Abstract

Body Mass Index is an important indicator of a person's health. Traditionally, BMI is calculated by measuring the height and weight of a person, which takes manual effort. Deep learning neural networks can help automate this task by predicting a person's BMI with one's facial features. Previous work had used many approaches including computer vision, deep learning, and computational. Since the dataset I have is very small, I adopted the transfer learning approach. I used three pre-trained models including VGG16, ResNet-50, and SENet-50, and finetuned them on the VisualBMI dataset. The ResNet-50 model showed the best performance with an MAE score of 7.5. The model was deployed in Streamlit so that users can upload images or use a webcam for BMI prediction.

1. Introduction

The body mass index (BMI) is an essential indicator of a person's health status. BMI is calculated using a person's weight in kilograms divided by the square of the person's height in meters. The National Institute of Health uses BMI as a metric to evaluate whether a person is underweight, normal weight, overweight, or obese. Generally, a person with BMI lower than 18.5 is considered underweight; A person with BMI between 18.5 and 24.9 is considered normal weight; A person with BMI between 25 and 29.9 is considered overweight; A person with BMI greater than 30 is considered obese (Weir & Jan, 2023). As for BMI for children, if a child's BMI is above the 95th percentile then the child is considered obese. According to the Centers for Disease Control and Prevention, the obesity prevalence in the U.S. is approximately 42%, which is almost half of the population. Obesity can significantly increase the risk of health issues including high blood pressure, high cholesterol, type 2 diabetes, breathing problems, joint problems, gallstones disease, and gallbladder disease. Obesity-related absenteeism also negatively impacts the nationwide productivity, costing the U.S. between 3.38 billion to 6.38 billion dollars every year (Trogdon et al., 2008).

A person's face can imply information regarding a person's BMI. Normally, people with high BMI tend to have a larger face which has more body fat while people with low BMI tend to have a smaller face which has low body fat. Therefore, we can use people's face images to predict a person's BMI, leveraging advanced machine learning techniques. Since BMI is an important metric to evaluate a person's health, researchers have been studying various deep learning techniques and testing on a variety of image datasets to evaluate the model performance. This technology can help people easily keep track of their health without knowing the person's weight and height. It can play a significant role in measuring, monitoring, and even improving people's health status in this era of the obesity epidemic. It can further help governments revise policies to improve the overall health of U.S. residents.

2. Literature Review Summary

Kocabey et al. used a computer vision approach to predict a person's BMI with face images from social media. Since the image dataset they used was very limited, they adopted transfer learning and build upon pre-trained models. They first used VGG-Net and VGG-Face to conduct deep feature extraction. They then used epsilon support vector regression models for BMI regression. Based on the Pearson correlation between the test set and the predicted BMI value, they found that VGG-Face performed significantly better than the VGG-Net in terms of feature extraction (Kocabey et al., 2017).

Sidhpura et al. used a deep learning based approach to predict a person's BMI with face images. Similar to Kocabey et al., they also adopted a transfer learning approach. Their pre-trained models included VGG-Face, Inception-v3, VGG19, and Xception. They first cropped and aligned faces to the center and then resized the images for transfer learning. They used 3 publicly available image datasets of prisoners and celebrities. Based on the MAE score, they found that VGG-Face performed the best on the prisoner images while Inception-v3 performed the best on the celebrity images (Sidhpura et al., 2022).

Wen and Guo used a computational approach to predict a person's BMI with face images. They extended the step of image preprocessing from face detection and face normalization to advanced facial feature extraction using Active Shape Model (ASM). These facial features included cheekbone-to-jaw width ratio, cheekbone width to upper facial height ratio, perimeter to area ratio, average eye size, lower face to face height ratio, face width to lower face height ratio, and average eyebrow height. They then used support vector regression, the Gaussian process, and least squares estimation to conduct BMI regression. By comparing the three models in terms of mean absolute errors, they found that the support vector regression method performed the best overall (Wen & Guo, 2013).

