UCL
SCHOOL OF
MANAGEMENT

Module code/name	MSIN0097 Predictive Analytics
Module leader name	Dr A P Moore
Academic year	2024/25
Term	2
Assessment title	Individual Coursework
Individual/group assessment	Individual
Candidate Number	MYBQ6
Word Count	1932

Table of Contents

1. Introduction

Demographic prediction from facial images remains a key challenge in modern computer vision, requiring advanced machine learning techniques to achieve accuracy and fairness. This study investigates whether a Convolutional Neural Network (CNN) can accurately predict age, gender, and geographic ancestry using the FairFace dataset, and how it compares to traditional models like Random Forest in terms of accuracy and computational efficiency.

The FairFace dataset comprises 97,698 images with five key attributes: file path, age category, gender, race, and a service flag. The dataset is split into 86,744 training images and 10,954 validation images, ensuring a diverse representation across demographic groups. Sample entries include "train/1.jpg" (Male, 50-59, East Asian) and "val/4.jpg" (Female, 20-29, Latino_Hispanic), demonstrating its demographic diversity. A comprehensive data quality check confirmed no missing values, ensuring reliability for training deep learning models.

Rather than developing a single multi-output model, separate classifiers were trained for each attribute to optimize prediction accuracy. Two different approaches were implemented:

- Random Forest classifiers as baselines
- Convolutional Neural Networks (CNNs) to capture complex visual patterns in facial images

While Random Forests serve as a strong benchmark for structured data classification, CNNs excel at capturing intricate spatial patterns present in facial features. For the CNN implementation, both custom architectures and MobileNetV2 transfer learning models were explored to evaluate performance trade-offs.

Due to hardware limitations of using a MacBook M2 Air, certain compromises were necessary, including reducing the number of training epochs for CNNs and limiting decision trees in Random Forests to balance computational constraints with model performance.

Beyond technical implementation, accurate demographic classification has significant real-world applications, including:

- Social media analytics (personalized recommendations, targeted advertising)
- Security verification (identity authentication, fraud detection)
- AI-driven personalization (age-based content filtering, accessibility features)

```
Total Rows: 97698
Total Columns: 5
Training Data: 86744 rows, 5 columns
Validation Data: 10954 rows, 5 columns

Training Data Sample:
          file    age  gender        race  service_test
0  train/1.jpg  50-59    Male  East Asian          True
1  train/2.jpg  30-39  Female      Indian         False
2  train/3.jpg    3-9  Female       Black         False
3  train/4.jpg  20-29  Female      Indian          True
4  train/5.jpg  20-29  Female      Indian          True

Validation Data Sample:
        file    age  gender             race  service_test
0  val/1.jpg    3-9    Male       East Asian         False
1  val/2.jpg  50-59  Female       East Asian          True
2  val/3.jpg  30-39    Male            White          True
3  val/4.jpg  20-29  Female  Latino_Hispanic          True
4  val/5.jpg  20-29    Male  Southeast Asian         False

Column Names:
['file', 'age', 'gender', 'race', 'service_test']

Missing Values in Training Data:
file            0
age             0
gender          0
race            0
service_test    0
dtype: int64

Missing Values in Validation Data:
file            0
age             0
gender          0
race            0
service_test    0
dtype: int64

Data Types in Training Set:
file            object
age             object
gender          object
race            object
service_test      bool
dtype: object

Data Types in Validation Set:
file            object
age             object
gender          object
race            object
service_test      bool
dtype: object

Number of images in Training Folder: 86744
Number of images in Validation Folder: 10954
```

2. Data Cleaning & Pre-Processing

To ensure data quality and consistency, a structured data cleaning and preprocessing pipeline was implemented. The FairFace dataset was analysed for missing values, inconsistencies, duplicate entries, and formatting errors before being used for training. This process involved multiple steps, including label standardization, file path normalization, duplicate removal, and dataset integrity checks. Below is a breakdown of the key preprocessing steps along with their corresponding methods and outcomes.

Step	Description	Methods Applied	Outcome/Result
Dropping Unwanted Columns	Removed irrelevant columns (`service_test`) to streamline processing.	`df.drop(columns=['service_test'])`	Dataset size reduced, keeping only necessary information.
Ensuring Label Consistency	Standardized text formatting for categorical labels (`gender` , `race`).	Applied `lower()` , `strip()` , and manual corrections.	Labels are now uniform (`male/female` → `Male/Female`).
Fixing File Paths	Normalized image filenames and paths for portability.	Extracted base filename, converted to lowercase.	Filenames changed from `train/1.jpg` to `1.jpg` , making them more accessible.
Finding Missing Images	Checked for missing images in the dataset and replaced them if necessary.	`os.path.exists()` validation for each image file.	**Train:** 86,744 images found (0 missing). **Validation:** 10,954 images found (0 missing).
Detecting & Removing Duplicates	Used perceptual hashing to detect and remove duplicate images.	`imagehash` for similarity checks, deleted redundant files.	179 duplicate images removed from training, 3 from validation.
Checking for Corrupt Images	Ensured all images are valid and loadable.	Used `PIL` to attempt loading each image and flagged corrupt ones.	No corrupt images detected.
Final Dataset Validation	Verified dataset integrity, ensuring alignment between CSV entries and available images.	Counted images and cross-checked against dataset entries.	All images matched expected counts, dataset integrity confirmed.

After completing the cleaning pipeline, a final check was performed to ensure alignment between the CSV records and the actual image files. The training set contained 86,744 valid images, and the validation set contained 10,954 images, with no missing or corrupted files. These steps ensured the dataset was clean, structured, and ready for feature engineering.

3. Feature Engineering

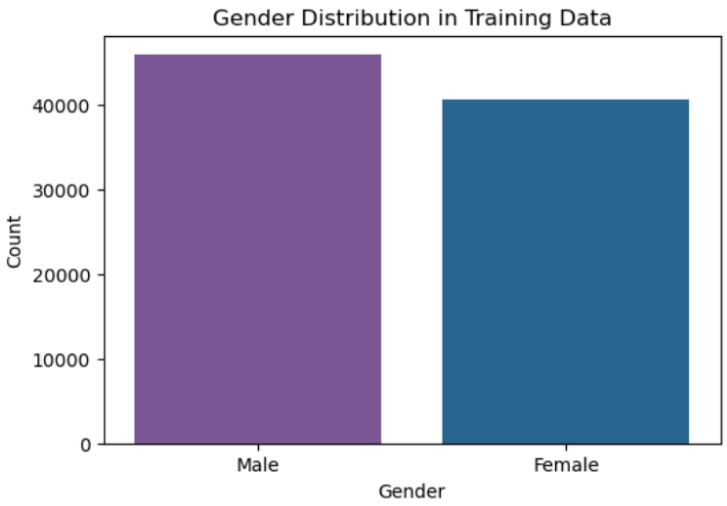
Feature engineering was an essential step in refining the dataset for machine learning models. Various transformations were applied to categorical and numerical attributes to enhance interpretability and efficiency. The table below outlines the key feature engineering steps, the methods used, and their outcomes.

Step	Description	Methods Applied	Outcome/Result
Converting Age Ranges to Midpoints	Transformed age group labels into numerical midpoints for model compatibility.	Extracted midpoints from predefined age bins.	Example: `"20–29"` → `24`.
Encoding Age Bins	Assigned numerical labels to age categories for classification tasks.	Used `LabelEncoder` to convert age bins into encoded values.	Age categories encoded into unique integers.
Encoding Gender	Converted gender labels into numerical values (`Male: 0, Female: 1`).	Used `LabelEncoder` for categorical encoding.	**Gender Encoding Applied** Unique values in 'gender' column (Train): `[0, 1]`.
Label Encoding for Geographical Ancestry	Assigned numerical labels to race categories for improved model interpretability.	Applied `LabelEncoder` on race column.	Example: `"East Asian"` → `1`, `"Black"` → `2`.
Adding Image Encoding in Dataset	Extracted grayscale pixel intensities from images to be used as features.	Loaded images, resized to `32x32`, converted to grayscale, and flattened pixel arrays.	Each image represented as a **1024-dimensional feature vector**.
Principal Component Analysis (PCA) for Dimensionality Reduction	Reduced image feature dimensions while retaining 99% variance.	Applied `PCA(n_components=300)`, reducing dimensionality from 1024 to 300.	**Before PCA:** `X_train: (86,559, 1024)`, `X_val: (10,951, 1024)`. **After PCA:** `X_train: (86,559, 300)`, `X_val: (10,951, 300)`.
Saving Final Dataset	Stored processed datasets for efficient training.	Saved structured CSV files and NumPy feature arrays for model input.	Ready-to-use **train and validation datasets** for model training.

After all feature engineering steps were applied, the dataset was fully transformed and optimized for machine learning. These enhancements significantly improved data efficiency, reduced storage requirements, and streamlined model input processing.

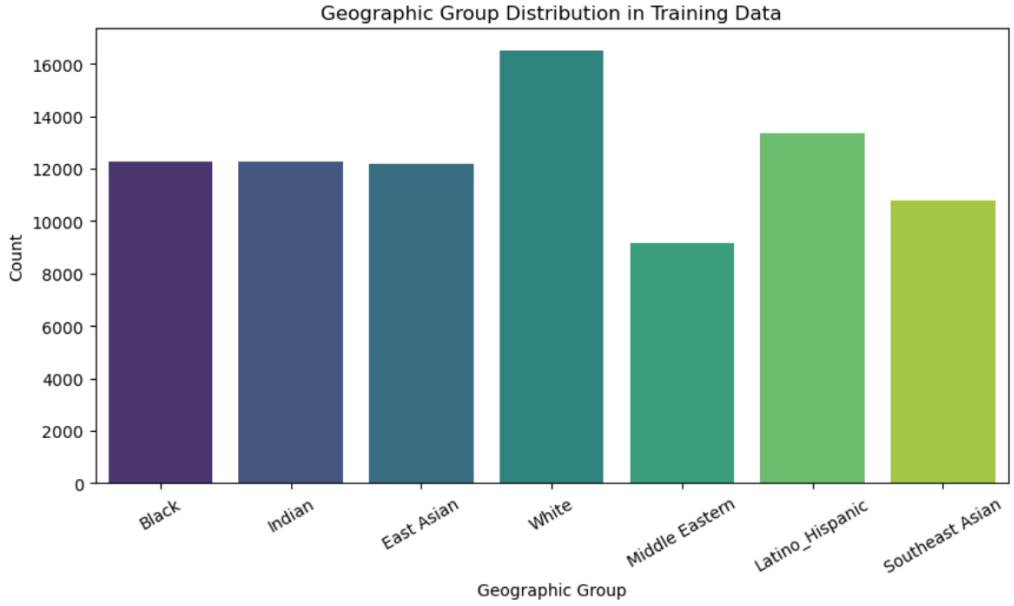
4. Exploratory Data Analysis

4.1 Gender Distribution in Training Data



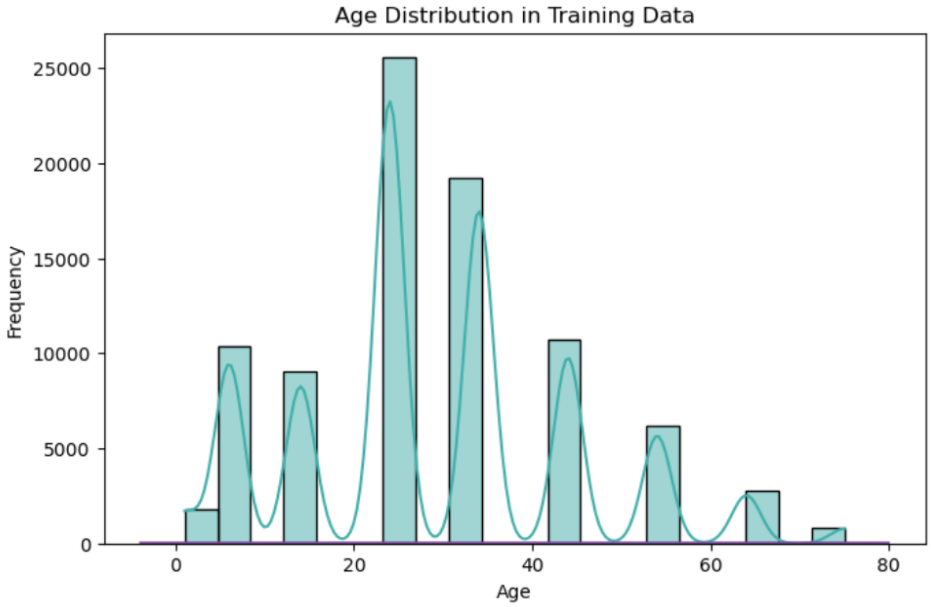
The dataset is balanced in terms of gender representation, with a slightly higher number of male images compared to female images.

