Exploring Different CNN Architectures for Predicting Body Measurements from 2D Images

Ling-Yau Lee

Word count: 7,338

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

2053925

COMMITTEE

Supervisor: dr. Çiçek Güven

Second reader: Prof. dr. Eric Postma

External supervisor: dr. Marleen Balvert

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, the Netherlands

DATE

January 14, 2022

ACKNOWLEDGEMENTS

I would like to extend my special appreciation to:

My supervisor Professor Çiçek Güven for providing guidance and support during the entire thesis trajectory. Besides her continued help throughout the semester, she also inspired me in her machine learning course to tack real-life societal problems.

My second reader Professor Erik Postma for providing direction and valuable expertise. It has been a privilege to work with a true expert in the field of deep learning.

Professor Marleen Balvert from the Zero Hunger Lab who has supervised me from the very beginning and has made this research possible.

My beloved wife Shun Yee Fong for standing by my decision to study again and to pursue my big data ambitions. And for giving never-ending support, even when relocating to the Netherlands for the last semester during the global pandemic.

Exploring Different CNN Architectures for Predicting Body Measurements from 2D Images

Ling-Yau Lee

ABSTRACT

This thesis explores different convolutional neural networks (CNNs) for predicting bodily measurements from 2D images with the intention of improving accuracy of automated malnutrition assessment. The research question is: How effective are EfficientNet and Big Transfer (BiT) architectures in predicting body measurements from 2D images compared to the baseline model of Mohammed Khan (2020) which uses Restnet50 and Inceptionv3 architectures? The data set consists of virtual images of human body shapes which are based on 3D human body meshes originating from the Civilian American and European Surface Anthropometry Resource (CAESAR) dataset. On a subset of this dataset principal component analysis was used to obtain the body measurements of height and waist circumference which acted as the labels to train our CNNs on. Both height and weight are essential dimensions for detecting malnutrition. This study explores EfficientNet and Big Transfer architectures ability to predict bodily measurements in comparison to the baseline models. Findings show that BiT has an average MAE of 68.1mm and 120.8mm while EfficientNet has 84.6mm and 120.2mm for height and weight respectively. For the baseline models, Inception has an average MAE of 64.7mm and 121.8 while ResNet has an average MAE of 74.2mm and 121.5mm for height and weight respectively. In conclusion, we find that BiT outperforms ResNet however it does not outperform Inception. EfficientNet is the worst performer, however due to the efficiency, it can still prove to be advantageous when used in mobile devices.

DATA SOURCE / CODE / ETHICS STATEMENT

The virtual images used in this research are anonymized. The images have been obtained with permission from Zero Hunger Lab for this study.

The code in this thesis was replicated using code from Mohammed Khan (2020) with permission. The author of this thesis acknowledges that they do not have any legal claim to this data or code.

The code used in this thesis is publicly available with the location provided in the appendix B.

1 INTRODUCTION

According to World Food Programme estimates, approximately 690 million people which is equivalent to 8.9% of the world population are currently suffering from hunger. In 2019, 6.9% or 47 million children under the age of 5 were affected by acute malnutrition. From a societal point of view, COVID-19 has only exacerbated the situation and immediate action is needed (World Food Programme, 2020).

The Zero Hunger Lab from Tilburg University is taking part in a project where the goal is to develop algorithms which will be used in automated malnutrition assessment. These algorithms are incorporated into an app known as the Child Growth Monitor ("Child Growth Monitor", 2021). The ultimate purpose of the Child Growth Monitor (CGM) is to detect malnutrition based on bodily measurements taken from images. Body measurements such as height, weight and waist circumference can be used as indicators of malnutrition. According to child growth standards, when a child is too short for his age, this could be due to chronic disease ("Child Growth Standards", 2021). The focus of the CGM is primarily aimed towards children considering they are more susceptible to chronic malnutrition (Bandsom, 2021). Detecting malnutrition will enable timely nourishment and prevent adverse health effects. Upon successful completion, the CGM can be deployed globally to automatically identify cases of malnutrition.

In the context of the CGM, the goal of this research is to explore different neural network architectures and how these configurations can lead to better prediction accuracy. First, this study will replicate the neural network architecture of Mohammed Khan (2020) and this will form the baseline to compare with, hereafter referred to as the baseline model. Subsequently, EfficientNet and Big Transfer (BiT) architectures are applied with the goal of improving prediction accuracy.

The work of Mohammed Khan (2020) was novel in nature as it involved predicting from full body images without any reference objects within the image. From the related works section, you can see there is research on neural networks. However, further research specific to the CGM context is required. Deeper understanding of this lays the foundation for better prediction accuracy which is an important prerequisite for launching the CGM app and enable it to achieve its prediction objectives.

On a personal level, my aspiration is to utilize my data science skills and its versatile tools to contribute to real-life societal problems. This thesis is a suitable platform to realize this aspiration.

The main research question will be:

- How effective are EfficientNet and Big Transfer architectures in predicting body measurements from 2D images compared to the baseline model of Mohammed Khan (2020) which uses Restnet50 and Inceptionv3 architectures?

2 RELATED WORK

This sections covers prior research explaining anthropometry, introduces similar studies within the academic community and how body shapes are formulated.

Anthropometry is defined as the science of measurement of the human body, which includes body height, body circumferences, size of body segments etc. (Fialho et al., 2021). The uses of anthropometric measurements are not limited to health purposes, instead the uses range from fashion industry, medical diagnosis, ergonomics production to security systems (Yan & Kämäräinen, 2021). Traditionally manual anthropometric techniques have been used such as tape measure, ruler, protractor directly taken from the patient. Due to curvatures in the human body shape, this may lead to inconsistent measurements. The measurement accuracy is also reliant on the individual performing the measurement. Another disadvantage of manual techniques is the time required. Healthcare increasingly demands faster and accurate measurement process. Furthermore, COVID-19 has increased the need for less physical contact between health professionals and patients (Fialho et al., 2021).

