Predicting the stability of distal radius fractures with a small dataset and machine learning

MSc Thesis Biomedical Engineering - Medical Physics - BM51035

by

Julia E. Wilbers

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended on Wednesday December 14, 2022 at 11:00 AM.

Student number:	4470249	
Project duration:	December 6, 2021 – December 14, 2022	
Thesis committee:	Dr. T. van Walsum,	Erasmuc MC, supervisor
	Dr. F. M. Vos,	TU Delft, BME supervisor
	Dr. Q. Tao,	TU Delft

TU Delft

Erasmus MC
Universitair Medisch Centrum Rotterdam

Predicting the stability of distal radius fractures with a small dataset and machine learning

J.E. Wilbers[1], M.Benmahdjoub[2], R. Su[2], F.M. Vos[1], T. van Walsum[2], A. Cohen[2], and J.W. Colaris[2]

[1]Technical University Delft
[2]Erasmus MC

Abstract

Purpose. Distal radius fractures are common fractures of the wrist. These fractures are often displaced and need reduction, after adequate reduction, the patients will have follow-up X-rays to check if the fracture stays stable. This is important because surgery might be required if the fracture becomes unstable. This can lead to delayed surgery which can worsen the treatment outcome. Therefore, it would be valuable to predict which distal radius fractures are likely to become unstable. Machine learning could help predict the stability of distal radius fractures based on CT. In addition, because there is only a small dataset available, this research also studies the effect of different machine learning methods.

Method. Two different methods were evaluated for the stability prediction of distal radius fractures: traditional machine learning (radiomics method) and a residual network with and without transfer learning (deep learning method). For the radiomics method, a python package called WORC was used, which automatically extracts radiomic features and optimizes machine learning models. For the deep learning method, the backbone of a residual network called Med3D with added layers for classification was used.

Results. The radiomics method combined with augmentation, gave the best results (AUC: 0.64 $\pm$ 0.01). Also, it was found that using an augmented data set in the radiomics method resulted in improved performance. This gives a slight indication that the radiomic method can learn to predict DRF when there would be more data available.

Conclusion. The radiomic method is the most promising method for predicting the stability of distal radius fractures, despite the small difference in performance compared to random guessing and the deep learning method. However, for further research, it is highly recommended to acquire a larger dataset.

I INTRODUCTION

Bone fractures, especially distal radius fractures (DRF) are regularly treated at the Emergency Department (1), (2). In the Netherlands, DRFs have an occurrence of around 45.000 patients yearly (3). Despite year-to-year differences, this number seems to be growing every year (4), (5). DRFs occur in every age group but it is more common in female elderly because of osteoporosis (6). The main causes of DRFs in adolescents are high energetic falls, for example in sports events. In the elderly, the most common causes are low energetic falls from standing or seated positions (7), (8).

A physician will evaluate every identified DRF to ascertain whether it is displaced, using X-ray imaging of the DRF and registered criteria (9). If the fracture is displaced, the fracture will be manually

reduced at the emergency department. After close reduction patients receive a cast and an X-ray to assess whether the reduction was successful. Besides an X-ray, a CT-scan is recommended for complex and / or intra-articular fractures or when the physician is doubting if the fracture is adequately reduced. A CT-scan gives the clinician usually a clearer representation of the fracture and it provides the clinician extra information for the treatment (9), (10). After adequate reduction, the patients will have follow-up X-rays starting every week until the third week after reduction to check if the fracture stays stable. This is important because the treatment plan depends on the stability of the fracture: unstable fractures are generally surgically treated, while stable fractures are conservatively treated with cast (3).

The current treatment strategy can result in delayed surgical treatment for patients with an unstable DRF. For the final bone alignment of the radius and function of the wrist, it is important that the surgical replacement is not too long after the injury (11). Although the best timing is not established yet, it is assumed that surgery should be performed within two weeks after the injury (12). With current strategies, however, the surgery is generally not performed within this timespan. Improved classification of patients with a DRF that are likely to become unstable could overcome the problem of delayed surgical treatment. Since CT-scans contain more information than X-rays, it is preferred to predict this based on CT-scans.

Machine learning could help in the identification of patients with DRFs that are most prone to become unstable. Moreover, Machine learning is already being applied successfully and increasingly in the other fields of medical imaging (12). Specifically, machine learning can be used for example in automatic diagnosis, or as a decision support system (13). In several applications, the use of machine learning has reduced the workload of clinicians (14), which can result in lower emotional stress levels for clinicians. Lowering clinician stress levels has been shown to improve the quality of patient treatment (15), (16). In addition, algorithms can perform consistently and tirelessly and can therefore also potentially improve patient outcomes (17). Furthermore, machine learning algorithms can detect patterns in data that are invisible to the human eye. Ultimately, this can help to prevent invasive treatment or improve diagnosis speed (18), (19).

If a machine learning algorithm could be trained to predict the stability of distal radius fractures based on a CT-scan the patient outcome might improve and the workload of the clinician could be reduced. Previous studies have already shown that machine learning gives good results in bone classification, such as in fracture detection and bone disease classification tasks (20), (21), (22). As such, there is reason to assume that a machine learning system can be developed to support the prediction of distal radius fracture stability.

Although distal radius fractures are common, there are several issues that hinder the collection of a large CT-scan dataset. For example, a specific issue is that patients often receive surgical treatment before it can be established if their fracture would remain stable or not. Using a strategy to overcome the small dataset problem could help create a good-performing prediction algorithm despite the limited dataset. Therefore, the goal of this research is twofold: (1) predict the stability of distal radius fractures on CT-scans using machine learning, and (2) determine which machine learning method is best suited for the limited data problem.

II RELATED WORK

A Application of machine learning in DRFs

The use of machine learning is rising in the diagnosis of DRF, for instance, in the automatic diagnosis of DRF on X-rays. There are already algorithms published that can detect DRF with a similar performance as a radiologist (23), (24), (25). Gan et al. (25) even found that the CNN performed better in distinguishing wrist X-rays with DRFs from normal images.

To the best of our knowledge, no research has been published that focuses on predicting the stability of DRFs or uses CT-scan data for automatic fracture recognition.

B Strategies for limited datasets in machine learning

Limited datasets are one of the biggest challenges in the application of machine learning, especially in 3D medical image analysis (26). Fortunately, there are various strategies to overcome the small dataset problem. For example, traditional machine learning, transfer learning, residual networks, and ensemble learning offer relevant methodologies. These strategies are, according to a systematic review of the literature (Appendix A), likely to be effective in the case of DRF stability prediction.

Machine learning can be defined as the training of algorithms to analyze, learn or make decisions for new data, based on raw data input (27). Machine learning is a broad term that includes both traditional machine learning and deep learning. These methods can be distinguished by how they handle input data. In deep learning, features will be automatically extracted from the input data, then processed, and eventually, an outcome will be produced (28), (29). In contrast, traditional machine learning only optimizes a decision-maker based on the given (derived) input features (30), (31). Therefore, deep learning is less reliant on the user because it optimizes its own feature extraction, at the expense of requiring more data to optimize. Consequently, it is possible that traditional machine learning is better suited to small datasets. Its usefulness for small medical 3D datasets was underlined by several successful applications (32), (33), (34), (35).

One method to overcome the problem of a small dataset is the use of transfer learning. In transfer learning, a model is trained on another dataset (source) before training it on the dataset of interest (target) (36). In this way, the obtained knowledge from training on the source domain is transferred to the target domain and less data is needed to train the network. Often ImageNet (37), a large dataset of natural images is used as a source dataset but the best transferring effect is observed with a source dataset with a similar data distribution to the target dataset (38). The relevance of a similar source domain was emphasized by Raghu et al. (39) who found that for medical imaging tasks transfer learning with ImageNet did not significantly affect the performance of their light weighted model. The literature supports the positive effect of transfer learning: several studies show that the use of transfer learning to overcome the problem of a small 3D medical dataset increases model performance compared with training from scratch (40), (41), (42), (43), (44). Most such studies used a source dataset similar to the target dataset (40), (42), (44). In contrast, Rajpukar et al. (43) used a large dataset of small videos called Kinetics and Clymer et al. (41) used a brain MRI data to pre-train the proposed shoulder classification network. Both studies found an increase in performance when the pre-trained weights were used.

Residual networks are introduced to overcome the overfitting problems in the training of deep neural networks (45). The most important feature of the residual network is the introduction of skip connections. With this skip connection, the output of the previous layer is added to the next layer without modification (46). The skip connection can overcome an overfitting problem called the vanishing gradient: the gradient becomes very small making it hard to train the network (47). This happens in particular when a network uses multiple activation functions that produce small gradients. Skip connections usually give a higher gradient and therefore help in preventing the vanishing gradient problem (48). Therefore, residual networks make it possible to train deeper networks without overtraining, which is especially a problem with small datasets. Satisfactory results with residual networks for 3D medical classification application were found (43), (42), (49). However, it should be noted that these studies did not compare the residual network to other methods so the difference in performance for 3D medical classification purposes remains unknown.

Lastly, ensemble learning could be used to cope with small amounts of data. Ensemble learning approaches combine multiple models to achieve higher prediction performance (50). It is possible to integrate various models in a variety of ways using ensemble learning approaches. For example, different models can be combined by using the mean of their outputs. Also, a single model that is trained on different subsets of the original dataset can be combined to produce one output (51). In the recent literature, several studies found that ensemble learning can be effective for addressing the problem of a small 3D medical dataset (52), (53).

III METHOD

In this study, three different methods were selected (based on a literature review (Appendix A)) for the stability prediction of distal radius fractures: traditional machine learning, a residual network with and without transfer learning, and ensemble learning. In the traditional machine learning method, Radiomic feature extraction will be used to derive image features from the CT-scans, therefore this method is further referred to as the 'radiomics method' (54). The method using a residual network is further referred to as the 'deep learning method'. The implementations of these methods are described in section D and E. Ensemble learning is implemented by averaging the posterior probabilities of the other methods. For all methods, the data pre-processing steps consisted of cropping and segmentation. In addition, the effect of data augmentation was evaluated by performing all methods with and without $10\times$ data augmentation.

