

**Bone Age Assessment with Less Human Intervention**

by

**Yi-Hsuan Lien**

Bachelor of Engineering, Chung Cheng Institute of Technology, 2014

Submitted to the Graduate Faculty of the  
Swanson School of Engineering in partial fulfillment  
of the requirements for the degree of  
Master of Science

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This thesis was presented

by

**Yi-Hsuan Lien**

It was defended on

July 19, 2021

and approved by

Liang Zhan, Ph.D., Assistant Professor, Department of Electrical and Computer Engineering

Zhi-Hong Mao, Ph.D., Professor, Department of Electrical and Computer Engineering

Jingtong Hu, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

Thesis Advisor: Liang Zhan, Ph.D., Assistant Professor, Department of Electrical and Computer Engineering

Copyright © by Yi-Hsuan Lien

2021

## **Bone Age Assessment with Less Human Intervention**

Yi-Hsuan Lien, MS

University of Pittsburgh, 2021

Biomedical imaging allows doctors to examine the condition of a patient's organs or tissues without a surgical procedure. Various modalities of imaging techniques have been developed, such as X-radiation (X-ray), Magnetic Resonance Imaging (MRI), and Computed Tomography (CT). For example, the Bone Age Assessment (BAA) evaluates the maturity in infants, children, and adolescents using their hand radiographs. It plays an essential role in diagnosing a patient with growth disorders or endocrine disorders, such that needed treatments could be provided. Computer-aided diagnosis (CAD) systems have been introduced to extract features from regions of interest in this field automatically. Recently, several deep learning methods are proposed to perform automated bone age assessment by learning visual features. This study proposes a BAA model, including image preprocessing procedures and transfer learning with a limited number of annotated samples. The goal is to examine the efficiency of data augmentations by using a publicly available X-ray data set. The model achieves a comparable MAE of 5.8 months, RMSE of 7.3 months, and accuracy (within 1 year) of more than 90% on the data set. We also study whether generating samples by a Generative Adversarial Network could be a valuable technique for training the model and prevent it from overfitting when the samples are insufficient.

## Table of Contents

<b>Preface.....</b>	<b>ix</b>
<b>1.0 Introduction.....</b>	<b>1</b>
<b>2.0 Background .....</b>	<b>3</b>
<b>2.1 Pediatric Bone Age .....</b>	<b>3</b>
<b>2.2 Image Processing .....</b>	<b>5</b>
<b>2.2.1 Thresholding.....</b>	<b>5</b>
<b>2.2.2 Histogram equalization.....</b>	<b>7</b>
<b>2.3 Machine Learning .....</b>	<b>8</b>
<b>2.3.1 Artificial Neural Networks .....</b>	<b>8</b>
<b>2.3.1.1 Convolutional Networks for Biomedical Image Segmentation .....</b>	<b>10</b>
<b>2.3.1.2 Densely Connected Convolutional Networks .....</b>	<b>11</b>
<b>2.3.2 Generative Adversarial Network.....</b>	<b>12</b>
<b>2.3.3 Training Generative Adversarial Networks .....</b>	<b>13</b>
<b>2.4 Performance Metrics.....</b>	<b>15</b>
<b>3.0 Methods.....</b>	<b>16</b>
<b>3.1 Materials.....</b>	<b>16</b>
<b>3.2 Data Preprocessing.....</b>	<b>17</b>
<b>3.3 Prediction network .....</b>	<b>20</b>
<b>3.4 Experiment Settings .....</b>	<b>22</b>
<b>4.0 Results .....</b>	<b>25</b>
<b>5.0 Discussion.....</b>	<b>28</b>

<b>6.0 Conclusions.....</b>	<b>30</b>
<b>Bibliography .....</b>	<b>31</b>

## List of Tables

<b>Table 1: Structure of the assessment network .....</b>	<b>21</b>
<b>Table 2: Parameters for transformation.....</b>	<b>22</b>
<b>Table 3: Performance of different settings .....</b>	<b>25</b>
<b>Table 4: Comparison with the published models on RSNA testing set.....</b>	<b>26</b>

## List of Figures

<b>Figure 1: The structure of a multilayer perceptron. ....</b>	<b>9</b>
<b>Figure 2: The architecture of U-Net.....</b>	<b>11</b>
<b>Figure 3: DenseNet Structure. ....</b>	<b>12</b>
<b>Figure 4: Overview of Differentiable Augmentation.....</b>	<b>14</b>
<b>Figure 5: Samples from the training set .....</b>	<b>16</b>
<b>Figure 6: The distribution of ages and genders.....</b>	<b>17</b>
<b>Figure 7: Examples of original images and annotated images .....</b>	<b>19</b>
<b>Figure 8: Examples at each preprocessing stage.....</b>	<b>19</b>
<b>Figure 9: Generated images and real images .....</b>	<b>23</b>
<b>Figure 10: Training loss and validation loss of the models. ....</b>	<b>27</b>



## Preface

Conducting research is more rewarding than I could have imagined, though it is more complicated than I thought. None of this would have been possible without the help from many people. I would like to express my most profound appreciation to my advisor, Dr. Liang Zhan, for enabling me with opportunities to this exciting project and provide resources to carry on the experiments. I would also like to extend my deepest gratitude to my committee for valuable advice and insightful suggestions. Special thanks to Dr. Heng Huang, Dr. Ahmed Dallal, and Dr. Azime Can-Cimino. Their courses shed light on machine learning, digital signal processing, and pattern recognition. They offered extensive knowledge and helped me to gain practical experience during lessons. I gratefully acknowledge the help that I received from my family, friends, and colleagues; they provided me with encouragement and support throughout my study at the University of Pittsburgh.

## 1.0 Introduction

Biomedical imaging plays a critical role not only in the healthcare process but also in communication, education, and research, and it can show the structure of the body in great detail. For example, the function of the tissues within the body can help doctors examine the condition of a patient's organs or tissues without a surgical procedure. There are various types or modalities of imaging techniques, such as X-radiation (X-ray), Magnetic Resonance Imaging (MRI), and Computed Tomography (CT).

Bone Age Assessment (BAA) evaluates the maturity in infants, children, and adolescents using their hand radiographs. It plays an essential role in diagnosing an individual with growth disorders or endocrine disorders, such that needed treatments might be provided [1]. Computer-aided diagnosis (CAD) systems have been introduced to automatically extract features from regions of interest in radiographs, which are generally based on either the Greulich-Pyle atlas method or Tanner-Whitehouse (TW) scoring method [2]–[6].

Recently, artificial neural networks have gained an incredible amount of attention because of their success in image classification [7]. Several deep learning methods are used to perform automated bone age assessment by learning visual features. For example, a model based on a convolutional neural network is developed to segment regions of interest, standardize images, and conduct classification tasks with a pre-trained network [8]. A model called BoNet+ adopts a regression method based on densely connected convolutional networks to address poor-quality images and discovers that mean absolute error is a better loss function in the BAA problem than mean square error [9]. The ensemble of the regression and classification models suggests the performance can be improved [10], [11]. An existing model employs a generative adversarial

network to enhance image quality and fine-tunes a pre-trained network for transfer learning by gradually tuning from the top layer to the bottom layer to prevent the model trap in a local optimum [12].

Utilizing gradient-based learning to the network to train a potentially complex learning model, specifically deep neural networks, is referred to as end-to-end learning [13]. The power of end-to-end learning has been demonstrated in computer vision and various domains, such as Natural Language Processing [14]. An end-to-end learning model learns all the paths between the input and the output, and parameters in the model are simultaneously trained. Features are automatically learned from the training data set. Prior domain-specific knowledge might not be required for solving a given task, but more training samples are needed [13]. Training a complex model with a limited number of data tends to result in overfitting and degradation in the performance. Acquiring a sufficient amount of training data set in biomedical imaging may be arduous due to the human annotation burden and medical ethics. Data augmentation is widely used to address the issue.

The purpose of this thesis is to examine the performance of data augmentations by constructing a BAA system and using a publicly available data set for validation. Furthermore, we aim to study whether generating samples by GAN could be a valuable technique for providing more data to train the model.

## **2.0 Background**

### **2.1 Pediatric Bone Age**

X-ray imaging, a projection technique, is beneficial for producing images of organs, like bones, traversed by X-ray beam with lower energy than Gamma rays and higher power than visible light. X-rays, which is a type of electromagnetic radiation and are first discovered by Wilhelm Conrad Roentgen, awarded the 1901 Nobel Prize in Physics for this achievement, collide with electrons when they interact with an object. There are more collisions if a thing is dense or is made of higher atomic numbers elements. For example, bones are full of calcium, which has a relatively high atomic number; therefore, they absorb X-rays. On the other hand, soft tissues mostly have lower atomic numbers, like hydrogen, carbon, and oxygen. These interactions are recorded on the film and produce various degrees of brightness and darkness on the image. More X-rays penetrate tissues resulting in darker film. The differential contrast of hard and soft matter on the picture is the source of identifying anatomic structures [15]. The X-ray images are referred to as radiographs.

