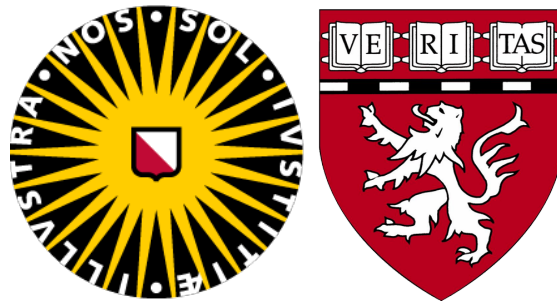


Enabling the creation of X-ray patient registries through Deep Neural Networks

Jose M. Acitores
(7007337)

Supervised by
Hugo Schnack
Soheil Ashkani



A thesis presented for the degree of
MSc Artificial Intelligence

Department of Information and Computing Sciences
Utrecht University - Harvard Medical School
2022

Contents

1	Introduction	2
2	Artificial intelligence in medicine	4
2.1	Traditional machine learning	4
2.2	Deep learning	4
2.3	Computer vision	4
2.3.1	Convolutional neural networks	5
2.3.2	Classification	5
3	Model architecture	5
3.1	ResNet101	5
3.2	EfficientNet	6
3.3	Vision Transformer	6
4	Data Analysis and Processing	6
4.1	Body part data	7
4.1.1	Class similarities	8
4.1.2	Processing	9
4.2	Hip fracture data	9
4.2.1	Data preparation	10
5	Methods	10
5.1	Body part classification	10
5.1.1	Architecture modifications	10
5.1.2	Training	10
5.2	Instrument detection	11
5.2.1	Model and training	11
5.3	Fracture detection	11
5.3.1	Model and training	11
5.4	App	11
6	Results	12
6.1	Body part classification	12
6.2	Instrument detection	14
6.3	Fracture detection	16
7	Explainability	20
8	Discussion	21
9	Acknowledgements	22
10	Code and data availability	22
11	Bibliography	22

Abstract

The **objective** is to implement a neural network based on previous models like the ResNet/ EfficientNet/ Vision Transformers in order to classify x-rays in three different sets of categories; nine different body part categories, the presence or absence of an instrument, and the presence or absence of a fracture. The **datasets** consists of 45000 x-rays from the nine different classes, labeled by physicians, for the body part classification matter as well as for the instrument detection. A hip X-ray dataset was used for fracture detection, consisting of 468 patients and also labeled by physicians. The **methods** used to solve this task consist of the implementation of three different neural networks, the use of transfer learning from the ImageNet dataset, and the fine-tuning of the networks to fit them to our problems. The data underwent normalization and augmentation for both datasets, as well as concatenation for the fracture dataset. After collecting the results, the final model is chosen based on the metrics collected after the testing stage. The **Results** show great performance for both problems that were trained on the larger dataset. The body part classification problem was solved by the three models almost. Equally, the best model being Efficient Net with a logarithmic loss of 0.1053 and a Cohen Kappa score of 0.9940. For instrument detection, the best model was ResNet101, achieving an AUC of 0.99 with a 95 CI of 0.95-1. Finally, in the proposed proof of concept for fracture detection, the results did not surpass the ones of a professional radiologist, achieving a sensitivity of 0.86, a specificity of 1, and an AUC of 0.93. In **conclusion**, the results show that the creation of an automated pipeline for the creation of x-rays patient registries is possible and achievable with a low error rate; the main limitation is the lack of labeled data to aid the creation of the given pipeline. Therefore the main challenge is the collaboration of medical staff for the creation of an initial database that can help to complete the work that is often overlooked and avoided. Enabling the expansion of scope for many possible applications of AI in the medical field.

1 Introduction

Registries are referred to as systems that gather and store information, and patient registry is focused on collecting, storing, and recording health-related data. The term “patient registry” is being used to distinguish registries aimed to advance the efficacy, safety, quality, and distribution of healthcare services and products; however, currently, a uniform definition of the word is still lacking [Gliklich et al., 2018]. World Health Organization (WHO) has described patient registries as “a file of documents containing uniform information about individual persons, collected systematically and comprehensively, to serve a predetermined purpose.” [Brooke and Organization, 1974] The Agency for Healthcare Research and Quality (AHRQ) has defined a patient registry as “an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves one or more predetermined scientific, clinical, or policy purposes.” [Gliklich et al., 2018] Furthermore, the registries were described as “an organized system for the collection, storage, retrieval, analysis, and dissemination of information on individual persons who have either a particular disease, a condition (e.g., a risk factor) that predisposes [them] to the occurrence of a health-related event, or prior exposure to substances (or circumstances) known or suspected to cause adverse health effects” by the National Committee on Vital and Health Statistics [Gliklich RE, Dreyer NA, Leavy MB, editors.]. The development of routine patient registries and the digitalization of health data can provide valuable insights into various health-related topics, leading to improved evidence-based medicine and healthcare quality. AHRQ has suggested a set of standards for developing patient registries [Gliklich RE, Dreyer NA, Leavy MB, editors.]. Firstly, the data must be collected naturally, meaning that the care patients receive should be determined by both physician and patient and not by the registry protocol. Secondly, the purposes of the registry must be defined before the data collection and analysis. Thirdly, the parameters in focus must be specified and consistently defined. Fourthly, the data collection for each patient must be carried out uniformly, regarding the type of data being collected and the frequency. Fifthly, the data must be reflective of patients’ clinical status. Finally, at least one of the elements of the data must be collected actively to meet the purpose of the registry (commonly being collected from patients or physicians) rather than to be inferred from other sources (administration, pharmacy database, etc.). However, these guidelines are flexible, and exceptions could be applied in certain circumstances[Gliklich RE, Dreyer NA, Leavy MB, editors.].

A patient registry could be a robust implementation to monitor the course of the disease, evaluate the effectiveness of medical interventions, find contributing factors to the prognosis of the disease and

quality of life, examine healthcare utilization, and determine the quality of care [Gliklich RE, Dreyer NA, Leavy MB, editors.]. Patient registries could produce a real-time, actual picture of the burden of the disease as well as therapeutic practices and outcomes of patients and provide valuable information for different stakeholders. A physician could use the registry to assess whether the disease is being managed according to evidence-based guidelines, pay attention to specific aspects of the disease that might be overlooked, or compare themselves to their peers. Furthermore, a payer could use the information gathered by registries to evaluate which procedures, devices, and pharmaceutical products have higher effectiveness. Registry-based studies could also benefit pharmaceutical and medical device companies by demonstrating the performance of a product, developing hypotheses, or identifying the patient population that the product could be helpful for them. The U.S. Food and Drug Administration (FDA) has implied that “through the creation of registries, a sponsor can evaluate safety signals identified from spontaneous case reports, literature reports, or other sources, and evaluate the factors that affect the risk of adverse outcomes such as dose, the timing of exposure, or patient characteristics.” [Gliklich RE, Dreyer NA, Leavy MB, editors.] Although the use of registries has been increasing during the past few decades, the areas that have gained the most attention were cancer and cardiovascular diseases, and thus the need for the development of registries in other fields, such as orthopedics, is still unmet. The burden of musculoskeletal (MSK) disease is forecasted to rise in the future, and it is estimated that the prevalence of arthritis among the adult population will be increased to 25% in the U.S. by 2030 [Magaway and Malanga, 2022]. This increasing incidence and prevalence of MSK conditions might have been affected by age and lifestyle [Oh et al., 2011]. Besides, MSK conditions could influence the economy directly because of the costs expended on the treatment and indirectly because of the loss of productivity [Woolf et al., 2012]. Given the growing prevalence of MSK conditions and the rising healthcare expenditure, a trend towards systematically collecting data in patient registries has appeared [Magaway and Malanga, 2022]. Different stakeholders need to gather adequate and appropriate information to facilitate strategies for preventing and managing MSK diseases. The development of computerized patient registries in orthopedics could achieve this.

Orthopedic registry platforms have primarily aimed to collect data on joint replacements and other orthopedic surgical procedures. The International Society of Arthroplasty Registries (ISAR) is currently being conducted in 31 countries [Magaway and Malanga, 2022]. Several U.S. national registries have also been utilized, such as the National Trauma Data Bank (NTDB), the Veterans Affairs and American College of Surgeons National Surgical Quality Improvement Programs (NSQIPs), the Kaiser Permanente National Total Joint Replacement Registry (TJRR), Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR), the American Joint Replacement Registry (AJRR), the Musculoskeletal Tumor Registry (MsTR), the Shoulder and Elbow Registry (SER), the American Spine Registry (ASR), and the Fracture and Trauma Registry (FTR)[Magaway and Malanga, 2022]. The last five branches of the registry were started in 2017 by the American Academic of Orthopedic Surgeons (AAOS), and our research group at Mass General Brigham is responsible for supplying data.

The need to collect patient data in the form of registries comes from the lack of standards and procedures in many institutions, where all of the data is collected but appears to be easily untraceable when needed for projects later. In many cases, data coming from the emergency room is not adequately stored or labeled due to the lack of time, and this problem often extends to general clinicians. Therefore the application of artificial intelligence to help with the automatization of these tasks can enable a large number of projects in the future.