4.2 Geographic Group Distribution in Training Data



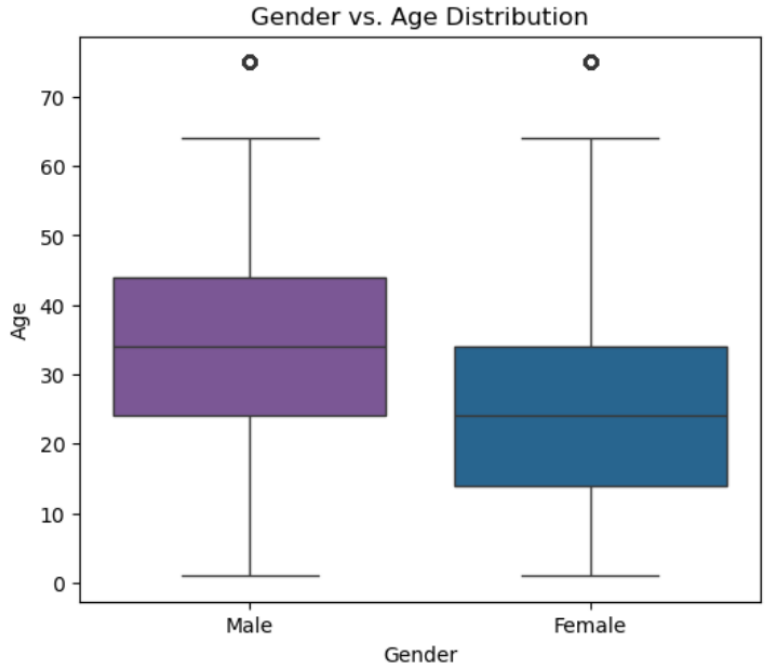
The dataset covers a diverse set of geographic ancestry categories, with the "White" group having the highest representation, while "Middle Eastern" has the lowest.

4.3 Age Distribution in Training Data



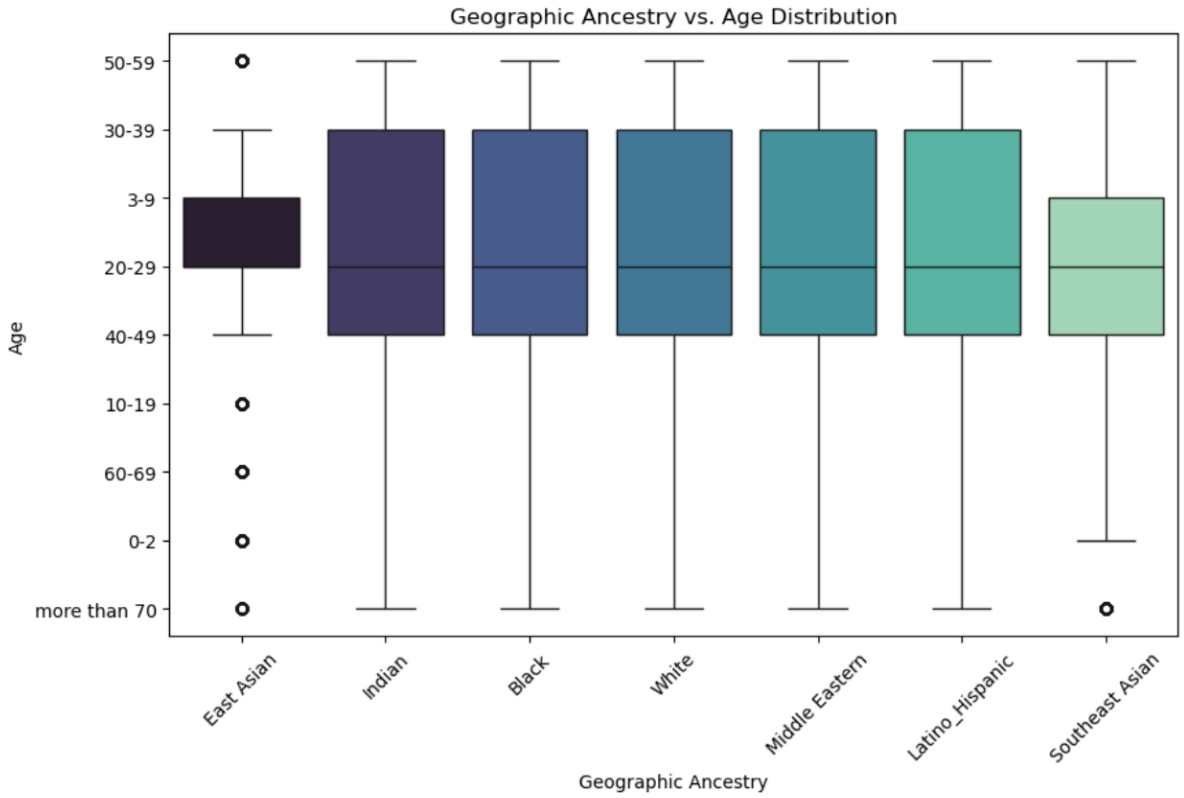
The dataset exhibits multiple peaks, indicating a skewed distribution with a higher concentration in younger age groups (especially around 20 and 30 years).

4.4 Gender vs. Age Distribution (Box Plot)



Males tend to have a slightly higher median age than females.
There are outliers in both male and female distributions, particularly in the older age ranges.

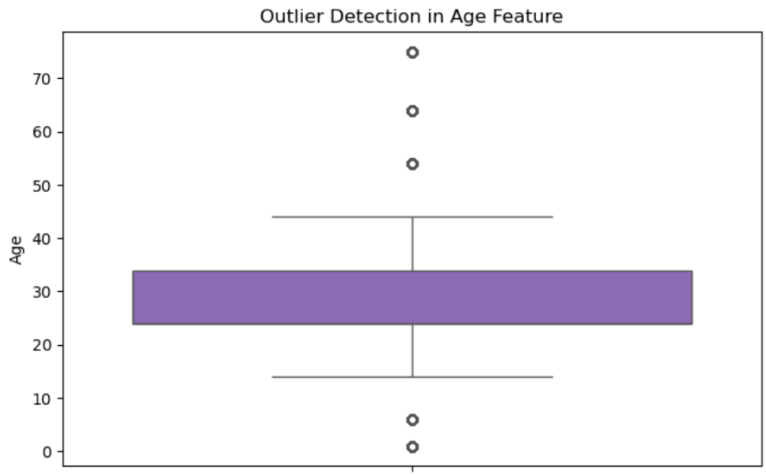
4.5 Geographic Ancestry vs. Age Distribution (Box Plot)



The age distribution across different geographic ancestries is consistent.

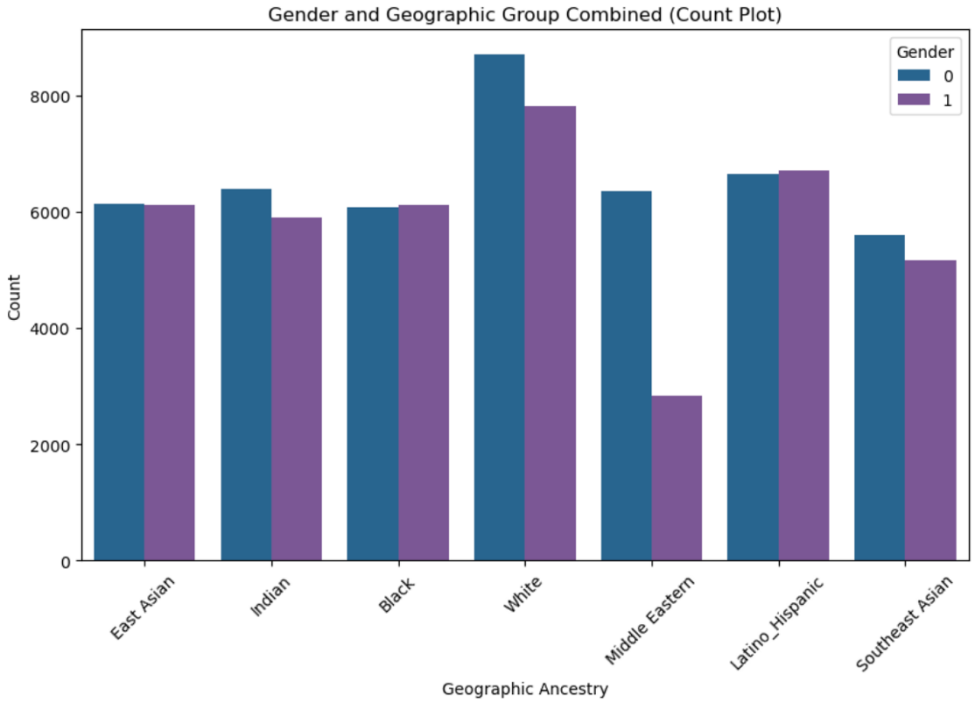
The interquartile range (IQR) is similar across groups, with some outliers in younger and older age brackets.

4.6 Outlier Detection in Age Feature



The dataset contains some extreme values in the age column, notably in the youngest (0-2) and oldest (more than 70) age groups.

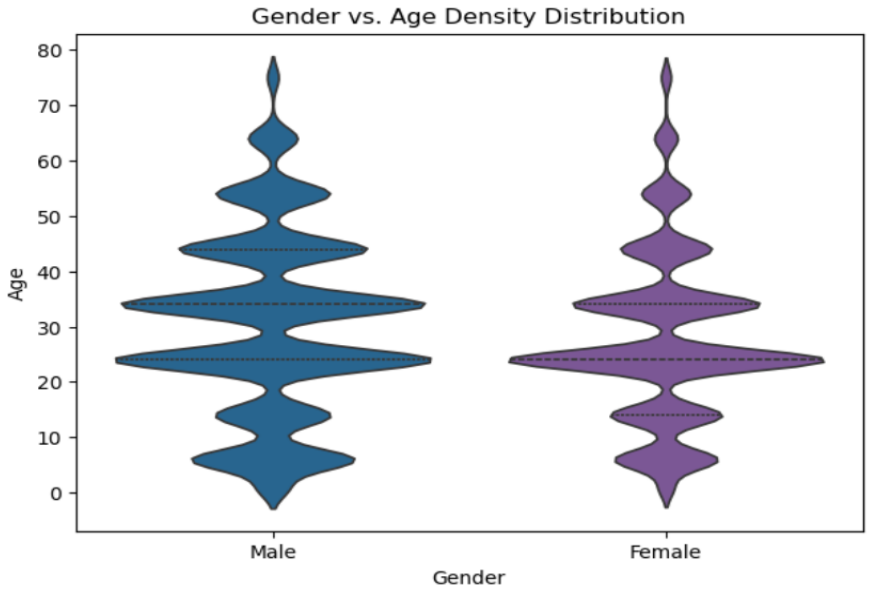
4.7 Gender and Geographic Group Combined (Count Plot)



Gender representation within geographic groups is relatively even.

White individuals have the highest representation, and Middle Eastern individuals have the lowest, with minor gender imbalances in certain groups.

4.8 Gender vs. Age Density Distribution (Violin Plot)



The age distribution for males and females follows a similar pattern, but females show a slightly denser concentration in younger age groups.

5. Traditional Baseline Model: Random Forest

5.1 Random Forest Regression to Predict Age

The first approach used regression, predicting a continuous age value instead of classifying into bins. Since the dataset contained age ranges (e.g., 30-39, 40-49, etc.), the midpoint of each range was used as the target variable.