Measuring the physical maturity of children by using events during puberty throughout adolescence, such as the breast development for girls, voice change for boys, and appearance of public hair, is not without deficiency. The event sequences are coarsely spaced, and the coverage of developmental age span is uneven. X-ray imaging enables experts to inspect skeletal development, indicating strong evidence to the degree of maturity and is suitable to assess maturity. It results from all the bones develop into constant shape along a pathway to physical maturity [16]. This inspection is known as the Bone Age Assessment. It is a kind of evaluation by examining the shreds of evidence from skeletal development of hand and wrist bones to deduce

the bone age of an individual. Skeletal age can be used to assess growth disorder and growth potential through the gap between the estimated age and the chronological age and is also a measurement of epiphyseal center development, a necessary procedure in diagnosing endocrine disorders, skeletal dysplasias, and maturation in various syndromes [1].

There are two types of assessment systems universally adopted by pediatricians: Greulich-Pyle (GP) atlas and Tanner-Whitehouse (TW) scoring method. GP method describes the sequences of changes of bones and epiphyses occurring during childhood, generally. It examines the distance between the fastest and slowest maturing centers of ossification in hand and wrist radiographs against a set of the atlas at a certain age [17]. This matching method has the advantages of simplicity and availability of evaluating multiple ossification centers; however, the method was developed based on middle-class white populations. Therefore, it is liable to be sensitive to the subjective nature of different observers [1]. TW method is a system by scoring twenty bones of a left hand, and each bone is provided with nine possible ratings of maturity. Among twenty bones, radius, ulna, and eleven short bones (RUS) are generally helpful. Weighted ratings enable this method to provide sensitive bone age and overcome racial differences during maturation [18]. Both assessment systems rely on physician background knowledge, and they are time-consuming [16]. In recent years, numerous computer-aided systems have been developed to address this issue [6], [8], [19], [20].

In 2017, the Radiological Society of North America (RSNA) conducted Pediatric Bone Age Machine Learning Challenge. It provided a data set of hand radiographs with corresponded bone age reviewed by multiple experts to rate the performance of computer algorithms in estimating the skeletal age. A total of 260 individuals or teams worldwide registered the challenge, the performance is assessed by the mean absolute difference (MAD) between the model's

estimates and reviewers' estimates, and the winning approach obtained the MAD of around 4.3 months [21], [22].

## 2.2 Image Processing

Most computer-aided systems for BAA perform background removal to eliminate the noise and an area outside the patient body [1], [5], [23]. This area contains no pertinent information and might adversely affect the image analysis system and degenerate the performance [16].

### 2.2.1 Thresholding

Thresholding is a method of segmenting images and is widely used because of its effectiveness and simplicity. It replaces each pixel in a picture with specific values if the image intensity is greater or less than some fixed constant. Thresholding can be categorized into two general types: Global thresholding and local thresholding [24].

Global thresholding is based on the idea that an object in an image can be extracted from the background by comparing image values of pixels intensity with a threshold value if an image has a bimodal histogram [25]. In the case of binarization, it can be represented as:

$$g(x, y) = \begin{cases} 1, & f(x, y) > T \\ 0, & f(x, y) \leq T \end{cases} \quad (2-1)$$

The  $f(x, y)$  is denoted as pixel intensity in coordination  $(x, y)$ , and  $T$  is a threshold value determining the intensity range of an object and the background. The result is a binary image, where the value of 1 corresponds to an object, and the value of 0 corresponds to the background.

An image is divided into sub-images in local thresholding, and a threshold value is selected based on local properties for each sub-image [25]. Sub-images and different threshold values allow local thresholding to resolve non-uniform illumination over the image [24]. However, the size of sub-images and threshold values are difficult to set since the size is chosen globally. Some regions might require a larger size of sub-images, while some might require a smaller size to optimize the thresholding [16].

Unlike local thresholding, local adaptive thresholding computes a threshold value for each pixel by sliding a window through an image. Sauvola's method is one of the popular techniques. The threshold  $T(x, y)$  is calculated by mean  $m(x, y)$  and standard deviation  $s(x, y)$  in a  $w \times w$  window centered around the pixel  $(x, y)$ .  $R$  is the maximum value of the standard deviation, and  $k$  is a positive constant ranging from 0.2 to 0.5 to control the threshold value in the local window [24].

$$T(x, y) = m(x, y) \left[ 1 + k \left( \frac{s(x, y)}{R} - 1 \right) \right] \quad (2-2)$$

When the threshold is computed, an image where  $f(x, y) \in [0, 255]$  at location  $(x, y)$  can be denoted as:

$$g(x, y) = \begin{cases} 255, & f(x, y) > T(x, y) \\ 0, & f(x, y) \leq T(x, y) \end{cases} \quad (2-3)$$

### 2.2.2 Histogram equalization

Histogram equalization is a method for image enhancement by adjusting the intensity distribution of an image. It aims to map one distribution to another distribution that has generally more uniform intensity values. Let  $f$  be an image ranging from 0 to  $L - 1$  with 0 representing black and  $L-1$  representing white.  $P_n$  is denoted as the normalized histogram of  $f$  with a bin for every possible intensity  $n$ , where  $n = 0, 1, \dots, L - 1$ , and is recognized as the probability density function of  $f$ .

$$P_n = \frac{\text{number of pixels with intensity } n}{\text{total number of pixels}} \quad (2-4)$$

The histogram equalized image  $s$  has the form:

$$s = T(k) = \text{floor} \left( (L - 1) \sum_{n=0}^k P_n \right) \quad (2-5)$$

where  $k$  is pixel intensities and output values are round down to the nearest integer. For simplicity, suppose  $T$  is invertible and differentiable; therefore,  $s$  defined by  $T(n)$  is uniformly distributed on 0 to  $L - 1$  [26].

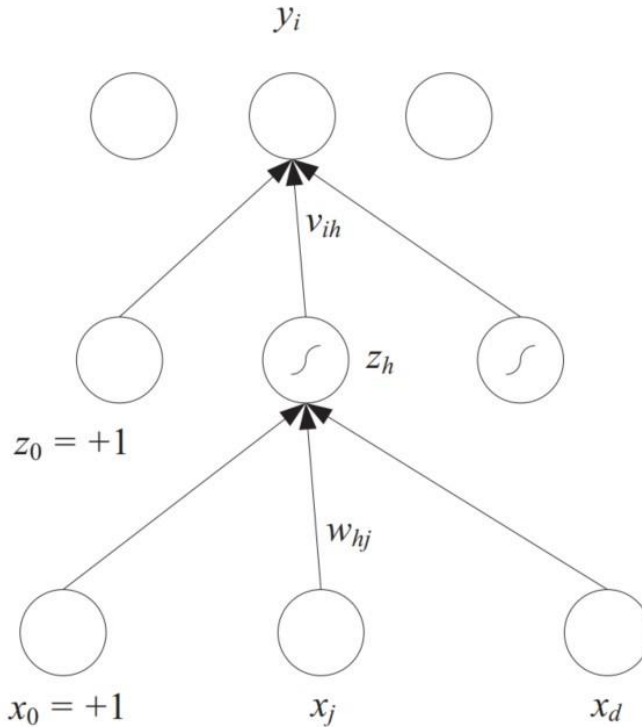
Contrast Limited Adaptive Histogram Equalization (CLAHE), a variation of adaptive histogram equalization, is formed based on splitting an image into several non-overlapping areas with almost equal sizes. It limits the contrast amplification to reduce noise amplification and renormalizes the histogram after the clipping limit [27].



## 2.3 Machine Learning

### 2.3.1 Artificial Neural Networks

Artificial Neural Networks (ANN) are inspired by the human brain's operating mechanism, mainly composed of neurons and synapses. Neurons are processing units that operate parallelly; synapses are connections among neurons, and information is transferred over them. The network composed of artificial neurons and synaptic connections was proposed in the 1960s, and it is called the perceptron model. The perceptron model calculates a value for each neuron by summing up activation values from all the connected neurons and multiplied them by their synaptic weights. Also, these neurons can be grouped into layers where neurons in a layer take input from neurons in the previous layer, and their outputs are fed to neurons in the following layers. This kind of model is called multilayer perceptron [28]. The structure of a multilayer perceptron is shown in Figure 1, where  $x_j$ ,  $j = 0, 1, \dots, d$  are the inputs,  $z_h$ ,  $h = 1, 2, \dots, H$  are the hidden units,  $z_0$  is the bias of the hidden layer,  $H$  is the dimensionality of the hidden space,  $y_i$ ,  $i = 1, 2, \dots, K$  are the output unit,  $w_{hj}$  are weights in the first layer, and  $v_{ih}$  are weights in the second layer [29].