3. Methodology

3.1 Data Collection

The VisualBMI dataset contains 4206 images. Each image is labeled by BMI, gender, and a training indicator which specifies whether the image is in the training set or not. Each image is in the BMP format with brightness, contrast, resolution, and shape variations. All the images are front-facing portraits. The image dataset I obtained only consists of 3962 images out of the 4206 images. I matched the data so that each image in my dataset has the corresponding labels including BMI, gender, and training indicator from the original VisualBMI dataset. I then split the data using the training indicator and obtained 3210 images for training and 752 images for testing. Then I created a training and a testing folder and placed the images in their corresponding folder. In this way, the training and testing images were ready to be preprocessed for modeling purposes.

3.2 Data Preprocessing

Since each image in the dataset has a different size, it is important to resize each image to a consistent size before feeding it into the model. Moreover, the pre-trained models I used are VGG16, ResNet-50, and SENet-50 contained in the VGGFace package on Github, which accepts images of size 224 x 224 pixels. Therefore, I resized every image to the size of 224 x 224 x 3 for transfer learning. This ensured that all images had the same input dimensions and were ready to be trained by the model. The two graphs below serve as an example of transforming the image from its original size of 176 x 164 x 3 (Figure 1) to the size of 224 x 224 x 3 (Figure 2).

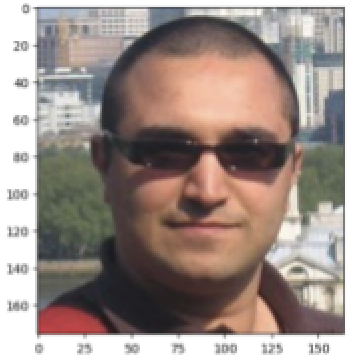


Figure 1

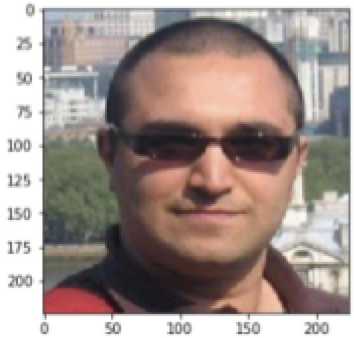


Figure 2

3.3 Transfer Learning

Due to the complexity of calculating BMI from facial images, it is impractical to learn all the necessary features from limited datasets. To overcome this challenge, transfer learning has been widely employed in computer vision tasks to enhance performance and minimize training time. Since the dataset I used for this paper only contains 3962 images, I applied transfer learning on the following cutting-edge pre-trained models:

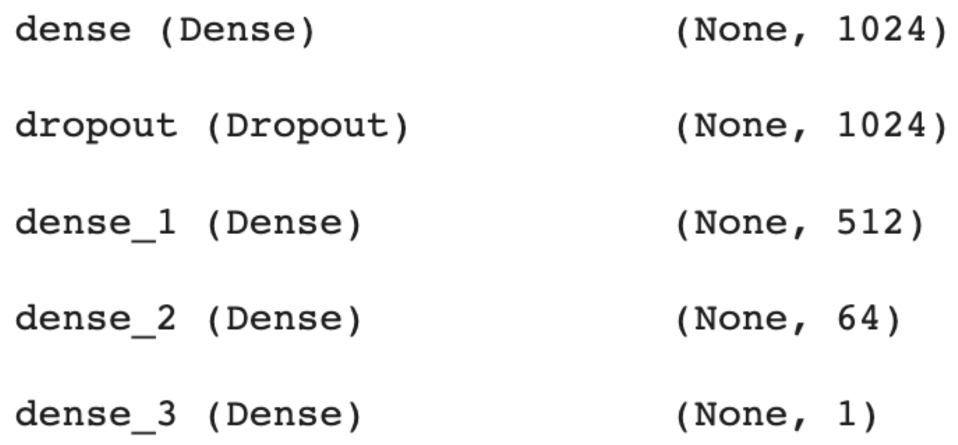
- VGG16: Visual Geometry Group 16 is a convolutional neural network (CNN) architecture developed by the Visual Geometry Group at the University of Oxford. It is widely recognized for its effectiveness in image classification tasks. The "16" in its name refers to the total number of layers in the network, which includes 13 convolutional layers and 3 fully connected layers (Simonyan & Zisserman, 2014)

- ResNet-50: The ResNet-50 model is a deep learning architecture with 50 layers that belongs to the ResNet (Residual Network) family. ResNet-50 addresses the challenge of training deep neural networks by introducing the concept of residual connections and enables the training of very deep neural networks (He et al., 2015).
- SENet-50: The SENet-50 (Squeeze-and-Excitation Network) model is a variant of the ResNet-50 architecture. By integrating the squeeze-and-excitation blocks into the ResNet-50 architecture, SENet-50 enhances the model's ability to adaptively focus on informative features and suppress irrelevant ones. The attention mechanism enables the network to allocate its resources more effectively and improve performance in various computer vision tasks (Hu et al., 2017).