The input consisted of PCA-transformed image embeddings, and the output was a numerical age prediction. The Random Forest Regressor was used with the following settings:

Hyperparameter	Value (Baseline Model)
n_estimators	30
max_depth	10
min_samples_split	10
Loss Function	Mean Squared Error (MSE)

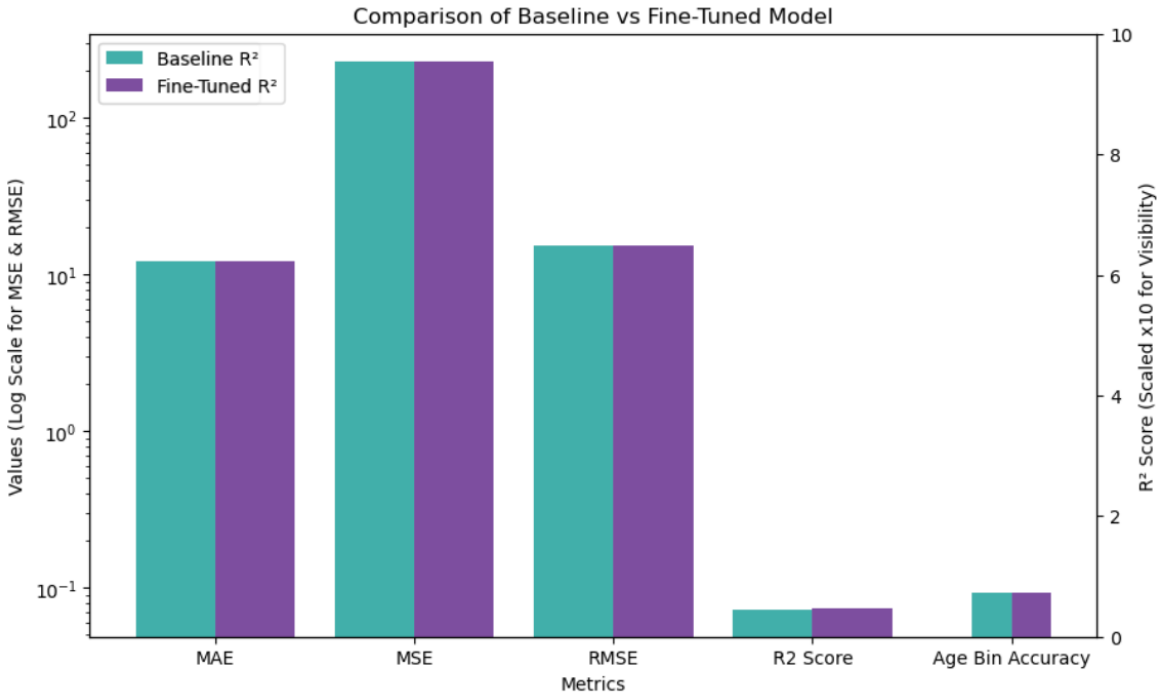
Fine-Tuning the Regression Model:

To improve performance, RandomizedSearchCV was used to optimize key parameters:

Hyperparameter	Search Space	Best Value (Fine-Tuned Model)
n_estimators	40, 50	40
max_depth	10, 15, 20	10
min_samples_split	2, 5	2
Cross-validation folds	3	3
Optimization Metric	Minimized MSE	Minimized MSE

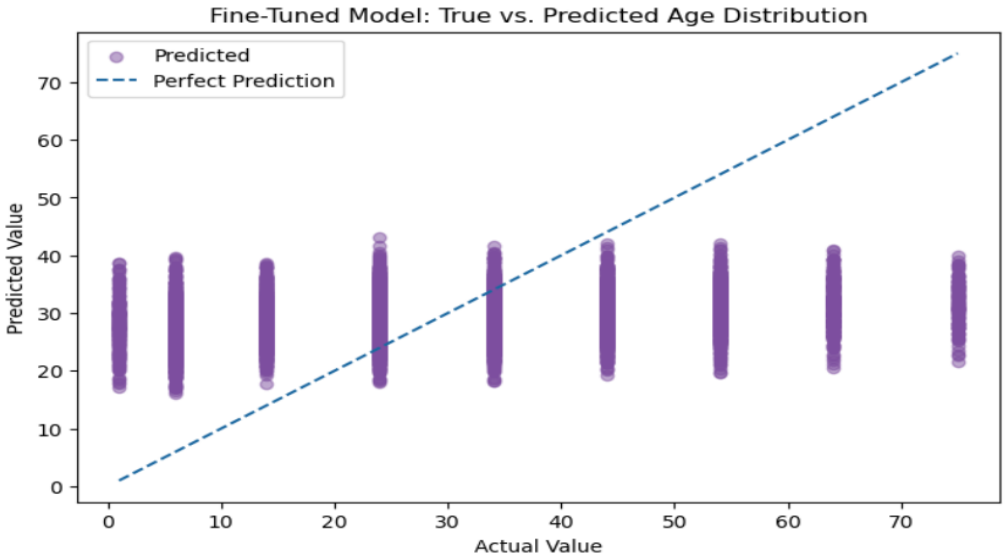
Baseline vs. Fine-Tuned Regression Model Performance:

Metric	Baseline Model	Fine-Tuned Model	Change
MAE	12.0717	12.0687	**-0.0030**
MSE	230.5436	230.2326	**-0.3110**
RMSE	15.1837	15.1734	**-0.0103**
R² Score	0.0728	0.0740	**+0.0012**
Age Bin Accuracy	24.86%	24.85%	**-0.0001**



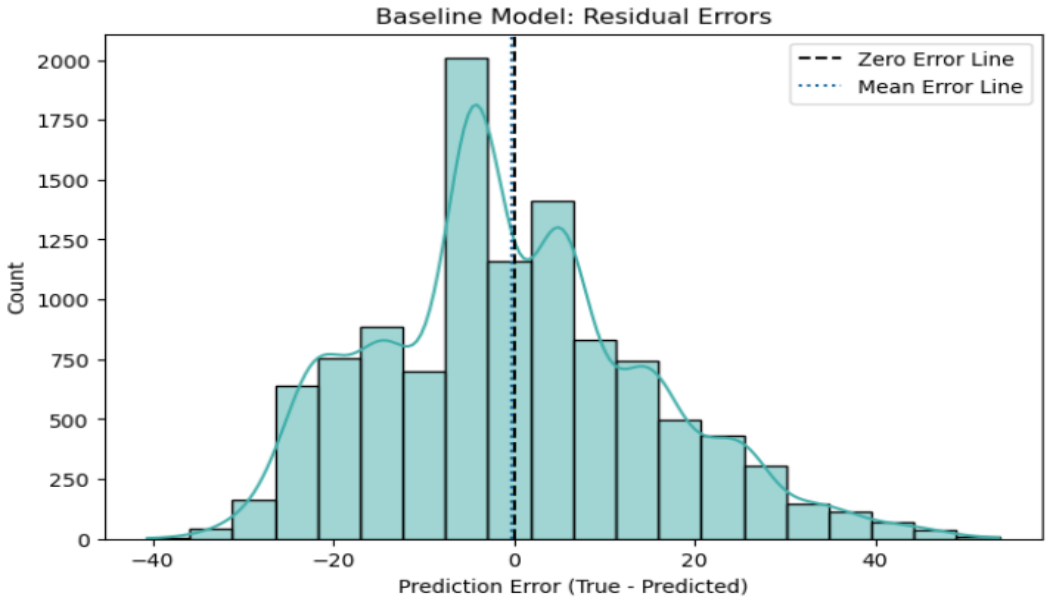
Key Takeaways from Graphs:

Actual vs Predicted Scatter Plot:



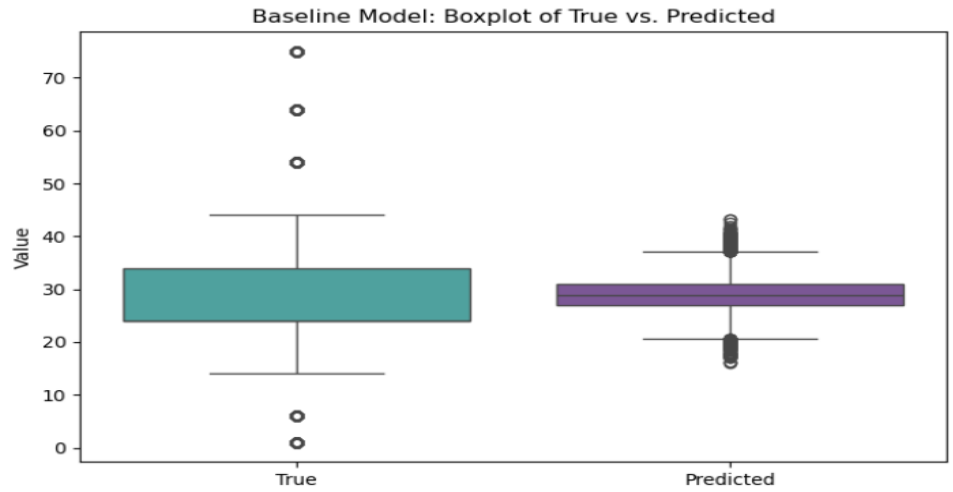
Observation	Insight
The predicted values are clustered and show gaps, meaning predictions are not well-distributed across the actual age values.	The model seems to **underfit**. It might not have enough **complexity** to capture the full range of age values, suggesting **bias** in the predictions.

Residual Error Plot:



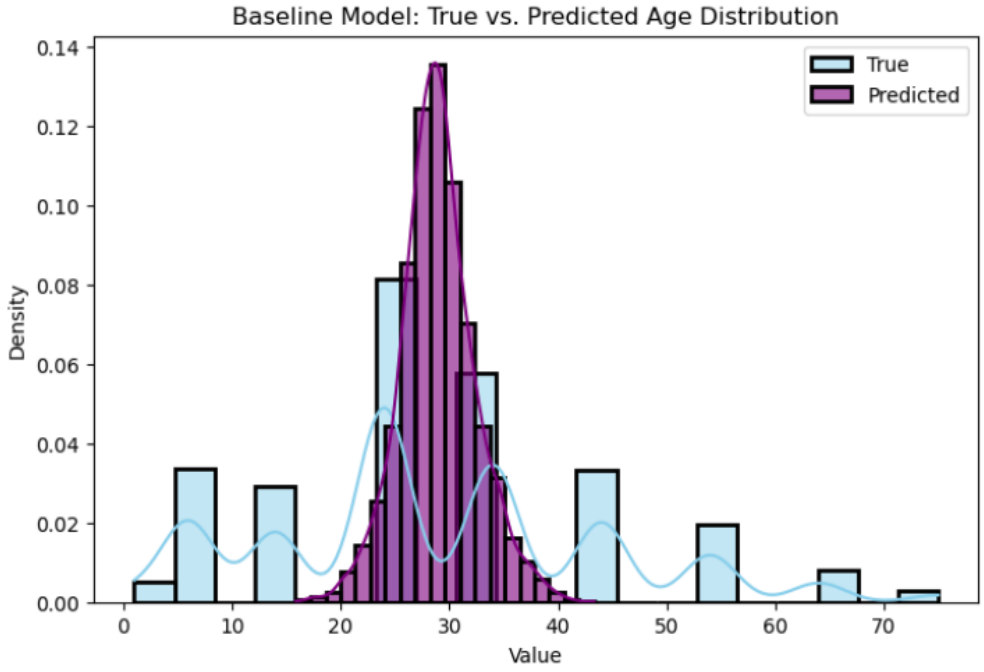
Observation	Insight
There are **outliers** in the residual errors, though most errors are concentrated around zero.	The presence of **large errors** indicates the model struggles with certain cases. The model might be **overfitting** or **not generalizing well**.

Boxplot (True vs Predicted):



Observation	Insight
The **predicted values** have a **wide spread** and show **outliers**, suggesting inconsistency in predictions.	The model is producing **extreme predictions** that are **far from the actual values**. This could be a sign of **underfitting** or **overfitting**.

Age Bin Accuracy:



Observation	Insight
The model correctly predicted the **age bin** only **24.85%** of the time, which is **low**.	The model is **struggling with classification** into age bins. Consider using a **classification approach** instead of regression for better bin predictions.

5.2 Random Forest Classification to Predict Age:

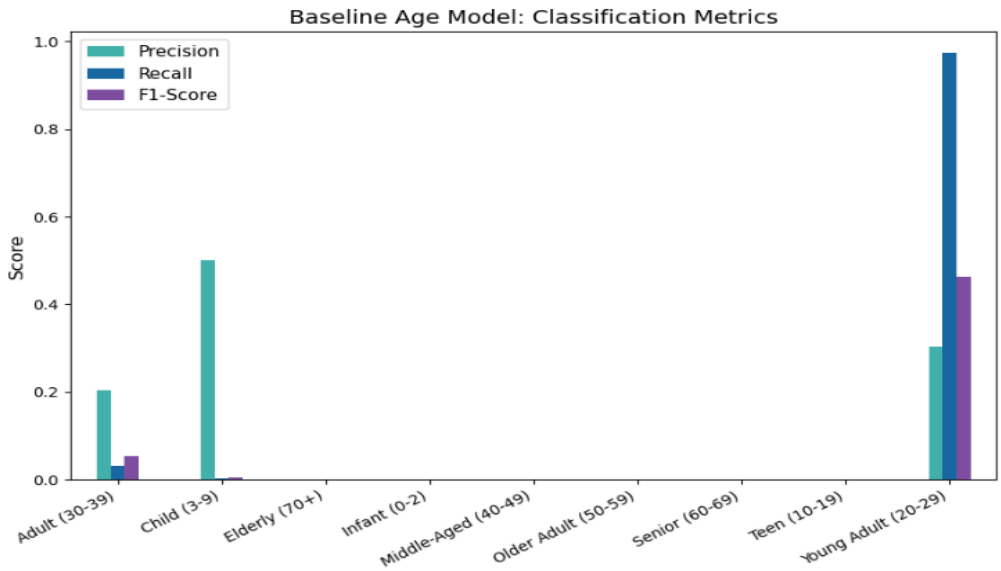
Since regression failed to predict age effectively, the problem was reframed as classification. Instead of predicting exact ages, individuals were assigned to one of nine predefined age groups based on the dataset.

