Human Body Shape Classification Using Deep Learning

Harshit Jain
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
harshitjaintata@gmail.com

Dhrubo Bhattacharjee
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
dhrubob026@gmail.com

Kedar Hardikar
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
hkedar0@gmail.com

Ekaksh Upadhyay
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
upadhyayekaksh@gmail.com

Prof. Prachi Kadam
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
prachi.kadam@sitpune.edu.in

Prof. Om Mishra
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
om.mishra@sitpune.edu.in

***Abstract*—In our research, we meticulously applied a series of data preprocessing techniques, which involved tasks such as resizing images, converting them to grayscale, applying high-pass filters, normalizing, and setting thresholds. These procedures were essential for standardizing and improving the image data, making it suitable for training deep learning models. However, a significant drawback of our study was the absence of an appropriate dataset with diverse body shape classes and uniform image quality. Instead, the dataset we had access to contained only four distinct body shape categories, and the images within each category exhibited considerable inconsistencies. This high level of image variation posed a challenge for our models, impeding their ability to effectively learn and extract features. Consequently, our models were more prone to overfitting, potentially memorizing specific instances within each class rather than acquiring generalizable features.**

***Index Terms*—Body Shape Classification, Image Preprocessing, Computer Vision, Deep Learning, Convolutional Neural Network, Transformers**

I. Introduction :

The categorization of human body forms from visual data is a challenging problem in computer vision and deep learning. There are several uses for the capacity to identify and classify body forms in a variety of fields, such as fashion, fitness, and medicine. Precise categorization not only has the potential to improve e-commerce experiences by offering tailored suggestions, but it also supports body positivity and self-worth by recognizing and appreciating the variety of human bodies. The central problem addressed by this deep learning project is the classification of human body shapes from images.
The purpose of this report is to present the findings and results of a deep learning project focused on body size classification in humans. The scope of this project is the development of a deep learning model, training and assessment for the classification of individuals into specific body size groups Key objectives include:

- To build a robust deep-learning model that can accurately classify body size based on input data.
- To evaluate the performance of the model on a representative data set of human body images.
- To see how this technology can be applied to areas such as fashion, fitness, and health.
- To complement existing knowledge in computer vision and deep learning techniques for image segmentation.

This research project seeks to address a specific problem in computer vision and deep learning, with potential applications across various domains, ultimately aiming to improve the accuracy and efficiency of body shape classification for the benefit of individuals and industries alike. The project's importance and motivation stem from its endeavor to tackle the intricate task of classifying human body shapes using advanced deep-learning techniques. This research is driven by a few key reasons:

- Practical Relevance: Accurate body shape classification holds practical value in industries like fashion and fitness, where it can lead to personalized recommendations and improved customer satisfaction, ultimately reducing returns. In addition, it has the potential to enhance fitness and healthcare guidance based on individual body types.

- Psychological Well-being: The project aligns with the broader societal goal of fostering body positivity and enhancing self-esteem by recognizing and celebrating the diversity of human physiques.
- Context in Technology: This project builds on the progress made in computer vision and deep learning, specifically employing Convolutional Neural Networks (CNNs) to address the complexities associated with classifying human body shapes.

II. Literature Review:

The research proposes an approach to enhance the online shopping experience by estimating human body measurements from 2D images captured with smartphone cameras. It involves photographing volunteers, manually measuring their bodies, and collecting reports of their actual clothing sizes. Pre-trained computer vision algorithms are utilized to detect major body parts. The methodology segments images into 40 parts, extracting focal points for estimating measurements such as shoulder width, bust circumference, waist circumference, and hip circumference. Multiple machine learning models are trained to predict clothing sizes based on estimated measurements. The study reveals discrepancies between predicted and reported sizes, highlighting potential limitations related to variations in body shapes, camera angles, and model accuracy. Future work aims to improve the detection of side images, enhance SVM models, and minimize error percentages, addressing these limitations. [1]The thesis presents a novel approach for estimating height and waist circumference from 2D body scan images using ResNet-50 and Inception V3. The methodology involves further developing existing datasets, adapting CAESAR dataset meshes for neural network training, and conducting experiments. The results show promising outcomes with MAE error losses of 9.119 mm for height and 58.46 mm for waist circumference. The impact of different viewpoints on height prediction is explored, indicating that front and back images suffice for accurate height estimation. However, more research is needed to solidify waist circumference predictions and address gender biases. Future work will incorporate age and additional body dimensions, with a focus on real human body scans, especially in children, for early malnutrition detection and health trend analysis. Limitations include the need for further investigation into waist predictions and addressing gender biases. [2]

