
5.5	Study III.....	77
5.5.1	Discussion of results.....	77
5.5.2	Strengths.....	80
5.5.3	Limitations	80
5.6	Study IV	81
5.6.1	Discussion of results.....	81
5.6.2	Strengths.....	83
5.6.3	Limitations	84
6	Conclusions.....	85
6.1	Study I.....	85
6.2	Study II.....	85
6.3	Study III.....	85
6.4	Study IV	85
7	Points of perspective.....	87
8	Acknowledgements	89
9	Declaration about the use of generative AI	91
10	References	93

List of abbreviations

AI	Artificial intelligence
ANN	Artificial neural networks
AO or AO/OTA	AO Foundation/Orthopedic Trauma Association
AUC	Area under the curve. Usually, the AUROC.
AUROC	Area under receiver operating characteristic curve.
AUPR	Area under the precision-recall curve
CI	Confidence interval
CNN	Convolutional neural network
CONSORT	Consolidated Standards of Reporting Trials
CT	Computed tomography
DICOM	Digital Imaging and Communications in Medicine
DL	Deep learning
EQ-5D	EuroQol 5 dimensions
EQ-5D-3L/5L	EQ-5D three or five levels in the answers
EQUATOR	Enhancing the QUALity and Transparency Of health Research
EVD	External validation dataset
GPU	Graphics processing unit
IOR	Intra-observer reliability
IRR	Inter-rater reliability
IVD	Internal validation dataset
LDA	Latent Dirichlet allocation
ML	Machine learning
MRI	Magnetic resonance imaging
NLP	Natural language processing
PACS	Picture archiving and communications system
PROM	Patient reported outcome measure

PROM0	Initial PROM survey, asking how function was before the injury.
PROM1	PROM after one year, asking how the function one year later.
PROM Δ	The one-year <i>change</i> in PROM
RIS	Radiology information systems
RNN	Recurrent neural network
ROI	Region of interest
SD	Standard deviation
SFMA	Short Functional Musculoskeletal Assessment
SPIRIT	Standard Protocol Items: Recommendations for Interventional Trials
TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
SFR	Swedish Fracture Registry (Svenska frakturregistret)
STARD	Standards for Reporting Diagnostic Accuracy Studies
STROBE	The Strengthening the Reporting of Observational Studies in Epidemiology
VAS	Visual analog scale
3L	EQ-5D-3L
5L	EQ-5D-5L

Introduction

This doctoral thesis focuses on using AI and ML for the analysis and interpretation of orthopedic trauma. The assessment of orthopedic trauma is heavily reliant on medical imaging. Computed tomography (CT), magnetic resonance imaging (MRI), and, to a much lesser extent, ultrasound are important in studying orthopedic extremity trauma. However, radiographs (X-rays) are still by far the dominant mode of study and decision-making for extremity trauma.

Radiographic interpretation of medical imaging can be challenging. Radiologists take years to train, and specialists in, e.g., musculoskeletal radiology, are not always available when needed. In addition, every radiograph must be examined by two radiologists. However, in practice, radiographs must also be interpreted by non-radiologists, for example, in emergency department settings. Usually, the orthopedic surgeons examine the radiographs themselves, and sometimes, there are discrepancies between what the radiologist reports and what the orthopedic surgeons want to know. In addition, from a global perspective, there is an even greater shortage of doctors in general, and radiologists are in even shorter supply.

Modern **artificial intelligence** (AI) excels at image analysis and interpretation via **machine learning** (ML). With the demand for radiographic imaging and interpretation outstripping the availability of interpretation, AI has been suggested as a solution. This thesis explores this idea.

The thesis aims to study AI applications on medical data. In developing methods for studying radiographic imaging, it investigates AI modeling for orthopedic trauma. The same techniques could also be adapted to other kinds of medical data. Study I examines fracture detection in ankle, wrist, and hand radiographs ¹. Study II uses the AO Foundation/Orthopedic Trauma Association (AO) classification ² to study fractures. Study III examines external validation and model tweaking to make AI models usable in other environments ³. In our final study, Study IV, we attempt to expand the use of fracture radiographs to predict patient-perceived outcomes over time or death within one year of the trauma.

This is a doctoral of medicine thesis, so mathematical formalism will be very sparse.

1 Literature review

1.1 Data

The foundation of learning is information and data, which is also true for ML. Data can be difficult and expensive to gather and then validate. However, data availability has been fundamental to allowing for iterative and empirical tweaking of algorithms. Much AI development has centered around the drive to tweak performance on open-source datasets where the desired outcomes are known. These datasets often serve as benchmarks and validation tools for new methods.

The MNIST dataset contains handwritten greyscale digits, although there are also versions with small images and hand-drawn letters ⁴. The CIFAR-10 and CIFAR-100 datasets each contain 60,000 labeled images ⁵. The ImageNet dataset ⁶ is a set of color photos taken from the Internet that contained approximately 3.2 million hierarchically labeled images (today, approximately 14 million ⁷). ImageNet functions as a development data set and has been used for a long time as an ML competition dataset. Today, many algorithms outperform humans at labeling images in this dataset. ImageNet, and the ImageNet challenge, is by many considered the catalyst for the AI and ML boom we see today ⁸.

Medical data is usually sensitive, i.e., personal information requiring strong privacy protection. It needs close vetting for personal information and trained experts to review it, and it is often ambiguous (i.e., is it a lung nodule or not). Even so, the availability and quality of medical data sets have also improved. CheXpert ⁹ and ChestX-ray8 ¹⁰ are two widely used datasets for studying chest radiographs. They contain information on the presence or absence of lung nodules, fluid, infiltrates, and other lesions. In 2018, the MURA dataset of orthopedic trauma radiographs was released and contained 14,863 studies of seven study types (e.g., elbow, wrist, hand, and others.) Each study includes information on the presence or absence of fractures ¹¹. Esteva et al. have released a dataset of skin lesions used to create a melanoma detector for public use ¹². The dataset was biased towards light-skinned patients ^{13,14} and was updated – showing the power of sharing data for development and validation by the scientific community. Our test dataset, used for performance testing in Study II and Study III, annotated according to the AO 2018 classification, is set to become publicly available, as are some additional datasets from our research group.

The data for the studies in this thesis project consisted of imaging and registry data.

1.1.1 Imaging data

Medical images are stored in a specialized Picture Archiving and Communication System (PACS), usually in the Digital Imaging and Communications in Medicine (DICOM) imaging format. The DICOM standard was created to hold medical imaging and facilitate its transfer inside and between institutions. As such, it is a standard for communicating and storing DICOM images¹⁵. The DICOM images in PACS systems contain the radiographic examination (i.e., the image) and study-related metadata tags. DICOM has four hierarchical levels that help keep track of studies: (1) patient, (2) study (also known as exam or procedure), (3) series, and (4) image (or instance). Each patient can have performed multiple studies in their life. Each study consists of one or more series. A series can be different modalities, such as CT and radiographs, but it can also be an ankle and shoulder series. Each series comprises one or more images that make up a coherent picture, such as all the slices of a CT scan or the individual images of the projections in a scaphoid trauma examination^{16–18}. While PACS systems usually store DICOM images, they can also store other imaging data. However, they typically do not contain information on referrals or radiologist reports. These are usually stored in radiology information systems (RIS).

For example, in this project, we collected four different imaging datasets. We had a dataset of radiographic examinations collected from Danderyd Hospital's PACS. Subsets of this data have been used in three of the studies in this thesis project: studies I and II¹² and Study III. Other parts of that dataset have been used for other studies from the same research group^{19–22}. The dataset, which also contains radiologist reports but no referrals, will also serve as the foundation for future studies and models. For Study III, we collaborated with a research group from the Netherlands (Groningen University) that was also connected to Australia (Flinders University, Adelaide). Through them, we gained access to radiographic examinations from a trauma center in Adelaide, which were used in Study III.

In the fall of 2020, we collected radiographic examinations for approximately 3,100 fractures from Region Gotland's PACS. In the spring of 2023, we collected imaging on 41,000 fractures from Region Stockholm's PACS. Both datasets were based on a second important medical data source—registry-based data.

1.1.2 Registries – The Swedish Fracture Registry

The **SFR** is a Swedish national registry that tracks fractures in Sweden. It was created in 2011 to track fractures, their treatment, and patient outcomes. The registry was originally unique in that it tracks both those fractures that have undergone operative and non-operative treatment. It connects to and syncs data with national registries, such as the Swedish National Board of Health and Welfare registers and the Swedish Arthroplasty Register ²³.

As of December 2020, there were approximately 525,000 registered fractures; as of July 2024, there were 961,000 fractures. Each fracture is registered by the participating clinics, i.e., all emergency hospitals in Sweden dealing with orthopedic injuries. It is usually the treating physician who registers. Patient data, time of injury, type of injury, and mechanism of injury are registered. The type of injury is registered using ICD 10 and AO classification, including whether the fracture is open or closed, close to a prosthesis, and more. Treatment (operative and conservative) is tracked, and to a lesser extent, complications (e.g., infection and healing complications) are registered ²³.

1.1.2.1 *Using the SFR*

The data registration in the SFR, i.e., the type of fracture, has been validated by Juto et al., who studied 152 ankle fractures registered in the SFR. Three orthopedic surgeons examined the fractures and created a consensus standard of fracture classification according to the AO classification. They found excellent to near-perfect agreement between their observations and the registered fracture type in the SFR and almost as good results for the fracture group ^{24,25}. Wennergren et al. examined the validity of fracture classification in the SFR by examining 116 humerus fractures. Like Juto et al., a three-surgeon team assessed all radiographic examinations twice to create a ground truth according to the AO standard. They also found excellent inter-rater reliability between their standard and the data in the SFR. However, a caveat was that they had to make a series of assumptions to reach that accuracy, and without those assumptions, agreement was moderate ²⁶. Knutsson et al. performed a similar study on 118 femur fractures in the SFR, all from the same hospital. They found an almost perfect agreement for the AO type and a substantial agreement with the AO group ²⁷. Agreement between observers in these studies refers to the Landis and Koch scale regarding Cohen's Kappa ²⁸. These studies were primarily performed on data from Gothenburg and the clinics most closely associated with the SFR. However,

in a study by Sundkvist et al., looking at basocervical femoral neck fractures from the SFR, 868 out of 1185 fractures (73%) were excluded from the study due to misclassification²⁹. Thus, it is unclear how the previous studies apply to the SFR as a whole.

1.1.2.2 *Health outcomes after a fracture*

Health or patient functional outcomes can be measured as patient-reported outcome measures (PROMs). The SFR collects PROM using the EuroQuol 5 Dimensions (EQ-5D)³⁰ and the short musculoskeletal function assessment (SMFA) survey³¹. Each patient registered in the SFR receives a survey where they estimate their PROM just before the injury (PROM0). One year after the injury, they receive another survey where they estimate their current PROM (PROM1). An additional outcome, death, is added to the SFR, and some patients will die before answering the surveys. Death is collected via the PID from the National Population Register in Sweden.

EQ-5D™ is a self-reporting tool that measures five dimensions of health:

1. mobility
2. self-care
3. usual activities
4. pain/discomfort
5. anxiety/depression.

It also consists of a 0–100 VAS scale on which participants estimate their overall health status, where 0 is the worst and 100 is the best³⁰. The EQ-5D was developed in the 1980s as a non-disease-specific, standardized tool for measuring health. Specific versions have been developed and are standardized for different populations, e.g., Spain, Sweden, Japan, Algeria, etc.

The original EQ-5D consisted of three levels of answers and is now referred to as EQ-5D-3L (3L). The EQ-5D-5L (5L) is a recent version with five levels^{32,33}. The reason behind this upscaling was that it was difficult to distinguish between changes in health outcomes. There were attempts to solve this using unofficial 5L, but the current and official 5L have been extensively researched and validated^{34–39}. Indeed, Janssen et al. compared the 3L to the 5L for 3,919 individuals in six countries and found that 5L had more discriminative power³⁹. van Hout et al. showed that it was possible to translate 3L into 5L. However, the mapping can only reach the value space of the 3L; it needs to be updated for

each specific population and is only valid for the EQ-5D index ⁴⁰. The SFR moved from the 3L to the 5L in 2018–2019.

SFMA ³¹ is a simplified version of the original musculoskeletal function assessment ^{41,42}. It is a self-report health-status questionnaire designed to detect functional status differences in patients with common musculoskeletal disorders. The SFMA measures how bothered the patient is by these conditions. It consists of 34 items to measure dysfunction and 12 items to measure “bother,” i.e., how bothered they are by the dysfunction, on a 5-point scale. The answers are summed and transformed into an index on a scale of 0–100, where 0 is the best function, and 100 is the worst function.

The SFR does not contain information on comorbidities and other risk factors. Survey respondents answer questions regarding smoking in the PROM. However, information on diseases such as diabetes, cortisone-requiring diseases, alcohol consumption, etc., which is essential for patient outcomes, is not reported.

The registry also contains information on complications via surgeon reporting—e.g., reoperations—or is answered as part of the PROM1 survey. However, this data is incomplete or absent, e.g., if the patient died.

The PROM response rate, i.e., PROM0 to PROM1, is expected to decrease since there are two surveys to answer, one year apart. It is not easy to know if those who respond differ in any way from those who do not. Therefore, Juto et al. conducted a study in 2017 to answer this. Comparing SFR responders to non-responders, they found that non-responders and responders had similar functions ⁴³. As mentioned, some patients will pass away during the time between the two surveys, which is recorded in the SFR.

1.2 Fracture classification

Orthopedic decision-making is not straightforward. Kodama et al. ⁴⁴ and Neuhaus et al. ⁴⁵ studied the factors that influence treatment decisions for orthopedic surgeons. Both found that the appearance of the fracture in the radiographic image was the dominant factor. While both studies considered distal radius fractures, radiographic imaging is crucial in all orthopedic decision-making. For spine surgery, MRI and CT are more important. However, imaging is still vital.

Given that radiographs constitute a significant decision criterion in traumatology, it is unsurprising that radiologists and orthopedic surgeons try to understand

them better. This is usually done by grouping and attempting to use these groups to make decisions and fracture classification systems. Numerous classification systems are used in orthopedics and radiology. Audigé et al. examined 44 different classification systems for eight different localizations ⁴⁶, while Shehovych et al. noted 15 recognized classification systems for distal radius fractures alone ⁴⁷. Gilbert et al. studied three different classification systems for glenoid fracture classification ⁴⁸.

1.2.1 Classification systems

There are many classification systems. We present a few examples of interest to our discussion, but as we saw above, there are countless more.

The **Lauge-Hansen classification** of ankle fractures dates back to the 1950s ^{49,50}. It classifies ankle fractures according to the foot's position at trauma (the rotational mechanism) and the force that caused the fracture ^{51,52}.

The **Danis-Weber classification** divides ankle (i.e., malleolar) fractures based on the radiographic appearance and the lesion relative to the syndesmosis. It divides fractures into infrasyndesmotic (type A), intrasyndesmotic (type B), and suprasyndesmotic (type C) fractures ⁵². The Danis-Weber ankle classification is a simplified version of the AO ankle classification scheme.

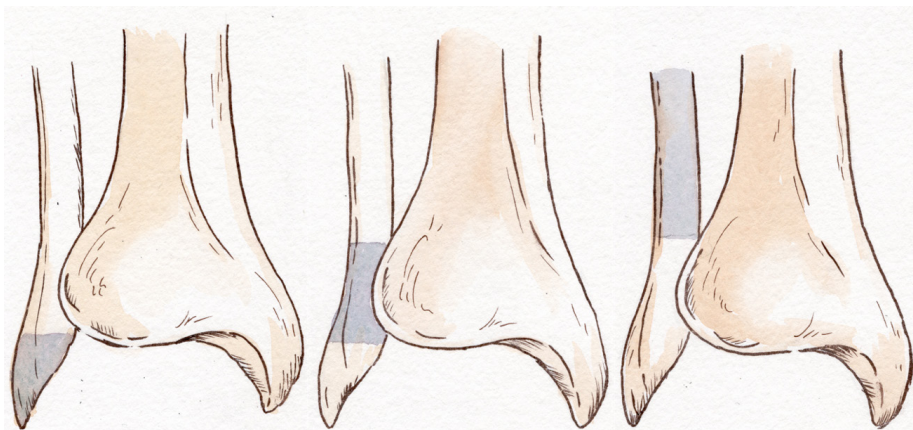


Figure 1. Danis-Weber classification. Infra- (type A), trans- (type B), and supra (type C) syndesmotic. Original artwork by Märta Nummelin

The **AO classification** is one of the most widely known and all-encompassing classification systems ^{24,53,54}. The latest update for long bone fractures came out in 2018 ⁵⁵. There are additional classifications for spine fractures ^{56,57} and pediatric fractures ^{58,59}. Unlike many fracture systems, it has been developed

over time and validated through multiple studies. However, it is generally considered complicated and cumbersome. Studies II and III focus on the AO 2018 ankle classification, and Study IV uses the AO classification implicitly, as lower extremity fractures are registered as AO fractures in the SFR ^{24,60}.

1.2.2 Utility and problems of fracture classification from radiographs

Few classification systems undergo validation before publication, and fewer are adequately validated. We expect a clinically used classification to be valid, reliable, and relevant ^{46,61,62}. The classification should be independent of the observer, say something about the injury, guide treatment, and positively impact the outcome. There are some crucial questions regarding the usefulness of classification systems, which will be addressed in turn.

A fundamental issue is the reproducibility between observers (IRR) and the same observer at different points in time (IOR) ^{46,61,62}. There are many examples, but we will mention a few. Neer's classification is a four-segment classification system based on the observation that humerus fractures tend to be displaced into four major segments: the lesser and greater tuberosity, the articular surface, and the humerus shaft ⁶³. Siebenrock and Gerber studied the reproducibility of classifications for humerus fractures and compared Neer's and the AO/ASIF classification (which developed into the current AO system). They found both systems had such poor reproducibility that they could not compare different studies ⁶⁴. Sidor et al. studied Neer's classification's reproducibility and found similarly poor agreement between observers ⁶⁵. Marongiu validated AO 2018 for humerus fractures compared to Neer's classification and the AO 2007 humerus classes. For AO 2018, they found an agreement similar to Neer's classification, which significantly improved the 2007 scheme ⁶⁶. Fonseca et al. compared the IRR of the major ankle fracture classification systems ⁶⁷: the Lauge-Hansen, the Danis-Weber, and the AO classifications. The Danis-Weber classification was the top performer, followed by the AO and the Lauge-Hansen. However, the Danis-Weber system had a moderate agreement ^{28,67}. At the same time, there are systematic ways to make the AO classification more reliable ^{68,69}. That we need these systematic ways to make AO more reliable, signals a problem of complexity – the common critique against the AO system. The reliability and validity of the Lauge-Hansen system were questioned by Lindsjö when it was clear that otherwise similar populations from different areas of the world had dissimilar fracture distributions in terms of the Lauge-Hansen class ⁷⁰. Later

attempts to reproduce the system by comparing injury footage or reproducing Lauge–Hansens experiments have generally failed, as we will see later.

Our 2017 study, Study I, found only moderate IRR for detecting fractures in radiographs ¹. However, in 2020, we found substantial IRR between human observers for the 2018 AO ankle classification. It varied more for individual subgroups (e.g., AO 44A1.1) and related to the number of cases ². In the same study, there was almost perfect agreement for detecting fractures. There are several reasons for this improvement. The reviewers had gained more experience with the review process and examining radiographs. In addition, the review process had improved, with labeling being performed on the original image in its original size, which was not in Study I.

Imagine creating a system that reliably and predictively reproduces and automates classification, removing IOR and IRR. It would enable wider usage, validation, and enhanced utility. In the long run, it would allow us to truly study whether fracture classification matters when we remove the human factor from the classification. It would enable us to determine whether the classification systems used and suggested in the future are relevant. One path towards this that we studied in this doctoral project was using AI and ML.

1.2.3 Ankle fracture classification – Lauge–Hansen vs. Danis–Weber vs. AO ankle

Studies II and III focus on detecting and classifying ankle fractures in radiographs. We repeat the discussion from Study III regarding ankle classification systems. There are three central classification systems for ankle fractures: the Lauge–Hansen, the Danis–Weber, and the AO classifications. In our studies, we used the Danis–Weber and expanded AO classifications ⁵⁵.

The Lauge–Hansen classification system is widely used to predict fracture patterns and ligamentous injuries based on injury mechanisms. Several studies have shown that Lauge–Hansen is only partially valid and reproducible. In 1985, Lindsjö raised the question of the poor reproducibility of the Lauge–Hansen system between different populations based on previous studies ⁷⁰. The findings of poor reproducibility have been repeated in several studies ^{67,71–74}. Gardner et al. performed an MRI study and found that Lauge–Hansen had limitations in predicting soft-tissue damage and ligamentous injuries ⁷³. Using actual injury footage, Kwon et al. replicated these findings in 2010 and 2012 ^{75–77}. Boszczyk et al. compared patient-reported injury mechanisms and radiographs and concluded the same, i.e., the reproducibility was poor ⁷⁴. Patton et al. concurred

based on CT findings and complete patient workups⁷⁸. Michelson et al. tried to replicate Lauge-Hansen's results physically, and in a separate study, so did Haraguchi and Arminger. Both failed and concluded that the Lauge-Hansen system could not be used to predict injury mechanisms or injury patterns^{79,80}. In the clinic, Lauge-Hansen and AO (complete or the simplified Danis-Weber) are often used together to guide treatment decisions.

The AO standard launched the Danis-Weber system. Danis-Weber bases its classification on the fracture's location in relation to the syndesmosis. This ligamentous joint holds the distal fibula and tibia together. In type A fractures, the fibula is broken below the syndesmosis (infra syndesmotic), type B fractures at the level of the syndesmosis (trans syndesmotic), and in type C fractures above the syndesmosis (supra syndesmotic).

The AO classification extends the Danis-Weber classification to include medial and posterior malleolus injuries. It grades fractures based on severity and physical appearance⁵⁵. The fracture types A-C are extended with a group number (A1-A3, B1-B3, and C1-C3) and then to a subgroup (A1.1-A1.3, etc.). Generally, a type A fracture is more stable than a type B fracture, which, in turn, is more stable than a type C fracture. The same goes for groups and subgroups; for example, A1.1 is more stable than C3.3.

The main criticism of the AO ankle system is that many consider it complex. Another criticism is that isolated medial malleolus fractures are treated as distal tibial fractures^{67,81}. The system also considers ligamentous injuries that are not readily visible in radiographs. However, they are visible in MR and during surgery.

Lauge-Hansen is mechanism-based and was created to solve the problem of deciding which ankle fractures to operate and how before imaging was widely available. The AO standard is based on injury appearance regardless of the mechanism. We aimed to develop AI models for easy, accurate, and rapid fracture classification and clinical decision-making. As we do not know the injury mechanism for each fracture in population-sized datasets, Lauge-Hansen is inappropriate for this task. As noted, Lauge-Hansen is not well suited to predicting injury mechanisms from radiographs in its current form, whereas AO is imaging-based. The classifications are similar, and conversions between the two systems have been suggested, but no fully agreed-upon or complete conversion exists^{77,82,69,83,52,84}.

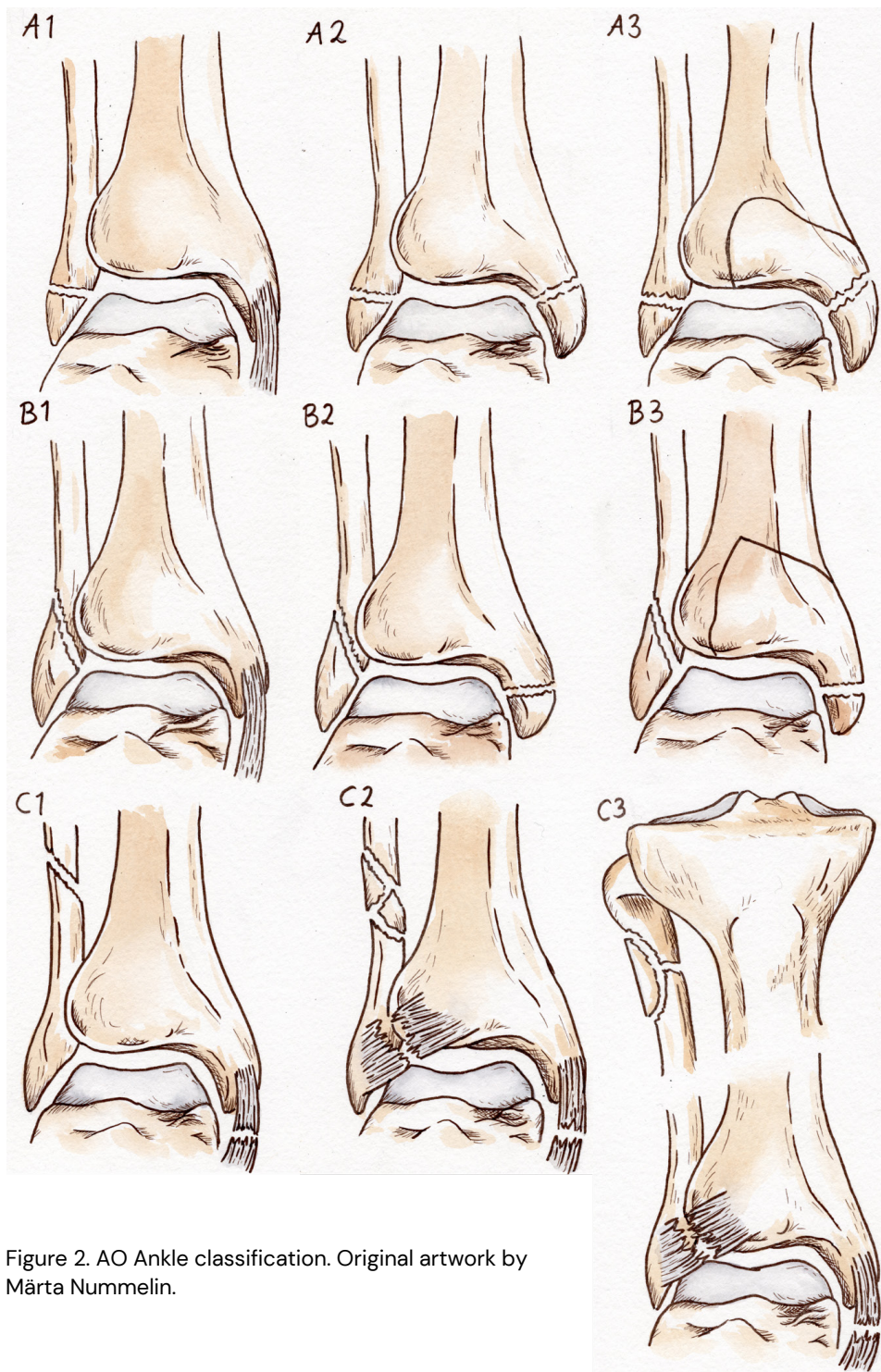


Figure 2. AO Ankle classification. Original artwork by Märta Nummelin.

1.3 Artificial intelligence, deep learning, and machine learning modeling

Intelligence is the ability to perform and learn techniques to solve problems and achieve goals appropriate to the context. Based on the original definition from 1956 by John McCarthy, AI is “the science and engineering of making intelligent machines.”⁸⁵ The EQUATOR network defines AI as “the science of developing computer systems which can perform tasks normally requiring human intelligence.”^{86–88}

Machine learning (ML), a field of AI, is the science that concerns the development of algorithms that learn patterns from data to solve problems rather than follow explicit rules^{89,90}. An **algorithm** is a sequential list of exact steps to solve a problem. However, ML algorithms **focus on parts of the problem, like computing rewards or learning**. It derives the way to combine the features, i.e., the weights, from interacting with the data⁸⁵. AI is about learning features, whereas e.g. statistics is about selecting features. We refer to a *trained* algorithm, i.e., the result of an AI or ML algorithm, as a **model**. ML draws from computer science, statistics, control theory, biology, and economics to study how computers can improve knowledge, perception, and thinking based on data experience. **Deep learning** (DL), currently the most successful approach, uses multilayer **artificial neural networks** (ANN) to compute continuous values (real numbers) over many iterations of the data.

