MS APPLIED DATA SCIENCE

Machine Learning II

Body Mass Index (BMI) Prediction Model
Project White Paper

Amulya Jayanti | Halleluya Mengesha | Hira Stanley | Sami Naeem | Vaishnavi Kokadwar

Table of Contents

I. Executive Summary

1. Project Overview

Body Mass Index (BMI) is a widely recognized metric for assessing an individual's overall health and well-being. It serves as an important indicator in studies related to obesity, metabolic disorders, and general physical health. However, much of the BMI data available in public health datasets is self-reported, which introduces significant inaccuracies due to the absence of clinical validation by licensed healthcare professionals. Such inaccuracies can lead to flawed insights in epidemiological studies and hinder the development of effective treatment strategies. This work builds upon the approach proposed by Kocabey et al., which utilizes computer vision techniques to predict BMI directly from facial images. The objective of this study is to improve upon their methodology by leveraging more recent advancements in deep learning architectures and transfer learning.

1.1. Problem Statement

Given the limitations of self-reported BMI data, there is a need for more objective and scalable methods of BMI estimation. The goal of this work is to predict BMI from facial images using computer vision and deep learning, thereby providing a non-invasive and potentially more reliable alternative. While Kocabey et al. demonstrated the feasibility of this approach using VGGFace and VGGNet architectures, the emergence of more sophisticated models and fine-tuning strategies presents an opportunity for enhanced predictive accuracy. This study explores the use of modern pre-trained architectures, combined with tailored fine-tuning techniques, to better align model performance with the specific task of BMI prediction.

1.2. Goal

The primary task is to develop a model capable of estimating a person's BMI from a single facial image. The training dataset consists of 3,210 labeled facial images paired with ground truth BMI values. The model must generalize effectively to unseen images, making it suitable for real-world deployment. Additionally, a user interface is developed to allow users to input static images or live video streams, with real-time BMI predictions.

The pipeline consists of two major stages: (1) facial feature extraction using a variety of pre-trained models such as VGGFace, EfficientNet, and VGGNet, and (2) regression modeling, employing both deep learning regressors and traditional machine learning methods. While the foundational work utilized only VGGFace and VGGNet, this study extends the methodology by incorporating additional feature extractors, fine tuning techniques and regression models to improve performance and explore broader applicability.

1.3. Methodology

1.3.1. Data sources

This study utilizes the dataset from the VisualBMI project, which comprises facial images sourced from publicly available posts on Reddit. The original authors performed extensive preprocessing on the dataset, including manual cleaning, cropping, and resizing of the images to 160×160 pixels, in order to reduce noise and remove extraneous visual information. The dataset consists of 3,210 training images and 752 test images, each annotated with the subject's gender and corresponding BMI. The training and test sets were curated to ensure balanced representation across different BMI ranges and genders, thereby supporting robust and unbiased model training and evaluation.

1.3.2. Image Models

- VGGFace
- VGG19
- ResNet50
- EfficientNet
- FaceNet

1.3.3. Evaluation Metrics

- Pearson correlation (r)
- Mean Absolute Error (MAE)

1.4. Results

- The EfficientNetB3 model achieved the lowest MAE of 4.72 and a Pearson r of 0.67 overall, outperforming both the VGGFace (r = 0.65) and VGGNet (r = 0.47) baselines referenced in the paper .

- Gender-specific performance showed EfficientNetB3 attained r = 0.69 for males and r = 0.65 for females, with only the male metric falling short of the VGG19 target of 0.71.

- An ensemble of top regressors (SVR, CatBoost, LightGBM) on VGGFace embeddings yielded an MAE of 5.04 and r = 0.64, demonstrating that stacking adds marginal gains over single-model CNN outputs .

- Overall, most fine-tuned architectures surpassed the reference paper’s VGGNet performance, confirming the value of modern backbones and transfer-learning strategies

II. Data Collection & Preprocessing

This section details the data preprocessing and feature engineering steps performed to create a comprehensive dataset for a machine learning model aimed at predicting a user's BMI in real-time. The process involves integrating image data sources of peoples' faces—including actual BMI values—and fine-tuning pre-trained image models to get the best performing one.