**Figure 1: The structure of a multilayer perceptron. Adapted from [29]**

The backpropagation algorithm or generalized delta rule is one of the most popular methods for training a multilayer perceptron. It gave rise to numerous applications in different domains and fields because operations of a multilayer network start from a raw input and gradually apply a more complex transformation until an abstract representation is obtained. For example, we feed handwritten digits as input to the network. The neurons in the hidden layer combine image pixels to find basic descriptors, and the following layer combines these to observe more complicated shapes, like rectangles and circles. Layers successively process these features to form the representations of handwritten characters [28].

A convolutional neural network (CNN) is a variation of a multilayered network. The units between layers are not fully connected to the input units in the network but are connected to a small subset of the inputs. The units are defined as a window over the input space, and the operation

matches its input and weight for each unit. The idea is to combine the features in a more significant segment of the input space until a layer can look at the entire input, and the features will get fewer in terms of number and more abstract [28].

### **2.3.1.1 Convolutional Networks for Biomedical Image Segmentation**

U-Net is a popular network for image segmentation. The architecture of U-Net is symmetric and is mainly composed of two parts: encoder and decoder (Figure 2), and encoder and the decoder follow the general structure of a CNN. In the encoder, a contracting path, the number of feature channels is doubled at each downsampling step. The decoder, an expansive approach, consists of an upsampling part that halves the number of feature channels. The corresponding feature map from the contracting path is concatenated with the output of an upsampling function. This connection enables the network to pass context information to a successive layer of the decoder. Furthermore, there are no fully connected layers in the network, and this strategy allows seamlessly segmenting any large images by the overlap-tile method [30].

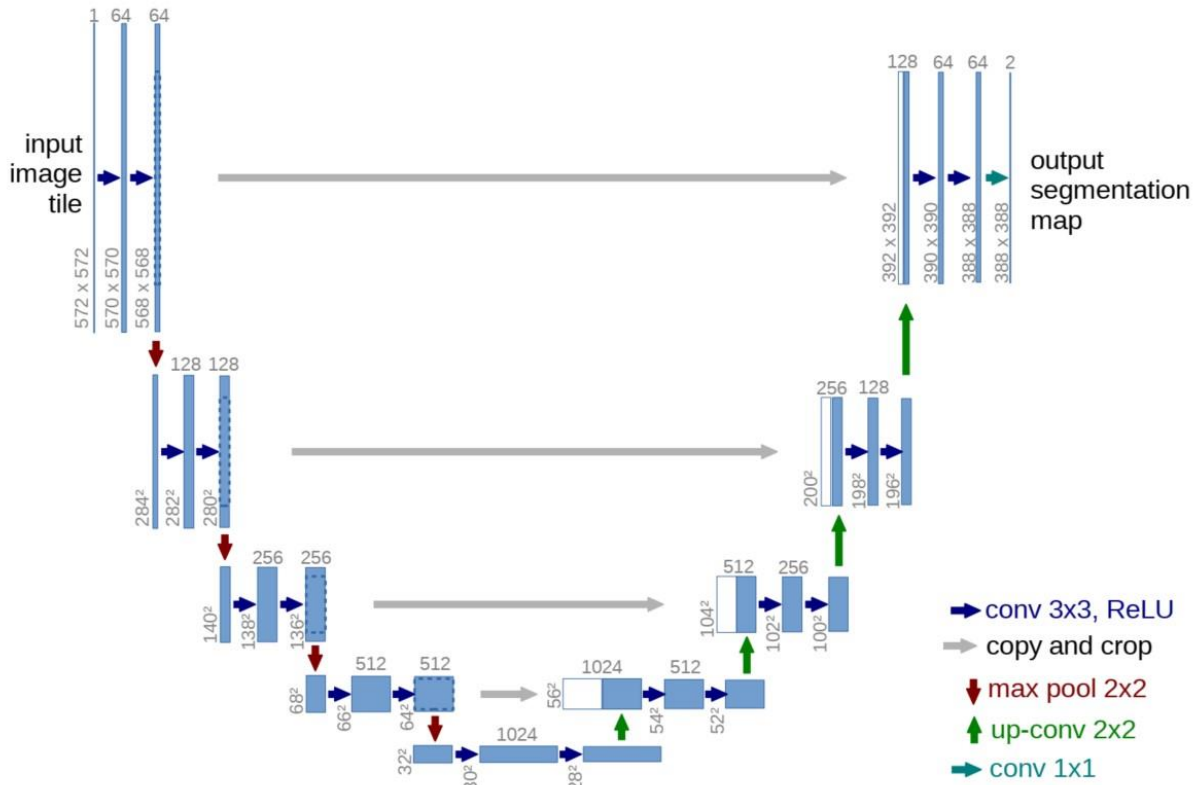


Figure 2: The architecture of U-Net. Adapted from [30]

### 2.3.1.2 Densely Connected Convolutional Networks

DenseNet is a type of CNN where layers in dense blocks directly are connected in a feed-forward fashion (as shown in Figure 3). A traditional  $L$  – layer CNN has  $L$  connections, while DenseNet has  $\frac{L(L+1)}{2}$  connections to improve the information flow and address the problem resulting from the vanishing gradient in deep neural networks. The densely connected method requires fewer parameters than conventional CNN by concatenating the preceding layer and the successive layer instead of adding them in terms of element-wise. The parameter efficiency also leads prevention of overfitting, especially on smaller training sets [31].

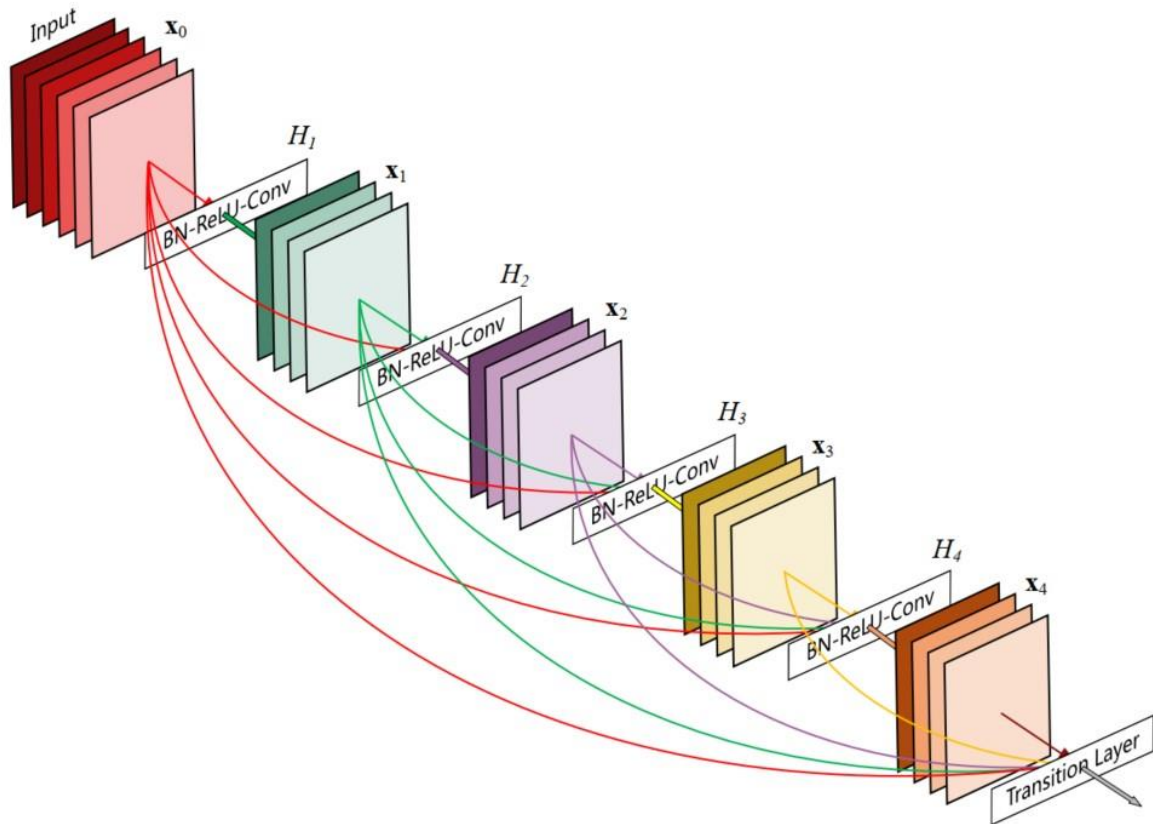


Figure 3: DenseNet Structure. Adapted from [31]

### 2.3.2 Generative Adversarial Network

Generative Adversarial Network (GAN) mainly comprises two models: a generative model  $G$  and a discriminative model  $D$ . It was proposed for generative modeling (i.e., a model that can generate augmented data). The model  $G$  aims to learn the distribution over data and produces fake data. The model  $D$  evaluates the probability of the input is actual data from the training set or counterfeit data from the  $G$ . These models are put against one another in a zero-sum game where the advantage won by one of the models is lost by the other [32].