Hierarchical structures, similar to communicating **neurons** in biological nervous systems, are the primary inspiration for ANNs. The visual cortex inspired the first ANN, so the simplest possible neural network – a single neuron – is called a **perceptron**. The neurons are mathematical functions, or computational nodes, that take an **input** (e.g., a single value, a vector, matrix, or tensor – the three or higher dimensional equivalent of matrices) and produce an **output**. By making the output from one neuron the input of another, we create a **layered** structure of information flow. We can have one neuron communicate with many other neurons, creating width. Connecting many layers in depth, layer after layer, leads to **deep networks**. Training these deep networks is the origin of DL. Initially, a network of 5–8 layers would be considered deep; today, networks can have hundreds or thousands of layers, and the output of one network can become the input of another ANN. Part of DL’s success comes from techniques to handle the increasing number of layers. By changing the mathematical functions in the neurons and the strength of signaling between nodes, we can have information

flow in the network in various ways. This allows the network to focus on different things in the data.

1.3.1 The case for ANNs – The Universal approximation theorem

The strength of the ANN approach is that theoretically and under certain constraints, there is an ANN that can approximate almost any mathematical function arbitrarily well. I.e., ANNs can be considered **universal function approximators** ^{91–96}. In this context, a mathematical function describes a relationship, via some form of computation, between input and output. For example, an image, sentence, or other data can be broken down into numbers and fed into a function. That function generates an output (new image, description of image contents, a translation, a prediction, or something else.) The relationships that can be modeled via ANNs are extensive. However, the universal approximation theorem for ANNs (several versions and extensions exist) only states that there is such an ANN, which is why ANNs are so helpful and popular. It does not say how to find it, which is a highly complex problem and is the reason why ANNs were not as widely used and popular before 2012.

The neurons in ANNs are composed of nonlinear mathematical functions (i.e., equations) called **activation functions**. There are two necessary conditions of ANNs: their activation functions are nonlinear and nonpolynomial ⁹¹. Without a non-linear activation function, a neural network is nothing more than a linear regression model. No matter how many neurons and layers there are, the output will be equivalent to a single linear regression perceptron. The analogy between biological and artificial neural networks has limitations. However, biological nerve cells are also nonlinear. They will only release a signal, or action potential, to the next neuron if the input matches a specific condition. Usually, this is a buildup in intracellular charge that exceeds a threshold potential, leading to a signal down the axon.

1.3.2 Training ANNs

The universal approximation theorem tells us that an ANN should exist that solves our problem arbitrarily well, but we do not know how to find it. The modern approach to ANNs is to train an ANN to approximate the theoretical function (or ANN) and produce a particular outcome. This is the machine learning part. There are different kinds of learning. In **supervised learning**, we know the desired learning outcome, and the ANN is trained to approximate the desired result. In **unsupervised learning**, the ANN gets data and finds patterns

independently, such as grouping a set of observations into groups that fit a pattern. However, it decides upon the pattern by itself. Some tasks fall in between and are often called **semi-supervised learning**. The simplest way to understand the learning process is to use supervised learning.

In the **training** phase, the ANN receives input with a known desired output, such as an image with a label stating what the image contains, a sentence in one language with a desired translation into another, or a question with the answer. The task could be for the model to produce the desired output. The input passes through a series of neurons, and each output serves as the input for the next neuron in a process called **forward propagation**. The last neuron produces the output from the ANN. The algorithm compares the output to the desired (true) outcome and computes an **error** via a **loss function**. The calculated error residual is propagated back through the ANN from the end to the beginning, making corrections in a process called **backpropagation**. The goal is to minimize the error (i.e., the difference between produced and desired output). The most significant change, i.e., the most correction, is found by following the **gradient**.

For this reason, backpropagation entails computing differential equations (i.e., of the activation function) and correcting by the magnitude of the gradient. The corrected network is then the starting point of the next training round. Theoretically, each neuron can communicate with every other neuron, but the importance of each neuron that sends it information is called the **weight**. This weight can be anything from zero (i.e., no information exchange or connection) to anything. The neuron's output can change by assigning different weights to different inputs. The training process aims to teach every neuron in every layer how much weight it is to assign to *each* input. The error correction, i.e., learning, is the gradient, and the model computes a correction between each successive network neuron. If the error is small, the gradient is small, the correction is minor, and vice versa. A final important note is that the training of an ANN is usually not deterministic. Introducing variability (variable data points) and randomness is central. This variability, enhanced by the nonlinear activation functions, allows different connections to form between neurons until they form stable connections (those that increase the likelihood of the correct output). This learning process from data – input, model, output, error, correction, repeat – is how most ML models conduct their training.

Convolutional neural networks (CNNs) are a class of networks well suited to processing grid-like data, such as images. They use **convolution layers**, a form of image processing filter layers that highlight the most essential features in images. Convolutions combine image features into successively more advanced features, e.g., combining pixels into lines, then lines into shapes, and the shapes into objects. CNNs are particularly common in image analysis tasks because they can learn much from relatively little information.

Recurrent neural networks (RNN) are particularly well-suited to sequential data, such as time series and text. Their neurons consist of a series of “hidden” states that act as a memory of previous states. The previous states are updated as a context to predict the next state—e.g., the next word or value in a time series.

A common problem with both these architectures in practical implementations is that information is lost downstream. The most recent neuron matters the most, and the signal is weaker the further you go. For CNNs, this usually entails losing spatial information (position and detail). For RNN, the training update signal is lost in a “vanishing gradient,” which is also a problem in CNNs.

Transformer neural networks (transformers) are another network architecture good at sequential data. Unlike RNNs and CNNs, which process data sequentially (meaning that the last part of the sequence will matter the most) and update their state, transformers look at the entire data sequence. Via an attention mechanism, the network can focus on different parts of the sequence regardless of where the information is in the sequence^{97–100}. Transformers are the foundation of the generative pre-trained transformers (GPT), which are currently in vogue. They make training problems easy to divide into subproblems (are easy to parallelize) and are excellent at capturing long-range relationships. They can also be applied to images and video but require much more data and computational and economic resources than CNNs or RNNs.

These three are the most widely encountered architectures today, but others exist, such as the multilayer perceptron, graph neural networks, generative adversarial neural networks (GANs), and autoencoders.

1.3.3 Recent advances

While the idea behind ANNs is not new, the current AI and DL innovation boom is relatively new. DL requires a lot of computational power, and only recent advancements in technology and algorithms have enabled it. Some examples are

improved software, utilizing the computational power of Graphics Processing Units (GPUs), and diving the problem into subproblems (distributed computing). Research has also evolved the solutions to the mathematical difficulties introduced by transferring theoretical mathematics to the constraints of the physical world.

Allowing for technological advancements that enabled efficient training, the great revolutions in CNN have consisted of empirical trial and error and minor tweaks to algorithms. E.g., what activation functions we should use, how to use them, how to process the input to each neuron, and how we choose how much weight each neuron puts on each input.

There are countless ML methods, but our studies are based primarily on CNNs. Therefore, we focus on these. The first genuinely successful CNN was LeNet-5, a seven-layer CNN ⁴. It could read handwritten characters and was intended for banks and the United States Postal Service to read bank checks and letters. However, due to technical limitations, CNNs did not achieve much further success for some time.

In 2012, a DL CNN won the ImageNet pattern recognition challenge, outperforming contending approaches ^{61,62}. In an instant, error rates fell from 25% to 10%. The most successful algorithm was the DL CNN AlexNet ¹⁰¹. Since then, AI and DL research has exploded. We mention some widely recognized milestones. An early milestone was the Network-in-Network (NIN) architecture ¹⁰², which built upon the ideas of AlexNet and added a small network within the network to allow for better information transference.

Chatfield et al.¹⁰³ designed VGG CNN S, an eight-layer CNN, in 2014. In the same year, Szegedy et al. ¹⁰⁴ introduced the Inception network (GoogleNet), which has since been updated multiple times ^{105,106}. Simoyan and Zisserman enhanced the VGG network with VGG-16 and VGG-19 CNNs, which were state-of-the-art at the time ¹⁰⁷.

LSTM networks ¹⁰⁸ (a form of RNNs) use “gates” to allow some information to pass through the network relatively unchanged (extended memory) and some gates to pass information from layer to layer (short memory). The long memory connections inspired Highway networks ¹⁰⁹. Highway networks took gated connections and introduced “skip connections” to allow the transfer of

information across the network. From the previous usable limit of 20 layers, Highway networks allowed 100 layers or more.

ResNet ¹¹⁰ was built upon the ideas of Highway networks, improving performance and stability. DenseNet ¹¹¹ tweaked ResNet by adding connections from every network layer to every other layer, improving performance with fewer parameters.

There are countless more CNN architectures. Those mentioned are a few of the most important or popular. They often serve as a starting point or reference for developing new applications or testing data.

1.3.4 AI in medicine

An **AI intervention** is an intervention that relies on an AI or ML component to serve its purpose ^{89,90}. Many interesting reviews examine AI and ML for various medical fields ^{112,113}. Hosny et al., to make a case for the increased adoption of AI, describe how the increased availability and need for medical imaging leads to an increased need for interpreting medical imaging and some of the advances being made ¹¹³. Liu et al. performed a review and meta-analysis of pathology detection in medical imaging in 2019 ¹¹⁴. They found that many DL studies reported results on par with healthcare professionals but that the level of reporting was generally poor and that results were difficult to verify.

While there is much hype about AI and ML, it is essential to note that it is one tool among others. ML models can fail because the more parameters they have, the more data they require. This is true, especially for CNNs, which can have thousands or millions of parameters. AI is not always better than modeling with other tools. For example, a review of 71 ML models found that they did not perform better than logistic regression models and were more prone to bias ¹¹⁵. Oosterhoff et al. tried eight different algorithms to predict outcomes after orthopedic trauma. All were trained on the same data (one logistic regression and seven different ML algorithms). They found that their non-ML algorithm tended to perform better than the rest. However, the training of the models was not explained ¹¹⁶.

Cary et al. looked at 30-day and 1-year mortality after hip fractures. They compared a multilayer perceptron (a form of ANN) and logistic regression on their dataset, and both performed similarly. Given that, they concluded that the logistic regression model was more accessible to clinicians to interpret and

required fewer computational resources. For that reason, it was the more reasonable tool ¹¹⁷. We fully concur.

1.3.5 ML and ANNs in orthopedics

Cabitza et al. provide a review of ML for orthopedics ¹¹⁸. The first paper in their review was from 2000 and used an ANN to control a trans-femoral prosthesis ¹¹⁹. In 2010, Pogorelc and Gam compared ANN to decision trees for gait analysis and found that ANN outperformed the other ¹²⁰. Nair et al. also studied gait analysis in patients with rheumatoid arthritis versus hip osteoarthritis ¹²¹. Prasoon et al. studied MRI scans of knee cartilage using ANN and found better performance than the state-of-the-art methods at the time ¹²². Thong et al. used ANN for the 3D reconstruction of an adolescent idiopathic scoliosis patient's spine ¹²³. Abidin et al. used ANN for chondrocyte pattern analysis to detect osteoarthritis in CT scans ¹²⁴. Shim et al. detected rotator-cuff tears in MRI studies ¹²⁵. To our knowledge, the first attempt to use ANN for fracture detection was by Al-Helo et al. 2013, who studied vertebrae fractures in CT scans with impressive accuracy ¹²⁶. The use of AI and ML has increased further.

1.3.6 ML and ANN for outcome prediction in Orthopedics

Dijkstra et al. conducted a systematic review of predictive ML models for orthopedic trauma and found 45 studies ¹²⁷. Most models were derived for hip fracture patients. Mortality and hospital stay were the most predicted outcomes. Some were ANNs, but none appear to have been CNNs. However, they excluded studies reporting on models analyzing diagnostic imaging, which is this thesis's foundation. The Machine Learning Consortium studied ANNs to predict infection risk after operative treatment using ANNs. They also tested different ML algorithms and one ANN but provided scarce information about modeling details

¹²⁸.

1.3.6.1 Mortality prediction after fracture using ANNs

In Study IV, we examine, among other things, whether the patient died within one year of the study. This has mostly been done for hip fractures because they have the highest mortality and worst post-fracture recovery. Lin et al. studied mortality after hip fractures and compared a logistic regression and ANN model to predict 1-year mortality, resulting in an AUC of 0.95 for the ANN vs an AUC of 0.78 for logistic regression ¹²⁹. While interesting, the excellent performance was probably due to overfitting. In a similar study by Shi et al., the results were an

AUC of 0.87 for the ANN vs. an AUC of 0.73 for logistic regression on a much larger dataset and testing many different networks.

Liu et al. systematically reviewed ML models for predicting mortality in hip fracture patients. For hip fractures, postoperatively, mortality is reported between 5% and 30% within one year. They found that ML models had an ideal mortality prediction after hip joint surgery. ANNs and random forest algorithms had the best performance and, in general, better accuracy than existing clinical scores¹³⁰. DeBaun et al. tried three models (LR, naive Bayes, and ANN) and found the ANN superior¹³¹.

In contrast, Oosterhoff et al., as mentioned previously, found no performance boost for ML algorithms over logistic regression¹¹⁶. Cary et al. looked at 30-day and 1-year mortality. They compared an ANN and logistic regression and found similar performance¹¹⁷. Chen et al. used an ANN to predict mortality after a hip fracture. They trained it on a national registry and found it worked better than Cox regression¹³². Cao et al. used all hip fracture patients registered in Sweden between 2008–2013 to model predictive preoperative characteristics for 30-mortality in traumatic hip patients after surgery. They cross-referenced with the Swedish National Board of Health and Welfare registers to get date of death and comorbidity data. Comparing logistic regression and ANN, the latter performed somewhat worse, but confidence intervals are not provided¹³³.

1.4 Ethical considerations and methodological biases

AI has many benefits and pitfalls, and we must consider its ethical implications. We elaborate on some common issues that clinicians should consider. The following discussion builds upon Olczak et al., 2021⁸⁶.

Outcome imbalance: Medical outcomes are often heavily skewed towards some specific and commonly occurring ones. A negative outcome is the most likely outcome for most disease tests – as most people are healthy for what is tested. However, this is not true under certain circumstances. We are unlikely to find ankle fractures if we randomly examine people's ankles in the street. We are considerably more likely to find fractures examining ankles after trauma in the ER. Where there are multiple outcomes, individual outcomes become less likely. We are more likely to find any ankle fracture than a fracture of the medial malleolus and the posterolateral rim (Volkman's fragment) – i.e., AO 44B3.3.

By emphasizing rare cases, we can alleviate the imbalance during training, also known as assigning weights to classes. We can also manipulate the images so that the network becomes less sensitive to particular features, known as data augmentation. This becomes more difficult during testing as the test examples are usually fewer than the training examples. We must also consider what algorithms we use in training and evaluating model performance depending on the dataset.

Missing data: We need many examples of the outcomes we are searching for to train a model. A rare outcome is not likely present in the data or can be so infrequent that we cannot get a good training result. The algorithm cannot learn a pathology if it is not in the training data. While we could write rules for an algorithm to follow, it is impractical to write rules for all possible outcomes that occur in the real world. Fundamentally, this differs from humans, who can understand a thing before encountering it, e.g., a Pipkin fracture in the hip. Thus, we can have unknown gaps in our models.

Overfitting: AI learns by studying examples. If the model learns the data too well, it learns training cases rather than the general features. An overfitted model will give a false sense of security and lead to a more biased model. This is a common concern in any form of statistical learning, and it is why data is split into at least a training and a test set, with no overlapping cases (e.g., patients). The test set contains examples the model never encounters during training and cannot learn. Another way is to compare the model to data from an independent location, i.e., external validation.

Data and privacy: ML models are powerful tools, but some are “data hungry.” ANNs can have hundreds, and sometimes millions, of parameters. They need a lot of data to optimize all parameters, learn the desired patterns, and capture unusual cases. Therefore, they thrive on large amounts of data during training, encouraging large-scale data collection. Large-scale data collection constitutes a risk to patient integrity and sometimes data ownership. Medical data cannot usually be shared due to its sensitive nature, which causes reproducibility, traceability, and reporting problems. It is also sometimes possible to extract the underlying training data from the ANN model, further complicating privacy and integrity.

Bias and fairness: Bias comes from input data and design decisions during development. Biases are mapped to the output. This means that the AI model

will learn and reproduce prejudice in the data ¹³⁴. Common confounders and biases are race, gender, and socioeconomic factors. For example, an AI melanoma detector was trained on a population dominated by fair skin and was shown to perform worse on dark-skinned patients ^{13,14}. A study by Zech et al. studied chest radiographs from different sites and found that the CNN could implicitly learn where the data points came from and adjust predictions accordingly ¹³⁵. In another study by the same research group, hip fractures from radiographs were studied with high accuracy. However, correcting for socioeconomic, logistical, and healthcare process data factors (e.g., different scanners, locations, age), they could show that their model performance fell to a random detector ¹³⁶. Recognizing, reflecting, and examining biases in AI studies is essential ¹³⁷.

Informed consent and autonomy: AI risks autonomy (the right to self-determination) and integrity. AI models return outcomes based on opaque datasets, and their results are difficult to explain to patients. There is also a risk that decision-making responsibility will be diverted from clinicians to algorithms that “perform superior to an expert.” Healthcare systems and clinicians might implicitly become forced to implement and follow AI recommendations, forcing patients to subject themselves to AI ¹³⁸.

Interpretability and safety: Transparency is crucial for clinical AI implementations, with the critical implications that can come with errors. The preferred option is to share data and as much information about the model as possible. A problem not unique to medicine is that data can be very sensitive. The data cannot always be legally shared. In addition, there are risks in releasing development and research models to the public, for example, for public scrutiny. They risk being used for other things than validating the model. This can cause considerable harm as an unfinished or unvalidated tool is used outside the intended context ¹³⁹.

AI models are often described as “black boxes.” The model's decision processes are largely unknown, and so is what happens inside. Enhanced transparency and interpretability of the algorithms could compensate for this. Understanding the inner workings of ML models is a field that is actively researched and is constantly evolving. However, we need to learn to interpret ML models, and some argue that we must create interpretable models from the start ¹⁴⁰. For example, a popular attempt to understand CNN models is to visualize the regions that

activate the model toward a specific classification decision. One popular method is activation (or heat/saliency) maps that show what areas of an image the model reacted to. Another method bounds the region of interest into boxes. However, whether the incorrect or correct region is emphasized, neither explains why the model reacted to that region, and their ability to explain the model is incomplete ¹⁴⁰.

Responsibility and liability: How to allocate responsibility and liability for an AI intervention is unclear. A model that is 99% correct is wrong 1% of the time. Even excellent AI models fail in obvious cases. Suppose an AI recommendation was accurate, and not following it harmed the patient. Are clinicians responsible for not following the “black box” recommendation of the model? Suppose the AI recommendation was followed, which resulted in a critical error, constituting malpractice. Who is responsible? Who is liable? Both legally and morally? This fundamental issue must be resolved before AI decision-making replaces clinicians’ judgment: the creators of AI models need to accept legal responsibility for the outcomes of AI models.

Reproducibility: Traditionally, machine learning has been presented to a non-medical community, but as the research has moved into medicine, it poses new challenges. While reproducibility is fundamental to all sciences, there are differences in focus between reporting traditional ML and medicine – resulting in problems specific to AI interventions.

The EQUATOR network ⁸⁷ is the originator of evidence-based reporting guidelines for medical research. Well-known guidelines include SPIRIT ¹⁴¹, CONSORT ¹⁴², and STROBE ¹⁴³. In recognition of the increasing prevalence of AI intervention research but poor reporting, the EQUATOR network has created the CONSORT-AI ⁸⁹ and SPIRIT-AI ⁹⁰ addendums ⁸⁸. Both focus on clinical trials (trial protocols and trial reporting) containing AI interventions.

Protocols for reporting on prognostic and diagnostic studies using ML and AI were published in 2024 via TRIPOD+AI ^{144,145}, while STARD-AI is still in development ⁸⁸. However, Olczak et al. ⁸⁶ also examined the different aspects of research, implementation, and reporting of AI interventions. Similar checklists exist for reporting AI and ML in medicine. Such guidelines should help avoid common problems specific to the medical domain.

Overdiagnosis: There are risks with the probable over-availability of AI models, where we can upload any data and easily and cheaply get a result. Routinely using a cheap and fast model can lead to overdiagnosis, or even correct diagnosis, of benign conditions. This can lead to unnecessary psychological suffering, overuse of healthcare resources, and unnecessary treatments, which in turn can lead to complications and more suffering.

1.5 Discussion and conclusion

Orthopedic trauma is a considerable part of the global health burden, and with an aging population, this will increase. This thesis project envisions a system that can help clinicians and researchers on multiple levels. We strive towards an automated fracture classification system to improve interrater and intra-observer reliability. It could allow for backward utility, i.e., application to previous data and studies, and forward utility, i.e., a clinical intervention as part of a computer-aided decision system. As such, we envision a system where we can predict patient outcomes, which could greatly improve patients' quality of care and aid in research.

2 Research aims

This thesis aimed to explore various facets of pathology detection and prediction using artificial neural networks.

The specific aims of this thesis were to:

1. Explore CNN for image analysis and classification in orthopedic medicine and transfer learning.
2. Develop complex fracture classification, particularly for ankle fracture classification, according to the AO standard.
3. Explore CNN model verification, transferability, and generalizability of an AI model to a clinical setting.
4. Explore using neural networks for patient outcome prediction using PROM with the purpose of personalizing orthopedic medicine.

3 Materials and methods

3.1 Study design

Studies I–III were cross-sectional studies, and Study IV was a cohort study. Study I and II were single-center studies, whereas Study III added an external validation site. Study IV was a multicenter study that used an additional external site for model evaluation.

3.2 Data sources

Studies I and II used data from a single site (Danderyd Hospital, Stockholm, Sweden). Study III used the same data source as studies I and II but added an external validation site (Flinders Medical Centre, Adelaide, Australia) as an external validation dataset (EVD). It used the test set of Study II as an internal validation dataset (IVD). Study IV used register data from the SFR and related imaging collected from major trauma hospitals in the Stockholm region of Sweden. We also collected imaging from Gotland, Sweden, for the EVD.

3.2.1 Danderyd Hospital, Stockholm, Sweden (DS)

IMAGING: Images were collected from DS PACS for all traumatic imaging at DS between 2002 and 2015.

REPORTS: Radiology reports were collected from the RIS system.

All data was anonymized upon collection.

3.2.2 Flinders Medical Centre, Adelaide, Australia (FMC)

IMAGING: We received 399 anonymous radiologic studies of post-ankle trauma emergency imaging. Studies were selected and provided by our research collaborators connected to Flinders University Medical Centre in Adelaide, Australia. We did not receive radiology reports or patient data.

REPORTS: We had no radiologist reports for the Flinders data.

All data was anonymized upon collection.

3.2.3 The Swedish Fracture Registry (SFR)

Data was collected on all fractures registered at the seven emergency hospitals in the greater Stockholm region between 2011-01-01 and 2019-06-30. We also collected data on all fractures from Region Gotland during the same period.

Data was pseudonymized upon collection, and the unique personal identification number (PID) of all patients was only used to ensure there was no overlap between different study populations and to associate imaging with patients and fractures.

3.2.4 Region Stockholm (RS)

Region Stockholm has a joint PACS database managed by "Bild och funktionstjänsten" (BFT) and SPECTRA AB. We collected radiographic imaging of the fractures registered in the SFR during 2023. The seven major emergency hospitals in the Stockholm Region were included.

1. Capio S:t Görans Hospital
2. Danderyd Hospital
3. Karolinska University Hospital, Huddinge
4. Karolinska University Hospital, Solna
5. Södersjukhuset Hospital
6. Södertälje Hospital
7. Tio Etthundra Norrtälje Hospital

Data was pseudonymized after collection, and the PID of all patients was only used to associate imaging with pseudonymous SFR data.

3.2.5 Region Gotland (RG)

In December 2020, we collected the radiographic imaging of fractures in the SFR for Region Gotland. Imaging was collected for all registered fractures during the study period and one year forward. The eighth hospital in Study IV was thus:

8. Visby Lasarett

Data was pseudonymized after collection, and the PID of all patients was only used to associate imaging with pseudonymous SFR data.

3.3 Neural networks

3.3.1 Images and imaging

The labeling evolved over the studies. When present, labels were extracted from the DICOM metadata.

For Study I, the primary outcome was fracture (yes/no). We used an unsupervised ML technique called Latent Dirichlet (LDA), a form of natural

language processing (NLP), to extract report topics from the radiologist reports^{146,147}. LDA is based on the idea that unique combinations of words are used in texts depending on topics. These topics create groups. We used these groups to create regular expressions that assigned class labels to studies based on the radiographic report. The fracture label was assigned for the entire study.

The gold standard/test set was randomly selected from the training data based on these autogenerated labels. We aimed for a 50/50 split between fracture and no fracture. We used manual labeling by human reviewers to assign labels to the gold standard set. The reviewers first labeled the test set independently. We then held a consensus session to determine the labels for images where there was reviewer disagreement, with a majority vote deciding the final label.

After the test set had been extracted, a subset of radiographs and labels in the training data were also manually reviewed. This was done to evaluate the automatic label generation and improve the labeling quality. We also held error review sessions, where we went through subsets of network classification errors that did not agree with the labels.

The CNN outcome labels were based on each image rather than the entire study, in contrast with the automatic label generation for fracture, which labeled the study. The secondary outcomes body part, laterality, and exam view were extracted from the DICOM metadata. When available in the metadata, they were unique to each image.

The labelers were a medical student, a resident radiologist, a senior consultant radiologist, one consultant orthopedic trauma surgeon, and two senior trauma consultants.

For studies II and III, labels were based on the AO 2018 standard. A manual review of all radiographs and studies was required to assign classes. Labeling was performed using the Raiddex platform, an in-house-developed labeling tool. The gold standard set ("test set" in Study II and IVD in Study III) was randomly selected from the labels generated for Study I. The goal was to get a set with two-thirds "fracture" and one-third "no fracture." After division into train and test sets, each set was independently labeled according to the AO ankle classification.

For Study IV, labels were based on data in the SFR register. The primary outcomes were PROM1 or death within the study period. When possible, we

calculated the one-year change in PROM, $PROM\Delta$. If the patient died during the study period, they could not have answered the PROM1 survey.

$$\text{One year change in PROM} = \text{PROM1} - \text{PROM0}$$

The one-year change in PROM was derived from PROM0, which patients answered weeks post-trauma. We believed it was less reliable than PROM1 and had too many confounders that we could not correct for in this study. Therefore, we considered Δ PROMs as important secondary outcomes. Less important secondary outcomes were AO class, as reported in the SFR. When assigning images to fractures, we used the DICOM metadata to match images with fractures, as multiple fractures could be present in the same study. We selected all series that studied the fractured region within seven days of the trauma to capture post-intervention and immediate follow-ups. We only looked at lower extremity fractures, i.e., from the femur and distally. We also included adjacent imaging when available. However, we generally defined adjacent as the most proximal and distal segments to the fractured segment. The reason was that we wanted to capture more facets of the fractured region.

3.3.2 Image transformations

As CNNs can only learn to detect outcomes they have seen, having a wide range of data is essential. The goal is to make the activation general yet specific to the outcome. We want to change the information content of the image but keep the vital information the same. Transformed images were passed to the network with the same training labels as the original image.

Rotation and reflections: By rotating and reflecting the image, we give the network different angles and perspectives of the same object or type of object, but the fundamental information remains the same.

Jittering: Altering the pixel values by other proximal pixels is called jittering. Jittering can make the image look more “grainy” and less sharp, enforcing other features and training the network to look at less sharp images.

Cropping or blocking: Cutting out smaller regions of images is a way to change the information content in the image. Usually, we hope that the relevant region remains, in our case, the fracture, but this is not guaranteed. But cropping is random; the same area will not be cropped, and often, the thing of most interest will remain. It could also enforce the learning of different class features, as the

most prominent feature might be hidden, and the network must depend on the less noticeable feature of the class.

