Deep Learning Project
Multitask Facial Analysis: Gender and Ethinicty Classification and Age Regression

Asia Grillo
ID: 5409650

Vincenzo Roberto Sillitti
ID: 5406466

Università Cattolica del Sacro Cuore

December 18, 2025

Outline

1. Introduction
2. Dataset
3. Model Architecture
4. First Model: Gender and Ethnicity Classification
5. Second Model: Age Regression
6. Conclusion and References

Outline

1. Introduction
2. Dataset
3. Model Architecture
4. First Model: Gender and Ethnicity Classification
5. Second Model: Age Regression
6. Conclusion and References

Project Goal

In this project we have implemented two models:

- **Multitask Classification Model** for predicting both ethnicity and gender;
- **Regression Model** specifically designed to estimate age.

Goal

The goal of our project is to predict gender, ethnicity, and age from facial images using a modified version of ResNet-18, specifically adapted to the properties and constraints of our dataset.

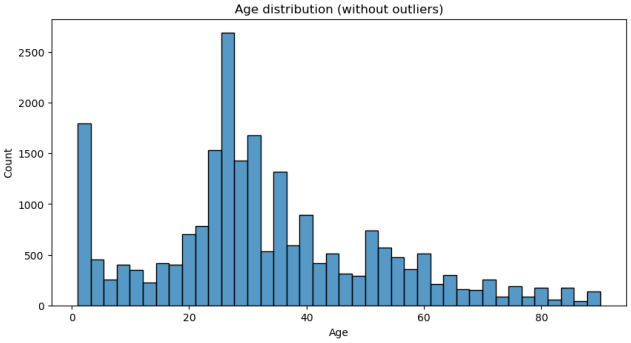
Outline

1. Introduction
2. Dataset
3. Model Architecture
4. First Model: Gender and Ethnicity Classification
5. Second Model: Age Regression
6. Conclusion and References

Data Description

The dataset is provided in CSV format, where each row corresponds to a single facial image and its labels. Images are stored as flattened pixel arrays, allowing reconstruction. The dataset consists of **23,705**[1] grayscale images of resolution 48×48 **pixels**.

- **Age range:** 1–116 years
- **Gender:**
  - 0 = Male (12,391)
  - 1 = Female (11,314)
- **Ethnicity:**
  - 0 = White (10,078)
  - 1 = Black (4,526)
  - 2 = Asian (3,434)
  - 3 = Indian (3,975)
  - 4 = Other (1,692)



[1]Since duplicates were present, 957 images were removed, identified through exact matches detected via pixel-array hashing and near-duplicates detected through cosine similarity between image vectors.

Data Preprocessing

Split into **80% training**, **10% validation**, **10% test**. All images are converted to $1 \times 48 \times 48$ tensors (grayscale). Pixel values standardised to the range $[-1, 1]$.

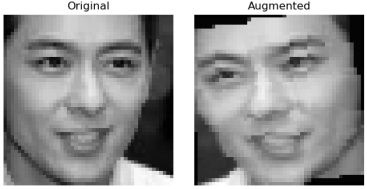
Data Augmentation

Base augmentations to all training images:

- **RandomCrop** 48 with padding = 3;
- **RandomHorizontalFlip** with 50% probability;
- **RandomRotation** by a random angle in the range $\pm 8°$;
- **RandomAffine** with translation $\leq$ 3% and scaling between 0.95 and 1.05;
- **RandomErasing** with 10% probability covering 2% -10% of the image.

Extra Augmentations (Minority Classes)

- **RandomResizedCrop** (scale = (0.85, 1.0));
- **ColorJitter** (brightness = 0.08, contrast = 0.1);
- **RandomRotation** in the range $\pm 10°$;
- **RandomErasing** with 30% probability.

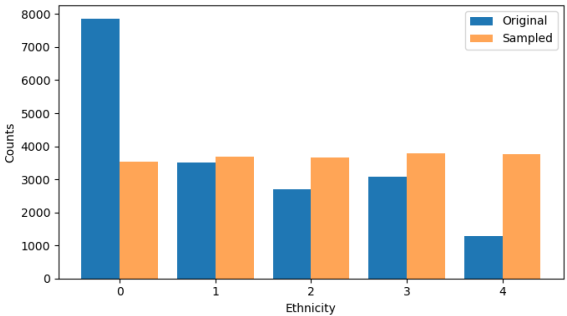


Weighted Random Sampler

It is applied to assign higher sampling probabilities to minority classes and lower probabilities to majority classes, to address the strong imbalance in the ethnicity classes.