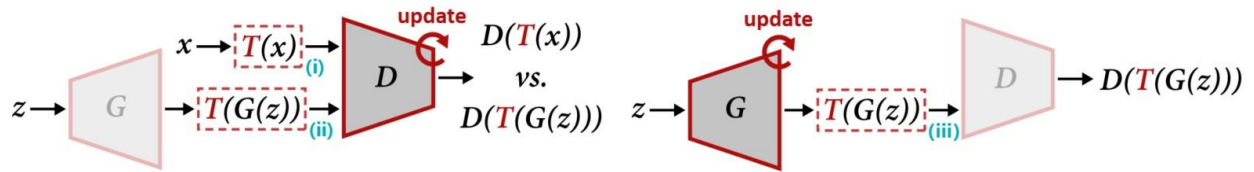
- The conditional generative adversarial network (cGAN): a type of GAN which is capable of learning a multi-modal model by adding the dependent labels as input to both  $G$  and  $D$ . It allows  $G$  to generate images of a given type based on a class label [33].
- The Auxiliary Classifier GAN (ACGAN): an extension of the conditional generative adversarial network by modifying  $D$  and adding a specialized loss function to predict the class label of input.  $D$  is trained without a class label as one of the inputs. It appears to stabilize the training process and enable the production of large-high-quality images [34].
- The Style-based GAN architecture (StyleGAN): one of the state-of-the-art networks in data-driven unconditional generative modeling for image synthesis. StyleGAN2 is an improved version of StyleGAN. The normalization used in the generator is redesigned, and a demodulation process, which is applied to the weights corresponded to each convolution layer, replaces the instance normalization. Moreover, the training method is modified by starting from low-resolution images then gradually shifting to images with higher resolution. The network topology remains unchanged during training [35].

### 2.3.3 Training Generative Adversarial Networks

Overfitting occurs when a model memorizes the training data and learns irrelevant information within samples, and the performance of unseen data degrades significantly. Training a model with less data is liable to overfitting. Data augmentation, which increases the diversity of examples, is one of the most popular solutions against overfitting. However, adding noisy data to a GAN while training not only deteriorates the model to learn the distribution over data but also interrupts the subtle balance between the generator and the discriminator. The adaptive discriminator augmentation method was proposed to address the issue. Data augmentation is

applied to all real images and generated images, and all these images are used to train either the generator or the discriminator. Standard differentiable primitives are used to differentiate the augmentations when training the generator. The discriminator outputs for the training set, validation set, and generated images are measured by the overfitting heuristics. The augmentation strength is adjusted based on the heuristics [36].

Another method called Differentiable Augmentation adopts a similar strategy, imposing data augmentations on both real and generated data for training the generator and discriminator. Still, the augmentation is differentiable such that the gradients of the augmented data are able to be propagated to the generator. Three types of transformations (i.e., Translation, Cutout, and Color) are chosen to demonstrate the performance. The overview of the method for update the generator and the discriminator is shown in Figure 4 [37].



**Figure 4: Overview of Differentiable Augmentation. Adapted from [37]**

## 2.4 Performance Metrics

Accuracy (ACC): represents the proportion of correctly predicted samples among a total number of pieces. It is used as a measurement of how well a binary classifier recognizes a condition. The formula is as follows:

$$ACC = \frac{\textit{True positive} + \textit{True negative}}{\textit{True positive} + \textit{True negative} + \textit{False positive} + \textit{False negative}} \quad (2-6)$$

Fréchet Inception Distance (FID): is the measurement of comparing the statistics of generated images to real images. It is the squared Wasserstein metric between two multivariate normal distributions, and lower FID is better because it means the distance of real and generated images between their activation distributions is closer. The formula of FID is:

$$FID = |\mu_r - \mu_g|^2 + \textit{tr} \left( \Sigma_r + \Sigma_g - 2 \sqrt{\Sigma_r \Sigma_g} \right) \quad (2-7)$$

where  $X_r$  denotes real images,  $X_r \sim \mathcal{N}(\mu_r, \Sigma_r)$ , and  $X_g$  denotes generated images,  $X_g \sim \mathcal{N}(\mu_g, \Sigma_g)$  [38], [39].



### 3.0 Methods

#### 3.1 Materials

The data sets are obtained from the RSNA Pediatric Bone Age Machine Learning Challenge and are provided by Children’s Hospital Colorado and Lucile Packard Children’s Hospital at Stanford. It contains 12,611 images for training, 1,425 images for validation, and 200 images for testing (Figure 5). For each image, the ground truth skeletal age ranges from 0+ to 19 years and is based on the estimates from six reviewers, and the GP standard, and sex, are provided. Reviewers’ evaluation determines the ground truth estimates for the testing set. They are corrected by calculating the mean of the inverse of the mean absolute difference between their estimates and the average of all reviewers’ estimates [21]. The distribution of ages and genders among data sets is shown in Figure 6.

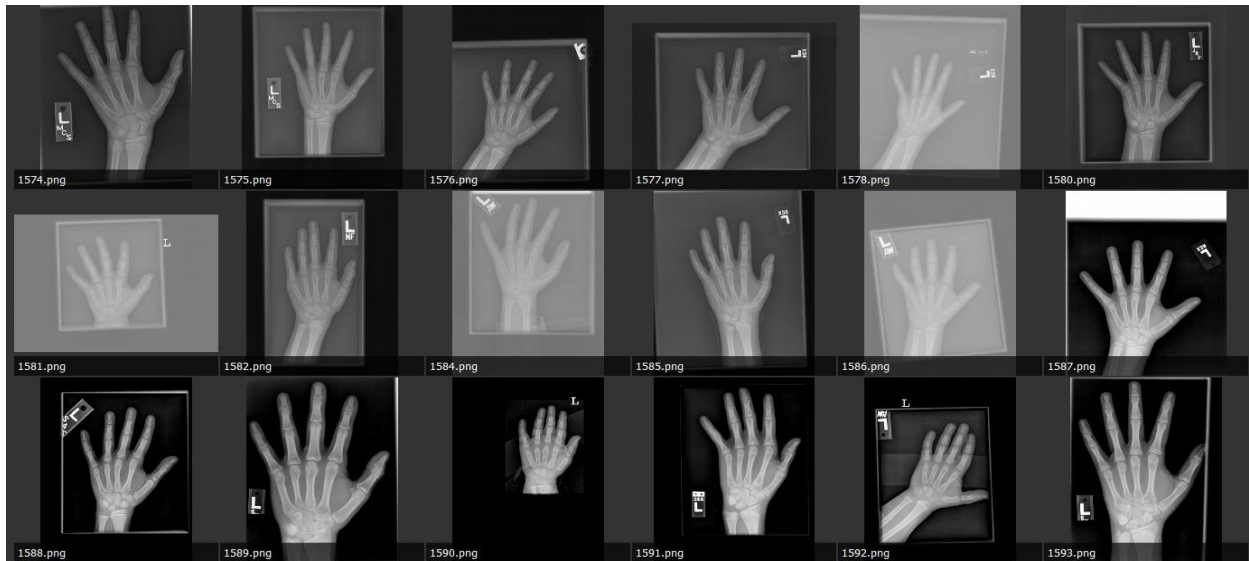
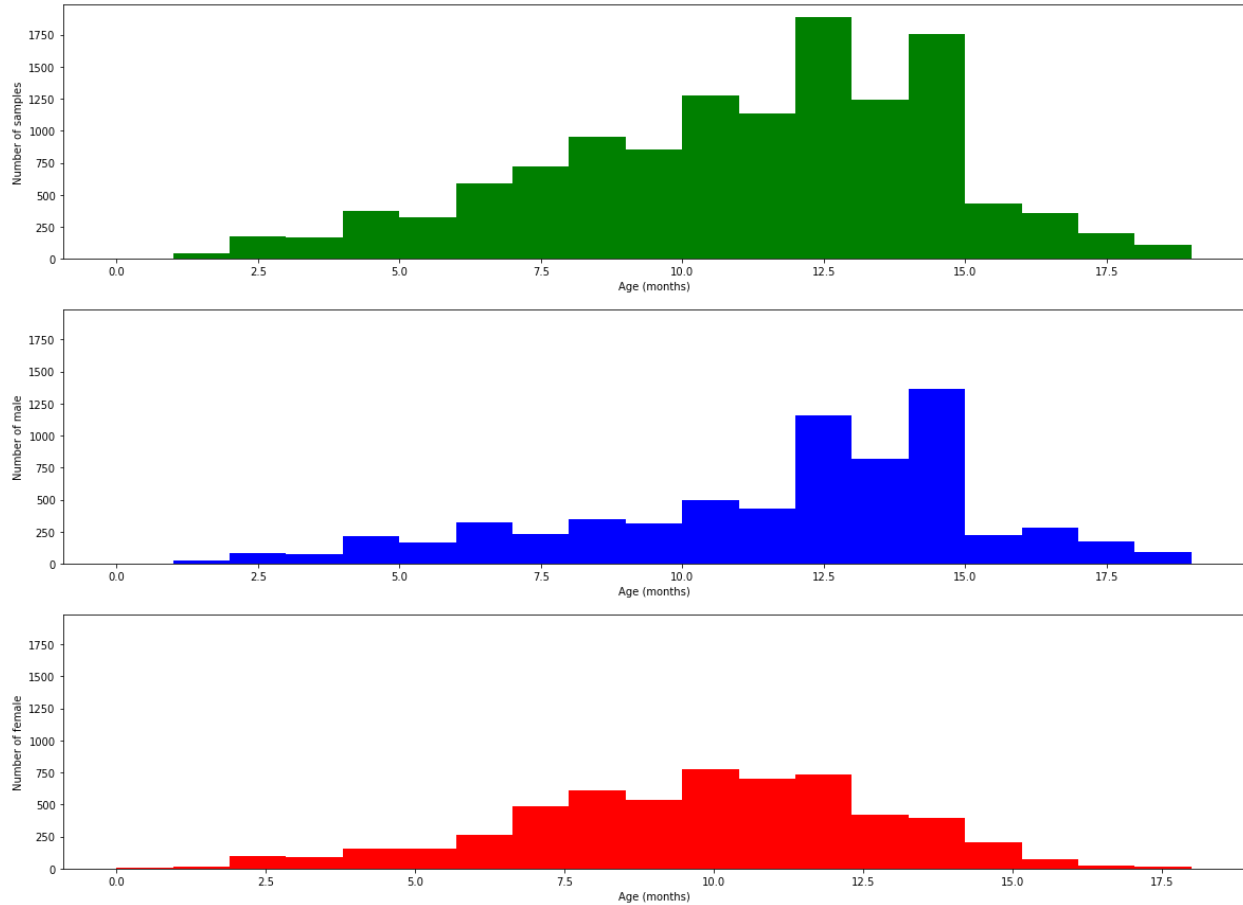


