Predicting Body Mass Index From Facial Pictures Using ResNet 18

By

Anusha Bhat

Beechui Koo

Jiayi Li

Yichuan Li

Zoey Hu

Machine Learning II Final Project

University of Chicago

Masters of Science in Applied Data Science

Division of Physical Sciences

May 2025

Abstract

This study investigates the feasibility of predicting Body Mass Index (BMI) from facial images using deep learning models, with a focus on enabling accessible, non-invasive health assessments. Leveraging a dataset of 3,963 annotated face images, we compare five prominent architectures—ResNet-18, FaceNet, ViT B-16, EfficientNet, and DenseNet-121—using both feature extraction and fine-tuning strategies. ResNet-18, adapted for regression, achieved the highest Pearson correlation ($r = 0.660$) between predicted and true BMI, outperforming more complex models. Our pipeline includes standardized image preprocessing, robust model training with early stopping and learning rate scheduling, and deployment via a mobile-friendly application. Despite challenges such as imbalanced data, facial variability, and limited resolution, our results show that facial features can serve as a meaningful proxy for BMI, particularly in settings where traditional measurements are impractical. Future improvements include incorporating full-body images, collecting more diverse demographic data, and leveraging video-based inputs for richer feature capture. This work highlights the potential for deep learning to support scalable, remote health monitoring.

Table of Contents

Introduction

Obesity and overweight rates have risen sharply worldwide, which drive an alarming increase in metabolic disorders, cardiovascular disease, and other chronic conditions. Despite this surge, many clinical settings lack efficient, non-invasive tools for early risk screening, especially where in-person assessments may be limited. Our project addresses this gap by exploring whether facial image analysis can serve as a reliable surrogate for traditional Body Measurement Index (BMI) measurement, enabling rapid risk assessment wherever a simple photograph is available.

BMI remains one of the most widely adopted metrics for gauging adiposity and long-term health risk. Its simplicity and ease of calculation make it indispensable for routine check-ups, yet traditional BMI assessment still requires in-person measurements and can miss subtle early indicators.

By applying deep-learning–derived facial feature extraction alongside regression, this project aims to estimate BMI directly and efficiently from standard portrait images, rather than estimating from a person's height and weight. Such a tool could seamlessly integrate into telehealth platforms, facilitate remote monitoring, and empower clinicians to identify at-risk individuals more quickly and unobtrusively.

Literature Review

Face-to-BMI Paper

Previously, research on predicting BMI from image data was initially presented in the paper "*Face-to-BMI: Using Computer Vision to Infer Body Mass Index on Social Media*" by Enes Kocabey et. al.[1] In this paper, researchers utilized VGG-Net and VGG-Face models to predict BMI from a dataset of images and self-reported BMIs collected from Reddit. The VGG-Face model outperformed the VGG-Net model on test data, achieving a pearson r correlation of 0.65 for the overall test data, 0.71 for the male set, and 0.57 for the female set. This project aims to develop a model that can outperform the original paper's VGG-Face model by achieving a pearson r correlation greater than 0.65 on the test set.

Feature Extraction v.s. Fine-Tuning

Feature extraction and fine-tuning are two widely used transfer learning strategies for adapting pretrained models to new tasks, particularly in image analysis. Feature extraction involves freezing the pre-trained model's weights and training only a new task-specific head, making it computationally efficient and well-suited for situations with limited data.[2] However, since the internal representations remain tailored to the original task, they may not fully capture the relevant features needed for the new domain. In contrast, fine-tuning updates some or all of the pretrained layers, allowing the model to better adapt to the specific characteristics of the new dataset and often resulting in improved performance.[3] This approach, however, demands more computational resources and careful tuning to avoid overfitting. Overall, feature extraction offers simplicity and speed, while fine-tuning provides greater flexibility and task-specific accuracy.