Digital anthropometric instruments (e.g., three-dimensional scanners) are more reliable comparing to manual tools, however they are still reliant on the measurer and takes time to perform. Fialho et al. (2021) introduces NLMeasurer, a smartphone application which automatically evaluates anthropometric measurements with deep learning models. A photo is made on which anatomical reference points (ARPs) are identified and the size of body segments are determined. NLMeasurer utilizes PoseNet which is a computer vision model from TensorFlow, and it is able to identify 17 ARPs. PoseNet uses either MobileNetV1 or Resnet50 architecture. The results only contained MobileNetV1 data as the Resnet50 configuration was not able to identify any ARPs hence no results were produced for Resnet50. This may be attributable due to the processing power limitations of a smartphone. MobileNetV1 requires less computational capacity than ResNet50 and therefore was able to deliver the results. This poses a risk to the CGM as it plans to use smartphone application in a similar manner to run ResNet50 architecture. The

ResNet50 tractability issue, highlights the need for testing the CGM with different architectures and bearing processing requirements in mind (Fialho et al., 2021).

One key limitation of the NLMeasurer is the extremely small sample size of 4 participants. The reason given was that COVID-19 did not allow safe researchers and patients to conduct this research without risk of infections. However, performing this research on a virtual dataset could have circumvented this problem. Datasets with virtual body dataset can be used instead and the details of our dataset are in method section. Another limitation mentioned was the clothing on participants which added some difficulty in identifying the ARPs. Since the virtual body images only has light clothing this problem can be minimized by using a virtual dataset (Fialho et al., 2021).

Kocabey et al. (2017) uses social media profile images to infer the BMI by adapting a two stage prediction process. In the first stage, feature extraction was performed by using VGG-Net and VGG-Face both of which are pre-trained CNNs. Subsequently, a regression model is trained using epsilon support vector regression because of the robust generalization qualities it provides. The performance of the algorithm was compared to a human evaluator. Both the human evaluator and the algorithm were presented with two profile images every instance and were required to predict which profile image was more overweight. Although the algorithm only performed on par with the human evaluator, this research does pave the way for future research to model BMI figures on a population and demographic level by processing profile images.

De Souza et al. (2020), predicts body dimensions from images in a similar fashion but employs a selection of eight different algorithms. The study presents a benchmark of prediction accuracies by testing and presenting MSE scores by algorithm and by body part. For the waist the top performing algorithm was the Expectation-Maximization algorithm. This algorithm calculates the multivariate probability density function parameters with the setup of a Gaussian mixture distribution where the number of mixtures has been determined beforehand (De Souza et al., 2020). From this study, it was interesting to note that the best MAE scores were obtained not necessarily as expected. This is an encouragement for this study to take on new architectures that were previously not attempted.

Yan & Kämäräinen (2021) estimates anthropometric measurements from images by directly regressing body images to body measurements. The method used leaves out the body reshaping phase. Nonetheless, a 3D body mesh is still used by learning a mapping derived from the body dimensions to the shape coefficient of a part-based shape model.

Mohammed Khan (2020) investigates to what extent principal component (PC) values can be predicted from body shape images and explores if viewpoint effects the prediction accuracy. Principal Component Analyses provides valuable insight on an individual's overall shape as it captures multiple dimensions. Hence the PC values are used as training targets and they represent height and the waist circumference which are key for detecting malnutrition. The images are trained with the respective PC values as labels by use of CNN. Viewpoint is found not to have significant effect on prediction performance. The study utilises Resnet50 and Inception V3 models both of which are CNNs. Resnet50 is often utilised as they allow many layers to be used in the CNN architecture. This is enabled by the skip connections between layers and addresses the vanishing gradient issue. Inception V3 varies the filters sizes which enables the model to capture the high detail in images (Khan et al., 2020). Equipped to capture high detail, it has a strong ability for pattern recognition and extracting features which is needed with training images (Mohammed Khan, 2020).

Dhaliwal et al. (2020) predicts facial anthropometric measurements (e.g., height of face, nose width) in order to differentiate features between sub-ethnic groups from 2D images. Deep learning architectures such as VGG16, ResNet50 and InceptionV3 were utilized and InceptionV3 was found to have the leading performance for accuracy. Features extraction was performed by using facial landmark detector which is composed of an ensemble of regression trees. This ensemble of regression trees was pre-trained on 68 facial landmark positions.

In this facial anthropometric measurements study (Dhaliwal et al., 2020), the experiment was run on both the raw dataset and also separately on the normalized dataset. Although the photographs were taken indoors where the lighting was controlled, other variations still exist. For instance, the face size and location can differ. Hence, normalization is essential in obtaining accurate predictions. Min-max normalization was chosen in line with facial health care research practice. Results indicated the normalized dataset had the stronger performance. Since normalization has been described to be an essential task to perform prior to feature extraction, this leads to the motivation of selecting Big Transfer architecture which will be elaborated upon in the methods section.

The process of selecting a deep learning network with its complex nature is of vital importance for improving the performance. Yet in practice, most often architectures are designed by experts in a trial and error manner (Devaguptapu et al., 2021). There are automated solutions such as neural architectures searches, however, these still operate on a trial and error basis to find the optimal architecture. In the method section, the two proposed architectures will be introduced and explained why better

performance is expected versus the baseline. However, given the fact that most architectures are improved by trial and error, to a certain extent this research will also be experimenting with new architectures.

Yang et al. (2014) uses a novel approach to reshape and estimate human pose by a process called Semantic Parametric Reshaping. The pose and shape variations were gathered by utilizing the Shape Completion and Animation for People (SCAPE) method. This was performed on the CAESAR data set and what this data set precisely entails will be explained in the methods section. The SCAPE method estimates and builds the human shapes which is a vital concept of this research worth elaborating on. Anguelov et al. (2005) is the original paper introducing this method by building a unified model of human shapes. This entails that the method separately learns different types of model deformation. One model is dedicated to learning the pose of the human while the other model learns on the differences in body shapes between different humans. Together these models are able to provide full body shapes with adequate level of detail. Even muscle deformation in various poses are embedded in the model.