The weighted random sampler ensures that each batch contains a balanced mix of ethnicity classes, regardless of the original distribution.

The **weights** are computed as the **inverse frequency** of each class. Then, the sampler draws the items according to the weights.



Outline

1. Introduction
2. Dataset
3. Model Architecture
4. First Model: Gender and Ethnicity Classification
5. Second Model: Age Regression
6. Conclusion and References

Model Architecture

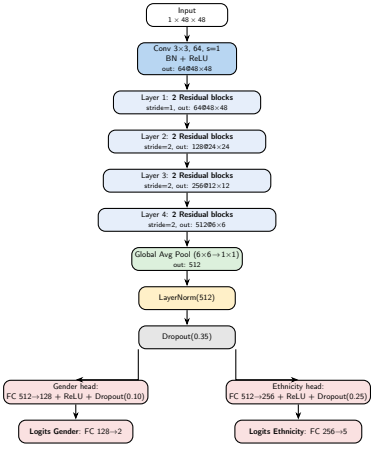
Why ResNet18? Robust learning, balanced complexity and generalization to noise.

Multitask ResNet18

A modified ResNet18 was used for 48×48 grayscale images, producing two outputs (ethnicity and gender) from a shared backbone.

Key architectural changes:

- Input layer adapted to 1 channel (grayscale);
- Initial max-pooling removed to preserve spatial detail;
- Convolution kernels adjusted for small images;
- LayerNorm + Dropout added to improve generalization;
- Two task-specific heads for multitask prediction (gender and ethnicity).



Note:
– Residual block = 2×(Conv3×3 + BN + ReLU) with a skip connection. A 1×1 conv is added when stride >1 or when the number of channels changes.

Model Architecture

ReLU activations,

$$\mathrm{ReLU}(x) = \max(0, x),$$

can reduce activation variance and destabilize deep training. Kaiming (He) initialization ensures stable variance:

$$\mathrm{Var}(w) = \frac{2}{\text{fan\_in}}, \qquad w \sim \mathcal{N}\left(0, \tfrac{2}{\text{fan\_in}}\right)$$

with

$$\text{fan\_in} = \text{input channels} \times \text{kernel height} \times \text{kernel width}.$$

Residual learning:

$$y = F(x) + x$$

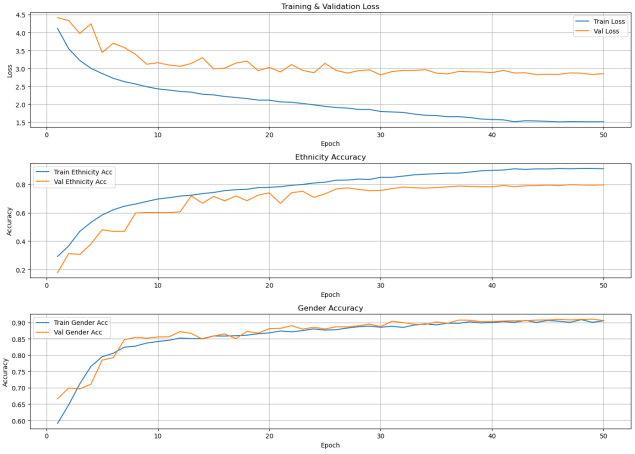
requires compatible activation scales in both branches, which Kaiming initialization preserves, keeping F(x) and x on comparable scales.

Outline

1. Introduction
2. Dataset
3. Model Architecture
4. First Model: Gender and Ethnicity Classification
5. Second Model: Age Regression
6. Conclusion and References

Training

- **Epochs**: 50
- **Optimizer**: AdamW
- **Dropout**: 0.35
- **Loss function**:
  $L = w_{gen} CE_{gen} + w_{eth} CE_{eth}$ [2]
- **Learning Rate**: 6×10^{-4}
- **Weight Decay**: 1×10^{-3}
- **Warmup Epochs**: 5
- **Cosine Scheduler**: True
- **Ethnicity Label Smoothing**: 0.12
- **Early stopping**[3]

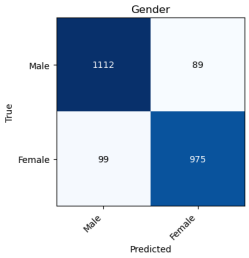


[2] Where the weights are: $w_{gen} = 0.4$ and $w_{eth} = 2.5$.