FaceNet Model

Yousaf et al. (2021) propose a semantic-segmentation-guided region-aware pooling (Reg-GAP) that extracts deep embeddings from FaceNet (InceptionResNetV1-VGGFace2) per facial subregion and uses ε-SVR to predict BMI.[4] This model leverages anatomical zones (eyes, cheeks, etc.) to capture localized adiposity cues, improves Pearson's R from ~0.65 to ~0.75 and reduces MAE by up to 63% across benchmark datasets, and model-agnostic design works with any deep embedding extractor. However, the model depends on accurate segmentation masks, which can fail under occlusion or extreme poses, it adds segmentation pre-processing overhead and complexity and its performance is sensitive to lighting variations and low-resolution inputs.

DenseNet-121 Model

DenseNet121 leverages dense connections to improve gradient flow and parameter efficiency, making it well-suited for image-based tasks like BMI prediction. The architecture was enhanced with Squeeze-and-Excitation (SE) blocks for channel attention and a regression head featuring batch normalization and dropout (0.5 to 0.3) to reduce overfitting. The regression head with a frozen backbone was trained first, then deeper layers (denseblock3/4) using differential learning rates and the OneCycleLR scheduler were fine-tuned.[5] A key limitation of DenseNet121 is its relatively high memory usage during training due to the dense connectivity structure, which can be a constraint on lower-end hardware.

EfficientNet Model

EfficientNet, introduced by Tan and Le (2019), is a CNN architecture that uses compound scaling to jointly adjust depth, width, and resolution, enabling high accuracy with fewer parameters.[6] This makes it well-suited for BMI prediction from facial images, as it can efficiently

capture subtle visual cues while maintaining strong generalization. However, its performance is sensitive to training data quality and diversity; models trained on demographically limited datasets may show bias.[7] Additionally, its complex structure requires careful tuning to avoid overfitting on small datasets. Despite these challenges, EfficientNet remains a strong candidate for facial BMI estimation due to its balance of efficiency and precision.

Vision Transformer Model

A Vision Transformer (ViT) is a deep learning model that processes images as sequences of patches, enabling it to capture global context better than CNNs.[8] For BMI prediction, ViTs can learn complex visual cues like facial adiposity or body shape.[9] They perform well with large datasets and are efficient when pretrained and fine-tuned. However, they require substantial labeled data and computing power.[10] In this project, a ViT B-16 pretrained on ImageNet was fine-tuned by unfreezing the final block to better capture task-specific patterns.[3]

ResNet-18 Model

ResNet18, a shallower variant of the residual network family introduced by He et al. (2016)[11], is well-regarded for its efficient training capabilities, enabled by identity-based skip connections that address the vanishing gradient problem. With only 11.2 million parameters, it offers a lightweight solution suitable for low-latency applications and resource-constrained environments, such as mobile inference or small-scale biomedical datasets.[12] However, its limited depth may constrain its ability to capture complex, high-level features, making it less competitive on tasks requiring deeper abstraction.

Data

Dataset Description

To construct a high performing model, the same BMI dataset collected by the researchers for the Face-to-BMI paper was utilized. This dataset contains a csv file as well as a folder of face pictures. The csv file contains 4,206 entries of photo names with metadata related to BMI value and gender. BMI ranges from approximately 17.716 to 85.986, with approximately 58% of records being male and 42% female. Figure 1 displays sample images from the dataset.

Fig. 1: *Sample portraits from the BMI dataset. Photos are of varying quality.*

There are only 3,963 valid images in the dataset, despite the 4,206 entries in the csv. For model construction, we only retained the 3,963 records that had both a valid picture and an entry in the csv.

Exploratory Data Analysis

Exploratory data analysis was conducted to assess the quality of the images and metadata, informing the choice of data augmentation methods. In Figure 2, we observe the distribution of BMI by gender. The median BMI for females is approximately 31.6, while the median for males is slightly lower at approximately 31.156. The distribution of BMI for both genders is

right-skewed with a unimodal peak between 25-35 BMI. This suggests that scaling of BMI may potentially be necessary when constructing the models. Additionally, the models could potentially use gender as a predictor.