The model used PCA-transformed image embeddings and encoded categorical features as input, with a categorical age bin as the target variable.

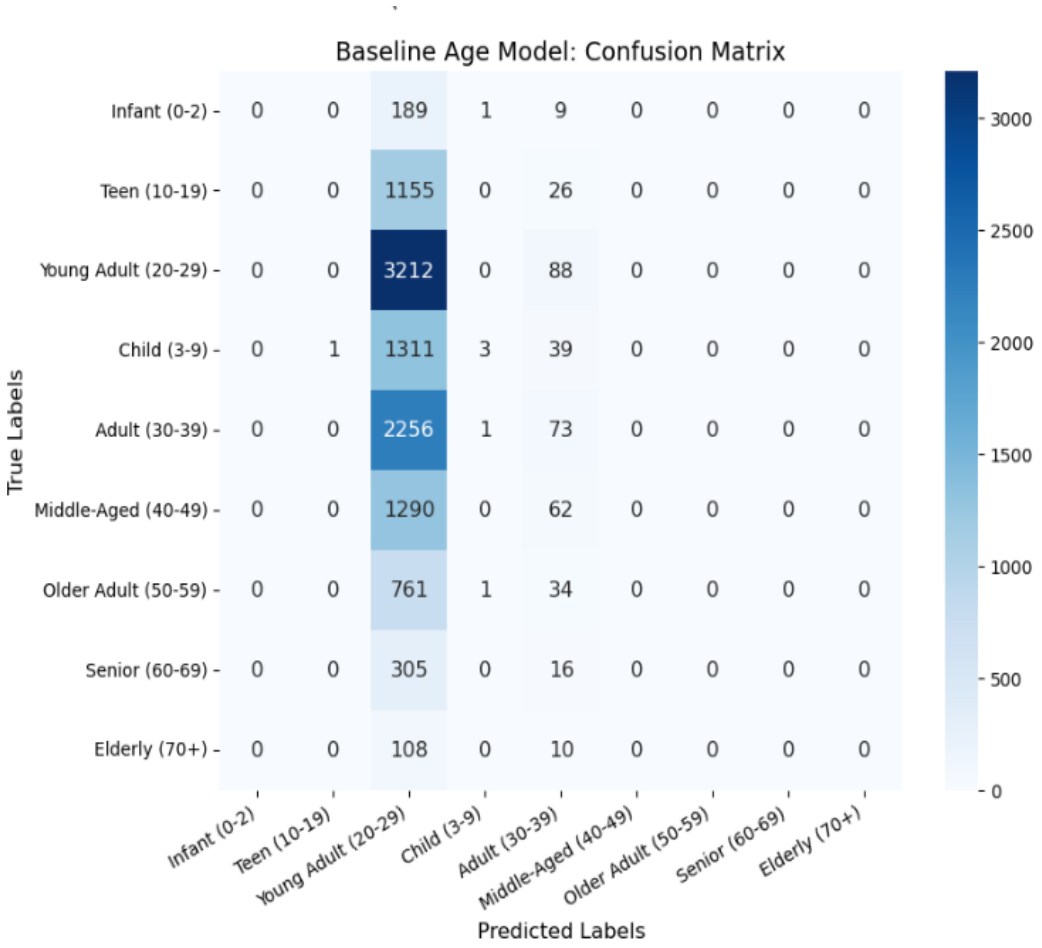
Baseline Model:

Hyperparameter	Baseline Model
n_estimators	30
max_depth	10
min_samples_split	10
Loss Function	Cross-Entropy

Baseline Model Performance:



Observation	Insight
The model performs well only in the **Young Adult (20-29)** category, with recall reaching 97.3%.	The classifier is **biased toward this category**, misclassifying many other samples into it.
Other age groups exhibit **low recall and precision**, with most close to zero.	The model fails to distinguish between classes, suggesting a lack of **informative decision boundaries**.



Observation	Insight
Most predictions are **concentrated in the Young Adult (20-29) class**, with **severe misclassification** for all other groups.	The model is **overgeneralizing** and failing to learn proper age distributions.
Older adults and infants are almost never predicted correctly.	The classifier does not capture the **distinct features of different age groups**.

Fine-Tuning with SMOTE & Hyperparameter Optimization:

To address the imbalance issue, SMOTE was applied, generating synthetic data to ensure all age groups had equal representation.

Hyperparameter Search Space for Fine-Tuning:

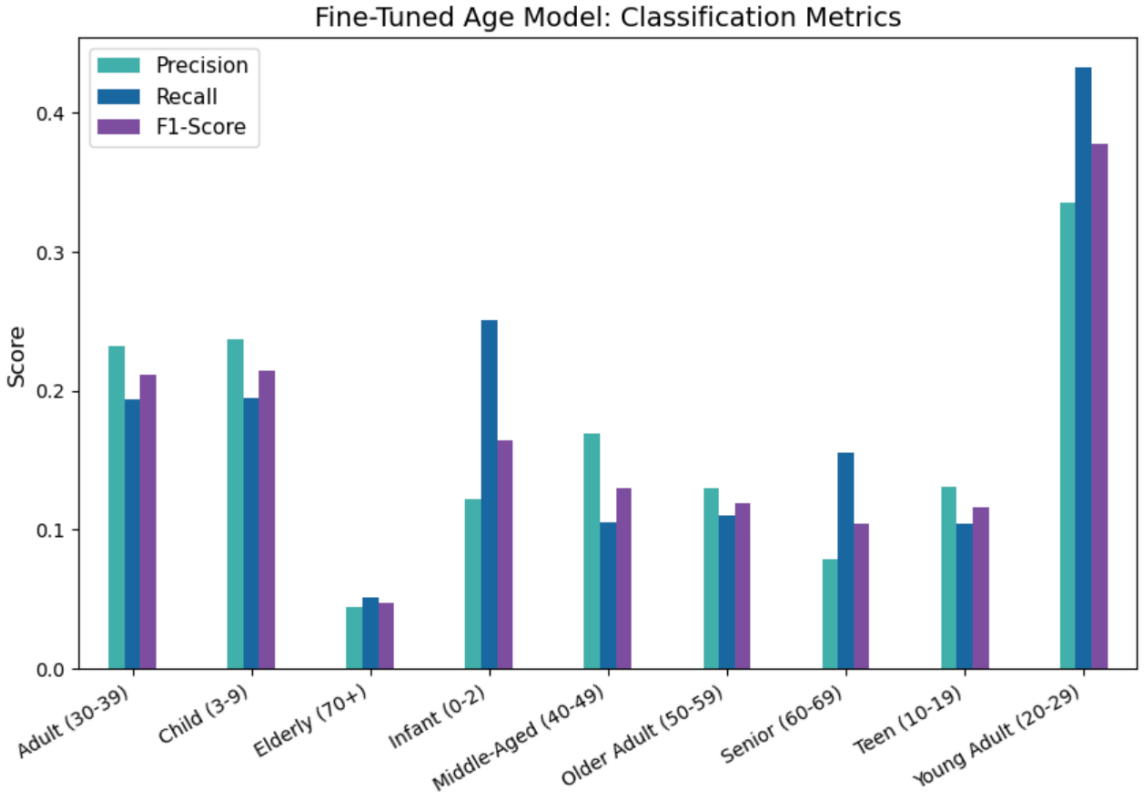
Hyperparameter	Search Space	Best Value (Fine-Tuned Model)
n_estimators	50, 100	100
max_depth	10, 15	15
min_samples_split	2, 5	2
Cross-validation folds	2	2
Optimization Metric	Accuracy	Accuracy

After tuning, the best hyperparameters found were:

- n_estimators: 100
- max_depth: 15
- min_samples_split: 2

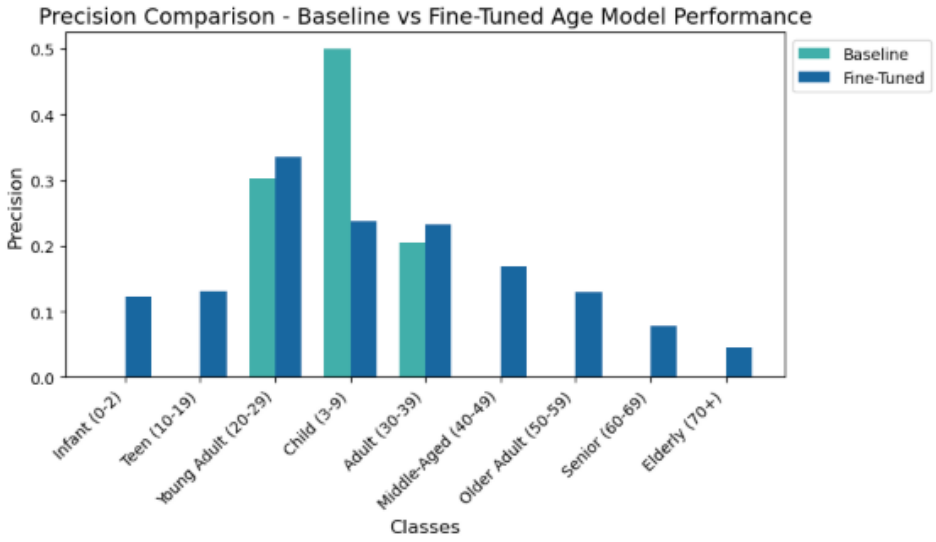
Fine-Tuned Model Performance:

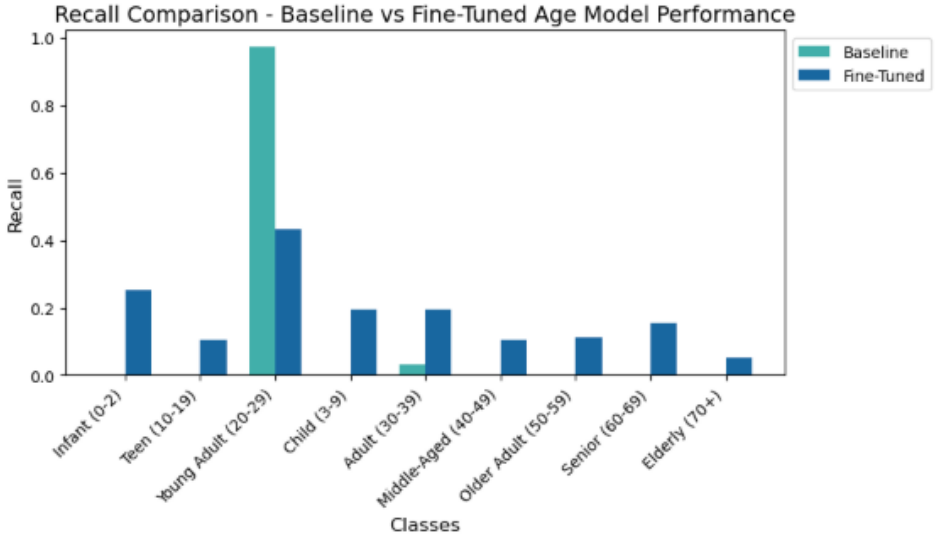
After applying SMOTE balancing and hyperparameter tuning, the model saw improvements across multiple age groups, though misclassification issues persisted.

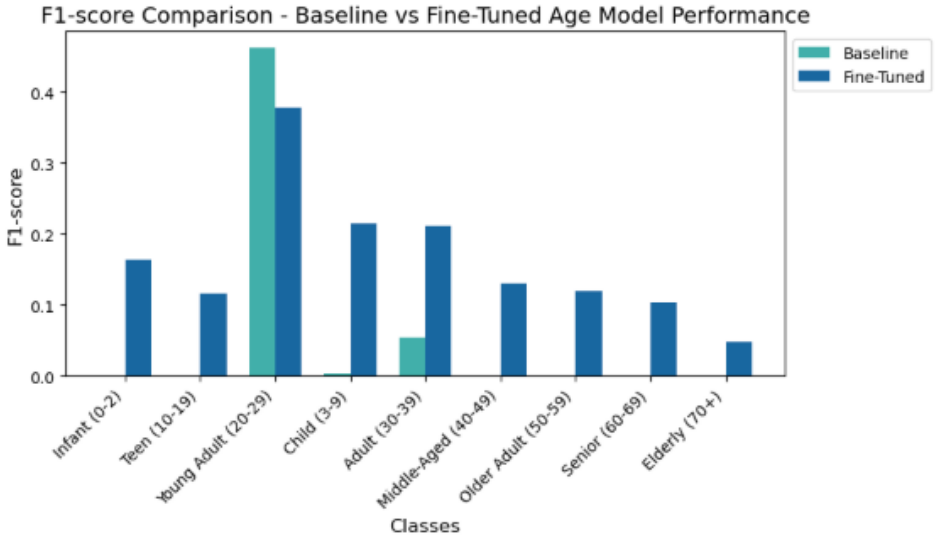


Age Group	Observation	Insight
Adult (30-39)	Precision and recall are moderately balanced.	The model identifies this group relatively well but still struggles with misclassification.
Child (3-9)	Precision is relatively high, but recall is lower.	The model can predict this group well but fails to capture all actual instances.
Elderly (70+)	All three metrics (precision, recall, F1-score) are very low.	The model struggles significantly with this group, likely due to dataset imbalance.
Infant (0-2)	Recall is higher than precision, but both are still low.	The model captures some true cases but still frequently misclassifies infants.
Middle-Aged (40-49)	Recall is significantly higher than precision.	The model correctly identifies more true cases but has a high rate of false positives.
Older Adult (50-59)	Precision is better than recall.	The model is more confident in its predictions but still misses many true cases.
Senior (60-69)	All metrics remain low.	The model does not perform well for this group, requiring further improvements.
Teen (10-19)	Low recall compared to precision.	The model is selective in predictions but misses many actual instances of teens.
Young Adult (20-29)	Highest recall and F1-score across all age groups.	The model is biased towards this group, frequently predicting young adults.