The application of artificial intelligence (AI) models has been increasing in both biomedical studies and clinics, showing their potential to improve different aspects of medicine, from screening to treatment[Castiglioni et al., 2021]. A growing body of evidence shows that using A.I. models to detect and classify bone fractures could save a considerable amount of labor, time, and resources[Meena and Roy, 2022]. The potential of AI-based technologies to monitor patients, identify disease clusters and estimate the prognosis of disease has made us propose an AI-aided registry database with which the clinical decision-making process could be improved drastically. Therefore, we question if creating an X-ray registry based on AI models is possible, where the task of clustering the images in different classes is done automatically. Based on the previous evidence, it should be possible to do, but testing it with specific patient data is necessary to validate its real impact and confirm the theoretical approach.

2 Artificial intelligence in medicine

Artificial intelligence has been defined in many ways. Still, in simple terms, we could explain it as the science branch that helps artifacts such as computers make intelligent decisions based on the input received. This comes with a set of technologies that allows machines to simulate human intelligence. In many cases, simple tasks that humans can develop can be hard to replicate in a machine, especially those involving our senses, like understanding words or recognizing objects. The advantages and disadvantages of AI in medicine have been highly discussed in previous review papers such as [Yu et al., 2018] or [Rajpurkar et al., 2022], showing that AI is not only feasible but can bring numerous advantages to several fields of medicine. AI would allow for personalized healthcare as well as improve diagnosis and help physicians to accelerate their work.

2.1 Traditional machine learning

Traditional machine learning models require patient data to make predictions, including age, sex, and previous diseases, but can also include test results, medication, or symptoms. The models can be supervised to classify or predict certain diseases or conditions present in the patient. Still, they can also be unsupervised to extract more specific features that are not visible at first. These models are highly relying on statistics, making them closer to traditional approaches and usually easier to understand. It is the most widespread branch of artificial intelligence since it can be helpful in endless situations. As long as there is patient data, and some kind of prediction or classification involved, there is always room to attempt to solve the problem in this fashion. Uses in medicine can vary from prognosis to diagnosis or even insurance approval which, whether we want it or not, forms a part of the medical world [Rajkomar et al., 2019]. In most cases, the machine learning models end up as tools for clinicians, who always have the last saying on every decision.

2.2 Deep learning

Stepping aside from the classical models and by the hand of neural networks, more complex models have emerged through the years; the complexity allows for the models to extract small details from the data, but at the same time, they require larger amounts of information to extract these details. This opened the path to exploring images through machine learning and deep learning. Images play a distinguished role in many fields of health, and nowadays, they are obtained through a wide range of varieties such as X-Ray, MRI, CT, or ultrasound as some of the multiple variations. Using convolutional neural networks (CNNs)[Shin et al., 2016], many features can be extracted from the data, helping to predict or classify elements like cancer cells or fractures. The main focus will be on imaging techniques for the detection of lesions as well as classification techniques in medicine. These tasks have been approached in many ways, from MRI to micro-CT, showing the capability of developing a fully functioning algorithm to predict fractures[Roth et al., 2016, Chmelik et al., 2018, Zhou et al., 2018, Liu et al., 2021] or conditions [Farooq and Hafeez, 2020]. Many previous advances have shown good theoretical results but need to be generalized and put into practice.

2.3 Computer vision

The main type of ML or DL used for computer vision and image data is CNNs. With the advance of technology and simultaneous increase in computational power, they have been popularized once again after remaining inefficient with old computing capabilities. The recent availability of large-scale annotated datasets has allowed advances as more accurate and generalizable models have been developed[Shin et al., 2016]. The tasks that can be achieved through this technology are countless, as well as the different possibilities within architectures [Chea and Mandell, 2020]. Just in medicine, they have been increasingly used for lesion detection, classification, segmentation, and non-interpretive tasks. Some specific uses stated by Pauley Chea et al. include cartilage segmentation, skeletal bone age assessment, or automated quantification of osteoarthritis[Chea and Mandell, 2020].

2.3.1 Convolutional neural networks

CNNs are a machine learning model that extracts features from an image, gathering those considered important and discarding the non-relevant ones. The output can vary from a classification to a prediction or regression. Many architectures have developed through the years to achieve the maximum potential for every specific problem. Some of the most important variations in health are classic CNNs, U-Net, and GANs.

As the most used structure for image processing, many variations and types of architectures originated from it. The basic shape consists of a series of convolutional layers that apply filters to the pixel data to extract the features from the desired image. Each filter will make the image smaller and extract more relevant features. The last layer can vary depending on the purpose of the model, but mainly it is a prediction of whether there is something in the image or if it belongs to a certain class. There are already many algorithms that help find bone fractures and lesions in different types of medical images, such as the one from Holger R. Roth et al. detecting fractures on spine CT where they achieved an area-under-the-curve of 0.857 while marking the location of 55 displaced posterior-element fractures in 18 trauma patients[Roth et al., 2016] or Jiri Chmelik et al. classifying difficult to define metastatic spinal lesions on 3D CT data[Chmelik et al., 2018].

2.3.2 Classification

From the previously mentioned task, we will focus on classification, as this work aims to enable the automatic classification of X-rays to enable a better way of dealing with the great amount of data generated every day.

Aside from medicine, where it has numerous purposes[Kim et al., 2019, Litjens et al., 2017], many innovative breakthroughs have been coming up in recent years in the classification task. Remarkable issues include the creation of large datasets in common objects such as IMAGENET[Deng et al., 2009, Russakovsky et al., 2015]. This has allowed the appearance of many pre-trained models that can be used in problems with smaller datasets. Some of the models that have made a difference in the field include AlexNet, which in 2014 changed the way of looking at deep convolutional neural networks[Krizhevsky, 2014], GoogleNet[Szegedy et al., 2014], ResNet[He et al., 2015].

More recently, several new ways of approaching CNNs came to the scene, bringing different perspectives and innovations in different niches. One of the main points in the newer methods is the efficiency in the training for the deployment of models in small devices. EfficientNet has it on the name, and MobileNet follows its steps to aim for an even smaller number of parameters[Howard et al., 2017, Tan and Le, 2020]. Finally, with the appearance of attention models and the spread of transformers from natural language processing to the rest of the fields of AI, both have made their entrance into the field of computer vision[Dosovitskiy et al. [2021], Dai et al. [2021].

3 Model architecture

From the main classification architectures ResNet101, EfficientNet, and Vision transformer were chosen to tackle the registry problem. The reason behind this decision was their performance on the ImageNet dataset for both Efficient Net and the transformers as well as the standard benchmarking purposes of ResNet101.

Another reason was the difference in architectures and purposes that these models entail, bringing variety to the experiment and achieving a greater scope on the possible final model.

3.1 ResNet101

Residual neural networks were designed in order to ease the process of training and learning of deep neural networks. One of the main challenges in training deep neural networks was the problem of vanishing gradients, where the gradients of the parameters with respect to the loss function become very small as the data flows through the network, making it difficult to optimize the parameters using gradient descent. This can be particularly problematic in very deep networks, where the gradients can become so small that they effectively "vanish" and the network is unable to learn. Unexpectedly, the gradient degradation is not caused by overfitting and adding more layers to the model lead

to higher training error, as reported in the paper by Christian Szegedy et al. "Going deeper with convolutions"[Szegedy et al., 2014]. To address this issue, the concept of "residual connections" was introduced, which allowed the gradients to bypass one or more layers in the network to learn even when it is very deep[He et al., 2015].

This type of architecture has been widely used in medicine for the classification of images, mainly in the detection of diseases or conditions[Zhang et al., 2020, Farooq and Hafeez, 2020], and being one of the most used models through all the fields, conforms a great benchmark tool. In many cases, smaller models have been used such as ResNet18 or ResNet50[Zhang et al., 2020, Farooq and Hafeez, 2020, He et al., 2015]

3.2 EfficientNet

EfficientNet is a family of convolutional neural network (CNN) models developed by Google that have achieved state-of-the-art accuracy on various image classification and object detection tasks while being smaller and faster than many other models. The goal of EfficientNet is to improve the efficiency of CNNs, which can be measured in terms of the number of parameters (weights and biases) in the model, the amount of computation required during training and inference, and the number of bits required to represent the model.

EfficientNet models are constructed using a compound scaling method, which scales the model's dimensions (such as the width and depth of the network) and the resolution of the input images. By scaling all of these factors together, the authors of EfficientNet were able to achieve better performance with fewer parameters and less computation than other state-of-the-art models.

EfficientNet was introduced in a paper published in 2019 by Mingxing Tan and Quoc V. Le: "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." [Tan and Le, 2020]. Since its introduction, EfficientNet has been widely adopted and has achieved state-of-the-art results on various tasks, including image classification on the ImageNet dataset, object detection on the COCO dataset, and face recognition. It has also been used in several applications, such as image classification in mobile devices and large-scale image search engines[Howard et al., 2017, Dosovitskiy et al., 2021].

3.3 Vision Transformer

Vision transformers are a type of neural network architecture that has recently gained popularity in computer vision. They are based on the transformer architecture, originally developed for natural language processing tasks such as machine translation and language modeling.

The transformer architecture is known for its ability to process sequential data in a parallelized manner, using self-attention mechanisms to capture long-range dependencies in the input data. Its architecture makes it particularly well-suited for tasks that require understanding context and relationships between words or tokens in a sequence.