The proposed BMnet introduces a novel method for estimating body measurements based on silhouettes.. The main innovation involves a differentiable adversarial training approach that generates challenging body shapes within the SMPL shape, highlighting potential training gaps. When the BMnet training is augmented with these adversarial shapes, measurement accuracy is enhanced, particularly in cases where real data is scarce, resulting in new state-of-the-art outcomes. The study also introduces the BodyM dataset, acquired from real human subjects, to promote advancements in body measurement research. Limitations may include potential challenges in accurately representing highly diverse body shapes, and the dataset's representativeness, which may impact the generalizability of the model. [3]The paper explores the impact of different data representations, including 2-dimensional images and 3-dimensional point clouds, in estimating anthropometric body measurements. It involves the generation of a large-scale synthetic dataset encompassing various data formats, providing valuable resources for research in human body analysis. Baseline end-to-end methods for accurate body measurement estimation from 2-dimensional and 3-dimensional inputs are presented. The findings emphasize the significance of both the grid structure and depth information in the input data, contributing positively to the estimation process. Grid-structured gray-scale images and unstructured 3D point clouds both yield competitive results, with a mean error of approximately 5 mm. Limitations may include potential challenges in handling highly diverse body shapes and variations, as well as the generalization of the model to real-world scenarios beyond the synthetic dataset. [4]The paper introduces the application of neural networks for body measurements, conducting experiments to assess the feasibility of predicting useful measurements from diverse body data. It compares the performance of three types of input data and their corresponding network models, demonstrating good numerical measurement outcomes despite limited data. However, there is room for improvement. More extensive and privacy-preserving data collection is needed, focusing on the even distribution of the data and selecting relevant body parts as input for neural network models. Data preprocessing is essential for achieving these objectives. While neural networks can aid in cost reduction for body measurements, challenges remain in data quality and quantity, as well as the selection of pertinent body parts. Further efforts are required to enhance the performance of neural networks and increase cost savings in body measurement using these models. Limitations include the need for privacy-conscious data collection, ensuring data distribution, and addressing data quality concerns, as well as refining the selection of relevant body parts in the input data. [5]

The work presents a comprehensive pipeline for 3D clothes simulation on human avatars, comprising body dimension recognition, clothes scanning, and 3-dimensional avatar simulation. The implementation and performance of these sub-approaches are detailed. However, limitations include high error rates in both approaches due to factors like background subtraction in the webcam approach, low-resolution issues in the Kinect approach, and sensitivity to pose variations. Standardization of pose using 3D scanners is suggested to address these issues. Additionally, calculated circumferences for hips, waist, and chest often exceed real measurements, prompting the need for standard deviation incorporation. The work has not explored certain measurement categories, and future work should consider anthropomorphic differences and clothing variations for improved body dimension measurements. The segmentation of subjects in clothes scanning could benefit from alternative methods such as point cloud fusion and the projection of binary pixel patterns to overcome issues related

