



Multi Locale Bone Fracture Radiographs and Localization

Bachelor's Thesis

to achieve the university degree of
Bachelor's degree programme: Software Engineering

Academic Year: 2020-2021

submitted to

Islamic University of Technology

Iftekhharul Abedeen, 170042083
MD Ashiqur Rahman, 170042085
Fatema Zohra Prottyasha, 170042039
BSc in SWE

Supervisor

Tareque Mohmud Chowdhury
Assistant Professor
Dept. of CSE, IUT

Co-Supervisor

Tasnim Ahmed
Lecturer
Dept. of CSE, IUT

Department of Computer Science and Engineering

Gazipur, 11 May 2022

Declaration of Authorship

We declare that we have authored this thesis under the supervision of Tareque Mohmud Chowdhury, Assistant Professor of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh, and Tasnim Ahmed, Lecturer of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh. We have not used other than the declared sources/resources, and that we have explicitly indicated all material that has been quoted either literally or by content from the sources used. It is also declared that neither this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma.

Authors

Abedeen

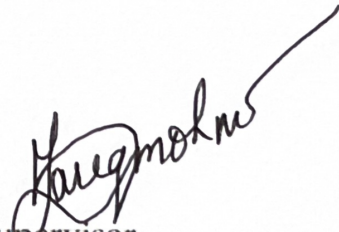
Iftexharul Abedeen, ID: 170042083

Ashiqur Rahman


MD Ashiqur Rahman, ID: 170042085

Fatima

Fatema Zohra Prottiyasha, ID: 170042039


Supervisor

Tareque Mohmud Chowdhury
Assistant Professor,
Dept. of CSE, IUT


Co-suprvisor

Tasnim Ahmed
Lecturer,
Dept. of CSE, IUT

Abstract

We introduce MLBFR, a varied radiographs dataset of human bone fractures. The dataset contains 2,583 radiographs, among which 410 have 575 fracture points. A radiologist manually labelled the dataset as "fractured" and "non-fractured" with masks for the fracture locations. The dataset was verified and approved by an expert medical officer to evaluate the radiologist's performance further. To precisely detect and localize the fracture areas, we experimented with several state-of-the-art object detection models, YOLOv5, maskRCNN, efficientDet and more, along with their ensemble. The trained models fell under two criteria, one being the full dataset and the other being only the fractured radiographs. The trained models managed to achieve a precision of 78.9% and 91.65% on combined and only fractured radiographs, respectively. The model performances were comparable to that of radiologists in detecting major abnormalities in the arm and shinbone area. With falling slightly behind in detecting fractures in the hip, thigh, and finger fractures. It is our belief that the task of improving this performance will be a good challenge for future research. To further encourage advancement in this area, we intend to make this dataset freely available in the future.

Acknowledgements

We would like to thank Dr. Md. Asaduzzaman Bhuiyan without whose expertise and supervision this whole endeavour could not have been possible.

A debt of gratitude is also owed to Mijanur Rahman for spending so much time and effort to help us out in the annotation process of the dataset.

We would also like to thank Effat jahan for the effort on helping us collect radiographs from different sources.

Contents

1. Introduction	1
1.1. Goals and Motivation	1
1.2. Methodology and Structure	2
2. Background Study and Related Works	4
2.1. Baseline Models	5
2.1.1. YOLO	5
2.1.2. Mask-RCNN	7
2.1.3. RetinaNet	7
2.2. SOTA Solutions	8
2.2.1. Hand crafted approaches	9
2.2.2. Use of Artificial Intelligence	9
2.3. Existing Datasets	11
2.3.1. MURA	11
2.3.2. Medpix	11
2.3.3. Radiopaedia	12
2.3.4. IEST	12
2.3.5. MOST	12
2.3.6. ChestX-ray8	12
3. MLBFR Dataset	13
3.1. Data Collection	13
3.2. Data Cleanup	13
3.3. Abnormality analysis	13
3.4. Dataset Properties	14
4. Design & Conceptual Model	16
4.1. Starting Point and Motivation	16
4.2. Conceptual Architecture	16
4.2.1. NMS	17
4.2.2. WBF	17
5. Implementation Details	18
5.1. Architectures	18
5.1.1. YOLOv5s	18
5.1.2. YOLOv5m	19
5.1.3. Mask R-CNN	19
5.1.4. RetinaNet	20

5.1.5. EfficientDet	20
5.2. Collective prediction	22
6. Evaluation	23
6.1. Results	23
6.2. Discussion	24
7. Challenges	26
8. Future Work	27
9. Conclusion	28
A. Data usage permission	30
Bibliography	31

List of Figures

1.1.	Workflow	2
1.2.	Localization of fractures. The red region represents fracture area masks in the radiographs.	2
2.1.	YOLO: 7 by 7 grid cells applied over a target image (Redmon et al. (2016))	6
2.2.	YOLO ₁ : Preliminary Architecture (Thuan (2021))	6
2.3.	RetinaNet Architecture (Lin et al. (2017))	8
2.4.	Performance of fracture prediction (Thian et al. (2019))	10
2.5.	DeepWrist pipeline. (Raisuddin et al. (2021))	11
4.1.	Architecture for ensemble of predicted regions from multiple models	16
5.1.	YOLOv ₅ PA-Neck architectures. (Solawetz (2020))	19
5.2.	YOLOv ₅ pretrained models and their sizes. (Solawetz (2020))	19
5.3.	Mask R-CNN head with 2 different backbones (He et al. (2017))	20
5.4.	Speed (ms) versus accuracy (AP) on COCO test-dev of RetinaNet versions (Lin et al. (2017))	21
5.5.	EfficientDet architecture (Tan et al. (2020))	21
6.1.	Localization pipeline	24
6.2.	Fracture predictions (A) inference using EffiecientDet, (B) inference using retinaNet, (C) inference using YOLOv ₅ m, (C) inference using YOLOv ₅ s	25

List of Tables

2.1.	Average Precision of Mask R-CNN on COCO test-dev related to previous SOTA solutions (He et al. (2017))	8
3.1.	MLBFR study distribution	14
6.1.	mAP, mAR and F ₁ /Dice score for the baseline models trained on MLBFR	23

1. Introduction

With the introduction of sufficient computational capacity, there has always been a strive to automate different medical aspects, may it be the detection of disease, physiological structure, or DNA sampling. This need rose due to the great effort, cost and time needed to do the said activities manually. Humans are also susceptible to errors due to fatigue or absence of mind. For this reason, there have been studies regarding the automation of bone fracture detection along with other diagnosis of disease and abnormalities since early 2000. One of such studies Kositbowornchai et al. (2001) tried automating fracture detection from x-ray scans by identifying bone structure edges and overlapping those to a set of predefined boundary shapes for a set of bones. Following this, there have been many attempts to better the performance of such automated systems. Around 2013 there was a sudden uprise in machine learning approaches to solve different problems. With this, the bio-medical sector also saw a rise in the study of disease detection automation through machine learning Ubaidillah et al. (2013).

1.1. Goals and Motivation

With the surge of new machine learning approaches, the need for Large, High-quality datasets proliferated. Because any solution that is proposed through machine learning has a significant correlation with the size and variance of data it is trained upon Deng et al. (2009). Though there have been many endeavours for a solution to fracture detection automation, the existing techniques face myriads of issues. Most can only classify an image as fracture and nonfracture, meanwhile failing to localize the fracture. Though there are solutions for localization of bone fracture, the solutions either require high-performance cost, lack accuracy or even fail to identify multiple fractures in the same image. But in reality it is possible to arise such situations where multiple fractures occur within a single radiograph. Failing to detect multiple fractures can lead to further complexity and discomfort to a patient.

Due to the nature of the medical domain, most of the data that has been used to devise the proposed solutions and SOTA methods are not openly available. This makes replicating them almost impossible. The lack of data also greatly hinders the development and progress of this field. The openly available data are also not suitable for most of the cases. Some lack proper annotation for localization, while others are poorly prepared and maintained. Facing this issue with the datasets, we

introduce MLBFR with 2,583 radiographs where 410 of which contain 575 fractures and 2,173 are healthy samples. Unlike most other datasets that focus on specific bones or body parts. Our dataset has scans representing the whole body with the exception of the spine, chest and skull.

1.2. Methodology and Structure

To evaluate how different models perform compared to a trained radiologist, we trained multiple baseline object detection and localization architectures on MLBFR. Each of the architectures takes one or a batch of radiographs as input. On each radiograph, state of the art object detection model predicts the probability of fracture and specifies the exact location of the fracture in the radiograph.