A Dataset

The image dataset was created by retrospectively searching the patient databases from the Erasmus MC (EMC) and the University Medical Center Groningen (UMCG). All adult patients (18+) with a primary, adequately reduced DRF treated with a cast, including a CT-scan for the DRF and follow-up X-rays, were eligible for this research (Fig. 1). Data was included from 40 female and 21 male cases. The mean age of the subjects was 48 (95% C 44 - 52) from which 40 were labeled as 'Stable' and 21 were labeled as 'Unstable'.

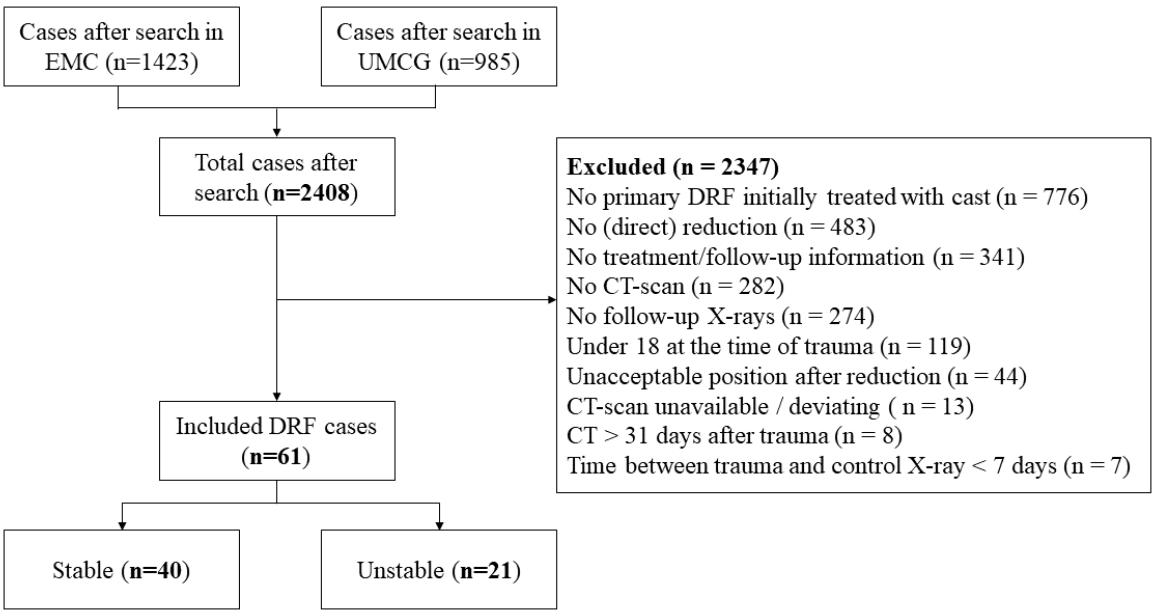


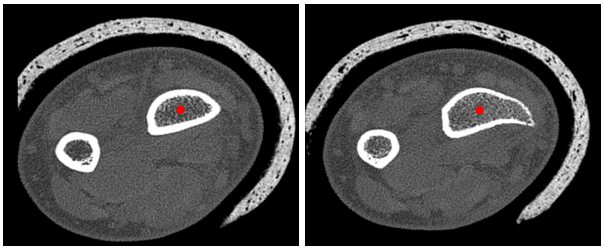
Figure 1: *Inclusion flowchart of DRF cases included in this research. The database search in the EMC and UMGC resulted in 2408 cases. After exclusion, 61 cases, regarding different patients were included.*

B Pre-processing: Cropping

To extract the most relevant features for stability prediction, the distal part of both the radius and ulna was cropped. All CT scans were cropped based on anatomical landmarks in Mevislab (55), a guideline was created (Appendix B) and followed to ensure repeatability. The anatomical landmarks defined a new coordinate system (X_n, Y_n, Z_n) that was in line with the radius and ulna. In the new coordinate system, a rectangle was drawn that cropped out both bones without selecting a large amount of surrounding tissue.

Z_n was formed by connecting two landmarks positioned at the proximal and distal end of the radius (Fig. 2). After defining Z_n the CT-scan was reformatted by resampling the original scan in planes perpendicular to Z_n. In this representation, two additional landmarks at the left and right palmar side of the radius side were placed (Fig. 3a). Similar to the first landmarks, these landmarks were connected and formed a vector: X_n.

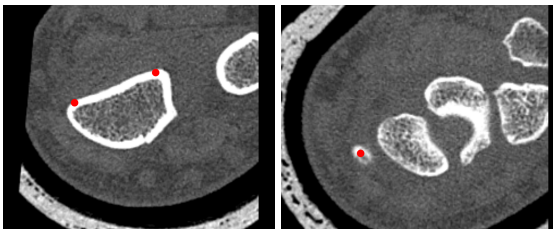
The two vectors Z_n and X_n were respectively vertically and horizontally in line with the radius and ulna. To compose orthogonality X_n could be adjusted to be perpendicular to Z_n and in this way, both vectors formed the basis of the new-coordinate system. To complete the new coordinate system Y_n was defined by placing it perpendicular to both Z_n and X_n (Fig. 4).



(a) (b)

Figure 2: *Visualisation of the landmarks placed for defining the Z_n axis. (a) shows the landmark placed at the proximal side of the radius, (b) shows the landmark placed on the distal side of the radius.*

In addition, a reference point (R_p) was selected to establish the origin of the rectangle. This R_p is an anatomical landmark: the styloid process of the radius, this is the highest point of the radius (Fig. 3b). R_p was chosen because it is easy to recognize on CT; nevertheless, it was not entirely appropriate to be the origin of the rectangle. Therefore, a fixed vector is defined that indicates the displacement from the rectangle's origin to R_p. Lastly, all scans were re-sampled with tri-linear interpolation.



(a) (b)

Figure 3: *Visualisation of the landmarks placed for defining the X_n axis and reference point. (a) shows the landmarks placed at right palmar side of the radius, (b) shows the landmark placed on the reference point (R_p).*

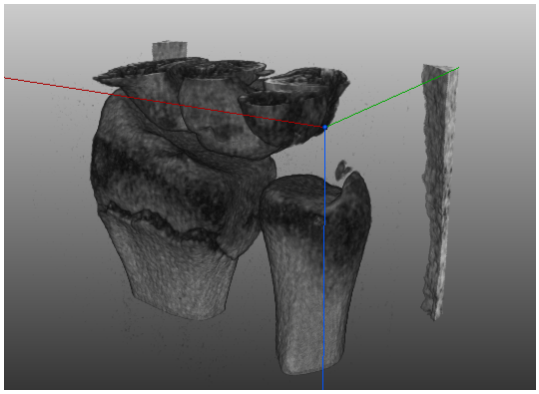


Figure 4: *Three-dimensional visualization of the new coordinate system used for cropping. The red line represents X_n, the green line Y_n, and the blue line Z_n*

C Pre-processing: Segmentation

Providing machine learning methods with a segmentation of the image can help focus on important areas and thus help to improve performance. Moreover, the radiomic method requires a region of interest for radiomic feature calculation that is defined by the segmentation of both the radius and ulna. Segmentation is done semi-automatically in Mevislab (55). First, markers were placed in the first slice roughly surrounding the bones. These markers are automatically connected and form a segmentation. This process is repeated in the last slice and in a random selection of $\sim$ 15 slices in between. The segmentations the in the remaining slices are created by Shape Based Interpolation (56). In this method, the shape of the segmented object is estimated by the use of a specific distance matrix and used for interpolation of the other slices. Finally, the resulting segmentation was visually checked to ensure that the radius and ulna are inside the segmentation.

D Radiomics method

In this method, an open-source python package called WORC is used (57). WORC automatically extracts around 500+ Radiomic features (Supplementary data C) and optimizes a traditional machine learning model (54). For the traditional machine learning model, a wide range of algorithms for feature pre-processing and classification can be used. It can be understood that finding the best model may be challenging, especially since model selection depends on the tuning of the algorithms' associated hyperparameters (58). Optimizing both the algorithms and associated hyperparameters is called the Combined Algorithm Selection and Hyperparameter (CASH) optimization problem (59). The aim of the CASH optimization problem is to find the algorithm set A^* and associated hyperparameter set λ^* that minimizes a loss L. The WORC package solves the CASH problem by introducing hyperparameters for model selection in the hyperparameter space (60). In this way, the CASH problem is redefined as a pure hyperparameter optimization problem and is it easier to solve.

In WORC the hyperparameter optimization problem is solved by randomly constructing workflows and selecting the best performing ones for an ensemble model. These workflows consist of a combination of different algorithms for feature pre-processing and a classifier. A detailed overview of the algorithms used in WORC is provided in Supplementary data D. The workflows (w) are created by randomly selecting hyperparameters (λ_w) from the hyperparameter space. This process is repeated and produces a total of N_{rs} workflows. All these workflows are trained and validated, k times. In each k round, the workflows are trained on 85% of the training set and evaluated on 15% of the dataset. The loss of a radiomic workflow is calculated as the average of all k rounds:

$$L_w = \frac{1}{k}\sum_{i=1}^{k} L(\lambda_w, D_{\text{train}^{(i)}}, D_{\text{valid}^{(i)}}) \quad .$$

The F1-score is the loss function used for the workflow optimization, which is a class-balanced performance metric:

$$F_{1,w} = 2 \sum_{c=1}^{N_{\text{classes}}} \frac{N_c}{N_{\text{total}}} \frac{PREC_c \times REC_c}{PREC_c + REC_c} \quad ,$$

in which w refers to the radiomic workflow, N_{classes} is the number of classes, N_c number of samples in class c, N_{total} the total number of samples, $PREC_c$ precision of class c and REC_C the recall of class c

After all k training rounds, all the Nrs workflows are ranked based on the average loss. To prevent over-fitting, the best-performing N workflows are then ensembled by averaging the posterior probabilities. Finally, the ensembled model is trained on the complete training set and evaluated on a 1000x bootstrap resampled test set (Fig. 5).