Data Import and Mounting

Facial images were stored in a directory and metadata, including image filenames, BMI values, and gender, were provided in a CSV file. These were loaded using Pandas and image paths were programmatically mapped to corresponding entries in the dataset. Invalid or missing file references were dropped to ensure data integrity.

Preprocessing Steps

- **Image Resizing**: All facial images were resized to match the input shape required by the respective pre-trained CNN models:
  - 224×224×3 for **VGGFace**, **VGG16/VGG19**, and **ResNet50**
  - 160×160×3 for **FaceNet**
  - 300x300x3 for **EfficientNet**
- **Pixel Normalization**: Pixel values were normalized to the [0, 1] range or standardized as per the base model requirements. For VGG-based models, this included mean subtraction aligned with ImageNet/VGGFace training protocols.
- **Dimensional Formatting**: Images were cast to float32 and reshaped into 4D batches (batch_size × height × width × channels) to support Keras inference.
- **Face Detection (FaceNet only)**: MTCNN was used to detect and align the most prominent face in each image before resizing.

Feature Engineering

VGGFace

- **Model Used**: VGG16-based VGG-Face model pre-trained on facial identity recognition tasks.
- **Feature Extraction**: Features were extracted from the fc6 layer (4096-dimensional vector) by removing the softmax head. These embeddings served as fixed-length representations of facial geometry.
- **Fine-Tuning**: Experiments involved unfreezing deeper layers (Blocks 4 and 5) sequentially, and data augmentations (rotation, flip, zoom, shift) were applied during training.
- **Additional Features**: Gender was concatenated with the extracted embeddings for model training and fed into regressors(Ridge, SVR, Random Forest, KNN, XGB, LightGBM and CatBoost) for enhanced interpretability and ensemble modeling.

VGG16/VGG19

- **Feature Extraction**: In baseline models, only the convolutional base was used (frozen) and coupled with a custom regression head.
- **Fine-Tuning**: In enhanced versions, top convolutional layers were unfrozen, and data augmentations (rotation, flip, zoom, shift) were applied during training.
- **Feature Transfer**: In some setups, the learned embeddings from the best VGG19 model were extracted and fed into traditional regressors for enhanced interpretability and ensemble modeling.

ResNet50

- **Feature Extraction**: The pre-trained ResNet50 was appended with a regression head (GAP + Dense). Later iterations involved unfreezing the top 25 layers to fine-tune deeper representations.
- **Data Augmentation**: Image augmentations such as random flip, rotation, and contrast enhancement were added to improve generalization.
- **Embedding Transfer**: Extracted deep features were used as input to external regressors, optionally enriched with gender labels.

FaceNet

- **Face Localization**: MTCNN localized the main facial region to improve embedding quality.
- **Feature Extraction**: A 512-dimensional vector was extracted from the pre-trained Inception-ResNet backbone.
- **Fine-Tuning**: Block8 modules and the final FC layer were unfrozen for further training with reduced learning rate.
- **Additional Features**: Gender information was appended to FaceNet embeddings to create richer inputs for the regression models.

EfficientNet

- **Model Used:** All 8 EfficientNet B0 - B7 models were tested by training the base model at 5 epochs. The EfficientNet B3 model had the best performing model statistics and was used for further fine tuning
- **Feature Extraction**: The pre-trained EfficientNet B3 was appended with a regression head. Later iterations involved unfreezing the top 30 layers to fine-tune deeper representations. Gender was converted to a binary variable and also used for training the model.
- **Data Augmentation**: Image augmentations such as random flip, rotation,width shift, height shift, zoom range and horizontal flip were added to improve generalization.
- **Embedding Transfer**: Extracted deep features were used as input to external regressors, enriched with gender labels.