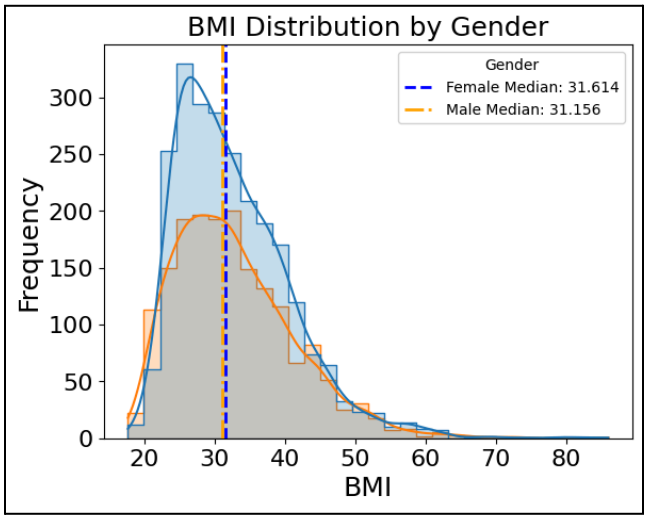

Fig. 2: *Distribution of BMI split by gender.*

In Figure 3, image resolutions are displayed. A majority of images have low resolutions, with widths and heights mainly less than 400 pixels. Additionally, there appears to be numerous images with rectangle dimensions rather than square. This insight reveals the need for image resizing when constructing the model.

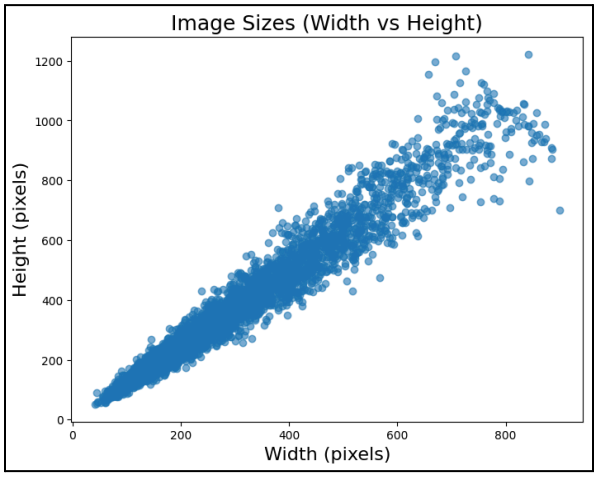

Fig. 3: *Scatterplot of image resolutions.*

Finally, RGB channels were examined across images. A majority of pixels are low intensity, with values less than 100. The green and blue channels have right-skews, whereas the red channel has slightly more uniformity. This suggests that pixel values must be normalized during the data augmentation process.