to low-texture areas in cloth material, enhancing measurement accuracy. [6]In this study, the researchers aimed to evaluate the stability of FFIT's body classification system in fields such as ergonomics, garment construction, and scientific research. They revealed that variations in the definitions of measurement placements within FFIT could lead to inconsistent body shape classifications for the same individuals. The methodology involves analyzing measurement placement definitions and their impact on body shape classifications. The primary limitation is that FFIT's classification system lacks consistent measurement definitions, which hinders its reliability. This highlights the need for standardized anthropometric practices and more comprehensive inclusion of shoulder anthropometrics in body classification systems, potentially enhancing their utility in various applications, including clothing development. [7]
The study developed polyhedral models representing 122 Japanese adult female body shapes and examined angular defects in those. Principal Component Analysis revealed that the first and second principal components primarily described body shape characteristics, with the first component associated with the waist-to-bust girth ratio and the second with the bust-to-back length ratio. The body shapes were classified into four quadrants based on these angles, indicating the ease or difficulty of creating clothing patterns for different shapes. The research also explored age-related changes in body shape. However, the findings are limited to Japanese adult females, and practical applications for garment construction and ergonomics require further investigation and broader applicability. [8]The study introduces a program for semi-automatic quantitative morphological analysis of neurons, enabling the evaluation of cell bodies and dendrite trees through various methods, including automatic extraction and manual editing. A wide range of morphological parameters is utilized to better characterize normal neuron morphology and identify structural and functional changes resulting from experimental or pathological conditions. The program also offers a reliable automatic method for classifying cell body shapes within neuron populations. While this method serves as a precise and practical tool for neuronal research, its limitations may include challenges in accurately evaluating certain dendritic features and potential restrictions when applied to cell populations other than neurons, particularly if staining specificity is a concern. [9]
The research employs topological data analysis to extract significant shape-related information from anthropometric point clouds. The study reveals that homologies in the persistence diagram of human body points have anatomical interpretations, aiding in the accurate detection of scan anomalies through clustering algorithms. Gender differentiation is achieved effectively by focusing on trunk body points and employing hierarchical clustering methods, with Ward-linkage and K-Medoids clustering showing superior performance over complete-linkage hierarchical clustering. The study ultimately identifies eight morphotypes for men and seven for women, characterized by weight classes, circumferential ratios, torso sizes, and lower body shapes. While this approach is promising for anomaly detection and classification, limitations may include challenges in handling point clouds in various contexts and the need for further research to extend the method to other human body-related problems, such as measurement extraction using supervised machine learning algorithms. [10]The paper tackles the task of estimating human 3D shape from orthogonal human body masks, framing it as a chunk wise regression mapping problem guided by the input mask's body type segmentation. To achieve accurate 3D shape estimation, the paper introduces a regression network architecture based on the multi-channel Swin-T framework. This design aims for simplicity and efficiency, benefiting from Swin-T's feature extraction capabilities. However, potential limitations might include the need for large and diverse training datasets and addressing variations in body poses, clothing, and environmental factors, which could impact the model's generalization and real-world applicability. [11]The study addresses the challenge of accurately selecting characteristic factors for predicting human body shape, impacting clothing mass customization. It introduces an improved GA-BP comprehensive prediction model, combining genetic algorithms (GA) with backpropagation neural networks (BP) and principal component analysis (PCA). This model outperforms other algorithms in verifying known body shape data, distinguishing shapes, and predicting unknown body shapes. GA optimizes the BP neural network by eliminating input variable correlations and fine-tuning weights and thresholds, mitigating local optima issues and enhancing prediction accuracy. While this model provides valuable insights into human body shape, potential limitations may arise when extending it to diverse age and gender groups, necessitating further research and validation. [12]

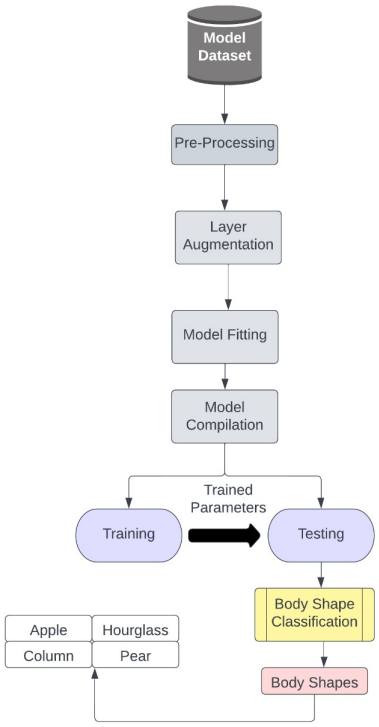
III. Methodology:

A. Data Preprocessing -

In the pursuit of accurately classifying human body shapes, an essential aspect of the project involves meticulous data preprocessing. The primary objective of these preprocessing steps is to ensure that input images are suitably prepared for training deep learning models, ultimately enhancing the quality and relevance of the data. Below, we describe the key data preprocessing techniques applied in this project[Fig 8.]:

- Image Resizing: The first step in the preprocessing pipeline involves resizing the input images to a standardized target size. This operation ensures that all images are of uniform dimensions, which is vital for deep learning models to process the data consistently.
- Grayscale Conversion: Following the resizing operation, the images are converted to grayscale. Grayscale images are represented using a single channel, simplifying data and computational complexity. Grayscale images provide the essential structural information while reducing the amount of information the model needs to process.
- High-Pass Filtering: High-pass filtering is applied to the grayscale images using the "FIND_EDGES" filter from the Python Imaging Library's ImageFilter module. This

Fig. 1. Work Flow



filter enhances the edges within the image by accentuating regions where pixel values undergo rapid changes. This process effectively highlights contours and shapes within the images, facilitating improved feature extraction.

