**Age Prediction Using Deep Learning**

**Author**: Purnesh Mani Pandey  
**Company**: Nullclass Edtech Pvt Ltd  
**Batch Size**: 32  
**Epochs**: 10  
**Train-Test Split**: 80:20  
**References**: [IMDB-WIKI Dataset](https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/)

*1. Introduction*

Age estimation from facial images is a crucial task in computer vision with applications in security, healthcare, entertainment, and social media. Accurate age prediction can be beneficial for targeted advertisements, identity verification, and age-based content recommendations. However, predicting age from images is inherently challenging due to variations in facial expressions, lighting conditions, ethnic diversity, and age-related features such as wrinkles and skin texture.

In this project, we have developed a deep learning model to predict a person’s age from images using the **IMDB-WIKI dataset**. The dataset consists of images labeled with ages and associated metadata. Our goal is to build a convolutional neural network (CNN) model that generalizes well across different age groups while minimizing prediction errors.

***2. Dataset Description***

**2.1. IMDB-WIKI Dataset**

The dataset consists of images collected from IMDB and Wikipedia, primarily featuring celebrities. Each image is labeled with the individual's birth year, and the dataset includes metadata such as gender, name, and year the image was taken.

**Key Details of the Dataset:**

* **Total Images**: Over 500,000
* **Age Range**: 0 - 100 years
* **Metadata**: Name, birth year, gender, and year photo was taken
* **Data Format**: Images with corresponding age labels

**2.2. Data Preprocessing**

To ensure high-quality data for training, the dataset was preprocessed with the following steps:

1. **Face Detection & Filtering**
   * The dataset includes a **face\_score** metric, indicating the reliability of face detection.
   * We **filtered out** images with a face score **below 2.5** to remove poorly detected faces.
2. **Age Calculation**
   * The dataset provides birth year and photo year. The actual age is calculated as: Age=Photo Taken Year−Birth Year\text{Age} = \text{Photo Taken Year} - \text{Birth Year}Age=Photo Taken Year−Birth Year
3. **Resizing & Normalization**
   * All images were resized to **128\*128 pixels** for model consistency.
4. **Train-Test Split**
   * **80% Training Data**
   * **20% Test Data**

**3. Exploratory Data Analysis (EDA)**

**Before training the model, a detailed exploratory data analysis (EDA) was conducted to understand the dataset's characteristics, including gender distribution, age distribution, and overall data quality.**

**3.1. Gender Distribution**

**Exploratory Data Analysis (EDA)**

Before training the model, a detailed exploratory data analysis (EDA) was conducted to understand the dataset's characteristics, including gender distribution, age distribution, and overall data quality.

**3.1. Gender Distribution**

The dataset contains more male images than female images, with the distribution as follows:

* **Male:** 72%
* **Female:** 27%

This gender imbalance means that the model might be slightly biased towards male facial features, leading to better predictions for men compared to women.

**3.2. Age Distribution**

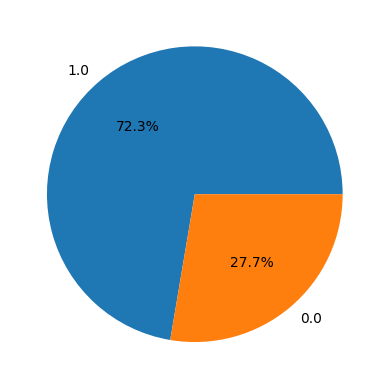
The dataset primarily consists of individuals aged **20 to 100 years**, with the highest concentration of data points in the **20 to 40-year age range**. This distribution affects model training because the model learns better from age groups with more data, potentially leading to lower accuracy for older age groups (above 60 years), where the number of samples is lower.

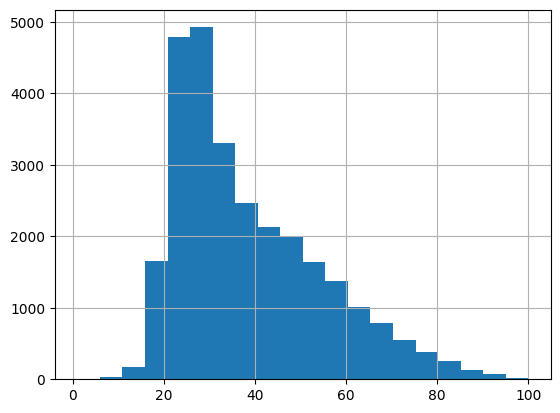
**3.3. Observations**

* **The dataset does not contain images for individuals younger than 20 years.**
* **There was no explicit removal of outliers.**
* **More images exist for people aged 20 to 40, leading to higher accuracy for these age groups.**
* **The dataset contains variations in lighting, poses, and facial expressions, making the task more challenging.**

**The dataset contains more male images than female images, with the distribution as follows:**

* **Male: 72%**
* **Female: 27%**

****

****

**3.3. Observations**

* **The dataset does not contain images for individuals younger than 20 years.**
* **There was no explicit removal of outliers.**

***3. System Configuration***

The project was developed and trained on a **CPU-based system**, which introduced computational constraints.

**Hardware Specifications**

| **Component** | **Specification** |
| --- | --- |
| **Processor** | Intel i3 |
| **RAM** | 4GB |
| **GPU** | No GPU used (CPU-based training) |

Since no dedicated GPU was available, training was computationally expensive and required **longer processing times**.

***4. Model Architecture & Training Details***

The deep learning model used for age prediction is a **Convolutional Neural Network (CNN)**.

**4.1. Model Architecture**

The model follows a **CNN-based regression approach**, designed to extract age-related features from facial images and predict continuous age values.