In computer vision, vision transformers have been used to perform a wide range of tasks, including image classification, object detection, and segmentation[Atito et al., 2021, Paul and Chen, 2021]. They have also been used for video recognition, visual question answering, and reasoning. The key concept is the processing of image patches sequentially imitating sentence processing to achieve an understanding of the input[Dosovitskiy et al., 2021].

Several vision transformer models have been developed and shown to achieve state-of-the-art performance on various tasks in computer vision. Some examples include ViT (Visual Transformer) by Google Research, DeiT (Distilled Education-aware Image Transformer) by Facebook AI, and ResNet-Transformers by Microsoft Research[Dosovitskiy et al., 2021, Touvron et al., 2021, Zhang et al., 2021].

4 Data Analysis and Processing

Two datasets were used for the creation of our patient registry pipeline. The main one was the body part dataset, allowing us to classify each X-ray into its body part and also to identify the presence of an instrument in the X-ray. The second dataset was the hip fracture dataset, serving as a proof of concept. Our pipeline's final goal is to be able to classify the X-rays into a body part, the presence of instruments, and the presence of a fracture. But due to the lack of fracture data, this had to be done separately for the moment.

4.1 Body part data

The dataset for body part classification and instrument detection consists of a set of 51262 high-quality X-rays from 9 different categories, as seen in table 1: ankle, elbow, foot, hand, hip, knee, shoulder, spine, and wrist. In each of the categories, the data is divided into two different categories, Instrument vs. no instrument; in the instrument category, some element has been placed, such as nails, screws, or prostheses. Some exclusion criteria were applied, removing images including body parts with a great level of deformity, amputations, or operation room X-rays that include many pieces of material such as scissors. The images were screened by both physicians and medical students to ensure that every image corresponded to its category, both for body parts and the presence of instruments.

Body part	Instrument	No Instrument	Total
Ankle	915	5948	6863
Elbow	579	5157	5736
Foot	498	5960	6458
Hand	197	5323	5520
Hip	1426	2647	4073
Knee	1369	5002	6371
Shoulder	591	5038	5629
Spine	950	3755	4705
Wrist	523	5384	5907
Total	7048	44214	51262

Table 1: Summary of all the images present in the system.

All the images selected for the study were deidentified so that the identity of the patients could remain private. As visible in the distribution, the categories are not too different in numbers, which made it easier for division of data between train, test, and validation. After removing images that were not useful, the final dataset was composed of 4320 images for testing, 8640 for validation, and 30240 for training.













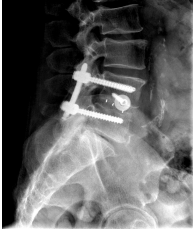


Body part	Instrument	No Instrument	Body part	Instrument	No Instrument
Ankle			Elbow		
				Hand	
Hip			Knee		
				Spine	
Wrist					

Figure 1: Sample of each category showing both hardware and no hardware images.

4.1.1 Class similarities

Some of the categories present similar cases of images. Due to this problem, medical professionals revised the data to come to a consensus on how to divide overlapping categories.

Ankle vs. Foot. In the foot and ankle case, the main point was lateral views that included some parts of the leg. For this problem, the chosen threshold was the fingers. If all of the fingertips are visible in the X-ray, it will be considered a foot X-ray. Otherwise, it will belong to the ankle category.

Hand vs. Wrist. The consensus for this problem was similar to the previous one. If all fingertips are included, the image will be classified as hand. On the contrary, an image that does not show the fingertips and usually shows more parts of the arm will be the wrist.

Hip vs. Spine. This case presents more difficulties to separate. The spine category shows several types of images, ranging from the head all the way to the sacrum, so the separation between frontal or AP hip X-rays and a lower spine one is ambiguous. The final medical approach was to classify as hip those images that show a complete view of the hip.

4.1.2 Processing

In the case of the body part classification, the data classes were sufficiently balanced to train the models. Therefore the split was straightforward, and the data were divided randomly into three sets, 10% for testing, 20% for validation, and the remaining 70% for training.

On the contrary, in instrument detection, the distribution of the data was not even. The division was 85% of the images without an instrument and only 15% with the presence of it. Therefore some strategies were considered in order to balance the dataset. Finally, and in order to keep as much information as possible, weighted random sampling provided by PyTorch was used to upsample the minority class, with no instrument. For this task, an array was created with the probability of each individual case to be selected by the training data loader.

To achieve a greater generalization of the models and better performance, data augmentation was performed on the training dataset, generating more data points for the training phase. Simple methods were used, random flip on the vertical axis and a random rotation of 90 degrees. Other methods, such as cropping, were not viable since images would change the category if some parts were removed. The training data was also used to extract the mean and std for the channels of the dataset. After this was done, normalization was performed to standardize every image.

To ensure the correct input to the model, images were resized to 224x224. The choice was made based on the architectures selected for the problem, the standard model input for the three models is 224x224, and the tasks were not complex enough to need more information.

4.2 Hip fracture data

The Hip fracture data were collected from a study based only on the detection of fractures in the The exclusion criteria were 1) the presence of hardware, 2) missing one view, 3) having any artifacts on the images, fractures in the shaft without any fracture in the proximal part, and lesions other than fractures that change the normal texture of the bone including tumors, union deformities after a previously detected fracture, cysts, masses, etc. Images were screened by three expert physicians who were trained and familiar with the detection of hip fractures on radiographs. In cases where there was a lack of consensus among the reviewers, patients' records, including CT and/or MRI if available, examinations, and reports from outside the center were screened to reassure the diagnosis. In total, 220 patients (440 radiographs) with proximal femoral fractures were detected and assigned to the case group. From the same data source, 261 individuals (522 radiographs) who did not have any history of fracture or trauma to the femur were assigned to the control group.

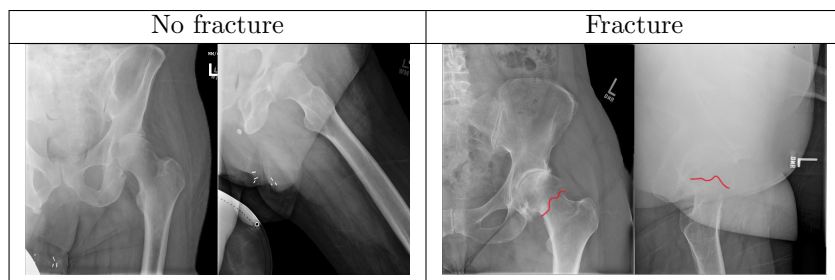


Figure 2: Hip X-rays combining lateral and AP views.

4.2.1 Data preparation

The dataset comprised high-quality radiographs that were directly obtained from the data management software used in our institution (EPIC, Ver. 2021) in a .jpeg format. Data augmentation was performed. This included a random rotation between -90 and 90 degrees and a random flip of the image. The dataset had a total of 962 image sets (AP and Lateral combined) afterward. The outcome of our algorithm was the detection of hip fracture; thus, it was a dichotomous outcome of yes/no fracture. Our CNNs (convolutional neural networks) were trained using the combined views, AP, and Lateral together, which could enable us to improve the accuracy of diagnosis compared to using only one view in the training. The dataset was randomly split with a 70:20:10 ratio into a training set, validation set, and test set, respectively.

Images were merged to contain both the AP view and lateral view. Before feeding the model with the images, they were normalized following the training dataset mean and std and resized to (224,224).

To ensure that the model would fit the data, the input layer and the last layer were modified to accept the size of the combined radiographs. The result was a dropout layer followed by a single-unit dense layer to perform the classification, 1 for fracture and 0 for no fracture present in the radiograph. The models were then trained on the training set of each of the combinations, using pretrained weights provided by Torchvision, and validated on the validation set. For this process, Adam Optimizer, with a learning rate of 0.001, was used. The batch size was kept at 32, and the epochs were limited to 100 with no convergence stop before the 100 epochs. Finally, the performance was measured on the test set, completely independent from the previous two training and validation sets.

5 Methods

For each of the three problems, the approach was the same; the three mentioned models were used to tackle the challenges. The model’s weights were loaded from torchvision available weights from ImageNet training. After that, the last layers were modified to fit each of the problems and models. Finally, the models were retrained on the new data with a low learning rate to keep all the learned information.

5.1 Body part classification

For the initial problem, which was classifying body part X-rays into 9 different classes, several models were considered. The final decision was EfficientNet, ResNet, and Vision Transformers, entailing different branches of computer vision architectures that have been proven to achieve outstanding results in multi-class classification problems, such as in the ImageNet dataset. To implement these models, the approach selected was transfer learning, and the framework was PyTorch, as torchvision provides the weights for the three of them trained on IMAGENET. Pytorch also provides a CUDA connection to make the training faster through the use of GPU.

5.1.1 Architecture modifications

The architecture of the models was modified to fit the 9-class classification, as mentioned before. For ResNet 101, the last fully connected layer was substituted by a ReLU layer, a Dropout layer, and a fully connected linear layer with 9 classes. For the other two models, the same modules were added. In the EfficientNet case, the classifier module was removed, and in the Vision transformer, it was the heads. This way, we achieved a customized outcome for our specific challenges.