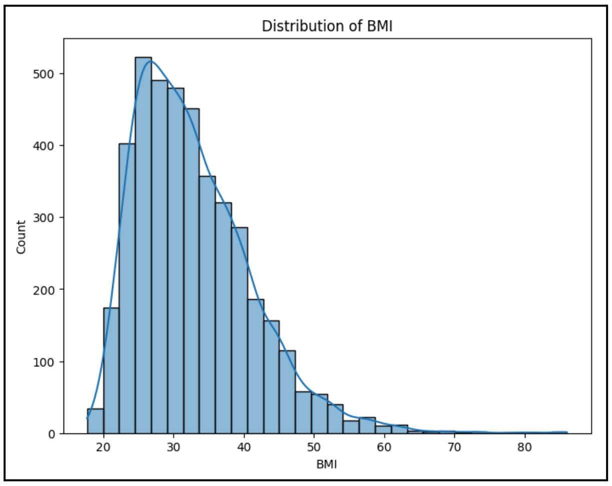
III. Exploratory Data Analysis (EDA)

Data Integrity and Structure

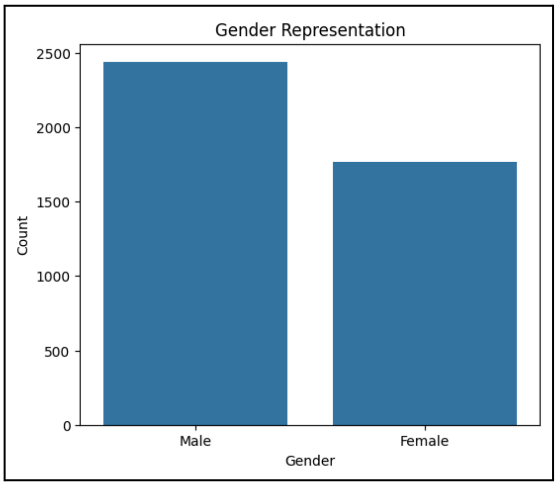
The CSV dataset has **4206** entries and **5** columns of data. We appended an additional sixth column called "filepath" that represents the file path of the corresponding image in the directory.

Column	Data Type	Example Value
[index]	int	0
bmi	float	34.207396
gender	object	Male
is_training	int	1
name	object	img_0.bmp

There are no null or missing values in the original dataset. However, not all records in the CSV file had a corresponding image. Before filtering, there were 3,368 train records and 838 test records. Below are the exploratory data visualizations from this dataset:



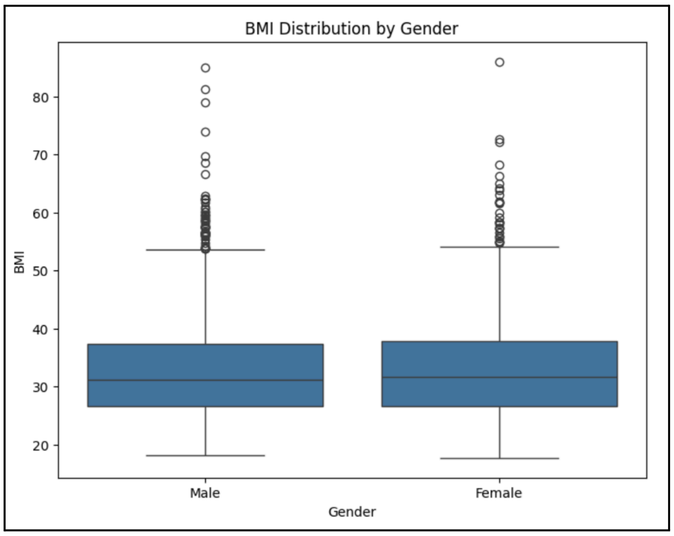
The average BMI in this dataset is 32.80.