**Layer Structure:**

1. **Convolutional Layers**
   * Extract features like edges, wrinkles, and facial structures.
   * Uses **ReLU activation** for non-linearity.
2. **Batch Normalization**
   * Normalizes activations to speed up training.
3. **MaxPooling Layers**
   * Reduces feature map dimensions while retaining essential features.
4. **Fully Connected Layers**
   * Maps extracted features to age predictions.
5. **Output Layer**
   * Single neuron with **linear activation** (since age is a continuous variable).

**4.2. Training Configuration**

| **Parameter** | **Value** |
| --- | --- |
| **Batch Size** | 32 |
| **Epochs** | 10 |
| **Optimizer** | Adam |
| **Loss Function** | Mean Squared Error (MSE) |
| **Train-Test Split** | 80:20 |

***5. Performance Evaluation***

The model's performance was evaluated using **two primary metrics**:

**5.1. R² Score (Coefficient of Determination)**

* **R² Score**: **57.8%**
* Measures how well the model explains variance in age.

**5.2. Mean Absolute Error (MAE)**

* **MAE**: **7.8 years**
* Average absolute error in predicted ages.

***6. Results & Analysis***

**6.1. Actual vs. Predicted Age**

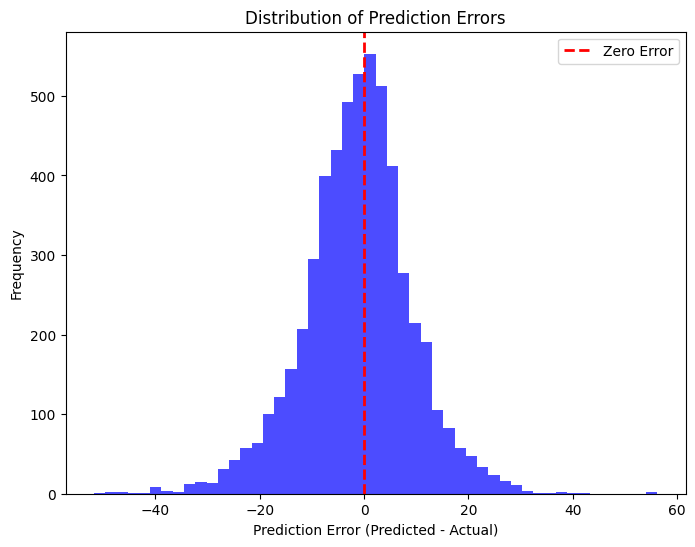
The following scatter plot compares **actual age vs. predicted age**:



* The red dashed line represents the **perfect prediction line**.
* The blue points indicate individual predictions.
* Observations:
  + Predictions **closely align** for younger age groups.
  + Larger deviations for older individuals.

**6.2. Distribution of Prediction Errors**

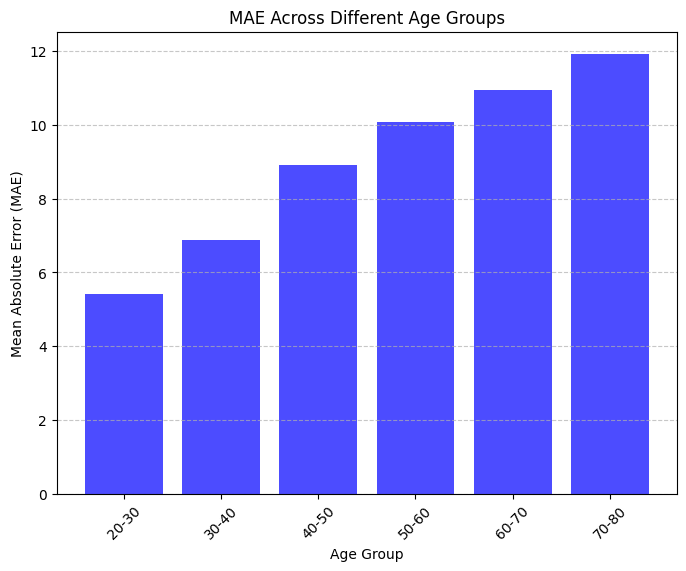
The histogram below visualizes the **error distribution**:



* The error distribution is roughly **Gaussian** (centered around zero).
* However, a few extreme errors are present, especially for **older age groups**.

**6.3. MAE Across Different Age Groups**

To understand prediction accuracy across different age brackets:



* **Lowest MAE**: **20-30 years**
* **Highest MAE**: **70-80 years**
* Prediction errors **increase with age**, highlighting challenges in estimating older individuals.

***7. Challenges & Potential Improvements***

**7.1. Key Challenges**

1. **Limited computational resources**
   * No GPU available → Longer training times.
2. **Higher error for older age groups**
   * Likely due to **dataset imbalance** (fewer older individuals).
3. **Dataset noise**
   * Some images may contain incorrect labels, affecting accuracy.

**7.2. Potential Improvements**

1. **Loss Reweighting**
   * Assign **higher weight** to older individuals to improve accuracy.
2. **Data Augmentation**
   * Increase older age group representation.
3. **Experimenting with Different Loss Functions**
   * **Huber Loss** or **Quantile Loss** may improve robustness.

***8. Conclusion***

This project successfully implemented a **deep learning-based age prediction model** using the IMDB-WIKI dataset. Despite hardware limitations, the model achieved a **57.8% R² score** and a **7.8-year MAE**.

**Key Findings:**

* Predictions are **more accurate for younger individuals**.
* Errors **increase for older individuals**, likely due to dataset imbalance.
* Using **data augmentation and loss function tuning** may further enhance performance.

**Future Work:**

* **Fine-tuning the model** for better generalization.
* **Using a GPU-based system** for efficient training.
* **Exploring transformer-based architectures (ViTs)** for improved feature extraction.