The pose deformation part of the model is derived from a collection of 3D scans of a single individuals in various poses. Within the pose model, deformation is further sub-divided into rigid and non-rigid components. The rigid component of deformation is represented with a low degree of freedom body shape. While the non-rigid component is responsible for modelling the residual deformation (e.g., flexing of muscles). Another characteristic of this model is that deformation is only dependent on adjacent joints. Due to this restriction, the data is relatively low dimensional and enabling the shape deformation to be learned with relatively smaller datasets (Anguelov et al., 2005).

The SCAPE model also creates variations in shape between patients. This is done by using 3D scans of various people in various poses. The shape variation is modelled using Principal Component Analysis (PCA). One of the useful attributes of PCA is reduction of dimensionality. In this context, the subspace of human shape deformations is low dimensional. In addition, the variations in shape between patients does not confound the other pose and residual deformations. Maintaining isolation between these deformation types is essential for the model to perform (Anguelov et al., 2005).

By isolating the modelling of pose and body shape deformation, this model simplification in turn leads to simplification of the mathematics. This is useful since the algorithms can be learned more efficiently enabling the computations to be run in shorter lead times. On the contrary, some limitations require mentioning. By isolating types of deformation, this also prevents the model capturing trends where body

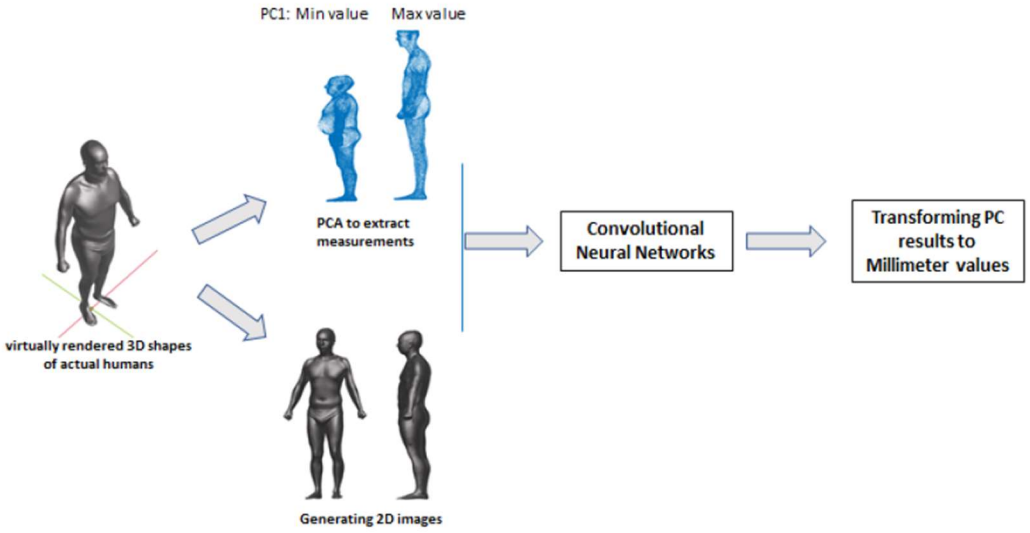
shape and muscle deformation show a strong correlation. For instance, patients who are more muscular usually have more muscle deformation. However, this aspect is not captured by the SCAPE model (Anguelov et al., 2005).

3 METHOD

This section will provide a overview of the overall research process and the PCA technique is explained. The first step of the research process involves virtually rendering 3D shapes from actual humans which is depicted on the very left of figure 1. Afterwards, these 3D human shapes were used separately for two steps. The first step was to generate 2D images from these 3D human shapes which are later used as inputs to the CNNs (lower arrow in figure 1). Secondly, PCA was used on the 3D body shapes to calculate measurements which will be used as labels for the CNN (upper arrow in figure 1). After we have the PC values as our targets and the corresponding 2D images, we can train the CNN with these two inputs. Finally, the PC unit predictions from our CNN need to be transformed into millimetres for comparability with other studies.

Figure 1

Overview of the research process taken from the baseline study of Mohammed Khan (2020)



3.1 PCA

PCA is a technique that has been used in both the body shape formulation and PCA was also separately used to calculate the PC values that served as labels for our training. As this research relies extensively

on the PCA method, it will be explained in this section. Datasets are increasingly large and complex to interpret. PCA is effective in reducing the dimensionality which increases the interpretability while the variability is preserved as much as possible. This is achieved by creating new variables, or principal components, which are uncorrelated to one another. These new variables are defined by the current dataset being used as opposed to determined beforehand, making PCA an adaptive data analysis method. This adaptive exploratory method can be applied to various data types and is suitable for this thesis which is also exploratory in nature (Jolliffe & Cadima, 2016).

The motivation of selecting PCA for this research resides with the ability to reduce dimensionality of complex data sets such as body shapes. Reducing the body shapes into low-dimensional PC values allows the CNNs to efficiently process and predict PC values. The PCA is capable of providing comprehensive insight on a human shapes as it captures data across multiple dimensions. Furthermore, this method provides uncorrelated PC values each representing a distinct body measurement. Enabling this study to select the relevant body parts such as height and waist circumference which are indicators of malnutrition. The fact that PCs are uncorrelated and treated in isolation translates into useful interpretability and can explain different PCs for different purposes for future research as well.