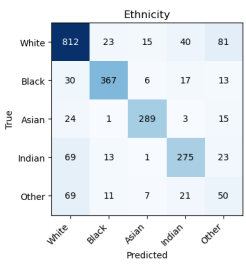
[3] With patience = 8, based on a combined score of ethnicity and gender macro-F1.

Model Performance

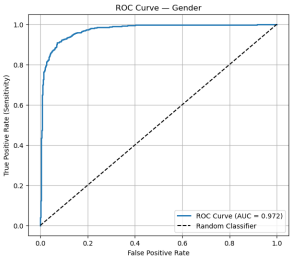
Confusion Matrix for Gender



Confusion Matrix for Ethnicity

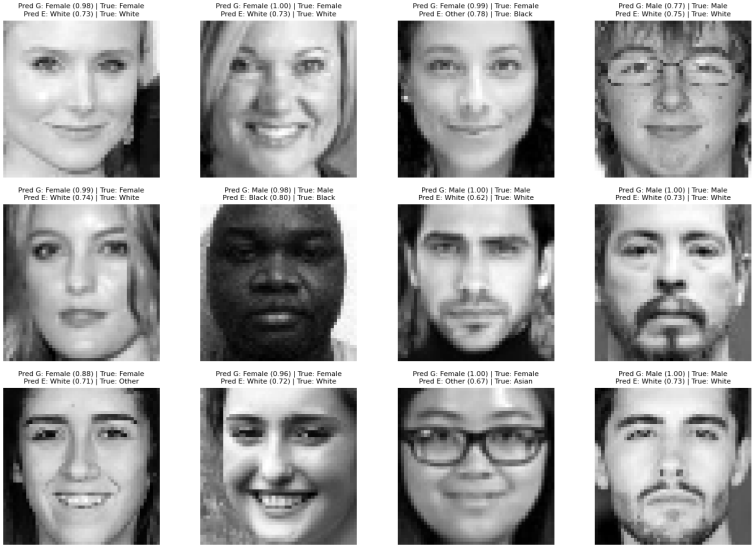


ROC Curve for Gender



- Classification for **gender** achieves an accuracy of **91.7%** and a macro-F1 score of **0.917**. The AUC value is over **0.97**.
- Classification for **ethnicity**, more challenging, has an accuracy of **79%** and a macro-F1 score of **0.723**.

Prediction on Test Set

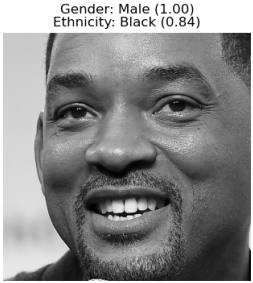


Grad-CAM

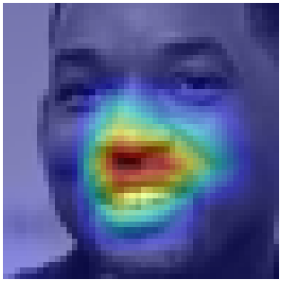
Gradient-weighted Class Activation Mapping

Grad-CAM highlights the regions of the input image that contribute most to the model's decision, allowing us to visually inspect whether the network focuses on meaningful facial features or on irrelevant artifacts.