- Conversion to NumPy Array: The high-pass filtered images are then converted into NumPy arrays. NumPy arrays are a fundamental data format for numerical operations in Python. This step allows for efficient manipulation and processing of the image data, a crucial prerequisite for deep learning.
- Normalization: The pixel values within the NumPy array are normalized to a specific range. Normalization is a standard practice in machine learning to scale the values within the range of [0, 1] or [-1, 1]. In this project, the normalization process involves subtracting the minimum value and dividing by the range between the minimum and maximum values, making the data compatible with machine learning algorithms.
- Image Thresholding: Image thresholding is employed to convert the normalized image into a binary image. This process sets pixel values above a predefined threshold to 1.0 (white) and pixel values below the threshold to 0.0 (black). Thresholding simplifies the image further, making it binary and facilitating feature extraction.

B. Models used-

1) ResNet-: In 2015 the ResNet (Residual Network) was introduced as an innovative deep-learning architecture. This specific form of convolutional neural network (CNN) addresses the vanishing gradient problem commonly encountered in extremely deep neural networks.ResNet utilizes residual blocks, which allow the training of exceptionally deep networks while maintaining high performance. In traditional deep networks, as the network becomes deeper, it becomes more challenging for gradients to propagate through all layers during training. This can lead to the vanishing gradient problem, making it difficult to successfully train deep networks. However, ResNet introduces skip connections as a solution to this problem. Our model utilized the ResNet50 variation of the ResNet architecture. It has a deep architecture and 50 convolutional layers in total. We used pre-learned weights that were trained on the ImageNet dataset to initialize the model because it has millions of annotated images spread across many branches. We chose to use ResNet50 as a feature extractor for our custom classifier on top of it rather than using its output layer, which has fully linked layers. (244,244,3) was the input shape for the images.[Fig 9.] The custom layers added on top of the base model are:

- Global Average Pooling(GAP) layer was added to minimize the spatial dimensions of the feature map.
- A Dense layer with 128 units along with ReLU activation function was used to memorize the high-level representation from the features map received from the GAP layer.
- Dropout Layer with 0.2 Dropout was used for regularization with help with overfitting.
- Output Dense Layer with 4 units and Softmax Activation was used as the final layer responsible for making predictions. Softmax activation function was used for multi-class classification with 4 classes('apple', 'column', 'hourglass', 'pear').

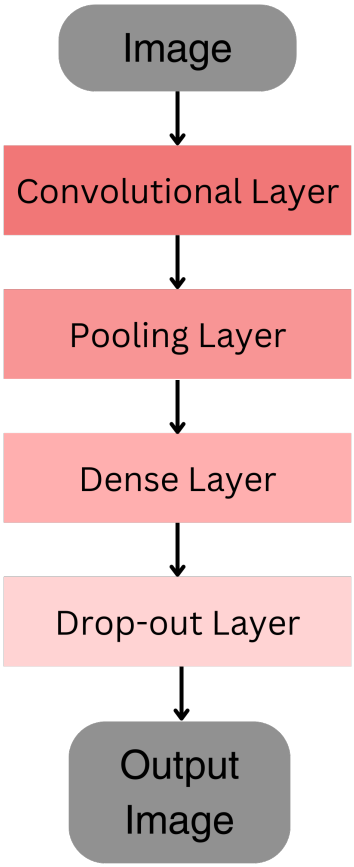
The model was finally compiled using the Adam Optimizer. Adam combines both RMSprop and momentum-based optimization methods, which makes it effective for a wide range of tasks. For the Loss Function, 'sparse_categorical_crossentropy' was used for multi-class classification as the labels were integer labels and not one hot encoded vector. ResNet is a well-established deep learning architecture known for its ability to handle very deep networks while mitigating the vanishing gradient problem. It was chosen as the initial model because of its impressive feature extraction abilities, advantages in transfer learning, and, its ability to overcome vanishing gradient concerns. The additional layers were added to adapt the model to the specific classification task and to extract relevant features. The model was trained for 5 epochs.