3.3.3 Convolutional neural networks

For Study I, we evaluated a series of different neural networks (AlexNet, VGG 8, VGG-16, VGG-19, and Network In Network). For studies II through IV, we used the ResNet architecture. Similarly to Study I, we initially tested different network architectures (ResNet, DenseNet¹⁴⁸, and Inception^{105,106}) and found ResNet performed best for our task. All network architectures were freely available, open-source networks.

3.3.3.1 AlexNet (BVLC reference net)

AlexNet¹⁰¹ was the original neural network implementation that sparked the AI and CNN boom. It uses rectified linear units (ReLU) nonlinear functions instead of others that were popular when it was introduced. It has eight layers; the first five are convolutional, and the remaining three are fully connected. AlexNet was one of the CNNs studied in Study I.

3.3.3.2 VGG 8, 16, 19 layers

The Visual Geometry Group (VGG) S (8 layers)¹⁰³ and VGG-16 and VGG-19 networks¹⁰⁷ are CNNs that improved upon AlexNet's architecture and performance by making the neural networks "deeper." Sixteen and nineteen layers were the deepest that still allowed for proper training. Deeper networks were "too" deep as the training signal (the gradient) became too small for model training. VGG networks were evaluated in Study I.

3.3.3.3 Network In Network

Network in Network was an attempt to improve CNN's ability to study local image patches¹⁰². It was evaluated in Study I.

3.3.3.4 ResNet

ResNet is built upon the VGG architecture with up to 50 or 100 neurons. It used skip connections for residuals, which allowed better network training as the gradient update could pass deeper down the network¹⁴⁹.

While there are newer CNN architectures, these are robust and still widely used today.

3.3.3.5 Classification using one-hot encoding

Classification evaluation was done using one-hot encoding. This means that each outcome is trained and tested independently of the others. In our models, fracture, 44A, 44A1, and 44A1.1 are four different and independently determined classes. Each study is evaluated against all possible classes.

3.4 Statistics

Evaluation metrics are important to model building, particularly in machine learning, where the model inputs and outputs are complex. Evaluation metrics tell us if a model does what it purports to do. Table 1 displays various performance metrics used in medicine and machine learning that were most relevant to our studies. As Olczak et al. 2021⁸⁶ discussed, we must balance our presentation between absolute correctness and the intended audience. We follow the recommendations therein.

Table 1. Evaluation metrics.

Measure	Calculation or description
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$
Recall, true positive rate (TPR), Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{FP + TN}$
Youden J	sensitivity + specificity – 1
False positive rate (FPR)	$\frac{FP}{FP + TN} = 1 - \text{specificity}$
Precision, Positive predictive value (PPV)	$\frac{TP}{FP + TP}$
Negative predictive value (NPV)	$\frac{TN}{TN + FN}$
Model performance curves	
Receiver operating characteristic (ROC) curve	sensitivity (y-axis) against (1–specificity) (x-axis), i.e., TPR against FPR
Precision–recall (PR) curve	Precision (y-axis) against sensitivity (x-axis)
Area under the curve (AUC)	
AUC of the ROC curve (AUROC)	Statistic of model performance
AUC of the PR-curve (AUPR)	Statistic of model performance
Regression or ordinal data modeling errors	
Standard deviation (SD)	$\sqrt{\frac{1}{\text{samples} - 1} \sum (\text{prediction} - \text{mean value})^2}$

Table 1. Evaluation metrics.

Measure	Calculation or description
Means squared error (MSE)	$\frac{\sum(\text{true value} - \text{prediction})^2}{\text{number of cases}}$
Root mean squared error (RMSE)	$\sqrt{\text{MSE}}$
Mean absolute error (MAE)	$\frac{\sum \text{true value} - \text{prediction} }{\text{number of cases}}$
Multiple measurements	
Frequency weighted average	$\frac{\sum_{i=1}^{\text{categories}} \text{measurement}_i \cdot n_i}{\sum_{i=1}^{\text{categories}} n_i}$

TP (True positive), FP (False positive), TN (True negative), FN (False negative). Table adapted from Olczak et al. 2021⁸⁶.

3.4.1 Balanced vs. imbalanced problems

If there are two outcomes to a model (e.g., fracture yes/no), the model is a binary classifier. We call it a multilabel classifier if we have more than two outcomes. It is common practice in classification tasks to reformulate a multilabel outcome as a binary task. If we have three classes, e.g., ankle fracture Weber A, B, and C, a classifier will often translate the problem into three separate tasks: Webber A/not A, Weber B/not B, and C/not C. As the number of outcomes increases, the “not” class will become more prevalent relative to the individual classes. A dataset can be both balanced and imbalanced at the same time. We can have a 50/50 distribution for fracture yes/no, but a subgroup, e.g., Pipkin fracture, can be on in a ten thousand¹⁵⁰.

3.4.2 Dataset size selection

A random classifier should always be able to reach the accuracy of the most dominant class by simply guessing that outcome. Therefore, selecting a 50/50 positive/negative outcome in the test set is customary. If the classifier obtained an accuracy of 40%, we could flip the labels and have 60% accuracy. Some argue that you should always select 50/50 data distribution, where one-half is no finding. This is not always possible or reasonable. For example, there is no default option in our Webber example where the outcomes are Webber A, B, and C. We should have a 33/33/33 test class split. If we were to add “no fracture” and make that 50% of the test, there would be a 17.5/17.5/17.5/50. Any model could always reach at least 50% accuracy by guessing no pathology, making a lower modeling accuracy unlikely.

As the number of classes increases, this becomes more difficult, and as we work with real-world problems, this can become impractical, impossible, and perhaps

unethical. Studies II and III have approximately 40 possible outcomes for ankle classes. If we wanted to balance the training or test set, we would have had to peek at the dataset before selecting them – introducing bias and perhaps losing important information. Some outcomes are not presented at the necessary proportion, and some are not present at all. In addition, if we have many outcomes and want them represented, the amount of data separated for testing might further remove the few existing cases or fail to include any of the cases in the test set. For this reason, studies II and III aimed at 2/3 fractures, and the rest had no fracture in the test data.

3.4.3 Accuracy

Accuracy is commonly used to measure the proportion of correct guesses compared to all guesses. Each instance is equally important, including the TN. Accuracy can be misleading for imbalanced problems when the TN can dominate and is generally not recommended for imbalanced data ⁸⁶. Take the AO ankle classification down to subgroup classification, as in studies II and III, in a perfectly balanced dataset (44A1.1 – 44C3.3 and no fracture, i.e., 27 outcomes). If we have a model that says no to whatever outcome, the classifier will have approximately 96% accuracy and perfect specificity (probability that the test returns negative if the thing tested is negative). However, sensitivity (the probability of a positive test result given that the condition is positive) would be zero. The expected random accuracy of a classifier with n classes is

$$ACC_{random} = \sum_{i=1}^n p_i^2$$

3.4.4 Precision and recall

Unlike accuracy and specificity, precision and sensitivity (i.e., recall) do not consider TN, making them better suited to imbalanced problems. Precision and recall are defined in Table 1. Precision is the likelihood that a positive prediction is truly positive, while recall is the proportion of actual positives that are correctly identified.

3.4.5 Area under the curve (AUC)

Area under the curve (AUC) measures are a way to evaluate the overall performance of models and not specific instances of the model. When constructing a model, we must select a threshold (or cut-off) value (often a probability) at which the test returns positive or negative. We might, for example,

decide that the test will return positive if it is more than 50% likely true, i.e., $p > 0.5$; otherwise, it will return negative. For a screening test, we usually prefer a lower threshold. We can get different performance metrics for the model depending on whether we set this probability high or low. Performance measures, such as specificity, sensitivity, accuracy, precision, etc., rely on this probability threshold and are threshold-dependent. Selecting a threshold can be arbitrary, based on experience, or one can try to optimize it as a parameter based on experiments.

A way around this is to look at different thresholds and performances across a range of thresholds. Plotting the threshold-specific measures into a diagram for different thresholds will result in a curve. The area under the positive outcome guesses is the area under the curve (AUC). This is a summary statistic that speaks to overall model performance. However, it does not say anything about the specific components. It will not capture the actual best possible or worst possible performance but an average over the range of all thresholds. We need to look at the actual curves to understand each component properly. Also, we must decide on a particular threshold when implementing the model.

3.4.6 Area under the receiver operating characteristic curve (AUROC)

The area under the receiver operating characteristic curve (AUROC) looks at the true positive rate vs the false positive rate. It measures the overall performance of a model over all thresholds or independently of thresholds. The AUROC (often abbreviated as AUC in the literature) is widely used. AUROC measures the ability of the model to assign a higher probability to a randomly chosen true positive case than a true negative case. Random AUC is always 0.5, or 50%. It considers true negative instances equal to positives, making it poorly suited for imbalanced datasets.

3.4.7 The area under the precision-recall curve (AUPR)

A precision (positive predictive rate) and recall (sensitivity, the true positive rate) curve (PR curve) focuses on performance in the positive class. It ignores true negative cases, making it useful for imbalanced datasets. The area under the PR curve (AUPR) measures the trade-off between precision and recall across all thresholds. A random classifier will give the AUPR as the prevalence of the class in the data.

$$AUPR_{\text{random}} = \text{number of cases for the class} / \text{total number of cases}$$

A random classifier should give AUPR 0.2 for a class that makes up 20% of the data. Anything above is better than chance ¹⁵¹.

Deciding how to measure model performance can be challenging. There are different schools of thought, and there is no best way for all situations, but it is an active field of research. We prefer AUC performance measures for model development, but once you intend to deploy a model, you need to decide on a decision threshold. There are ways to extract the “optimal threshold” for both curves, but the optimal performance is not always the desired outcome.

3.4.8 Bootstrapping confidence intervals (CI)

Bootstrapping is a statistical sampling procedure often used to generate probability distributions, such as ninety-five % confidence intervals (95% CI) ^{152,153}. Bootstrapping consists of randomly sampling data points from the dataset with replacement. We generate a distribution by repeatedly sampling the same number of points as our original dataset. Repeating this many times allows us to assess variability and derive confidence intervals. Therefore, we can estimate how representative our outcome is compared to chance. We used bootstrapping to calculate confidence intervals.

3.4.9 Top-N performer

Top-N performer means that, for a multilabel classifier, we look at the N most likely outcomes, and if your outcome is one of those, you are partially correct. We only looked at the top-1 performers, i.e., the most likely (highest probability) outcomes in our studies.

3.4.10 Weighted average

There are different ways to calculate aggregate average performance for a multilabel classifier. Averaging the performance metric over the number of classes (i.e., the arithmetic mean or macro average) gives equal weight to all outcomes. A rare class performing exceedingly well or poorly will have a disproportionate influence. In an imbalanced set, this can matter a lot. If we instead weigh the outcome based on the number of instances in the class, we get the micro average, also known as the frequency-weighted average.

$$\text{frequency-weighted average} = \frac{\sum_{\text{case}=1}^{\text{last}} n_{\text{case}} \cdot \text{measure}_{\text{case}}}{\sum_{\text{case}=1}^{\text{last}} n_{\text{case}}},$$

where n is the number of cases ⁸⁶.

3.5 Data and population

3.5.1 Study I

3.5.1.1 *Study population*

The data represented a random subset of patients who had ankle, foot, wrist, or hand fractures and were radiographically examined at DS between 2002 and 2015. Pediatric patients, i.e., having open physes, were excluded. All data was collected anonymously, and there was no way to identify patients or derive population statistics.

3.5.1.2 *Images, radiology reports, and labels*

The primary labeling outcome was the presence or absence of a visible fracture in the radiograph (fracture yes/no). Fracture classification labels were first generated from the radiologist reports associated with each study. LDA is a form of unsupervised machine learning for NLP, i.e., text analysis^{154–157}. The radiology reports were analyzed using LDA to generate report topics. These topics were manually refined and used to extract labels from radiologist reports^{146,147}.

Secondary outcomes – side/laterality, body part, and exam view – were extracted from the DICOM image metadata.

3.5.1.3 *Training data*

The original data consisted of 256,458 radiographs. We divided the patients into an 80/20/10 train/training validation/test split. The training and validation data were used for model training.

3.5.1.4 *Test data/gold standard*

A random test set of 400 images (from the same number of patients) was selected from the patient test dataset. Two senior orthopedic consultants independently reviewed and labeled the radiographs. The radiographs were reviewed at the same resolution as the network (256x256 pixels), with all available views and the radiologist's report. Afterward, a consensus session was held for all radiographs on which the reviewers disagreed, resulting in a fracture/no fracture gold standard. The review process was inspired by Audigé et al.^{46,61}

3.5.1.5 *Network performance review*

We tested multiple CNNs, as described below. After the best-performing network for the fracture detection task had been determined, a network error review was conducted. We selected 200 radiographs where the CNN had misclassified the exam view, 200 radiographs where it had misclassified the laterality, and all radiographs where the body part had been incorrectly classified. We manually reviewed all these images for the respective category. All images were also examined for fracture presence. A senior radiologist consultant reviewed the exam view outcome alone, whereas JO (medical student) and MG (consultant orthopedic trauma surgeon) reviewed fracture, body part, and laterality.

3.5.2 **Study II**

3.5.2.1 *Study population*

The study population was a subset of the same collected dataset as in Study I, i.e., the population of patients from DS between 2002 and 2015 without any population parameters or identifiable information. Only ankle imaging was included, and studies with open physes were excluded.

3.5.2.2 *Images, radiology reports, and labels*

Study II only included ankle-level imaging. Pediatric images (i.e., open physes) were excluded because they are classified differently from adult fractures^{55,59,158}. As in Study I, initial study labeling (fracture/no fracture) was performed with automated text analysis based on radiology reports. Studies were then separated into training and test sets before AO classification.

Reviewers looked at the full-resolution images and labeled the entire study using all images. Labelers had access to the radiologist's report during labeling. Radiologist reports never contained the AO classification; however, sometimes, the location was mentioned according to the Danis-Weber classification (infra, trans, or supra syndesmotic)⁵⁵. To the extent that tibia or fibular fractures were visible, they were also labeled according to the AO 2018 classification standard. If fractures were visible in the foot, these fractures were labeled according to bone location (e.g., os talus, os calcaneus, os cuboid, etc.).

Primary outcomes were AO ankle fractures (i.e., segment 44). Other fractures were secondary outcomes, e.g., fibula (4F2), tibia (42 or 43), etc.). During image

classification and training, studies were labeled using a purpose-built labeling platform Raiddex created in-house.

3.5.2.3 Training data

Study II only examined fractures visible in ankle imaging. Compared to Study I, ankle fracture data was expanded with additional ankle studies. The training data was labeled by a group of five reviewers consisting of a senior consultant orthopedic trauma surgeon (AS), a consultant orthopedic trauma surgeon (MG), an emergency medicine specialist (TA), a junior doctor (JO), and a fifth-year medical student (FE). TA, JO, and FE were specially trained for the labeling task and labeled between 2000 and 4000 exams each.

At least two out of five reviewers reviewed each study included in the training set. If there were any discrepancies between reviewers, MG reviewed the exam and decided on the final class. The training set included only outcomes with at least five cases.

3.5.2.4 Test set/gold standard

The test set consisted of 400 randomly selected patients to ensure no overlap between the training and the test set. To accommodate the large number of classes and the non-specificity of the initial automated labeling, 2/3 of the studies were selected to have a fracture label. All studies (409) of the selected patients were included in the test set.

Two orthopedic trauma surgeons (MG and AS) independently reviewed all studies in the test set. For cases where the reviewers disagreed on labeling, a consensus session was held to decide the classification.

3.5.3 Study III

3.5.3.1 Study population

We used the same data set, and thus the general population, from DS for training and modeling, as in studies I and II. Study II, like Study III, only examined ankle fractures. Additionally, 399 ankle exams of random patients from Flinders Medical Centre in Adelaide, Australia, were included as EVD. As population data did not exist for the DS dataset, the dataset from Flinders was provided anonymized and without any population data or radiologist reports.

3.5.3.2 Images, radiology reports, and labels.

Study III used the same datasets as Study II, but additional studies were included in the training dataset. The Flinders data was labeled similarly to the DS test data, except it was provided without radiologist reports.

3.5.3.3 Training data

The same training data as in Study II was extended with 2664 additional labeled studies. At least two reviewers reviewed all newly included studies: MG (senior consultant orthopedic trauma surgeon), JO (medical doctor), and FW (medical student). As part of active learning, described later, many studies in the original training dataset from Study II were re-reviewed.

3.5.3.4 Internal validation data

Study III used the gold standard derived in Study II as an IVD.

3.5.3.5 External validation data

Three hundred ninety-nine studies were obtained from Flinders. Four orthopedic trauma surgeons (MG, JD, FIJ, and EA) independently classified images according to the AO 2018 standard. Two surgeons classified the entire dataset, and two surgeons classified half of the data each. Once classification was performed independently, a consensus session was held. During the consensus session, disagreements in classification between reviewers were resolved by a majority vote. The result was the EVD.

3.5.4 Study IV

Figure 3 shows the data collection and design of Study IV.

3.5.4.1 Study population

All fractures registered in the SFR by one of the emergency hospitals in the greater Stockholm Region from the start of the SFR (2011-01-01) until 2019-06-30 were eligible for inclusion. The seven hospitals included were Danderyd Hospital, Karolinska University Hospital in Solna and Huddinge, Norrtälje Hospital, S:t Göran Hospital Södersjukhuset Hospital, and Södertälje Hospital. Visby Hospital in Gotland, Sweden, was included as a control or external validation site. Inadvertently, a subset of the patients in the previous Danderyd dataset (studies I-III) were included in this study. However, we could not say which patients were reintroduced as that data was anonymous.

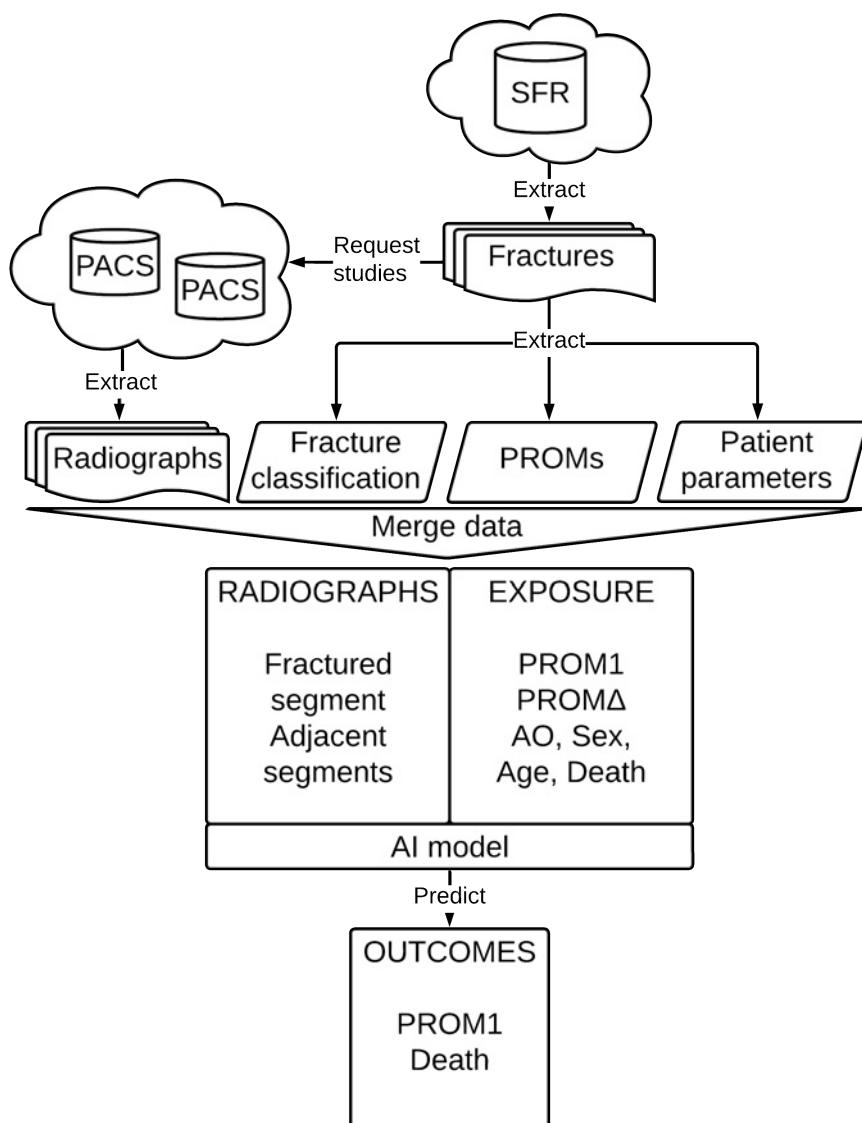


Figure 3. Schematic overview of data sources, retrieval, exposures, and study outcomes. Data was collected on each fracture from the SFR. All radiographs visualizing a fractured and adjacent segments within seven days of trauma were collected from various PACS systems. Variables and imaging studies were merged. 3) The AI model was trained to predict outcomes and individual PROM scores one year after the trauma from the merged data. PACS – Picture Archiving and Communications Systems. PROMs – self-reported patient-reported outcome measures. SFR – Swedish Fracture Registry. PROMΔ – one year change in PROM.

Included were all patients where there was a PROM1 registered or who died within one year of the fracture date. If both were missing, the patient was excluded. In addition, fractures were excluded if there was no initial radiographic imaging of the fracture (i.e., no radiographic imaging within <7 days). Only fractures of the lower extremities were eligible for inclusion, i.e., proximal femur/hip and distal. Pelvic/acetabular fractures were not included. Any patients with any fracture or treatment registered at any Stockholm site *and* Gotland were excluded to ensure no patient overlap between training and test data.

3.5.4.2 Images, radiology reports, and labels.

Imaging was collected independently from the PACS of Region Stockholm and Region Gotland. Studies were matched to the SFR fracture using DICOM metadata. As described above, all studies possibly visualizing each fracture were mapped to the fracture.

Outcomes were collected from the SFR. The **primary outcomes** were PROM1 outcomes. **Secondary outcomes** were fracture segment, AO class (as per the SFR), and the change in PROM. We intended to predict complications (reoperation and infections); however, these were difficult to derive reliably from the SFR. It would require a manual review of all imaging (including MRI and CT) of the fractured area within one year of the trauma for all fractures.

3.5.4.3 Training data and validation data

We split the Stockholm dataset 80/20 between training and training validation data. Due to too few data points, we did not create a local test set (IVD). Instead, we used an EVD with patients from Gotland as test data.

3.5.4.4 Test data – external validation data

As stated, the test data consisted entirely of the Gotland EVD.

3.6 Modelling

3.6.1 Study I

3.6.1.1 Outcomes

The **primary outcomes** were 1) the top-performing neural network on fracture detection and 2) fracture detection accuracy (fracture yes or no) in terms of accuracy. **Secondary outcomes** were prediction accuracy on extremity (hand, foot, ankle), side (left or right), exam view, and previous/old fracture (yes/no). A

secondary outcome was also to assess if transfer learning was a reasonable strategy for training on orthopedic trauma. While ascertaining if transfer learning worked was of primary interest, it was a secondary outcome. We did not extensively compare training from scratch to transfer learning.

3.6.1.2 Modeling and neural networks

Study I compared five network architectures: BVCG reference net (“AlexNet”) ¹⁰¹, VGG 8/16/19-layers ¹⁰⁷, and Network In Network ¹⁵⁹. We used pre-trained networks, i.e., trained for other tasks, and they were then retrained for the fracture prediction task. The idea was that the network had learned a set of primary properties and shapes, which could then be remodeled and adapted for fracture detection. This is called “transfer learning”. Transfer learning helped these networks to manage more with the limited data set we provided.

After training, the best-performing network on the primary task (fracture detection) was selected for final evaluation on the gold standard. Performance was evaluated using top-1 accuracy.

3.6.2 Study II

3.6.2.1 Outcomes

The **primary outcome** was ankle fracture classification according to the AO 2018 ankle fracture classification ⁵⁵ in terms of AUROC. **Secondary outcomes** were fracture detection (yes/no), fibular and tibial fractures (AO 2018 classification), and foot fractures (bone localization). IRR was also a secondary outcome.

3.6.2.2 Modeling and neural networks

We used the ResNet neural network architecture ¹⁴⁹, and training details are described in Table 2.

Table 2. Neural network architecture and training strategies for Study II.

Layer type	Blocks	Kernel size	Filters	Group
ResNet block	1x2	5x5	32	Image
ResNet block	1x2	3x3	64	Image
ResNet block	4x2	3x3	64	Core
ResNet block	2x2	3x3	128	Core
ResNet block	2x2	3x3	256	Core
ResNet block	2x2	3x3	512	Core
Image max	1	-	-	Pool
Convolutional	1	1x1	72	Classification
Fully connected	1	-	4	Classification

Table 2. Neural network architecture and training strategies for Study II.

Fully connected	1	-	4	Classification
Session	Epochs	Internal learning rate	Noise	Teacher-student pseudolabels
Initialization	70	0.025	None	No
Noise	80	0.025	5%	No
Teacher-student	40	0.005	5%	Yes
Regularization	20	0.025	10%	No
SWA	20x5	0.01	5%	No
Overfitting strategy	Description			
Image jittering	Each image was randomly flipped, cropped and rotated during training.			
Random noise	A denoising autoencoder was employed to regularize the visual representation manifold. The encoder and decoder have identical layers and parameters.			
Teacher-student network using alternate data	Semi-supervised training where a co-existing teacher network learned the labels from both the report and image. This allowed us to use the teacher's labels when images had none. As these labels were less certain than the manually labeled images, the teacher label's loss was reduced by 10%. During the teacher-student session the data set was augmented unlabeled exams using a ratio of 1:2. During all sessions we switched between the ankle dataset and a similarly labeled dataset with wrist images that consisted of 17,511 exams. These were also augmented with unlabeled images with the same proportion between unlabeled as labeled in the ankle dataset ¹⁶⁰ .			
Stochastic weight averaging (SWA)	A cosine function was used for decreasing the learning rate. It was reset between each section of training. Once the learning rate leveled off, we trained for 5 series using stochastic weight averaging ¹⁶¹ . In Study III we had five series of 20 epochs.			
Active learning	Poorly performing categories, during training, were actively reviewed. We also added more training examinations to further improve accuracy, more examinations were added to improve those categories. Highest entropy over predictions was used as the sampling strategy for active learning. I.e., we selected the cases that were closest to 50% probability for an outcome and focused on labeling those. This also called uncertainty sampling ¹⁶²⁻¹⁶⁵ .			

Adapted from Olczak et al. 2021 ².

3.6.3 Study III

3.6.3.1 Outcomes

The **primary outcome** was model performance (AUPR and AUROC) on 1) the external validation set and 2) the internal validation set. **Secondary outcomes** were fracture detection (yes/no), fibular and tibial fracture classes (according to AO 2018 classification), foot fractures (bone localization), and IRR.

As we concluded in Study II, the data was imbalanced with many possible outcomes. For that reason, we followed the recommendations of Olczak et al. 2021⁸⁶ and focused on other performance measures (AUPR), which are better suited to imbalanced data. However, we also reported AUROC as it is more widely used.

3.6.3.2 Modeling and neural networks

While the network was not pre-trained, other anatomies and outcomes were included during training. This was done to introduce noise and randomness, hoping the model would be perturbed sufficiently to find a better optimum. However, the network's training data is expanded with other features that can be transferable to the actual task. We have seen this enhance network performance, and it is related to the transfer learning concept in Study I. Unlike Study II, we did not use teacher-student augmentation during training. Other than that, the modeling parameters are described in Table 2.

3.6.4 Study IV

3.6.4.1 Outcomes

The **primary outcome** was model performance at predicting the PROM1 or prediction of death within the study period on the EVD. **Secondary outcomes** were the one-year change in PROM and fracture classification according to the AO classification used in the SFR.

We used the RMSE vs. the SD to evaluate model performance on ordinal and numerical outcomes. The RMSE needed to be lower than the SD for the model to be helpful. If RMSE was greater than or equal to the SD, we could have guessed the most frequent outcome (mode) or the mean, and our expected error would be approximately the SD. If RMSE was equal to, or very close to, the SD, this suggested that the model has learned the mode or mean – i.e., it was considered a sign of overfitting.

For non-ordinal classification tasks, we used accuracy if they were balanced and binary. We used AUROC and AUPR as performance measures for complex and imbalanced classification.