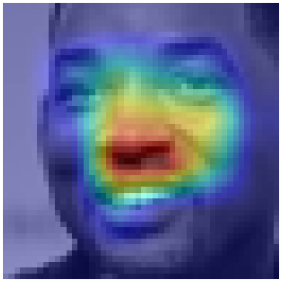
Prediction on real photo



Gender



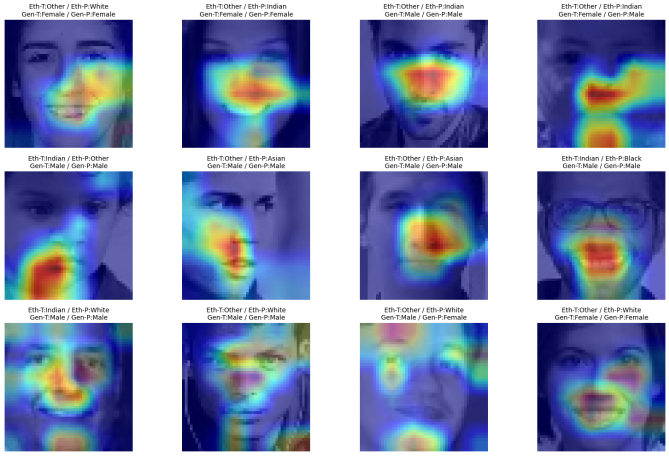
Ethnicity



Grad-CAM

We applied Grad-CAM also to test-set images, focusing specifically on **misclassified samples** belonging to the Indian and Other ethnicity categories.

If the heatmaps highlight inconsistent or irrelevant regions this suggests that the model has not learned a stable, discriminative pattern for these categories.



Error clustering

We also performed an **error clustering analysis** focused on the ethnicity prediction task to uncover patterns in the types of errors the model makes and identify systematic failure modes.

Error clustering works by extracting the internal embeddings (feature representations) of the misclassified samples and applying an unsupervised clustering algorithm: **k-means** with 3 clusters.

- **Cluster 0**: Mixed ethnicities, uniform lighting, smooth appearance;
- **Cluster 1**: Medium-tone faces, soft contrast, many Indian misclassifications;
- **Cluster 2**: Darker faces, high contrast, and challenging poses.

Outline

1. Introduction
2. Dataset
3. Model Architecture
4. First Model: Gender and Ethnicity Classification
5. Second Model: Age Regression
6. Conclusion and References

Age Regression Model

To predict the exact age, a dedicated regression pipeline is introduced to handle the strong imbalance in the age distribution.

Two-stage approach

1. **Coarse classification**
   The ages are grouped into **6 bins**[4] with approximately equal sample sizes; images are assigned to an age bin using a classifier that reuses the pretrained multitask classification backbone.
2. **Fine regression**
   A regression model is applied inside each bin to refine the prediction and get as close as possible to the true numerical age.

We applied the same data augmentation strategy used for the previous model.

[4] Bins are created as follow: Bin 0 = [0,16], bin 1 = [16,26], Bin 2 = [26,29], Bin 3 = [29,37], Bin 4 = [37,53], Bin 5 = [53,90].

Coarse Classifier

Coarse classifier training:

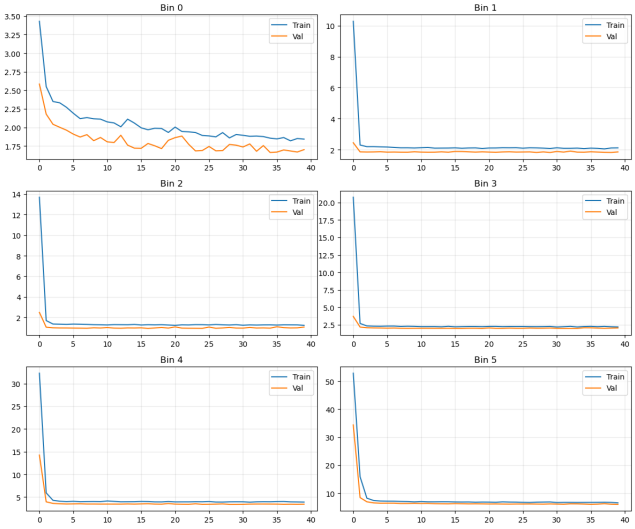
- **Head-only training (train only final layers)**: the pretrained multitask backbone already learned rich features (faces, structure, and so on);
- **Full fine tuning (entire network)**: after the classifier head learns to use the backbone features, we unfreeze the entire model to extract features specifically to predict age.

The coarse classification model achieves an accuracy and an F1 of approximately **0.58** on both validation and test sets. Given the difficulty of the task, the strong imbalance in the age distribution, and the low resolution (48×48 grayscale), this performance is reasonable. However, misclassifications remain common, especially in adjacent bins where visual differences are subtle, but the coarse classifier provides a solid foundation for the second-stage regression.