ResNet-50 stands as a highly appropriate foundation for utilizing deep learning in human body shape classification, thanks to its impressive feature extraction abilities, advantages in transfer learning, ability to overcome vanishing gradient concerns, exceptional performance, widespread use, adaptability to varying image conditions, straightforward customization, and ample support and resources within the community.

2) CNN- : Convolutional Neural Networks, or CNNs, are deep learning model that are particularly made for processing and analyzing visual input, such as images and videos. CNNs

have completely modernized the field of computer vision and have excelled in a lot of tasks like image classification, object identification, facial recognition, and others. Different

Fig. 2. CNN Architecture



characteristics of the CNN are[Fig 2.]:

- Convolutional Layers: These are used to take the input images with small filters. These filters move over each pixel of the input images depending on the stride and they produce feature maps that are used to capture patterns, edges, and textures.
- Pooling Layers: These are used to reduce the spatial dimension of the feature maps by only selecting the relevant features. It helps to scale down the feature map and make the network more efficient.
- Fully Connected Layers: These are used at the end of multiple convolutional and pooling layers. They connect all the neurons from previous layers and are used for high-level feature learning.
- Dropout: It is a regularization technique used to prevent overfitting. It randomly deactivates some fraction of neurons to prevent the model from over learning.

The input shape provided for the input layer of the CNN was (244,244,3), which accepts images with dimensions 244*244 pixels and 3 color channels(RGB). For the convolutional layers, 5 Conv2D layers were used initial layer consisted of 96 filters with a kernel size of 11*11 and 4*4 stride the following layers consisted of 256 and 384 filters with a kernel size of 5*5 and 3*3 respectively. ReLU activation function was used in each layer. ReLU introduces nonlinearity to the network to help it learn complex features. L2 regularizer was used for regularization which helped in preventing the overfitting, its strength was set to 0.0005. Three MaxPooling layers were also used alongside with the pool size of 3*3 to reduce the spatial dimension of the feature map. A single flatten layer was used to flatten the output from the convolutional layer to a 1D vector. For the fully connected layers, 2 dense layers with 4096 units alongside ReLU activation function and L2 kernel regularization were used to memorize the high-level representation from the features map received from the flattened layer.[Fig 10.] The model was finally compiled using Adam optimizer with an initial learning rate defined as 0.001. Adam combines both RMSprop and momentum-based optimization methods, which makes it effective for a wide range of tasks. An exponential learning rate decay schedule was used as it decays exponentially with a decay rate of 0.9 every 10,000 steps. For the Loss Function, 'sparse_categorical_crossentropy' was used for multi-class classification as the labels were integer labels and not one hot encoded vector. The total trainable parameters were 46763396. No. of epochs were set to 50 and batch size was set to 32. Custom CNNs offer flexibility in designing architectures specifically tailored to the problem at hand. In this case, the CNN was constructed to allow experimentation with different layer sizes and depths to capture intricate features associated with body shape.

3) Vision Transformer-: The "Vision Transformer" (ViT) model architecture applies the principles of the Transformer architecture to computer vision tasks. Originally developed for natural language processing, Transformers have proven effective in tasks like language understanding and machine translation. The extension of this concept to image analysis gives rise to Vision Transformers, enabling them to handle various computer vision problems, including image classification, object detection, and image segmentation. Features of the Vision Transformer are:

- Vision transformer has a self-attention mechanism which allows it to capture long range dependencies from data which helps it to understand the images.
- In contrast to Convolutional Neural Networks (CNNs), which operate on small local regions of an image, Vision Transformers split the input image into smaller patches. Each patch is treated as a separate "word" in the context of the Transformer. This allows the model to process images in a more structured and scalable manner.

- Vision Transformers consists of several layers of Transformer blocks, each of which contains feedforward and self-attentional neural networks. These building elements give the model the ability to learn abstract and hierarchical representations of the input image.
- Dropout: For tasks like image classification, Vision Transformers use a classification head that takes the output embeddings from the Transformer layers and produces class predictions.