3.6.4.2 *Modeling and neural networks*

Study IV is an experimental study that, like Study I, consisted of multiple experiments to obtain the best model for predicting patient outcomes. We used a ResNet-based model with the same design as in Study III but experimented with many different hyperparameters. The model was trained sequentially on all tasks. There were three types of training tasks: classification (fracture AO class, sex), ordinal scale prediction (most PROM outcomes), and regression tasks (age, PROM indices, VAS score, etc.). Primary outcomes were included in all models, and secondary outcomes depending on the experiment.

For this study, we want to predict PROM1, a patient-centric measure. Therefore, we experimented by including adjacent radiographic imaging within seven days of the trauma. The idea is that we have a patient and not just a fracture. Our initial approach is to look at each study separately, even where there should not be a fracture but pass the same training PROM or death within 1 year. We hope the model will “overfit” and recognize that it is the same patient. Other experiments will include combining all imaging into one patient/training case, as Study II combined all radiographs in one study, whereas Study I looked at individual radiographs. Other experiments will only look at the actual fracture location imaging and ignore other imaging for the patient, except if the patient has multiple fractures.

Classification

During our training, we learned that a model that trains on various outcomes tends to perform better at individual outcomes. The model becomes richer and learns more patterns. Therefore, we included classification tasks as secondary outcomes. Also, classifications were taken from the SFR and not generated by us.

Ordinal classification

Ordinal parameters have an order, but the steps are not necessarily evenly spaced. For example, a scale with “bad,” “neutral,” “good,” and “best.” The distance between “good” and “best” is arbitrary, but “best” is always better. We implemented the rank-consistent ordinal regression (CORN) loss¹³³ as in the coral-pytorch package¹⁶⁶ for ordinal outcomes. Nearly all PROM1 outcomes are

ordinal, except for indices, and we additionally construct the one-year change in PROM, i.e., $\text{PROM}\Delta$, outcomes, which are ordinal or numerical, depending on the parameters they were derived from.

The CORN Loss uses binary classification (comparing two outcomes) to check if the outcome is greater than the previous class. For example, is the outcome ≥ 0 , ≥ 1 , ≥ 2 , etc., where 0, 1, 2, ... are the ordered outcomes?

We also experiment with modifications to the CORN Loss, in line with the Focal Loss introduced by Lin et al. in 2018¹⁶⁷. The Focal Loss was introduced into an object detection scenario with extreme class imbalance. It introduced a *modulating factor* $f(\gamma, p_t) = (1 - p_t)^\gamma$ to the Cross-Entropy Loss, commonly used in object detection. Here, $p_t = p$ if we are looking at the correct ("true") class, and $p_t = 1 - p$ otherwise, where p is the predicted probability of the class. γ is a focusing parameter that downweights easy examples. The more likely a class is, i.e., $p \rightarrow 1$, the smaller $(1 - p_t)^\gamma$ will get, as long as $\gamma \geq 1$. The easier classes will influence the loss and training less.

We experiment with an implementation of the modulating factor with the CORN Loss as

$$\text{Focal CORN Loss} = f(\gamma, p_t) \cdot \text{CORN}(X, y),$$

where γ and p_t are as before, and $\text{CORN}(X, y)$ was defined as in Cao et al¹³³.

Linear regression

We had previously found regression modeling of parameters, like distance and positions in radiographs, difficult with MSE loss. This is likely because CNNs have difficulty retaining spatial information¹⁶⁸. However, we were not looking to model spatial relationships. Therefore, we experimented with alternate loss functions to see if we could train the network to perform better at regression tasks.

For regression outcomes, we experimented with MSE loss and robust general loss¹⁶⁹. The robust general loss was implemented using the `robust_loss_pytorch` package¹⁷⁰. We chose whichever performed best and did not report the other.

Experiments

We conducted several experiments varying:

- Image size

- “Study” vs. “patient” data combinations
- Loss functions
- Optimizers
- Model input (secondary parameters)
- Various regularizers, such as L2 regularization, drop out

Testing and model evaluation

To avoid model selection and presentation bias, we test only the best-performing model from the experiments on the EVD after all experiments are concluded.

4 Results

4.1 Study I

4.1.1 Primary outcome – Fracture detection

We trained five pre-trained neural network models on a dataset of 256,458 radiographs, of which 56% had been labeled as having a fracture (See Table 3).

Table 3. Image and label data.

Table 3a		Table 3b	
Label	n (%) *	Label error	n (%)
Fracture			
No	111,275 (43)	Correctly classified	276 (69)
Yes	143,183 (56)	Misclassified	124 (31)
Missing	2,000 (1)		
Laterality			
Left	120,377 (47)	Correct laterality	52 (26)
Right	132,511 (52)	Misclassified	8 (4)
Missing	3,570 (1)	Marker missing	140 (70)
Exam body part			
Finger	390 (0.2)	Correct body part	17
Thumb	76 (0)	Related body part	51
Scaphoid	27,962 (11)	Unrelated body part	15
Hand	5,614 (2)	Invalid image	3
Wrist	65,264 (25)		
Ankle	98,002 (38)		
Missing	59,150 (23)		
Exam view			
Distal	7,136 (3)	Correct view	110 (55)
AP	55,916 (22)	Misclassified	90 (45)
Oblique	44,962 (18)	Unrelated view	12 (6)
Proximal	6,776 (3)	Closely related view	78 (39)
Radial	6,946 (3)	Ankle: mix-up between AP and mortise	22 (11)
Lateral	67,465 (26)	Ankle: mix-up between oblique and lateral	23 (12)
Ulnar	7,014 (3)	Scaphoid: mix-up between supination and pronation	14 (7)
Missing	60,243 (24)	Scaphoid: mix-up between distal and proximal	7 (4)
		Miscellaneous	12 (6)

3a shows raw image and label data and 3b the results of the manual review of classifications and labels. These were labels from the training set. Olczak et al. 2017 ¹.

* 70% were reserved for training, 20% for validation, and 10% for testing.

56% of images were labeled as having a fracture, and 43% as not having one. Only 1% of radiographs could not be labeled. Ankles were the most studied body part (38%), followed by wrists (25%). 23% of images were missing information about the body part, and 24% lacked information on the exam view.

The VGG-16 model performed best in training validation for the primary outcome, and we selected it as our evaluation model. VGG-19 was a close second, and the differences in performance for the two models were minimal. All networks performed excellently for the exam body part and similarly for the exam view. Laterality had the most significant spread between the networks. See Figure 4.

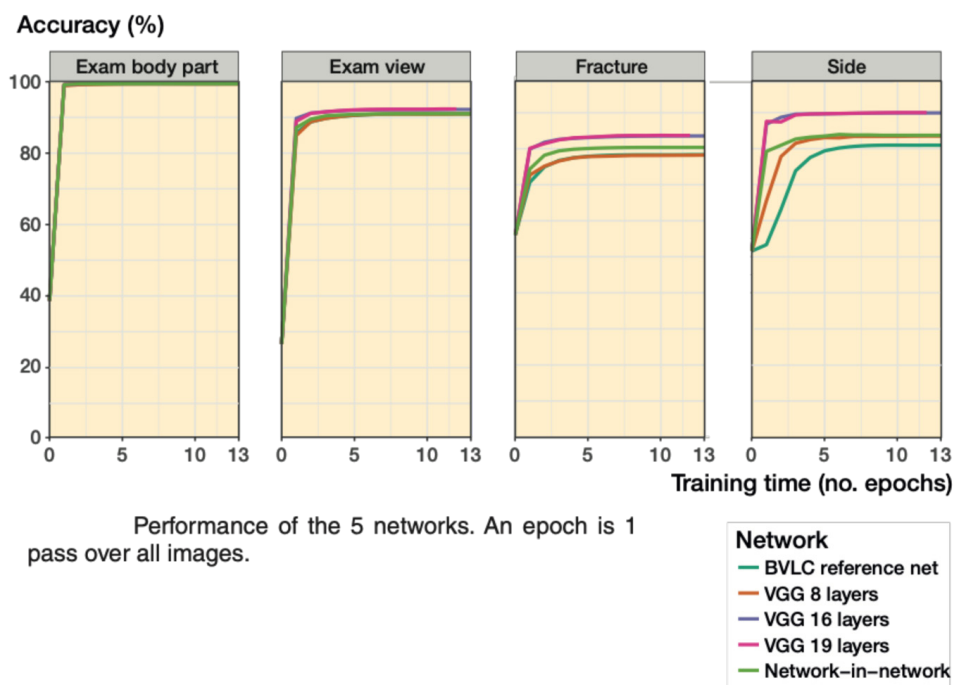


Figure 4. Performance of the five networks during validation. The best performer at fracture detection, VGG-16, was selected for further analysis. Image from Olczak et al. 2017¹.

VGG-16 had a fracture detection accuracy of 83% (95%CI 80–87%) on the gold standard. This was on par with the accuracy of human reviewers, who were 82% (95%CI 78–86) accurate for reviewer 1 and 82% (95%CI 78–85) for reviewer 2 (see Table 4).

Table 4. Outcomes compared between observers.

Observer	Label ^a	Network ^b	Reviewer 1	Reviewer 2	Gold standard
Label ^a	–	80 (0.6)	76 (0.5)	74 (0.5)	83 (0.7) 79–87
Network ^b	80 (0.6)	–	84 (0.7)	86 (0.7)	83 (0.7) 80–87
Reviewer 1	76 (0.5)	84 (0.7)	–	90 (0.8)	82 (0.6) 78–86
Reviewer 2	74 (0.5)	86 (0.7)	90 (0.8)	–	82 (0.6) 78–85
Gold standard	83 (0.7)	83 (0.7)	82 (0.6)	82 (0.6)	–

Performance is the % of outcomes where both observers agree reported as accuracy % (kappa) 95% CI. Olczak et al. 2017 ¹.

^a Four of the radiographs were missing and were excluded from the analysis for this category.

^b VGG-16, the best performing network during training and validation.

4.1.2 Secondary outcomes

The best model's performance was impressive for the secondary outcomes. The accuracy of identifying the exam body part was near perfect, 100%. The accuracy of determining the exam view was >95%, and for identifying the laterality, it was 90%. These results underscore the reliability and robustness of our model. A subsection of misclassifications and images was manually studied for causes of error. For fracture misclassification, the study was often labeled as a fracture, but a fracture was not visible in that view. In 69% of cases, the model correctly classified the radiographs, while the label was incorrect. For the exam views, the view was frequently mistaken for a similar view. It was also clear that, for example, scaphoid images were often taken in non-standard views. See Table 3b for details.

4.2 Study II

Out of 5495 radiographic ankle exams, 400 patients (409 exams) were assigned to the test set. The remaining 5086 examinations were used for model training and validation. No patients were present in the test and training sets. As studies had been vetted in Study I, none were excluded now. See Figure 5.

Table 5. Case distribution in training and test set.

	Train (n=4,941)			Test (n=409)		
	Yes	Maybe	No	Yes	Maybe	No
Fracture	2,156 (44)	121 (2)	2,664 (54)	306 (75)	13 (3)	90 (22)
Malleolar (44)	1,696 (34)	63 (1)	3,182 (64)	210 (51)	6 (1)	193 (47)
Tibia distal (43)	254 (5)	6 (0)	4,681 (95)	63 (15)	2 (0)	344 (84)
Fibula (4F2–3)	129 (3)	3 (0)	4,809 (97)	37 (9)	0 (0)	372 (91)
Tibia diaphyseal (42)	88 (2)	0 (0)	4,853 (98)	27 (7)	0 (0)	382 (93)
Other bone	210 (4)	47 (1)	4,684 (95)	35 (9)	5 (1)	369 (90)

"Other bone" generally indicates a visible fracture of the foot. It was possible for an examination to have multiple fracture labels. Percentages of dataset in parenthesis. Table from Olczak et al. 2021 ².

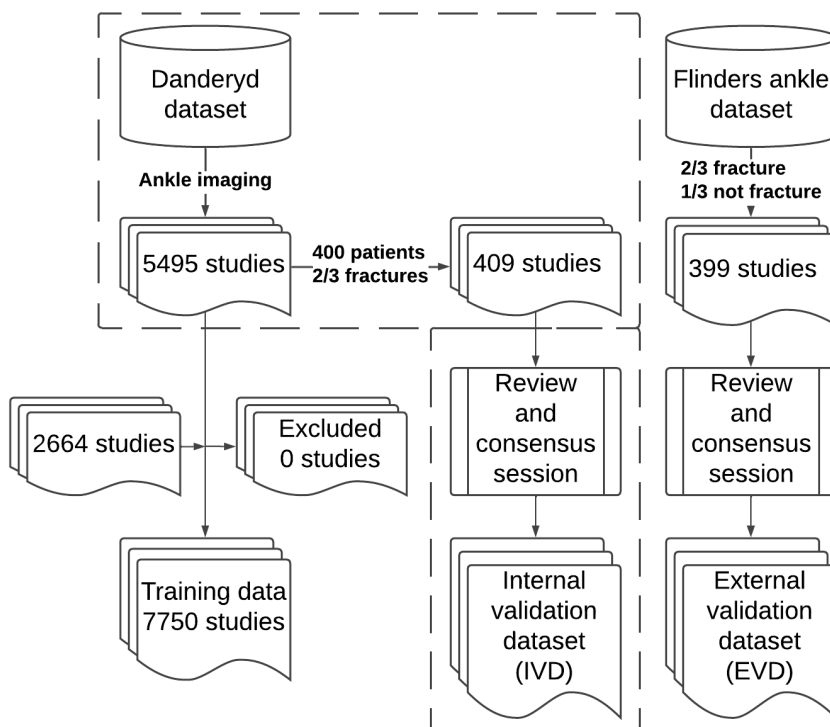


Figure 5. Combined Study II and III flowchart. The dashed line demarcates the data used in Study II. The data outside the demarcations are the extensions made for Study III. Image adapted from Olczak et al. 2024 ³.

The training set had 44% fractures and 54% without, whereas the test set had the desired distribution of 75% fractures. However, only 210 (51%) were malleolar fractures; the rest were fractures of the tibia, fibula, and foot bones. See Table 5. The distribution of fractures according to the AO 2018 ankle classification for the training and test sets is displayed in Figure 6 and Table 6. All types of ankle fractures were represented in the training set except for A3.2 and only one A2.2 in the entire data. Type B fractures were twice as many as type A fractures, and the training set had twice as many type C fractures as type B fractures. The test set had mostly type B fractures, with more type C than type A fractures.

The network could detect a malleolar fracture with an $AUC_{\text{malleolar}}$ 0.92 (0.89–0.95). However, the weighted mean AUC (wAUC) was $wAUC_{\text{malleolar}}$ 0.90, with $wAUC_A$ 0.84, $wAUC_B$ 0.90, and $wAUC_C$ 0.87. These, along with individual AO outcomes, are presented in Table 7. We only reported outcomes (32 out of 39) with ≥ 2 cases, the minimum for computing confidence intervals.

Table 6. Distribution of AO outcomes in the malleolar fracture data.

AO type	Train (n=4,941)	Test (n=409)
44A (483 train & 31 test cases)		
A1.1	78 (22)	6
A1.2	165 (46)	7
A1.3	114 (32)	9
A2.1	105 (93)	5
A2.2	1 (1)	-
A2.3	7 (6)	2
A3.1	11	-
A3.3	2	2
44B (1,015 train & 136 test cases)		
B1.1	385 (74)	39
B1.2	132 (25)	26
B1.3	6 (1)	2
B2.1	99 (44)	20
B2.2	105 (47)	16
B2.3	19 (9)	2
B3.1	76 (28)	12
B3.2	152 (57)	13
B3.3	41 (15)	6
44C (255 train & 47 test cases)		
C1.1	85 (67)	17
C1.2	20 (16)	5
C1.3	22 (17)	2
C2.1	30	6
C2.2	21	3
C2.3	39	9
C3.1	10	3
C3.2	9	1
C3.3	19	1

% of the AO group is reported in parenthesis after the count, if there were more than 100 cases in the group. Olczak et al. 2021².

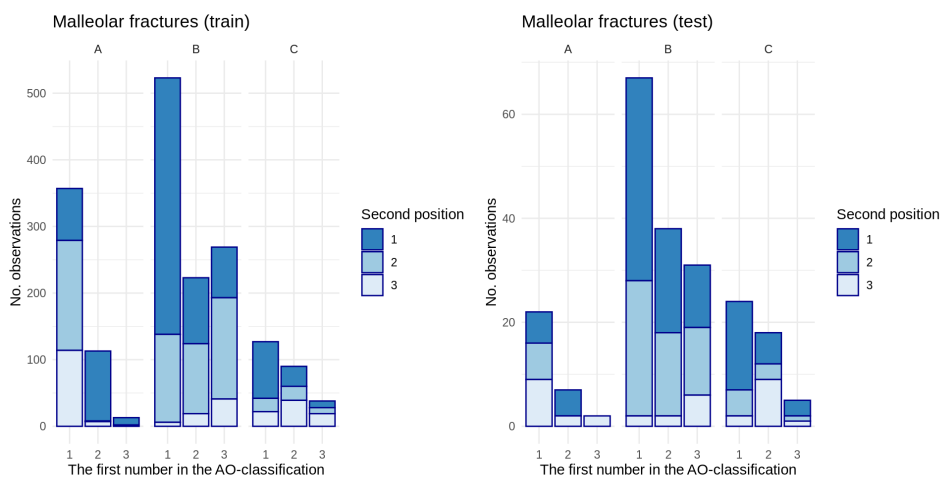


Figure 6. Distribution of AO classes in the malleolar fracture data. Diagram from Olczak et al. 2021².

Figure 7 and Figure 8 illustrate examples of classification errors the model performs.



Figure 7. Type A fractures the network incorrectly classified as a type C fracture. Image from Olczak et al. 2021².

Table 7. Modeling outcomes for AO ankle (44) fractures in the test set.

AO	Cases (n=409)	Sensitivity (%)	Specificity (%)	Youden's J ^a	AUC (95% CI)
Malleolar	216	86	90	0.76	0.92 (0.89–0.95)
44A	32	73	81	0.54	0.81 (0.72–0.88)
1	22	88	75	0.63	0.87 (0.77–0.94)
1.1	6	75	93	0.68	0.87 (0.70–0.98)
2.1	7	80	83	0.63	0.79 (0.54–0.94)
3.1	9	75	88	0.63	0.84 (0.70–0.95)
2	7	100	74	0.74	0.91 (0.83–0.97)
2.1	5	100	74	0.74	0.89 (0.80–0.97)
3	2	100	86	0.86	0.90 (0.83–0.96)
44B	137	89	88	0.77	0.93 (0.90–0.95)
1	67	90	88	0.77	0.93 (0.88–0.96)
1.1	39	87	84	0.71	0.89 (0.85–0.93)
2.1	26	92	85	0.77	0.90 (0.81–0.96)
2	38	82	84	0.65	0.87 (0.80–0.92)
2.1	20	100	72	0.72	0.87 (0.83–0.92)
2.2	16	88	74	0.62	0.82 (0.68–0.91)
2.3	2	100	98	0.98	0.99 (0.97–1.00)
3	32	78	90	0.68	0.90 (0.85–0.94)
3.1	12	83	75	0.58	0.79 (0.63–0.90)
3.2	13	92	82	0.74	0.91 (0.84–0.96)
3.3	6	100	91	0.91	0.96 (0.93–0.98)
44C	47	74	90	0.65	0.86 (0.79–0.92)
1	24	75	79	0.54	0.83 (0.72–0.91)
1.1	17	76	85	0.61	0.86 (0.74–0.94)
1.2	5	80	92	0.72	0.89 (0.77–0.97)
1.3	2	100	88	0.88	0.92 (0.86–0.97)
2	18	100	72	0.72	0.91 (0.86–0.95)
2.1	6	83	93	0.76	0.91 (0.79–0.98)
2.2	3	100	88	0.88	0.96 (0.88–1.00)
2.3	9	100	77	0.77	0.88 (0.84–0.92)
3	5	100	88	0.88	0.95 (0.90–0.98)
Weighted mean AUC					
A					0.84
B					0.90
C					0.87
Malleolar ^b					0.90

^a Criterion based on Youden's Index ^{171–174} defined as $YI(c) = \max_c(Se(c) + Sp(c) - 1)$. This is maximizes the sum of Sensitivity and Specificity ^{175,176} and to the criterion that maximizes concordance, which is a monotone function of the AUC. Adapted from Olczak et al. 2021 ².

^b Weighted mean of malleolar classes in the table.



Figure 8. The fracture is a malleolar type C fracture. The network predicted a type B fracture. Image from Olczak et al. 2021².

4.3 Study III

Study III initially contained the same data distributions as Study II (see Figure 5 and Table 5). We added 2,664 training cases for 7,750 training and validation cases, focusing on type A fractures, for active training but do not report the resulting training distribution.

There were considerable differences between the IVD and EVD. The EVD had three projections, whereas the IVD had ≥ 4 . The EVD was focused on lateral malleolus fractures with a higher proportion of type A fractures (94 out of 274 malleolar fractures in the EVD vs 32 out of 216 in the IVD). The EVD included one-week follow-ups and weight-bearing studies, which are not immediate studies at the ER, as all were in the IVD. The exclusion criteria for the EVD were images or views of poor quality and severely displaced fractures, whereas none had been excluded from the training data or IVD. This amounted to less severe fractures in the EVD, i.e., a higher proportion of type A1 and B1 fractures. See Table 8 for details.

Table 8. Properties of the IVD and EVD.

Dataset properties	IVD		EVD	
Cases	409		399	
Projections	≥4		3	
Focus	Ankle study		Lateral malleolar fracture	
Timing	Initial imaging		Initial imaging, one-week follow-up, weight-bearing	
Implants & casts	Yes		No	
Open physes	No		No	
Excluded on imaging quality	None		Insufficient quality views Poor quality images Severely displaced fractures	
Fracture	Cases	Percent (%)	Cases	Percent (%)
Base	253	61,9%	277	69,4%
Malleolar	216	52,8%	274	68,7%
Fibula *	37	9,0%	3	0,8%
Previous fracture/other*	134	32,8%	15	3,8%
Foot*	57	13,9%	2	0,5%

Numbers are based on ground truth labelling by reviewers after the consensus session. The IVD is the internal validation dataset and the EVD the external validation dataset. Table from Olczak et al. 2024 ³.

* Denotes fractures and outcomes that were flagged as fractures during study selection but are not malleolar fractures but secondary outcomes.

4.3.1 Flinders data (EVD)

Type A fractures were the second most numerous, and all but one (dropped) were type A1. Only for type A1.1 did performance not exceed chance, but the decrease was not statistically significant. Type B fractures performed well, but three cases did not perform better than a random classifier. Type C fractures were either C1.1 or C2.1, and the classifier performed well on both. Figure 9 shows an incorrectly classified type B fracture with network activation. Figure 11 illustrates examples of type A fractures where the network classified them incorrectly, where one is shown in Figure 10 with an activation map.

An AUROC of 0.83 is good for fracture detection, and an AUPR of 0.93 is excellent. The change in wAUC for the EVD after active learning was +0.06 to wAUC 0.83, and wAUPR was +0.07 to 0.64. Twenty-one outcomes were represented in the EVD, and 17 were statistically significantly better than chance. Type C fractures decreased performance (Δ AUPR -0.06), resulting from the network losing understanding of type C2.1 fractures (Δ AUPR -0.38). The model could perform better than random for all outcomes with >5 test cases. For classes with fewer cases, except B2.2, the 95% CI could not be bounded to

indicate that the performance was significantly better than chance. Table 9 shows model performance and improvement on the EVD.

Table 9. Flinders external validation dataset (EVD) performance. 399 Cases.

AO	Cases	AUC (95% CI)	Δ AUC	AUPR (95% CI)	Δ AUPR
Malleolar	274	0.86 (0.82–0.89)	0.03	0.93 (0.91–0.96) *	0.00
44A	94	0.74 (0.68–0.80)	0.12	0.52 (0.43–0.61) *	0.20
1	93	0.75 (0.69–0.81)	0.14	0.57 (0.46–0.64) *	0.25
1.1	5	0.63 (0.33–0.94)	–0.07	0.04 (0.00–0.16)	0.02
1.2	28	0.78 (0.69–0.87)	0.15	0.26 (0.11–0.39) *	0.14
1.3	60	0.68 (0.61–0.76)	0.08	0.30 (0.21–0.42) *	0.10
44B	142	0.90 (0.87–0.93)	0.03	0.84 (0.78–0.88) *	0.03
1	116	0.84 (0.80–0.88)	0.03	0.68 (0.60–0.76) *	0.07
1.1	87	0.80 (0.75–0.85)	0.05	0.47 (0.37–0.57) *	0.06
1.2	27	0.80 (0.72–0.88)	0.02	0.19 (0.11–0.31) *	0.03
1.3	2	0.60 (0.17–1.02)	–0.30	0.01 (0.00–0.02)	–0.01
2	21	0.85 (0.75–0.94)	0.10	0.32 (0.18–0.49) *	0.19
2.1	18	0.85 (0.75–0.95)	0.12	0.33 (0.16–0.56) *	0.24
2.2	3	0.93 (0.88–0.99)	0.00	0.05 (0.01–0.16) *	–0.03
3	5	0.82 (0.61–1.04)	–0.06	0.19 (0.01–0.47)	0.11
3.1	5	0.82 (0.63–1.02)	–0.05	0.12 (0.01–0.30)	0.07
44C	38	0.89 (0.82–0.96)	0.04	0.63 (0.45–0.77) *	–0.06
1	28	0.90 (0.84–0.96)	0.08	0.42 (0.26–0.62) *	0.07
1.1	27	0.90 (0.84–0.97)	0.07	0.44 (0.22–0.64) *	0.10
2	9	0.92 (0.82–1.01)	–0.04	0.19 (0.04–0.37) *	–0.40
2.1	9	0.90 (0.79–1.02)	–0.04	0.16 (0.05–0.31) *	–0.38
		Weighted mean		Weighted mean	
		AUC	Δ	AUPR	Δ
		0.83	+0.06	0.64	+0.07

Performance reported with the area under the receiver operating characteristic curve (AUC) and the area under the precision–recall curve (AUPR). 95% confidence intervals (CI) were computed using bootstrapping. Outcomes with ≤ 1 instance were not reported. Δ AUC and Δ AUPR was the difference in AUC and AUPR comparing the actively trained network to the pre-active training network. Δ Implies a change or difference. Table modified from Olczak et al. 2024 ³.

* Indicates that the AUPR with 95% CI exceeded random AUPR.

4.3.2 Danderyd (IVD)

Performance was adequate for type A fractures, but only four classes were shown to be significant. See Table 10 for model performance on the IVD after active learning. Type B and type C fractures were overall better than chance.

The network had an AUC of 0.95 and an AUPR of 0.96, both excellent. The wAUC improved by +0.04 to 0.93, and the wAUPR improved by +0.08 to 0.65. wAUC differed by 0.10 between the EVD and IVD (0.83 vs. 0.93), but the wAUPR for the

EVD and IVD were similar (0.64 vs. 0.65). Thirty-four outcomes and 26 were statistically significantly better than chance, except for one type B outcome. Once again, all performance seemed substantially better than chance, but the lack of cases made bounding the errors difficult.

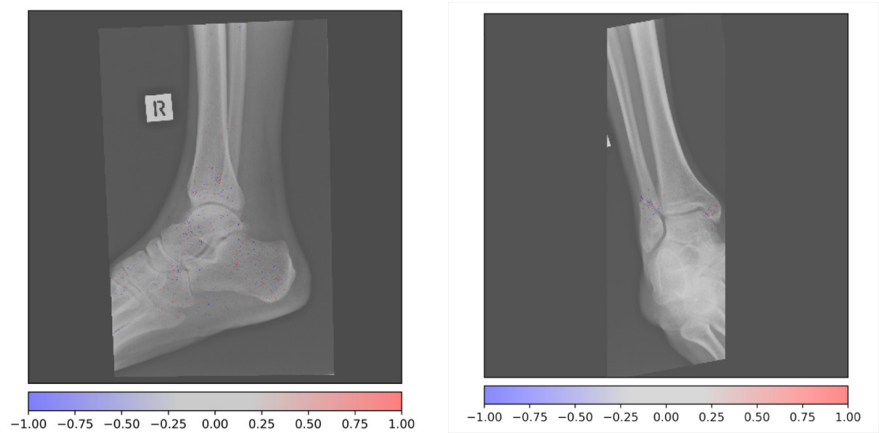


Figure 9. Activation heatmap of a type 44B1.2 fracture, incorrectly classified as a type C fracture. The activations show what the model reacts to when classifying fractures. Study from the external validation dataset and Olczak et al. 2024 ³.