To generate the labels for our data set, PCA was used on the human meshes to calculate the first ten PC values. For instance, the first PC values corresponds to the height, the second PC value is the belly. The S-SCAPE method from Pishchulin et al. (2017) was utilized which is an updated version of SCAPE. The approach is not novel but it improved the SCAPE model by applying bootstrapping and posture normalization. In addition, the 3D human shapes were created based on the largest commercially available datasets. Previously, models were only based on small datasets which had limited shape variations. S-SCAPE has demonstrated improved performance versus SCAPE, in terms of rebuilding shapes and being more robust towards variations in posture caused by movement or tilting of the body.

Besides the PCA, anthropometric measurements also can be predicted by detecting fiducial points on the contours from frontal and lateral body images. Anthropometric measurements can be calculated by finding the difference between the two relevant fiducial points. Circumference measurements are calculated by use of the ellipsoid model (Aslam et al., 2017). This method was not chosen in order to align methodologies with the baseline study which also used PCA. For the baseline comparison to make sense, the same method needs to be implemented. Beyond this restriction, PCA is able to provide comprehensive human shape insight and is efficient to process for the CNNs as mentioned earlier.

4 EXPERIMENTAL SET-UP

This section is dedicated to describing the detail the dataset and the experimental procedure. Including an explanation of the architectures models and how they work. Lastly, the evaluation criterion is discussed.

4.1 Dataset

Our data set contains virtual images of human bodies which act as input to our CNN and also the body shape measurements which are the labels to train with. In this section, the virtual body set will be explained, followed by the body shape labels and finally the image data set.

4.1.1 Virtual body data set

We used a dataset from a secondary source to acquire a virtual body dataset from the Semantic Parametric Reshaping of Human Body Models provided by Yang et al. (2014). Which originates from the CAESAR database which stands for Civilian American and European Surface Anthropometry Resource. The original CAESAR database could have provided actual body measurements and would be more accurate to use these instead of predicted PC values. However, the financial cost to obtain this dataset made this not a feasible option. Hence, we used the body shapes provided by Yang et al. (2014) as our virtual dataset. The virtual dataset consists of 1,517 males meshes and 1,531 females meshes and is a subset of the original dataset. Patients were standing upright with both arms at their sides hanging freely. In the virtual dataset, virtual persons were anonymized as the face is not recognizable and all patients have grey skin. The meshes were not labelled therefore requires the next step in obtaining body shape labels outlined in the next paragraph.

4.1.2 Body Shape Labels

In order to get the PCs, PCA was applied to the 3D body shapes. The PC is the axes with the most variation. In the case of body shapes, the PC value (e.g., height) resembles the measure of variation in shape for that component. The first PC represents height and contains the most variation, the second PC value represents the belly measurement, the third and fourth PC value represent the female and male waist circumference respectively. Matlab R2019 was used read the human shapes into a male and female matrix. PCA was then applied to both matrices in order to obtain the first ten PC values from both matrices. These measurements served as the labels for our training.

4.1.3 Image data set

Blender 2.82 was used to transform 3D virtual objects into 2D images. Within Blender 2.82, a virtual studio was set up and the camera captured images rotating 360 degrees around the patient. This process created 185 images per patient with one image was taken for approximately every 2 degrees. The distance from the person to camera is 1 meter and also one meter from the ground.

In the figure below, images of one female is taken at various viewpoints. With 3K male and female individuals and 185 images per individual, this leads to approximately 564K images in total. This research was performed on a 100 male and 100 female individuals which consists of 37K images in total due to computational constrains of Google Colab. Images were cropped and rescaled to 224 x 224 x 3 for the same computational reason.

Figure 2

Female sample images used for training

4.2 Pre-Processing

The advantage of using CCNs is the need for pre-processing is limited since it can recognize patterns well and extract the relevant features without having to alter the image. Input images were re-shaped to 224 by 224. In addition, the input was also normalized to values between 0 to 1 by diving by 255.

Further pre-processing included ensuring every image is aligned with the correct PC label. Secondly, confirming no data leakage exists between patients as there can be no overlap of images between patients. From the processing, it was found that one male image (“SPRING0001”) was contained in the female images collection. Further checking revealed no other issues with the same problem. Due to the fact that this is an isolated event and one image will not have any significant distortion on our results as one image accounts only 1% of the female dataset where training was performed on. No rectification was performed and there is a preference to not to alter the original dataset.

4.3 Models

The research uses pre-trained CNNs which require labelled data and are supervised learning methods. CNNs are designed specifically for working with images and maintain the spatial and temporal dependencies (Duong et al., 2020). By using a kernel that convolves around the input image, the CNN also minimizes the number of parameters. This section discusses the two models this research is implementing.

4.3.1 Big Transfer

BiT is a newer variant and based on Resnet152. BiT model utilises transfer learning for visual tasks where it pre-trains on very large supervised datasets which forms the starting point for fine tuning for other visual tasks. In addition to scale, BiT replaces batch normalization (BN) with group normalization (GN) and weight standardization (WS). GN divides the channels into groups and normalizes by group. In doing so, the computations of GN are independent of batch size hence the accuracies are more stable when batch size can come in a wide spectrum (Kolesnikov et al., 2019). In practice, small batch sizes are used for tractability. Batch sizes often vary between train and test set and between pre-train and fine-tune etc. The architecture my research is replicating is use Resnet50 and Inception v3 both of which use BN. As both architectures in prior studies have used BN, this research uses BiT to apply GN which should bring about an improvement in performance. Furthermore, as BiT is a newer version based on Resnet, it would be reasonable to expect this updated variant to perform better. Resnet50 is from 2015 while BiT is from 2019.

Qiao et al. (2019) elaborates on GN and finds that GN only outperforms BN for small batch sizes. Normalization is done at every neural network layer by mean centering. The mean is derived from the batch and is only meaningful is the batch is not too small. Although the batch size of 32 as used in the basemodel this is still sufficient, GN adds the possibility of using smaller batch sizes when scaling the size of the network. Larger networks can adopt a smaller batch size for computational reasons. Furthermore, when GN is used in conjunction with WS the results outperform the BN. Just as GN normalizes the data at every layer of the network, WS normalizes the weights at every layer in a similar manner. Earlier in related works section, Dhaliwal et al. (2020) established that normalized dataset performed more accurately than the raw dataset in the facial anthropological study. In this research, we expect to obtain improved prediction results by normalizing with GN and WS at every layer.