5.1.2 Training

For the training process, there are several aspects to have in mind; the optimizer selected was Adam, and the learning rate is an important factor to take into account. As the models are already loaded with ImageNet weights, we did not want to use a large learning rate, but since it is transfer learning and not feature extraction, the weights still need to be modified. A learning rate scheduler was used

with several steps, starting at 2×10^{-5} and multiplying the learning rate by 0.5 after several milestones. The batch size was set to 16, and the number of epochs was set to 100, which was a sufficient number for convergence, but also a convergence threshold was established to avoid overtraining and overfitting the model to the training set.

Finally, the loss function used was cross-entropy loss with a label smoothing of 0.1 to achieve better performance and cover the small imbalances in the dataset. This method for training was used for the three proposed models.

5.2 Instrument detection

The problem of instrument detection was more straightforward than the body part classification, as it is a simple task without any type of conflict, whether there is or not. For this problem, tubes, syringes, lines, intravenous (IV) lines, and casts are not considered instruments. On the other hand, metallic hardware or instruments will be included in the positive category. Some examples include but are not limited to prostheses, nails, screws, plates, spacers, or k-wire.

5.2.1 Model and training

Instrument detection can be categorized as a binary classification task. In this category, models such as the previously used ones also tend to perform well. Therefore we decided to use the same three models. The approach of transfer learning was the same as from the previous problem, keeping the learning rate to a small number in order to not modify too much of the information learned in the previous dataset. But in this case, the final classification was binary, so only two classes needed to be predicted.

During the training, the approach was the same as the previous case for most of the parameters. The batch size was 16, the maximum number of epochs was 100, and a convergence threshold was set to avoid train-set overfitting. The learning rate was again 2×10^{-5} and multiplied by 0.5 every 10 epochs. In this case, there was no label smoothing, but as in mentioned in the data analysis section, the data was not balanced, and a weighted random sample had to be implemented in the training data loader.

5.3 Fracture detection

Finally, for fracture detection, only the model for Hip fracture was developed, serving as a placeholder and proof of concept for the future steps in the project. The problem presented itself as a binary classification but can be extended in the future to include several types of fractures in the classification. Therefore, the models used were really similar to the previous ones.

5.3.1 Model and training

In this case, the input of the model was two images. To keep the proportion of the images, the resize was done to 448×224 . Therefore, the input layer of the models had to be modified. In the case of the ViT, there was a problem when trying to set the input to the one mentioned, as the model provided by torchvision would not allow for the change. Therefore, for the ViT, the input remained as in the previous approaches, 224×224 , and results are expected to be worse than in the cases of ResNet and EfficientNet.

For the remainder of the process, the same approach as in instrument detection was used in every way.

5.4 App

To combine the functionalities of the three problems into one framework or tool, a web application was developed using Django in Python. The main purpose was to showcase the results and allow the potential user to test the 'model'.

The use of the app was developed to be easy for the end user, several images could be uploaded to the application. The images were transformed and normalized following the training dataset distribution and fed into the BPC model, yielding a result for the total of the images. When the body part class is selected, the images are then processed for the detection of instruments.

Finally, for fracture detection, images can be uploaded by pairs conformed by a front and an AP view. If this is the case, the images are joined into one and then processed to be fed into the

corresponding fracture detection model. The only fracture detection model implemented is the hip, with the possibility to include future developed models for the rest of the body parts, either for fracture detection or any other purposes.

An alternative local application was also developed to organize folders by body part and the presence of instruments. This app can be run through the terminal with Python, and a folder to organize would be the input, and the application will organize the files in a different given folder.

6 Results

To perform tests on the different models, we need to differentiate the body part classification and the instrument detection, which both were trained in a large dataset in comparison to the fracture detection algorithm, which dataset was notably smaller, but also the difference between multi-class classification and binary classification.

6.1 Body part classification

The results for the body part classification were similarly excellent for the three of the models, achieving accuracies up to 99%. This is due to the fact that the body parts that were chosen to classify present great differences in their anatomy. The classes that were closer in anatomy are clearly the ones that presented more misclassifications. As seen in figures 3, 4 and 5, the class with the most misclassifications for the three models is the ankle, where 8-10 images, 0.017%-0.021% of the ankle images were classified as a foot. This was one of the main class similarities reported previously. As for the other class similarities, the three models were capable of differentiating the classes with a maximum of 3 misclassified images and only one in most cases.

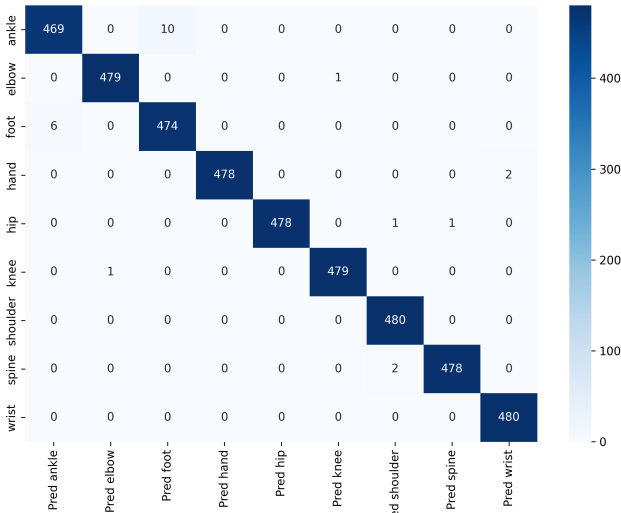


Figure 3: ResNet101 confusion matrix of body part classification.

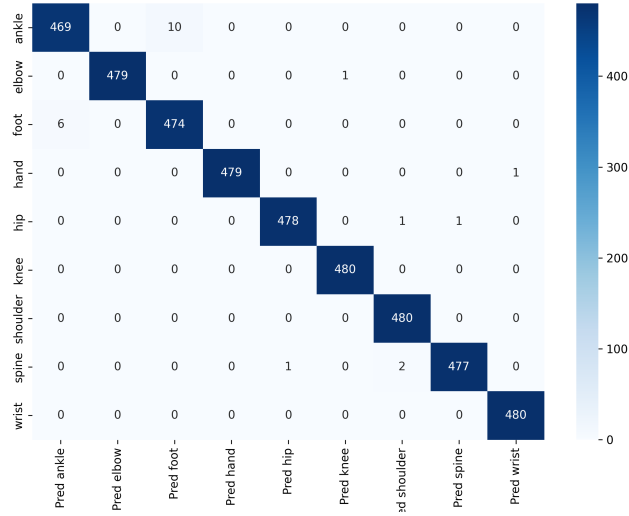


Figure 4: Efficient Net confusion matrix of body part classification.

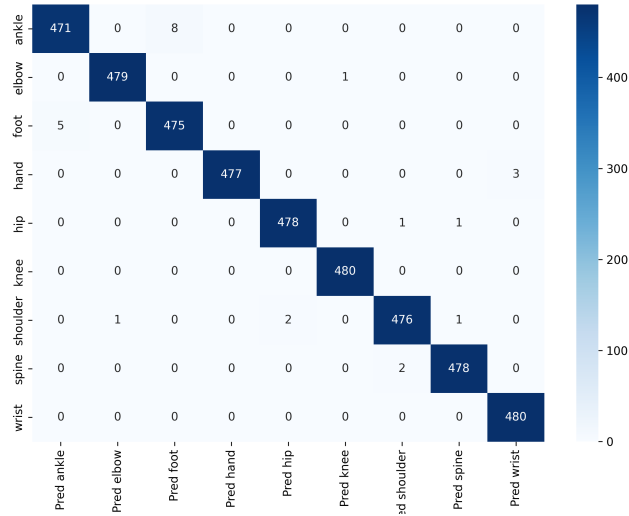


Figure 5: Vision transformer confusion matrix of body part classification.

To compare the models, calculating one vs. all ROC-AUC was intended, but it is almost impossible to distinguish them as they are all too close to 1.00. Therefore to compare the models, other metrics were selected to detect differences in an easier manner.

Model	Logarithmic loss	Matthew's correlation coefficient
ResNet101	0.1098	0.9937
EfficientNet	0.1053	0.9940
ViT	0.1097	0.9934

Table 2: Results of body part classification.

The metrics chosen in this case were Logarithmic loss, also referred to as cross-entropy loss. The interesting side of this metric is its inherent use for probabilistic cases, which matches our scenario.

The binary formula can be found ahead, but it can be extended for multiple classes:

$$L_{\log}(y, p) = -(y \log(p) + (1 - y) \log(1 - p))$$

The other metric included is Matthew’s correlation coefficient (MCC), which involves true positives, true negatives, false positives, and false negatives. It is a symmetric measure unlike precision and recall and ranges between -1 and 1, being 1 a perfect classifier. In our case, similarly to the AUC, it is close to 1 but reassures the fact that the EfficientNet is the best model. Matthew’s correlation coefficient formula is the following.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

As seen in the table 2, the metrics are close to each other, but if we have to choose a model, EfficientNet would be in the first place. It performs slightly better on both metrics, but we have to go to the third decimal place in order to find the difference.

6.2 Instrument detection

Instrument detection is again a simple task for humans but can present some more difficulties for the machine. The main reason behind this statement is the presence of miscellaneous elements in the images, such as IV lines, tubes, syringes, or even rings in hand X-rays. This can generate confusion as many of the mentioned items can appear in similar shapes or brightness as the items we want to detect, such as nails, plates, or screws.