In contrast to traditional methods which trained models using the ImageNet dataset, the pre-trained models I used were trained on the VGGFace dataset, which contains 2.6 million face images of 2,622 people, capturing various poses, expressions, and lighting conditions (Parkhi et al., 2015). While the ImageNet dataset with more than 14 million annotated images is significantly larger than the VGGFace dataset, the ImageNet covering a wide range of object categories beyond faces is commonly used for general object recognition tasks. On the other hand, the VGGFace dataset with detailed annotations and labels specific to faces is specifically curated for face recognition tasks. Therefore, I chose to use pre-trained models on the VGGFace dataset. I believe this will lead to better performance in satisfying the face detection goals of this study.

3.4 Training

The pre-trained models served as feature extractors by removing the last fully connected layers responsible for task-specific predictions. In terms of transfer learning, I added several fully connected layers at the end of pre-trained models. I used the same fully connected layers for the three pre-trained models. The layers I added at the end are shown in Figure 3. The first layer has 1024 neurons. I then added one dropout layer with a dropout rate of 0.5 to prevent overfitting. The third layer has 512 neurons. The fourth layer has 64 neurons. For these 4 layers, I used Rectified Linear Unit (Relu) as the activation function. The last layer has only 1 neuron with a linear activation function to conduct the regression task. I used the Adam optimizer and Mean Absolute Error (MAE) as the loss function.



```
dense (Dense)                (None, 1024)

dropout (Dropout)            (None, 1024)

dense_1 (Dense)              (None, 512)

dense_2 (Dense)              (None, 64)

dense_3 (Dense)              (None, 1)
```

(Figure 3)

3.5 Cross-Validation

Since the data has a training indicator label, I split the data into training and testing images, which was discussed in the data collection section. The training dataset contains 3210 images. The testing dataset contains 752 images. During the training, I set the validation split to

0.2 in order to separate a portion of the training data into validation data to evaluate the model performance on the validation dataset at each epoch.

4. Results and Discussion

4.1 Evaluation Metrics

The evaluation metrics used in this paper are mean absolute error (MAE) and root mean square error (RMSE). The MAE function is defined as below

$$MAE = (\frac{1}{n}) \sum_{i=1}^{n} \left| y_i - \tilde{y}_i \right| \tag{1}$$

where n is the total number of samples, y_i is the actual BMI, and $\tilde{y}_i$ is the predicted BMI. The RMSE function is defined as below

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2} \tag{2}$$

where n is the total number of samples, y_i is the actual BMI, and $\tilde{y}_i$ is the predicted BMI.

4.2 Results

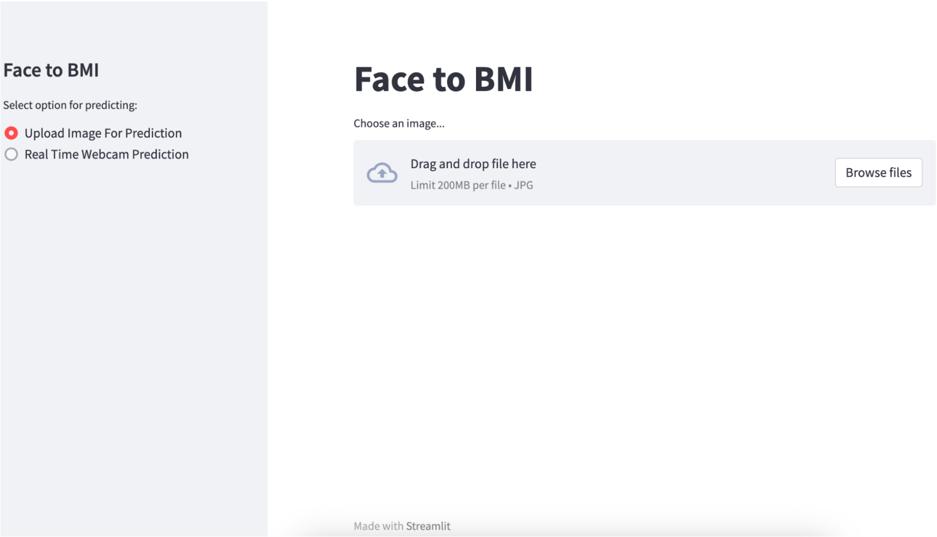
Figure 4 shows a comparison of model performance in terms of the above-mentioned metrics. The ResNet-50 model performs the best with the lowest scores in both MAE and RMSE metrics. Therefore, I choose the ResNet-50 model as my final model and deploy it via Streamlit.