4.3.2 EfficientNet

EfficientNet is a CNN architecture with a scaling method that uniformly scales up all three dimensions. Tan et al. (2019) introduces a systematic method of scaling up CNN since there are numerous ways to scale up. Scaling up has not yet been well understood and mostly is done arbitrarily. The paper establishes that different scaling dimensions (width, depth, resolution of CNN) are dependent on each other. For example, if width is increased, accuracy and efficiency are improved if the input resolution is also increased. Compound scaling method is used to balance dimensions when upscaling and scaling with a constant ratio. In general, compound scaling improves accuracy by 2.5% compared to single-dimension scaling methods. Intuitively this concept makes sense because the larger the input image, the more layers and channels are needed to capture the detail of the larger image. The EfficientNet B7 model can attain a state-of-the-art accuracy of 84.3% on ImageNet. At the same time, this CNN is 8.4x smaller and 6.1x faster compared to the best CNNs. This model is faster because there are less parameters to compute.

The compound scaling technique provides the scaling ratios to be applied to the dataset. How can we be confident that these ratios will continue to work for the dataset of this research as well? To answer this question, it should be noted the wide spectrum of application of this model. The application of EfficientNet ranges from detecting COVID-19 from chest X-rays (Chowdhury et al., 2020) to the automatic classification of fruit to enhance productivity of traditional farming (Duong et al., 2020).

For the purposes of this study, EfficientNet is a scaling method and is able to obtain improved accuracy and efficiency (Duong et al., 2020), therefore selected as an architecture to experiment with for this research. This study is exploring architectures with the primary aim to increase prediction accuracy of body measurements, however gains in efficiency can also prove to be valuable. With shorter processing duration, computationally it can be afforded to experiment with more hyper parameter settings. Although fine-tuning is not within the scope of this research, this could be valuable for future research. This also simplifies any future fine-tuning as the number of possible configurations among width, depth and resolution parameters are vastly reduced by adhering to the compound scaling formula which has a track record in terms of prediction results.

4.4 Experiment Procedure

Keras deep learning platform was used and online servers were used in order to accommodate for the large data size. The online platform used for this research was Google Colab Pro which has an Intel Xeon CPU of 2.2GHz with 2 cores. GPU memory and RAM are respectively 16GB and 24.4GB. Google. Resources are not guaranteed and actual specifications may vary when running due to availability and memory limits.

The activation of neurons followed the Relu function and global average pooling was used in the fully connected layers. The output layer is linear function in order to change from classifier to a regression model. Learning rate initially set at 1e-3 and reduced by factor of 0.1 and a batch size of 32. Male and female subjects were trained separately to isolate gender bias.

As this research is replicating the CNNs from Mohammed Khan (2020), the first part of the experiment was to replicate the code from the baseline model. In replicating the code, the majority of the original code was used. At certain places, minor adjustment needed to be made to get the code to work. The original code was not performed on Colab, the dataset is also slightly different and some code was added to split the images into train, validation and test set which was not provided by the original code. I will not elaborate in detail about these minor code adjustments as they will be different for the person replicating the code of this research depending on what platform will be used. The code for this research is made available in appendix B. Lastly, the parameters of the models were kept the same as the original study only the size of the input shape was adjusted to 224 by 224 to fit the images of this research.

In order to use a subset of 100 images for male and female images, the file images had to be selected manually. The labels data frame also needs to be manually changed in excel to ensure only those 100 patients are included. Afterwards, the 100 images are compressed using RAR format and placed on a google drive where Colab can access the images.

Colab uses Python 3.6 has all the packages this research needs installed already. The packages only need to be imported as they are installed already, the following packages were used: Tensorflow Keras, ImageDataGenerator, Sequential, Sklearn, matplotlib, cv2, pandas, numpy, shutil.

In line with the replicated research, mean absolute error (MAE) is used for evaluation purposes and is defined by the following:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

where n is the number of images y*i* is the predicted PC value and xi is the target label.

The PC values are converted into millimetres in order to compare with other studies, by using the below formula obtained from Mohammed Khan (2020) for converting height.

$$Y_{new} = \frac{\sigma_{St}}{\sigma_{PC}} Y_{old} + \mu_{St}$$

- Y_{new} = new prediction value in millimetres
- Y_{old} = old prediction value in millimetres
- σ_{St} = standard deviation of height from CAESAR dataset
- σ_{PC1} = standard deviation of PC1 representing height
- μ_{St} = mean of height in CAESAR dataset

The conversion of PC units to millimeters of waist circumference is given by the below formula (Mohammed Khan, 2020):

$$Y_{new} = \frac{\sigma_{Wc}}{\sigma_{PC}} Y_{old} + \mu_{Wc}$$

- Y_{new} = new prediction value in millimetres
- Y_{old} = old prediction value in millimetres
- σ_{Wc} = standard deviation of waist circumference from CAESAR
- σ_{PC} = standard deviation of PC value which represents waist circumference
- μ_{Wc} = mean of waist circumference in CAESAR dataset

5 RESULTS

In this section, the results from height will be presented first along with a benchmark comparison with other papers. Thereafter, weight results are presented.