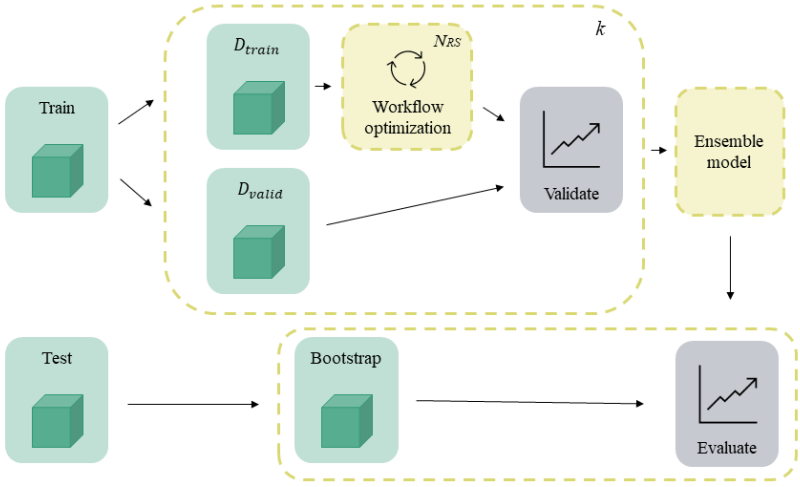


Figure 5: *Schematic overview of the train/test protocol in WORC. This figure was adapted from (60).*

E Deep learning method

The Med3D backbone will be used as the basis for the residual network in the deep learning method (61). Med3D is similar to the ResNet family but it has been modified to manage 3D medical data (46). The authors of Med3D trained and tested multiple versions of their network. However, a relatively shallow residual network is desired for this project due to the limited dataset, therefore the most shallow version of Med3D, which is based on Resnet-10, is used.

For this project's classification problem, several layers were added to the Med3D backbone. Fig. 6 shows the architecture of the Med3D backbone and additional layers for classification. The first layer of the Med3D backbone is a 3D convolution layer, followed by batch normalization, ReLU, and max-pooling. After that, there are 4 BasicBlocks (Fig. 7), which are residual blocks with a 3D convolution layer followed by batch normalization, ReLU, a 3D convolution layer, and batch normalization. All the BasicBlocks have a skip connection in which the input is downsampled, by adaptive average pooling, and added to the output of the last batch normalization layer before a ReLU layer. The classification layers are added after the BasicBlocks. First, the data is multiplied elementwise with the downsampled segmentation of the mask. Followed by a: max-pooling, flatten, dropout (with a probability of p), linear, and sigmoid layer. The network produces an outcome between 0 and 1, with 0 indicating a stable fracture and 1 indicating an unstable fracture.

Binary cross-entropy loss is used as loss function:

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i * log(p(y_i)) + (1 - y_i) * log(1 - p(y_i))$$

Where y_i and $p(y_i)$ are the label and the posterior probability of sample i, respectively and N is the number of samples.

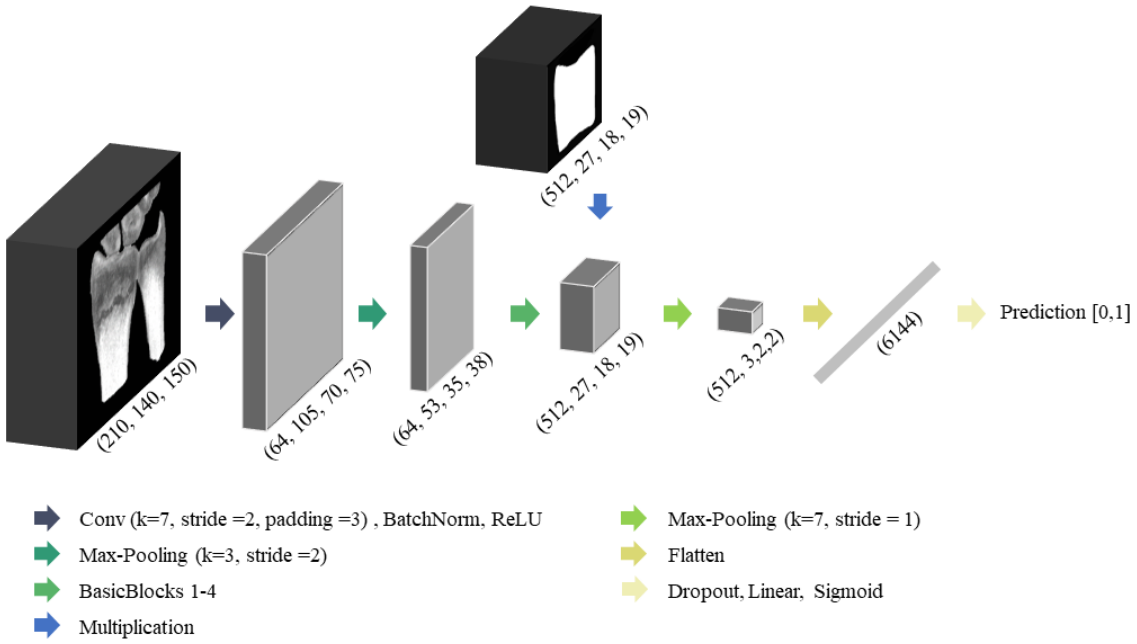


Figure 6: *Architecture of the adapted Med3D backbone. The different layers are described with different colored arrows, size of the data is specified between each layer. The Med3D backbone consists of all the layers before the multiplication with the segmentation. The architecture of the BasicBlocks is shown in Fig. 7*

.

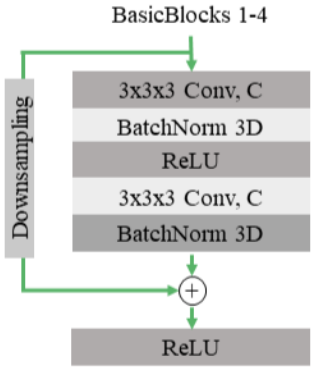


Figure 7: *Visualisation of the BasicBlocks. With C = 64, 128, 256, 512 in BasicBlock 1 - 4 respectively. In BasicBlock 2, the stride of the convolutional layers is 2, in the other BasicBlocks, the stride is 1.*

The weights of the Med3D network, which was trained on a large collection of medical images, were stored and made publicly available. In addition to training with random initialization, these

weights will be used to initialize the weights of the adapted residual network. In this way, the obtained knowledge from training on the medical dataset is transferred to DRF prediction task.

F Evaluation metrics

The performance of all methods is established by 5-fold cross-validation. Therefore, the data is split into 5 random, balanced, and equal groups. In each fold the data is trained on 4 groups an tested on 1 group, this is repeated 5 times and finally, the mean performance of all folds is calculated. Performance metrics to compare the different methods are the sensitivity, specificity, accuracy, confusion matrix, Receiver Operator Curve (ROC), and, Area Under the Curve (AUC). In addition, the ROC curves of the experiments were compared with each other by the DeLong test (62). All the experiments are repeated 5 times, these are called sub-experiments. The sensitivity, specificity, and accuracy are calculated based on all the sub-experiments, by concatenating the posterior probabilities of each sub-experiment. The AUC is calculated based on the mean ROC and given with the standard deviation. Finally, the best workflows from the radiomic method will be evaluated to analyze which classifiers and feature groups are mostly used.

IV EXPERIMENTS AND RESULTS

A Implementation details

All included scans were reoriented, cropped and segmented. This resulted in 61 cropped scans and masks with a size of $150, 140, 210$ pixels with a pixel size of 0.3 mm and stored in compressed nifti format. For the augmentation experiments, all images were augmented 10 times with TorchIO 0.18.82 (63) by randomly scaling, translating, and zooming.

All methods were performed with Python, version 3.6.8 and 3.7.2, an AMD Opteron 2378 CPU and an AMD EPYC 7451 GPU, for the radiomics method and deep learning method, respectively. Additionally, the radiomics method used WORC 3.6.0 (57) and the deeplearning method Pytorch 0.4.1 (64). Statistical analysis was performed with Python 3.7.2, Seaborn 0.12.0 (65), Scikit-learn 0.23.1 (66).

A.1 Experiments radiomic method

According to the DRF dataset the N_{rs}, N and, k parameters were set to 1000, 100, and 5, respectively. Experiments with $k = 10$ were also conducted, these results are shown in Supplementary data E.

A.2 Experiments deep learning method

In all experiments the model was trained during 100 epochs with a batch size of 5. The residual network has a total of 14.187.841 trainable parameters, from which 14.181.696 originates from the Med3D backbone, and 6145 parameters, from the added layers for classification.

A selection of hyperparameters is optimized during training: learning rate (lr), optimizer, dropout rate (p), and, freezing of the Med3D backbone. Experiments to determine the best hyperparameters values were performed in that order. As a result, once a hyperparameter value was chosen, it was used in all future experiments. Lastly, the architecture of the last BasicBlock (4) was changed in order to enhance performance. It was decided to only adapt the architecture of the last BasicBlock because the last BasicBlock is expected to learn the most specific features for the DRF application. (67). The weights before this BasicBlock were frozen to see the effect of the adapted BasicBlock. Figure 8, depicts the adapted structure of BasicBlock 4, it can be seen that the BatchNorm layers were removed and two dropout layers were added. Furthermore, optimization was performed with a validation set: a randomly selected balanced group of 20% of the training data.