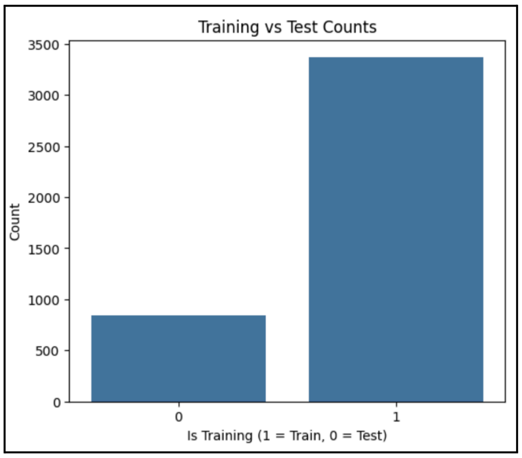
Male: 2438, Female: 1768.

Gender	Min	Max	Mean	St. Dev.
Female	17.7	86.0	32.8	8.6
Male	18.1	85.0	32.8	8.1

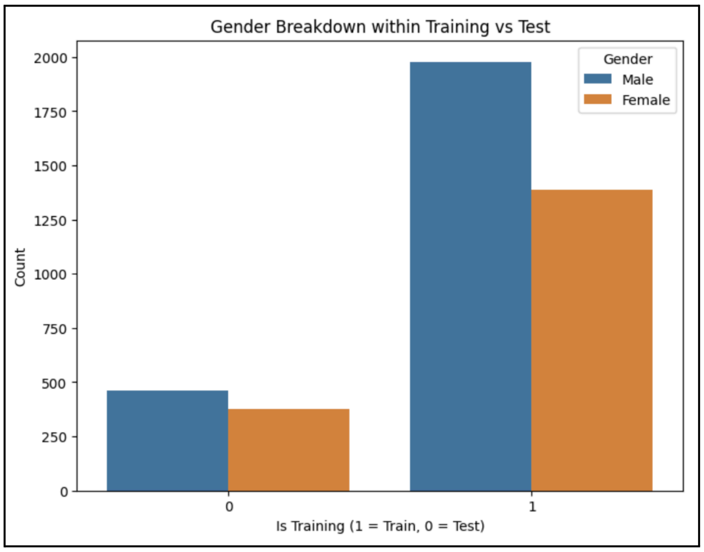
The BMI statistical summary by gender shown above.



BMI boxplot by gender showcasing outliers.



There is an 80/20 train/test split.

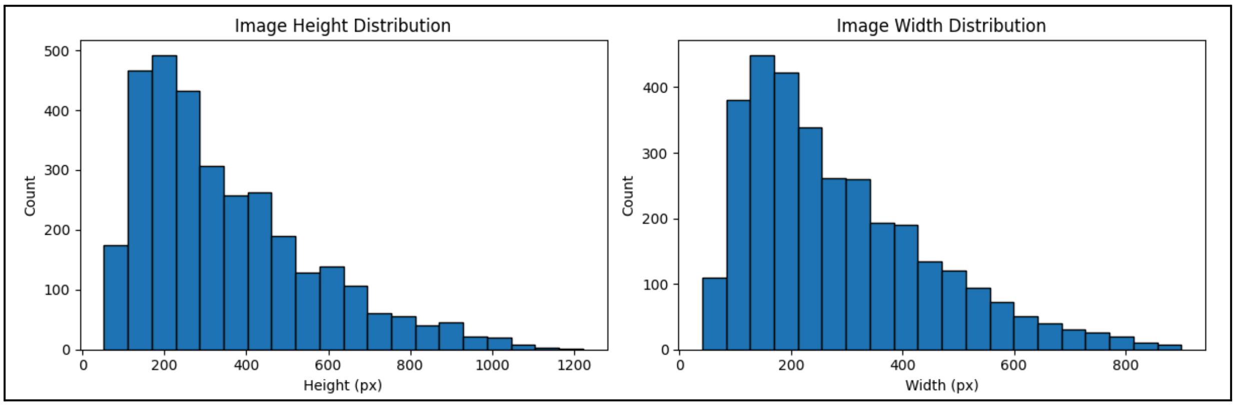


In training, females represent 41% of total, and men represent 58% of total. However, in testing, females represent 45% of total, and men represent 55% of total.