5.1 Height Results

Figure 3 Height MAE in mm by CNN Model

	Female MAE in mm		Male MAE in mm		Average MAE in mm	
CNN Model	This Thesis	Baseline Model	This Thesis	Baseline Model	This Thesis	Baseline Model
Resnet	89.7	12.2	58.8	7.8	74.2	10.0
Inception	70.8	11.2	58.7	7.0	64.7	9.1
EfficientNetB0	90.0	n/a	79.2	n/a	84.6	n/a
BiT	71.0	n/a	65.2	n/a	68.1	n/a

Inception performs the best for predicting both male and female height and has an average MAE of 64.7mm (average of both sexes). Also in the baseline study, Inception was also the top performer outperforming ResNet. The second best performer in terms of average MAE is BiT with MAE of 68.1mm. Followed by ResNet at 74.2mm and the worst performance is from EfficientNet at 84.6mm. This is not in line with expectations as EfficientNet was expected to perform better versus the baseline models. BiT outperformed on Resnet by 6.1mm, however performed worse by 3.4mm compared to Inception on average MAE.

From figure 3, the male MAE is performing significantly better compared to the female counterpart in this thesis and this trend is also evident in the baseline model.

Figure 4 Height MAE in mm by Various Papers

Paper	Model	MAE in mm
Dantcheva, Bremond, and Bilinski 2018	Resnet	77-78
Haritosh et al. 2019 ResNet	Resnet	73
Mohammed Khan 2020	Resnet	10.3
Mohammed Khan 2020	Inception	9.1
this thesis	Resnet	74.2
this thesis	Inception	64.7

Earlier from figure 3, the MAE of the baseline model was significantly better than the proposed models on all fronts. However, when adding other studies to this comparison as shown in figure 4, the MAE of this thesis is closer to other studies such as Dantcheva et al. (2018) and Hartiosh et al. (2019). When comparing results across different studies, it needs to be pointed out that these other studies are using facial images to predict height and not using full body images. Although they are also using the Resnet

model, there are many factors that could lead to such differences as they are using facial images instead of fully body images to begin with.

5.2 Waist Results

The waist results were almost the same for all models except some minor variations depicted in figure 5 below. For the average MAE, the two proposed models performed slightly better versus the two baseline models with the differences ranging between 0.8mm to 1.6mm and arguably insignificant. The difference comes from the male results since female results virtually performed equally across all models at 120mm.

Figure 5 Waist MAE in mm by CNN Model

	Female MAE in mm		Male MAE in mm		Average MAE in mm	
CNN Model	This Thesis	Baseline Model	This Thesis	Baseline Model	This Thesis	Baseline Model
Resnet	120.2	75.6	122.9	44.3	121.5	60.0
Inception	120.7	72.9	122.9	44.1	121.8	58.5
EfficientNetB0	120.8	n/a	119.5	n/a	120.2	n/a
BiT	120.9	n/a	120.7	n/a	120.8	n/a

When carrying out this research, it was noticeable that the results were significantly better when using less epochs. The baseline model uses 50 epochs and all the parameters had to kept consistent for the baseline comparison to be meaningful. However, it is worth noting that the models performance deteriorated rapidly as epochs was increased as shown in Figure 7 below. The reasons for this will be treated in the next discussion section.

Figure 6 MAE by Epochs for EfficientNet Females

Epoch	Height
1	73.3
2	77.3
3	82.5
5	93.5
50	84.6

6 DISCUSSION AND RECOMMENDATIONS

The goal of this research is to study how effective the two proposed models are compared to the two baseline models. The results of height are discussed followed by waist circumference and some recommendations for future research.

From our height results, Inception model is the top performing model, followed by BiT, ResNet and EfficientNet in order of ranking performance. The fact that Inception outperformed ResNet is in line with prior research results (Mohammed Khan, 2020; Dhaliwal et al., 2020). However, it was not expected in this research for EfficientNet to be the worst performer. EfficientNet, equipped with compound scaling technique was in theory expected to obtain improved accuracy and efficiency (Tan et al., 2019). One possible explanation could be that introducing scaling and more complexity is not the solution in this context. Instead feature extraction could be more vital as Inception and ResNet are both strong for such qualities.

As mentioned in the methods section, the intention of selecting EfficientNet is to experiment with and find architectures that are efficient and therefore can be deployed on mobile devices. In terms of result rankings EfficientNet does rank last, however, this thesis would argue that this model could still be useful. As outlined in related studies section, other studies have attempted to use ResNet50 (Fialho et al., 2021), and failed to generate any results on mobile devices for Resnet50 due to computational constraints of mobile devices. Therefore, even though ResNet outperforms EfficientNet on prediction accuracy in this research, it could be possible that Resnet will not work for the CGM on mobile phone applications. In that case, EfficientNet could be an option which sacrifices some accuracy for computational efficiency required on mobile devices.

Furthermore, it is important to note that the results explained earlier are based on the exact architecture configuration of the baseline models. EfficientNet did not have the benefit of tuning or adjusting any parameters unlike the baseline models. Proposed models adapted the exact same configuration in order to allow meaningful comparability between the models. However, when performing a some tuning on the EfficientNet, the female MAE was able to improve from 90.0mm to 63.7mm with only a few adjustments. This tuning was only a preliminary analysis by applying dropout function and experimenting with different learning rates and early stopping configurations with Keras tuner. The results of the initial tuning analysis for all models are included in the appendix B. Further research can follow this starting point to study fine

tuning of these models and finding a method to maintaining comparability as different tuning methods will have different impacts on different architectures.

BiT lagged behind Inception, however, BiT still significantly outperformed ResNet for height by 6.1mm MAE. These results indicate that the normalizing with GN and WS at every layer has had a positive impact on prediction accuracy. This exploratory research has found indications that normalizing techniques help improve the prediction accuracy of these models. This does require further testing, however, future research involving selection of CNN models can consider choosing models with normalisation components. In addition, as BiT is a updated version of Resnet, the results indicate that adopting newer versions are beneficial to the prediction accuracy. Note that this is for the height results.