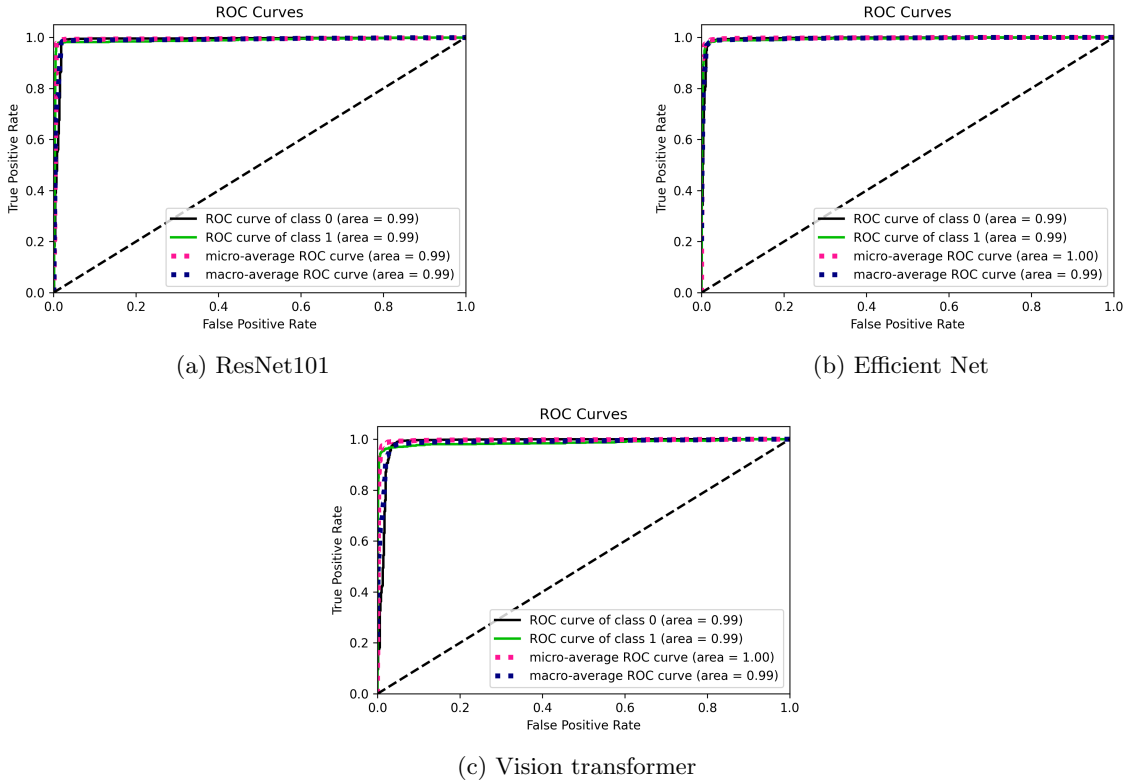


Figure 6: ROC AUC for instrument detection.

The results achieved were remarkable, as seen in table 3, achieving AUC values of 0.99, a sensitivity of 0.97-0.95, and a specificity of 0.99-0.98 for the three models. The misclassification varies slightly between the three of them. Still, the overall result follows a similar fashion, with the best results being achieved by the ResNet101, getting almost identical results to the EfficientNet, but both are slightly

better than ViT. The amount of data presented to learn and solve the problem was enough to solve the task in a great way, and this allowed all of the models to learn well the differences between the two desired classes.

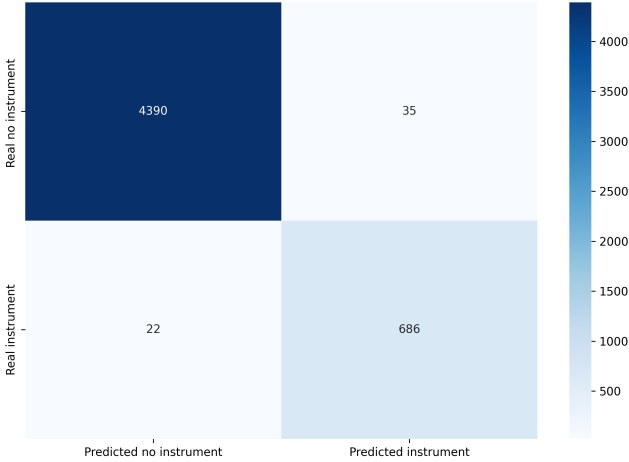


Figure 7: ResNet101 confusion matrix for instrument detection.

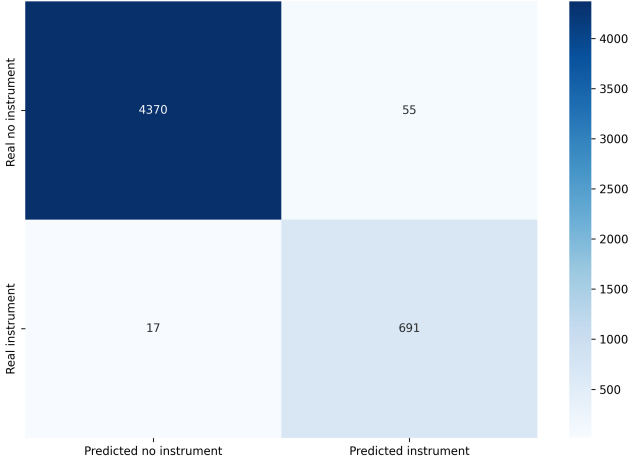


Figure 8: Efficient Net confusion matrix for instrument detection.

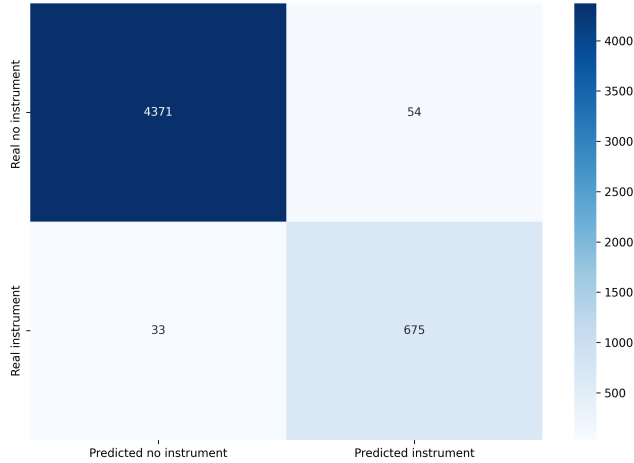


Figure 9: Vision transformer confusion matrix for instrument detection.

Model	AUC	Sensitivity	Specificity	Youden In.	F1 Score	Accuracy	PPV	NPV
ResNet	0.99 (0.97-0.99)	0.97 (0.96-0.98)	0.99 (0.99-0.99)	0.96 (0.95-0.97)	0.96 (0.95-0.97)	0.99 (0.99-0.99)	0.95 (0.94-0.97)	0.99 (0.97-1.00)
EffNet	0.99 (0.97-0.98)	0.97 (0.95-0.98)	0.99 (0.99-0.99)	0.96 (0.94-0.97)	0.95 (0.94-0.96)	0.99 (0.98-0.99)	0.94 (0.92-0.96)	0.99 (0.99-1.00)
ViT	0.99 (0.96-0.98)	0.95 (0.94-0.97)	0.99 (0.98-0.99)	0.94 (0.93-0.96)	0.94 (0.93-0.95)	0.98 (0.98-0.99)	0.93 (0.92-0.96)	0.99 (0.99-0.99)

Table 3: Results of the models in instrument detection.

After getting the results, we performed some analysis of the miss classified images. For the false positives, in almost every case, some other miscellaneous element was present in the x-ray, while thin screws or strange angles mostly covered the false negatives where the instrument was barely visible. Some misclassifications were also detected, and the errors were divided into the main classes found.

Model	Misslabeled	FP	FN	HighSatFP	SpineFP	KneeFP	SpineFN	FootHandFN	PoorQ
ResNet101	12-9	13	2	15	5	6	4	7	2
EfficientNet	14-8	13	4	5	7	3	5	11	2
ViT	14-10	13	8	3	4	2	15	14	4

Table 4: Missclassifications on instrument detection.

We can see in table 4 that depending on the classifier, the errors were divided in different ways, EfficientNet has an even distribution of the errors, making it the most reliable model when generalizing for different scenarios. On the other side Vision transformer shows more problems detecting instruments in spine X-rays and foot or hand X-rays. Finally, the ResNet has difficulties distinguishing high-saturation images from instrument ones.

6.3 Fracture detection

The fracture detection problem is presented in the project as a potential pipeline application; hence it is not supported by the same dataset or amount of data. This, in addition to the jump in complexity from the other parts of the system, makes it the worst-performing section. Although this task’s performance is lower than a professional radiologist’s, it is typically trained with just over 450 patients,

making it too complex for the machine to learn in such a low amount of cases. It would be interesting to investigate the accuracy of a novel radiologist with as little experience as 450 images would achieve and compare it with the one our model achieves. Some studies have been carried on previously, showing that postgraduate year physicians show a median accuracy of 0.70, a sensitivity of 0.58, and a specificity of 0.82 for predicting hip fracture in X-rays. On the other hand, experienced physicians such as orthopedic surgeons or radiologists show median accuracies of 0.94-0.96, median sensitivities of 1.00-0.99, and median specificities of 0.87-0.94, respectively [Cheng et al., 2020]. The results of the novice physicians show a great improvement when accompanied by the model. also shown in Duron et al. study of 2021, Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study [Duron et al., 2021]

Our results show that there is room for improvement to reach the level of experienced physicians, but also show the need for more images, as our model with a lower amount of training images outperformed low-experience physicians.