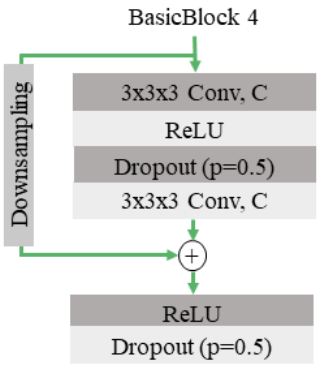


Figure 8: *Visualisation of the adapted BasicBlock 4, with C = 512.*

B Results

B.1 Hyperparameter selection

The selected hyperparameters to optimize were the learning rate (lr), optimizer, dropout rate (p), freezing of Med3D backbone parameters, and adaption of last BasicBlock. Fig. 1, shows the various hyperparameter values that were tested. Different experiments were performed and value selection was based on an analysis of the learning curve. When analyzing the learning curve, the following factors were taken into account (in order of importance): validation loss, training loss, and, the fluctuations and slopes of both curves. The learning rate and optimizer were chosen based on learning curves of 1 cross-validation fold, while the dropout rate and freezing of the Med3D backbone were chosen based on all cross-validation folds. All learning curves are provided in In Supplementary data F. The hyperparameter selection resulted in: an initial lr of 0.0001, an ADAM optimizer, a dropout rate p of 0.2, partial unfreezing of the Med3D backbone (BasicBlock 4), and using the adapted version of BasicBlock 4. In addition, a ReduceLROnPlateau schedular decreased the learning rate by 0.1 when the validation loss was not decreasing for 10 epochs.

Table 1: *Hyperparameter values used during optimization of the deep learning method, the selected values are highlighted by bold text. The abbreviation 'BB' refers to BasicBlock.*

Learning rate	Optimizer	Dropout rate	Freezing Med3D backbone	Adapting BB 4
0.001	**ADAM**	0.1	Unfrozen	**Yes**
0.0001	SGD	**0.2**	Partially unfrozen[1]	No
0.00001		0.3	**Partially unfrozen**[2]	
		0.4	Frozen	
		0.5		

[1] The Med3D backbone was frozen except from BasicBlock 3 and 4
[2] The Med3D backbone was frozen except from BasicBlock 4

B.2 Classification outcomes

Results of all the classification outcome measures are provided in Table 2. For the radiomic method, two experiments were performed, one with augmented data and one without. The AUC of the experiment without augmented data was AUC of 0.58 ± 0.07 and the AUC of the experiment with augmented data was 0.64 ± 0.01. The ROC curves and confusion matrices of both experiments are shown in Fig. 9a, Fig. 10a, and, Fig. 10b.

For the deep learning method, three experiments were performed with the adapted Med3D residual network: trained from scratch, trained from scratch with augmented data, and, transfer learning and augmented data. The AUCs of these experiments were 0.63 ± 0.01, 0.55 ± 0.02 and 0.51 ± 0.05 for

the experiments without augmentation, with augmentation, and with both augmentation and transfer learning, respectively. The ROC curves and confusion matrices of all experiments are shown in Fig. 9b, Fig. 10c, Fig. 10d, and, Fig. 10e.

Based on the results of the deLong test (Table 3. It was found that the radiomic experiment with augmented data was significantly different ($p < 0.05$) from both the deep learning experiment with augmentation and with augmentation and transfer learning ($p = 0.014$). The deep learning experiment without augmentation was found to be significantly different from the deep learning experiment with augmentation and transfer learning ($p = 0.014$).

The classification outcomes of a random guess were included in the results for reference. Random guess was performed by randomly generating values between 0 and 1.

Because the performance of the radiomic and deep learning methods were close to random guessing, it was decided not to perform ensemble learning. Therefore, it is excluded from the results.

Table 2: *Classification results of the different methods. The AUC is given with the standard deviation of all sub-experiments. Abbreviations: RM: radiomic method; DML: deep learning method. The 'a' and the 't' behind the experiment name indicate the use of data augmentation and/or transfer learning.*

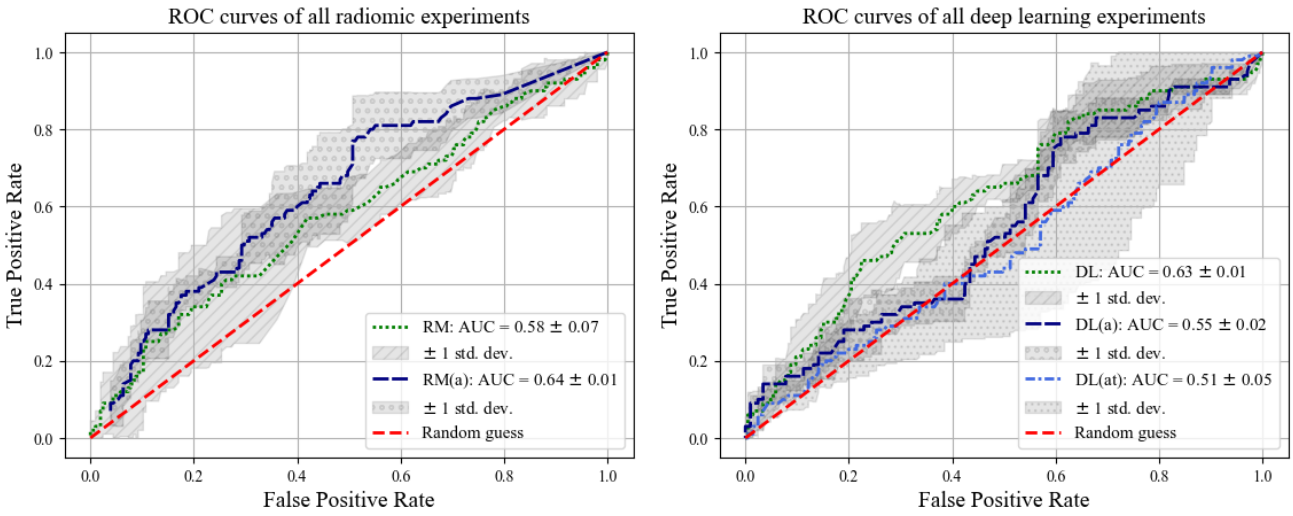
Experiment	Sens. [%]	Spec. [%]	Acc. [%]	AUC
RM	0.11	0.95	0.67	0.58 ± 0.07
RM (a)	0.35	0.83	0.67	0.64 ± 0.01
DLM	0.22	0.89	0.67	0.63 ± 0.01
DLM (a)	0.28	0.78	0.62	0.55 ± 0.02
DLM (at)	0.7	0.31	0.43	0.51 ± 0.05
Random	0.44	0.52	0.49	0.50 ± 0.08

Table 3: *Results of the deLong test. The delong test was used to compare the ROC curves of all experiments and establish if the curves are statistically different ($p < 0.05$). This table shows the p-values obtained after comparing each experiment to the others. Abbreviations: RM: radiomic method; DML: deep learning method. The 'a' and the 't' behind the experiment name indicate the use of data augmentation and/or transfer learning.*

Experiment	RM	RM(a)	DL	DL(a)	DL(at)
RM	-	0.18	0.31	0.63	0.15
RM(a)	-	-	0.78	0.073	0.014
DL	-	-	-	0.13	0.014
DL(a)	-	-	-	-	0.34
DL(at)	-	-	-	-	-

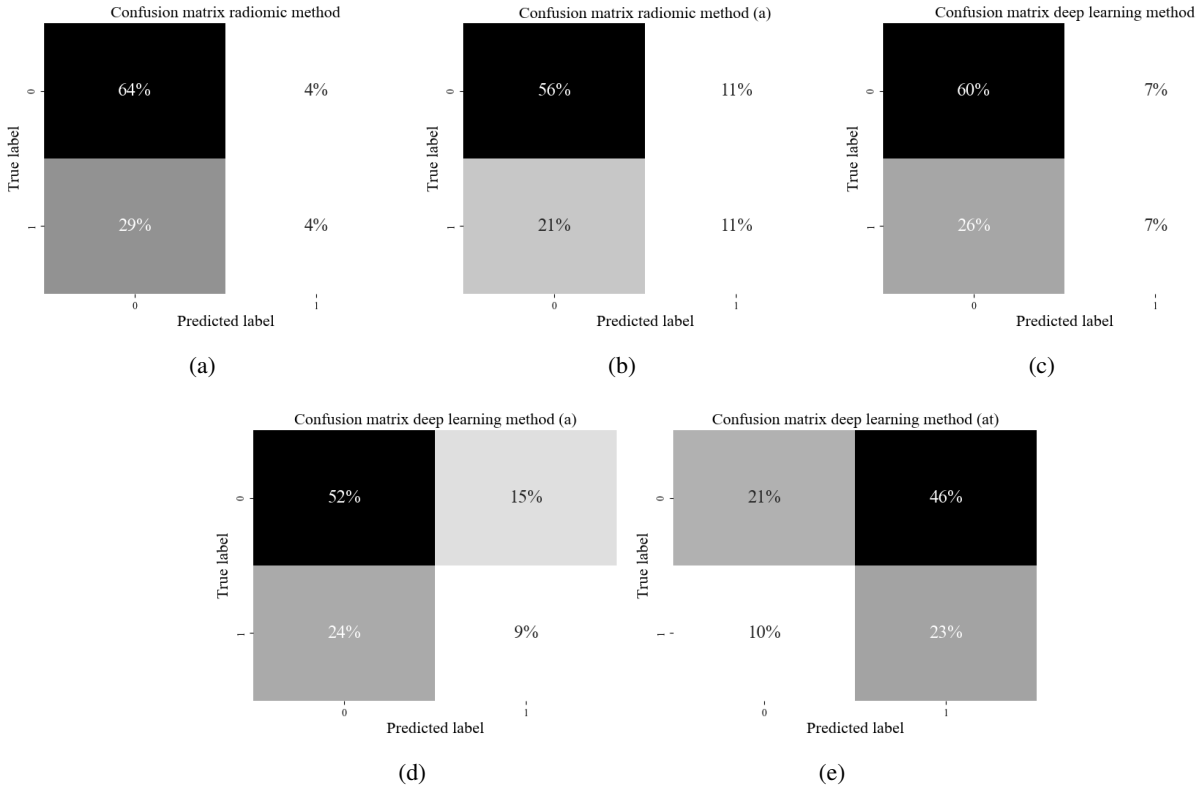
B.3 Additional analysis of the radiomic method

In addition to the classification results, it was determined which classifiers and feature groups were most commonly used in the WORC ensemble model. This was done by selecting the best best-performing experiment, and then selecting the 10 best radiomic workflows of all cross-validation folds, resulting in 50 workflows. After analyzing these workflows it was found that the SVM was used in all workflows (50/50) and that the histogram features and texture Gray-Level-Co-Occurrence Matrix (GLCM) features were used most often in all the workflows (48/50) and (34/50), respectively.