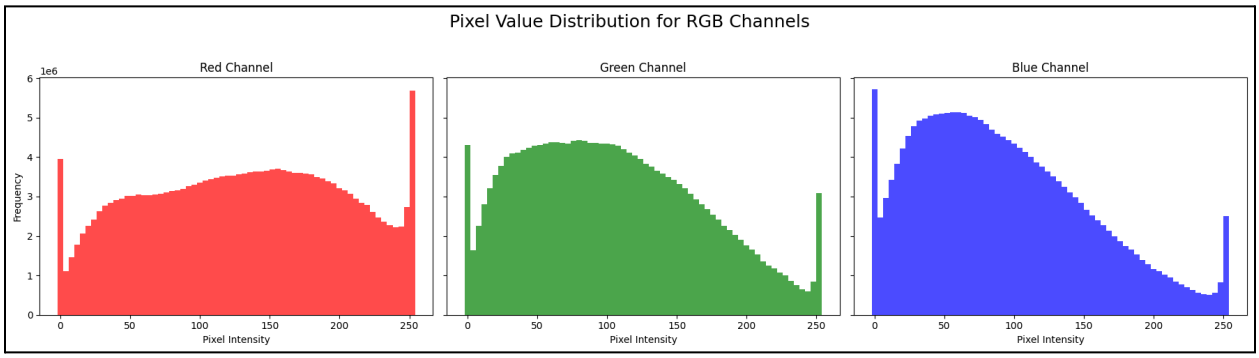

Fig. 4: *Distribution of pixel intensities for the Red, Green, Blue channels.*

Methodology

Model Architecture

The project employed a convolutional neural network architecture based on ResNet18 to predict BMI from facial photographs. Table 1 lists each major component, its corresponding output tensor shape, and the number of learnable parameters. The architecture includes four sequential residual layers (each containing two Basic Blocks), followed by a global average pooling layer and a final linear layer adapted for regression. ResNet18 was adapted for regression by replacing its final fully connected classification layer with a single-node linear output layer. ResNet18 was chosen for its balance between computational efficiency and expressive power. The residual connections in the architecture also supported more stable training by reducing gradient vanishing.

Layer Type	Output Shape	Parameters
Conv2d	[-1, 64, 112, 112]	9,408
BatchNorm2d + ReLU	[-1, 64, 112, 112]	128
MaxPool2d	[-1, 64, 56, 56]	0
Layer 1	[-1, 64, 56, 56]	147,456
Layer 2	[-1, 128, 28, 28]	525,312
Layer 3	[-1, 256, 14, 14]	2,099,712
Layer 4	[-1, 512, 7, 7]	8,393,728
AdaptiveAvgPool2d	[-1, 512, 1, 1]	0
Linear (FC)	[-1, 1]	513

Table 1: Layer-wise summary of the ResNet18 architecture used for BMI prediction.

Image Processing

All images were processed using a consistent pipeline to ensure input standardization. Cropped face images were resized to 224×224 pixels to match the input dimensions expected by ResNet18. After conversion to RGB format, images were normalized using the mean ([0.485, 0.456, 0.406]) and standard deviation values ([0.229, 0.224, 0.225]) of the ImageNet dataset documented on official PyTorch tutorials.

Training Procedure

The dataset consisted of a total of 3,958 facial images labeled with corresponding BMI values. It was split into a training set and a testing set with a ratio of 3210 to 748. The model was trained to minimize Mean Squared Error (MSE) using the Adam optimizer with an initial learning rate of 1×10^{-4}, and weight decay of 1×10^{-5}. The training was conducted for 10 epochs, and to prevent overfitting and improve convergence, both early stopping and learning rate scheduling strategies were employed based on the validation loss. The training pipeline was implemented to support efficient GPU usage and ensure reproducibility.

Deployment Architecture

To demonstrate practical utility, the trained model was deployed in a full-stack application that combined a Flask-based backend with a mobile front-end developed using Expo Go (React Native). The mobile application allowed users to capture or upload a photo, which was then transmitted to the Flask server as a base64-encoded string. Upon receipt, the server decoded and preprocessed the image, following the same steps used during model training. The processed image was passed through the ResNet18 model to generate a BMI prediction, which was then returned to the mobile interface and displayed in real time. This cross-platform deployment strategy ensured compatibility with both Android and iOS devices and enabled lightweight, responsive inference.

Results

Evaluation Metrics

Model name	Fine-tuned/feature extraction	Regression head	Pearson r (overall)
ResNet-18	Fine-tuned	Linear	0.660
FaceNet (InceptionResnetV1)	Pretrained embedding extraction	ε-SVR	0.637
ViT B-16	Fine-tuned last transformer block	Linear	0.605
EfficientNet	Fine-tuned last 2 transformer blocks	Manually Designed MLP	0.561
DenseNet-121	Fine-tuned denseblock3 & 4	Manually Designed MLP	0.542