For the waist results, the improvement in performance of the proposed models versus baseline are so small and considered insignificant. The waist results do not produce the confidence to claim an improved model has been found. The reason why the waist results show so little variation between models is not fully understood. One possible reason is for waist a circular circumference has been assumed and reality would be more accurately reflected by an ellipse instead (Mohammed Khan, 2020). Compared to the height results, waist error is also significantly higher for this thesis and also the baseline study. Since the waist error was also relatively larger in prior studies, future work can explore the underlying factors of why waist MAE is structurally and significantly larger.

In the final part of the results, figure 7 depicts the MAE deteriorating rapidly as the number of epochs increases. This is most likely due to overfitting which a common issue with large neural networks. Some measures to counter this overfitting are methods such as adding dropout function, tuning the early stop parameter to cease earlier, and regularization.

De Souza et al. (2020) explored with eight different algorithms to study accuracy on predicting bodily measurements. ResNet50 and Inceptionv3 have been chosen for their known ability to extract features from images useful in the baseline study (Mohammed Khan, 2020). And although the results of these two baseline models have shown improvement against other studies, other algorithms should not be disregarded. Alternative approaches may exist for the same problem and other algorithms may perform better than expected. In the related works section (Devaguptapu et al., 2021) pointed out the complex nature of deep learning and noted in practice architectures are designed in trial and error manner. The same trial and error mind set can be applied in finding different algorithms. Future research could be done by testing other algorithms in the context of the CGM.

Our results also noted that male outperforms females. The reasons for this gender bias is unclear as this research processed male and females separately and underwent the exact same process. Further research could be done to explore why such variation occurs.

6.1 Limitations

The proposed models have been implemented based on the exact configuration of the baseline model and unlike the baseline model, it did not have the benefit of tuning parameters. A little change in the hyper parameter settings can have a sizeable effect on the performance of the model. The configuration was kept completely identical for comparability but this introduces an unlevelled playing field. This was mentioned earlier in detail in the discussion section and it is a limitation of this research. At the same time, the results are still valid as all models have been subjected to the exact same process and parameters. This allows comparability between models, hence in the results and analysis comparisons and rankings have been used.

From comparing the results of the proposed models against the baseline model, results indicate that the proposed model is lagging behind the baseline model results significantly for both male and females. It is worth mentioning that although this research is replicating the baseline model, there are still differences to be accounted for. For instance, the structure of dataset is different to begin with. The baseline model has 100 images for every patient as opposed to 185 images per patient for this research. The total number of images also is different as the baseline study used 66K images versus 37K images in this thesis in total for both males and females. Any changes in the complex landscape in CNNs can lead to a sizeable change in overall performance. The inter-study comparison of heights MAEs should be understood with caution since different studies have used entirely different methods. However, we have included such benchmarks in order to have an approximation of where our results lie in relation with other academic studies.

The difference in images per patient could also be an advantage for the baseline model. Since Mohammed Khan (2020) established that viewpoint does not matter for the prediction performance of body measurements, having more images per patient should in theory not be advantageous and only add to the computational workload. For future studies, it would be interesting to study only several images per patient. Doing this analysis would allow the results to incorporate far greater number of patients within the computational constraints. Increasing the number of patients studied would help alleviate overfitting.

Lastly, our training targets we have used estimations instead of actual measurements. Purchasing the CAESAR dataset with the actual human measurements would be the solution, unfortunately the large cost makes this an unfeasible option.

7 CONCLUSION

This thesis explored different CNN architectures for predicting body measurements from 2D images. In particular, the effectiveness of EfficientNet and Big Transfer architectures compared to the baseline models of ResNet50 and Inceptionv3 were assessed.

The newly proposed BiT architecture outperformed Resnet by 6.1mm at 68.1mm for average MAE for height. However, BiT performed worse by 3.4mm and Inception is the strongest performer at 64.7. Even though BiT is still lagging slightly behind Inception, it has managed to outperform Resnet considerably. In doing so, future research can adopt BiT as one of the models or other models which contain normalization features. This is only for the height and more studies need to be carried out for the weight before reaching any conclusion.

EfficientNet ranked last at 84.6mm for average MAE, however can still be useful for mobile devices where computational constraints play a vital factor. Further research should also take into account the efficiency of models since the CGM is eventually deployed on a smartphone.

REFERENCES

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer: General visual representation learning. arXiv preprint arXiv:1912.11370, 2019.

Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., & Davis, J. (2005). SCAPE: shape completion and animation of people. *ACM Trans. Graph., 24*, 408-416.

Aslam, M., Rajbdad, F., Khattak, S., & Azmat, S. (2017). Automatic measurement of anthropometric dimensions using frontal and lateral silhouettes. *IET Comput. Vis., 11*, 434-447.

Bandsom, K. (2021). A Healthy Diet for Mother and Child. Retrieved 28 December 2021, from https://www.welthungerhilfe.org/our-work/countries/ethiopia/a-healthy-diet-for-mother-and-child/

Child Growth Monitor. (2021). Retrieved 28 December 2021, from https://www.tilburguniversity.edu/research/impact/creating-value-data/zero-hunger-lab/child-growth-monitor

Chowdhury, N.K., Rahman, M.M., Rezoana, N., & Kabir, M.A. (2020). ECOVNet: An Ensemble of Deep Convolutional Neural Networks Based on EfficientNet to Detect COVID-19 From Chest X-rays. *ArXiv, abs/2009.11850*.

Dantcheva, A., Brémond, F., & Bilinski, P.T. (2018). Show me your face and I will tell you your height, weight and body mass index. *2018 24th International Conference on Pattern Recognition (ICPR)*, 3555-3560.

Devaguptapu, C., Agarwal, D., Mittal, G., Gopalani, P., & Balasubramanian, V.N. (2021). On Adversarial Robustness: A Neural Architecture Search perspective. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 152-161.

Duong, L.T., Nguyen, P.T., Sipio, C.D., & Ruscio, D.D. (2020). Automated fruit recognition using EfficientNet and MixNet. *Comput. Electron. Agric., 171*, 105326.