(a) (b)

Figure 9: *Receiver operator curves (ROC) of radiomic method (a) and deep learning method (b). The plotted curves are the mean ROC curves of all sub-experiments shown with the standard deviation in grey. Abbreviations: RM: radiomic method; DML: deep learning method. The 'a' and the 't' behind the experiment name indicate the use of data augmentation and/or transfer learning.*



(a) (b) (c)

(d) (e)

Figure 10: *Confusion matrices of all experiments. (a) and (b) show the confusion matrix of the radiomic experiments. (c), (d), and, (e) show the confusion matrices of the deep learning experiments. The labels '0' and '1' refers to the prediction of a stable DRF and prediction of an unstable DRF, respectively. The 'a' and the 't' behind the experiment name indicate the use of data augmentation and/or transfer learning.*

V DISCUSSION

In this study, we studied the use of machine learning for predicting the stability of DRF based on a small CT-scan dataset. In order to solve the problem of having a small dataset, different machine-learning methods were selected: traditional machine learning (radiomics method), residual network (deep learning method) and ensemble learning. After evaluation of the results, the radiomics method combined with augmentation, gave the best results (AUC: 0.64 ± 0.01). Although the difference between random guessing and the other experiments is small it is the most promising method. Also, it was found that using an augmented data set in the radiomics method resulted in improved performance. This gives a slight indication that the radiomic method can learn to predict DRF when there would be more data available. In addition, the use of data augmentation resulted in a lower standard deviation which indicates that experimenting with more data will result in more stable performance.

Surprisingly, the deep learning experiment without augmentation performed best of all the deep learning-based experiments while it was expected that the use of transfer learning and/or data augmentation would improve performance. The AUC is even decreasing with the use of augmentation and transfer learning which we do not attribute to random fluctuations because a larger overlap in ROC curves would then be expected. Moreover, the deep learning method without augmentation was significantly different from the method with augmentation and transfer learning. The higher performance of the experiment without augmentation or transfer learning may be an accidental finding resulting from the noise in the experiment. This was less likely to occur in the augmented dataset because interpolation was used for rotation and scaling and therefore the experiments that used augmented data performed worse. These effects probably occurred because the network had too many trainable parameters and a lack of training data. Therefore the results of the deep learning method should be evaluated with care, especially regarding the 'good' performance of the experiment without data augmentation.

To the best of our knowledge, physicians are not able to distinguish stable cases from non-stable ones by looking at the CT-scan. However, the radiomic experiments indicated that there lies some information in image intensities or texture that can help to predict DRF stability. For example, dorsal comminution, a specific fracture pattern, might be a predicting factor for the stability of DRFs (68). This fracture pattern might be 'recognized' by the GLMC texture features. It remains, however, uncertain why the radiomic method could extract features from CT-scans to predict the stability of DRF and the deep learning method could not. This might be because the WORC algorithm, used in the radiomic method, had fewer trainable parameters compared to the deep learning method and therefore overfitted less quickly on the training set. In addition, also the feature extraction in the radiomic method was less tuned to the dataset and therefore more robust.

Finally, a notable finding in this study was that completely freezing the Med3D network gave the best validation loss and resulted in less overtraining (Supplementary data E Fig. 18). Contrary to expectations, this did not enable learning specific DRF features (67). One possible explanation could be, again, that the network had too many trainable parameters for the small dataset. A common solution is to add a dropout layer to reduce the number of parameters. However, with an increased dropout rate the training loss did not decrease, and there appeared to be not much improvement in the validation loss (Supplementary data E Fig. 17). Changing the architecture of the last BasicBlock did however improve the learning curve slightly (Supplementary data E Fig. 19). Since the change mainly involved using dropout instead of batch normalization layers it can be concluded that dropout has a positive influence on the performance but the location of the dropout layer is of great importance.

Limitations

One of the main limitations of this study is the small dataset, although several approaches were applied to solve this issue. It is important to keep this in mind when looking at the results because methods that are currently unsuccessful might perform better with a larger dataset. Furthermore, the method used to create the segmentations was not very precise and might have included some bias, especially since the segmentation is important for the radiomic feature extraction (69). Moreover, the segmentations were accomplished by roughly selecting the bones with some surroundings. This method was chosen because there might be some information in the surrounding tissue that is important for the stability prediction of DRF. However, it might have been better to segment strictly the bones without surrounding tissue and later dilate the mask to include some surroundings. Another limitation is the hyperparameter selection in the deep learning method: it was limited to four parameters and was done in a certain order. Therefore, it might be that there are more hyperparameters, or different combinations of hyperparameters that perform better. Lastly, the dataset selection might have introduced bias. For example, many cases were excluded because there was insufficient follow-up information or X-rays to determine the progress of these cases. If this follow-up information was available they could have been suitable for this research. In view of this project, a hospital protocol where all patients with DRF receive a CT-scan would have been better.

Future work

First, since the performance was improved in the radiomics method when using simple data augmentation to increase the dataset, it would be very relevant to experiment with the methods in this project with other data augmentation techniques or a larger dataset. Secondly, although it is not yet established, gender and age might be predictors for the secondary displacement of DRFs (68), therefore it may be worthwhile to include these factors in future research. Moreover, since all patients also have follow-up X-rays it might be interesting to include those X-rays for the prediction of DRF. In addition, it might be interesting to also include information about the shape of the bone and fracture, for example by supplying a segmentation that follows the bone cortex and fracture lines. Lastly, other research could further investigate feature extraction in the radiomic method. In this study, only the larger feature (sub)groups were evaluated, but it is interesting to see go a bit more into detail to see if there exist features that are good predictors for DRF stability.

VI CONCLUSION

The radiomic method is the most promising method for predicting the stability of distal radius fractures, despite the small difference in performance compared to random guessing and the deep learning method. However, for further research, it is highly recommended to acquire a larger dataset.

SUPPLEMENTARY DATA

A Literature Review

Use this link to the Literature study. Please contact the author if it is not possible to access the file.

B Guideline for cropping CT scans in Mevislab.

1. Check if the patient's fracture involves the left or right wrist (excel document)
    a. If the fracture is in the right wrist: go to the "Switch" module and press on the right arrow.
    b. If the fracture is in the left wrist: check if the input of the "Switch" module is "0"
2. Double click on the "DirectDicomImport" module
3. Click "Browse" and select the folder with the patient's CT
4. Click "Import"
5. Close "DirectDicomImport" module
6. Double click on "X_Markers" module
7. Scroll down to slice 0 (left bottom corner)
8. Slowly scroll up, as soon as the whole proximal end of the radius is in the plane, move the cursor to the middle of the shaft hold ALT and click with the left mouse button.
9. Check if you see an ORANGE rectangle
10. Scroll up until you last see a not fractured radius, so the fracture needs to be whole. BUT make sure that you are at least 20 slices away from the first point.
11. Move the cursor to the middle of the shaft hold SHIFT and click with the left mouse button
12. Check if you see a GREEN rectangle
13. Close "X_Markers" module
14. Double click on "Reference_Point" module
15. Scroll up until you see the processus Styloideus radii i.e. the highest point of the radius articular plane.
16. Hold SHIFT and left-click on the processus styloideus radii
17. Check if you see a RED rectangle
18. Close "ReferencePoint" module
19. Double click on "Y_Markers" module
20. Scroll down until you see a complete part (without fracture) of the radius.
21. Move to the point of the radius closest to the ulna on the palmar side of the radius
22. Hold ALT and click on the left proximal side of the radius
23. Check if you see a BLUE rectangle
24. Move over the oblique palmar side of the radius in a somewhat straight line following the bone until you can no longer follow the bone in a straight line.
25. At this point hold SHIFT and click on the right proximal side of the radius
26. Check if you see a YELLOW rectangle
27. Close the "Y_Markers" module
28. Double click on "3DCroppedView" module
29. Check if the whole distal part of the radius is cropped out, if not return to step 4 and try to select more precisely the indicated points
30. Close the "3DCroppedView" module
31. Open "DicomTool"
32. Click "Browse" select the folder "patientnumber_cropped_radius" click select folder
33. Click "Save"
34. Close "DicomTool"

C Algorithms used in WORC

Table 4: *Categories of algorithms used in WORC (60).*

Algorithm group	Algorithms
Feature selection	Group-wise selection, Variance Threshold, SelectFromModel, Univariate testing, RELIEF
Feature Imputation	Mean, Median, Mode, Constant (zero), K-nearest neighbors
Feature Scaling	Robust z-scoring
Dimensionality Reduction	Principal component analysis
Resampling	RandomUnderSampling,RandomOverSampling, NearMiss, NeighborhoodCleaningRule, SMOTE, Adaptive synthetic sampling
Classification	Support vector machine, Logistic regression, Linear discriminant analysis, Quadratic discriminant analysis, Random Forest, Gaussian Naive Bayes, Adaptive boosting, Extreme gradient boosting