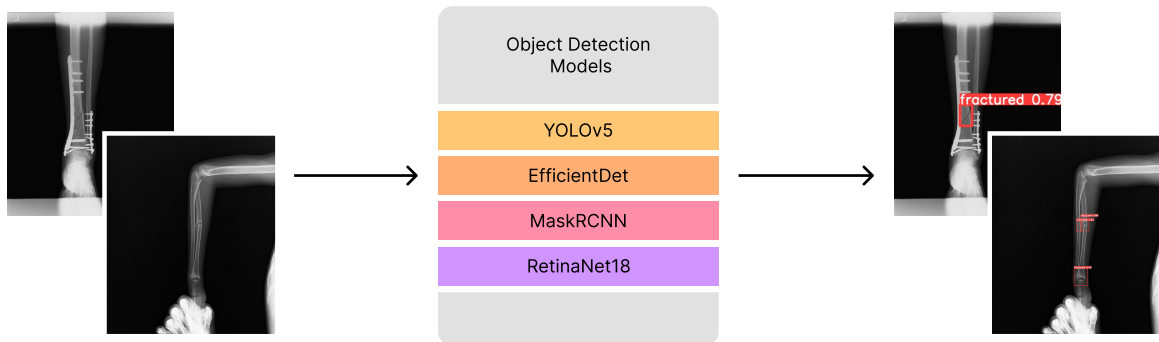


Figure 1.1.: Workflow

To evaluate the model performance and decrease false negatives, we used multiple layers of Consensus and Affirmative ensemble with NMS and WBF. We are planning to make our dataset freely available to encourage advances in medical imaging models.

The radiographs were examined and polygon masks were generated for the fractured areas by an expert radiologist. There were several radiographs with multiple fracture locations in them. The produced masks were later verified by a medical officer in order to validated the work.

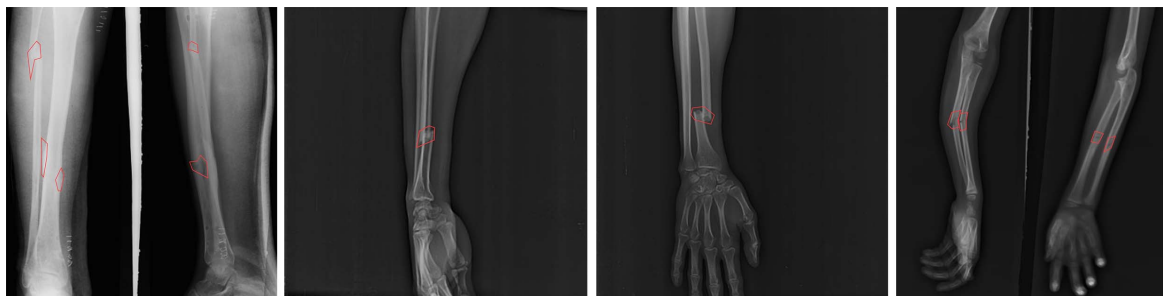


Figure 1.2.: Localization of fractures. The red region represents fracture area masks in the radiographs.

The subjects of the radiographs have an age range of 8 months to 78 years. The gender distribution in the radiographs containing fractures is 85.4% to 14.6% among males and females, respectively. For the complete dataset, the gender distribution is 62% male and 38% female. The scans contain 1856 anteroposterior views and 1152 lateral views. The maximum number of fractures marked in one image ranges from 0 to 5, where 0 represents a nonfractured or healthy scan, and any value larger identifies that scan as a fractured one. Along with the number of fractures in a scan, the annotations also contain masks and bounding boxes corresponding to each of the fractures present in that image.

2. Background Study and Related Works

Throughout the last few decades, there has been significant development in the automation of bone fracture detection. Firstly to illustrate the problem we are dealing with, the physiological and economic aspects are to be examined.

The type of fractures can differ from minor breaking and detachment of major bones, e.g., femur, radio, ulna, to hair-like minor surface fractures. Also, fractures can exist in a cluster of small bones, which are hard to isolate between actual bone joints (Outram (2002)). Failing to diagnose and identify complications such as osteonecrosis, nonunion, and degenerative arthritis can result in persistent pain and functional compromise for the rest of one's life (Tentori et al. (2014)).

There are several methods for fracture diagnosis. Though Magnetic resonance imaging (MRI), X-ray, and computerized tomography (CT) are available options for musculoskeletal scans to find bone structure abnormalities, MRI and CT scans are cost-prohibitive in terms of acquisition and operation. This makes these tools out of reach to most medical diagnostic centers and clinics in developing countries. The most common and widely available form of diagnosis is 2D X-Ray radiology. So, it is crucial that the automation of bone abnormality detection is done considering this accessibility concern. According to Raisuddin et al. (2021) the number of fracture cases for hand and wrist is around 18 million, among which cases regarding radius and ulna fractures are diagnosed 100,000 inhabitants of United States on average (Karl et al. (2015)). The treatment for such injury can vary depending on the type of fracture and its severity. According to a study by De Putter et al. (2012) in the Netherlands, the annual cost regarding hand and wrist lesions was over €540,000,000. Besides economic hurdles, such injuries can cause loss in Health-related quality of life. The fractures can also take a long time to recover, and sometimes past a period after injury; recovery becomes almost impossible.

For bone-related injuries, the first form of diagnostic tool that is used is conventional radiography (x-Ray imaging) (Basha et al. (2018)). In some situations, it is the only form of diagnosis conducted and available. So, it is crucial to identify the abnormalities at this stage, or they may go undetected and untreated causing future complexities and discomfort.

2.1. Baseline Models

There are a number of machine learning architectures available for object detection. The problem of localization encompasses both the classification of an image or a part of image. It also finds out the locale for that object within an image. YOLO, MaskRCNN, FasterRCNN, RetinaNet, and EfficientDet can be mentioned, if we are to name a few. We trained and evaluated our models based on these architectures and their ensemble.

2.1.1. YOLO

In the early days of object detection, the task was usually done in 2 stages. At first, they use a sliding window to separate out different sections of the target image. They used different sizes of windows to separate out multiple sets of image segments. After doing so, in the second stage, they apply classification models to identify what that sort of object is present in that segment (Thuan (2021)). Though this method is straightforward, the computation needed for the operations is extensive. This makes those solutions prohibitively slow and costly. There is also the problem of the operations being done in stages. This causes different stages to be dependent on each other and hard to optimize for speed.

YOLOv1

Redmon et al. (2016) introduced the You Only Look Once (YOLO) algorithm. This algorithm brought all the stages under a single neural network, combining multiple bounding box coordinates and class probability. This model simultaneously predicts multiple bounding boxes and the chance of finding an object within each of those boxes. This drastically increased the speed and accuracy of object detection compared to other models of that time. At the time of its release most common way of object detection was through Convolutional Neural Networks (CNN), Region-Convolutional Networks (R-CNN), and Regions Proposal Networks (RPNs). The YOLO architecture incorporates multiple Convolutional Networks within its single neural network to generate prediction vectors corresponding to each object. This deviation from the multi-frame multi-stage detection procedure of R-CNN made the YOLO system compute all feature vectors in one go. It is to be noted that the objects may or may not appear in a target image.

Along with marked out bounding box for a detected object (Thatte (2020)), Yolo also generates confidence score for that region of being a that certain object. The confidence score is calculated with

$$\text{Confidence score} = p(\text{Object}) \times IOU_{pred}^{\text{truth}}$$

Here, $p(Object)$ represents probability of an object being in the predicted area, and IOU_{pred}^{truth} is for finding intersection over union of predicted region and ground truth,

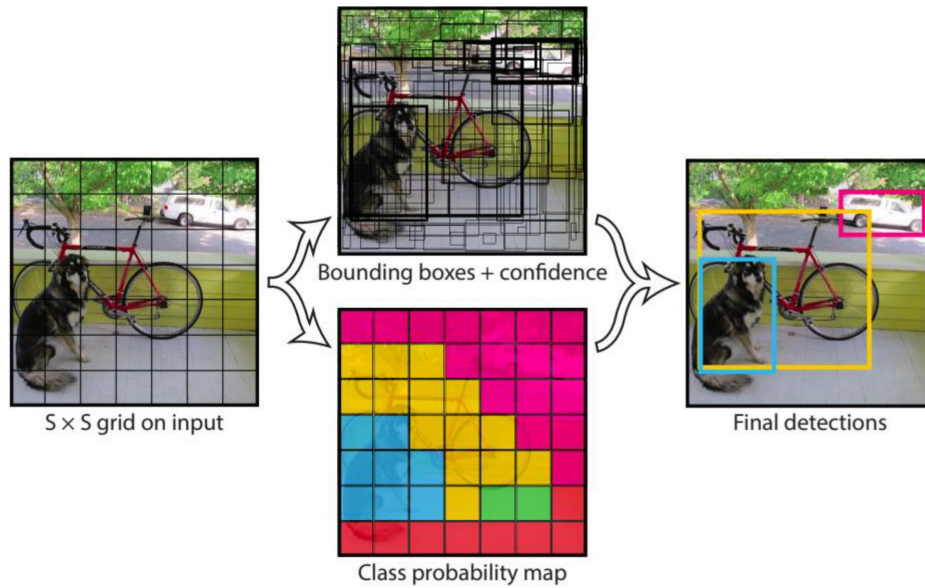


Figure 2.1.: YOLO: 7 by 7 grid cells applied over a target image (Redmon et al. (2016))

YOLO introduced Darknet architecture, its job is to process all the image features. The output of the DarkNet layer goes to 2 fully connected layer for the bounding box prediction (Figure 2.2).