Table 2: *Comprehensive Results of Various Models*

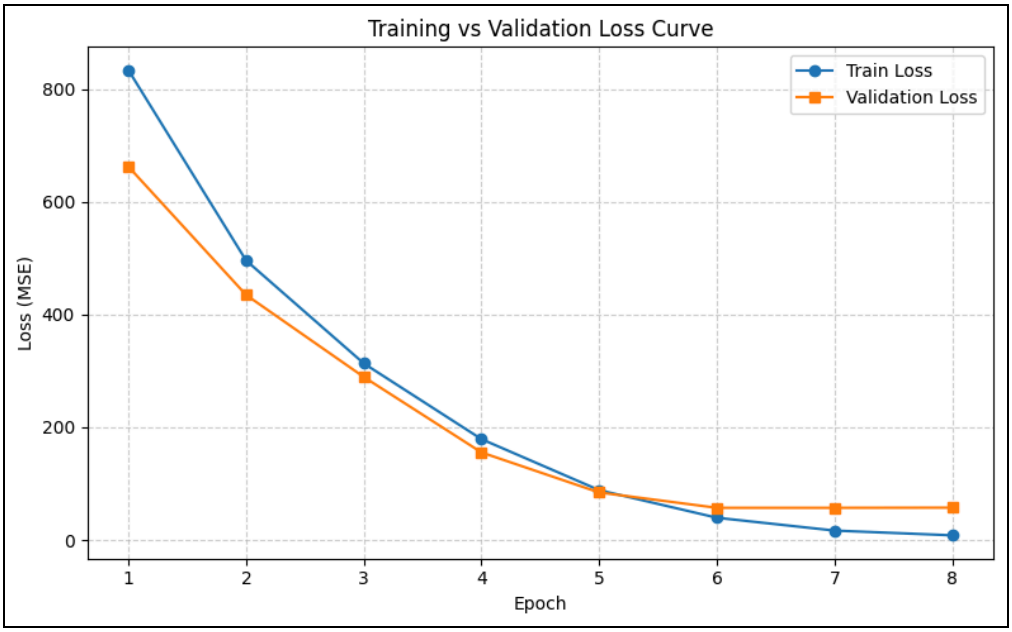

Fig. 5: *Loss Curve of ResNet Model*

The ResNet-18 model outperforms the other models considered for this project, therefore, it was selected as the final model for deployment. Figure 5 shows continuous decrease in the training and validation MSE over 8 epochs, indicating effective learning and minimal overfitting since the training and validation MSE do not have a large difference in value. The primary evaluation metric for our BMI prediction system is the Pearson correlation coefficient (r), measuring the linear relationship between predicted and actual BMI values. Our final model achieved an overall Pearson r of 0.6526, with gender-specific performance showing r = 0.6928 for males and r = 0.6013 for females. The R^2 scores indicate that our model explains approximately 36.4% of the variance in BMI, a reasonable achievement given the challenging nature of inferring body composition from facial features alone.

Challenges

The most significant challenge in preprocessing was achieving consistent facial feature extraction across our dataset of 4,206 images. Faces in the dataset exhibited substantial variation

in pose, lighting conditions, image quality, and capture angles. Initial attempts using basic face detection often failed on images with extreme angles or poor lighting, leading to other strategies like center-cropping when face detection failed, which caused inconsistencies and introduced noise into our training data. Also, the dataset contained images captured under vastly different lighting conditions - from professional portraits to casual selfies. Finding the optimal parameters required extensive experimentation to balance enhancement without introducing artifacts.

Also, there are imbalances in gender. Our dataset contained fewer female samples (1,768 females vs 2,438 males), and facial features correlating with BMI may manifest differently across genders due to differences in, for instance, facial fat distribution patterns.

Remarks and Future Directions

For the ResNet-18 model, the loss curve analysis in Figure 5 reveals successful convergence without significant overfitting, validating our regularization strategies. However, the relatively high final loss values (MSE ≈ 54) suggest inherent limitations in predicting exact BMI from facial features alone. The model appears to capture general trends rather than precise values, which aligns with the biological reality that facial appearance provides only partial information about overall body composition. And the persistent gender gap in performance ($\Delta r \approx 0.09$) indicates that future work should explore gender-specific architectures or additional preprocessing steps tailored to each gender's facial characteristics.