For our model, Firstly Data augmentation is applied to input images, including normalization, resizing to the specified image size of 64*64, horizontal flipping, random rotation, and random zoom. This helps increase the diversity of training data and improves the model's generalization. A custom patch layer is used with a patch size of 6 to extract non-overlapping patches from the augmented images. Then another custom layer, 'PatchEncoder' encodes the extracted patches using a dense layer for projection and position embeddings for each patch.[Fig 3.] 8 transformer layers are created, Each block includes:

- Layer normalization.
- Multi-Head Self-Attention: This layer allows the model to capture relationships between patches.
- Skip connection.
- Layer normalization.
- Multi-Layer Perceptron (MLP): It applies feedforward neural networks to the patches.

Fig. 3. Patched Image

After the Transformer blocks, a layer of layer normalization is applied, and the representation is flattened and passed through a dropout layer.[Fig. 11] Additional dense layers (MLP) are applied to further process the representation before classification. And finally, for the Classification head, a dense layer with 4 units is used for the classification of 4 different classes. The optimizer used was AdamW optimizer from TensorFlow Addons (tfa). The "AdamW" optimizer is a variant of the popular Adam optimizer that includes a form of weight decay regularization directly in the optimization algorithm. The optimizer uses a learning rate of 0.001 and has a weight decay of 0.001. Sparse categorical cross-entropy is used as the loss function for classification tasks. No. of epochs were set to 100 for training and batch size was set to 256. Vision Transformers have shown great promise in various computer vision tasks, particularly in image classification. The choice of ViT reflects its capacity to model long-range dependencies and its potential for capturing complex patterns in the data. The architecture's strong performance is in part due to its suitability for image classification tasks, as it doesn't rely on handcrafted features.

IV. Experimental Setup:

A. Dataset Description:

The dataset used in this project comprises images representing different human body shapes, categorized into four distinct classes: "apple," "hourglass," "column," and "pear." To facilitate data handling and labeling, a mapping has been established between class labels and class prefixes. Below is an overview of the dataset:

- Data Collection: The dataset we used is of fashion dataset which consists of images depicting diverse human body shapes. Details regarding the collection process were not provided at the source.
- Size and Characteristics: The dataset consists of 1064 images of the female human body of different body shapes divided into 4 classes. The image file is labeled as "BodyShape_CountOfThatClass". For example "apple_204".

Fig. 4. Apple Shaped Body

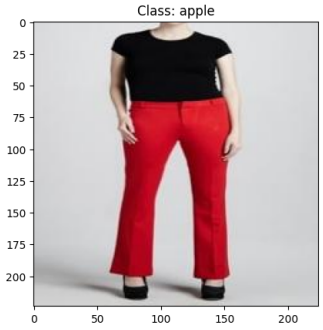
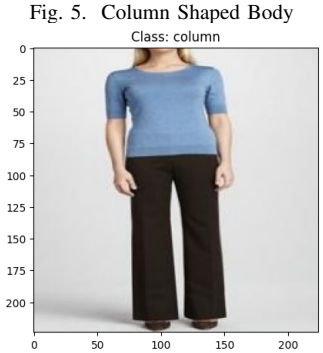


Fig. 5. Column Shaped Body



B. Training Procedure:

The training procedure is outlined with essential details:

- Data Split: The code defines a training-test split ratio of 0.7, signifying that 70% of the data is allocated for training, with the remaining 30% reserved for testing.

Fig. 6. Hourglass Shaped Body

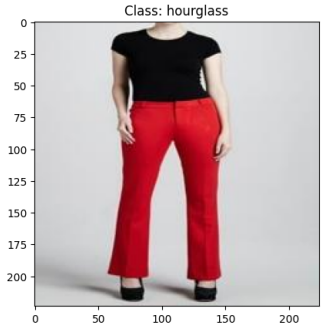
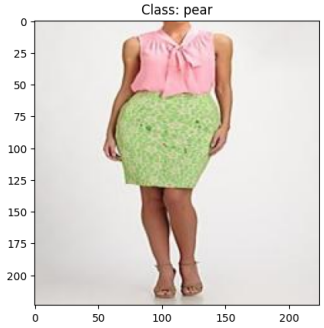