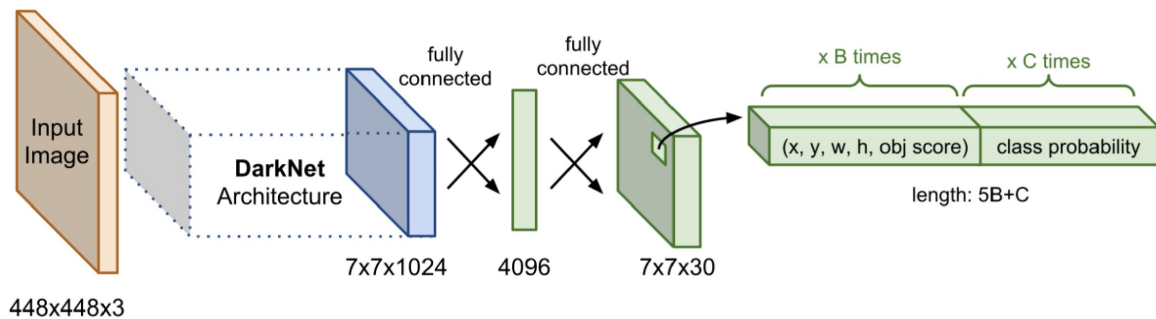


Figure 2.2.: YOLO1: Preliminary Architecture (Thuan (2021))

YOLOv2-3

After version 1 version 2 and 3 brought significant changes on multiple front. in Redmon and Farhadi (2017) Improved the existing architecture of version one by introducing Batch Normalization. They also introduces a Higher resolution classifier and Convolutional anchor box along with cutting of some minor features that didn't perform as expected.

in version 3 (Redmon and Farhadi (2018)) further improvement was made. The improvement can in a form of bigger network with ResNet and Multi-scale detector.

YOLOv4

YOLOv4 (Bochkovskiy et al. (2020)) saw some major changes with the leaving of the first author Redmon. The changes were made in Object detection architecture, use of Backbone-CSPDarknet53, reshuffling of Neck and SPP block, feature aggregation model to name a few.

YOLOv5

Though named version 5, it was developed almost in parallel to version 4 by a different team Jocher et al. (2022). As they were developed almost in the same time period, the SOTA methods used were more or less the same. The most significant change came with switching from C-based Darknet to Python-based framework Pytorch. This led to more widespread adoption and accessibility. Backbone (Focus structure and CSP network), Neck (SPP block, PANet), and Head (YOLOv3 Head usingGloU-loss) are to be mentioned if we are to point out the significant changes in terms of architecture from v4 to v5.

2.1.2. Mask-RCNN

Mask-RCNN does object detection and segmentation introduced by He et al. (2017). In a segmentation process, the system needs not only to find an object and mark it with a bounding box, but it is also required to classify each pixel in the image as a certain object or background (He et al. (2017)). The Mask R-CNN extends upon the existing Faster R-CNN along with the localization by rectangular boundaries. It introduces a predictor for generating segmentation masks on the region of interest (ROI). Mask R-CNN does the operation in 2 stages. In the first stage, it runs a baseline Faster R-CNN over the image to localize the ROI. Upon getting the RIO within the bounding box, a small Fully Convolutional Network (FCN) is run over the ROI in the second stage. The second stage generates the mask in a pixel-to-pixel manner. The use of Faster R-CNN makes the segmentation process faster with a small overhead for FCN.

2.1.3. RetinaNet

Introduced by Lin et al. (2017); Retina net uses 2 stage predictor. The first stage uses a backbone network followed by two task-specific subnetworks of the second stage. The backbone network of the first stage computes a convolutional feature map on

Architectures	Backbone	AP	Remark
Mask R-CNN	ResNet-101-C4	33.1	Introduced in 2017
	ResNet-101-FPN	35.7	
	ResNeXt-101-FPN	37.1	
MCN	ResNet-101-C4	24.6	Winner of of the COCO 2015 segmentation challenge
FCIS+OHEM	ResNet-101-C5-dilated	29.2	Winner of of the COCO 2016 segmentation challenge

Table 2.1.: Average Precision of Mask R-CNN on COCO test-dev related to previous SOTA solutions (He et al. (2017))

the whole image. This layer can be chosen from a set of off-the-shelf convolutional networks. The preliminary subnet from the second stage does object classification on the output of the backbone. As for the second subnet, it performs convolutional bounding box regression.

The authors used Feature Pyramid Network (FPN) to generate a multi-scale feature pyramid from the input image in the proposed solution. With each layer on the pyramid, different scales of objects are detected as depicted in Figure 2.3.

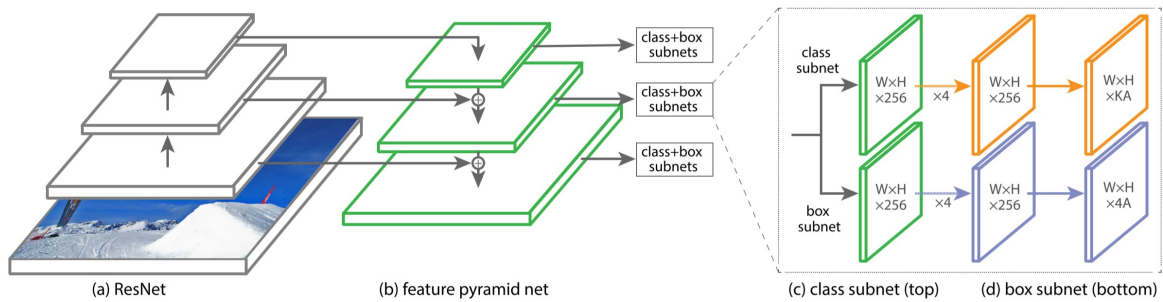


Figure 2.3.: RetinaNet Architecture (Lin et al. (2017))

2.2. SOTA Solutions

According to Welling et al. (2008) the main and most challenging one is to identify fractures that happen between a small cluster of bones, e.g., Metacarpals, .etc. Even very experienced radiologists can fail to detect those fractures. Furthermore, on top of that, these fractures are prevalent as they can happen in sports or other light injuries. In this section we are going to discuss about, how the automation of bone fracture detection progressed over the past years. To draw out the timeline we are mentioning the major research and developments in this field.

2.2.1. Hand crafted approaches

There have been endeavors to automate bone fracture detection with computer vision from the early years of this century. Kositbowornchai et al. (2001) tried to automate the process of fracture detection by generating outlines for the edges in an x-ray image and then overlapping those edges to a set of predefined bone outlines to find anomalies. This process had a flaw as one had to keep outlines for every possible angle for all the bones in a human body. Also, with the change in age, human bones tend to change their surface pattern due to wearing and other diseases over the years of one's life. This made the process inaccurate and not widely adaptable. The detection task was also prolonged due to the lack of processing power of the computers at that time.

Fast forward to the early years of the previous decade, computer-aided fracture detection techniques were still not assumed as successful because of many difficulties such as the variety of the fracture types, sensitivity differences of x-ray devices, imaging errors, and the proximity of bone and tissue color tones (Eksi and Cakiroglu (2012)). The success rate of differentiation of bone and tissue from each other directly affected the performance of fracture detection. In general, the tones of tissue and bone were in shades close to each other, which interfered with the detection process. The most prominent solutions around the year 2012 were different modes of clustering.

K-Means: it was a famous classification algorithm. First, the user-defined a K value. This K value indicated the number of clusters in the image. The system then went through the data features and extracted K number of clusters. The clustering was done by gathering similar features from the image, and the locale was pointed to the center of each cluster. For it to work, there had to be a high similarity in the intra-cluster features and very low to no similarity between the inter-cluster features.

Fuzzy C-Means (FCM): This was also a clustering algorithm. It was considered a powerful image segmentation technique. It performed the segmentation by dividing two or more clusters. It separated each cluster by assigning membership values to each item in a cluster.

OTSU: Named after Nobuyuki Otsu, the algorithm also worked by clustering. However, here the clustering was done in a different method. Rather than depending on locality, it was clustered by distributing values of pixels in the image and thresholding.