Comparison Between Baseline and Fine-Tuned Models:







Observation	Insight
Overall recall improved in the fine-tuned model	The fine-tuned model captures more true positives, reducing the number of missed classifications, particularly in underrepresented age groups.
Precision decreased slightly for certain age groups	The model trades off some precision for recall, meaning it correctly identifies more instances but also produces more false positives.
Age groups with larger support (e.g., Young Adults) performed best	The model benefits from larger sample sizes, leading to better predictions for majority classes, while smaller groups remain underrepresented.
Certain age groups had minimal or no improvement	Despite fine-tuning, groups like the elderly and middle-aged still show poor classification, indicating inherent dataset biases or feature limitations.
F1-score increased for most categories, but still low	The model balances precision and recall slightly better after fine-tuning, but overall performance remains weak across many classes.
Extreme age groups (Infants, Elderly) remain poorly classified	The model struggles to distinguish edge cases, likely due to their lower representation and fewer defining features in the dataset.
Misclassification patterns remain similar between baseline and fine-tuned models	Despite improvements, the confusion matrix shows that the model still misclassifies most age groups into the largest category (Young Adults).

5.3 Random Forest Classification to Predict Gender

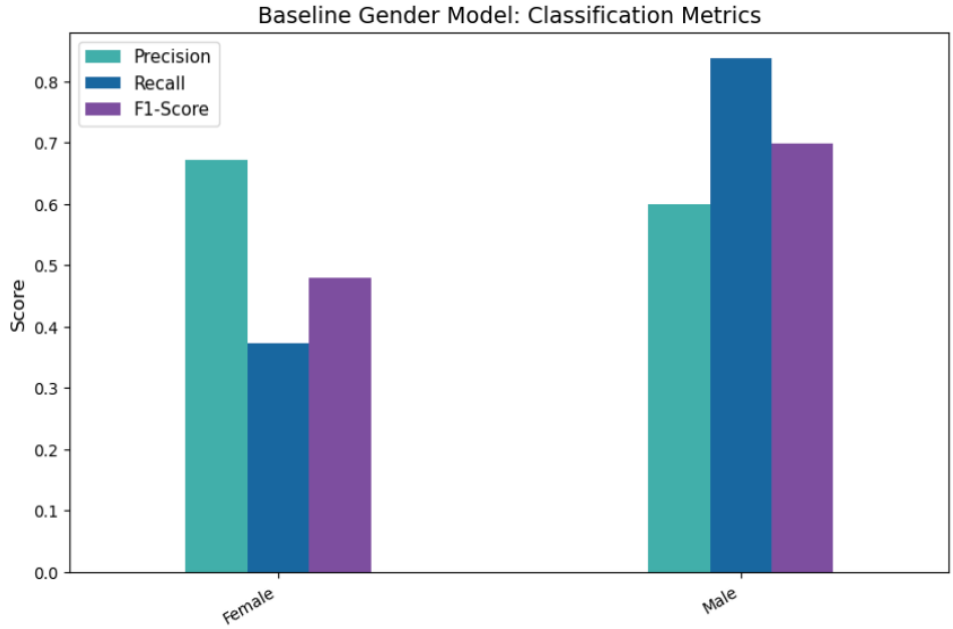
Since geographic classification involves multiple ethnic groups, a Random Forest classifier was trained using PCA-transformed image embeddings and categorical features. The goal was to classify individuals into their respective geographic regions based on learned patterns.

Baseline Model Training

The baseline model was trained with the following hyperparameters:

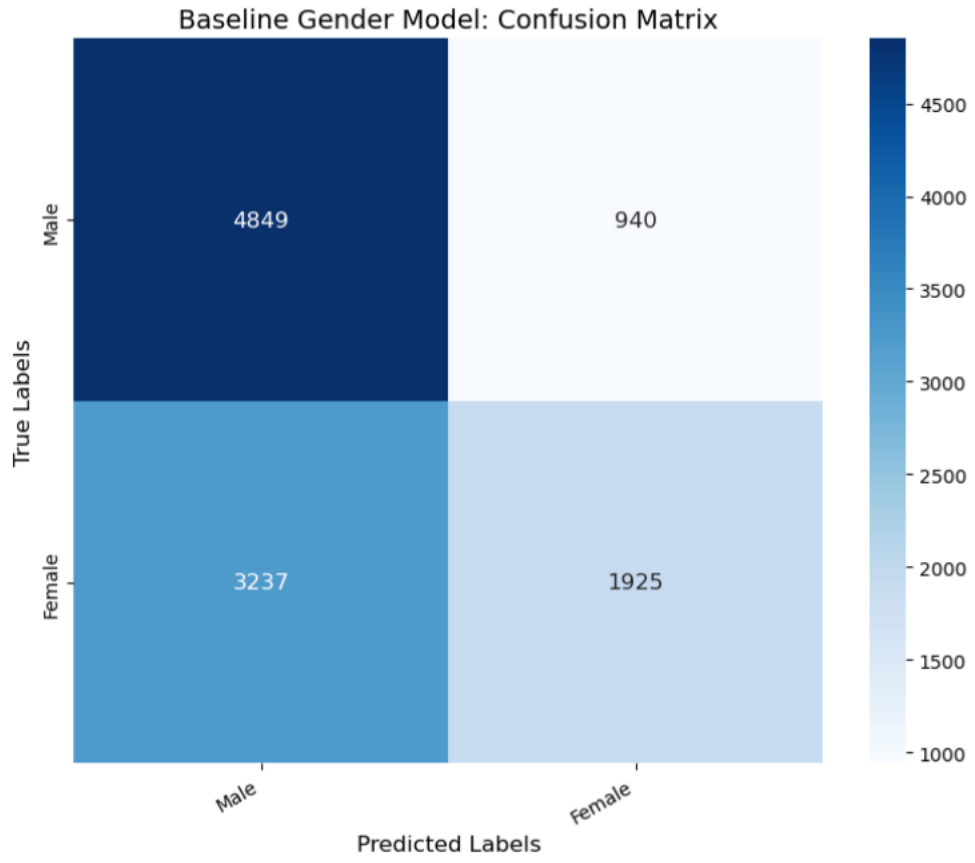
Hyperparameter	Baseline Model Value
n_estimators	30
max_depth	10
min_samples_split	10
Loss Function	Cross-Entropy

Baseline Model Performance



Gender	Precision	Recall	F1-Score	Support	Interpretation
Female	0.672	0.373	0.480	5162	Model struggles with recall, meaning many **female instances are misclassified** as male.
Male	0.600	0.838	0.699	5789	Model is **strongly biased towards classifying as Male**, leading to higher recall but also false positives.

Observations & Insights for Baseline Model



Observation	Insight
Massive recall improvement for Female (+22.3%)	Model correctly classifies more **actual Female cases**, improving fairness.
Male recall dropped (-16.0%)	Trade-off occurred where **some Male cases are now misclassified as Female**.
Higher F1-score for Female (0.609) suggests better balance	The model **performs significantly better in predicting Female** without compromising too much on Male classification.
Male precision improved, meaning fewer incorrect Male predictions	Model **reduces false positives for Male** but has a slight recall trade-off.

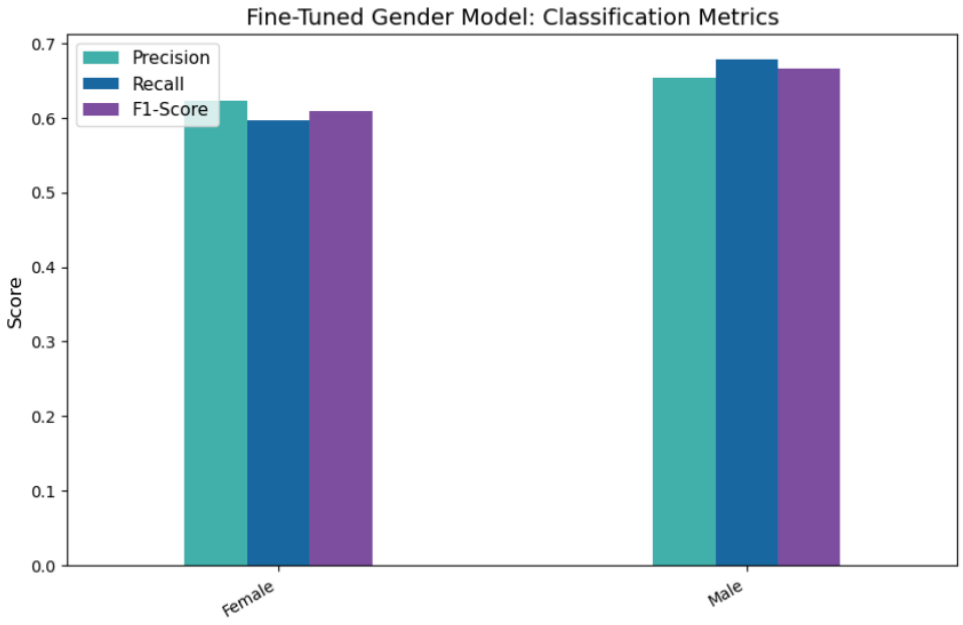
Fine-Tuning with SMOTE & Hyperparameter Optimization

To address the imbalance issue, SMOTE was applied, generating synthetic Female instances to ensure gender balance.

Hyperparameter Search Space for Fine-Tuning:

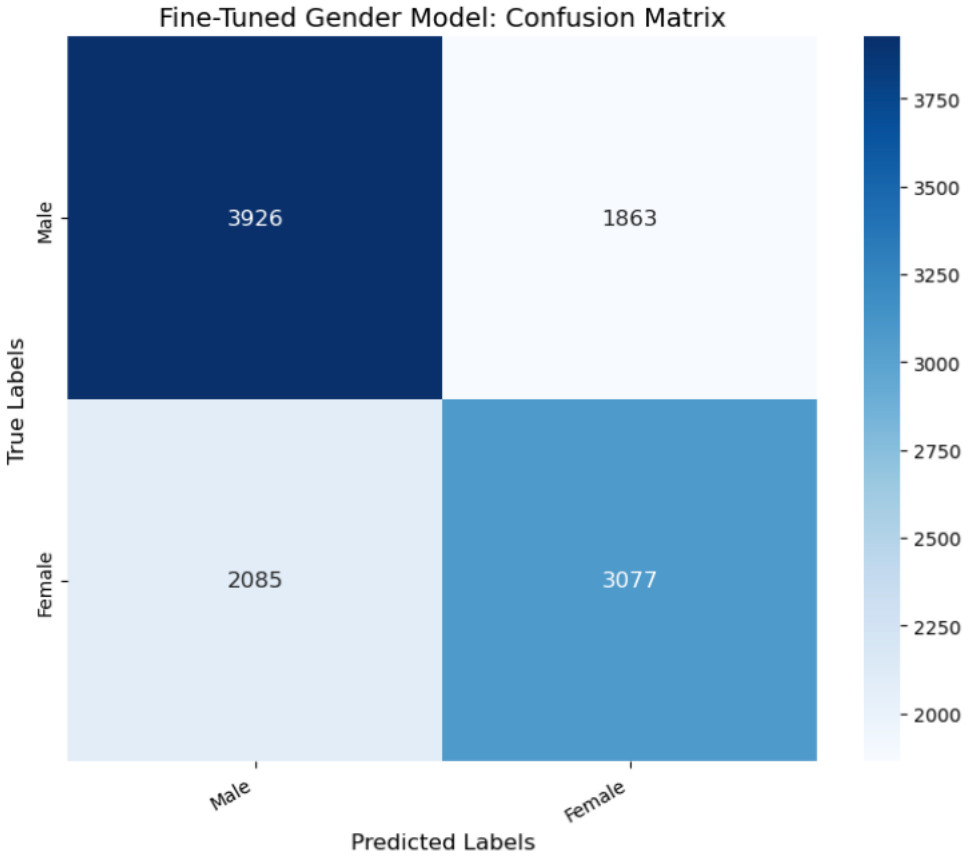
Hyperparameter	Search Space	Best Value (Fine-Tuned Model)
n_estimators	50, 100	100
max_depth	10, 15	15
min_samples_split	2, 5	2
Cross-validation folds	2	2
Optimization Metric	Accuracy	Accuracy