Fig. 7. Pear Shaped Body



- Validation Set: During training, the model's performance is evaluated on a validation set using the test images that is 30% of the total images and their corresponding labels.
- Number of Classes: The dataset encompasses four classes: "apple," "hourglass," "column," and "pear", representing different human body shapes. This information is crucial for configuring the model's output layer with the appropriate number of units.
- Input Shape: For CNN and ResNet the input images are expected to have a resolution of 224x224 pixels with three color channels (RGB) and for ViT the input image has a resolution of 512*512. This information is pivotal for configuring the neural network's input layer correctly.
- Learning Rate: For CNN and ViT The training process is set with a learning rate of 0.001. The learning rate is a hyperparameter that influences the step size during the optimization process of the model.
- Weight Decay: For CNN and ViT, A weight decay value of 0.0001 is defined. Weight decay acts as a regularization technique, contributing to the prevention of overfitting during training.
- Batch Size: For CNN the batch size was set to 32 and that of ViT was set to 256. The batch size determines the number of examples processed during each iteration of the training process.
- Number of Epochs: The training procedure for CNN is configured to run for a total of 50 epochs and for ViT, it is configured to run for a total of 100 epochs. An epoch represents a single complete pass through the entire dataset during model training.

C. Evaluation Metrics:

1) F1 Score- : F1 score combines recall and precision into a single value. When working with datasets that are unbalanced or when both false positives and false negatives are important, it is extremely helpful. The formula for the F1 score is the harmonic mean of Precision and Recall. An F1 score of 1.0 means all the the positive predictions are correct and the model performance is very good whereas an F1 score of 0.0 indicates that the model performance is very poor.[Fig 14.]

2) Confusion Matrix- : A confusion matrix is a tabular representation that provides detailed information about the model's performance by breaking down predictions into categories such as true positives, true negatives, false positives, and false negatives.[Fig 13.] True Positive: count of correctly predicted positive classes. True Negative: count of correctly predicted negative classes. False Positive: count of negative classes incorrectly predicted as positive. False Negative: count of positive classes incorrectly predicted as negative. The confusion matrix provides insights into the model's performance for each class in the classification problem.

3) Accuracy Score- : The accuracy score is a simple metric that measures the correctly classified classes in the entire dataset. It is expressed as a percentage. The formula for checking accuracy is the ratio of the number of correct predictions of classes to the total predictions.The higher the accuracy score, the better the model. [Fig 12.]

V. Results and Discussion:

A. Result:

Model Used	Accuracy Score
ResNet	42%
CNN	60.9%
ViT	34%

TABLE I
ACCURACY SCORE OF DIFFERENT MODELS.

The project explored three distinct deep learning models - ResNet, a custom CNN, and a Vision Transformer. The CNN model, in particular, exhibited the highest accuracy, 60.9%. The CNN model achieved the highest accuracy among the three models, as it was better at capturing complex patterns and relationships in the images. ResNet's a well-established choice for image classification tasks, and a 42% accuracy suggests that the model can learn some discriminative features for body shape classification. The Vision Transformer model achieved the lowest accuracy among the three models 34%.

Model Used	F1 Score
ResNet	0.4123
CNN	0.6067
ViT	0.3307

TABLE II
F1 SCORE OF DIFFERENT MODELS.

B. Discussion:

A suite of data preprocessing techniques was diligently applied, encompassing image resizing, grayscale conversion, high-pass filtering, normalization, and thresholding. These techniques served as pivotal steps in standardizing and enhancing the image data, ensuring its readiness for deep learning model training. The main limitation was that we could not find a proper dataset that consisted of multiple body shape classes, without any inconsistency in the images of a particular class. Rather we had a dataset that consisted of only 4 different classes for body shapes and there was a high inner class variance, all the images under a single class were very inconsistent and the model was unable to learn and extract features correctly for a particular class. With this limited data and high inconsistency, the models are more prone to overfitting. They might have memorized specific instances within each class rather than learning generalizable features.