2.2.2. Use of Artificial Intelligence

Kim and MacKinnon (2018) was the first paper to introduce the concept of machine learning in bone fracture detection. They used transfer learning to bring a pre-trained on non-medical image model of Convolutional Neural Network (CNN) to

detect fracture. This model could classify fractured radio graphs. They retained the top layer of inception v3 using lateral wrist radiographs to classify new studies into 'fracture' and 'no fracture'. 11,112 images were used to train the model with eightfold data augmentation. They used an initial set of 1,389 radiographs where 695 were of 'fracture' and 694 of 'no fracture' class. They used a total of 100 radiographs being, 50 for 'fracture' and 50 for 'no fracture.' to evaluate the model. Their model performed comparable to SOTA and worked as a proof of concept for the use of transfer learning for the use bone fracture classification task.

Following those work Thian et al. (2019) introduced localization of fracture beside classification. The previous works used binary classification to predict if an image has any fracture. This lacked the location information for those predicted fractures. It is difficult for physicians to trust such broad categories of prediction by a 'black-box' method. This lacked the explainability for such a sensitive and crucial field as the medical sector. Localization gives physicians visual evidence to verify the result from this sort of automated system.

The proposed workflow by Thian et al. has a base CNN layer comprising a convolutional layer and a max-pooling layer followed by Inception-ResnetV2 for localization and caption. The Inception-Resnet had been pretrained with the COCO dataset. Their model detected 91.2% of the fractured images; AKA classified them. And 96.3% of another image set was correctly localized.

To depict the latest STOA solution, we take a look at the solution by Raisuddin et al. (2021). They named their solution DeepWrist. What sets them apart from the prior SOTA methods is that most of the previous ones based their prediction on a general dataset while ignoring the very minute and hair-like fractures. If those go undetected, it may cause osteoarthritis and other problems. For these types of fractures, the regular X-Rays are not enough for expert radiologists to identify them. For those, It requires computed tomography or CT Scan, for which one needs to face a significant amount of radiation and is very costly to operate. The DeepWrist proposed here can detect those challenging fractures only from X-Ray images with an accuracy of 99%, whereas other methods can achieve an average of 64% accuracy.

Image Projection	No. of Ground-Truth Fracture Marks	No. of CNN Fracture Marks	No. of True-Positive CNN Marks	Per-Mark Sensitivity (%)	Per-Image Sensitivity (%)	Per-Image Specificity (%)	AUC (%)
Frontal	340	370	310	91.2 (87.6, 94.0)	95.7 (92.4, 97.8)	82.5 (77.4, 86.8)	0.918 (0.894, 0.941)
Lateral	245	276	236	96.3 (93.1, 98.3)	96.7 (93.6, 98.6)	86.4 (81.9, 90.2)	0.933 (0.912, 0.954)

Note.—Data in parentheses are 95% confidence intervals. AUC = area under the receiver operating characteristic curve, CNN = convolutional neural network.

Figure 2.4.: Performance of fracture prediction (Thian et al. (2019))

The Proposed DeepWrist Pipeline has mainly two parts. A wrist radiograph is passed to the ROI (region of interest) localization block to predict landmark points. After detecting the Landmark point, a bounding box is set around the points, and the image is cropped around that box. The final block is called the fracture detection block. It takes the cropped image from the previous block and sees if there is any fracture in the image. In addition to the predicting fracture, the final layer in the block generates a probability distribution graph and overlays on the radiograph. This graph shows the probability of fracture in that radiograph and its locality. This is done by using GradCAM.



Figure 2.5.: DeepWrist pipeline. (Raisuddin et al. (2021))

2.3. Existing Datasets

There are a handful of publicly available Datasets related to bone fracture and bone abnormalities; those are MURA, Medpix, Radiopaedia, IEST, ChestX-ray8, and MOST.

2.3.1. MURA

MURA is a dataset of 2D musculoskeletal radiographs consisting of 40,561 multi-view radiographic images. The dataset contains radiographs of the elbow, finger, forearm, hand, humerus, shoulder, and wrist regions. It comes with manual labeling. Board-certified radiologists from Stanford Hospital have done the labeling, and each study is labeled as "normal" or "abnormal." Though it is a robust dataset and works well for classifying the images as normal or abnormal, it does not provide any localization information. Hence not suitable for our needs. Rajpurkar et al. (2017)

2.3.2. Medpix

Secondly, we have gone through the Medpix dataset, an online database of 2D and 3D medical images of all sorts of diseases. To extract the images for our purpose, we filter the dataset with the keyword "fracture," which gives us a total of 1954 images that consist of x-rays, real images, MRI, ct, and Ultrasound. We could not utilize this dataset because it is unorganized, some images are falsely labeled, and

several spam images were found within. Also, it only provides fractured data, and since we can use only 2D images, usable images from this dataset are very small in number. HHS (2016)

2.3.3. Radiopaedia

“Radiopaedia” (2006) is an openly usable and editable educational radiology resource that has been rapidly growing since 2005. Radiologists and radiology trainees primarily compiled it from across the world. Its mission is to create the best radiology reference and make it available for free, forever. Filtering the images with the ‘fracture’ keyword, we get a dataset of around 4314 cases containing multiple images. The dataset contains x-rays, real images, MRI, ct, and Ultrasound. The images are labeled as fractured but do not provide the localization of the fractures. There is no localization information, and only one ‘fractured’ class is found, making the dataset unusable for both classification and segmentation tasks.

2.3.4. IEST

IEST Yadav and Rathor (2020) is a small dataset of 2D x-rays containing 217 images. There are 49 healthy, 99 fractured and 69 cancerous bone X-ray images. The dataset is small and inadequate and does not serve our purpose well.

2.3.5. MOST

MOST S.Gornale (2020), consists 4446 X-ray and MRI images, labeled by the KL grading system having five classes **Grade 0**- No Radiographic features of OA present **Grade 1**- Doubtful OA(narrowing of joint space) **Grade 2**- Mild OA(definite narrowing of joint space) **Grade 3**- Moderate OA (multiple osteophytes, sclerosis) **Grade 4**- Severe OA (large osteophytes, severe sclerosis, bone deformity). MOST Online/UCSF is no longer involved in public sharing of MOST Public Use Limited Datasets and Images due to the end of funding and closeout. It has a shortcoming of only containing knee joint cases. The dataset is expected to be available in 2023.

2.3.6. ChestX-ray8

ChestX-ray8 Wang et al. (2017), a chest X-ray database published in 2017, comprises 108,948 frontal-view X-ray images of 32,717 unique patients with the text-mined eight disease image labels. Each image has multi-labels from the associated radiological reports using natural language processing. The dataset has only chest images. Since there are not many fracture cases in the chest, the dataset is for various chest and lung diseases. So, the fracture data is inadequate in quantity.

3. MLBFR Dataset

With the aim to develop a faster and more robust bone fracture detection system, we reached out to different authors of the published research works for access to the used datasets. However, we could not get hold of such datasets for regulatory reasons, confidentiality, and medical domain data's sensitive nature. Also, as discussed in 2.3 the existing public datasets did not meet our requirement. We had to resort to creating our dataset from scratch for this shortcoming.

3.1. Data Collection

As the necessity arose, we contacted several Hospitals and diagnostic centers throughout the country. Some agreed to cooperate, while others did not due to safety and privacy concerns. With the approval and confirmation from the bio-informatics lab and department of computer science and engineering, we reach out to LabAid Diagnostic Center, Brahmonbaria, Prime Diagnostic Center, Barishal, and New Anupam Hospital, Bogra. After collecting X-Ray scans from these sources, we had 9,512 image samples in our hands. As we promised not to take any personal information, the source authorities ran queries in their database to provide us with different matrices.

3.2. Data Cleanup

All the samples we collected were general-purpose diagnosis scans for various purposes. This meant with the studies of bone fracture and related conditions; there were also studies of lung diseases and other chest and skull complication-related studies. As the cases of chest fracture, skull, and spine fractures were sparse in our collected data, with the supervision of a medical officer, we removed scans of these places. This clean-up operation left us with 2,598 study images in the end.

3.3. Abnormality analysis

With the help of an expert radiologist, we identified and marked down locations of abnormalities in the X-Ray images. Later those locations were annotated with

the use of makesense.ai (Skalski (2019)) in COCO JSON format. The annotation was done as polygon masks which enabled us to convert these to PascalVOC, Darknet annotation, and CSV format as necessity had arisen. The annotations were cross-checked and verified by an expert medical officer of Kasba Govt. Hospital.

3.4. Dataset Properties

Our dataset was collected from general-purpose X-Ray studies; it provides us with a varied and diverse dataset. The variation comes in the form of age, gender, the locale of the study, scan plane, and anomaly ratio. The age of the subjects of our dataset is in the range of 8 months to 78 years old. The age difference in our dataset has great significance. Due to low bone density, the cap of each bone seems like separate disc pads for younger patients. This can lead untrained models to make false-positive assumptions for almost all young patients. Having such a diverse range of subjects helps us eliminate those issues.