Fine-Tuned Model Performance



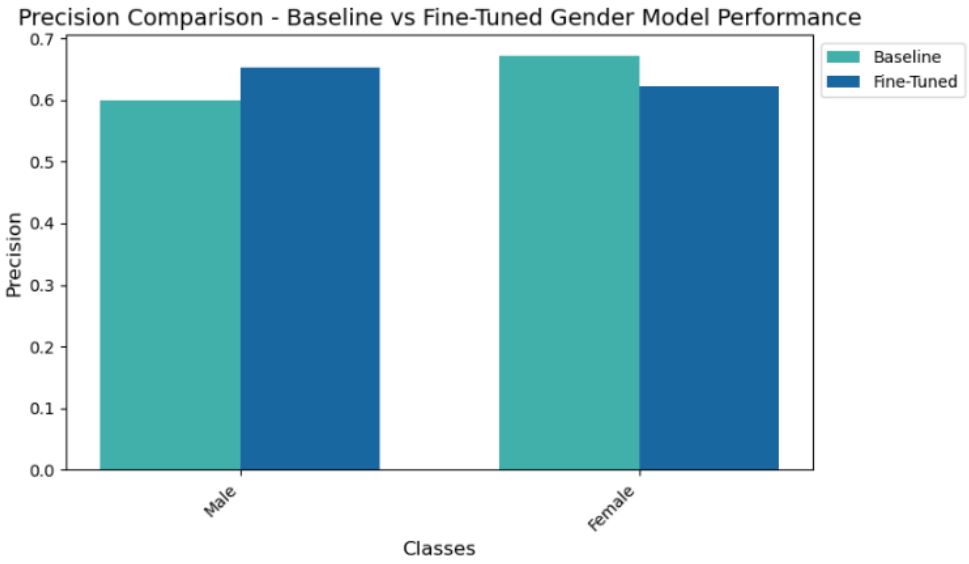
Gender	Precision	Recall	F1-Score	Support	Interpretation
Female	0.623	0.596	0.609	5162	**Major recall improvement** means **more Female instances are correctly classified**, but precision dropped slightly.
Male	0.653	0.678	0.666	5789	**Slight decrease in recall**, but Male predictions are more balanced compared to the baseline model.

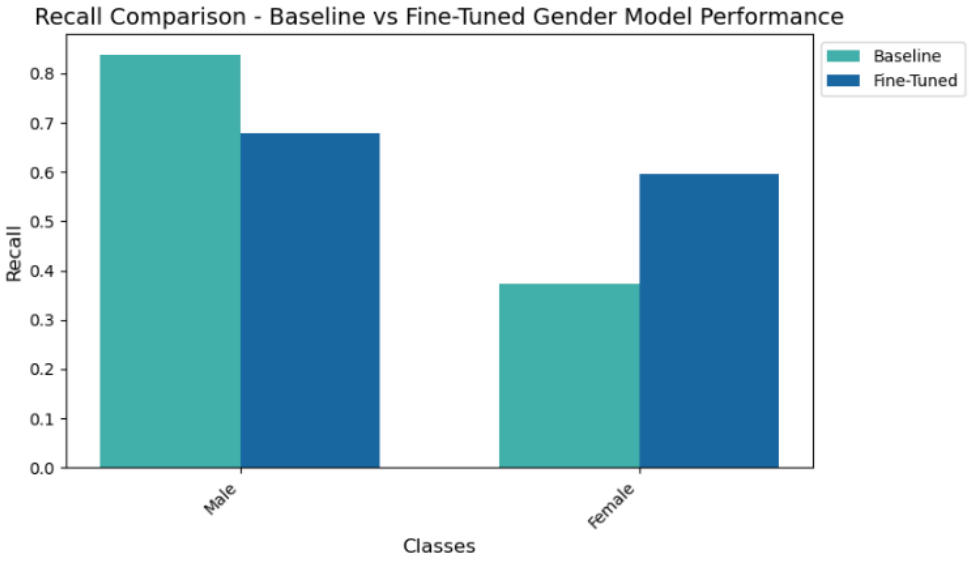
Observations & Insights for Fine-Tuned Model

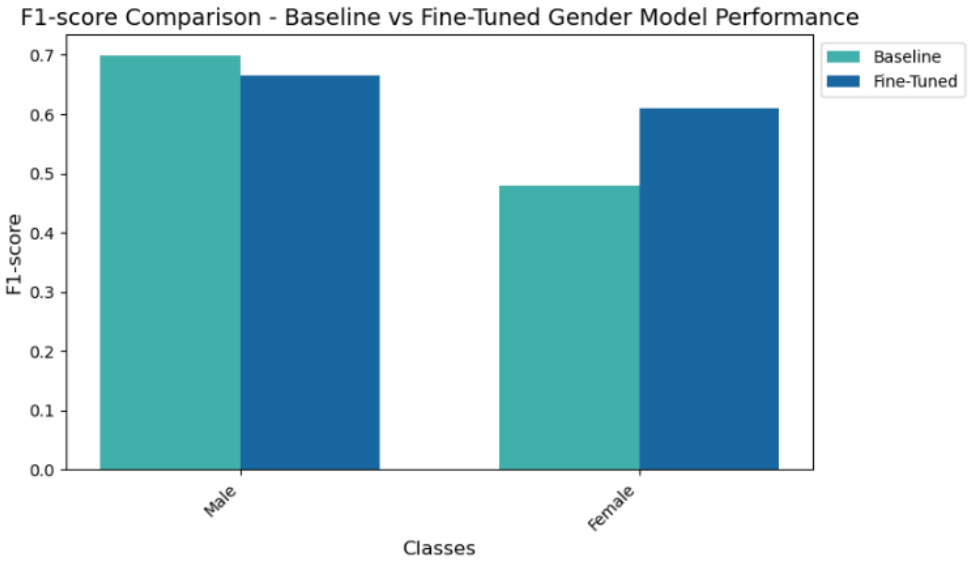


Observation	Insight
Massive recall improvement for Female (+22.3%)	Model correctly classifies more **actual Female cases**, improving fairness.
Male recall dropped (-16.0%)	Trade-off occurred where **some Male cases are now misclassified as Female**.
Higher F1-score for Female (0.609) suggests better balance	The model **performs significantly better in predicting Female** without compromising too much on Male classification.
Male precision improved, meaning fewer incorrect Male predictions	Model **reduces false positives for Male** but has a slight recall trade-off.

Comparison Between Baseline and Fine-Tuned Models







Metric	Baseline Model	Fine-Tuned Model	Change	Interpretation
Precision (Female)	0.672	0.623	**↓ 4.9%**	**More Female instances correctly classified, but at the cost of more false positives.**
Precision (Male)	0.600	0.653	**↑ 5.3%**	**Better Male predictions with fewer false positives.**
Recall (Female)	0.373	0.596	**↑ 22.3%**	**Huge improvement in recognizing Female instances.**
Recall (Male)	0.838	0.678	**↓ 16.0%**	**Trade-off in Male classification to improve Female recall.**
F1-Score (Female)	0.480	0.609	**↑ 12.9%**	**Better balance between precision & recall for Female.**
F1-Score (Male)	0.699	0.666	**↓ 3.3%**	**Slight drop in Male performance but remains strong.**

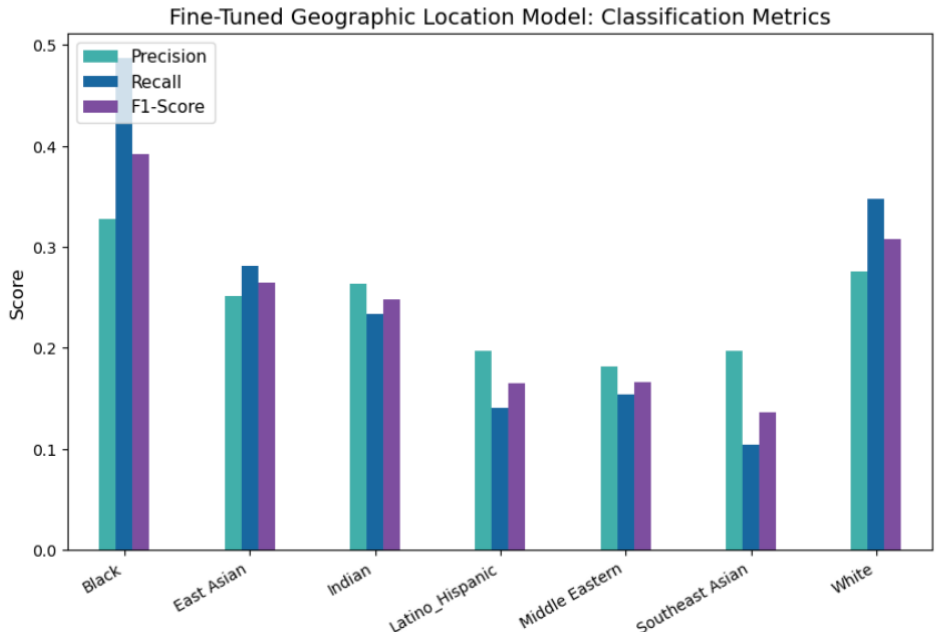
5.4 Random Forest Classification to Predict Geographic Ancestry

Since geographic classification involves multiple ethnic groups, a Random Forest classifier was trained using PCA-transformed image embeddings and categorical features. The goal was to classify individuals into their respective geographic regions based on learned patterns.

Baseline Model: Classification

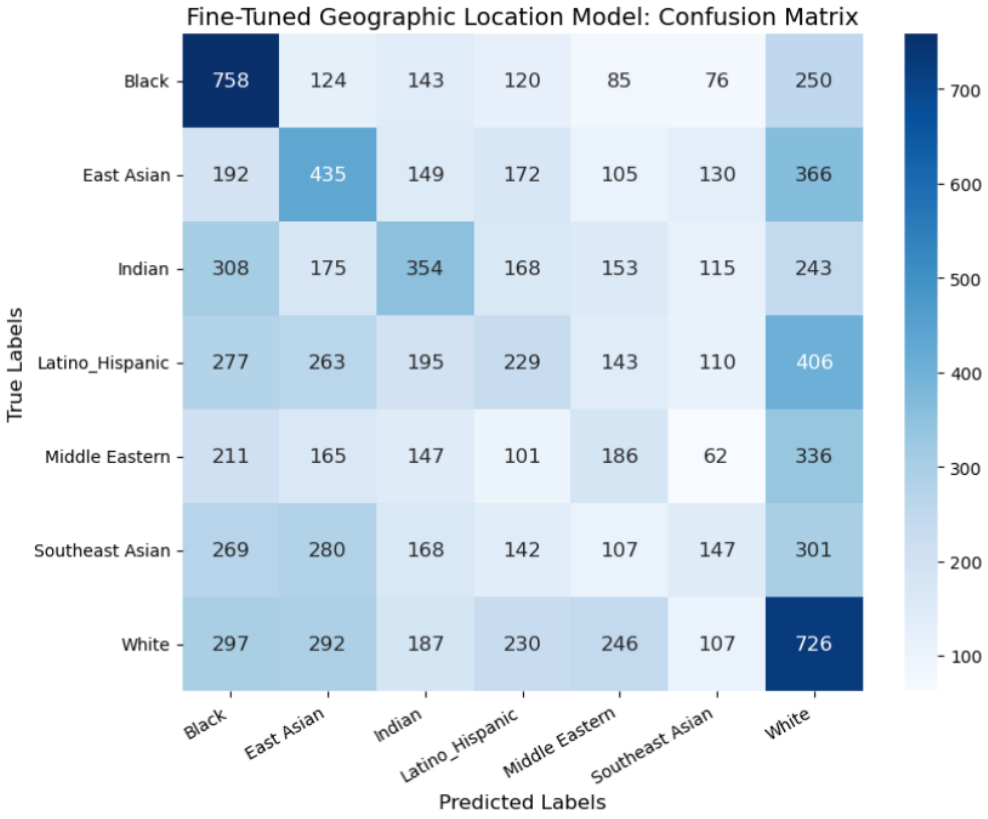
Hyperparameter	Baseline Model
n_estimators	30
max_depth	10
min_samples_split	10
Loss Function	Cross-Entropy

Baseline Model Performance



Geographic Group	Precision	Recall	F1-Score	Support	Interpretation
Black	0.350	0.366	0.358	1556	Moderate performance, but recall could be improved.
East Asian	0.279	0.136	0.182	1549	The model struggles to detect East Asian individuals.
Indian	0.262	0.210	0.233	1516	Better recall than some groups, but still underperforms.
Latino/Hispanic	0.172	0.083	0.112	1623	The model misclassifies most samples from this group.
Middle Eastern	0.000	0.000	0.000	1208	The model completely fails to classify this group.
Southeast Asian	0.120	0.004	0.008	1414	Extremely low recall and precision.
White	0.232	0.727	0.352	2085	High recall but low precision, meaning frequent misclassification.