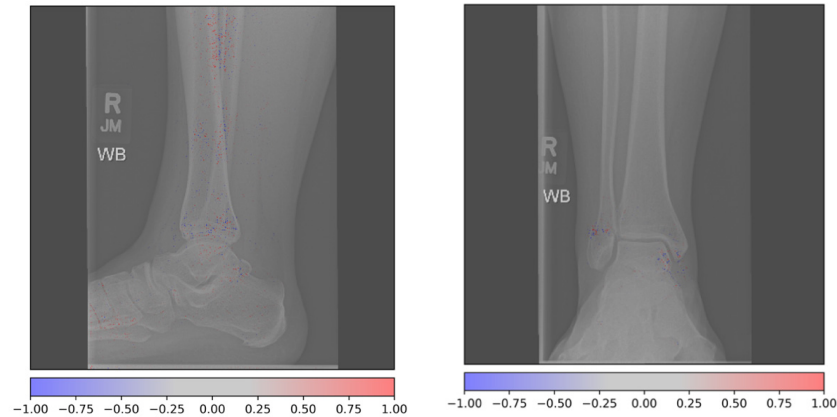


Figure 10. Activation heatmaps where a type 44A1.3 fracture is incorrectly classified as a type B fracture. The activations show what the model reacts to in the radiograph. Study from the external validation data and Olczak et al. 2024 ³.



Row no.	Options (%)			Max	True categories
	A	B	C		
1	1	86	21	B	A1.3
2	2	5	89	C	A1.3
3	2	2	37	C	A1.3
4	3	4	5	C	A1.3

Figure 11. Incorrectly classified cases where the network failed to detect Type A, sorted from lowest probability to highest. Studies from taken from the EVD. Table and images from Olczak et al. 2024 ³.

Table 10. Danderyd internal validation dataset (IVD) performance. 409 cases.

	Cases	AUC (95% CI)	Δ AUC	AUPR (95% CI)	Δ AUPR
Fracture	216	0.95 (0.94–0.97)	0.03	0.96 (0.94–0.97) *	0.03
44A	32	0.84 (0.76–0.92)	0.04	0.46 (0.11–0.61) *	0.23
1	22	0.84 (0.76–0.92)	–0.03	0.37 (0.15–0.56) *	0.19
1.1	6	0.88 (0.79–0.97)	–0.01	0.04 (0.01–0.10)	0.00
1.2	7	0.84 (0.69–1.00)	–0.02	0.30 (0.01–0.61)	0.22
1.3	9	0.82 (0.69–0.96)	0.03	0.18 (0.01–0.45)	0.11
2	7	0.99 (0.97–1.00)	0.15	0.52 (0.04–0.75)	0.28
2.1	5	0.99 (0.97–1.00)	0.09	0.41 (0.00–0.65)	0.15
2.3	2	0.99 (0.99–1.00)	0.14	0.25 (0.00–0.50)	0.23
3	2	0.95 (0.86–1.04)	–0.02	0.08 (0.03–0.17) *	0.01
44B	137	0.96 (0.93–0.92)	0.04	0.92 (0.88–0.95) *	0.05
1	67	0.95 (0.93–0.98)	0.05	0.77 (0.67–0.86) *	0.14
1.1	39	0.90 (0.87–0.94)	0.07	0.37 (0.25–0.51) *	0.06
1.2	26	0.94 (0.91–0.97)	0.07	0.40 (0.22–0.60) *	0.15
1.3	2	0.96 (0.90–1.02)	0.04	0.06 (0.01–0.23) *	0.03
2	38	0.86 (0.80–0.92)	0.01	0.40 (0.25–0.56) *	0.04
2.1	20	0.91 (0.85–0.97)	0.05	0.37 (0.20–0.55) *	0.14
2.2	16	0.88 (0.77–1.00)	–0.01	0.35 (0.15–0.53) *	0.13
2.3	2	0.87 (0.68–1.07)	–0.05	0.03 (0.00–0.11) *	0.00
3	32	0.92 (0.89–0.96)	0.06	0.50 (0.27–0.59) *	0.03
3.1	12	0.90 (0.83–0.97)	0.04	0.18 (0.06–0.34) *	0.02
3.2	13	0.92 (0.88–0.96)	0.08	0.20 (0.08–0.35) *	–0.04
3.3	6	0.96 (0.93–0.99)	0.02	0.16 (0.03–0.30) *	0.06
44C	47	0.93 (0.89–0.97)	0.05	0.73 (0.61–0.82) *	0.20
1	24	0.90 (0.84–0.97)	0.05	0.42 (0.27–0.63) *	0.18
1.1	17	0.93 (0.87–0.99)	0.03	0.39 (0.21–0.60) *	0.16
1.2	5	0.86 (0.75–0.97)	–0.01	0.05 (0.01–0.12)	0.01
1.3	2	0.93 (0.83–1.02)	0.02	0.04 (0.01–0.14) *	0.02
2	18	0.93 (0.90–0.97)	–0.02	0.40 (0.16–0.58) *	–0.05
2.1	6	0.86 (0.74–0.99)	–0.08	0.22 (0.01–0.51)	0.07
2.2	3	0.99 (0.99–1.00)	0.08	0.32 (0.00–0.62)	0.28
2.3	9	0.92 (0.88–0.96)	0.03	0.11 (0.04–0.21) *	0.00
3	5	0.98 (0.97–1.00)	0.07	0.29 (0.02–0.67) *	0.21
3.1	3	0.96 (0.90–1.03)	0.29	0.16 (0.00–0.50)	0.15
		Weighted mean AUC	Δ	Weighted mean AUPR	Δ
		0.93	+0.04	0.65	+0.08

Performance reported with the area under the receiver operating characteristic curve (AUC) and the area under the precision–recall curve (AUPR). 95% confidence intervals (CI) were computed using bootstrapping. Outcomes with ≤ 1 instance were not reported. Δ AUC and Δ AUPR was the difference in AUC and AUPR comparing the actively trained network to the pre-active training network. Δ Implies difference. Table modified from Olczak et al. 2024 ³. * Indicates that the AUPR with 95% CI exceeded random AUPR.

4.4 Study IV

The SFR had 41,043 unique fractures from Stockholm and Gotland during the study period and 41,004 after excluding patients that overlapped both regions. After exclusion and inclusion, we had 297 fractures (275 patients) in the EVD (Gotland in Study IV) and 6,161 fractures (5,430 patients) remaining from the seven clinics in the Stockholm training data. No fractures were excluded due to missing imaging. See Figure 12 for a flowchart and Table 11 for more details.

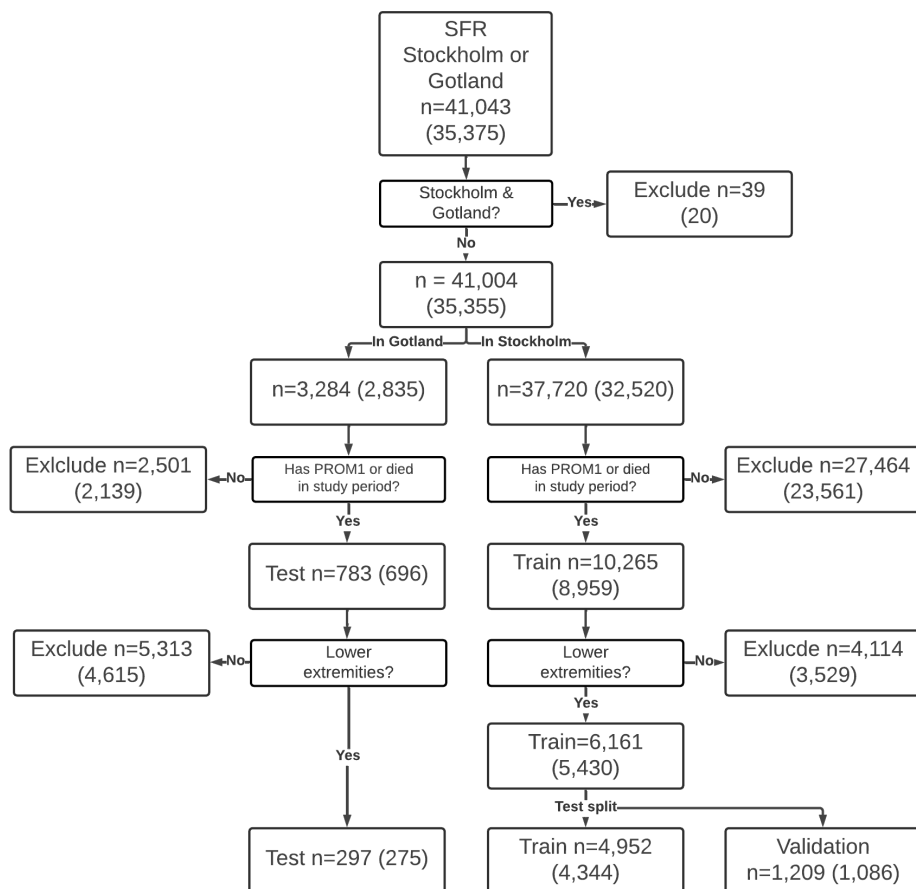


Figure 12. Data flowchart of Study IV, with the number of unique fractures reported. Numbers in parenthesis is the number of unique patients.

Table 11. Population statistics for fractures in the training and test sets.

Parameter	Train (Stockholm)	Test (Gotland)
Unique patients	4344	275
Unique fractures	4952	297
Died in study period	1591	89
PROMO *	3658	128
PROMI *	3367	208
Gender female/male (%)	64.5% / 34.5%	61.3% / 35.7%
Age (min: mean±sd: max; median; mode)	14: 71.2±18.9: 108; 75; 89	18: 70.3±17.8: 102; 72; 90
Number of AO classes	149	75

* Not necessarily complete PROM.

Table 12 lists training parameters derived from the SFR for classification outcomes and regression variables, and Table 13 gives the same for ordinal variables. Both tables show the training data from Stockholm.

Table 12. Classification and regression outcomes in the Stockholm data.

CLASSIFICATION	N _{fractures}	N _{outcomes}	Mode			
Died in study period	1,591	2	No			
Injury sex	4,952	2	Female			
Body part *	4,952	6	–			
Segment *	4,952	17	–			
AO Class *	4,952	149	–			
REGRESSION VARIABLES	N	Mean	SD	Median	Min	Max
PROM1 EQ5D index	2781	0.68	0.31	0.73	–0.6	1.00
PROM1 Daily activity index	3306	28.86	31.18	15.00	0	100
PROM1 Emotional index	3296	31.07	21.83	28.57	0	96.4
PROM1 Arm-hand function index	3302	11.83	21.15	0.00	0	100
PROM1 Mobility index	3303	27.80	24.79	22.22	0	100
PROM1 Function index	3304	24.86	23.09	17.65	0	100
PROM1 Bother index	3141	23.26	22.52	16.67	0	100

For scores, a higher number means more decrease in function or more problems, except for EQ5D Index where the reverse is true. N is the count of the parameter.

* Secondary outcome. Primary and secondary outcomes are treated identically by the network.

Table 13. Ordinal outcomes in the Stockholm training data.

ORDINAL VARIABLES	N	Mean	SD	Median	Min	Max	Mode
PROMO Recovery expected	1747	1.51	0.96	1	1	5	1
PROMO Smoker	3367	1.71	0.87	2	1	4	1
PROMI EQ5DAnxiety	2832	1.39	0.57	1	1	3	1
PROMI EQ5DPain	2825	1.75	0.56	2	1	3	2
PROMI EQ5DUusualAct	2816	1.45	0.67	1	1	3	1
PROMI EQ5DSelfCare	2837	1.23	0.53	1	1	3	1
PROMI EQ5DMobility	2827	1.55	0.55	2	1	3	2
PROMI EQ5DVAS *	3296	2.01	1.13	2	1	5	1

Table 13. Ordinal outcomes in the Stockholm training data.

ORDINAL VARIABLES	N	Mean	SD	Median	Min	Max	Mode
PROM1 DifficChair	3289	1.62	1.08	1	1	5	1
PROM1 DifficOpenMedBottle	3289	1.62	1.08	1	1	5	1
PROM1 DifficShop	3272	1.85	1.34	1	1	5	1
PROM1 DifficStairs	3287	2.30	1.28	2	1	5	1
PROM1 DifficTightFist	3273	1.35	0.80	1	1	5	1
PROM1 DifficShower	3284	1.86	1.20	1	1	5	1
PROM1 DifficComfortSleep	3290	1.68	0.94	1	1	5	1
PROM1 DifficBendKneelDown	3291	2.61	1.40	2	1	5	1
PROM1	3294	1.52	1.00	1	1	5	1
DifficUseButtonsZippers							
PROM1 DifficCutFingernails	3290	1.56	1.14	1	1	5	1
PROM1 DifficDressYourself	3282	1.56	0.98	1	1	5	1
PROM1 DifficWalk	3280	2.10	1.15	2	1	5	1
PROM1 DifficGetMoving	3267	2.06	1.02	2	1	5	1
PROM1 DifficGoOutYourself	3282	1.82	1.38	1	1	5	1
PROM1 DifficDriveCar	3248	2.03	1.53	1	1	5	1
PROM1	3291	1.38	0.91	1	1	5	1
DifficCleanAfterBathroom							
PROM1 DifficUseHandle	3290	1.36	0.87	1	1	5	1
PROM1 DifficWriteType	3298	1.42	0.96	1	1	5	1
PROM1 DifficTurning	3294	1.59	0.97	1	1	5	1
PROM1	3282	2.62	1.42	2	1	5	1
DifficPhysRecreaActivity							
PROM1 DifficUsualLeisureAct	3283	2.00	1.26	1	1	5	1
PROM1 DifficSexAct	3032	2.18	1.61	1	1	5	1
PROM1 DifficLightHousework	3300	1.76	1.23	1	1	5	1
PROM1	3294	2.52	1.55	2	1	5	1
DifficHeavyHousework							
PROM1 DifficUsualWork	3288	2.11	1.41	1	1	5	1
PROM1 OftenLimp	3253	2.64	1.45	2	1	5	1
PROM1	3274	2.17	1.24	2	1	5	1
OftenAvoidUsingPainful							
PROM1 OftenLegLock	3261	1.80	1.03	1	1	5	1
PROM1	3269	1.92	1.07	2	1	5	1
OftenProblConcentration							
PROM1 OftenOverworkAffect	3260	2.57	1.29	2	1	5	1
PROM1 OftenActIrritable	3283	1.98	0.96	2	1	5	1
PROM1 OftenTired	3290	2.90	1.11	3	1	5	3
PROM1 OftenFeelDisabled	3283	2.55	1.41	2	1	5	1
PROM1	3293	2.45	1.27	2	1	5	1
OftenAngryFrustrated							
PROM1	3292	2.39	1.19	2	1	5	2
BotherUseHandArmLeg							
PROM1 BotherUseBack	3266	1.72	1.06	1	1	5	1
PROM1 BotherWorkHome	3283	1.98	1.21	1	1	5	1
PROM1 BotherPersonalCare	3293	1.70	1.12	1	1	5	1
PROM1 BotherSleepRest	3293	1.76	1.03	1	1	5	1

Table 13. Ordinal outcomes in the Stockholm training data.

ORDINAL VARIABLES	N	Mean	SD	Median	Min	Max	Mode
PROM1	3274	2.56	1.39	2	1	5	1
BotherLeisureRecreAct							
PROM1 BotherFriendsFamily	3286	1.41	0.86	1	1	5	1
PROM1 BotherThinkConcRem	3293	1.70	1.01	1	1	5	1
PROM1 BotherAdjustCope	3287	1.94	1.11	2	1	5	1
PROM1 BotherUsualWork	3276	1.97	1.27	1	1	5	1
PROM1 BotherFeelDepend	3297	1.82	1.21	1	1	5	1
PROM1 BotherStiffPain	3286	2.41	1.18	2	1	5	2
PROM1 Recov	2516	2.56	1.36	2	1	5	2
PROM1 Reoperated	3240	0.18	0.38	0	0	1	0

For scores, a higher number means more decrease in function or more problems, EQ5D VAS where the reverse is true. N is the count of the parameter. The mode is the most common (highest frequency) value.

* VAS can be treated as numerical or ordinal variable, and some sources argue that it acts more like an ordinal than continuous variable ¹⁷⁷.

Table 14 and Table 15 lists parameters of the Gotland data for comparison.

Table 14. Classification and regression outcomes in the Gotland set.

CLASSIFICATION	N _{fractures}	N _{outcomes}	Mode			
Died in study period	89	2	No			
Injury sex	297	2	Female			
Body part *	297	6	–			
Segment *	297	17	–			
AO Class *	297	75	–			
REGRESSION VARIABLES	N	Mean	SD	Median	Min	Max
PROM1 EQ5D index	158	0.72	0.29	0.80	−0.17	1.0
PROM1 Daily activity index	207	22.05	29.49	8.33	0	100
PROM1 Emotional index	207	25.98	21.12	21.43	0	82.14
PROM1 Arm-hand function index	205	10.42	19.86	0.00	0	96.875
PROM1 Mobility index	205	22.10	24.14	13.89	0	100.0
PROM1 Function index	205	19.93	22.17	12.50	0	95.59
PROM1 Bother index	196	17.79	20.54	8.33	0	87.5

For scores, a higher number means more decrease in function or more problems, except for EQ5D Index where the reverse is true. N is the count of the parameter. The mode is the most common (highest frequency) value.

* Secondary outcome. Primary and secondary outcomes are treated identically by the network.

Table 15. Ordinal outcomes in the Gotland set.

ORDINAL VARIABLES	N	Mean	SD	Median	Min	Max	Mode
PROMO Recovery expected	120	1.53	0.93	1	1	5	1
PROMO Smoker	123	1.71	0.80	2	1	4	2
PROM1 EQ5DAnxiety	161	1.27	0.47	1	1	3	1
PROM1 EQ5DPain	162	1.70	0.57	2	1	3	2
PROM1 EQ5DUusualAct	162	1.33	0.59	1	1	3	1

Table 15. Ordinal outcomes in the Gotland set.

ORDINAL VARIABLES	N	Mean	SD	Median	Min	Max	Mode
PROM1 EQ5DSelfCare	163	1.20	0.51	1	1	3	1
PROM1 EQ5DMobility	161	1.47	0.55	1	1	3	1
PROM1 EQ5DVAS *	180	76.56	19.02	80	10	100	90
PROM1 DifficChair	203	1.80	1.06	1	1	5	1
PROM1 DifficOpenMedBottle	205	1.60	1.08	1	1	5	1
PROM1 DifficShop	205	1.60	1.18	1	1	5	1
PROM1 DifficStairs	205	1.97	1.21	2	1	5	1
PROM1 DifficTightFist	204	1.42	0.86	1	1	5	1
PROM1 DifficShower	205	1.61	1.04	1	1	5	1
PROM1 DifficComfortSleep	205	1.56	0.85	1	1	5	1
PROM1 DifficBendKneelDown	205	2.33	1.40	2	1	5	1
PROM1	205	1.51	1.04	1	1	5	1
DifficUseButtonsZippers							
PROM1 DifficCutFingernails	203	1.43	0.99	1	1	5	1
PROM1 DifficDressYourself	203	1.42	0.86	1	1	5	1
PROM1 DifficWalk	205	1.84	1.08	1	1	5	1
PROM1 DifficGetMoving	205	1.80	0.93	2	1	5	1
PROM1 DifficGoOutYourself	205	1.60	1.23	1	1	5	1
PROM1 DifficDriveCar	201	1.74	1.41	1	1	5	1
PROM1	204	1.26	0.76	1	1	5	1
DifficCleanAfterBathroom							
PROM1 DifficUseHandle	204	1.24	0.69	1	1	5	1
PROM1 DifficWriteType	204	1.42	0.93	1	1	5	1
PROM1 DifficTurning	203	1.49	0.91	1	1	5	1
PROM1	205	2.24	1.40	2	1	5	1
DifficPhysRecreaActivity							
PROM1 DifficUsualLeisureAct	203	1.73	1.17	1	1	5	1
PROM1 DifficSexAct	190	1.94	1.58	1	1	5	1
PROM1 DifficLightHousework	206	1.54	1.09	1	1	5	1
PROM1	206	2.21	1.50	2	1	5	1
DifficHeavyHousework							
PROM1 DifficUsualWork	206	1.87	1.31	1	1	5	1
PROM1 OftenLimp	202	2.45	1.45	2	1	5	1
PROM1	205	1.88	1.14	1	1	5	1
OftenAvoidUsingPainful							
PROM1 OftenLegLock	204	1.62	0.95	1	1	5	1
PROM1	206	1.82	1.01	1	1	5	1
OftenProblConcentration							
PROM1 OftenOverworkAffect	205	2.32	1.25	2	1	5	1
PROM1 OftenActIrritable	207	1.79	0.97	1	1	5	1
PROM1 OftenTired	207	2.70	1.06	3	1	5	3
PROM1 OftenFeelDisabled	207	2.20	1.32	2	1	5	1
PROM1	207	2.20	1.20	2	1	5	1
OftenAngryFrustrated							
PROM1	206	2.12	1.05	2	1	5	2
BotherUseHandArmLeg							
PROM1 BotherUseBack	205	1.63	1.01	1	1	5	1
PROM1 BotherWorkHome	202	1.68	1.04	1	1	5	1

Table 15. Ordinal outcomes in the Gotland set.

ORDINAL VARIABLES	N	Mean	SD	Median	Min	Max	Mode
PROM1 BotherPersonalCare	206	1.48	0.95	1	1	5	1
PROM1 BotherSleepRest	205	1.67	1.00	1	1	5	1
PROM1 BotherLeisureRecreAct	205	2.15	1.31	2	1	5	1
PROM1 BotherFriendsFamily	205	1.20	0.62	1	1	5	1
PROM1 BotherThinkConcRem	205	1.64	1.00	1	1	5	1
PROM1 BotherAdjustCope	204	1.67	0.97	1	1	5	1
PROM1 BotherUsualWork	204	1.66	1.02	1	1	5	1
PROM1 BotherFeelDepend	204	1.63	1.13	1	1	5	1
PROM1 BotherStiffPain	204	2.25	1.09	2	1	5	2
PROM1 Recov	205	2.31	1.28	2	1	5	1
PROM1 Reoperated	205	0.14	0.34	0	0	1	0

For scores, a higher number means more decrease in function or more problems, except for EQ5D VAS where the reverse is true. N is the count of the parameter. The mode is the most common (highest frequency) value.

* VAS can be treated as numerical or ordinal variable, and some sources argue that it acts more like an ordinal than continuous variable ¹⁷⁷.

Table 16 and Table 17 reports the PROM Δ data, secondary outcomes, for the training set. We see that all scores, on average, show a decrease in function, whether they are directly associated with lower extremities or not. Examples are PROM Δ DifficOpenMedBottle or PROM Δ DifficWriteType. However, even the most affected SFMA parameter, PROM Δ OftenLimp, does not, on average, increase one step on the scale. I.e., the mean change is <1 for all ordinal values except EQ5D VAS.

Table 16. PROM Δ regression outcomes in the Stockholm training data.

REGRESSION VARIABLES	N	Mean	SD	Median	Min	Max
PROM Δ EQ5DIndex	2,672	-0.10	0.30	-0.07	-1.2	1.4
PROM Δ DailyActIndex	3,213	9.76	21.18	5	-100	100
PROM Δ EmotionalIndex	3,200	9.40	20	7.14	-82	79
PROM Δ ArmHandFuncIndex	3,238	3.01	12.74	0	-100	94
PROM Δ MobilityIndex	3,244	11.99	19.02	8.33	-72	89
PROM Δ FunctionIndex	3,243	8.56	15.67	5.88	-85	85
PROM Δ BotherIndex	2,949	9.82	18.71	6.25	-83	88

For scores, a higher number means more decrease in function or more problems, except for EQ5D Index where the reverse is true. N is the count of the parameter. The mode is the most common (highest frequency) value. PROM Δ is the one-year change in PROM.

Table 17. PROM Δ ordinal outcomes in the Stockholm training data.

ORDINAL VARIABLES	N	Mean	SD	Median	Min	Max
PROM Δ EQ5D VAS *	2,393	-8.75	22.24	-5	-100	98
PROM Δ EQ5D Anxiety	2,761	0.07	0.59	0	-2	2
PROM Δ EQ5D Pain	2,765	0.29	0.69	0	-2	2
PROM Δ EQ5D Usual Act	2,737	0.12	0.66	0	-2	2
PROM Δ EQ5D SelfCare	2,785	0.05	0.46	0	-2	2

Table 17. PROMΔ ordinal outcomes in the Stockholm training data.

ORDINAL VARIABLES	N	Mean	SD	Median	Min	Max
PROMΔ EQ5D Mobility	2,770	0.23	0.56	0	-2	2
PROMΔ DifficChair	3,229	0.42	0.97	0	-4	4
PROMΔ DifficOpenMedBottle	3,207	0.12	0.80	0	-4	4
PROMΔ DifficShop	3,185	0.27	1.02	0	-4	4
PROMΔ DifficStairs	3,209	0.56	1.03	0	-4	4
PROMΔ DifficTightFist	3,201	0.07	0.71	0	-4	4
PROMΔ DifficShower	3,203	0.32	1.00	0	-4	4
PROMΔ DifficComfortSleep	3,208	0.26	0.95	0	-4	4
PROMΔ DifficBendKneelDown	3,224	0.59	1.18	0	-4	4
PROMΔ	3,227	0.13	0.71	0	-4	4
DifficUseButtonsZippers						
PROMΔ DifficCutFingernails	3,214	0.12	0.80	0	-4	4
PROMΔ DifficDressYourself	3,205	0.21	0.73	0	-4	4
PROMΔ DifficWalk	3,210	0.52	0.98	0	-4	4
PROMΔ DifficGetMoving	3,184	0.49	0.92	0	-4	4
PROMΔ PROMΔ	3,209	0.26	0.99	0	-4	4
DifficGoOutYourself						
PROMΔ DifficDriveCar	3,142	0.29	1.09	0	-4	4
PROMΔ	3,216	0.10	0.68	0	-4	4
DifficCleanAfterBathroom						
PROMΔ DifficUseHandle	3,209	0.11	0.69	0	-4	4
PROMΔ DifficWriteType	3,222	0.09	0.68	0	-4	4
PROMΔ DifficTurning	3,224	0.17	0.79	0	-4	4
PROMΔ	3,175	0.75	1.32	1	-4	4
DifficPhysRecreaActivity						
PROMΔ DifficUsualLeisureAct	3,169	0.38	1.11	0	-4	4
PROMΔ DifficSexAct	2,784	0.20	1.16	0	-4	4
PROMΔ DifficLightHousework	3,193	0.20	0.93	0	-4	4
PROMΔ DifficHeavyHousework	3,183	0.43	1.16	0	-4	4
PROMΔ DifficUsualWork	3,160	0.37	1.10	0	-4	4
PROMΔ OftenLimp	3,134	0.93	1.51	1	-4	4
PROMΔ	3,142	0.53	1.30	0	-4	4
OftenAvoidUsingPainful						
PROMΔ OftenLegLock	3,153	0.33	1.04	0	-4	4
PROMΔ	3,169	0.22	0.98	0	-4	4
OftenProblConcentration						
PROMΔ OftenOverworkAffect	3,143	0.63	1.34	0	-4	4
PROMΔ OftenActIrritable	3,172	0.12	0.97	0	-4	4
PROMΔ OftenTired	3,187	0.34	1.12	0	-4	4
PROMΔ OftenFeelDisabled	3,181	0.69	1.27	0	-4	4
PROMΔ OftenAngryFrustrated	3,184	0.54	1.24	0	-4	4
PROMΔ	3,202	0.85	1.29	1	-4	4
BotherUseHandArmLeg						
PROMΔ BotherUseBack	3,176	0.18	0.98	0	-4	4
PROMΔ BotherWorkHome	3,185	0.35	1.02	0	-4	4
PROMΔ BotherPersonalCare	3,210	0.23	0.92	0	-4	4
PROMΔ BotherSleepRest	3,192	0.12	1.04	0	-4	4
PROMΔ BotherLeisureRecreAct	3,159	0.70	1.39	0	-4	4

Table 17. PROMΔ ordinal outcomes in the Stockholm training data.