Model	MAE	RMSE
VGG16	8.2618	11.143
ResNet-50	7.5022	10.0137

SENet-50	8.3787	11.6940

(Figure 4)

Figure 5 shows the user interface. The user will have two options for BMI prediction. The first option is uploading an image. An example of image prediction is shown in Figure 6. The second option is using a webcam to conduct real-time BMI prediction. An example of real-time prediction is shown in Figure 7. For either option, users will have a rectangle around their face indicating that a face is successfully detected by the app and receive a predicted BMI value.



Face to BMI

Select option for predicting:

● Upload Image For Prediction

○ Real Time Webcam Prediction

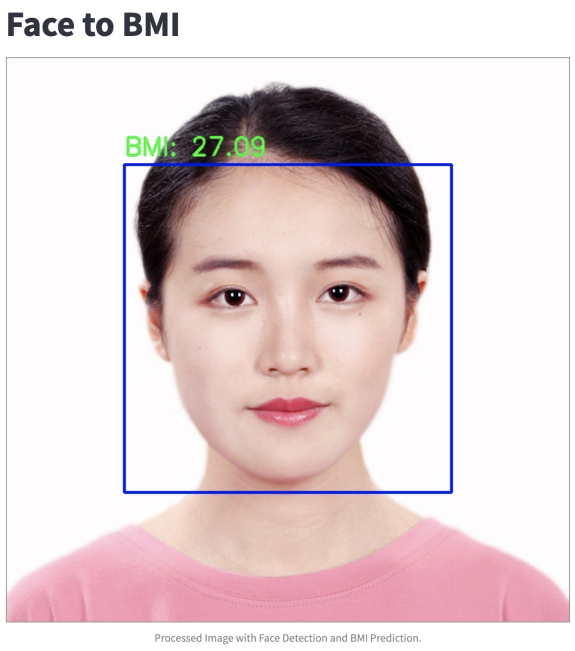
Face to BMI

Choose an image...

Drag and drop file here

Limit 200MB per file • JPG

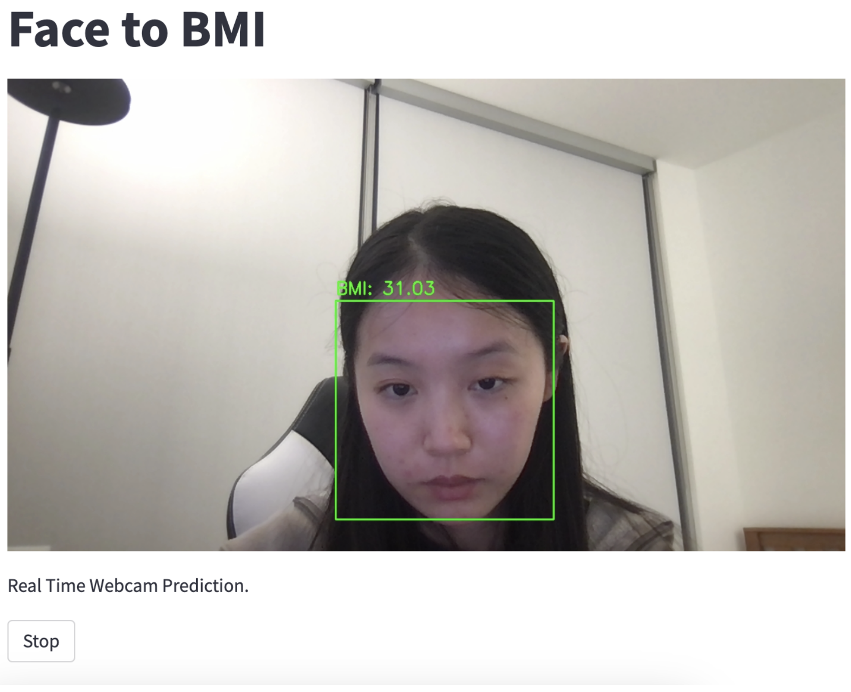
(Figure 5)