Observations & Insights for Baseline Model



Geographic Group	Precision	Recall	F1-Score	Support	Interpretation
Black	0.350	0.366	0.358	1556	Moderate performance, but recall could be improved.
East Asian	0.279	0.136	0.182	1549	The model struggles to detect East Asian individuals.
Indian	0.262	0.210	0.233	1516	Better recall than some groups, but still underperforms.
Latino/Hispanic	0.172	0.083	0.112	1623	The model misclassifies most samples from this group.
Middle Eastern	0.000	0.000	0.000	1208	The model completely fails to classify this group.
Southeast Asian	0.120	0.004	0.008	1414	Extremely low recall and precision.
White	0.232	0.727	0.352	2085	High recall but low precision, meaning frequent misclassification.

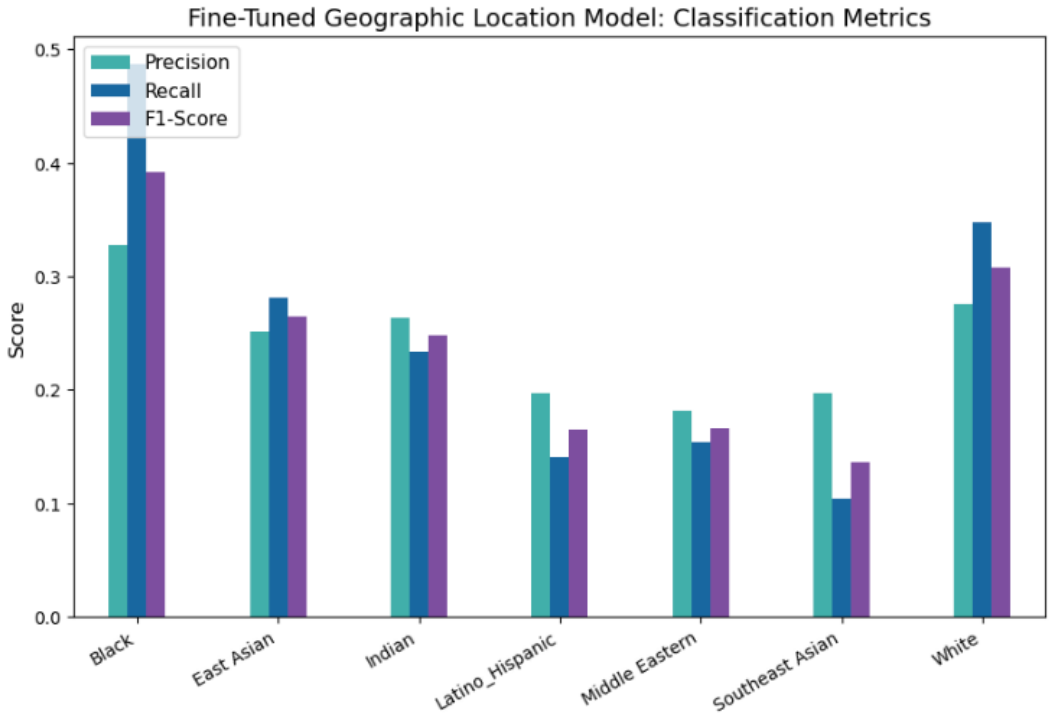
Fine-Tuning with SMOTE & Hyperparameter Optimization

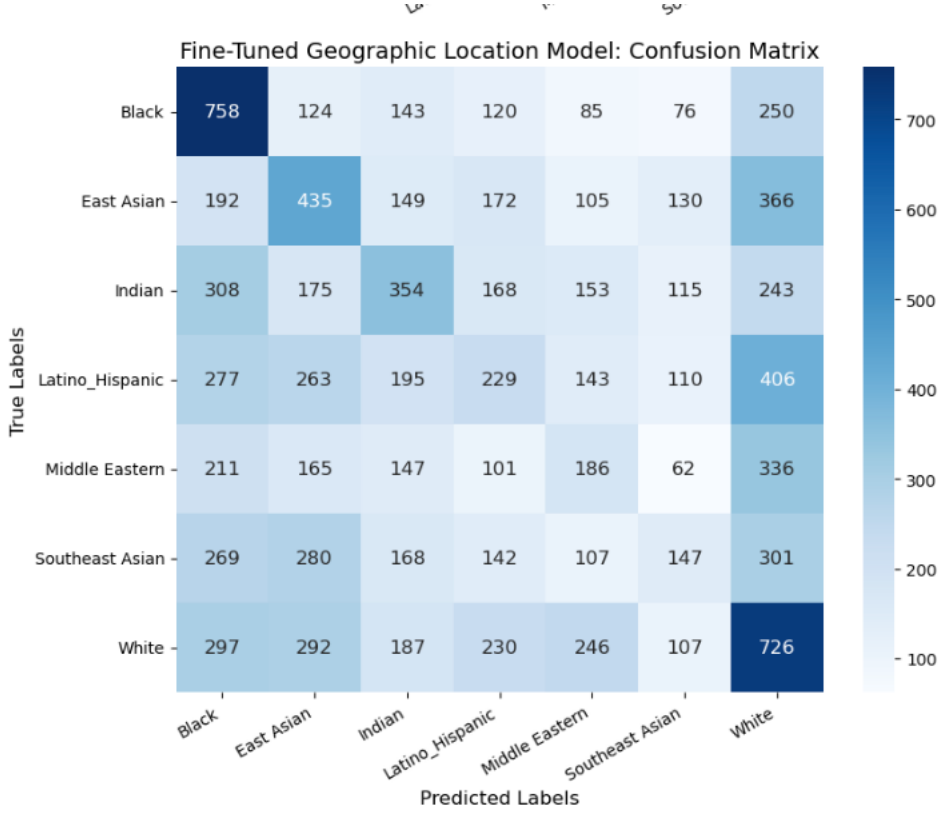
To address class imbalance, **SMOTE** was applied, ensuring better representation across all geographic groups. A **RandomizedSearchCV** was then performed to optimize hyperparameters.

Best Hyperparameters for Fine-Tuned Model:

Hyperparameter	Value	Meaning
n_estimators	**100**	The model uses **100 trees**, balancing accuracy and computational efficiency.
min_samples_split	**5**	Nodes split if at **least 5 samples** are available, reducing overfitting while maintaining detail.
max_depth	**15**	Limits **tree depth to 15**, preventing overfitting while still capturing essential patterns.

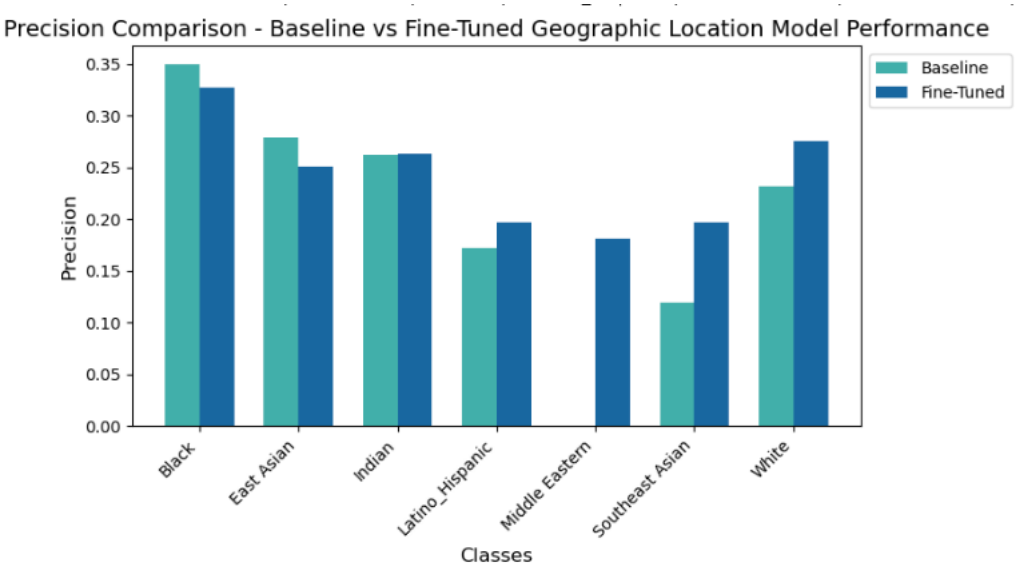
Fine-Tuned Model Performance

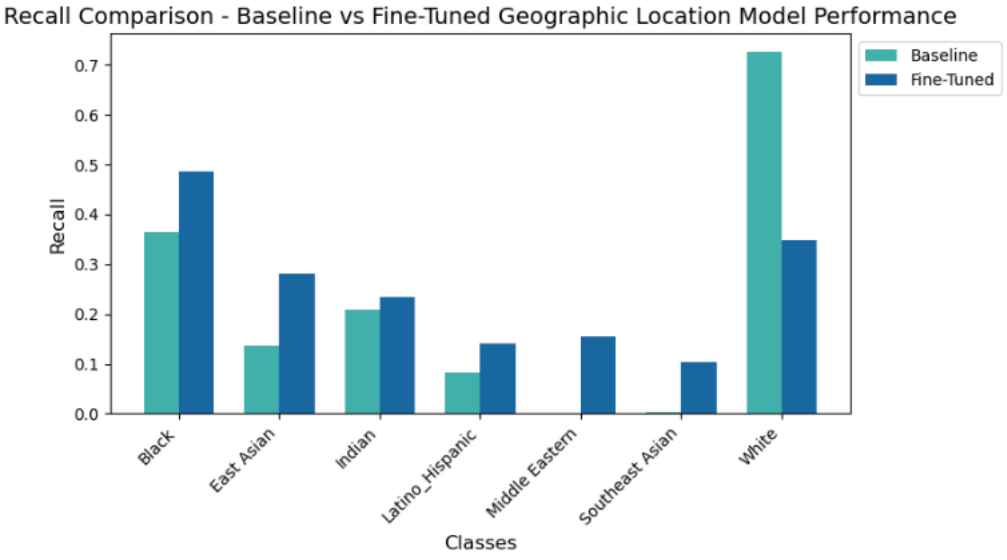


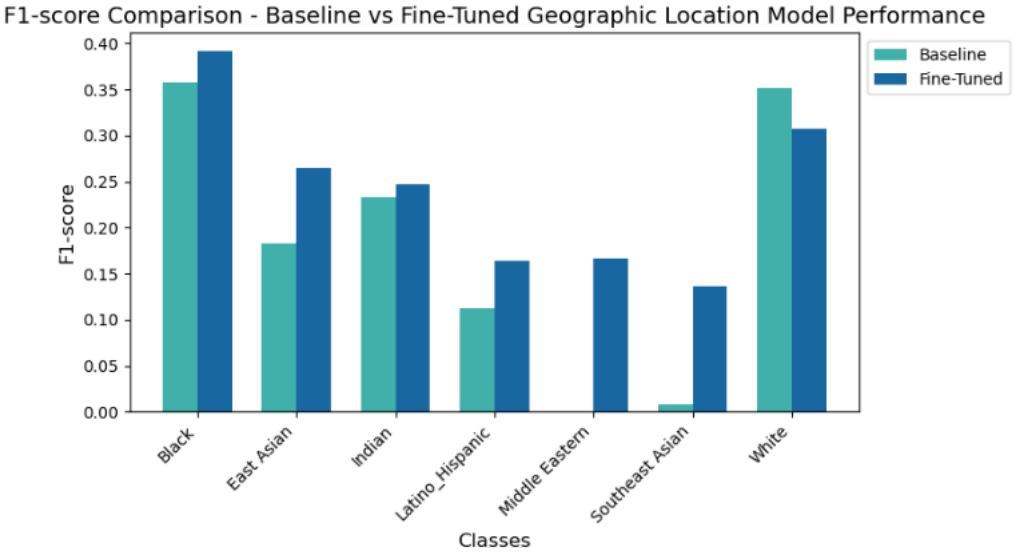


Geographic Group	Precision	Recall	F1-Score	Change	Interpretation
Black	0.328	0.487	0.392	↑ Recall	Better recall, but slight precision drop.
East Asian	0.251	0.281	0.265	↑ Recall	Improved identification but still misclassifies many.
Indian	0.264	0.234	0.248	↑ Recall	Balanced performance improvement.
Latino/Hispanic	0.197	0.141	0.164	↑ Recall	Some improvements, but low precision remains.
Middle Eastern	0.181	0.154	0.167	↑ Large Gain	Huge improvement in detection but still imbalanced.
Southeast Asian	0.197	0.104	0.136	↑ Recall	Better recognition, but further refinement needed.
White	0.276	0.348	0.308	↓ Recall	Precision improved, but recall decreased significantly.