The current dataset's limitations suggest several aspects that could improve model performance. Most importantly, the dataset would benefit from systematic augmentation with controlled variations in facial expressions, as our analysis revealed that neutral expressions dominated the training data, potentially limiting the model's ability to generalize to natural,

expressive faces where facial fullness might be masked or accentuated by different expressions. Additionally, incorporating full-body images alongside facial images could enable multi-modal learning, where the model learns to correlate facial features with actual body composition more accurately. The dataset should also be expanded to include demographic metadata such as age, ethnicity, and height, as BMI manifests differently across populations, and age significantly affects facial fat distribution. Furthermore, collecting longitudinal data from the same individuals at different BMI levels would help the model learn person-specific BMI variations rather than just population-level correlations, potentially improving individual prediction accuracy. Also, the dataset would benefit from professional-quality standardized photos taken under controlled conditions with consistent lighting, pose, and distance, as well as depth information from 3D facial scans that could capture subtle volumetric changes in facial structure that 2D images miss.

Conclusion

This project demonstrates the feasibility of automated BMI estimation from facial images using deep learning, offering strong potential for healthcare applications. Our app enables instant BMI prediction via any smartphone or computer camera, removing the need for physical measurements. This has immediate use in telemedicine, school screenings, and fitness apps for tracking body composition over time.

The system could integrate into healthcare workflows: emergency rooms could triage patients based on obesity-related risks, insurance companies could streamline health assessments, and researchers could analyze social media photos for large-scale epidemiological studies. Rural clinics and nursing homes could also benefit from low-touch, equipment-free health screenings.

Several lessons emerged from this project. We learned the importance of selecting the right models and data augmentation strategies—though augmentation remains largely trial-and-error. While fine-tuning deeper layers showed promise, outperforming the VGG-Face baseline was challenging, emphasizing the need for domain-specific feature learning. Some models benefited from fine-tuning 1–2 top layers, while others performed better with feature extraction. Preventing overfitting was essential, and notably, non-linear regression models did not always outperform simpler linear approaches in this BMI task.

Future improvements include using validated datasets, as self-reported BMI values may be unreliable. Expanding the dataset to include diverse demographics and multi-angle facial views could improve generalization. Additionally, incorporating temporal video data may help capture dynamic facial features relevant to BMI. With further refinement, this system could support scalable public health monitoring and personalized wellness tracking.

References

1. 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 Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, 572–575. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15651
2. Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, *35*(5), 1285–1298. https://doi.org/10.1109/TMI.2016.2528162
3. Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). *How transferable are features in deep neural networks?* In *Advances in Neural Information Processing Systems* (Vol. 27). https://proceedings.neurips.cc/paper_files/paper/2014/file/375c71349b295fbe2dcdca9206f20a06-Paper.pdf
4. Yousaf, N., Hussein, S., & Sultani, W. (2021). Estimation of BMI from facial images using semantic segmentation based region-aware pooling. *Computers in Biology and Medicine*, 133, 104392. https://doi.org/10.1016/j.compbiomed.2021.104392
5. Smith, L. N., & Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 11006, 369-386.
6. Tan, M., & Le, Q. V. (2019). *EfficientNet: Rethinking model scaling for convolutional neural networks*. arXiv preprint arXiv:1905.11946. https://arxiv.org/abs/1905.11946
7. Siddiqui, H., Rattani, A., Ricanek, K., & Hill, T. (2022). *An examination of bias of facial analysis based BMI prediction models*. arXiv preprint arXiv:2204.10262. https://arxiv.org/abs/2204.10262
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929. https://arxiv.org/abs/2010.11929
9. Kaya, Y., Hong, X., & Pietikäinen, M. (2021). Estimating BMI from face images using attention-based multi-task learning with transformer networks. *Pattern Recognition Letters*, *148*, 12–19. https://doi.org/10.1016/j.patrec.2021.04.001
10. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). *Training data-efficient image transformers & distillation through attention*. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 10347–10357). https://arxiv.org/abs/2012.12877
11. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

12. Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. Expert Systems with Applications, 149, 113378. https://doi.org/10.1016/j.eswa.2020.113378