Model	AUC	Sensitivity	Specificity	Youden Index	F1 Score	Accuracy	PPV	NPV
Resnet101	0.93 (0.82-1.00)	0.86 (0.65-1.00)	1.00 (1.00-1.00)	0.86 (0.65-1.00)	0.92 (0.79-1.00)	0.94 (0.84-1.00)	1.00 (1.00-1.00)	0.90 (0.75-1.00)
EffNet	0.90 (0.77-1.00)	0.79 (0.65-1.00)	1.00 (1.00-1.00)	0.79 (0.65-1.00)	0.88 (0.75-1.00)	0.91 (0.78-1.00)	1.00 (1.00-1.00)	0.88 (0.81-1.00)
ViT	0.81 (0.65-0.91)	0.57 (0.41-0.82)	1.00 (1.00-1.00)	0.57 (0.41-0.82)	0.73 (0.52-0.90)	0.81 (0.67-0.94)	1.00 (1.00-1.00)	0.75 (0.57-0.91)

Table 5: Results of the models for hip fracture detection.

As seen in figures 10,11,12 and 13,14,15 the ViT model performs the worse out of the three proposed methods, as mentioned in the Model architecture section. The input of this model had to be resized to smaller dimensions than the other two due to the fixed architecture of the model provided by torchvision. This causes information loss and therefore is one of the causes of the lower performance in the model.

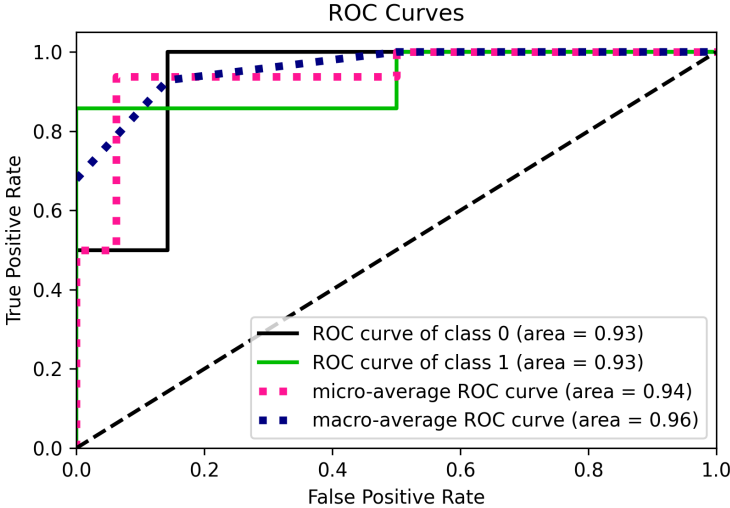


Figure 10: ResNet101 ROC AUC for hip fracture detection.

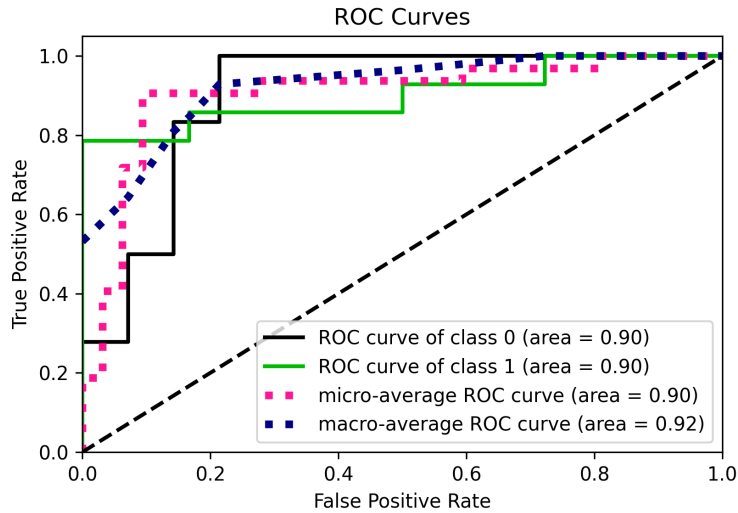


Figure 11: Efficient Net ROC AUC for hip fracture detection.

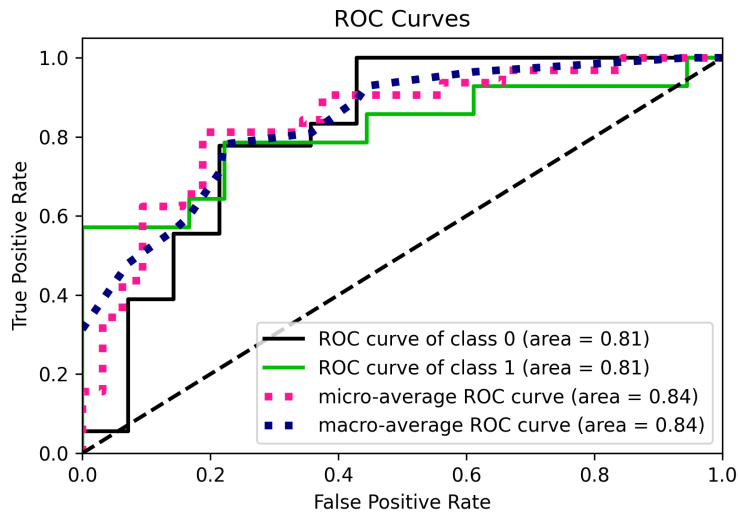


Figure 12: Vision transformer ROC AUC for hip fracture detection

Again, in the binary classification task of detecting a fracture, the ResNet101 performed the best with an AUC of 0.93. It is important to note that the ResNet model never misclassified a real fracture, which is the most important issue in medical problems such as this one, as missing a fracture could potentially cause a worse scenario in the patients' health. Although the ResNet presents itself as the best performing model, it has to be mentioned that Efficient net achieved almost identical results, with only one more error, in almost half of the epochs of training(45 vs. 87).

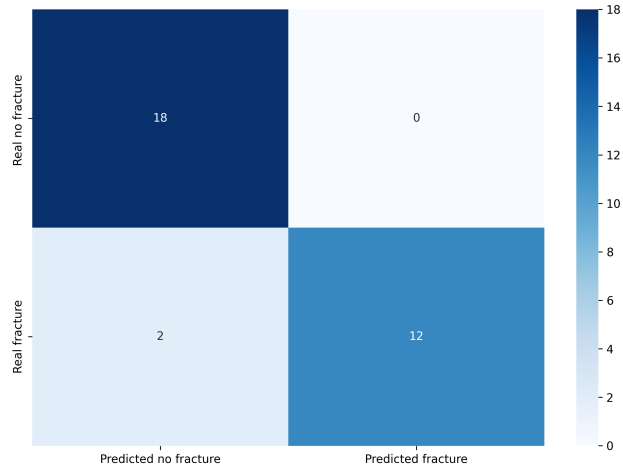


Figure 13: ResNet101 confusion matrix for instrument detection.

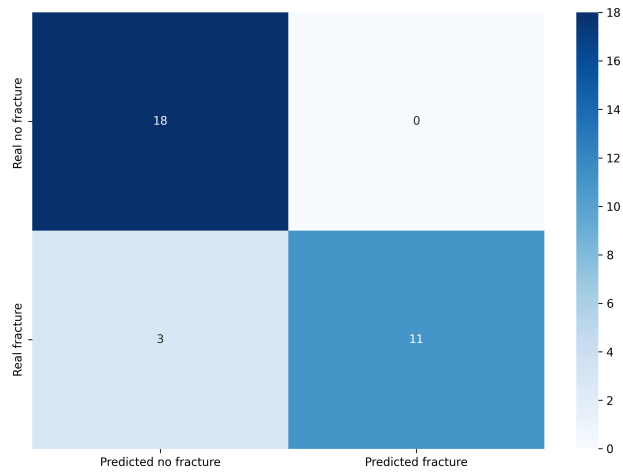


Figure 14: Efficient Net confusion matrix for instrument detection.

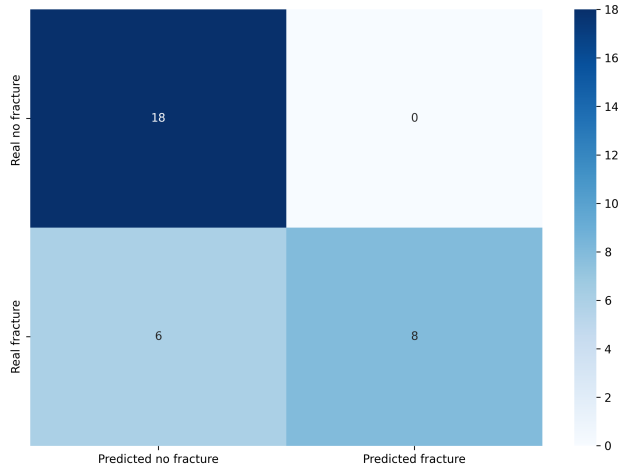


Figure 15: Vision transformer confusion matrix for instrument detection.