We also conducted some EDA on the .bmp image files.

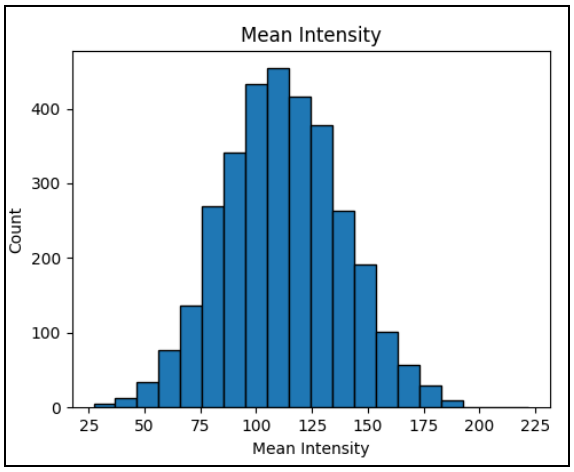
After appending a file path to each record, it was found that several records in the dataset did not have corresponding image files, so they were filtered. The resulting dataset was 3,210 train and 752 test records with actual .bmp files.

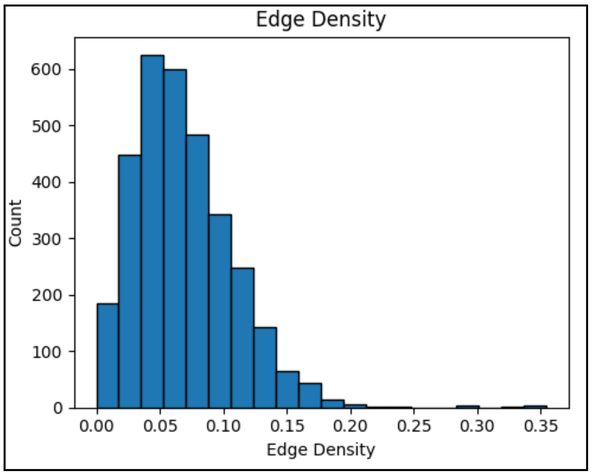
The images had height and width distributions as below before preprocessing.



	Min	**Max**	**Mean**
Height	52	1222	351
Width	41	900	286

The intensity and edge density histograms are shown below.





IV. Machine Learning Models

To analyze an image and predict BMI from a facial image, we made a deep learning-based regression pipeline using pre-trained models like VGG Face, VGG Net, ResNet50, Efficient Net and FaceNet to extract its features, and traditional regressors.

Our BMI prediction pipeline leverages a transfer learning approach using the pre-trained VGG-Face model, followed by classical regression techniques. This two-stage modeling process enables us to extract semantically rich facial features and map them to continuous BMI values with high interpretability and efficiency.

1. Model Types Implemented:

1.1. VGGFace

Baseline (Frozen VGGFace + Regressors):
We first evaluated the VGGFace model—a VGG16-based architecture pre-trained on a large facial identity dataset—by extracting features from its `fc6` layer (4096-dimensional vector). The classification head was removed, and the remaining network was used as a fixed feature extractor. Images were resized to 224×224, normalized, and embeddings were extracted. These embeddings were then passed to traditional regression models. Among the baseline regressors, **CatBoost and LightGBM** performed best, with **MAE of 5.12 and 5.14**, and **Pearson correlation of 0.625 and 0.610**, respectively.

Fine-Tuning of VGGFace:
To extract deeper, BMI-specific features, we progressively unfroze layers from Block 4 and Block 5 of the VGGFace model. Training was conducted over multiple epochs using a reduced learning rate and standard data augmentations including rotations, shifts, zoom, and horizontal flips. However, fine-tuning provided only marginal improvements over the frozen backbone, suggesting that the pre-trained identity features already carried meaningful structure for BMI estimation.