ORDINAL VARIABLES	N	Mean	SD	Median	Min	Max
PROMΔ BotherFriendsFamily	3,189	0.09	0.91	0	-4	4
PROMΔ BotherThinkConcRem	3,201	0.11	0.92	0	-4	4
PROMΔ BotherUsualWork	3,161	0.41	1.18	0	-4	4
PROMΔ BotherFeelDepend	3,197	0.24	1.11	0	-4	4
PROMΔ BotherStiffPain	3,187	0.54	1.26	0	-4	4

For scores, a higher number means more decrease in function or more problems, except for EQ5D VAS where the reverse is true. N is the count of the parameter. The mode is the most common (highest frequency) value. PROMΔ is the one-year change in PROM.

* VAS can be treated as numerical or ordinal variable, and some sources argue that it acts more like an ordinal than continuous variable ¹⁷⁷.

Below, we report the training and validation of RMSE for some outcomes. Figure 13 shows changes in the RMSE for two models. We look at the curves in general and do not focus on individual models. We compare the RMSE to the SD from Table 13. The validation error for OftenAvoidUsingPainful approached an RMSE of 1.25 vs. SD 1.24. OftenFeelDisabled had RMSE 1.45 vs. SD 1.41, whereas OftenProblemConcentration had RMSE 2.00 vs. SD 1.07 for the best model. OftenLimp is decreasing towards RMSE 1.55 vs. SD 1.45. OftenLegLock at best performs at RMSE 2.40 vs SD 0.95.

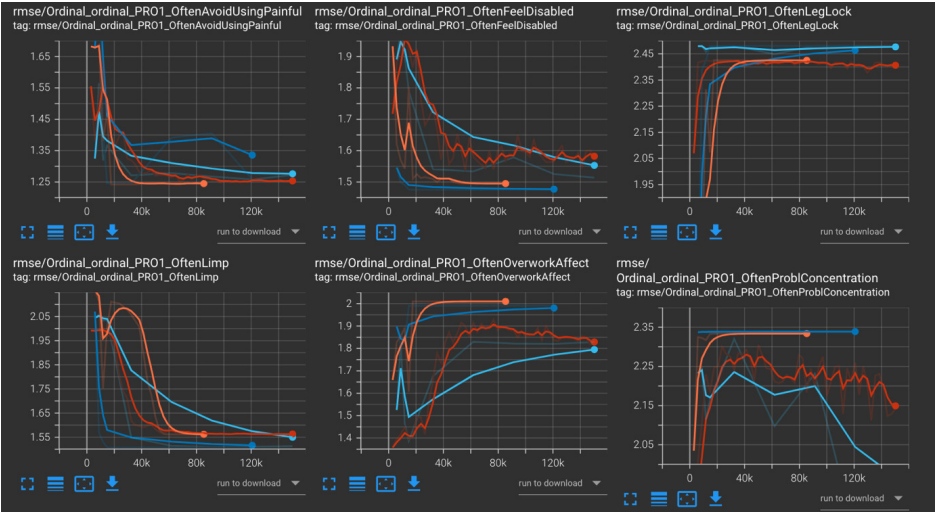


Figure 13. Comparing the root mean squared error (RMSE) for PROM1 parameters. The x-axis is batch iterations, and the y-axis is the RMSE.

We find the same pattern for the outcomes in Figure 14, which also reports additional variables. In Figure 14, we also find that the best model's prediction accuracy for "Died in study period" was 67.5% on average, which is the percentage of people who did not die in the study period. We see a similar

pattern for “Injury Sex.” For most parameters, we saw training performance tapering off and little additional benefit from additional training.

Figure 15 shows that the one-year change in PROM, $PROM\Delta$, outcomes deviated considerably for these same models. The graphs are representative of $PROM\Delta$ training performance. Preliminary results show that after seven epochs, the Focal CORN Loss (with $\alpha=1$ and $\gamma=2.0$) seems to perform better than other losses. However, we started to see the network clipping the losses, i.e., the losses are so small that they are set to zero and ignored to prevent exploding gradients – but these cases do not contribute to learning.

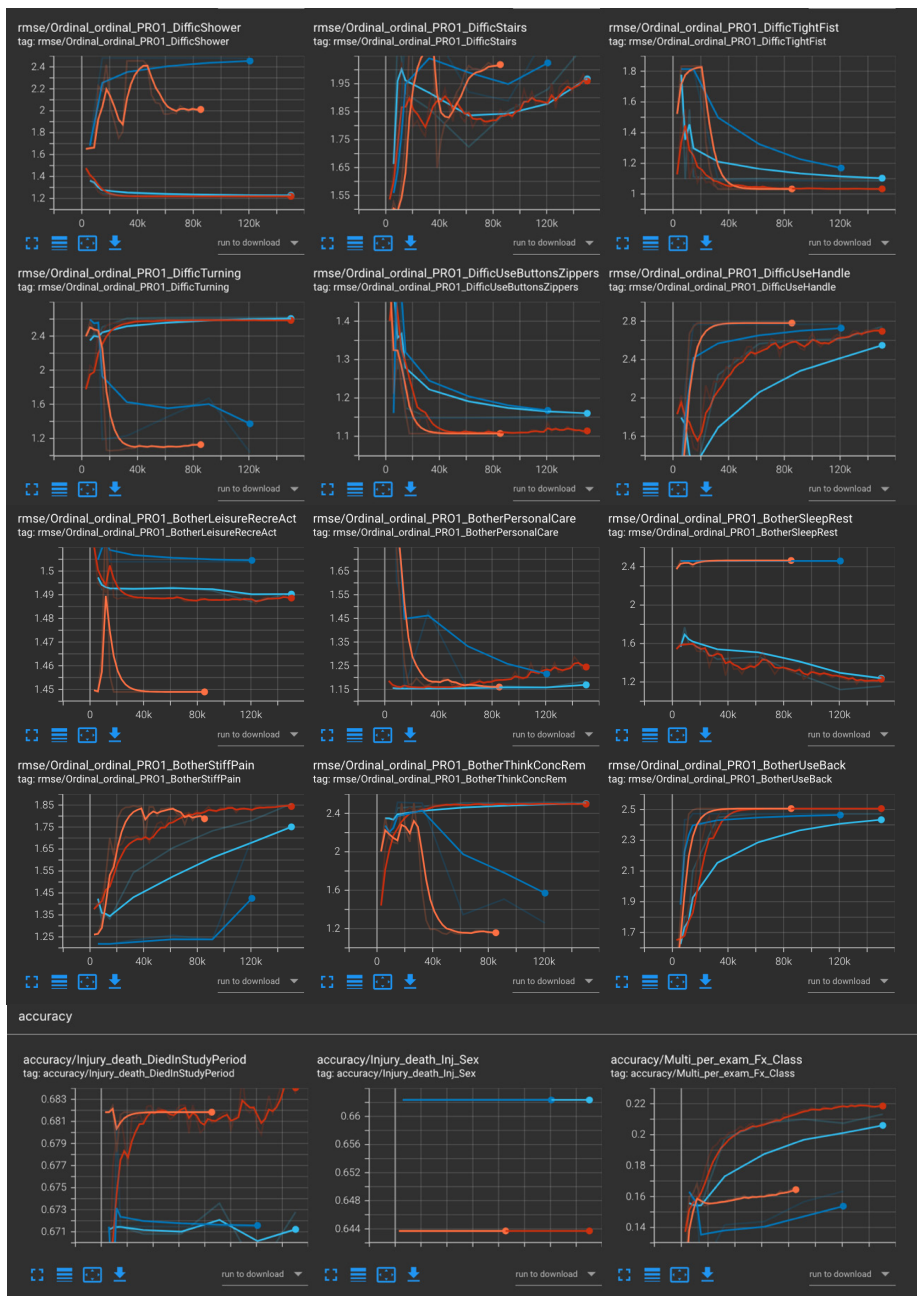


Figure 14. The performance for some PROMIS outcomes and classification tasks

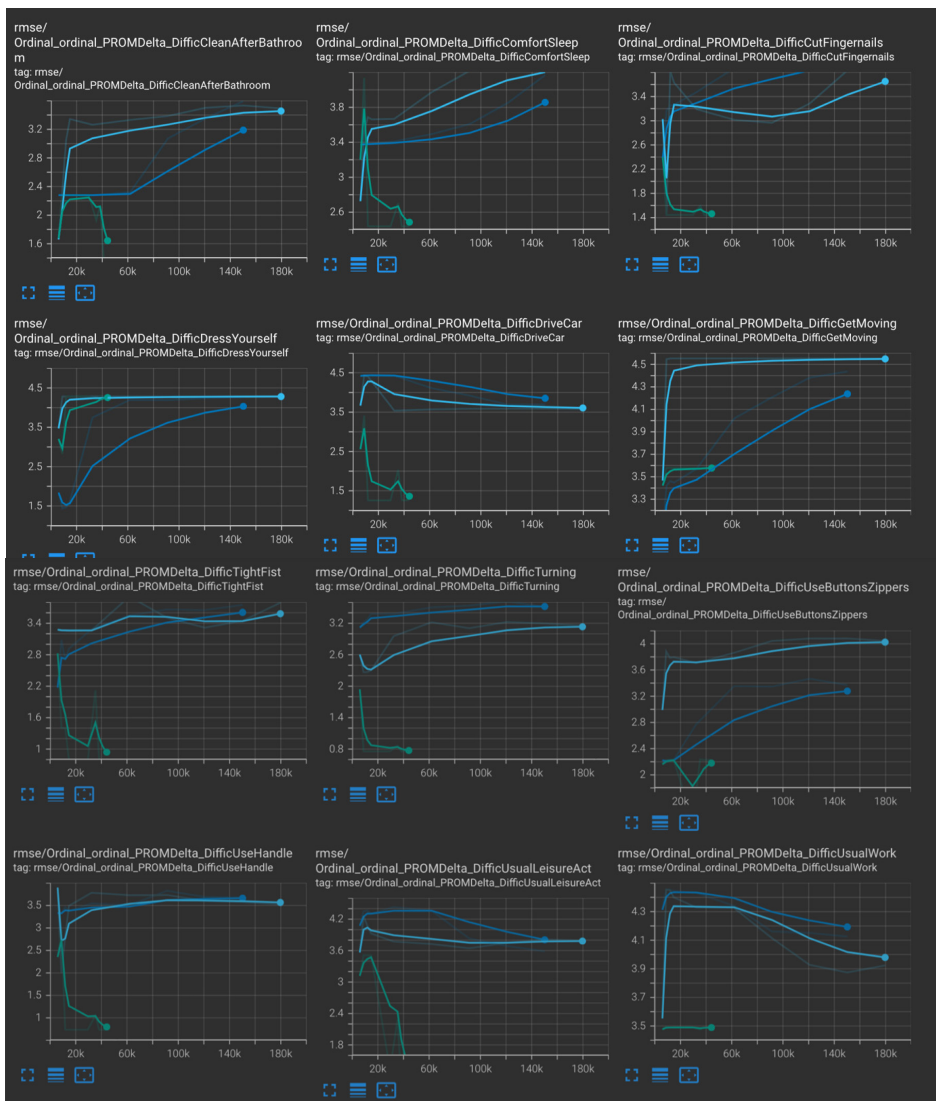


Figure 15. Training RMSE for PROMDelta parameters. The Focal CORN Loss (green) seven epochs compared to models using MSE and robust adaptive loss at approximately 50 epochs.

5 Discussion

5.1 Fracture detection using CNNs

The kind of fracture detection from radiographs, as concerns this thesis, is a relatively recent phenomenon. In Olczak 2017 (Study I), we applied artificial intelligence to fracture detection using CNNs¹. We built upon the ideas of Shin et al.¹⁷⁸ and Tajbakhsh et al.¹⁷⁹, who used transfer learning on medical images of chest radiographs. Both used pre-trained CNN and retrained them to classify chest radiographs. We applied a similar strategy to orthopedic trauma radiographs of hands, wrists, and ankles. We reached orthopedic surgeons' detection performance for several different outcomes; however, these were on the downscaled image¹¹⁴⁷. An additional feature, only implicitly stated in the original article, was that unsupervised learning with **natural language processing** (NLP) – language and text analysis – was used to derive the labels. NLP is another form of ML, and the goal was to create a workflow from report to label to classification – as there were over 250,000 radiographs. However, the NLP method caused problems with label and classification accuracy^{1146,147}. Kim and MacKinnon also used the transfer learning approach to study radiographs of distal radius fractures. They performed better than our study on a much smaller, more curated, and less clinically relevant dataset¹⁸⁰.

Urakawa et al. studied intertrochanteric hip fractures and achieved high performance in fracture detection¹⁸¹. Gale et al. predicted the presence of hip fractures with expert-level accuracy¹⁸². Still, the results were not peer-reviewed, and no peer-reviewed version has been presented. However, Badgeley et al., from the same research groups, presented results from a study of hip fractures, which reached a very high accuracy. However, they could also show that their accuracy was random once they accounted for logistic and healthcare parameters¹³⁶. In essence, the model overfitted to other data parameters than fracture detection. Nicolaes, in turn, studied vertebrae fractures and were able to localize fractures in a CT scan¹⁸³.

The described papers illustrate essential pathology detection. Just detecting a fracture (e.g., fracture is present or not) is not always a clinically relevant task. It is something most doctors can quickly learn. This triviality is also suggested by the excellent AUC values that many studies report. We believed that it is more relevant to determine the properties of the fracture, whether we need to do something about the fracture, and what that intervention should be. As

discussed, the commonly used criteria for selecting interventions are, for example, Neer's classification for humerus fractures or the Lauge-Hansen or the AO classification for ankle fractures. These are more complex classification tasks than just detecting the fracture, and the results are more complicated to learn and present to the end user.

Qi et al. trained a CNN to detect and classify femoral radiographs according to the AO classification and to place bounding boxes around the fracture location¹⁸⁴, i.e., the **region of interest** (ROI). They attained an area under the receiver-operating characteristic curve (AUC) of 0.71 for the double task¹⁸⁵, which is not considered useable in a clinical context. It was also unclear if the ROI, the classification, or the combined task attained that level of accuracy. In 2021 (Study II)², we studied ankle fractures, classifying them according to the AO 2018 classification without drawing an ROI. We reached a weighted AUC of 0.90 for all classes. We also approached prediction differently. Qi et al. approached classification in order of severity, i.e., C3 to A1. If the algorithm detected a positive outcome (>50% likely) in one class, it stopped, and later classes were ignored – even if that class would be more correct. Instead, we predicted all outcomes simultaneously and selected the most likely outcome. We also examined the complete study – i.e., where the fracture might be visible in one projection but not another, whereas Qi et al. studied individual images/radiographs. Gan et al. examined the presence or absence of distal radius fractures and located the ROI but did not classify the fractures further¹⁸⁶.

Chung et al. studied humerus fractures according to Neer's classification¹⁸⁷, whereas Heimer et al. used cadaveric CT scans to study skull fractures¹⁸⁸. Choi et al. studied fracture detection in pediatric elbows¹⁸⁹. Blüthgen et al. studied wrists¹⁹⁰. More complex outcomes also occur. Dreizin et al. studied CT slices of pelvic studies to classify fractures according to the AO standard¹⁹¹. Lind et al. classified knee fractures¹⁹, Qi et al. identified femur fractures¹⁸⁴, Tanzi et al. identified hip fractures⁹⁸, and Akbarian et al. studied hip fractures according to the AO 2018 system²².

While these studies use CNNs to study fractures, other outcomes are also explored. For example, Jang et al. used a CNN to predict osteoporosis from radiographs¹⁹², Magnéli et al. studied glenohumeral osteoarthritis and avascular necrosis, and Olsson et al. classified knee osteoarthritis according to the Kjellgren-Lawrence system²⁰.

We see that tradeoffs need to be made during algorithm creation and implementation. Some consider a bounding box (i.e., ROI prediction) highly valuable. In contrast, others consider it a risk that will draw attention to the ROI but cause reviewers to miss the whole picture. The idea behind these studies is that if we can classify fractures and medical image data accurately and consistently, we can use that information to agree upon treatment. As we saw, the critique of many classification systems is their (a) difficulty of application and (b) questions about their reproducibility between observers, leading to their (c) poor utility. The development of consistency, reproducibility, and reliability will produce utility. In addition, any system can be trained to report the class of several models, allowing for comparison and usage of what best serves the situation.

None of the mentioned studies examined outcome prediction from imaging using CNNs.

5.2 Imaging-based patient outcome prediction

While studies exist that use image-based CNNs to predict outcomes, these are usually for chest radiographs, chest CT, or skull imaging. They derive predictions for COVID-19, pneumonia, ICU admission, etc.^{193–197}. Shin et al.¹⁹⁷ built a model to predict pneumonia outcomes on an existing imaging data analysis platform. A clinical software model examined chest radiographs and calculated a severity score. This severity score was then part of a multivariate Cox-regression model to predict pneumonia outcome. Kim et al. used a similar approach but trained the image analysis CNN themselves. They used a pre-trained CNN on chest radiographs to predict 30-day mortality from the radiographs and compared it to a clinical score. The CNN performed better, but not significantly better. Then, similarly to Shin, they combined the CNN output with the clinical score in a logistic regression model and used the score output instead of the clinical score components. The combined model performed significantly better than the clinical score or CNN alone¹⁹⁶.

Pease et al. developed a model for predicting outcomes after traumatic brain injury from CT scans. They created a separate linear discriminant analysis model and combined the models using an ensemble stacking model to create a superior model¹⁹⁵. Gordeau et al.¹⁹⁴ trained a network to predict mechanical ventilation outcomes in COVID-19 patients. Instead of ensemble stacking, they used a similar feature enhancement approach. They pre-trained a model and

then selected the (two out of 1024) most distinctive features of the classifier. They used those features and the linear model output that predicted mortality to train various classifiers. Kwon et al. taught a CNN to predict COVID-19 outcomes. Unlike the other models, they added clinical variables as input to the last fully connected layer, i.e., the classification layer¹⁹³. We have found no imaging-based outcome predictor for orthopedic trauma. The closest was Alfraihat et al., who used radiographic features to predict future radiographic features. They used features of the images but not the actual imaging¹⁹⁸.

5.3 Study I

5.3.1 Discussion of results

We showed that the CNN could classify radiographs on par with human reviewers regarding the presence or absence of fractures. The most common errors were due to image ambiguity or missing data. “Fracture” was a label for the entire study, whereas the network looked at individual images. This confused the training and performance assessment we concluded in our manual review.

We showed that CNNs trained for other tasks could be retrained to detect fractures. In addition, they could be retrained to detect exam views, body parts, and laterality in a skeletal radiograph. This had previously been tested for other medical domains, such as lung nodules in chest radiographs¹⁹⁹, spine MRI²⁰⁰, and CT slices^{178,179,201}. However, this was the first study to show this for skeletal trauma radiographs.

We also showed that deeper layered models, with more features and nodes, outperformed shallower models, which indicated that the extra computational effort to train them was worthwhile. Neither did we see tendencies toward overfitting the data during training. We believed this was because we had a large data set, that training and validation sets were resampled at each epoch, and perhaps the automatic labeling created noise in the data. Overfitting means the model learns the individual data points, e.g., recognize the image and label rather than the features that define the label.

Surprisingly, the best networks could capture laterality better than the others. As training images were randomly mirrored and rotated, laterality effects from the scanner—e.g., the right hand appearing to the left—should be largely eliminated. Our interpretation was that the network captures other indicators that we did not. Perhaps the dominant hand's bone structure or tissue differs from the non-

dominant hands. This indicates that ML models can find patterns and predictors of which we are unaware. For example, an automated algorithm found that the stroma around breast cancers had value in prognosis ²⁰².

5.3.2 Strengths

We used a large dataset of 256,000 radiographs, which we believe assisted it in not overfitting.

With excellent performance on secondary outcomes, many of which have multiple possible outcomes, we showed that the ML model learned to interpret the data.

5.3.3 Limitations

The primary outcome, fracture, was automatically extracted from radiologist reports, though the extraction criteria had been manually coded via key phrases. Trained specialists with many years of experience generated the reports. The language was not always easy to interpret, and the same report could refer to different fractures or features in the same exam. We concluded that radiologist reports were unsuitable for labeling orthopedic trauma radiographs.

The classification, fracture/no fracture, has limited utility. Improved extraction using improved natural language processing might provide more helpful information.

Radiologist reports have limitations. They answer specific questions in the referral. Since we did not have access to referrals, the reports were taken out of context. Information in the image might have been omitted, which limits their utility.

The fracture was labeled for the study, whereas the network looked at individual images. While a fracture might be visible in one projection, it could be hidden in another. However, the network and gold standard studied each radiograph independently and did not consider this.

We did not have population data for the dataset, so we could not infer how general the results were.

5.4 Study II

5.4.1 Discussion of results

We classified fractures according to the AO standard for ankle fractures using AI, with better than random performance. Unlike Study I, we did not use a pre-trained network for Study II. In addition, unlike Study I, the network in Study II looked at the entire examination, including all images and projections.

One of our stated limitations with Study I was that we believed that classification needed to be more complex than just detecting fracture. For that reason, we implemented the AO 2018 ankle classification. As implemented in our study, we used the AO classification down to the subgroup level. We looked at type, group, and subgroup independently. Theoretically, but unlikely, a study could be classified as type 44A, group 44B2, and subgroup 4F2C3.3. We looked at all the images in the study for all possible outcomes and selected the most likely outcome. Some outcomes could co-occur, such as fracture yes/no, and other types of fractures, such as foot and tibia fractures. A different approach, chosen by Qi et al., was to select a priority order for the network. While they were only classified into the AO group, they looked sequentially at the fractures in order of severity, i.e., 44C3 > 44C2, 44C1, 44B3, etc. If 44C3 was positive, the model stopped and never checked if another, less “serious” injury group had a higher probability¹⁸⁴.

A problem we encountered with our classification, which became more pronounced the more finely granular the class, was missing data and imbalanced data. As we saw in Table 6, some outcomes are not represented in the training or testing data (such as 44A3.2) or are very rare in the training data and not present in the test data (e.g., 44A2.2). The model will thus never be able to detect a 44A3.2 fracture and is highly unlikely to learn what is specific for a 44A2.2 fracture. In addition, 44A2.2 occurs once in the entire dataset of 409 studies. Several other outcomes are about nearly as rare. These outcomes pose a problem when measuring performance. We cannot compute statistics for non-present classes or reliable statistics for classes with too few cases. In addition, as we saw in our discussion of accuracy and AUROC, accuracy can be over-optimistic.

5.4.2 Strengths

Our complex and granular model meaningfully and comprehensively represented the AO classification for lower extremity trauma. We focused our reporting on the primary outcome of malleolar fractures. The article supplement reports performance for secondary outcomes for fibular, tibial, and foot fractures and IRR.

5.4.3 Limitations

The model performance was difficult to assess, with few cases for many outcomes. Even while significant training had occurred, the actual utility of the classifier was challenging to determine. Many classes were missing or had very few training cases. This introduced classes similar to a more prevalent class but gave a tiny training signal. Excluding rare cases from training might give better model learning for remaining outcomes.

We did not perform external validation, so it was not possible to assess how representative the model was in different clinics or scenarios.

The training and test sets were not extensive enough to fully capture all possible malleolar fracture outcomes.

5.5 Study III

5.5.1 Discussion of results

The study aimed to validate a fracture classification model externally and to study strategies to deal with the difficulties that arise from the change in environment – dataset shift. Despite having very few training cases for some outcomes, the model appeared to perform better than chance at all individual outcomes. However, it was sometimes impossible to show due to low prevalence. As expected, the model appeared to perform better for the IVD. Comparing our classifier to other classifiers was difficult, as model external validation is rare and even rarer for complex classifiers.

5.5.1.1 *Model training*

ML training often comes down to learning hidden factors, and ML models are usually considered “black boxes.” Hopefully, the parameters learned are related to the appearance of the actual pathologies. As mentioned previously, the study by Badgeley et al. found that healthcare and logistic parameters were often responsible for prediction, i.e., a form of overfitting. Correcting their models for

those factors, the performance of a well-performing classifier fell to that of a random classifier¹³⁶. Exposing an ML model to a dataset from a different location subjects it to a different distribution – also called a dataset shift²⁰³. It helps to correct for logistic factors. While data is not readily available, external validation should be integral to the more mature model training and development stage. If a model only performed well on the data it was trained on or from one hospital, we could quantify this and seek ways to amend it.

In this study, the EVD had properties other than those of the IVD. There were three times as many type A fractures in the EVD. All EVD studies had three images compared to Danderyd, which had at least four views. The CNN had never been exposed to follow-ups during training, but such follow-up studies were present in the EVD, as signaled by “weight-bearing.” A human reviewer who sees a non-displaced “weight-bearing” fracture understands this as less alarming. The network was not trained to recognize this signal. AI models are rarely validated. This makes it difficult to assess how transferable or general they are. It also made it difficult to determine what performance we could expect in our study or whether our results were good or bad. For the three external validation studies, Oliveira e Carmo et al. found that performance was not dramatically affected by the EVD²⁰⁴ (see Table 18a.) Those studies evaluated models with just a few outcomes. Our classifier had 40 outcomes for ankle fractures – not all mutually exclusive. The review found two similarly complex classifiers^{19,20} (see Table 18b) with model-wide AUC on the internal validation data, similar to our model. However, they were not externally validated and are from the same dataset as the ankle subsets in Study II and Study III (radiographs from Danderyd Hospital between 2002 and 2016) were taken from.

We were dissatisfied with the model during external validation and wanted to try ways to improve model performance without overfitting data. Increasing image resolution during training, on its own, did not affect EVD performance. Dropping views in the exams to make the training data resemble the EVD more was ineffective. We believed that type A fractures only provided a discrete training signal for the network. We therefore concentrated on active training (i.e., additional data for training that focused on the problematic class and on predicting edge cases for that class). In combination with improved resolution, we saw improved performance. However, performance did not improve beyond 400x400px.

Table 18. Comparable studies to Study III.

Study	Anatomy	Outcomes	Exclusion	Performance
18a External validation studies				
Choi 2020 ¹⁸⁹	Elbow	Supracondylar/ no fracture	Dislocation, not supracondylar fracture, bone dysplasia	IVD: AUC 0.98 EVD: AUC 0.99
Blüthngen 2020 ¹⁹⁰	Wrists	Intact/defect	–	IVD: AUC 0.93 EVD: AUC 0.80 IVD: mean F1- score 0.84 EVD: F1-score 0.73
Zhou 2020 ²⁰⁵	Ribfractures CT-slices	Old, healing, and fresh	No fracture	
18b Complex classifiers				
Dreizin 2021 ¹⁹¹	Pelvic, CT- scans	AO Type A–C No. outcomes: 3	Any operative treatment	ACC 56–85% AUC 0.87 for proximal tibia; 0.89 for patella; 0.89 distal femur
Lind 2021 ¹⁹	Knee	AO knee No. outcomes: 49	–	
Qi 2020 ¹⁸⁴	Femur	AO femur No. outcomes: 11	Any disagreement between reviewers	ACC 72%
Tanzi 2020 ²⁰⁶	Hip	AO No. outcomes: 5	Type B and C	AUC 86%
Yoon 2020 ²⁰⁷	Inter- trochanteric 3D CT	AO type A No. outcomes: 10	No separation of patients between training and test	ACC 97% and 90% AUC 0.87, F1- score 0.86, vs AUC 0.75, F1- score 0.5 depending on configuration
Lee 2020 ²⁰⁸	Femur	AO A1–B3 No. outcomes: 9	Type C (too rare)	
Olsson 2021 ²⁰	Osteoarthritis	Kellgren & Lawrence No. outcomes: 5	–	AUC 0.92
Chung 2018 ¹⁸⁷	Shoulder	Neers' No. outcomes: 5	Reviewer disagreement	ACC 65–86%; AUC 0.90–0.98

The IVD is from the original training data location. The EVD is any data from a different site. Table 18a compares external validation studies found by Oliveira e Carmo et al. 2021. Table 18b compares studies that evaluate complex classifiers with many outcomes (multinomial classifiers) comparable to our study, where none is externally validated. ACC is accuracy. F1 is the F1-score. AUC is the area under the receiver-operator characteristic curve (AUROC). Table from Olczak et al. 2024³.