Figure 5: Samples from the training set



**Figure 6: The distribution of ages and genders, total (green bins), male (blue bins), and female (red bins).**

### 3.2 Data Preprocessing

First, the U-Net segmentation network is used to conduct background removal. A significant number of labeled images tend to result in good segmentation results, but it involves the laborious effort of labeling. Therefore, we randomly select 500 images from the training set and manually annotate the hand for each image using LabelMe [40] (Figure 7). Also, the encoder of the U-net is replaced with the pre-trained RegNetX backbone trained on ImageNet to improve

training efficiency [41]. Then all the images in the data sets are segmented using masks generated from the trained U-Net model and have the most significant connected components in each mask.

Histogram equalization and contrast limiting (the grid size is 7 by 7, and the threshold value for contrast limiting is 9.0) is applied on the segmented images to enhance contrast. During the equalization, bilinear interpolation is used to remove artifacts in the borders.

After the equalization, the contours of the hand are retrieved from the binary image using the border following algorithm [42]. Then the bounding rectangle of the convex hull is constructed to compute an approximation of the center of the hand. The hand contour and convex hull are used to locate the fingertips. Next, the images are rotated by calculating the angle between the center of the hand, the farthest fingertips from the center of the hand, and the reference point ( $x=0$ ,  $y$ =the  $y$ -axis value of the center of the hand) for each image. The rotated images are cropped and padded with a constant value of 0, such that the magnification ratio of images is 1:1. A classification model requires the input with consistent resolution; thus, images are scaled to 256 by 256 using bilinear interpolation to alleviate the computational cost, and images are normalized by subtracting the value of zero and dividing by the value of one (Figure 8).

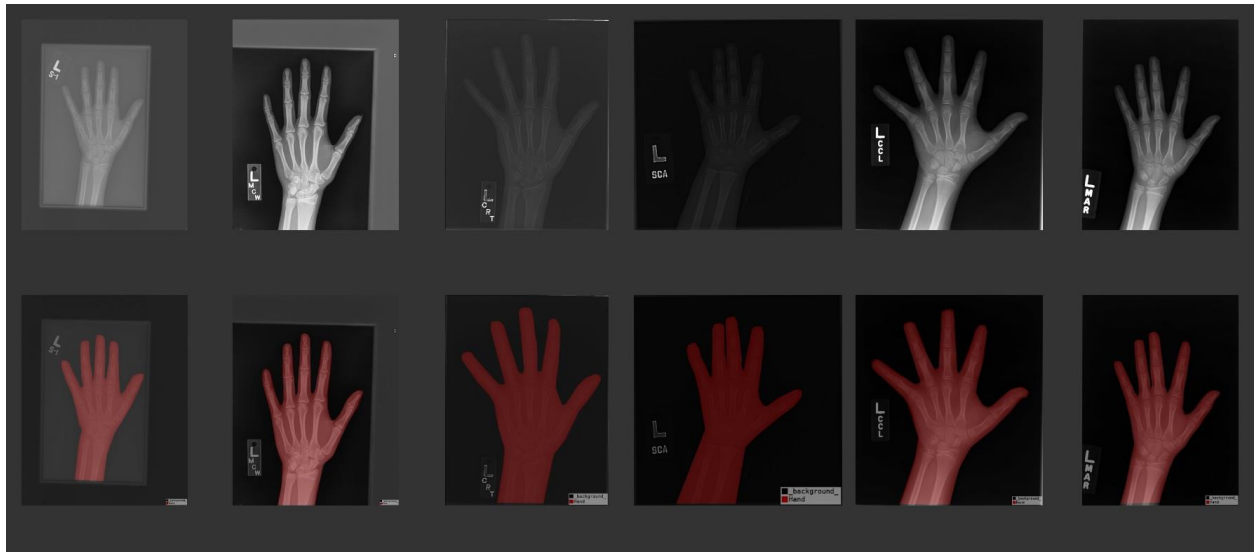


Figure 7: Examples of original images (top) and annotated images (bottom)

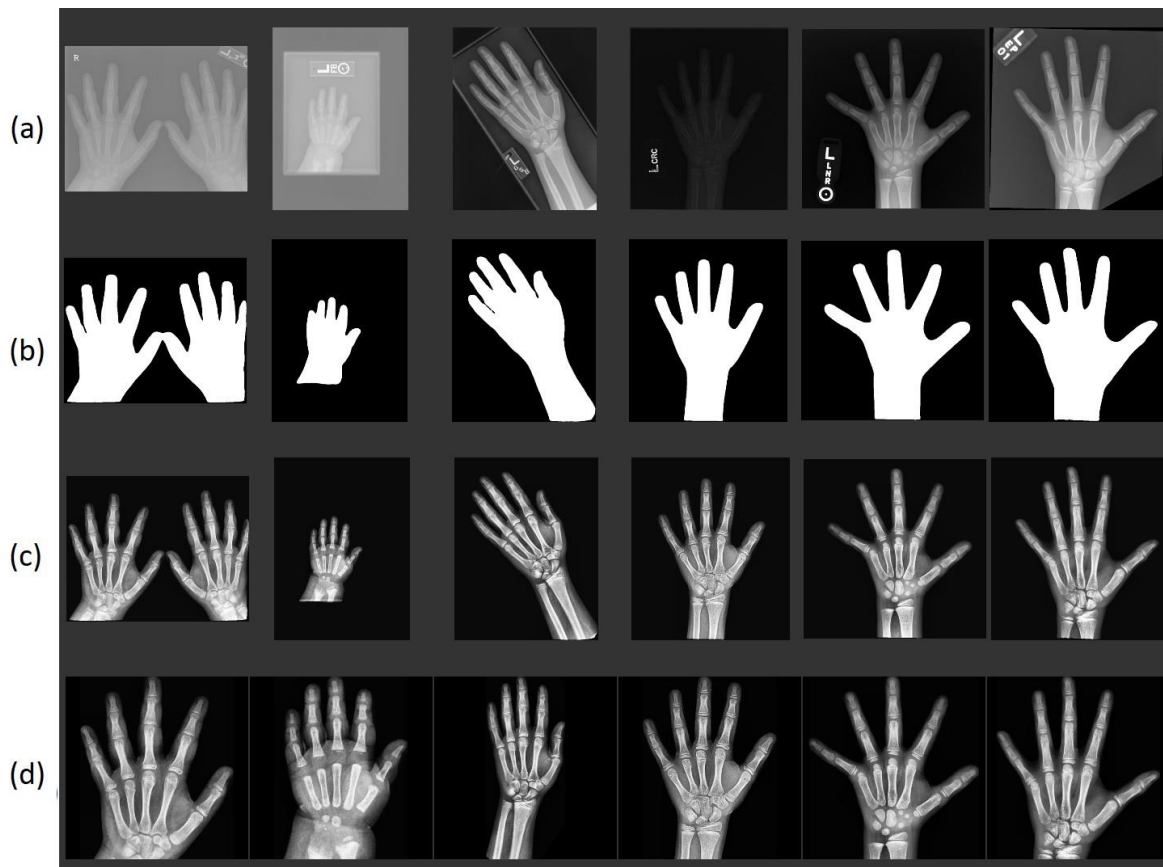


Figure 8: Examples at each preprocessing stage. (a) The original images. (b) Masks generated by the U-Net model. (c) Applying Histogram equalization (d) Processed images

### 3.3 Prediction network

The structure of the bone age assessment network is based on DenseNet (the growth rate is 32, the number of layers is 12) and ACGAN [31], [34], and the overview is shown in Table 1. All the bottleneck blocks and transition blocks are equipped with batch normalization [43], followed by the rectification nonlinearity [7]. The inputs are a fixed-sized 256 by 256 image and a gender label. Labels are embedded (the size of the diction of embeddings is 1, and the size of each embedding vector is 1). They are concatenated with the output of the Average Pooling layer as the input to the first linear layer. Two outputs are generated in the network: one is used to determine the input image is real or fake, the other is the predicted classes. The total number of parameters in the model is around 17.6M.