Study	Fractured	Nonfracture	Details
Wrist	19	21	targeted at wrist only
Hand	103	372	forearm + wrist + finger
Forearm	114	158	targeted at radius ulna
Humerus	19	38	targeted at humerus
Shoulder	3	116	targeted at shoulder bones
Elbow	39	67	targeted at elbow
Leg	95	47	targeted at lower leg (tibia febula)
Foot	10	510	targeted at ankle and feet
Thigh	35	17	targeted at Femur
Hip	22	220	targeted at hip bone and joint
Knee	28	815	targeted at knee joint
Total	487	2381	

Table 3.1.: MLBFR study distribution

The gender distribution in the abnormal set is 85.4% is to 14.6% among the male and female subjects, respectively. The gender ratio is 62% male and 38% female for the entire dataset. The number of anteroposterior scans in fractured class is 289, and lateral scans are 184 within 410 fracture images. These numbers are 1,856 and 1,152 respectively, for the whole dataset of 2,583 images. The Count may not add up as there are multiple films with bilateral views. There are 410 fracture images with 575 fracture points. The distribution of fracture to nonfracture is 410:2173. Each image in the dataset has 0-5 fracture locations marked in it. Here, 0 means the image has no fracture, and anything larger than 0 means that number of fractures localized in it. Our dataset also takes account of scans for all significant body parts like wrist, hand, forearm, humerus, shoulder, elbow, leg, foot, thigh, hip, and

knee. The chest, spine, and skull scans were not included because the number of abnormal scans was significantly low in those regions.

None of the patient data i.e. name, age, gender, address, were handed over to the research team and had an agreement to use the data only for research purposes with the permission of each of the medical center authorities.

4. Design & Conceptual Model

4.1. Starting Point and Motivation

It is our aim to make system as precise as possible in order to properly identify a fracture and reduce human intervention and After consultation with the medical officer and personnel, we realized that it is crucial that our predictions are optimistic. That means it is better to have a false positive in our prediction than to have a false negative. The reason behind this is that, If our model predicts a portion as fractured, a physician will analyze the result anyway. So, even if our system makes a false assumption of fracture being in a location, it will be verified afterward. However, it can be problematic the other way around. Let us say our system fails to recognize a fracture, and it gets overlooked by the physician. This can lead to a fracture not being addressed and cause future complications.

In response to the suggestion provided by the experts, we developed a voting system that takes prediction results from multiple models and ensembles those prediction values based on predicted location and confidence.

4.2. Conceptual Architecture

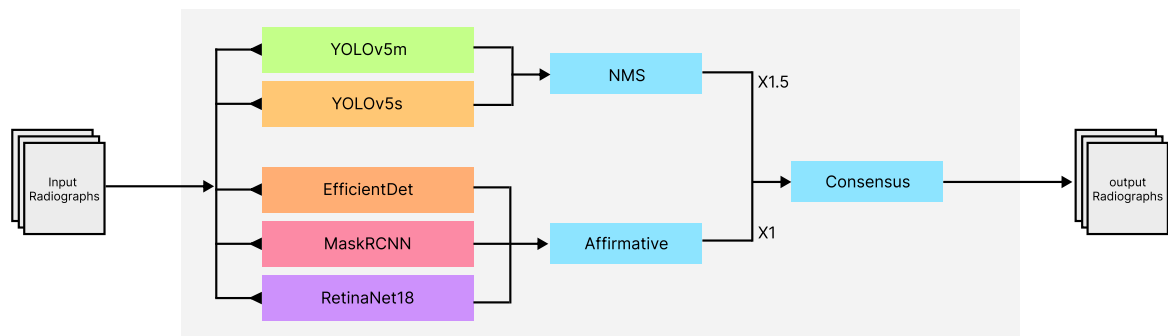


Figure 4.1.: Architecture for ensemble of predicted regions from multiple models

To our observation, The results of YOLOv5m and YOLOv5s tend to be more sensitive and create a lot of predicted areas. To ensure our system does not make too many false assumptions, we pass the predicted bounding areas and confidence score to an NMS vote evaluator. This helps us remove low confidence nonoverlapping boundaries. At the same time, another Affirmative vote evaluator looks if any

model among EfficientDet, MaskRCNN, and RetinaNet predict a region. It works like a union operation. This is done so for the said three methods because they tend to be more conservative at predicting fractures. The vote from NMS and the affirmative section goes into a consensus layer as input. The results of 2 YOLO predictors are emphasized more as they are optimistic compared to the bottom three models in Figure 4.1. Finally, we use the prediction windows as the output.

The above system can be a bit slower than running individual models, and voting adds a negligible calculation overhead relative to the prediction stages. Nevertheless, our system is fast enough for real-life use, and this voting system ensures false negatives are kept to a minimum, which is crucial in medical studies.

4.2.1. NMS

Non Maximum Suppression(NMS) is a very common method and underlying algorithm of object detection pipelines. The idea is to select the best scored box prediction with the maximum IoU(pre defined Intersection Over Union) from all the selected predictions from a pipelineBodla et al., 2017.

To optimize the false positive and false negative balance, we used soft-NMS following the equation:

$$\text{Where, } s_i = \begin{cases} s_i, & iou(M, b_i) < N_t \\ 0, & iou(M, b_i) \geq N_t \end{cases}$$

4.2.2. WBF

Weighted Box Fusion(WBF) is the idea of giving importance or impact factor to a certain prediction. From our study, YOLO architectures gave us the better precision and detection over the fractures. To keep the impact on the final result, we added an impact factor of 1.5 on the YOLO models predictions. For the rest of the models, we kept the impact factor as 1 as they have smaller precision and detection rate.

5. Implementation Details

We trained multiple object detection architectures to get a baseline model using our dataset. However, for variance and accuracy, we tried both single-stage architectures YOLOv5s, YOLOv5m, EfficientDet & RetinaNet with ResNet-18, and two-stage architectures such as MaskRCNN.

For each of the models, we split our dataset into train-test-validation with a ratio of 60-20-20, respectively, for both single and multi-class data.

5.1. Architectures

The used architectures for experimentation and result generations are discussed in the following sections.

5.1.1. YOLOv5s

To train the yolov5s architecture with our model, we trained both the multi-class and single-class data with 300 epochs. Here the image sizes were normalized to 640x640. YOLOv5s is a single-stage state-of-the-art object detection architecture. It has its own annotation convention. The annotations were prepared from original COCO JSON format masks with the help of makesense.ai (Skalski, 2019).

The YOLOv5 has three major stages that make the most significant changes to the result. The first one is Backbone, which is a convolutional neural network that extracts image features by aggregating the image at different levels of granularity. The Neck is the second one, which combines and mixes layers of feature vectors and passes them to the predictor. Lastly, there is the Head. It takes input features from the Neck section and creates the box, and predicts the classes in an image. (Solawetz (2020))

The reason behind choosing YOLOv5s for our dataset is the Adaptive Anchor Box selection algorithm. The datasets' abnormality analysis is critical and sensitive in the medical sector. We trained the architecture from scratch rather than using the pretrained weights. Thus the architecture automatically learned the best anchor boxes for the dataset and used them during training. (Solawetz (2020))

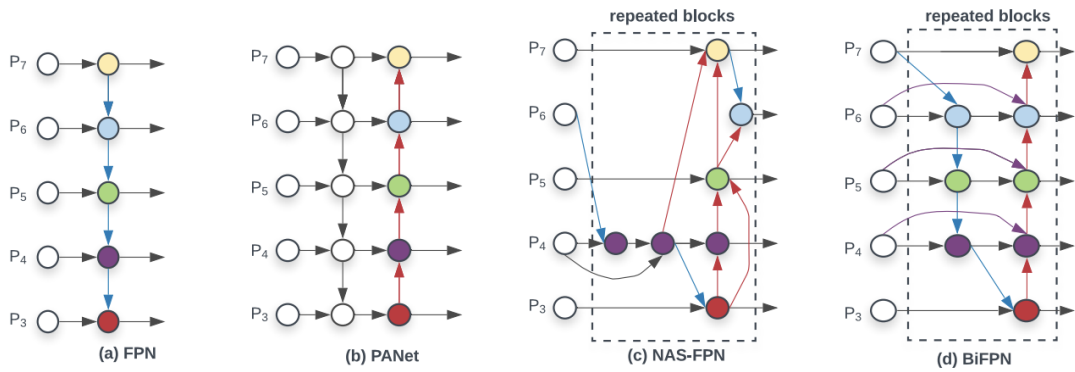


Figure 5.1.: YOLOv5 PA-Neck architectures. (Solawetz (2020))

5.1.2. YOLOv5m

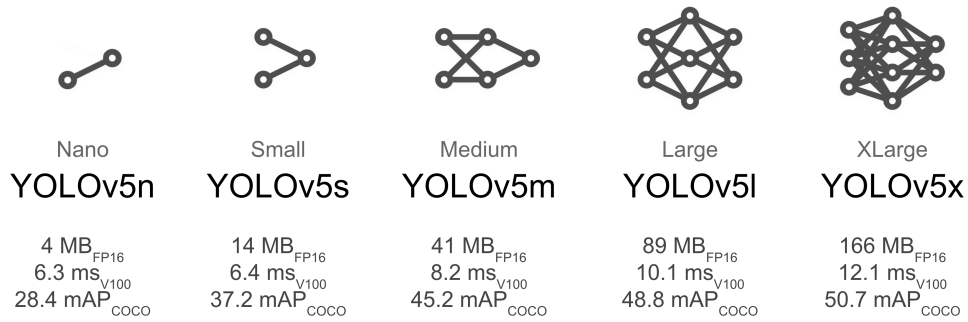


Figure 5.2.: YOLOv5 pretrained models and their sizes. (Solawetz (2020))

The model size of yolov5s is only 14MB, whereas yolov5m is 41MB. Due to the model size difference, we used the yolov5m model’s pretrained weight and the similar configuration of hyperparameters as of yolov5s. However, we considered models with better mean Average Precision(mAP). As of yolov5m has a significant boost in mAP compared to yolov5s. We used both pretrained weights of yolov5m architecture on the COCO 2017 dataset.