Face to BMI

Processed Image with Face Detection and BMI Prediction.

(Figure 6)



Face to BMI

Real Time Webcam Prediction.

(Figure 7)

4.3 Limitation and Future Work

The pre-trained models were trained using the VGGFace dataset. The identities in the VGGFace dataset are mostly celebrities, actors, and actresses living in the U.S. and the U.K. This will introduce bias to the model because the model was only trained on images of Western people and is completely not exposed to people with very different facial features from other parts of the world such as Asia.

Moreover, the dataset I used in this study only contains 3962 images, which is very small and limited for a face detection and BMI prediction task. I believe if the pre-trained models are finetuned on a larger dataset with people from different countries and of different ages, the models will yield better results. Future work can extend this study by experimenting the models with larger training datasets or trying to unfreeze some pre-trained layers.

Another way to enlarge the training dataset is to conduct image augmentation. Image augmentation is a technique to artificially increase the size and diversity of a training dataset by applying various transformations and modifications to the original images. Some common modifications include flipping, rotation, distorting resolution, and changing contrast and brightness of the images. These transformations can help improve the generalization and robustness of the pre-trained models by introducing variations in the training data, enhancing their capability of handling real-world scenarios.

Lastly, future work can also extend this study by experimenting with more advanced models such as Xception. Xception has 71 layers and reaches a Top-1 accuracy of 79% and a Top-5 accuracy of 94.5%, which are significantly higher than the pre-trained models used in this study.

5. Conclusion

In this study, I applied transfer learning to three state-of-art pre-trained models (VGG16, ResNet-50, and SENet-50) to develop a model using face images to predict BMI. The dataset I used is from the VisualBMI dataset which contains approximately 4000 images. The best model reached an MAE score of 7.5. I have also developed an API based on the best model which enables users to upload images or use a webcam for BMI prediction. The study displays great potential for future improvement. In future work, I will apply image augmentation techniques and train the model on larger and more comprehensive datasets with more advanced deep learning pre-trained models.

6. References

Weir CB, Jan A. BMI classification percentile and cut off points. https://www.ncbi.nlm.nih.gov/books/NBK541070/

Trogdon, J. G., Finkelstein, E. A., Hylands, T., Dellea, P. S., & Kamal-Bahl, S. J. (2008). Indirect costs of obesity: A review of the current literature. *Obesity Reviews*, *9*(5), 489–500. https://doi.org/10.1111/j.1467-789x.2008.00472.x

Kocabey, E., Camurcu, M., Ofli, F., Aytar, Y., Marin, J., Torralba, A., & Weber, I. (2017). Face-to-BMI: Using computer vision to infer body mass index on social media. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 572–575. https://doi.org/10.1609/icwsm.v11i1.14923

Sidhpura, J., Veerkhare, R., Shah, P., & Dholay, S. (2022). Face to BMI: A Deep Learning Based Approach for computing BMI from face. *2022 International Conference on*

Innovative Trends in Information Technology (ICITIIT). https://doi.org/10.1109/icitiit54346.2022.9744191

Wen, L., & Guo, G. (2013). A computational approach to body mass index prediction from face images. *Image and Vision Computing*, *31*(5), 392–400. https://doi.org/10.1016/j.imavis.2013.03.001

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv. /abs/1409.1556

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. ArXiv. /abs/1512.03385

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2017). Squeeze-and-Excitation Networks. ArXiv. /abs/1709.01507

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. *Proceedings of the British Machine Vision Conference 2015*. https://doi.org/10.5244/c.29.41