In the model, the loss function is composed of discrimination loss and classification loss.

$$\begin{aligned}\mathcal{L}_D &= \mathcal{L}_{real}^{dis} + \mathcal{L}_{real}^{cls} \\ &= E[\log P(\mathcal{S} = real|x)] + E[\log P(\mathcal{C} = class|x)]\end{aligned}\tag{3-1}$$

where  $x$  are images with labels,  $\mathcal{S}$  is the discrimination output, and  $\mathcal{C}$  is the classification output. The discrimination loss is measured by the mean squared error between the discrimination output and the target (a value of 1 representing real). The classification loss is measured by Cross Entropy Loss which is a combination of Log SoftMax and negative log-likelihood loss, described as:

$$loss(x|class) = -x[class] + \log\left(\sum_j \exp(x[j])\right)\tag{3-2}$$

The input is the raw scores for each class, and the target is a class index in the range  $[0, N - 1]$ , where  $N$  is the number of classes.

**Table 1: Structure of the assessment network**

Layers	Kernel	Strides	Padding	Output size	Activation
Convolution	$7 \times 7$	$2 \times 2$	$1 \times 1$	[batch, 64, 126, 126]	-
MaxPooling	$3 \times 3$	$2 \times 2$	0	[batch, 64, 62, 62]	-
Bottleneck Block (1)	$\begin{bmatrix} 1 \times 1 \\ 3 \times 3 \end{bmatrix}$	$\begin{bmatrix} 1 \times 1 \\ 1 \times 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \times 1 \end{bmatrix}$	[batch, 448, 62, 62]	ReLU
Transition Block (1)	$1 \times 1$	$1 \times 1$	0	[batch, 448, 31, 31]	ReLU
Bottleneck Block (2)	$\begin{bmatrix} 1 \times 1 \\ 3 \times 3 \end{bmatrix}$	$\begin{bmatrix} 1 \times 1 \\ 1 \times 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \times 1 \end{bmatrix}$	[batch, 832, 31, 31]	ReLU
Transition Block (2)	$1 \times 1$	$1 \times 1$	0	[batch, 832, 15, 15]	ReLU
Bottleneck Block (3)	$\begin{bmatrix} 1 \times 1 \\ 3 \times 3 \end{bmatrix}$	$\begin{bmatrix} 1 \times 1 \\ 1 \times 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \times 1 \end{bmatrix}$	[batch, 1216, 15, 15]	ReLU
Transition Block (3)	$1 \times 1$	$1 \times 1$	0	[batch, 1216, 7, 7]	ReLU
Bottleneck Block (4)	$\begin{bmatrix} 1 \times 1 \\ 3 \times 3 \end{bmatrix}$	$\begin{bmatrix} 1 \times 1 \\ 1 \times 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \times 1 \end{bmatrix}$	[batch, 1600, 7, 7]	ReLU
Transition Block (4)	$1 \times 1$	$1 \times 1$	0	[batch, 1600, 3, 3]	ReLU
Bottleneck Block (5)	$\begin{bmatrix} 1 \times 1 \\ 3 \times 3 \end{bmatrix}$	$\begin{bmatrix} 1 \times 1 \\ 1 \times 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \times 1 \end{bmatrix}$	[batch, 1984, 3, 3]	ReLU
Average Pooling	$3 \times 3$	$3 \times 3$	0	[batch, 1984, 1, 1]	ReLU
Linear (1)	-	-	-	[batch, 1]	Tanh
Linear (2-1)	-	-	-	[batch, 1000]	LeakyReLU
Linear (2-2)	-	-	-	[batch, 500]	LeakyReLU
Linear (2-3)	-	-	-	[batch, 229]	-

### 3.4 Experiment Settings

The network is trained using Adam optimizer with a learning rate of 0.0002, and the first and the second-moment estimates are 0.5 and 0.999, respectively [44]. The number of mini-batch sizes is 32, the number of CPU threads to use during batch generation is 8, the number of image channels is set to 1, and the number of classes is 229. All the training sets, validation sets, and testing sets are used. The experiment is implemented with the Pytorch framework on NVIDIA Tesla P100 GPU.

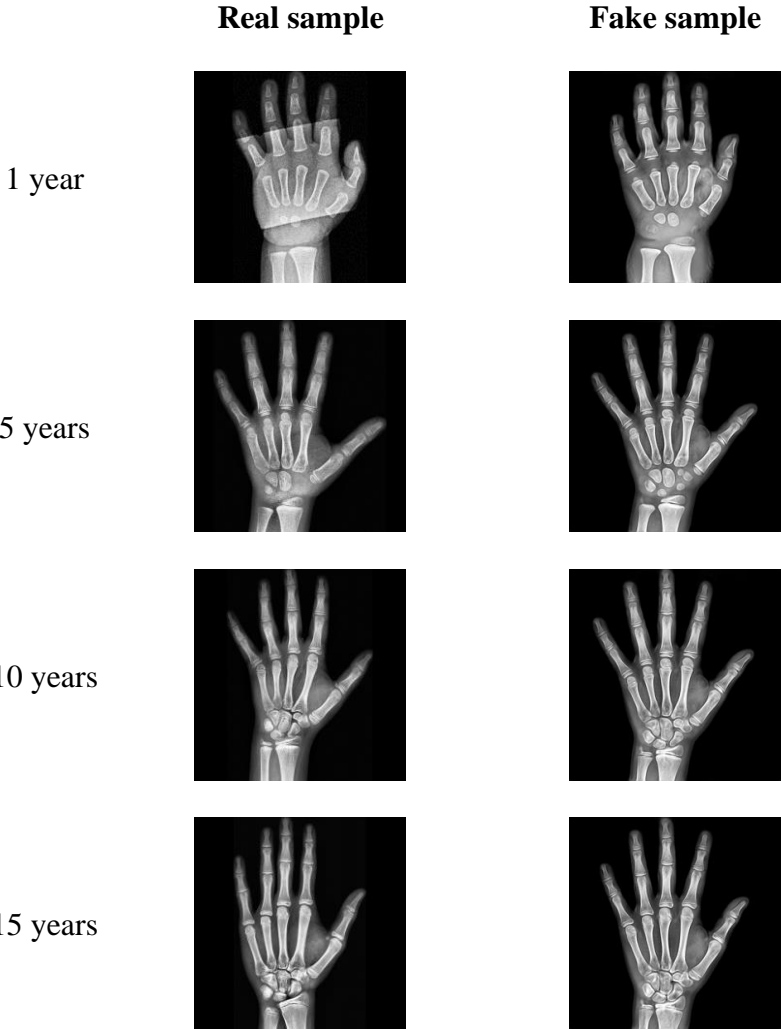
Table 2 shows the parameters for transformations, including horizontal flipping, vertical flipping, rotation, cutout, and blurring. They are performed with the probability of applying: 0.25 to improve the resiliency of the model.

**Table 2: Parameters for transformation**

<b>Methods</b>	<b>Parameters</b>
Horizontal flipping	0 / 1 (flip or not)
Vertical flipping	0 / 1 (flip or not)
Rotation	angle: [-90°, 90°]
Cutout	number of regions: 8 pixels maximum height, width: 32 pixels
Blurring	maximum kernel size: 7 pixels

To simulate the situation where training sets are not adequate, we split 10 percent of the preprocessed training images in a stratified fashion using bone ages rounded to year as the class labels (i.e., from age 0 to 19). Because the number of samples in the different age groups is not

even, the maximum number of examples of age is constrained to 50. A total of 917 images are selected out of 1,262 images (10% of training data) and are used to train the conditioned StyleGAN2 network with differentiable augmentation techniques, including color, translation, and cutout [37]. We used 50k samples for FID calculation. The model achieves an FID of 51.47 during the training for 160k images. Then the trained StyleGAN2 model is further used to be a generator to provide a continuously supported probability density function for training the assessment network (Figure 9).