It is visible from the results that the mistakes in the classification come when finding fractures, as the specificities shown in the three tables are 1.00. The models tend to have problems finding the fracture rather than confusing other features with a fracture. This problem can be solved by imputing higher-resolution images into the pipeline instead of downsizing the images to 448x224. Still, also, with the addition of extra x-rays, the models could improve, as 468 cases are barely enough to learn for the computer.

7 Explainability

When it comes to explainability and interpretability, machine learning and especially deep learning models indeed entail more complex problems than the classical models or statistical approaches. Even more, as the models have evolved and improved, their complexity has increased too. This leads to a clear trade-off between the performance of the models and our ability to explain and interpret them. This would not be a problem in other fields, but in medicine, it is necessary to know how the decisions are made, as the health and life of the patients are in our hands.

But what do we mean when we talk about interpretability and explainability? There have been several papers trying to define these widely used concepts, especially in the field of medical technology. However, they are vaguely understood. F. Doshi and B. Kim establishes the idea of interpretability as "the ability to explain or to present in understandable terms to a human", while on the other side, explainability can be defined as the ability to explain what the model does. This can vary depending highly on the type of model or problem that it is being faced [Doshi-Velez and Kim, 2017]. In the case of black box models, explainability has been addressed in many ways. M. Fox, D. Long, and D. Magazzeni surveyed a wide variation of black-box models to "open" the box and understand every model separately and to be used for similar approaches[Fox et al., 2017].

A different view on the problem has led to the development of algorithms that try to explain the existing models. Although there are machine learning models such as decision trees or statistical analysis that are easily interpreted, black box machine learning and deep learning are the best when it comes to performance, and trying to explain them is not easy. Therefore models like saliency maps or GRAD-Cam have been implemented.

Many methods have been reviewed by P. Linardatos et al. [Linardatos et al., 2020], but we will be focusing on those which would be useful for the previously discussed models.

Several methods have been proposed to understand the complexities of machine learning, and the most succeeding among those tend to be the ones that assign different values to the different inputs. The gradient is a methodology that quantifies how an input dimension could change the output values [Simonyan et al., 2013]. Those methods were improved and used as a base for the forthcoming.

Another type of validated method for CNN explanation is the Grad+CAM, coming from the gradients and the class activation map that can produce visual explanations for the important areas in the model choice [Selvaraju et al., 2020].

In essence, the trust in the models by the physicians is key to the progress of AI in medicine. A model or technology which is not understood can sometimes be overlooked as it is, thus, untrusted. All these methods bring some clarity to the decision-making, which can bring the physicians together to accept in an easier way new technologies in healthcare.

After performing some analysis of the models developed, we can see as an example several images and how they could come to help the decision-making, emphasizing the areas of the image that influence the models' decisions.

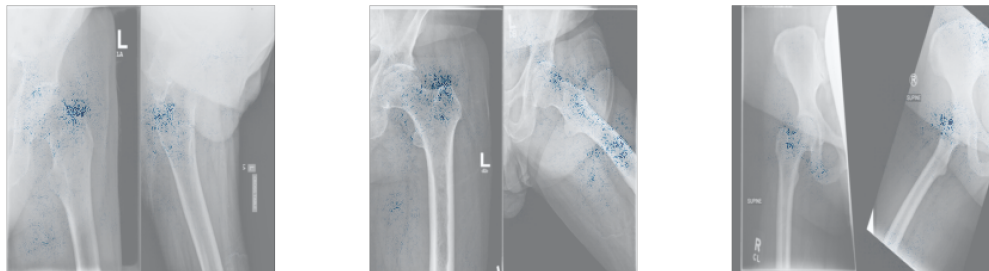


Figure 16: Saliency maps for hip fracture detection.

In figure 16, the images show different examples of what the original images and the saliency maps look like. This conforms to an essential part of the project's next steps; it is already achievable and can be used to speed the process and decision-making of physicians for those images that are not clear which class they belong to. Therefore this tool could not only become a trust source for the implementation of the automatic patient registry in hospital software but also a complementing tool for many physicians, especially the inexperienced ones, to perform their daily tasks in a more accurate manner.

Explaining what happens inside black box models such as neural networks is always complicated, especially for non-technical audiences, and even for knowledgeable audiences, it can bring some concern and lack of trust, depending on the input sources. Therefore, this topic must be addressed and could be one of the potential improvements for patient registries, where each image is stored with its respective "explanation" or heatmap of attention. This would help their quality assurance, becoming a second check in selecting the category for every image.

8 Discussion

The study shows exceptional results in the three sections, achieving remarkable metrics, especially in the body part classification task; no more than 30 images were misclassified in a 4320 image dataset, which is insignificant in this context. The same conclusions can be extracted from the instrument detection task, where the results were slightly worse, mainly due to the presence of miscellaneous objects in the X-rays. Even though the physician's selection process removed X-rays that were not standard, different objects can be present in normal X-rays that can lead the model to incorrect predictions. This step could be potentially added to the next stage of the pipeline, where undesirable images are removed from the pipeline. This would be done in a similar fashion to the other problems. The main limitation of this study is due to the origin of the dataset, as all of them are X-rays standardized by the Massachusetts General Hospital system. Therefore, the model can be overfitted to the dataset. Although being overfitted to the dataset provided by the hospital can seem like a problem, it can be solved by externally validating the system with new data and adding this external data to the training pipeline. The idea behind these models is to showcase the improvement that current systems would undergo by adding intelligent autonomous systems like the one proposed. Another possible downfall of the project is the thresholds established for the classification of each

part, as mentioned in the similarities subsection 4.1.1. Different institutions or countries can establish different rules, creating a confusing data input for algorithms as the one developed. Therefore, it would be necessary to reach a consensus on details like this. Nevertheless, this sometimes is close to impossible, and, in this case, training the models according to the available data categories can be an option. Adjusting to successful model categorizations would also work to take advantage of models trained by larger institutions that have access to more data. On the other side, the potential is clear. It is only left for institutions to implement and start using Deep Neural Networks to improve their data pipeline and create patient registries to enable future studies.

Finally, the proof of concept for hip fracture detection shows one of the multiple extensions that our algorithm could bring into the creation of patient registries. Creating endless opportunities and improvements in how we see research and data access at the moment. It has also opened the door to discussing whether it can be applied to different problems and challenges, given the great performance of the models. The simplest next step would be identifying different conditions in the X-ray domain, but why not take it further and test its functionality in the 3D spectrum? These models, including specific variations to adapt to the input, are worth trying and implementing to detect several conditions in images such as MRIs or CT scans.

Based on the performance of different algorithms, a different one could be implemented for each classification task, for example, dividing the data into the well-known ICD codes, which nowadays serve the purpose of looking for patient data, but if this could be automatized and become independent of the willingness of physicians to annotate every single case, new patterns, and diseases could be studied in the field of medicine.

After showing the feasibility of task automation, we can only conclude that the addition of the three models now, but many in the future, to a single platform is necessary to keep making progress in a field where in many cases, research fails due to lack of data, or data accessibility.

Other possible obstacles worth mentioning are the regulations in every country and the ethics behind such products. It has to be treated as an aid to physicians to guarantee the best practice and research possible, and minimum standards should be established to keep a high-quality health system. Approval by the government is necessary, as well as for the institutions in which it will be implemented. This can lead to a long process in many cases, especially with the low amount of regulations and trust in Artificial Intelligence. Regardless of the drawbacks, Artificial Intelligence and, more specifically, Deep Neural Networks have arrived in our times with the increase in computing power, and it is only in our hands to use them to improve our life quality and power our research to make the most out of the resources we have. A world with close to 8 billion people will keep creating data, and it is necessary to use it in the best way possible, and this starts with having it well organized.

9 Acknowledgements

I want to thank all the people who helped to make this project possible, working with the large amount of data collected. Special mention to Sanne Hoeksema, Nour Nassour, Huub de Klerk, and Alireza Ebrahimi. Also, my supervisors, who supported and helped me through the project, Soheil Ashkani and Hugo Schnack.

10 Code and data availability

The source code for the project can be found here: https://github.com/joseacitores/MSC_Registry

The data used for this project is confidential as it contains patient data owned by Massachusetts General Brigham and cannot be disclosed.