5.1.3. Mask R-CNN

The only 2-stage detector as well as a segmentation architecture used on our Dataset. Mask R-CNN has comparatively higher AP in keypoint detection (He et al. (2017)). In this approach, Mask R-CNN efficiently detects and classifies fractured radiographs from an image. It also generates a high-quality segmentation mask for each instance at the same time. Mask R-CNN has comparatively higher AP in keypoint detection.

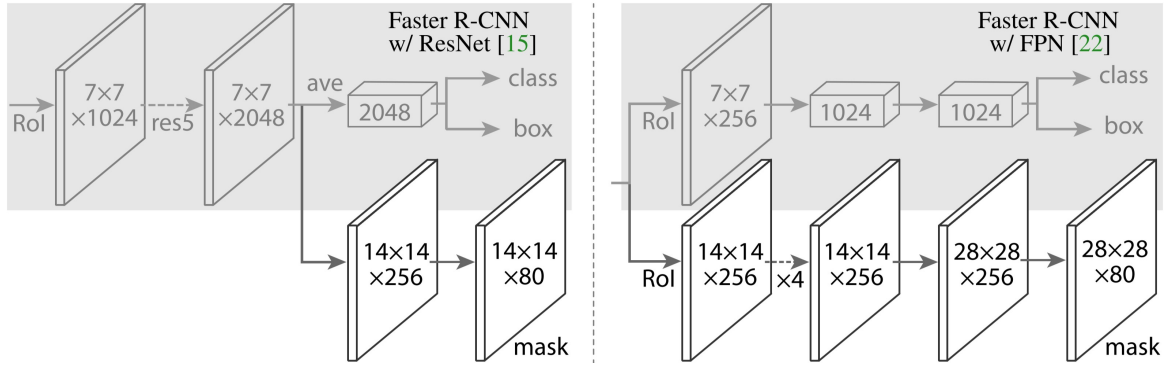


Figure 5.3.: Mask R-CNN head with 2 different backbones (He et al. (2017))

The annotation convention for Mask RCNN is quite different from other models as this approach requires polygon segmentation masks on the radiographs for fracture locations.

5.1.4. RetinaNet

In our situation, accuracy stands before speed. Moreover, RetinaNet introduced Focal Loss (Lin et al. (2017)), which mitigates the issue of extreme imbalance between foreground and background classes during training.

As we did not control or temper the ratio of appearing normal and abnormal scans in our dataset, the number of fracture cases is relatively small compared to the healthy scans. This creates a class imbalance which is typical in real-life situations. RetinaNet explicitly tries to solve this class imbalance problem.

$$CE(p, y) = \begin{cases} -\log(p), & \text{if } y = 1 \\ -\log(1 - p) & \text{otherwise} \end{cases}$$

We used balanced Cross-Entropy and the shallowest ResNet architecture ResNet-18, as the backbone for training with our dataset. To gain faster inference and better accuracy, we chose ResNet-18 as the backbone. However, most images are under 1024 pixels in either dimension, so this operating point improves performance over the ResNet-50 backbone.

5.1.5. EfficientDet

Introduced by Tan et al. (2020), EfficientDet proposed several key optimizations for hands-on efficiencies, such as BiFPN and Compound Scaling Method. From the feature extraction level of EfficientDet's architecture EfficientDetD7, which takes level 7 features from the backbone, achieved 55.1% AP on the COCO test-dev dataset.

5. Implementation Details

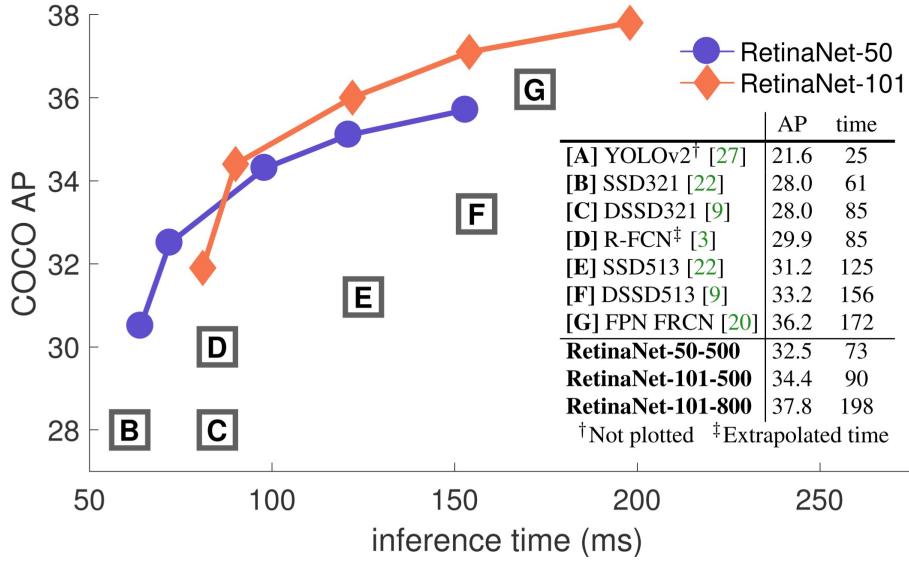


Figure 5.4.: Speed (ms) versus accuracy (AP) on COCO test-dev of RetinaNet versions (Lin et al. (2017))

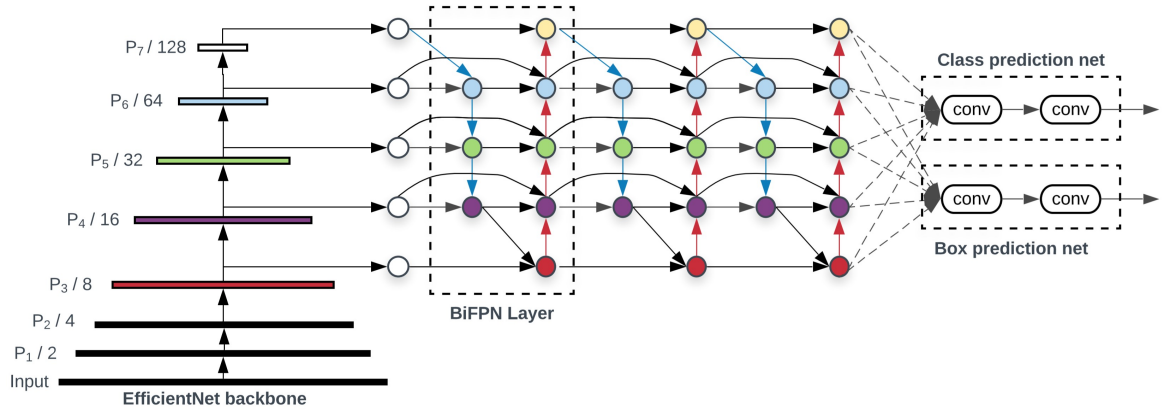


Figure 5.5.: EfficientDet architecture (Tan et al. (2020))

Backbone Network : EfficientDet resues the same backbone structure from EfficientNet -Bo to B6

BiFPN: BiDirectional Feature Pyramid Network is a rework of FPN where the conventional FPN aggregates multi-scale features in a Top-Down approach:

$$\begin{aligned}
 P_7^{out} &= Conv(P_7^{in}) \\
 P_6^{out} &= Conv(P_6^{in} + Resize(P_7^{out})) \\
 &\dots \\
 P_3^{out} &= Conv(P_3^{in} + Resize(P_4^{out}))
 \end{aligned} \tag{5.1}$$

Tan et al. (2020) recognized an issue of conventional FPN: it inherently limits the information flow by keeping the approach one way. Although PANet addresses this

issue, it requires much computational power and is difficult to modify. However, to mitigate this issue and optimize this approach by fusing more features, EfficientDet removes the nodes with one input edge that have less contribution to the feature network. They also add an extra edge in same-level nodes to infuse more features. As an average resolution our image size in the dataset is approximately 640, we chose $\varnothing = 1$ for the backbone, we chose EfficientNet-Bo weights with 0.0001 initial learning rate.