**Figure 9: Generated images and real images**



In this setting, the loss function in the assessment network is modified by adding discrimination loss of fake images and is denoted as:

$$\mathcal{L}_{GD} = \mathcal{L}_{real}^{dis} + \mathcal{L}_{real}^{cls} + \mathcal{L}_{fake}^{dis} \quad (3-3)$$

## 4.0 Results

Table 3 provides the performance of the proposed model along with two variant settings on the testing set. When training with 100% data set, the model with transformations achieves the best MAE of 5.8, RMSE of 7.3, and accuracy (within 1 year) of more than 90%. When training with only a 10% data set, the model Base outperforms its two variants and obtains MAE of 9.5, RMSE of 12.5, and accuracy (within 1 year) of 72%. Although the model, +GAN, does not outperform the baseline, its performance is better than the model, + Transform.

**Table 3: Performance of different settings**

Model	MAE		RMSE		ACC		ACC	
	(months)		(months)		(Within 1 year)		(Within 2 years)	
	Training data							
	(100%)	(10%)	(100%)	(10%)	(100%)	(10%)	(100%)	(10%)
Base (baseline)	7.425	<b>9.48</b>	9.459	<b>12.5</b>	82.5%	<b>72%</b>	99%	<b>96%</b>
+Transform	<b>5.805</b>	14.62	<b>7.296</b>	18.81	<b>92%</b>	50.5%	<b>99.5%</b>	80.5%
+GAN	-	12.5	-	15.79	-	55.5%	-	89%

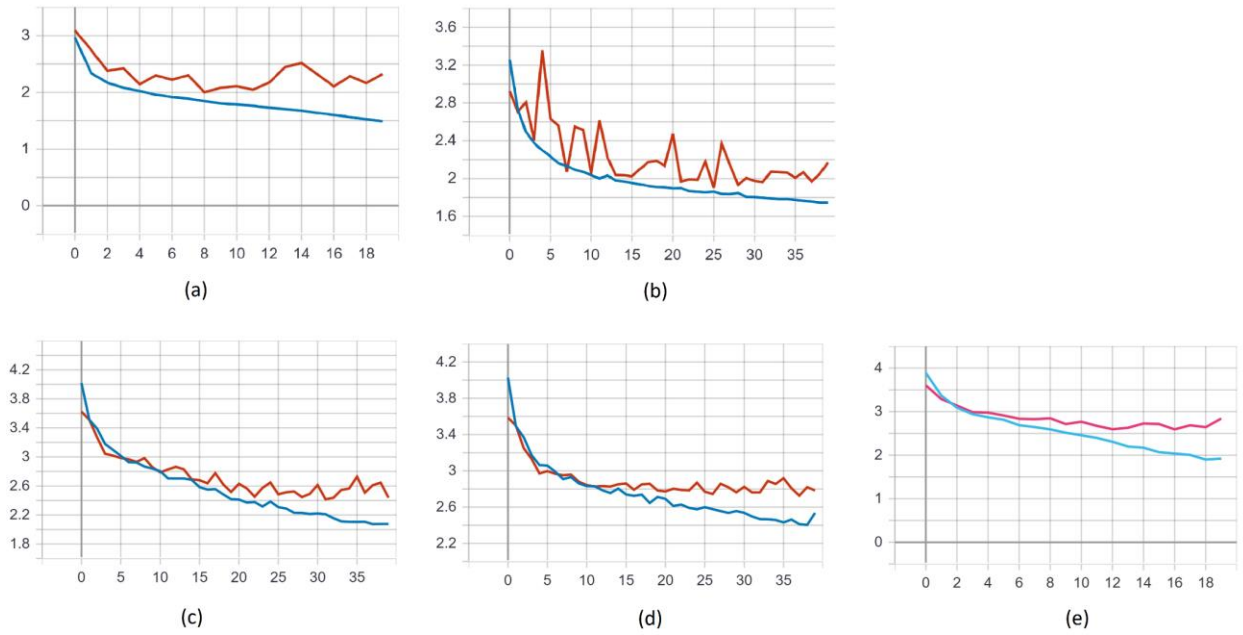
*MAE=Mean Absolute Error; RMSE=Root Mean Square Error; ACC=Accuracy*

Compared with existing models in BAA using the same testing data (Table 4), our method achieves a comparable result and further improves the performance of a DenseNet-based network.

**Table 4: Comparison with the published models on RSNA testing set**

<b>Item</b>	<b>Method</b>	<b>MAE</b>
1. Alexander et al. [21]	Inception V3	4.2 months
2. Iglovikov et al. [10]	VGG	5 months
3. Proposed	DenseNet+Transformation	5.8 months
4. Zhao et al. [12]	DenseNet	male: 6 months female: 6.3 months
5. Pan et al. [11]	CNN	7.4 months

The training loss and validation loss for the 100% training set and 10% training set are shown in Figure 10. Except for the model, +Transform, using 10% data, the training loss of others decreases gradually.



**Figure 10: Training loss (blue line) and validation loss (red line) of the models. (a) Base using 100% data. (b) Base+Transform using 100% data. (c) Base using 10% data. (d) Base+Transform using 10% data. (e) Base+GAN using 10% data**

## 5.0 Discussion

The hand segmentation is conducted via transfer learning. It demonstrates the effectiveness and applicability of transferring pre-trained deep learning weights from a different data set to a similar field. Furthermore, it allows a model to better find global optima with limited labeled data.

The efficiency of data augmentation is assessed in this study. The results support the finding by other studies that transformation is a standard solution against overfitting and improves the performance by providing relevant data to stabilize a model. In addition, it leads to a better converge.

In the experiments using 100% training data, we observe that the performance on testing data increase compared to the performance measured during validating models on the validation data. Though the number of the data sets is not the same, the distribution on age groups and gender are similar. One explanation for this improvement could be that testing sets have more accurate labeling of estimated age than the validation sets. According to the description of the data set, the ground truth skeletal age of testing sets is based on the estimates from reviewers. Therefore, it is corrected by the mean absolute difference among reviewers' assessments.

Differentiable Augmentation shows improvement on data efficiency of GAN in BAA when training the GAN with no more than 50 samples for each class. However, a sufficient number of high-quality training samples are still needed. Training a complex model with a limited number of high-quality data can overfit and degrade the model's generalization. They could not be replaced entirely with synthesized data, even though GAN-generated images are visually real-like and share close distribution over training data. The baseline has the best performance in the experiments

using 10% training data. The reason might be too many irrelevant samples are added and disturbs the model to parse out the relationship between the input and the output.

A deep learning model is powerful, but it still could be improved by several methods. For example, we did not consider gender while training a conditional GAN, so fake genders are generated to train the prediction model without gender labels. In clinical practice, male and female cohorts are determined with different standards. Therefore, the real distribution over samples might not be learned correctly. Also, the region of interest in this study is the whole hand of radiographs. Implementing feature selection based on clinical practice, such as GP atlas and TW method, could develop a more robust model, even when samples are not adequate. These procedures might potentially give rise to the development of a more generalized and state-of-the-art BAA system.

## 6.0 Conclusions

In this thesis, we reviewed several BAA models based on traditional methods and deep learning, implemented image processing techniques to enhance the image quality of the data set, and presented a BAA model. The model has obtained comparable performance. We also investigated the efficiency of data augmentations from the data-efficient perspective. Further techniques are necessary for a deep learning model with limited data.

## Bibliography

- [1] F.Cao, H. K.Huang, E.Pietka, and V.Gilsanz, “Digital hand atlas and web-based bone age assessment: System design and implementation,” *Comput. Med. Imaging Graph.*, vol. 24, no. 5, pp. 297–307, 2000, doi: 10.1016/S0895-6111(00)00026-4.
- [2] A.Zhang, A.Gertych, and B. J.Liu, “Automatic bone age assessment for young children from newborn to 7-year-old using carpal bones,” *Comput. Med. Imaging Graph.*, vol. 31, no. 4–5, pp. 299–310, 2007, doi: 10.1016/j.compmedimag.2007.02.008.
- [3] E.Pietka, A.Gertych, S.Pospiech, F.Cao, H. K.Huang, and V.Gilsanz, “Computer-assisted bone age assessment: Image preprocessing and epiphyseal/metaphyseal ROI extraction,” *IEEE Trans. Med. Imaging*, vol. 20, no. 8, pp. 715–729, 2001, doi: 10.1109/42.938240.
- [4] N.Shobha Rani, C. R.Yadhu, and U.Karthik, “Chronological age assessment based on wrist radiograph processing - Some novel approaches,” *J. Intell. Fuzzy Syst.*, vol. 40, no. 5, pp. 8651–8663, 2021, doi: 10.3233/JIFS-190779.
- [5] J. M.Sotoca, J. M.Iñesta, and M. A.Belmonte, “Hand bone segmentation in radioabsorptiometry images for computerised bone mass assessment,” *Comput. Med. Imaging Graph.*, vol. 27, no. 6, pp. 459–467, 2003, doi: 10.1016/S0895-6111(03)00053-3.
- [6] J.Liu, J.Qi, Z.Liu, Q.Ning, and X.Luo, “Automatic bone age assessment based on intelligent algorithms and comparison with TW3 method,” *Comput. Med. Imaging Graph.*, vol. 32, no. 8, pp. 678–684, 2008, doi: 10.1016/j.compmedimag.2008.08.005.
- [7] A.Krizhevsky, I.Sutskever, and G. E.Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May2017, doi: 10.1145/3065386.
- [8] H.Lee *et al.*, “Fully Automated Deep Learning System for Bone Age Assessment,” *J. Digit. Imaging*, vol. 30, no. 4, pp. 427–441, 2017, doi: 10.1007/s10278-017-9955-8.
- [9] J.Guo, J.Zhu, H.Du, and B.Qiu, “A bone age assessment system for real-world X-ray images based on convolutional neural networks,” *Comput. Electr. Eng.*, vol. 81, 2020, doi: 10.1016/j.compeleceng.2019.106529.
- [10] V. I.Iglovikov, A.Rakhlin, A. A.Kalinin, and A. A.Shvets, “Paediatric bone age assessment using deep convolutional neural networks,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11045 LNCS, pp. 300–308, 2018, doi: 10.1007/978-3-030-00889-5\_34.