Feature Extraction + Regressor Head:
From the best VGGFace variant (with Block 4 and 5 optionally unfrozen), we extracted the fc6 features and concatenated them with gender as an additional input feature. These were then used to train a broad set of regressors including **Ridge, SVR, Random Forest, KNN, XGBoost, LightGBM, and CatBoost**. Finally, we constructed a **stacked ensemble** using the top-performing regressors (**SVR, CatBoost, LightGBM**). This ensemble achieved the best results across all VGGFace-based models, with **MAE: 5.04, Pearson r (overall): 0.641, Pearson r (Male): 0.653, Pearson r (Female): 0.631**

1.2. VGGNet

Baseline (Frozen VGG16/VGG19 + Regression Head): We first evaluated VGGNet architectures with frozen convolutional layers and a lightweight regression head (Flatten, Dense(256, ReLU), Dropout(0.5), Dense(1, linear)). The input images were resized to 224x224. We ran 10 epochs in both cases.

- **VGG16** produced a test MAE of 9.14 with an overall Pearson *r* of 0.174 (Male: 0.247, Female: 0.070).
- **VGG19** yielded a slight improvement—test MAE 8.82 and *r* 0.176 (Male: 0.237, Female: 0.094) but errors remained high.

Fine-Tuning of VGG19: To extract deeper feature representations, we progressively unfroze convolutional layers and introduced data augmentations such as rotations, shifts, zoom, horizontal flips. Our best VGG-19 variant unfroze the top 8 layers, kept augmentation, and employed a learning rate scheduler over 15 epochs. This achieved MAE 5.03 with Pearson *r* 0.641 (Male: 0.691, Female: 0.574).

Feature Extraction + Regressor Head: Using our best VGG19 model, we extracted the features from that fine-tuned model and plugged them into traditional regressors. We trained a variety of regressors including (Ridge, RF, SVR, KNN, MLP, XGB, LightGBM, CatBoost) and ultimately stacking an ensemble for the best performing regressors (SVR, MLP and Random Forest). The ensemble attained a test ***MAE of 4.99*** and ***Pearson r 0.649*** overall (Male: 0.699, Female: 0.583), outperforming any single traditional regressor and the best fine-tuned model we tried.

1.3. ResNet50

Baseline: For a baseline model, we started with ResNet50, a CNN not trained on facial images, but rather the ImageNet dataset. The input images were resized to 224x224. The first iteration, the baseline model, we added a regression head that included Global Average Pooling, Dropout, and a final Dense layer with linear activation. With a learning rate of 1e-4 and 20 training epochs, it produced a training MAE of 6.85, and validation MAE of 7.44.

Fine-Tuning: The next few iterations involved fine-tuning this model. We added some data augmentation that included image flip, rotation, zoom, and contrast to add variety in the training process and help bring down the validation MAE. Additionally, since ResNet50 has 175 layers, we unfroze more layers and trained the last 25. We also added an L2 regularizer to the final Dense layer. Lastly, we lowered the learning rate to 5e-6. These changes brought down the training MAE to 6.09 and validation MAE to 6.74.

Regressor Head: The last version we tried with ResNet was extracting the features from that fine-tuned model and plugging them into traditional regressors. After separating for gender in the data set, the three best performing models (CatBoost, XGBoost, and SVR) were ensembled to produce the final MAE of 6.00 and Pearson correlations of 0.50 for female and 0.40 for male.

1.4. EfficientNet

Baseline Model: For the baseline, we employed **EfficientNetB3** for convolutional feature extraction. Leveraging compound scaling, EfficientNetB3 balances network depth, width, and resolution to maximize accuracy per FLOP. We loaded the model pre-trained on ImageNet (excluding its top classification layers), resized all inputs to 300×300 pixels, and normalized pixel values using ImageNet statistics. The backbone produces a 1 536-dimensional feature vector for each image.