External validation ensures the model's generality. However, we believe that reversing a model's generalization process in a clinical application can be desirable. We could gain a more locally accurate model by actively retraining the externally valid model to be more specific with data from the clinic or scanner where it will be used, like the transfer learning we studied in Study I.

We need to be careful when applying an algorithm to a new setting. Lim et al. concluded that many common orthopedic procedures had poor evidence-based medicine support and were unnecessary²⁰⁹. Audgé et al. found that many fracture classification schemes used in the clinic were not validated^{46,61}. Oliveira e Carmo et al. found that many ML models were not externally validated²⁰⁴. As far as we know, this was the first study to raise the question of what we can expect from such a complex fracture classification model in terms of external validity. Comparing our model to other multinomial classifiers, it transferred well (see Table 18b)^{19,20,184,187,191,206–208}. Tools like our model and its improved iterations could be part of the solution toward a more evidence-based and stringent form of medicine.

5.5.2 Strengths

This was an external validation study, which is rare in orthopedic ML. We found no external validation study of such a broad classification scheme.

The EVD differed from the IVD but still focused on the same problem. Introducing new problem domains (weight-bearing, one-week follow-ups) strengthens the reliability of the results. Even after active training, there was no real risk of model overfitting.

5.5.3 Limitations

We did not have population data for the IVD dataset, so none was collected for the EVD. Having an external validation dataset compensated for this, compensated for this somewhat.

Since we found no similarly complex classifiers to be externally validated, we had nothing against which to compare and assess our classifier.

Though there was considerable learning, and practically all outcomes were better than chance, many were too rare and difficult to bound. This made model performance challenging to assess for rare outcomes.

As with all CNN models, we do not know the actual decision algorithm and cannot know why it works/does not work. While we can use heatmaps—or activation maps—they do not provide rules or guides for improving the model.

5.6 Study IV

5.6.1 Discussion of results

Using imaging data and registry parameters, we have modeled patient outcomes after a fracture in the lower extremities. Compared to previous studies, Study IV shifted the focus to modeling numerical and ordinal outcomes. Therefore, we returned to the experimental trial approach of preceding studies, such as Study I, and related studies, such as the one documented in Olczak 2024^{146,147}. We studied 154 different PROM (primary and secondary) outcomes. At a 95% confidence level, we expect approximately eight outcomes to appear as good fits randomly. This is less likely to be a random chance if the same outcome performs well during training and external validation.

The validation performance, i.e., our proxy of model performance until all experiments are trained, showed learning of PROM1 parameters but less so for PROM Δ . Initially, our best-performing models, just under 50% of outcomes, showed learning in PROM1 but very few in PROM Δ . The changes in PROM over the year were small. We, therefore, experimented with different loss functions. The robust general loss was designed to deal robustly with outliers, i.e., to smooth them out, and there were tendencies to underperform compared to the standard MSE loss. However, after we implemented the Focal CORN Loss, we also started seeing learning in PROM Δ . The goal of implementing the Focal Loss was to capture uncommon and incorrectly classified examples that deviate from the mean and mode, i.e., to increase the importance of the outliers. It can be that we are pushing the model to rely less on the images and even more on the SFR data, and that is why we are overfitting further.

As discussed in the literature review, we found orthopedic studies that used imaging directly to predict patient outcomes, nor had they been reported as failures or successes. The closest was Alfraihat et al., who used radiographic features derived from the radiograph to predict future radiographic features but did not use the actual imaging¹⁹⁸. Pease et al. developed a CNN for predicting outcomes after traumatic brain injury from CT scans. Their model predicted mortality or the value on a brain injury outcome scale¹⁹⁵. Like the approach in Study IV, Kwon et al. trained a CNN to predict COVID-19 outcomes using patient

parameters and imaging. Unlike our model, they passed the clinical variables as input to the last fully connected layer, i.e., the classification layer, along with the CNN image data ¹⁹³. Many studies combine the output of CNNs with regression models to improve performance. Shin et al. ¹⁹⁷ used an existing CNN tool to predict a severity score for pneumonia, and this severity score was funneled into a regression model. Kim et al. used a similar approach but trained the network themselves to predict 30-day mortality. They combined the CNN output with the clinical score in a logistic regression model ¹⁹⁶. Gordeau et al. trained a network to predict mechanical ventilation outcomes. They pre-trained a model and then selected the (two out of 1024) most distinctive features of the image classifier (i.e., most variable nodes in the last layer before the classification layer). For each prediction, they passed those features to a linear model, and that output was to train other models. We have found no imaging-based outcome predictor for orthopedic trauma ¹⁹⁴. Pease et al. used their CNN outcomes as input to a regression model to improve predictions ¹⁹⁵.

Study IV was a pilot study to determine the feasibility of using a combination of radiographs and patient parameters to predict patient outcomes. We provide some considerations for improving the modeling in the future.

Kwon et al. combined different patient parameters with the imaging, as we did. They passed the clinical variables and the image data as input to the classification layer. Missing values were imputed ¹⁹³. In this study, given the many PROM parameters, the different anatomies and injuries, and the possibility of several inputs missing simultaneously, creating a linear model for imputation would require several different imputation models. We would have had to develop different imputation strategies for various types of fractures, possibly with fewer cases than there are parameters to model. Therefore, we ignore missing values and train each outcome separately. However, value imputation could improve modeling in future studies as it could strengthen the relationship between variables.

We experimented with predicting complications, such as reoperation and infection, within one year. While some complications are registered in the SFR, they are not complete or well-defined. In addition, it would have required us to manually study and label all imaging for each fracture within one year of the trauma, including MRI and CT scans, from all possible locations the patient might have visited during that year. We would have to look for signs of reoperation,

infections, and other signs of a complication. This was beyond the scope of this study. In addition, we did not attain sufficient quality in our primary outcomes to focus on secondary outcomes.

Our network examined each series individually, and each study associated with the injury got the same patient parameters. The model received the same PROM parameters in some experiments for different studies. For example, a hip fracture with multiple adjacent studies, e.g., long femur and knee imaging, could get the same PROM. However, if there were no fractures in the knee, that imaging would look uncomplicated, giving confusing signals to the model. A different strategy would have been to combine all imaging from the same time interval into a single series for a complete picture. This would have been similar to how we went from looking at single images in Study I to looking at complete series in later studies.

Our study focuses on studying multiple fracture types in one model, including all the lower extremities. We did this in part because of the low PROM answer rate. It could have been better to focus on one segment. In the Stockholm data, hips, wrist, and third ankle fractures were the most common, two of which we captured in this study. We can hypothesize that lower extremity injuries will affect patients similarly for many PROMs. A wrist fracture will affect patients very differently and would likely not make the model learn better.

We used a standard ResNet. Other CNN architectures, or indeed transformer networks, could have been used. We could have experimented with different architectures, similar to the approach in Study I, to select the best suited. However, unlike Study I, we did not have a clear-cut outcome to determine the “best” model. In addition, it was a good strategy to start with (or calibrate on) a known architecture and see what performance we could reach there. The risk is that ResNet is insufficient to capture the relationships well, as for Network In Network compared to VGG-16 in Study I.

5.6.2 Strengths

Our patient data comes from the SFR, a large, validated national register that includes non-operatively treated fractures.

This is a multi-center study with external validation.

5.6.3 Limitations

The data's biases and limitations correspond with the selection bias generally expected from registries and PROMs. However, there are indications that non-responders in the SFR are much like responders regarding PROM.

The switch from EQ5D-3L to 5L limits the generalization and prospective power of EQ-5D outcomes until a validated mapping between the two is established.

We did not have information on the patient's other health parameters, which are known to affect outcomes, such as smoking, diabetes, cardiovascular health, etc. Co-morbidities are essential in recovery, in deciding treatment and overall outcomes.

6 Conclusions

6.1 Study I

We showed the utility of artificial neural networks in detecting the presence or absence of fractures in trauma radiographs and various anatomies with high accuracy. We also showed the viability of transfer learning for orthopedic fracture detection. We also found that we could detect other features from radiographs, such as body parts, exam views, and laterality. In part, this was due to better labels from which to train.

6.2 Study II

We developed a fracture classifier for the AO 2018 ankle classification system to classify fractures to the subgroup level (44A1.1–44C3.3). Performance fell with complexity. For example, the AO type was more accurate than the group. This was expected as the task was difficult for human reviewers to perform on a radiograph.

6.3 Study III

In Study III, we underscored the critical need for external validation of AI models, as it is a crucial factor in assessing their utility. Our exploration of external validation revealed that our initial model did not perform as desired on external data. This was highlighted by the impact of unexpected logistic factors that reflected different clinical practices. We refined our model using active learning and concluded that while AI models should be trained to be general, they will later benefit from being honed for the specific task and setting to which they are applied.

6.4 Study IV

In Study IV, we conducted a multicenter study to train a CNN to predict patient outcomes using fracture radiographs and patient-reported outcomes. After focusing on classification tasks in previous studies, Study IV shifted towards modeling numerical, categorical, and ordinal variables, usually the domain of regression models. We experimented with different ways to construct models. We found that the network could learn to predict PROM1, but the indications were that it had learned the mode. We found a low response rate to the SFR, which caused problems with both the training and external validation set and data size.

7 Points of perspective

- Since the beginning of the projects that amounted to this thesis, new technologies have been introduced for AI and ML. Attention networks and transformers are powerful image analysis tools. They are the foundation of the generative pre-trained transformers (GPT) models in vogue today. During the thesis process, using transformers for this research was contemplated. However, the amount of data, processing power required, and potential ethical implications were prohibitive. While still a computationally massive undertaking, it is now feasible with the growth of computational power and resources available.
- The tasks that CNN and AI models perform for the user must be more comprehensive and practical. However valuable, identifying the presence of a fracture is not an exceedingly challenging task. The ultimate goal would be to have an AI model that reliably and coherently gives properties of the fracture, which could be tuned and honed as the knowledge of the field changes. In our studies, we have chosen the AO standard as an imperfect proxy. If we can create a model to predict the outcome after a fracture, we could attempt to reverse engineer a “classification” scheme that finds predictive features in the fracture appearance of which we were unaware.
- Even more, using ML, CNNs, and other types of ANNs to predict long-term outcomes after a fracture from imaging and patient data would be a step along personalized precision medicine in orthopedics. We can imagine an enhanced model that takes the imaging, patient parameters such as comorbidities and PROM at injury – and accurately predicts the most likely outcomes for the patient conditioned on different treatment options. No single option might give perfect outcomes for all functions, but we could select the one that optimizes the patient’s desired outcome. Given the nature of ANNs, they can be retrained and updated as time goes on, treatment evolves, and the model impacts future patients.

8 Acknowledgements

Thank you to all who have made this work possible.

Max Gordon. You have been wonderfully supportive and truly inspirational. Your breath and depth of competence is amazing, and so is your warmth and patience.

Olof Sköldenberg, your positivity and support has been instrumental.

Ali Sharif Razavian, without your groundwork this thesis would not have been possible.

Thank you, all my **collaborators** and **co-authors**. Many of you I have never met in person but all of you have contributed to making this work possible and made me better and smarter.

Gustaf Drevin thank you for all our talks and your steadfast inspiration with your grit and determination. This would have been a so boring without you.

To **Märta Nummelin** for bringing art and color to my life.

The **Karolinska Institute** for supporting me with the CSTP. To the individuals at the Karolinska Institute, **KIDS**, **Danderyd Hospital** and the **Department of Orthopedics** at Danderyd Hospital who have helped me to navigate the machinery where I would otherwise be stuck indefinitely.

The **Swedish Fracture Registry**, **Bild- och Funktionstjänsten** (in particular **Johan Söderström** – who got us unstuck), **Region Gotland**, **Region Stockholm** and the **hospital** that have shared their data, **SECTRA** and all the **patients** in Sweden and Australia who enabled this research, and the progress and betterment of medicine for all.

Ann Hovland-Tänneryd who has been very supportive and helped speed this work along, in a way that has allowed me to keep my health and sanity.

To all **my colleagues**, I feel privileged working with you all, for your skill and your wonderful friendships.

The **countless scientists, authors** and **contents creators** whom I have sourced for knowledge and understanding. You do not know me, but I know so many of you.

To **my family**. Without you I would never have gotten through this.

To my **friends** and **loved ones** throughout the years. Thank you! You are too many to mention by name. Some of you have been inspirations and mentors long after are paths crossed.

My **grappling buddies**. You keep me grounded and pounded. My **inspirations and higher powers**. You keep me flying.

Thank you to my opponent **Hans Berg** and the members of examination committee **Li Felländer-Tsai**, **Maria Cöster**, and **Mattias Rantalainen**.

All those I have crossed paths with along the way, who have been less supportive, I thank you too. You have helped me reflect and learn. You have made me wiser and more patient. Through your resistance, I have grown.

9 Declaration about the use of generative AI

The author has authored the comprehensive summary/"kappa" and all papers without generative AI. The author has used generative AI for the following purposes:

- ChatGPT models 3, 4, 4o, 4o mini, and Google Bard, assisted with language, such as proofreading self-authored texts, detecting inconsistencies in passages, and helping with formulations. They were also used to screen, clarify, understand, and translate texts and journal papers.
- ChatGPT models 4 and 4o and Github Copilot have been used to generate, improve, and analyze software code.
- Spell correction software (Grammarly) has been used to proof the text.

10 References

1. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthopaedica*. 2017 Nov 2;88(6):581–6.
2. Olczak J, Emilson F, Razavian A, Antonsson T, Stark A, Gordon M. Ankle fracture classification using deep learning: automating detailed AO Foundation/Orthopedic Trauma Association (AO/OTA) 2018 malleolar fracture identification reaches a high degree of correct classification. *Acta Orthopaedica*. 2021 Jan 2;92(1):102–8.
3. Olczak J, Prijs J, IJpma F, Wallin F, Akbarian E, Doornberg J, et al. External validation of an artificial intelligence multi-label deep learning model capable of ankle fracture classification. *BMC Musculoskelet Disord*. 2024 Oct 4;25(1):788.
4. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. 1998 Nov;86(11):2278–324.
5. Krizhevsky A. Learning Multiple Layers of Features from Tiny Images. 2009 Apr 8;60.
6. Jia Deng, Wei Dong, Socher R, Li-Jia Li, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE; 2009* [cited 2016 Sep 11]. p. 248–55. Available from: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5206848>
7. ImageNet. ImageNet [Internet]. ImageNet. [cited 2020 Dec 12]. Available from: <http://www.image-net.org/>
8. Beyer L, Hénaff OJ, Kolesnikov A, Zhai X, Oord A van den. Are we done with ImageNet? arXiv:200607159 [cs] [Internet]. 2020 Jun 12 [cited 2021 Dec 7]; Available from: <http://arxiv.org/abs/2006.07159>
9. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. arXiv:190107031 [cs, eess] [Internet]. 2019 Jan 21 [cited 2020 Oct 5]; Available from: <http://arxiv.org/abs/1901.07031>
10. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 Jul;3462–71.
11. Rajpurkar P, Irvin J, Bagul A, Ding D, Duan T, Mehta H, et al. MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. arXiv:171206957 [physics] [Internet]. 2018 May 22 [cited 2020 Dec 10]; Available from: <http://arxiv.org/abs/1712.06957>
12. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017 Feb;542(7639):115.
13. Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatol*. 2018 Nov 1;154(11):1247.

14. Kamulegeya LH, Okello M, Bwanika JM, Musinguzi D, Lubega W, Rusoke D, et al. Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning. *bioRxiv*. 2019 Oct 31;826057.
15. About DICOM- Overview [Internet]. DICOM. [cited 2024 Aug 8]. Available from: <https://www.dicomstandard.org/about>
16. Bidgood WD, Horii SC, Prior FW, Van Syckle DE. Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging. *J Am Med Inform Assoc*. 1997;4(3):199–212.
17. Graham RNJ, Perriss RW, Scarsbrook AF. DICOM demystified: a review of digital file formats and their use in radiological practice. *Clin Radiol*. 2005 Nov;60(11):1133–40.
18. Varma DR. Managing DICOM images: Tips and tricks for the radiologist. *Indian J Radiol Imaging*. 2012;22(1):4–13.
19. Lind A, Akbarian E, Olsson S, Näsell H, Sköldenberg O, Razavian AS, et al. Artificial intelligence for the classification of fractures around the knee in adults according to the 2018 AO/OTA classification system. *PLOS ONE*. 2021 Apr 1;16(4):e0248809.
20. Olsson S, Akbarian E, Lind A, Razavian AS, Gordon M. Automating classification of osteoarthritis according to Kellgren-Lawrence in the knee using deep learning in an unfiltered adult population. *BMC Musculoskeletal Disorders*. 2021 Oct 2;22(1):844.
21. Akbarian E, Mohammadi M, Tiala E, Ljungberg O, Razavian AS, Magnéli M, et al. Development and validation of an artificial intelligence model for the classification of hip fractures using the AO-OTA framework. *Acta Orthopaedica*. 2024 Jun 18;95:340–7.
22. Magnéli M, Axenhus M, Fagrell J, Ling P, Gislén J, Demir Y, et al. Artificial intelligence can be used in the identification and classification of shoulder osteoarthritis and avascular necrosis on plain radiographs: a training study of 7,139 radiograph sets. *Acta Orthopaedica*. 2024 Jun 17;95:319–24.
23. SFR. Svenska Frakturregistret [Internet]. 2020 [cited 2020 Nov 15]. Available from: <https://sfr.registercentrum.se/>
24. Marsh J, Slongo T, Agel J, Broderick J, Creevey W, DeCoster T, et al. Fracture and Dislocation Classification Compendium - 2007. *Journal of Orthopaedic Trauma* [Internet]. 2007 Nov 1 [cited 2019 Apr 16];21(10). Available from: insights.ovid.com
25. Juto H, Möller M, Wennergren D, Edin K, Apelqvist I, Morberg P. Substantial accuracy of fracture classification in the Swedish Fracture Register: Evaluation of AO/OTA-classification in 152 ankle fractures. *Injury*. 2016 Nov 1;47(11):2579–83.
26. Wennergren D, Stjernström S, Möller M, Sundfeldt M, Ekholm C. Validity of humerus fracture classification in the Swedish fracture register. *BMC Musculoskeletal Disord*. 2017 Jun 10;18(1):251.

27. Knutsson SB, Wennergren D, Bojan A, Ekelund J, Möller M. Femoral fracture classification in the Swedish Fracture Register – a validity study. *BMC Musculoskeletal Disorders*. 2019 May 8;20(1):197.
28. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977 Mar;33(1):159–74.
29. Sundkvist J, Hulenvik P, Schmidt V, Jolbäck P, Sundfeldt M, Fischer P, et al. Basicervical femoral neck fractures: an observational study derived from the Swedish Fracture Register. *Acta Orthopaedica*. 2024 May 22;95:250–5.
30. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. *Ann Med*. 2001 Jul;33(5):337–43.
31. Swiontkowski MF, Engelberg R, Martin DP, Agel J. Short musculoskeletal function assessment questionnaire: validity, reliability, and responsiveness. *J Bone Joint Surg Am*. 1999 Sep;81(9):1245–60.
32. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011 Dec;20(10):1727–36.
33. Oppe M, Devlin NJ, van Hout B, Krabbe PFM, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value Health*. 2014 Jun;17(4):445–53.
34. Ramos-Goñi JM, Pinto-Prades JL, Oppe M, Cabasés JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and Modeling of EQ-5D-5L Health States Using a Hybrid Approach. *Med Care*. 2017;55(7):e51–8.
35. Ramos-Goñi JM, Oppe M, Slaap B, Busschbach JJV, Stolk E. Quality Control Process for EQ-5D-5L Valuation Studies. *Value in Health*. 2017 Mar;20(3):466–73.
36. Ramos-Goñi JM, Craig B, Oppe M, van Hout B. Combining continuous and dichotomous responses in a hybrid model. *EuroQol Research Foundation*; 2016 Dec. (EuroQol Working Paper Series). Report No.: Number 16002.
37. Stolk E, Ludwig K, Rand K, Hout B van, Ramos-Goñi JM. Overview, Update, and Lessons Learned From the International EQ-5D-5L Valuation Work: Version 2 of the EQ-5D-5L Valuation Protocol. *Value in Health*. 2019 Jan 1;22(1):23–30.
38. Feng Y, Devlin NJ, Shah KK, Mulhern B, van Hout B. New methods for modelling EQ-5D-5L value sets: An application to English data. *Health Economics*. 2018 Jan;27(1):23–38.
39. Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Qual Life Res*. 2013 Sep;22(7):1717–27.
40. van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health*. 2012;15(5):708–15.

41. Engelberg R, Martin DP, Agel J, Obremsky W, Coronado G, Swiontkowski MF. Musculoskeletal function assessment instrument: Criterion and construct validity. *Journal of Orthopaedic Research*. 1996;14(2):182–92.
42. Martin DP, Engelberg R, Agel J, Snapp D, Swiontkowski MF. Development of a musculoskeletal extremity health status instrument: the Musculoskeletal Function Assessment instrument. *J Orthop Res*. 1996 Mar;14(2):173–81.
43. Juto H, Gärtner Nilsson M, Möller M, Wennergren D, Morberg P. Evaluating non-responders of a survey in the Swedish fracture register: no indication of different functional result. *BMC Musculoskelet Disord*. 2017 Dec;18(1):278.
44. Kodama N, Imai S, Matsue Y. A Simple Method for Choosing Treatment of Distal Radius Fractures. *The Journal of Hand Surgery*. 2013 Oct;38(10):1896–905.
45. Neuhaus V, Bot AG, Guitton TG, Ring DC. Influence of surgeon, patient and radiographic factors on distal radius fracture treatment. *J Hand Surg Eur Vol*. 2015 Oct 1;40(8):796–804.
46. Audigé L, Bhandari M, Kellam J. How reliable are reliability studies of fracture classifications? A systematic review of their methodologies. *Acta Orthop*. 2004 Jan 1;75(2):184–94.
47. Shehovych A, Salar O, Meyer C, Ford D. Adult distal radius fractures classification systems: essential clinical knowledge or abstract memory testing? *Ann R Coll Surg Engl*. 2016 Nov;98(8):525–31.
48. Gilbert F, Eden L, Meffert R, Konietzschke F, Lotz J, Bauer L, et al. Intra- and interobserver reliability of glenoid fracture classifications by Ideberg, Euler and AO. *BMC Musculoskelet Disord* [Internet]. 2018 Mar 27 [cited 2020 Nov 29];19. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5870213/>
49. Lauge-Hansen N. Fractures of the ankle. IV. Clinical use of genetic roentgen diagnosis and genetic reduction. *AMA Arch Surg*. 1952 Apr;64(4):488–500.
50. Lauge-Hansen N. Fractures of the ankle. III. Genetic roentgenologic diagnosis of fractures of the ankle. *Am J Roentgenol Radium Ther Nucl Med*. 1954 Mar;71(3):456–71.
51. Okanobo H, Khurana B, Sheehan S, Duran-Mendicuti A, Arianjam A, Ledbetter S. Simplified Diagnostic Algorithm for Lauge-Hansen Classification of Ankle Injuries. *Radiographics : a review publication of the Radiological Society of North America, Inc*. 2012 Mar 1;32:E71-84.
52. Tartaglione JP, Rosenbaum AJ, Abousayed M, DiPrea JA. Classifications in Brief: Lauge-Hansen Classification of Ankle Fractures. *Clin Orthop Relat Res*. 2015 Oct;473(10):3323–8.
53. AO Foundation. AO Foundation: Transforming Surgery–Changing Lives [Internet]. 2020 [cited 2020 Nov 29]. Available from: <https://www.aofoundation.org/>
54. Association Committee for Coding and Classification. Fracture and dislocation compendium. Orthopaedic Trauma Association Committee for Coding and Classification. *J Orthop Trauma*. 1996;10 Suppl 1:v–ix, 1–154.

55. Meinberg EG, Agel J, Roberts CS, Karam MD, Kellam JF. Fracture and Dislocation Classification Compendium-2018. *J Orthop Trauma*. 2018;32 Suppl 1:S1–170.
56. Vaccaro AR, Oner C, Kepler CK, Dvorak M, Schnake K, Bellabarba C, et al. AOSpine Thoracolumbar Spine Injury Classification System: Fracture Description, Neurological Status, and Key Modifiers. *Spine*. 2013 Nov 1;38(23):2028–37.
57. Vaccaro AR, Koerner JD, Radcliff KE, Oner FC, Reinhold M, Schnake KJ, et al. AOSpine subaxial cervical spine injury classification system. *Eur Spine J*. 2016 Jul 1;25(7):2173–84.
58. Slongo T, Audigé L, Schlickewei W, Clavert JM, Hunter J. Development and Validation of the AO Pediatric Comprehensive Classification of Long Bone Fractures by the Pediatric Expert Group of the AO Foundation in Collaboration With AO Clinical Investigation and Documentation and the International Association for Pediatric Traumatology. *Journal of Pediatric Orthopaedics*. 2006 Feb;26(1):43–9.
59. AO Pediatric Comprehensive Classification of Long Bone Fractures (PCCF). *Journal of Orthopaedic Trauma* [Internet]. 2018 Jan 1 [cited 2019 Apr 16];32. Available from: insights.ovid.com
60. Svenska Frakturregistret [Internet]. [cited 2023 Nov 24]. Available from: <https://sfr.registercentrum.se/>
61. Audigé L, Bhandari M, Hanson B, Kellam J. A concept for the validation of fracture classifications. *J Orthop Trauma*. 2005 Jul;19(6):401–6.
62. Burstein AH. Fracture classification systems: do they work and are they useful? *JBJS*. 1993 Dec;75(12):1743.
63. Carofino BC, Leopold SS. Classifications in Brief: The Neer Classification for Proximal Humerus Fractures. *Clin Orthop Relat Res*. 2013 Jan;471(1):39–43.
64. Siebenrock KA, Gerber C. The reproducibility of classification of fractures of the proximal end of the humerus. *J Bone Joint Surg Am*. 1993 Dec;75(12):1751–5.
65. Sidor ML, Zuckerman JD, Lyon T, Koval K, Cuomo F, Schoenberg N. The Neer classification system for proximal humeral fractures. An assessment of interobserver reliability and intraobserver reproducibility. *J Bone Joint Surg Am*. 1993 Dec;75(12):1745–50.
66. Marongiu G, Leinardi L, Congia S, Frigau L, Mola F, Capone A. Reliability and reproducibility of the new AO/OTA 2018 classification system for proximal humeral fractures: a comparison of three different classification systems. *J Orthop Traumatol*. 2020 Dec;21:4.
67. Fonseca L, Nunes I, Nogueira R, Martins G, Mesencio A, Kobata S. Reproducibility of the Lauge-Hansen, Danis-Weber, and AO classifications for ankle fractures. *Revista Brasileira de Ortopedia (English Edition)*. 2017 Dec 1;53.
68. Craig WL, Dirschl DR. Effects of binary decision making on the classification of fractures of the ankle. *J Orthop Trauma*. 1998 May;12(4):280–3.