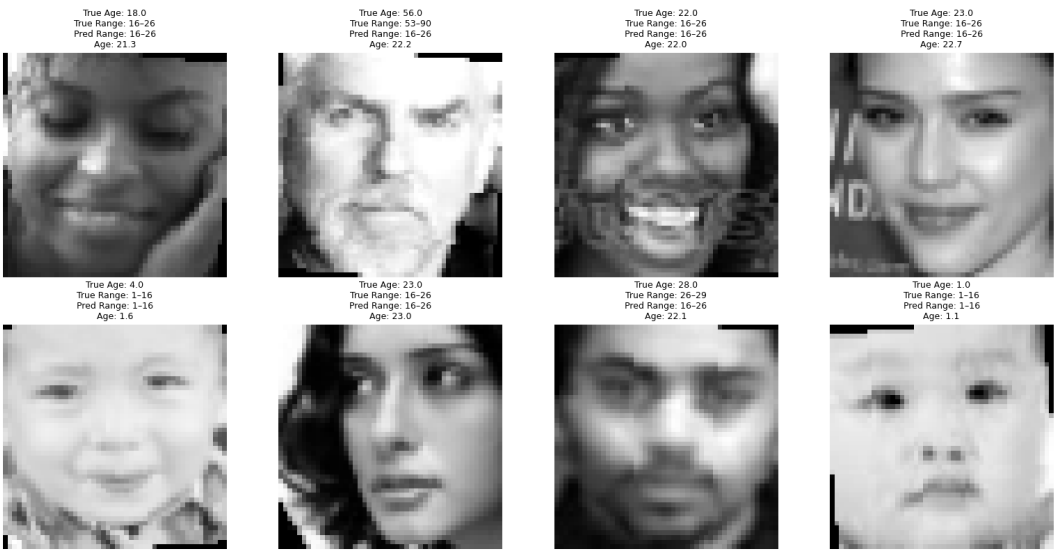
Bins Regression

After determining the age range, we train a separate regression model inside each bin.

- The model reuses the pretrained CNN backbone learned during coarse-bin classification. The backbone is frozen, providing stable 512-dimensional facial embeddings.
- Six independent regressors are trained, one for each coarse age bin. Each regressor adds a small fully connected head on top of the shared backbone (each regressor is trained only on images from its own bin).
- Only the regression heads are updated. Coarse classification learns global age features; bin-specific regressors refine age estimation locally.



Prediction on Test Set



Outline

1. Introduction
2. Dataset
3. Model Architecture
4. First Model: Gender and Ethnicity Classification
5. Second Model: Age Regression
6. Conclusion and References

Conclusion

The first model achieved solid accuracy on both gender and ethnicity, performing reliably even in the presence of several dataset limitations.

For age estimation the final regression model achieved a global MAE of roughly 3–4 years, which is competitive considering the quality of the images and the limited sample diversity.

In conclusion, this analysis achieved reasonably good results given the limitations of the dataset. However, using higher-quality images (RGB, better resolution, less noise) and a more balanced distribution of samples across ethnicity and age would likely lead to significantly better performance.

References

- Age, Gender and Ethnicity Face Data (CSV). (2020). *Kaggle Data Repository*.
- Amit, H. (2024). A guide to multi-task learning in machine learning. *Medium*.
- GeeksforGeeks. (2025). What is Fine-Tuning in Deep Learning?
- Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: An overview. *CRIF S.p.A., University of Bologna*.
- Hughes, C. (2022). Demystifying PyTorch's WeightedRandomSampler by example. *Medium*.
- Kumar, T., Mileo, A., Brennan, R., & Bendechache, M. (2023). Image data augmentation approaches: A comprehensive survey and future directions. *arXiv preprint*.
- Liu, N., Zhang, F., & Duan, F. (2020). Facial age estimation using a multi-task network combining classification and regression. *IEEE Access*, 8, 96365–96375.
- Van Otten, N. (2024). Cosine annealing in machine learning simplified: Understand how it works. *Spot Intelligence*.
- Wong, W. (2019). What is label smoothing? *Towards Data Science*.
- Wynn, F. (2024). The complete guide to data augmentation for computer vision. *Encord*.
- Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). *Dive into Deep Learning*.