Fine‑Tuning: To adapt the pre-trained backbone to our BMI‑prediction task, we added a lightweight regression head and initially froze all EfficientNet layers. The head consists of global average pooling, two fully connected layers (512 and 128 units, each followed by ReLU and dropout), and a final linear output node. We trained only the head for 20 epochs with an Adam optimizer (learning rate = 1×10^{-4}). Next, we unfroze the top half of the EfficientNetB3 layers, reduced the learning rate to 5×10^{-6}, and fine‑tuned the entire model for an additional 10–15 epochs, achieving improved validation performance without overfitting.

Regressor Head: Beyond the CNN head, we also evaluated classical regressors on features extracted from the frozen backbone. We concatenated auxiliary metadata (e.g., gender) to the 1 536‑dimensional embeddings and trained Ridge Regression, Support Vector Regression (RBF), XGBoost, LightGBM, and CatBoost models. An ensemble of the top three regressors yielded the lowest mean absolute error (MAE) on the hold‑out set and a strong Pearson correlation coefficient between predicted and true BMI values.

1.5. FaceNet

Baseline Model: For the baseline, we employed the FaceNet model for facial feature extraction. FaceNet utilizes detectors such as MTCNN to localize facial regions within input images. Feature extraction is subsequently performed using an Inception-ResNet architecture pre-trained on the VGGFace2 dataset. All input images were resized and standardized to a resolution of 160×160 pixels. In cases where multiple faces were detected, only the first detected face was retained for further processing to maintain consistency.

Fine-Tuning: To enhance performance, we fine-tuned the model by unfreezing the final fully connected layer and all Block8 modules of the Inception-ResNet backbone. Each Block8 module comprises approximately 6 to 10 internal layers, including convolutional, batch normalization, and activation layers, resulting in a total of around 48 to 50 trainable layers. The learning rate was decreased from 0.01 to 1e-4 to accommodate fine-tuning. However, this approach led to a degradation in performance, indicating that the original pre-trained parameters were already well-optimized for our specific task.

Regressor Head: The features extracted from FaceNet were also used as input to classical regression models. To enrich the embeddings, gender information was concatenated as an additional feature dimension. An ensemble of the top-performing regressors—Random Forest, Ridge Regression, and Support Vector Regression (SVR)—was constructed, which achieved a marginal performance improvement, yielding a mean absolute error (MAE) of 5.52 and a Pearson correlation coefficient of 0.577.

2. Evaluation Metrics Used

Model performance was assessed using **Mean Absolute Error (MAE)** and the **Pearson correlation coefficient (r)**. MAE provided an interpretable measure of prediction error in BMI units, while Pearson r quantified the strength of the linear relationship between predicted and actual BMI values. Metrics were reported both overall and separately for male and female subsets to evaluate gender-specific performance

V. Results

The results from the two models used in the reference paper are below. Our goal has been to beat the performance of these models.

Model: VGG Face	
Metrics	
Overall Pearson (r)	0.65
Male (r)	0.71
Female (r)	0.57

Model: VGG Net	
Metrics	
Overall Pearson (r)	0.47
Male (r)	0.58
Female (r)	0.36

The results of the five fine tuned computer vision models are below. We were successful in beating the Pearson r statistic for the Overall and Female classes, but were only able to achieve a Pearson r statistic of 0.69 compared to the 0.71 target from the reference paper.

Model	MAE	Pearson r (Overall)	Pearson r (Male)	Pearson r (Female)
VGGFace	5.04	0.64	0.65	0.63
VGGNet	4.99	0.65	0.7	0.58
ResNet50	6.04	0.47	0.41	0.51
FaceNet	5.52	0.58	0.62	0.53
EfficientNet	**4.72**	**0.67**	**0.69**	**0.65**

The table below provides detailed results of the five computer vision models, providing a breakdown of the Pearson r statistic and MAE for all the regression models. The Pearson r statistic is reported both for the overall dataset and by gender. The best results for each computer model have been highlighted. The EfficientNetB3 model has the best performance by all metrics except the Male Pearson r statistic, where the VGG19 model performs best.