D Radiomic features used in WORC

Table 5: *Categories of Radiomic features used in WORC (60).*

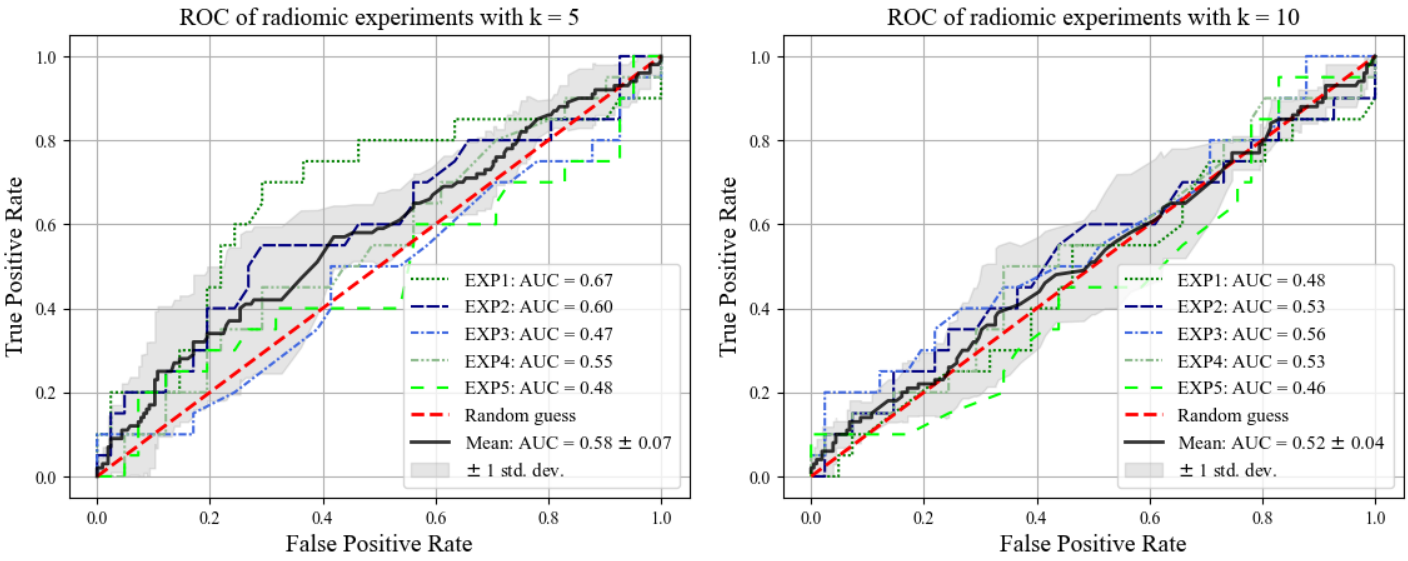
Feature group	Description
Histogram based features	Features based on the image intensities
Shape features	Morphological features based on the shape of the ROI
Orientation features	Features based on the orientation and location of the ROI
Texture features	Features based on image texture[1]
Vessel features	Features extracted after the application of vessel filter[2]
LoG filter features	Features extracted after the application of LoG [3] filters
Phase features	Features based on local phase
DICOM features	Features based on DICOM tags

[1] For example features extracted from the Gray-Level-Co-Occurrence Matrix (GLCM), or features extracted after the application of a filter such as a Gaussian filter

[2] Vessel filters are specific filters designed for tubular structures (70)

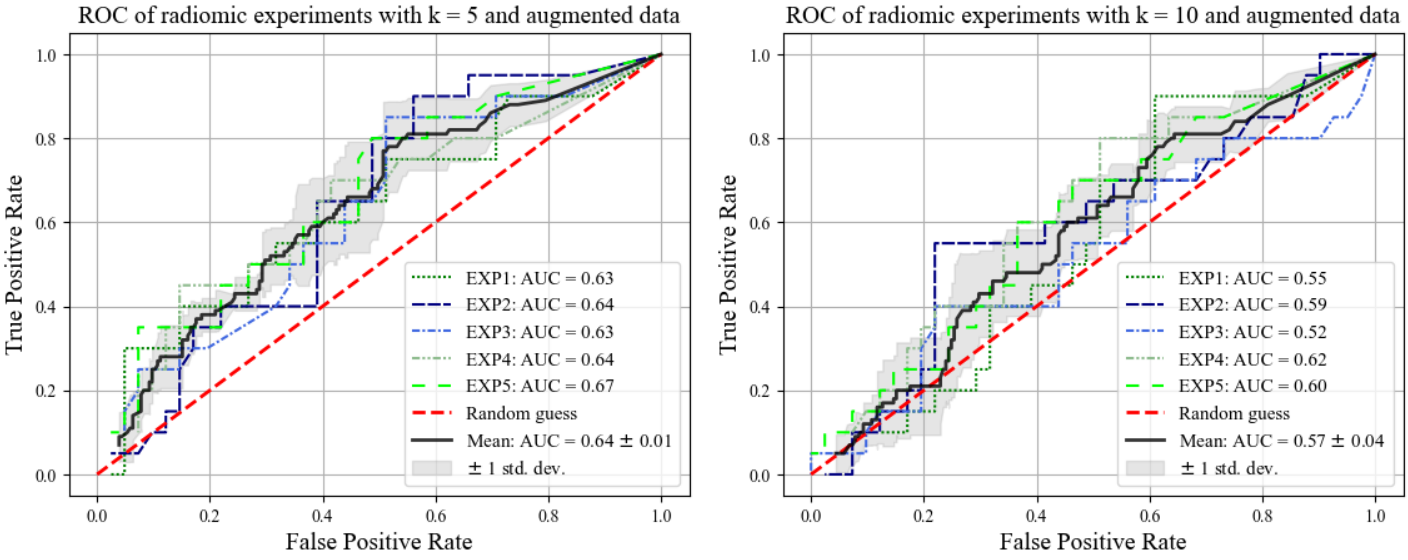
[3] Laplacian of Gaussian filters

E Additional ROC curves



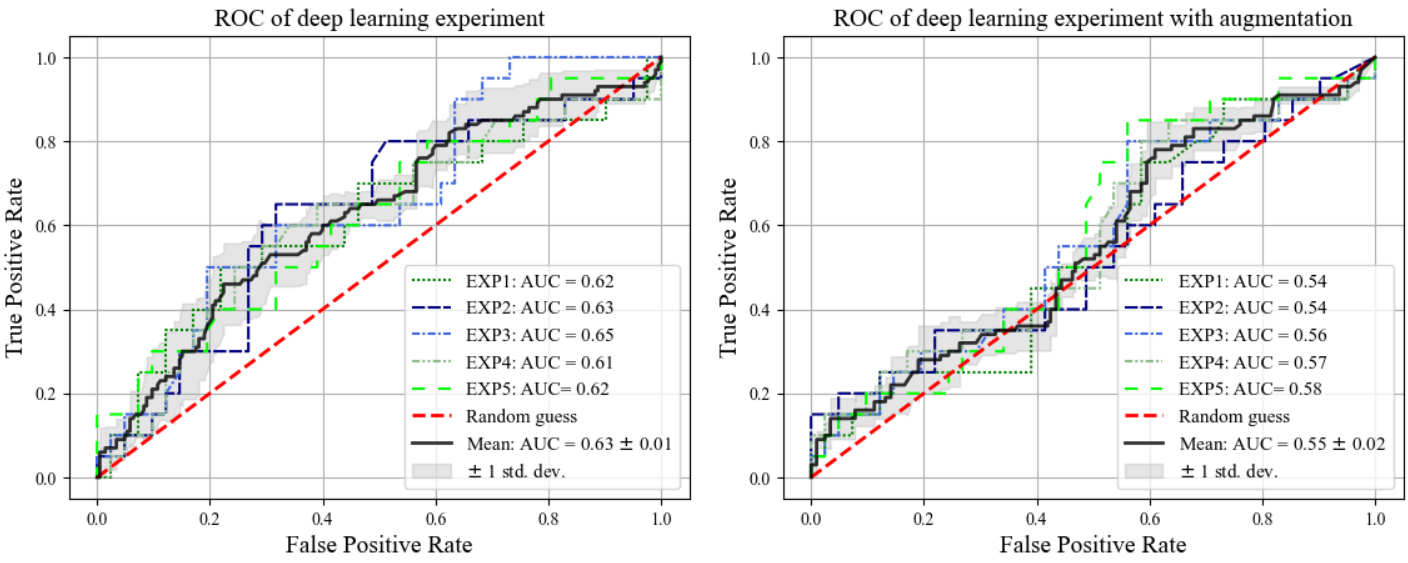
(a) (b)

Figure 11: *Receiver operator curves (ROC) of the sub-experiments from the experiment in the radiomic method without augmented data. (a) shows the results with 5 training rounds (k) and (b) the results with 10 training rounds.*



(a) (b)

Figure 12: *Receiver operator curves (ROC) of the sub-experiments from the experiment in the radiomic method with augmented data. (a) shows the results with 5 training rounds (k) and (b) the results with 10 training rounds.*



(a) (b)

Figure 13: *Receiver operator curves (ROC) of the sub-experiments from the experiment in the deep learning method without (a) and with augmented data (b).*