69. Budny AM, Young BA. Analysis of Radiographic Classifications for Rotational Ankle Fractures. *Clinics in Podiatric Medicine and Surgery*. 2008 Apr 1;25(2):139–52.
70. Lindsjö U. Classification of Ankle Fractures: The Lauge-Hansen or AO System? *Clinical orthopaedics and related research*. 1985 Oct;199:12–5.
71. Thomsen NO, Overgaard S, Olsen LH, Hansen H, Nielsen ST. Observer variation in the radiographic classification of ankle fractures. *J Bone Joint Surg Br*. 1991 Jul;73(4):676–8.
72. Nielsen JØ, Dons-Jensen H, Sørensen HT. Lauge-Hansen classification of malleolar fractures: An assessment of the reproducibility in 118 cases. *Acta Orthopaedica Scandinavica*. 1990 Jan;61(5):385–7.
73. Gardner MJ, Demetrakopoulos D, Briggs SM, Helfet DL, Lorch DG. The Ability of the Lauge-Hansen Classification to Predict Ligament Injury and Mechanism in Ankle Fractures: An MRI Study: *Journal of Orthopaedic Trauma*. 2006 Apr;20(4):267–72.
74. Boszczyk A, Fudalej M, Kwapisz S, Błoński M, Kiciński M, Kordasiewicz B, et al. X-ray features to predict ankle fracture mechanism. *Forensic Science International*. 2018 Oct 1;291:185–92.
75. Kwon JY, Chacko AT, Kadzielski JJ, Appleton PT, Rodriguez EK. A Novel Methodology for the Study of Injury Mechanism: Ankle Fracture Analysis Using Injury Videos Posted on YouTube.com. *Journal of Orthopaedic Trauma*. 2010 Aug;24(8):477.
76. Rodriguez EK, Kwon JY, Chacko AT, Kadzielski JJ, Lindsay H, Appleton PT. An Update on Assessing the Validity of the Lauge Hansen Classification System for In-vivo Ankle Fractures Using YouTube videos of Accidentally Sustained Ankle Fractures as a Tool for the Dynamic Assessment of Injury. *The Harvard Orthopaedic Journal*. 2012 Dec;14:40–3.
77. Rodriguez EK, Kwon JY, Herder LM, Appleton PT. Correlation of AO and Lauge-Hansen Classification Systems for Ankle Fractures to the Mechanism of Injury. *Foot Ankle Int*. 2013 Nov;34(11):1516–20.
78. Patton BK, Orfield NJ, Clements JR. Does the Lauge-Hansen Injury Mechanism Predict Posterior Malleolar Fracture Morphology? *The Journal of Foot and Ankle Surgery*. 2022 Nov 1;61(6):1251–4.
79. Michelson J, Solocoff D, Waldman B, Kendell K, Ahn U. Ankle fractures. The Lauge-Hansen classification revisited. *Clin Orthop Relat Res*. 1997 Dec;(345):198–205.
80. Haraguchi N, Armingier RS. A New Interpretation of the Mechanism of Ankle Fracture : *JBJS. J Bone Joint Surg Am*. 2009 Apr 1;91:821–9.
81. Glen LZQ, Wong JYS, Tay WX, Li TP, Phua SKA, Manohara R, et al. Weber Ankle Fracture Classification System Yields Greatest Interobserver and Intraobserver Reliability Over AO/OTA and Lauge-Hansen Classification Systems Under Time

Constraints in an Asian Population. *The Journal of Foot and Ankle Surgery*. 2023 May 1;62(3):505–10.

82. Harper MC. Ankle Fracture Classification Systems: A Case for Integration of the Lauge-Hansen and AO-Danis-Weber Schemes. *Foot & Ankle*. 1992 Sep 1;13(7):404–7.
83. Chen D wei, Li B, Yang Y feng, Yu G rong. AO and Lauge-Hansen Classification Systems for Ankle Fractures. *Foot Ankle Int*. 2013 Dec;34(12):1750–1750.
84. Rydberg EM, Zorko T, Sundfeldt M, Möller M, Wennergren D. Classification and treatment of lateral malleolar fractures - a single-center analysis of 439 ankle fractures using the Swedish Fracture Register. *BMC Musculoskelet Disord*. 2020 Aug 5;21(1):521.
85. Manning C. Artificial Intelligence Definitions [Internet]. 2020 [cited 2020 Nov 26]. Available from: <https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf>
86. Olczak J, Pavlopoulos J, Prijs J, Ijpma FFA, Doornberg JN, Lundström C, et al. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. *Acta Orthop*. 2021 May 14;1–13.
87. Pandis N, Fedorowicz Z. The International EQUATOR Network: enhancing the quality and transparency of health care research. *J Appl Oral Sci*. 2011;19(5):0.
88. Liu X, Rivera SC, Faes L, Keane PA, Moher D, Calvert M, et al. CONSORT-AI and SPIRIT-AI: New Reporting Guidelines for Clinical Trials and Trial Protocols for Artificial Intelligence Interventions. *Invest Ophthalmol Vis Sci*. 2020 Jun 10;61(7):1617–1617.
89. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*. 2020 Sep;26(9):1364–74.
90. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ, Ashrafian H, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *The Lancet Digital Health*. 2020 Oct 1;2(10):e549–60.
91. Cybenko G. Approximation by superpositions of a sigmoidal function. *Math Control Signal Systems*. 1989 Dec 1;2(4):303–14.
92. Hornik K, Stinchcombe M, White H. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*. 1989;2:359–66.
93. Hornik K. Approximation capabilities of multilayer feedforward networks. *Neural Networks*. 1991 Jan 1;4(2):251–7.
94. Guilhoto LF. An Overview Of Artificial Neural Networks for Mathematicians. :25.

95. Leshno M, Lin VYa, Pinkus A, Schocken S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*. 1993 Jan;6(6):861–7.
96. Kidger P, Lyons T. Universal Approximation with Deep Narrow Networks [Internet]. arXiv; 2020 [cited 2024 Jul 2]. Available from: <http://arxiv.org/abs/1905.08539>
97. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs] [Internet]. 2021 Jun 3 [cited 2021 Dec 7]; Available from: <http://arxiv.org/abs/2010.11929>
98. Tanzi L, Audisio A, Cirrincione G, Aprato A, Vezzetti E. Vision Transformer for femur fracture classification. arXiv:2108.03414 [cs] [Internet]. 2021 Oct 26 [cited 2022 Jan 6]; Available from: <http://arxiv.org/abs/2108.03414>
99. Bhojanapalli S, Chakrabarti A, Glasner D, Li D, Unterthiner T, Veit A. Understanding Robustness of Transformers for Image Classification. arXiv:2103.14586 [cs] [Internet]. 2021 Oct 8 [cited 2021 Dec 7]; Available from: <http://arxiv.org/abs/2103.14586>
100. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in Vision: A Survey. arXiv:2101.01169 [cs] [Internet]. 2021 Oct 3 [cited 2021 Dec 1]; Available from: <http://arxiv.org/abs/2101.01169>
101. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. *Advances in Neural Information Processing Systems 25* [Internet]. Curran Associates, Inc.; 2012 [cited 2020 Jul 2]. p. 1097–105. Available from: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
102. Lin M, Chen Q, Yan S. Network In Network. arXiv:1312.4400 [cs] [Internet]. 2014 Mar 4 [cited 2020 Nov 28]; Available from: <http://arxiv.org/abs/1312.4400>
103. Chatfield K, Simonyan K, Vedaldi A, Zisserman A. Return of the Devil in the Details: Delving Deep into Convolutional Nets. In: *Proceedings of the British Machine Vision Conference 2014* [Internet]. Nottingham: British Machine Vision Association; 2014 [cited 2020 Nov 28]. p. 6.1-6.12. Available from: <http://www.bmva.org/bmvc/2014/papers/paper054/index.html>
104. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going Deeper with Convolutions. arXiv:1409.4842 [cs] [Internet]. 2014 Sep 16 [cited 2020 Jul 3]; Available from: <http://arxiv.org/abs/1409.4842>
105. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. arXiv:1512.00567 [cs] [Internet]. 2015 Dec 1 [cited 2016 Sep 28]; Available from: <http://arxiv.org/abs/1512.00567>
106. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv:1602.07261 [cs] [Internet]. 2016 Aug 23 [cited 2020 Jul 2]; Available from: <http://arxiv.org/abs/1602.07261>

107. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs] [Internet]. 2015 Apr 10 [cited 2020 Dec 11]; Available from: <http://arxiv.org/abs/1409.1556>
108. Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Computation*. 1997 Nov 1;9(8):1735–80.
109. Srivastava RK, Greff K, Schmidhuber J. Highway Networks. arXiv:1505.00387 [cs] [Internet]. 2015 Nov 3 [cited 2020 Dec 10]; Available from: <http://arxiv.org/abs/1505.00387>
110. He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In 2015 [cited 2016 Aug 19]. p. 1026–34. Available from: http://www.cv-foundation.org/openaccess/content_iccv_2015/html/He_Delving_Deep_into_ICCV_2015_paper.html
111. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. arXiv:1608.06993 [cs] [Internet]. 2016 Aug 24 [cited 2019 Mar 14]; Available from: <http://arxiv.org/abs/1608.06993>
112. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. *Radiographics*. 2017 Apr;37(2):505–15.
113. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018 Aug;18(8):500–10.
114. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*. 2019 Oct 1;1(6):e271–97.
115. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Calster BV. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*. 2019 Jun 1;110:12–22.
116. Oosterhoff JHF, Gravesteijn BY, Karhade AV, Jaarsma RL, Kerkhoffs GMMJ, Ring D, et al. Feasibility of Machine Learning and Logistic Regression Algorithms to Predict Outcome in Orthopaedic Trauma Surgery. *JBJS*. 2021 Dec 17;10.2106/JBJS.21.00341.
117. Cary MP, Zhuang F, Draelos RL, Pan W, Amarasekara S, Douthit BJ, et al. Machine Learning Algorithms to Predict Mortality and Allocate Palliative Care for Older Patients With Hip Fracture. *J Am Med Dir Assoc*. 2021 Feb;22(2):291–6.
118. Cabitza F, Locoro A, Banfi G. Machine Learning in Orthopedics: A Literature Review. *Front Bioeng Biotechnol* [Internet]. 2018 Jun 27 [cited 2019 Oct 29];6. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6030383/>
119. Kalanovic VD, Popovic D, Skaug NT. Feedback error learning neural network for trans-femoral prosthesis. *IEEE Transactions on Rehabilitation Engineering*. 2000 Mar;8(1):71–80.

120. Pogorelc B, Gams M. Diagnosing health problems from gait patterns of elderly. *Annu Int Conf IEEE Eng Med Biol Soc.* 2010;2010:2238–41.
121. Nair SS, French RM, Laroche D, Thomas E. The Application of Machine Learning Algorithms to the Analysis of Electromyographic Patterns From Arthritic Patients. *IEEE Transactions on Neural Systems and Rehabilitation Engineering.* 2010 Apr;18(2):174–84.
122. Prasoon A, Petersen K, Igel C, Lauze F, Dam E, Nielsen M. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. *Med Image Comput Comput Assist Interv.* 2013;16(Pt 2):246–53.
123. Thong W, Parent S, Wu J, Aubin CE, Labelle H, Kadoury S. Three-dimensional morphology study of surgical adolescent idiopathic scoliosis patient from encoded geometric models. *Eur Spine J.* 2016 Oct 1;25(10):3104–13.
124. Abidin AZ, Deng B, DSouza AM, Nagarajan MB, Coan P, Wismüller A. Deep transfer learning for characterizing chondrocyte patterns in phase contrast X-Ray computed tomography images of the human patellar cartilage. *Computers in Biology and Medicine.* 2018 Apr 1;95:24–33.
125. Shim E, Kim JY, Yoon JP, Ki SY, Lho T, Kim Y, et al. Automated rotator cuff tear classification using 3D convolutional neural network. *Scientific Reports.* 2020 Sep 24;10(1):15632.
126. Al-Helo S, Alomari RS, Ghosh S, Chaudhary V, Dhillon G, Al-Zoubi MB, et al. Compression fracture diagnosis in lumbar: a clinical CAD system. *Int J CARS.* 2013 May 1;8(3):461–9.
127. Dijkstra H, van de Kuit A, de Groot T, Canta O, Groot OQ, Oosterhoff JH, et al. Systematic review of machine-learning models in orthopaedic trauma. *Bone Jt Open.* 2024 Jan 16;5(1):9–19.
128. Machine Learning Consortium, on behalf of the SPRINT and FLOW Investigators. A Machine Learning Algorithm to Identify Patients with Tibial Shaft Fractures at Risk for Infection After Operative Treatment. *J Bone Joint Surg Am.* 2021 Mar 17;103(6):532–40.
129. Lin CC, Ou YK, Chen SH, Liu YC, Lin J. Comparison of artificial neural network and logistic regression models for predicting mortality in elderly patients with hip fracture. *Injury.* 2010 Aug 1;41(8):869–73.
130. Liu F, Liu C, Tang X, Gong D, Zhu J, Zhang X. Predictive Value of Machine Learning Models in Postoperative Mortality of Older Adults Patients with Hip Fracture: A Systematic Review and Meta-analysis. *Archives of Gerontology and Geriatrics.* 2023 Dec 1;115:105120.
131. DeBaun MR, Chavez G, Fithian A, Oladeji K, Van Rysselberghe N, Goodnough LH, et al. Artificial Neural Networks Predict 30-Day Mortality After Hip Fracture: Insights From Machine Learning. *JAAOS - Journal of the American Academy of Orthopaedic Surgeons.* 2021 Nov 15;29(22):977.

132. Chen CY, Chen YF, Chen HY, Hung CT, Shi HY. Artificial Neural Network and Cox Regression Models for Predicting Mortality after Hip Fracture Surgery: A Population-Based Comparison. *Medicina (Kaunas)*. 2020 May 19;56(5):243.
133. Cao W, Mirjalili V, Raschka S. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*. 2020 Dec 1;140:325–31.
134. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: Mapping the debate. *Big Data & Society*. 2016 Dec 1;3(2):2053951716679679.
135. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*. 2018 Nov 6;15(11):e1002683.
136. Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*. 2019 Apr 30;2(1):1–10.
137. Beil M, Proft I, van Heerden D, Sviri S, van Heerden PV. Ethical considerations about artificial intelligence for prognostication in intensive care. *Intensive Care Med Exp* [Internet]. 2019 Dec 10 [cited 2020 Oct 17];7. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6904702/>
138. Lupton M. Some ethical and legal consequences of the application of artificial intelligence in the field of medicine. *Trends Med* [Internet]. 2018 [cited 2020 Oct 17];18(4). Available from: <https://www.oatext.com/some-ethical-and-legal-consequences-of-the-application-of-artificial-intelligence-in-the-field-of-medicine.php>
139. Bohannon M. Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions [Internet]. *Forbes*. [cited 2024 Aug 10]. Available from: <https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/>
140. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019 May;1(5):206–15.
141. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials | *The BMJ* [Internet]. [cited 2024 Jul 3]. Available from: <https://www.bmj.com/content/346/bmj.e7586.full?ijkey=QpAJnYI57zIwVr3&keytyp=ref>
142. Schulz KF, Altman DG, Moher D, the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC Medicine*. 2010 Mar 24;8(1):18.
143. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med*. 2007 Oct 16;4(10):e296.

144. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015 Jan 6;162(1):55–63.
145. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024 Apr 16;385:e078378.
146. Olczak J. Artificial Intelligence for Understanding Medicine. Machine Learning for Orthopaedic Trauma Radiographs [Master]. [Department of Clinical Sciences, Danderyd Hospital]: Karolinska Institutet; 2017.
147. Olczak J, Gordon M. From Radiologist Report to Image Label: Assessing Latent Dirichlet Allocation in Training Neural Networks for Orthopedic Radiograph Classification [Internet]. arXiv; 2024 [cited 2024 Aug 27]. Available from: <http://arxiv.org/abs/2408.13284>
148. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks [Internet]. arXiv; 2018 [cited 2023 Oct 25]. Available from: <http://arxiv.org/abs/1608.06993>
149. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs] [Internet]. 2015 Dec 10 [cited 2021 Dec 7]; Available from: <http://arxiv.org/abs/1512.03385>
150. Enocson A, Wolf O. Pipkin fractures: epidemiology and outcome. *Eur J Trauma Emerg Surg*. 2022;48(5):4113–8.
151. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One* [Internet]. 2015 Mar 4 [cited 2020 Aug 6];10(3). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349800/>
152. Efron B, Tibshirani R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy on JSTOR. *Stat Sci*. 1986 Feb;1(1):54–77.
153. Efron B. Bootstrap Methods: Another Look at the Jackknife. In: Kotz S, Johnson NL, editors. *Breakthroughs in Statistics: Methodology and Distribution* [Internet]. New York, NY: Springer; 1992 [cited 2022 Jun 3]. p. 569–93. (Springer Series in Statistics). Available from: https://doi.org/10.1007/978-1-4612-4380-9_41
154. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 2003;3(Jan):993–1022.
155. Blei DM, Lafferty JD. A Correlated Topic Model of Science. *The Annals of Applied Statistics*. 2007 Jun 1;1(1):17–35.
156. Blei DM. Probabilistic topic models. *Communications of the ACM*. 2012 Apr 1;55(4):77–84.
157. Ponweiser M. Latent Dirichlet Allocation in R [Internet]. 2012 [cited 2016 Aug 21]. Available from: <http://epub.wu.ac.at/3558/>

158. Joeris A, Lutz N, Blumenthal A, Slongo T, Audigé L. The AO Pediatric Comprehensive Classification of Long Bone Fractures (PCCF). *Acta Orthop*. 2017 Apr;88(2):123–8.
159. Lin M, Chen Q, Yan S. Network In Network [Internet]. arXiv; 2014 [cited 2024 Jun 22]. Available from: <http://arxiv.org/abs/1312.4400>
160. Guillaumin M, Verbeek J, Schmid C. Multimodal semi-supervised learning for image classification. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition [Internet]. San Francisco, CA, USA: IEEE; 2010 [cited 2020 Jan 14]. p. 902–9. Available from: <http://ieeexplore.ieee.org/document/5540120/>
161. Izmailov P, Podoprikin D, Garipov T, Vetrov D, Wilson AG. Averaging Weights Leads to Wider Optima and Better Generalization. arXiv:180305407 [cs, stat] [Internet]. 2019 Feb 25 [cited 2019 Nov 16]; Available from: <http://arxiv.org/abs/1803.05407>
162. Culotta A, McCallum A. Reducing Labeling Effort for Structured Prediction Tasks: [Internet]. Fort Belvoir, VA: Defense Technical Information Center; 2005 Jan [cited 2024 Sep 5]. Available from: <http://www.dtic.mil/docs/citations/ADA440382>
163. Settles B. Active Learning Literature Survey [Internet]. University of Wisconsin-Madison Department of Computer Sciences; 2009 [cited 2024 Sep 5]. Available from: <https://minds.wisconsin.edu/handle/1793/60660>
164. Ramirez-Loaiza ME, Sharma M, Kumar G, Bilgic M. Active learning: an empirical study of common baselines. *Data Min Knowl Discov*. 2017 Mar 1;31(2):287–313.
165. Liu S, Li X. Understanding Uncertainty Sampling [Internet]. arXiv.org. 2023 [cited 2024 Sep 5]. Available from: <https://arxiv.org/abs/2307.02719v3>
166. Raschka-research-group/coral-pytorch [Internet]. Raschka Research Group; 2024 [cited 2024 Jun 22]. Available from: <https://github.com/Raschka-research-group/coral-pytorch>
167. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection [Internet]. arXiv; 2018 [cited 2024 Sep 17]. Available from: <http://arxiv.org/abs/1708.02002>
168. Sabour S, Frosst N, Hinton GE. Dynamic Routing Between Capsules [Internet]. arXiv; 2017 [cited 2024 Jul 3]. Available from: <http://arxiv.org/abs/1710.09829>
169. Barron JT. A General and Adaptive Robust Loss Function [Internet]. arXiv; 2019 [cited 2024 Apr 15]. Available from: <http://arxiv.org/abs/1701.03077>
170. jonbarron. jonbarron/robust_loss_pytorch [Internet]. 2024 [cited 2024 May 24]. Available from: https://github.com/jonbarron/robust_loss_pytorch
171. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32–5.
172. Aoki K, Misumi J, Kimura T, Zhao W, Xie T. Evaluation of Cutoff Levels for Screening of Gastric Cancer Using Serum Pepsinogens and Distributions of Levels

- of Serum Pepsinogen I, II and of PG I / PG II Ratios in a Gastric Cancer Case-Control Study. *Journal of Epidemiology*. 1997;7(3):143–51.
173. Shapiro DE. The interpretation of diagnostic tests. *Stat Methods Med Res*. 1999 Apr 1;8(2):113–34.
 174. Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*. 2000 May 30;45(1):23–41.
 175. Albert A. *Multivariate Interpretation of Clinical Laboratory Data*. CRC Press; 1987. 332 p.
 176. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*. 1993 Apr 1;39(4):561–77.
 177. Devlin PDN, Parkin D, Janssen B. Analysis of EQ VAS Data. In: *Methods for Analysing and Reporting EQ-5D Data* [Internet] [Internet]. Springer; 2020 [cited 2024 Oct 11]. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK565683/>
 178. Shin HC, Lu L, Kim L, Seff A, Yao J, Summers RM. Interleaved Text/Image Deep Mining on a Very Large-Scale Radiology Database. In 2015 [cited 2016 Aug 19]. p. 1090–9. Available from: http://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Shin_Interleaved_TextImage_Deep_2015_CVPR_paper.html
 179. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*. 2016 May;35(5):1299–312.
 180. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clinical Radiology*. 2018 May 1;73(5):439–45.
 181. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. *Skeletal Radiol*. 2019 Feb 1;48(2):239–44.
 182. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. *arXiv:1711.06504 [cs, stat]* [Internet]. 2017 Nov 17 [cited 2020 May 18]; Available from: <http://arxiv.org/abs/1711.06504>
 183. Nicolaes J, Raeymaeckers S, Robben D, Wilms G, Vandermeulen D, Libanati C, et al. Detection of vertebral fractures in CT using 3D Convolutional Neural Networks. *arXiv:1911.01816 [cs, eess]* [Internet]. 2019 Nov 5 [cited 2020 Oct 9]; Available from: <http://arxiv.org/abs/1911.01816>
 184. Qi Y, Zhao J, Shi Y, Zuo G, Zhang H, Long Y, et al. Ground Truth Annotated Femoral X-Ray Image Dataset and Object Detection Based Method for Fracture Types Classification. *IEEE Access*. 2020;8:189436–44.

185. Wang N, Zeng NN, Zhu W. Sensitivity, Specificity, Accuracy, Associated Confidence Interval And ROC Analysis With Practical SAS Implementations. In 2010. p. 9.
186. Gan K, Xu D, Lin Y, Shen Y, Zhang T, Hu K, et al. Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. *Acta Orthopaedica*. 2019 Jul 4;90(4):394–400.
187. Chung SW, Han SS, Lee JW, Oh KS, Kim NR, Yoon JP, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop*. 2018 Jul 30;89(4):468–73.
188. Heimer J, Thali MJ, Ebert L. Classification based on the presence of skull fractures on curved maximum intensity skull projections by means of deep learning. *Journal of Forensic Radiology and Imaging*. 2018 Sep 1;14:16–20.
189. Choi JW, Cho YJ, Lee S, Lee J, Lee S, Choi YH, et al. Using a Dual-Input Convolutional Neural Network for Automated Detection of Pediatric Supracondylar Fracture on Conventional Radiography. *Investigative Radiology*. 2020 Feb;55(2):101–10.
190. Blüthgen C, Becker AS, Vittoria de Martini I, Meier A, Martini K, Frauenfelder T. Detection and localization of distal radius fractures: Deep learning system versus radiologists. *European Journal of Radiology*. 2020 May 1;126:108925.
191. Dreizin D, Goldmann F, LeBedis C, Boscak A, Dattwyler M, Bodanapally U, et al. An Automated Deep Learning Method for Tile AO/OTA Pelvic Fracture Severity Grading from Trauma whole-Body CT. *J Digit Imaging*. 2021 Feb;34(1):53–65.
192. Jang R, Choi JH, Kim N, Chang JS, Yoon PW, Kim CH. Prediction of osteoporosis from simple hip radiography using deep learning algorithm. *Sci Rep*. 2021 Oct 7;11(1):19997.
193. Kwon YJ (Fred), Toussie D, Finkelstein M, Cedillo MA, Maron SZ, Manna S, et al. Combining Initial Radiographs and Clinical Variables Improves Deep Learning Prognostication in Patients with COVID-19 from the Emergency Department. *Radiol Artif Intell*. 2020 Dec 16;3(2):e200098.
194. Gourdeau D, Potvin O, Biem JH, Cloutier F, Abrougui L, Archambault P, et al. Deep learning of chest X-rays can predict mechanical ventilation outcome in ICU-admitted COVID-19 patients. *Sci Rep*. 2022 Apr 13;12(1):6193.
195. Pease M, Arefan D, Barber J, Yuh E, Puccio A, Hochberger K, et al. Outcome Prediction in Patients with Severe Traumatic Brain Injury Using Deep Learning from Head CT Scans. *Radiology*. 2022 Aug;304(2):385–94.
196. Kim C, Hwang EJ, Choi YR, Choi H, Goo JM, Kim Y, et al. A Deep Learning Model Using Chest Radiographs for Prediction of 30-Day Mortality in Patients With Community-Acquired Pneumonia: Development and External Validation. *American Journal of Roentgenology*. 2023 Nov;221(5):586–98.
197. Shin HJ, Lee EH, Han K, Ryu L, Kim EK. Development of a new prognostic model to predict pneumonia outcome using artificial intelligence-based chest radiograph results. *Sci Rep*. 2024 Jun 22;14(1):14415.

198. Alfraihat A, Samdani AF, Balasubramanian S. Predicting radiographic outcomes of vertebral body tethering in adolescent idiopathic scoliosis patients using machine learning. *PLOS ONE*. 2024 Jan 12;19(1):e0296739.
199. Bar Y, Diamant I, Wolf L, Lieberman S, Konen E, Greenspan H. Chest pathology detection using deep learning with non-medical training. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI) [Internet]. 2015 [cited 2024 Jul 5]. p. 294–7. Available from: <https://ieeexplore.ieee.org/document/7163871>
200. Jamaludin A, Lootus M, Kadir T, Zisserman A, Urban J, Battié MC, et al. ISSLS PRIZE IN BIOENGINEERING SCIENCE 2017: Automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. *Eur Spine J*. 2017 Feb 6;1–10.
201. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *arXiv:160203409 [cs]* [Internet]. 2016 Feb 10 [cited 2016 Aug 19]; Available from: <http://arxiv.org/abs/1602.03409>
202. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med*. 2011 Nov 9;3(108):108ra113.
203. MIT Press. Dataset shift in machine learning. Quiñonero-Candela J, editor. Cambridge, Mass: MIT Press; 2009. 229 p. (Neural information processing series).
204. Oliveira e Carmo L, van den Merkhof A, Olczak J, Gordon M, Jutte PC, Jaarsma RL, et al. An increasing number of convolutional neural networks for fracture recognition and classification in orthopaedics. *Bone Jt Open*. 2021 Oct 20;2(10):879–85.
205. Zhou QQ, Wang J, Tang W, Hu ZC, Xia ZY, Li XS, et al. Automatic Detection and Classification of Rib Fractures on Thoracic CT Using Convolutional Neural Network: Accuracy and Feasibility. *Korean J Radiol*. 2020 Jul;21(7):869–79.
206. Tanzi L, Vezzetti E, Moreno R, Aprato A, Audisio A, Massè A. Hierarchical fracture classification of proximal femur X-Ray images using a multistage Deep Learning approach. *European Journal of Radiology*. 2020 Dec 1;133:109373.
207. Yoon SJ, Hyong Kim T, Joo SB, Eel Oh S. Automatic multi-class intertrochanteric femur fracture detection from CT images based on AO/OTA classification using faster R-CNN-BO method. *J Appl Biomed*. 2020 Dec 14;18(4):97–105.
208. Lee KM, Lee SY, Han CS, Choi SM. Long bone fracture type classification for limited number of CT data with deep learning. In: Proceedings of the 35th Annual ACM Symposium on Applied Computing [Internet]. New York, NY, USA: Association for Computing Machinery; 2020 [cited 2022 Jan 6]. p. 1090–5. Available from: <http://doi.org/10.1145/3341105.3373900>
209. Lim HC, Adie S, Naylor JM, Harris IA. Randomised Trial Support for Orthopaedic Surgical Procedures. *PLOS ONE*. 2014 Jun 13;9(6):e96745.