11 Bibliography

References

Richard E. Gliklich, Nancy A. Dreyer, Michelle B. Leavy, and Jennifer B. Christian, editors. *21st Century Patient Registries: Registries for Evaluating Patient Outcomes: A User's Guide: 3rd Edition*,

- Addendum.* AHRQ Methods for Effective Health Care. Agency for Healthcare Research and Quality (US), Rockville (MD), 2018. URL <http://www.ncbi.nlm.nih.gov/books/NBK493818/>.
- Eileen M Brooke and World Health Organization. The current and future use of registers in health information systems / Eileen M. Brooke, 1974. Pages: 43 p. Series: WHO offset publication ; no. 8.
- Gliklich RE, Dreyer NA, Leavy MB, editors. *Registries for Evaluating Patient Outcomes: A User's Guide [Internet]. 3rd edition. Rockville (MD): Agency for Healthcare Research and Quality (US); 2014 Apr. 1, Patient Registries. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK208643/>.*
- Cedric Lester Magaway and Gerard Malanga. Patient registries in orthopedics and orthobiologic procedures: a narrative review. *BMC musculoskeletal disorders*, 23(1):543, June 2022. ISSN 1471-2474. doi: 10.1186/s12891-022-05416-4.
- In-Hwan Oh, Seok-Jun Yoon, Hye-Young Seo, Eun-Jung Kim, and Young Ae Kim. The economic burden of musculoskeletal disease in Korea: a cross sectional study. *BMC musculoskeletal disorders*, 12:157, July 2011. ISSN 1471-2474. doi: 10.1186/1471-2474-12-157.
- Anthony D. Woolf, Jo Erwin, and Lyn March. The need to address the burden of musculoskeletal conditions. *Best Practice & Research. Clinical Rheumatology*, 26(2):183–224, April 2012. ISSN 1532-1770. doi: 10.1016/j.berh.2012.03.005.
- Isabella Castiglioni, Leonardo Rundo, Marina Codari, Giovanni Di Leo, Christian Salvatore, Matteo Interlenghi, Francesca Gallivanone, Andrea Cozzi, Natascha Claudia D'Amico, and Francesco Sardanelli. AI applications to medical images: From machine learning to deep learning. *Physica medica: PM: an international journal devoted to the applications of physics to medicine and biology: official journal of the Italian Association of Biomedical Physics (AIFB)*, 83:9–24, March 2021. ISSN 1724-191X. doi: 10.1016/j.ejmp.2021.02.006.
- Tanushree Meena and Sudipta Roy. Bone Fracture Detection Using Deep Supervised Learning from Radiological Images: A Paradigm Shift. *Diagnostics (Basel, Switzerland)*, 12(10):2420, October 2022. ISSN 2075-4418. doi: 10.3390/diagnostics12102420.
- Kun-Hsing Yu, Andrew L. Beam, and Isaac S. Kohane. Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10):719–731, October 2018. ISSN 2157-846X. doi: 10.1038/s41551-018-0305-z. URL <https://www.nature.com/articles/s41551-018-0305-z>.
- Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. AI in health and medicine. *Nature Medicine*, 28(1):31–38, January 2022. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-021-01614-0. URL <https://www.nature.com/articles/s41591-021-01614-0>.
- Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine Learning in Medicine. *New England Journal of Medicine*, 380(14):1347–1358, April 2019. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMra1814259. URL <http://www.nejm.org/doi/10.1056/NEJMra1814259>.
- Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, May 2016. ISSN 1558-254X. doi: 10.1109/TMI.2016.2528162.
- Holger R. Roth, Yinong Wang, Jianhua Yao, Le Lu, Joseph E. Burns, and Ronald M. Summers. Deep convolutional networks for automated detection of posterior-element fractures on spine CT. page 97850P, San Diego, California, United States, March 2016. doi: 10.1117/12.2217146. URL <http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2217146>.
- Jiri Chmelik, Roman Jakubicek, Petr Walek, Jiri Jan, Petr Ourednicek, Lukas Lambert, Elena Amadori, and Giampaolo Gavelli. Deep convolutional neural network-based segmentation and classification of difficult to define metastatic spinal lesions in 3D CT data. *Medical Image Analysis*, 49: 76–88, October 2018. ISSN 1361-8423. doi: 10.1016/j.media.2018.07.008.

- Zhaoye Zhou, Gengyan Zhao, Richard Kijowski, and Fang Liu. Deep convolutional neural network for segmentation of knee joint anatomy. *Magnetic Resonance in Medicine*, 80(6):2759–2770, December 2018. ISSN 1522-2594. doi: 10.1002/mrm.27229.
- Xiang Liu, Chao Han, Yingpu Cui, Tingting Xie, Xiaodong Zhang, and Xiaoying Wang. Detection and Segmentation of Pelvic Bones Metastases in MRI Images for Patients With Prostate Cancer Based on Deep Learning. *Frontiers in Oncology*, 11:773299, 2021. ISSN 2234-943X. doi: 10.3389/fonc.2021.773299.
- Muhammad Farooq and Abdul Hafeez. COVID-ResNet: A Deep Learning Framework for Screening of COVID19 from Radiographs, March 2020. URL <http://arxiv.org/abs/2003.14395>. arXiv:2003.14395 [cs, eess].
- Pauley Chea and Jacob C. Mandell. Current applications and future directions of deep learning in musculoskeletal radiology. *Skeletal Radiology*, 49(2):183–197, February 2020. ISSN 1432-2161. doi: 10.1007/s00256-019-03284-z.
- Mingyu Kim, Jihye Yun, Yongwon Cho, Keewon Shin, Ryoungwoo Jang, Hyun-jin Bae, and Namkug Kim. Deep Learning in Medical Imaging. *Neurospine*, 16(4):657–668, December 2019. ISSN 2586-6583, 2586-6591. doi: 10.14245/ns.1938396.198. URL <http://e-neurospine.org/journal/view.php?doi=10.14245/ns.1938396.198>.
- Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, December 2017. ISSN 1361-8423. doi: 10.1016/j.media.2017.07.005.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL, June 2009. IEEE. ISBN 978-1-4244-3992-8. doi: 10.1109/CVPR.2009.5206848. URL <https://ieeexplore.ieee.org/document/5206848/>.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, January 2015. URL <http://arxiv.org/abs/1409.0575>. arXiv:1409.0575 [cs].
- Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks, April 2014. URL <http://arxiv.org/abs/1404.5997>. arXiv:1404.5997 [cs].
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions, September 2014. URL <http://arxiv.org/abs/1409.4842>. arXiv:1409.4842 [cs].
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL <http://arxiv.org/abs/1512.03385>. arXiv:1512.03385 [cs].
- Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, April 2017. URL <http://arxiv.org/abs/1704.04861>. arXiv:1704.04861 [cs].
- Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, September 2020. URL <http://arxiv.org/abs/1905.11946>. arXiv:1905.11946 [cs, stat].
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL <http://arxiv.org/abs/2010.11929>. arXiv:2010.11929 [cs].

- Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. CoAtNet: Marrying Convolution and Attention for All Data Sizes, September 2021. URL <http://arxiv.org/abs/2106.04803>. arXiv:2106.04803 [cs].
- Qingchen Zhang, Changchuan Bai, Zhuo Liu, Laurence T. Yang, Hang Yu, Jingyuan Zhao, and Hong Yuan. A GPU-based residual network for medical image classification in smart medicine. *Information Sciences*, 536:91–100, October 2020. ISSN 00200255. doi: 10.1016/j.ins.2020.05.013. URL <https://linkinghub.elsevier.com/retrieve/pii/S0020025520304059>.
- Sara Atito, Muhammad Awais, and Josef Kittler. SiT: Self-supervised vIision Transformer, November 2021. URL <http://arxiv.org/abs/2104.03602>. arXiv:2104.03602 [cs].
- Sayak Paul and Pin-Yu Chen. Vision Transformers are Robust Learners, December 2021. URL <http://arxiv.org/abs/2105.07581>. arXiv:2105.07581 [cs].
- Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention, January 2021. URL <http://arxiv.org/abs/2012.12877>. arXiv:2012.12877 [cs].
- Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-Scale Vision Longformer: A New Vision Transformer for High-Resolution Image Encoding, May 2021. URL <http://arxiv.org/abs/2103.15358>. arXiv:2103.15358 [cs].
- Chi-Tung Cheng, Chih-Chi Chen, Fu-Jen Cheng, Huan-Wu Chen, Yi-Siang Su, Chun-Nan Yeh, I-Fang Chung, and Chien-Hung Liao. A Human-Algorithm Integration System for Hip Fracture Detection on Plain Radiography: System Development and Validation Study. *JMIR Medical Informatics*, 8(11):e19416, November 2020. ISSN 2291-9694. doi: 10.2196/19416. URL <http://medinform.jmir.org/2020/11/e19416/>.
- Loïc Duron, Alexis Ducarouge, André Gillibert, Julia Lainé, Christian Allouche, Nicolas Cherel, Zekun Zhang, Nicolas Nitche, Elise Lacave, Aloïs Pourchot, Adrien Felter, Louis Lassalle, Nor-Eddine Regnard, and Antoine Feydy. Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study. *Radiology*, 300(1):120–129, July 2021. ISSN 0033-8419, 1527-1315. doi: 10.1148/radiol.2021203886. URL <http://pubs.rsna.org/doi/10.1148/radiol.2021203886>.
- Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. 2017. doi: 10.48550/ARXIV.1702.08608. URL <https://arxiv.org/abs/1702.08608>. Publisher: arXiv Version Number: 2.
- Maria Fox, Derek Long, and Daniele Magazzeni. Explainable Planning. 2017. doi: 10.48550/ARXIV.1709.10256. URL <https://arxiv.org/abs/1709.10256>. Publisher: arXiv Version Number: 1.
- Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1):18, December 2020. ISSN 1099-4300. doi: 10.3390/e23010018. URL <https://www.mdpi.com/1099-4300/23/1/18>.
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. 2013. doi: 10.48550/ARXIV.1312.6034. URL <https://arxiv.org/abs/1312.6034>. Publisher: arXiv Version Number: 2.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359, February 2020. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-019-01228-7. URL <http://link.springer.com/10.1007/s11263-019-01228-7>.