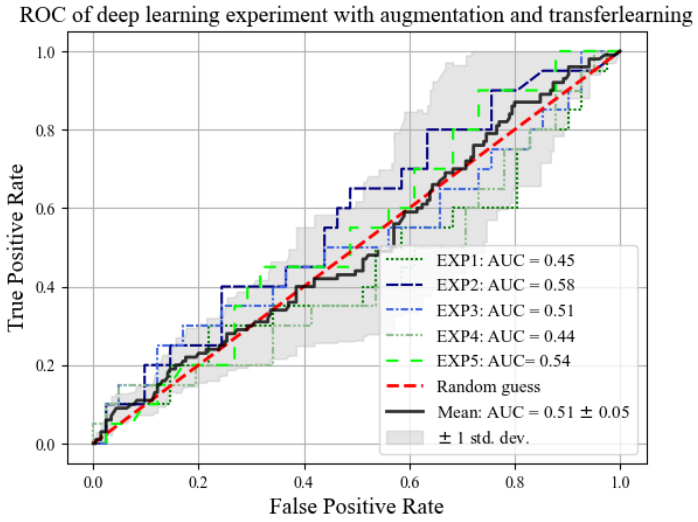


Figure 14: *Receiver operator curves (ROC) of the sub-experiments from the experiment in the deep learning method with augmentation and transfer learning.*

F Hyperparameter selection: learning curves

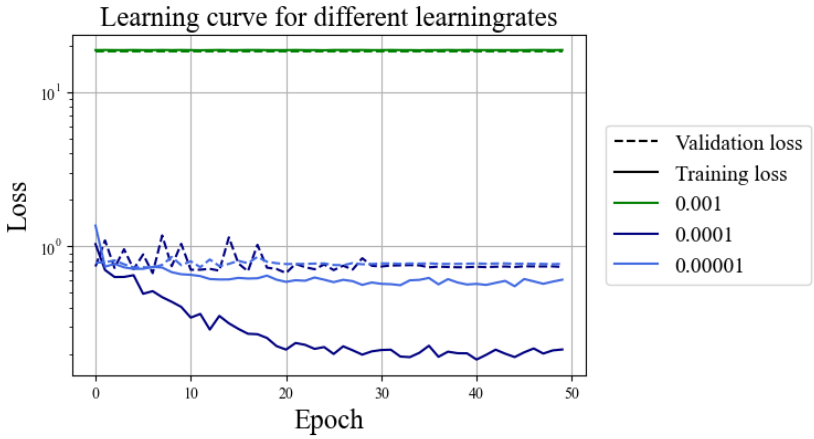


Figure 15: *Learning curves for experiments with different learning rates (lr).*

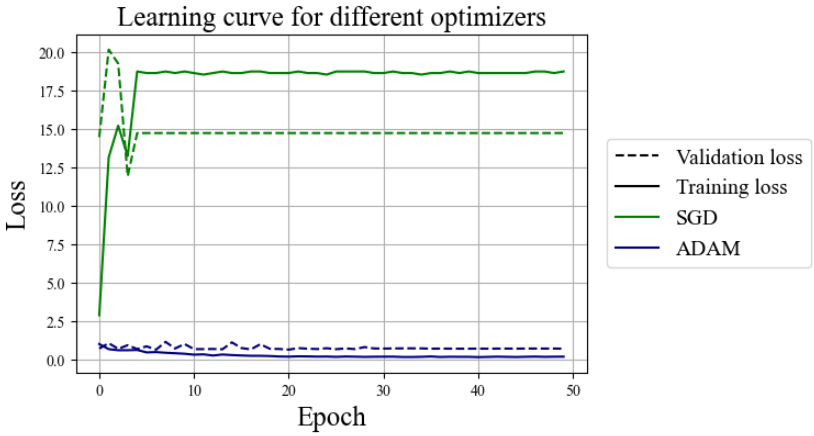


Figure 16: *Learning curves for experiments with different optimizers.*

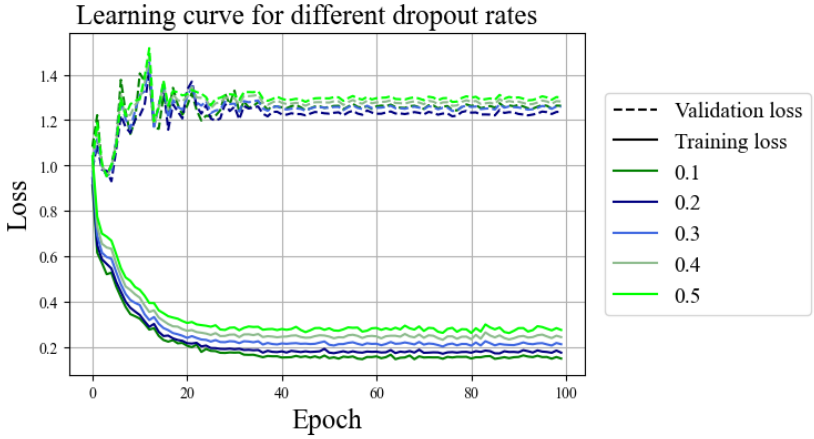


Figure 17: *Learning curves for experiments with different dropout rates (p).*

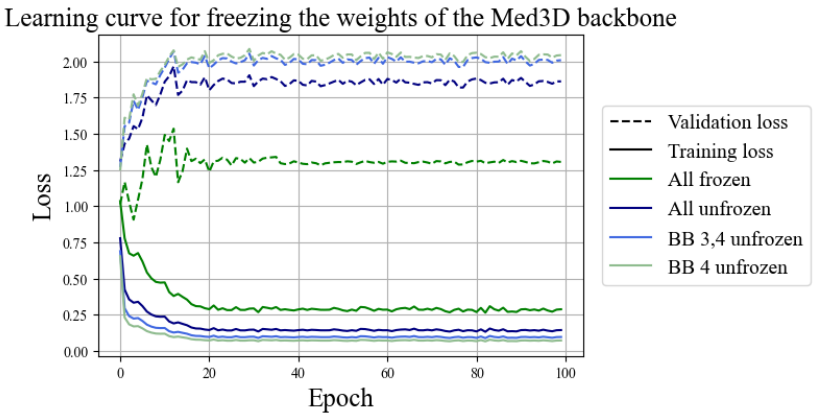


Figure 18: *Learning curves for experiments with different freezing Med3D backbone. 'BB' refers to BasicBlock.*

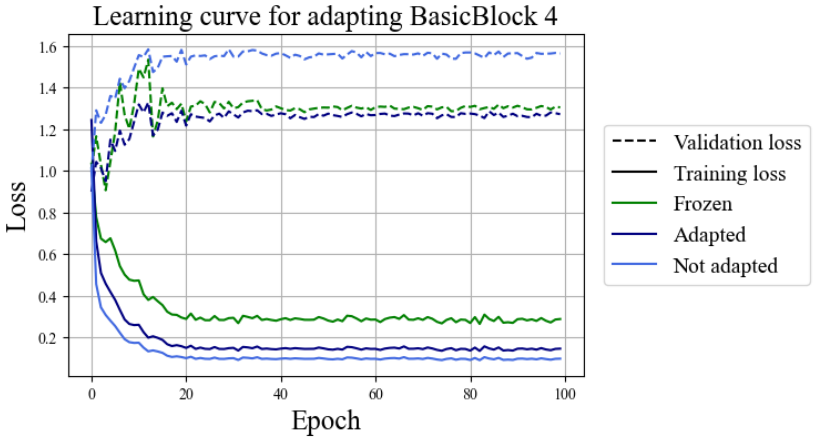


Figure 19: *Learning curves for experiments with and without altered architecture of BasicBlock 4. In both 'adapted' and 'not adapted' experiments, only the Med3D weights of BasicBlock 4 were unfrozen. For reference, the learning curve of the experiment with completely freezing the Med3D was also shown.*

REFERENCES

[1] N. H. A. M. Care, "Emergency department summary tables," 2008.

[2] N. J. MacIntyre and N. Dewan, "Epidemiology of distal radius fractures and factors predicting risk and prognosis," *Journal of Hand Therapy*, vol. 29, pp. 136–145, Apr. 2016.

[3] "Richtlijn distale radiusfracturen initiatief nederlandse vereniging voor heelkunde," 2021.

[4] T. van Staa, E. Dennison, H. Leufkens, and C. Cooper, "Epidemiology of fractures in england and wales," *Bone*, vol. 29, pp. 517–522, Dec. 2001.

[5] K. W. Nellans, E. Kowalski, and K. C. Chung, "The epidemiology of distal radius fractures," *Hand Clinics*, vol. 28, pp. 113–125, May 2012.

[6] J. Jack A. Porrino, E. Maloney, K. Scherer, H. Mulcahy, A. S. Ha, and C. Allan, "Fracture of the distal radius: Epidemiology and premanagement radiographic characterization," 2014.

[7] C. B. Corsino, R. A. Reeves, and R. N. Sieg, "Distal radius fractures," 2021.

[8] S. Meena, P. Sharma, A. Sambharia, and A. Dawar, "Fractures of distal radius: An overview," 2014.

[9] N. V. v. Heelkunde, "Distale radiusfracturen," 2021.

[10] F. Deng and J. Jones, "Distal radial fracture," *Radiopaedia.org*, 2015.

[11] C. Pailthorpe *et al.*, "Best practice for management of distal radial fractures (drfs)," 2018.

[12] S. H. Tajmir and T. K. Alkasab, "Toward augmented radiologists: Changes in radiology education in the era of machine learning and artificial intelligence," *Acad Radiol*, vol. 25, no. 6, pp. 747–750, 2018.

[13] B. Erickson, P. Korfiatis, Z. Akkus, and T. Kline, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, pp. 505–515, 2017.

[14] U. Schoepf and P. Costello, "Ct angiography for diagnosis of pulmonary embolism: State of the art," *Radiology*, vol. 230, no. 2, p. 329–337, 2004.

[15] K. Nieuwenhuijsen, D. Bruinvels, and M. Frings-Dresen, "Psychosocial work environment and stress-related disorders, a systematic review," *Occupational Medicine (London)*, vol. 60, no. 4, pp. 277–86, 2010.