Dhaliwal, J., Wagner, J., Leong, S.L., & Lim, C.H. (2020). Facial Anthropometric Measurements and Photographs — An Interdisciplinary Study. *IEEE Access, 8*, 181998-182013.

Fialho, R., Moreira, R., Santos, T.C., Vasconcelos, S.S., Teixeira, S., Silva, F., Rodrigues, J.J., & Teles, A.S. (2021). Can computer vision be used for anthropometry? A feasibility study of a smart mobile application. *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 119-124.

Fu, Y. (2020, June 30). *Image Classification via fine-tuning with EfficientNet*. Keras. Retrieved November 9, 2021, from https://keras.io/examples/vision/image_classification_efficientnet_fine_tuning/.

Haritosh, A., Gupta, A., Chahal, E.S., Misra, A., & Chandra, S. (2019). A novel method to estimate Height, Weight and Body Mass Index from face images. *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 1-6.

J. W. M. de Souza *et al*., "Predicting body measures from 2D images using Convolutional Neural Networks," *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1-6, doi: 10.1109/IJCNN48605.2020.9207330.

Jolliffe, I.T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*.

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, *53*(8), 5455-5516.

Kocabey, E., Çamurcu, M., Ofli, F., Aytar, Y., Marín, J., Torralba, A., & Weber, I. (2017). Face-to-BMI: Using Computer Vision to Infer Body Mass Index on Social Media. *ArXiv, abs/1703.03156*.

Li, S., Nguyen, V.H., Ma, M., Jin, C., Do, T.D., & Kim, H. (2015). A simplified nonlinear regression method for human height estimation in video surveillance. *EURASIP Journal on Image and Video Processing, 2015*, 1-9.

MohammedKhan, H. H. (2020). Predicting Human Body Dimensions from Single Images (thesis). Tilburg University.

Pishchulin, L., Wuhrer, S., Helten, T., Theobalt, C., & Schiele, B. (2017). Building statistical shape spaces for 3D human modeling. *Pattern Recognit., 67*, 276-286.

Qiao, S., Wang, H., Liu, C., Shen, W., & Yuille, A. (2019). Micro-batch training with batch-channel normalization and weight standardization. *arXiv preprint arXiv:1903.10520*.

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable Deep Learning Models in Medical Image Analysis. Journal of Imaging, 6(6), 52. MDPI AG. Retrieved from http://dx.doi.org/10.3390/jimaging6060052

Tan, M. & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning, in Proceedings of Machine Learning Research 97:6105-6114.

Varol, G., Ceylan, D., Russell, B.C., Yang, J., Yumer, E., Laptev, I., & Schmid, C. (2018). BodyNet: Volumetric Inference of 3D Human Body Shapes. *ECCV*.

WHO Child Growth Standards. (2021). Retrieved 5 December 2021, from https://www.who.int/toolkits/child-growth-standards

World Food Programme. (2020). The State of Food Security and Nutrition in the World 2020. Transforming food systems for affordable healthy diets. Rome; Food and Agriculture Organization of the United Nations.

Wu, Y., & He, K. (2018). Group normalization. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3-19).

Xiaohui, T., Xiaoyu, P., Liwen, L., & Qing, X. (2018). Automatic human body feature extraction and personal size measurement. *J. Vis. Lang. Comput., 47*, 9-18.

Yan, S., & Kämäräinen, J. (2021). Learning Anthropometry from Rendered Humans. *ArXiv, abs/2101.02515*.

Yang, Y., Yu, Y., Zhou, Y., Du, S., Davis, J., & Yang, R. (2014). Semantic Parametric Reshaping of Human Body Models. *2014 2nd International Conference on 3D Vision, 2*, 41-48.

APPENDIX A

Preliminary analysis of tuned vs non-tuned models

Female Height MAE in mm		
CNN Model	Tuned	Non Tuned
Resnet	61.9	89.7
Inception	71.2	70.8
EfficientNetB0	63.7	90.0
BiT	65.2	71.0

Female Waist MAE in mm		
CNN Model	Tuned	Non Tuned
Resnet	94.884	120.191
Inception	106.100	120.651
EfficientNetB0	107.25673	120.846
BiT	118.152	120.863

APPENDIX B

Below are the links to Google Colab links containing the code used in this research. For every model male and female have separate code.

ResNet CNN Male:

https://colab.research.google.com/drive/12yz-qmsc4ixN6MK4bXHWlDo5FuY-cH9h?usp=sharing

Inception CNN Male:

https://colab.research.google.com/drive/1ab2DU14mXKJv2td7BV2sE2EKoYZO4OnX?usp=sharing

EfficientNet CNN Male:

https://colab.research.google.com/drive/1NomPKjZRg0ZFwuekDsnZM-e_4RspMGmQ?usp=sharing

Big Transfer CNN Male:

https://colab.research.google.com/drive/1Gja5YkXJYMDSA4cEr54PAa5gZPEHOgjq?usp=sharing

ResNet CNN Female:

https://colab.research.google.com/drive/1QMmXA0QataAliuJvuIvzlVjmJ-_AZdjj?usp=sharing

Inception CNN Female:

https://colab.research.google.com/drive/1q8GpslKnF_jr8C3PWhxJRoFzOnJl_Z1A?usp=sharing

EfficientNet CNN Female:

https://colab.research.google.com/drive/1Ea_HWt49ipL1BdOpAMNncEnMg6yvRfbK?usp=sharing

Big Transfer CNN Female:

https://colab.research.google.com/drive/1sDr2JYybJ0JISzuJ5nzbJlZVHxFhJuIE?usp=sharing

Example of Tuned EfficientNet CNN Female:

https://colab.research.google.com/drive/1ufHsZ27PH1cmjVtTts4e4iLa7PF1S_FT?usp=sharing