5.2. Collective prediction

After getting prediction form each of the aforementioned models, we used the voting operation in 2 stages as discussed in chapter 4.

6. Evaluation

To evaluate and run inference of all the models with our dataset, we kept 40% of our dataset unseen to the models. We present experimental results on the bounding box detection on our validation and test set. For evaluation we considered mAP(Mean Average Precision), mAR(Mean Average Recall), F1 score and classification loss metrics. We also trained the models on single class and binary class(fractured and non-fractured). For faster runtimes and inference, we chose the shallow models on EfficientDet and RetinaNet.

6.1. Results

Model	Backbone	F1	$mAP_{0.5}$	$mAP_{0.5:0.95}$	mAR
<i>Combined</i>					
Yolov5s	EfficientNet-FPN	63.09	56.15%	22.09%	54.1%
Yolov5m	EfficientNet-FPN	51.2	45.4%	18.63%	40.98%
EfficientDet	EfficientNetB0	50.1	49.86%	14.51%	39.6%
RetinaNet	ResNet18	35.4	46.53%	29.85%	28.55%
MaskRCNN	ResNet-101-FPN	9.6	17.02%	5.01%	6.67%
<i>Fractured</i>					
Yolov5s	EfficientNet-FPN	69	66.65%	22.09%	55.73%
Yolov5m	EfficientNet-FPN	67	66.01%	21.87%	52.66%
EfficientDet	EfficientNetB0	54.2	53.20%	16.66%	55.33%
RetinaNet	ResNet18	44.3	47.20%	26.6%	41.66%
MaskRCNN	ResNet-101-FPN	20.31	33.02%	15.67%	14.67%

Table 6.1.: mAP, mAR and F1/Dice score for the baseline models trained on MLBFR

We evaluated the results both way, the models performance and predictions were cross checked with our radiologist and vice versa.

In both cases, there were false negatives that radiologist couldn't identify and models' couldn't predict. From our test set, there were 3 cases where the radiologist couldn't identify the fracture location from the first scan but they were localized by our model. We then verified those results with medical officer.

Also all the models' inferences were individually assessed by our radiologist.

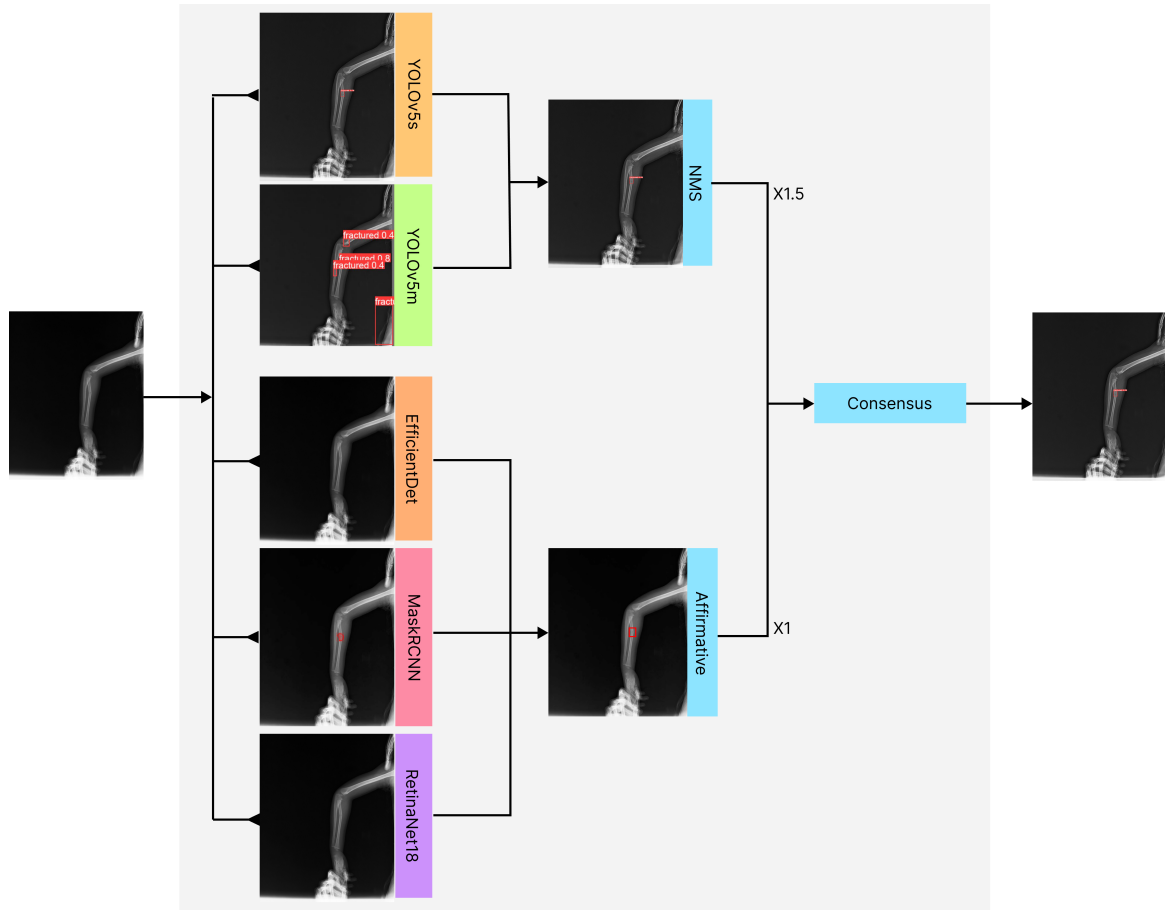


Figure 6.1.: Localization pipeline

6.2. Discussion

From our study, the YOLO models perform better in classification and localization of the fractures. We also assessed the inference results of each model and their performances with our radiologists. The single stage detectors perform best in this case. The two stage detectors fall behind because of the data filtration in the classification layer and no Adaptive Anchor boxes. On the other hand, RetinaNet performs significantly well because of the class imbalance in our dataset and their Focal Loss approach to balance the Cross Entropy.

In comparison to deep learning models and human level performance, our voting system tends to give us more False positives in terms of fracture detection. However, the study shows us that although the voting system is giving us more False Positives but it reduces the False Negatives. In such critical and sensitive field like Fracture detection in , False positives are more likely to get reduced as much as possible. False negatives can raise awareness but it can also provide the patient a double check on the diagnosis and treatment.

We used the shallowest available model and backbone for each architecture to see the base level performance and to decrease computational cost.

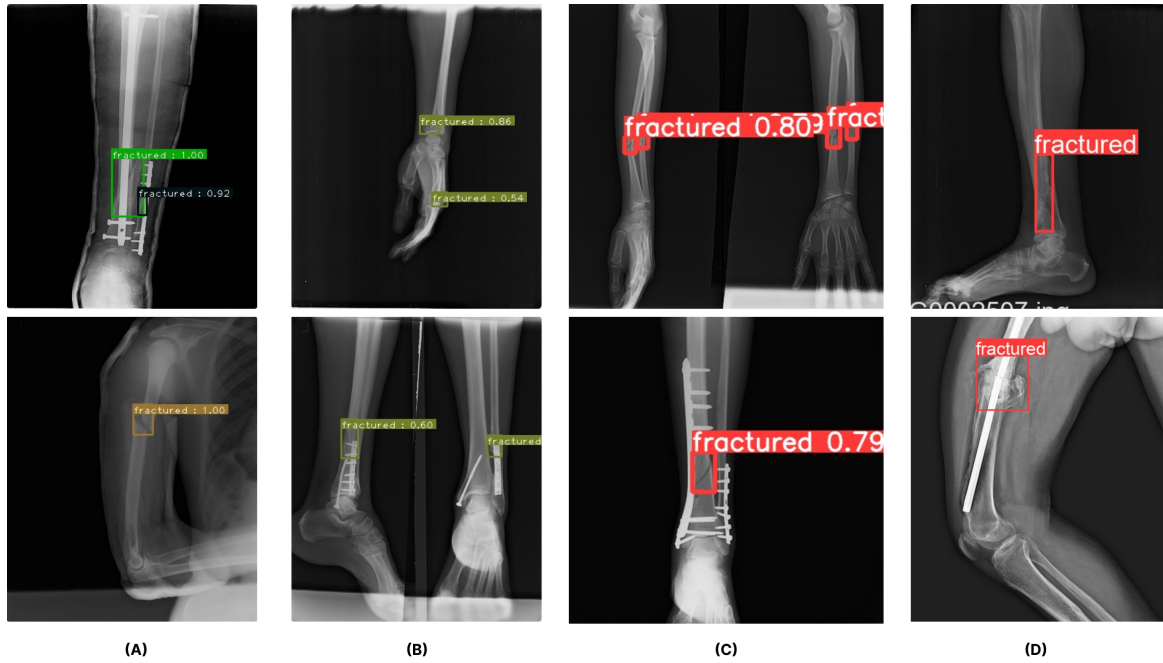


Figure 6.2.: Fracture predictions (A) inference using EfficientDet, (B) inference using retinaNet, (C) inference using YOLOv5m, (D) inference using YOLOv5s

For weighted box fusion, we added 1.5 weights on the YOLOv5 architecture emphasizes YOLO for better precision and performance. 6.1 shows that the YOLO models are given slightly more weight than the other models. We also used the idea of NMS (discussed in Bodla et al. (2017)) to generate the final result after ensemble of two of the outputs.