Regressor	Metrics	Feature Extractor				
		ResNet50	VGGFace	VGG19	EfficientNet	FaceNet
Ridge	MAE	6.122	5.54	5.15	4.78	5.52
	Overall (r)	0.35	0.59	0.63	0.66	0.57
	Male (r)	0.38	-	0.68	0.68	0.61
	Female (r)	0.48	-	0.58	0.65	0.53
Random Forest	MAE	6.08	5.59	4.99	**4.72**	5.7
	Overall (r)	0.45	0.56	0.65	**0.67**	0.55
	Male (r)	0.39	-	0.7	**0.69**	0.6
	Female (r)	0.5	-	0.58	**0.65**	0.49
SVR	MAE	6.08	5.5	**5**	4.88	5.52
	Overall (r)	0.45	0.61	**0.65**	0.66	0.58
	Male (r)	0.4	-	**0.7**	0.67	0.62
	Female (r)	0.5	-	**0.58**	0.64	0.53
KNN	MAE	6.49	6.17	5.2	4.89	6.4
	Overall (r)	0.39	0.42	0.63	0.65	0.4
	Male (r)	0.32	-	0.67	0.67	0.32
	Female (r)	0.45	-	0.57	0.63	0.45
MLP	MAE	6.27	-	4.98	4.89	5.6
	Overall (r)	0.46	-	0.65	0.65	0.54
	Male (r)	0.41	-	0.7	0.68	0.57
	Female (r)	0.5	-	0.59	0.64	0.52
XGB	MAE	6.06	5.51	5.08	4.82	-
	Overall (r)	0.44	0.56	0.64	0.66	-
	Male (r)	0.38	-	0.68	0.68	-
	Female (r)	0.49	-	0.58	0.63	-
LightGBM	MAE	6.11	5.14	5.15	4.78	5.6
	Overall (r)	0.45	0.61	0.62	0.66	0.55
	Male (r)	0.4	-	0.68	0.68	0.57
	Female (r)	0.5	-	0.55	0.64	0.52
CatBoost	MAE	6.07	5.12	5.05	4.76	5.9
	Overall (r)	0.46	0.63	0.65	0.66	0.48
	Male (r)	0.4	-	0.7	0.68	0.53

	Female (r)	0.51	-	0.58	0.64	0.46
Ensembling optional)	MAE	**6.04**	**5.04**	**4.99**	4.73	5.52
	Overall (r)	**0.47**	**0.64**	**0.65**	0.67	0.58
	Male (r)	**0.41**	**0.65**	**0.7**	-	0.62
	Female (r)	**0.51**	**0.63**	**0.58**	-	0.53

VI. Enhancements & Future Work

There are several enhancements that can be made to improve the above outcomes and results, particularly in terms of the data, modeling, and ethics.

Strengthen the dataset

- Broaden demographic coverage by adding more subjects across ages, ethnicities, and gender.
- Capture a wider range of facial expressions and real-world conditions (e.g. varied lighting, masks, glasses).
- Merge in complementary public and private datasets to boost sample diversity and model robustness.

Refine Modeling Strategies

- Fine-tune ensemble methods—experiment with stacking or weighted averaging to push Pearson *r* higher and MAE lower.
- Conduct systematic hyperparameter adjustments (learning rates, epoch counts, layers to unfreeze) to identify optimal configurations.
- Investigate attention-based backbones (such as Vision Transformers or hybrid CNN-Transformer models) to let the network learn which facial regions matter most.

Embed Fairness & Transparency

- Evaluate performance across subgroups—gender, age brackets and skin tones—to uncover and correct any biases.
- Integrate interpretability tools (such as SHAP) so you can see exactly which facial features drive each BMI estimate.
- Regularly audit outlier predictions to guard against unintended errors in deployment environments.

By integrating these enhancements, we anticipate surpassing our current best model's performance. Our systematic exploration of architectures, ensemble methods, and hyperparameter tuning has demonstrated that facial imagery can reliably estimate BMI. With these refinements, facial-based BMI estimation can serve as a seamless, noninvasive screening tool that complements traditional health assessments and acts as an early indicator ahead of clinical consultations