- [11] X.Pan, Y.Zhao, H.Chen, D.Wei, C.Zhao, and Z.Wei, “Fully Automated Bone Age Assessment on Large-Scale Hand X-Ray Dataset,” *Int. J. Biomed. Imaging*, vol. 2020, 2020, doi: 10.1155/2020/8460493.
- [12] C.Zhao, J.Han, Y.Jia, L.Fan, and F.Gou, “Versatile Framework for Medical Image Processing and Analysis with Application to Automatic Bone Age Assessment,” *J. Electr. Comput. Eng.*, vol. 2018, 2018, doi: 10.1155/2018/2187247.
- [13] T.Glasmachers, “Limits of end-to-end learning,” *J. Mach. Learn. Res.*, vol. 77, pp. 17–32, 2017.
- [14] M.Lewis, D.Yarats, Y. N.Dauphin, D.Parikh, and D.Batra, “Deal or no deal? End-to-end learning for negotiation dialogues,” *EMNLP 2017 - Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 2443–2453, 2017, doi: 10.18653/v1/d17-1259.
- [15] D. L.Rubin, H.Greenspan, and J. F.Brinkley, “Biomedical imaging informatics,” *Biomed. Informatics Comput. Appl. Heal. Care Biomed. Fourth Ed.*, pp. 285–327, 2014, doi: 10.1007/978-1-4471-4474-8\_9.
- [16] C. H.Yan, *Segmentation of Hand Bone for Bone Age Assessment*. 2013.
- [17] L. M.Bayer, “RADIOGRAPHIC ATLAS OF SKELETAL DEVELOPMENT OF THE HAND AND WRIST: Second Edition,” *Calif. Med.*, vol. 91, no. 1, p. 53, Jul.1959, [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1577858/>.
- [18] J. O.Forfar, “Reviewed Work(s): Assessment Of Skeletal Maturity And Prediction Of Adult Height by J. M. Tanner, R. H. Whitehouse, W. A. Marshall, M. J. R. Healy and H. Goldstein,” *Br. Med. J.*, vol. 1, no. 6018, p. 1153, 1976, doi: 10.1136/bmj.1.1203.118.
- [19] M.Mansourvar, M. A.Ismail, T.Herawan, R.Gopal Raj, S.Abdul Kareem, and F. H.Nasaruddin, “Automated bone age assessment: Motivation, taxonomies, and challenges,” *Comput. Math. Methods Med.*, vol. 2013, 2013, doi: 10.1155/2013/391626.
- [20] C.Spampinato, S.Palazzo, D.Giordano, M.Aldinucci, and R.Leonardi, “Deep learning for automated skeletal bone age assessment in X-ray images,” *Med. Image Anal.*, vol. 36, pp. 41–51, 2017, doi: 10.1016/j.media.2016.10.010.
- [21] S. S.Halabi *et al.*, “The RSNA pediatric bone age machine learning challenge,” *Radiology*, vol. 290, no. 3, pp. 498–503, 2019, doi: 10.1148/radiol.2018180736.
- [22] A.DelBimbo *et al.*, *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part IV*, vol. 12664. Cham: Springer International Publishing AG, 2021.
- [23] L.Yang, Y.Zhang, J.Chen, S.Zhang, and D. Z.Chen, “Suggestive annotation: A deep active learning framework for biomedical image segmentation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10435 LNCS, no. 1, pp. 399–407, 2017, doi: 10.1007/978-3-319-66179-7\_46.

- [24] F.Shafait, D.Keysers, and T.Breuel, *Efficient implementation of local adaptive thresholding techniques using integral images*, vol. 6815. 2008.
- [25] S. U.Lee, S.Yoon Chung, and R. H.Park, “A comparative performance study of several global thresholding techniques for segmentation,” *Comput. Vision, Graph. Image Process.*, vol. 52, no. 2, pp. 171–190, 1990, doi: 10.1016/0734-189X(90)90053-X.
- [26] R. C.Gonzalez and R. E.Woods, *Digital image processing*. Upper Saddle River, N.J.: Prentice Hall, 2008.
- [27] S. M.Pizer *et al.*, “Adaptive Histogram Equalization and Its Variations.,” *Comput. vision, Graph. image Process.*, vol. 39, no. 3, pp. 355–368, 1987, doi: 10.1016/S0734-189X(87)80186-X.
- [28] E.Alpaydin, *Machine learning: The new AI*. Cambridge: MIT Press, 2016.
- [29] E.Alpaydin, *Introduction to machine learning*, Third edit. Cambridge, Massachusetts ; The MIT Press, 2014.
- [30] O.Ronneberger, P.Fischer, and T.Brox, “U-net: Convolutional networks for biomedical image segmentation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4\_28.
- [31] G.Huang, Z.Liu, L.Van DerMaaten, and K. Q.Weinberger, “Densely connected convolutional networks,” *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 2261–2269, 2017, doi: 10.1109/CVPR.2017.243.
- [32] I.Goodfellow *et al.*, “Generative adversarial networks,” *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Jun.2014, doi: 10.1145/3422622.
- [33] M.Mirza and S.Osindero, “Conditional Generative Adversarial Nets,” pp. 1–7, 2014, [Online]. Available: <http://arxiv.org/abs/1411.1784>.
- [34] A.Odena, C.Olah, and J.Shlens, “Conditional image synthesis with auxiliary classifier gans,” *34th Int. Conf. Mach. Learn. ICML 2017*, vol. 6, no. 1, pp. 4043–4055, Oct.2017, Accessed: Jun.28, 2021. [Online]. Available: <https://arxiv.org/abs/1610.09585>.
- [35] T.Karras, S.Laine, M.Aittala, J.Hellsten, J.Lehtinen, and T.Aila, “Analyzing and improving the image quality of stylegan,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 8107–8116, 2020, doi: 10.1109/CVPR42600.2020.00813.
- [36] T.Karras, M.Aittala, J.Hellsten, S.Laine, J.Lehtinen, and T.Aila, “Training Generative Adversarial Networks with Limited Data,” *NeurIPS*, no. NeurIPS, 2020, [Online]. Available: <http://arxiv.org/abs/2006.06676>.
- [37] S.Zhao, Z.Liu, J.Lin, J.-Y.Zhu, and S.Han, “Differentiable Augmentation for Data-Efficient GAN Training,” no. NeurIPS, 2020, [Online]. Available: <http://arxiv.org/abs/2006.10738>.

- [38] D. C.Dowson and B.V.Landau, “The Fréchet distance between multivariate normal distributions,” *Journal of Multivariate Analysis*, vol. 12, no. 3. pp. 450–455, 1982, doi: 10.1016/0047-259X(82)90077-X.
- [39] M.Heusel, H.Ramsauer, T.Unterthiner, B.Nessler, and S.Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 6627–6638, 2017.
- [40] K.Wada, “labelme: Image Polygonal Annotation with Python.” 2016.
- [41] P.Yakubovskiy, “Segmentation Models Pytorch,” *GitHub repository*. GitHub, 2020.
- [42] S.Suzuki and K. A.be, “Topological structural analysis of digitized binary images by border following,” *Comput. Vision, Graph. Image Process.*, vol. 30, no. 1, pp. 32–46, 1985, doi: 10.1016/0734-189X(85)90016-7.
- [43] S.Ioffe and C.Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *32nd International Conference on Machine Learning, ICML 2015*, Feb. 2015, vol. 1, no. 6, pp. 448–456, [Online]. Available: <http://arxiv.org/abs/1502.03167>.
- [44] D. P.Kingma and J. L.Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015, pp. 1–15.