7. Challenges

The biggest challenge to our research was the lack of a properly annotated public dataset. The domain of medical science is very critical and sensitive. This makes sharing of patient information and datasets difficult. We also faced hurdles in the annotation process as we had to depend on medical experts for what we required heavily. As we had to invest a large portion of our time in the collection and preparation of the datasets, there was less time in our hands for experimentation and development of a State-of-the-art workflow to detect bone fractures in radiographs. The lack of appropriate hardware also proved the work of training models difficult as it took much time to iterate over our solutions.

Due to inherent properties, the number of fractured images compared to normal ones is small in our dataset. Though it is natural, this turned out to be a challenge as our models had a tendency to get biased toward nonfracture samples.

8. Future Work

In our dataset the number of samples for some location are small, which may not be enough for a machine learning model to generalize. We intend to expand our proposed dataset with more samples and scans in future. Also there are room for improvement in term of detection rate and precision of the models. We hope to develop a more robust architecture by ensemble of the best solutions we got so far.

9. Conclusion

To our observation, The field of medical science and its automation is very sensitive and restrained. Due to these situations, the development and automation of many aspects need to become robust and fault-tolerant before something is widely adopted. This sometimes hinders progress and creates steep carve to climb before anything is accepted and recognition is given. The number and scope of freely accessible datasets are meager in the medical domain due to their restrictive nature. With the introduction of MLFBR, we hope the horizon for medical research regarding bone abnormalities will be broadened. We also hope the solutions provided here will help researchers gain more in-depth knowledge of the needs in the medical domain.

Appendix

Appendix A.

Data usage permission

The data has been collected with the permission of (LabAid, Brahmonbaria. Prime diagnostic center, Barishal. New Anupam Hospital, Bogra) and their respective authority to only be used in research work. The data can not be used in any form or method for financial gains. The distribution of the data in public domain is allowed, but anyone using the data will be fully responsible for their work and none of the Hospital, diagnostic center or distributor will be held responsible for any misuse.

Bibliography

- Basha, M. A. A., Ismail, A. A. A., & Imam, A. H. F. (2018). Does radiography still have a significant diagnostic role in evaluation of acute traumatic wrist injuries? a prospective comparative study. *Emergency Radiology*, 25(2), 129–138 (cit. on p. 4).
- Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (cit. on p. 7).
- Bodla, N., Singh, B., Chellappa, R., & Davis, L. (2017). Improving object detection with one line of code. 2017. *CoRR*, abs/1704.04503. <http://arxiv.org/abs/1704.04503> (cit. on pp. 17, 25)
- De Putter, C., Selles, R., Polinder, S., Panneman, M., Hovius, S., & van Beeck, E. F. (2012). Economic impact of hand and wrist injuries: Health-care costs and productivity costs in a population-based study. *Jbjs*, 94(9), e56 (cit. on p. 4).
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (cit. on p. 1).
- Eksi, Z., & Cakiroglu, M. (2012). Performance evaluation of the popular segmentation algorithms for bone fracture detection. *Global Journal on Technology*, 1 (cit. on p. 9).
- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*, 2961–2969 (cit. on pp. 7, 8, 19, 20).
- HHS, N. (2016). Medpix. <https://medpix.nlm.nih.gov/search?allen=false&allt=false&alli=true&query=fracture>. (Cit. on p. 12)
- Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., TaoXie, Fang, J., imyhxy, Michael, K., Lorna, V, A., Montes, D., Nadar, J., Laughing, tkianai, yxNONG, Skalski, P., Wang, Z., ... Minh, M. T. (2022). *ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference* (Version v6.1). Zenodo. <https://doi.org/10.5281/zenodo.6222936>. (Cit. on p. 7)
- Karl, J. W., Olson, P. R., & Rosenwasser, M. P. (2015). The epidemiology of upper extremity fractures in the united states, 2009. *Journal of orthopaedic trauma*, 29(8), e242–e244 (cit. on p. 4).
- Kim, D., & MacKinnon, T. (2018). Artificial intelligence in fracture detection: Transfer learning from deep convolutional neural networks. *Clinical radiology*, 73(5), 439–445 (cit. on p. 9).

- Kositbowornchai, S., Nuansakul, R., Sikram, S., Sinahawattana, S., & Saengmontri, S. (2001). Root fracture detection: A comparison of direct digital radiography with conventional radiography. *Dentomaxillofacial Radiology*, 30(2), 106–109 (cit. on pp. 1, 9).
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, 2980–2988 (cit. on pp. 7, 8, 20, 21).
- Outram, A. K. (2002). Bone fracture and within-bone nutrients: An experimentally based method for investigating levels of marrow extraction. McDonald Institute for Archaeological Research. (Cit. on p. 4).
- Radiopaedia. (2006). <https://radiopaedia.org/search?lang=us&q=fracture>. (Cit. on p. 12)
- Raisuddin, A. M., Vaattovaara, E., Nevalainen, M., Nikki, M., Järvenpää, E., Makkonen, K., Pinola, P., Palsio, T., Niemensivu, A., Tervonen, O., et al. (2021). Critical evaluation of deep neural networks for wrist fracture detection. *Scientific reports*, 11(1), 1–11 (cit. on pp. 4, 10, 11).
- Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R. L., et al. (2017). Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957* (cit. on p. 11).
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788 (cit. on pp. 5, 6).
- Redmon, J., & Farhadi, A. (2017). Yolo9000: Better, faster, stronger. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263–7271 (cit. on p. 6).
- Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (cit. on p. 7).
- S.Gornale, S., U.Patravali. (2020). A comprehensive digital knee x-ray image dataset for the assessment of osteoarthritis. *IEEE Transactions on Biomedical Engineering*, 56(2), 407–415 (cit. on p. 12).
- Skalski, P. (2019). Make sense. <https://www.makesense.ai/>. (Cit. on pp. 14, 18)
- Solawetz, J. (2020). Yolov5 new version - improvements and evaluation. <https://blog.roboflow.com/yolov5-improvements-and-evaluation/>. (Cit. on pp. 18, 19)
- Tan, M., Pang, R., & Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10781–10790 (cit. on pp. 20, 21).
- Tentori, F., McCullough, K., Kilpatrick, R. D., Bradbury, B. D., Robinson, B. M., Kerr, P. G., & Pisoni, R. L. (2014). High rates of death and hospitalization follow bone fracture among hemodialysis patients. *Kidney international*, 85(1), 166–173 (cit. on p. 4).
- Thatte, A. V. (2020). Evolution of yolo-yolo version 1. <https://towardsdatascience.com/evolution-of-yolo-yolo-version-1-afb8af302bd2>. (Cit. on p. 5)

- Thian, Y. L., Li, Y., Jagmohan, P., Sia, D., Chan, V. E. Y., & Tan, R. T. (2019). Convolutional neural networks for automated fracture detection and localization on wrist radiographs. *Radiology: Artificial Intelligence*, 1(1), e180001 (cit. on p. 10).
- Thuan, D. (2021). Evolution of yolo algorithm and yolov5: The state-of-the-art object detection algorithm (cit. on pp. 5, 6).
- Ubaidillah, S. H. S. A., Sallehuddin, R., & Ali, N. A. (2013). Cancer detection using artificial neural network and support vector machine: A comparative study. *Jurnal Teknologi*, 65(1) (cit. on p. 1).
- Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2097–2106 (cit. on p. 12).
- Welling, R. D., Jacobson, J. A., Jamadar, D. A., Chong, S., Caoili, E. M., & Jebson, P. J. (2008). MdcT and radiography of wrist fractures: Radiographic sensitivity and fracture patterns. *American Journal of Roentgenology*, 190(1), 10–16 (cit. on p. 8).
- Yadav, D., & Rathor, S. (2020). Bone fracture detection and classification using deep learning approach. *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, 282–285 (cit. on p. 12).