VI. Conclusion and Future Work:

The deep learning project aimed to classify human body shapes into four distinct classes. The key findings are the achieved accuracies of 60.9% with CNN, 42% with ResNet, and 34% with Vision Transformer. These results indicate that the CNN model outperformed the other two models in the classification task. The varying accuracies among the models mean that the choice of model architecture has a substantial impact on the project's success. The CNN model, known for its ability to capture complex patterns and relationships, demonstrated the best performance, while ResNet, with its deep architecture, lagged behind. The dataset also presented challenges due to the inconsistency and limited variability of images within each class. This inconsistency likely contributed to the suboptimal model performances. The implications of your results and their significance in the broader context of deep learning:

- Model Selection: The project emphasizes how crucial it is to choose the appropriate deep-learning model for the given task. Due to their special attention processes, Vision Transformers are particularly good at detecting subtle elements in image classification tasks.
- Data Quality: The outcomes highlight how important high-quality data is to deep learning. Poor or inconsistent data might cause poor model performance. It is crucial to guarantee varied and properly labeled training data.
- Generalizability: The discoveries have greater implications for computer vision and deep learning. Practical applications require an understanding of model generalizability and how models behave on real-world, erratic data.

For Future work, we will explore advanced data augmentation techniques that can help the models handle inconsistency and improve their ability to generalize. We will also look for new datasets with less inconsistency within classes and will ensure that they cover a broader range of variations in body shape, poses, lighting conditions, and orientations. We will also invest in further hyperparameter tuning and model fine-tuning to extract better features from the data. We will also try to find the body measurements using computer vision and will classify the measurements into different categories to use to recommend fashion clothing. In further future, we would like to deploy this model in a mobile application through which user can just click their picture and get fashion recommendations based on their body shape and measurements.

References

[1] S. Ashmawi, M. Alharbi, A. Almaghrabi, and A. Alhothali, "Fitme: Body measurement estimations using machine learning method," *Procedia Computer Science*, vol. 163, pp. 209–217, 01 2019.

[2] H. Mohammedkhan, M. Balvert, C. Guven, and E. Postma, "Predicting human body dimensions from single images: a first step in automatic malnutrition detection," 01 2021.

[3] N. Ruiz, M. Bellver, T. Bolkart, A. Arora, M. C. Lin, J. Romero, and R. Bala, "Human body measurement estimation with adversarial augmentation," in *2022 International Conference on 3D Vision (3DV)*, (Los Alamitos, CA, USA), pp. 219–230, IEEE Computer Society, sep 2022.

[4] D. Skorvankova, A. Riečický, and M. Madaras, "Automatic estimation of anthropometric human body measurements," 12 2021.

[5] J. Zhao, X. He, and Q. Kemao, "Automatic body measurement by neural networks," *Proceedings of the 2019 2nd International Conference on Robot Systems and Applications*, 2019.

[6] D. Siegmund, T. Samartzidis, N. Damer, A. Nouak, and C. Busch, *Virtual Fitting Pipeline: Body Dimension Recognition, Cloth Modeling, and On-Body Simulation*. 09 2014.

[7] K. B. Christopher J. Parker, Steven George Hayes and S. Gill, "Assessing the female figure identification technique's reliability as a body shape classification system," *Ergonomics*, vol. 64, no. 8, pp. 1035–1051, 2021. PMID: 33719914.

[8] W. Lee and H. Imaoka, "Classification of body shape characteristics of women's torsos using angles," *International Journal of Clothing Science and Technology*, vol. 22, pp. 297–311, 08 2010.

[9] M. Masseroli, A. Bollea, and G. Forloni, "Quantitative morphology and shape classification of neurons by computerized image analysis," *Computer Methods and Programs in Biomedicine*, vol. 41, no. 2, pp. 89–99, 1993.

[10] S. de Rose, P. Meyer, and F. Bertrand, "Human body shapes anomaly detection and classification using persistent homology," *Algorithms*, vol. 16, no. 3, 2023.

[11] X. Li, G. Li, M. Li, K. Liu, and P. Mitrouchev, "3d human body modeling with orthogonal human mask image based on multi-channel swin transformer architecture," *Image and Vision Computing*, vol. 137, p. 104795, 2023.

[12] P. Cheng, D. Chen, and J. Wang, "Clustering of the body shape of the adult male by using principal component analysis and genetic algorithm–bp neural network," *Soft Comput.*, vol. 24, p. 13219–13237, sep 2020.