[16] J. H. Ruotsalainen, J. Verbeek, A. Mariné, and C. Serra, "Preventing occupational stress in healthcare workers," *Sao Paulo Medical Journal*, vol. 134, no. 1, p. 92, 2016.

[17] E. Topol, "High performance medicine the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, pp. 44–56, 2019.

[18] C. G. B. Yogananda, B. R. Shah, M. Vejdani-Jahromi, S. S. Nalawade, G. K. Murugesan, F. F. Yu, M. C. Pinho, B. C. Wagner, B. Mickey, T. R. Patel, *et al.*, "A novel fully automated mri-based deep-learning method for classification of idh mutation status in brain gliomas," *Neuro-Oncology*, vol. 22, no. 3, pp. 402–411, 2020.

[19] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. van der Laak, B. van Ginneken, and C. I. Sanchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[20] A. Akgundogdu, R. Jennane, G. Aufort, C. Laurent, L. Benhamou, and O. Ucan, "3d image analysis and artificial intelligence for bone disease classification," *Journal of Medical Systems*, vol. 34, pp. 815–828, 2010.

[21] K. Ukai, R. Rahman, N. Yagi, K. Hayashi, A. Maruo, H. Muratsu, and S. Kobashi, "Detecting pelvic fracture on 3d-ct using deep convolutional neural networks with multi-orientated slab images," *Scientific Reports*, vol. 11, no. 1, p. 11716, 2021.

[22] X. Yao, K. Zhou, B. Lv, L. Wang, J. Xie, X. Fu, J. Yuan, and Y. Zhang, "3d mapping and classification of tibial plateau fractures," *Bone Joint Research*, vol. 9, no. 6, pp. 258–267, 2020.

[23] C. Blüthgen, A. S. Becker, I. Vittoria de Martini, A. Meier, K. Martini, and T. Frauenfelder, "Detection and localization of distal radius fractures: Deep learning system versus radiologists," *European Journal of Radiology*, vol. 126, 2020.

[24] T. Suzuki, S. Maki, T. Yamazaki, H. Wakita, Y. Toguchi, M. Horii, T. Yamauchi, K. Kawamura, M. Aramomi, H. Sugiyama, *et al.*, "Detecting distal radial fractures from wrist radiographs using a deep convolutional neural network with an accuracy comparable to hand orthopedic surgeons," *Journal of Digital Imaging*, vol. 35, pp. 39–46, Dec. 2021.

[25] K. Gan, D. Xu, Y. Lin, Y. Shen, T. Zhang, K. Hu, K. Zhou, M. Bi, L. Pan, W. Wu, and Y. Liu, "Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments," *Acta Orthopaedica*, vol. 90, pp. 394–400, Apr. 2019.

[26] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, June 2017.

[27] R. Ramachandran, D. Rajeev, S. Krishnan, and P. Subathra, "Deep learning an overview," *IJAER*, vol. 10, no. 10, pp. 25433–25448, 2015.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[29] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," *International Conference on Communication and Signal Processing (ICCSP), Chennai*, pp. pp. 0588–0592, 2017.

[30] W. L. Hosch, "Machine learning," Retrieved 2022-11-14 from: http://www.britannica.com/EBchecked/topic/1116194/machinelearning.

[31] M. Awad and R. Khanna, "Machine learning," in *Efficient Learning Machines*, pp. 1–18, Apress, 2015.

[32] A. Pezeshk, N. Petrick, W. Chen, and B. Sahiner, "Seamless lesion insertion for data augmentation in CAD training," *IEEE Transactions on Medical Imaging*, vol. 36, pp. 1005–1015, Apr. 2017.

[33] C. L. Phillips, M.-A. Bruno, P. Maquet, M. Boly, Q. Noirhomme, C. Schnakers, A. Vanhaudenhuyse, M. Bonjean, R. Hustinx, G. Moonen, *et al.*, ""relevance vector machine" consciousness classifier applied to cerebral metabolism of vegetative and locked-in patients," *NeuroImage*, vol. 56, pp. 797–808, May 2011.

[34] L. Ubaldi, V. Valenti, R. Borgese, G. Collura, M. Fantacci, G. Ferrera, G. Iacoviello, B. Abbate, F. Laruina, A. Tripoli, A. Retico, and M. Marrale, "Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples," *Physica Medica*, vol. 90, pp. 13–22, Oct. 2021.

[35] L. Wei, C. Cui, J. Xu, R. Kaza, I. E. Naqa, and Y. K. Dewaraja, "Tumor response prediction in 90y radioembolization with PET-based radiomics features and absorbed dose metrics," *EJNMMI Physics*, vol. 7, Dec. 2020.

[36] L. Y. Pratt, "Discriminability-based transfer between neural networks," in *Advances in Neural Information Processing Systems* (S. Hanson, J. Cowan, and C. Giles, eds.), vol. 5, Morgan-Kaufmann, 1992.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[38] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[39] M. Raghu, C. Zhang, J. M. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning with applications to medical imaging," *CoRR*, vol. abs/1902.07208, 2019.

[40] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," 2019.

[41] D. R. Clymer, J. Long, C. Latona, S. Akhavan, P. LeDuc, and J. Cagan, "Applying machine learning methods toward classification based on small datasets: Application to shoulder labral tears," *Journal of Engineering and Science in Medical Diagnostics and Therapy*, vol. 3, Oct. 2019.

[42] Y. Onishi, A. Teramoto, M. Tsujimoto, T. Tsukamoto, K. Saito, H. Toyama, K. Imaizumi, and H. Fujita, "Investigation of pulmonary nodule classification using multi-scale residual network enhanced with 3dgan-synthesized volumes," *Radiological Physics and Technology*, vol. 13, pp. 160–169, May 2020.

[43] P. Rajpurkar, A. Park, J. Irvin, C. Chute, M. Bereket, D. Mastrodicasa, C. P. Langlotz, M. P. Lungren, A. Y. Ng, and B. N. Patel, "AppendiXNet: Deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining," *Scientific Reports*, vol. 10, Mar. 2020.

[44] K. C. Wong, T. Syeda-Mahmood, and M. Moradi, "Building medical image classifiers with very limited data using segmentation networks," *Medical Image Analysis*, vol. 49, pp. 105–116, Oct. 2018.

[45] A. Zaeemzadeh, N. Rahnavard, and M. Shah, "Norm-preservation: Why residual networks can become extremely deep?," *CoRR*, vol. abs/1805.07477, 2018.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[47] S. Hochreiter, "Untersuchungen zu dynamischen neuronalen netzen," *Diploma, Technische Universität München*, vol. 91, 1991.

[48] A. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Advances in Neural Information Processing Systems*, vol. 29, Curran Associates, Inc., 2016.

[49] Y. Chen, Y. Ren, L. Fu, J. Xiong, R. Larsson, X. Xu, J. Sun, and J. Zhao, "A 3d convolutional neural network framework for polyp candidates detection on the limited dataset of CT colonography," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, July 2018.

[50] Z. Z. H., *Ensemble Methods: Foundations and Algorithms*. Machine Learning Pattern Recognition Series, Chapman and Hall, CRC, 2012.

[51] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, pp. 1–39, Nov. 2009.

[52] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, and S. M. Smith, "Accurate brain age prediction with lightweight deep neural networks," *Medical Image Analysis*, vol. 68, p. 101871, Feb. 2021.

[53] X. Xu, C. Wang, J. Guo, Y. Gan, J. Wang, H. Bai, L. Zhang, W. Li, and Z. Yi, "MSCS-DeepLN: Evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks," *Medical Image Analysis*, vol. 65, p. 101772, Oct. 2020.

[54] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. Aerts, "Radiomics: Extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, pp. 441–446, Mar. 2012.

[55] "MeVisLab: https://www.mevislab.de."

[56] G. Herman, J. Zheng, and C. Bucholtz, "Shape-based interpolation," *IEEE Computer Graphics and Applications*, vol. 12, no. 3, pp. 69–79, 1992.

[57] M. P. A. Starmans, T. Phil, S. R. van der Voort, and S. Klein, "Workflow for optimal radiomics classification (worc)," 2018.

[58] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," 2018.

[59] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-weka: combined selection and hyperparameter optimization of classification algorithms," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, Aug. 2013.

[60] M. P. A. Starmans, S. R. van der Voort, T. Phil, M. J. M. Timbergen, M. Vos, G. A. Padmos, W. Kessels, D. Hanff, D. J. Grunhagen, C. Verhoef, *et al.*, "Reproducible radiomics through automated machine learning validated on twelve clinical applications," 2021.

[61] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.

[62] D. Klingelhöfer, M. Braun, R. K. Seeger-Zybok, D. Quarcoo, D. Brüggmann, and D. A. Groneberg, "Global research on fabry's disease: Demands for a rare disease," *Mol. Genet. Genomic Med.*, vol. 8, p. e1163, Sept. 2020.

[63] F. Pérez-García, R. Sparks, and S. Ourselin, "Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Computer Methods and Programs in Biomedicine*, p. 106236, 2021.

[64] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, , *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.

[65] M. L. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

[66] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[67] Y. B. J. Yosinski, J. Clune and H. Lipson, "How transferable are features in deep neural networks?," *Advances in neural information processing systems*, p. pp. 3320–3328, 2014.

[68] M. M. J. Walenkamp, S. Aydin, M. A. M. Mulders, J. C. Goslings, and N. W. L. Schep, "Predictors of unstable distal radius fractures: a systematic review and meta-analysis," *Journal of Hand Surgery (European Volume)*, vol. 41, pp. 501–515, Sept. 2015.

[69] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, *et al.*, "Radiomics: the process and the challenges," *Magnetic Resonance Imaging*, vol. 30, pp. 1234–1248, Nov. 2012.

[70] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI'98*, pp. 130–137, 1998.