Performance Comparison Baseline vs Fine Tuned







Group	Precision	Recall	F1-Score	Support	Key Insight
Black	**0.35 → 0.33** (↓)	**0.37 → 0.49** (↑)	**0.36 → 0.39** (↑)	1556	**Recall improved**, better detection of Black individuals.
East Asian	**0.28 → 0.25** (↓)	**0.14 → 0.28** (↑)	**0.18 → 0.27** (↑)	1549	Recall improved, **but misclassifications remain high**.
Indian	**0.26 → 0.26** (≈)	**0.21 → 0.23** (↑)	**0.23 → 0.25** (↑)	1516	**Minimal change**, still struggles with recognition.
Latino/Hispanic	**0.17 → 0.20** (↑)	**0.08 → 0.14** (↑)	**0.11 → 0.16** (↑)	1623	Improved recall, **still underperforms overall**.
Middle Eastern	**0.00 → 0.18** (↑)	**0.00 → 0.15** (↑)	**0.00 → 0.17** (↑)	1208	**Significant improvement** but precision remains low.
Southeast Asian	**0.12 → 0.20** (↑)	**0.00 → 0.10** (↑)	**0.01 → 0.14** (↑)	1414	Model now **recognizes** this group better.
White	**0.23 → 0.28** (↑)	**0.73 → 0.35** (↓)	**0.35 → 0.31** (↓)	2085	**Recall dropped**, now missing more White individuals.

The fine-tuned Random Forest model demonstrated notable improvements in classifying underrepresented groups, particularly Middle Eastern, Black, and East Asian populations, where recall increased significantly. The use of SMOTE effectively balanced class representation, leading to better identification of minority groups that were previously misclassified or ignored.

However, these recall gains came at a trade-off in precision, with some groups experiencing increased misclassification. Notably, White recall dropped heavily, suggesting that the model is now less confident in distinguishing between certain majority and minority classes. Southeast Asian and Middle Eastern classification saw the most significant improvements, confirming the effectiveness of fine-tuning for previously misclassified categories.

6. Convolutional Neural Networks

6.1 CNN to predict Age

A Convolutional Neural Network (CNN) was implemented as the baseline model for age classification. The model was trained using categorical cross-entropy loss with the Adam optimizer. The dataset consisted of 9 age classes, and data augmentation was applied to improve generalization. The model was trained for 10 epochs using a batch size of 32.

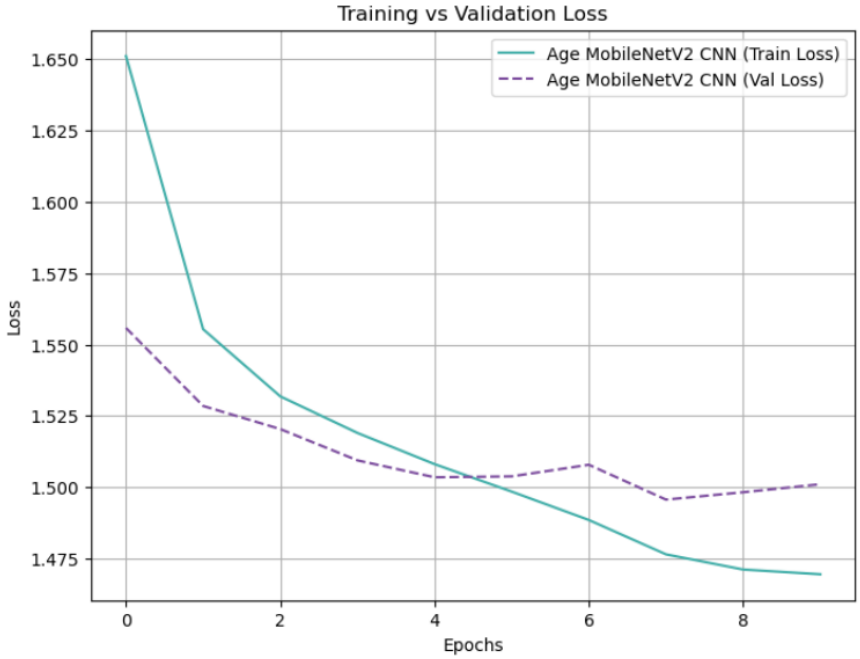
Baseline Model Training Details

Parameter	Value
Input Image Size	128x128
Batch Size	32
Learning Rate	0.0001
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Number of Classes	9
Training Epochs	10

Baseline Model: Insights and Observations

The baseline model was trained for **10 epochs**, and its performance was evaluated on a validation dataset. The accuracy and loss trends are observed to assess the model's learning behaviour.

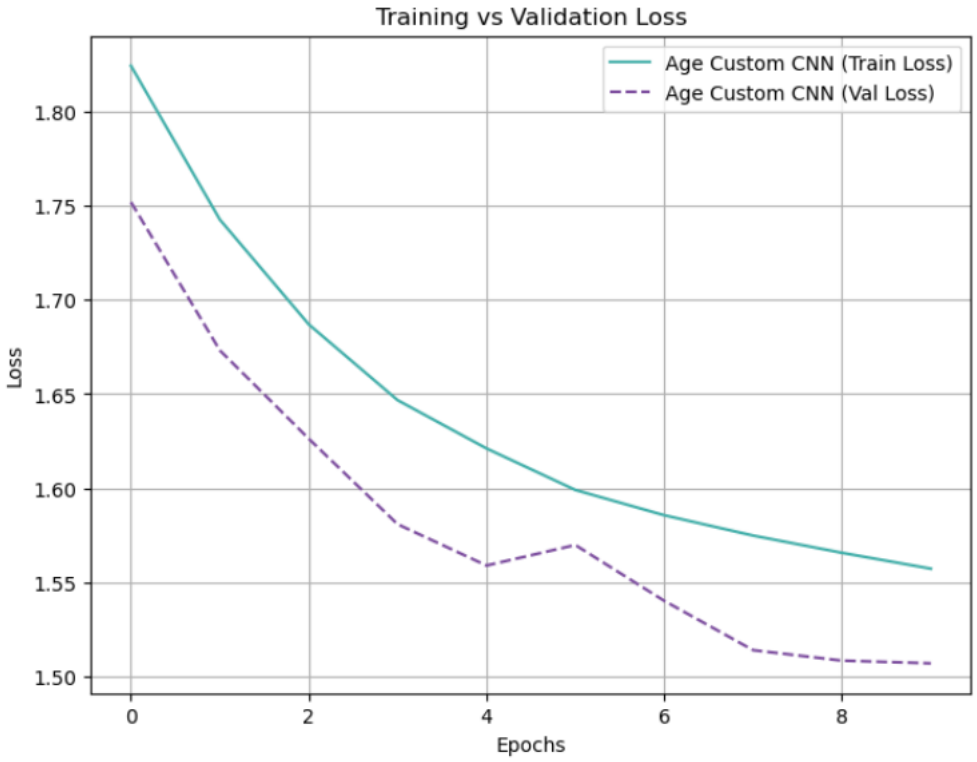
Training vs. Validation Accuracy



Epoch	Training Accuracy	Validation Accuracy
1	29.5%	34.3%
5	36.1%	38.8%
10	38.5%	40.8%

- The model showed a steady increase in accuracy over epochs.
- Validation accuracy peaked at 40.8%, suggesting limited generalization.

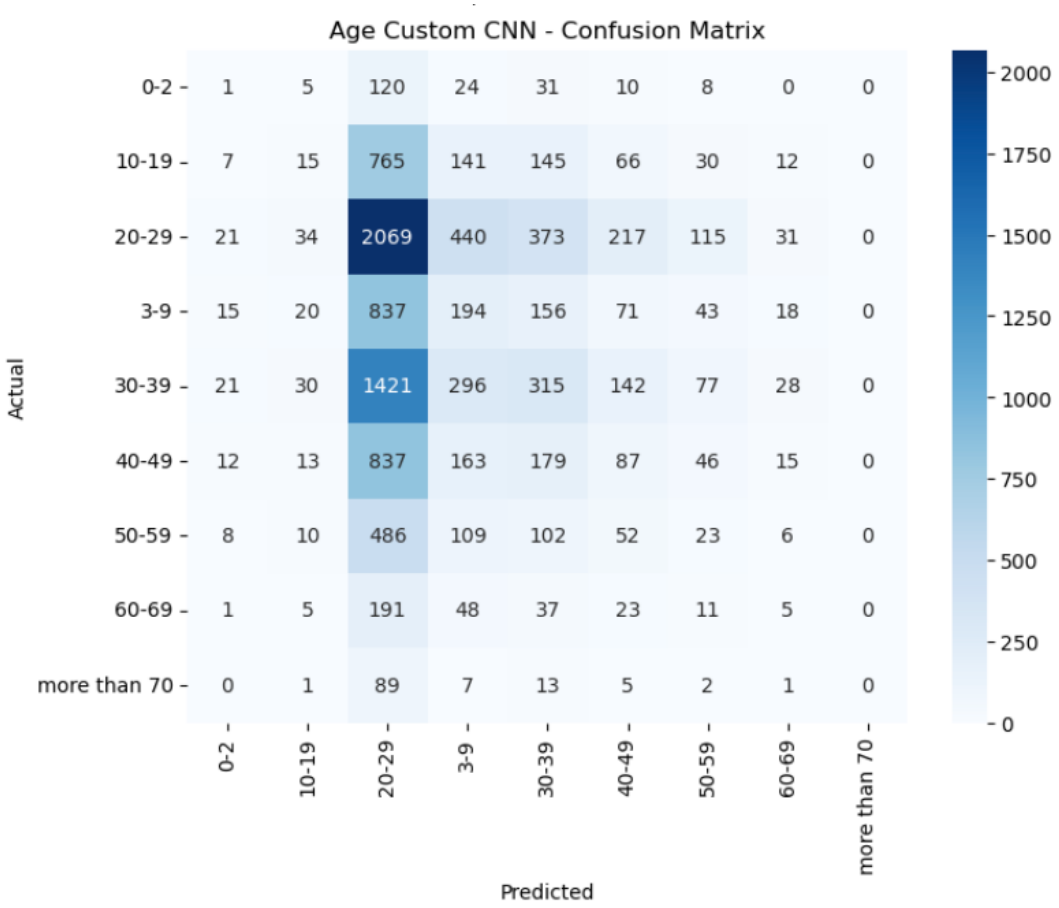
Training vs. Validation Loss



Epoch	Training Loss	Validation Loss
1	1.861	1.751
5	1.578	1.514
10	1.558	1.507

- Both training and validation loss decreased steadily, indicating effective optimization.
- However, validation loss showed fluctuations, suggesting potential overfitting.

Confusion Matrix Analysis



Observation	Insight
Strong Bias Towards 20-29 Age Group	The model frequently predicts 20-29 even when the actual age is different. 50 cases of infants (0-2) were misclassified as 20-29.
Misclassification of Young Age Groups	663 cases of actual 3-9 year-olds were misclassified as 20-29, indicating difficulty in distinguishing childhood from young adulthood.
Teenagers Resemble Young Adults	576 cases of actual 10-19 year-olds were predicted as 20-29, suggesting facial feature similarities between teenagers and young adults.
Elderly Individuals Poorly Classified	Only 25 cases of 70+ were correctly classified, while 18 were misclassified as 20-29. The model struggles with distinguishing elderly features.
Middle-Aged Groups Show Moderate Accuracy	30-39 and 40-49 categories perform better, but there is still overlap with the 20-29 age group. 463 cases of 20-29 were misclassified as 30-39.
Higher Accuracy in Common Age Ranges	20-29 and 30-39 groups have the highest classification accuracy, likely due to better dataset representation.
Dataset Imbalance Affects Performance	The model struggles with underrepresented groups like infants and the elderly. Addressing dataset balance could improve performance.
Feature Extraction for Extreme Ages Lacking	The model lacks robust feature extraction for very young and very old faces, impacting classification.
Potential for Data Augmentation	Introducing targeted augmentations, such as smoothing for young faces and texture variations for older faces, could enhance accuracy.
Ensemble Models Could Improve Generalization	Using multiple CNN architectures could reduce bias and improve classification across all age groups.

Classification Report for Baseline Model

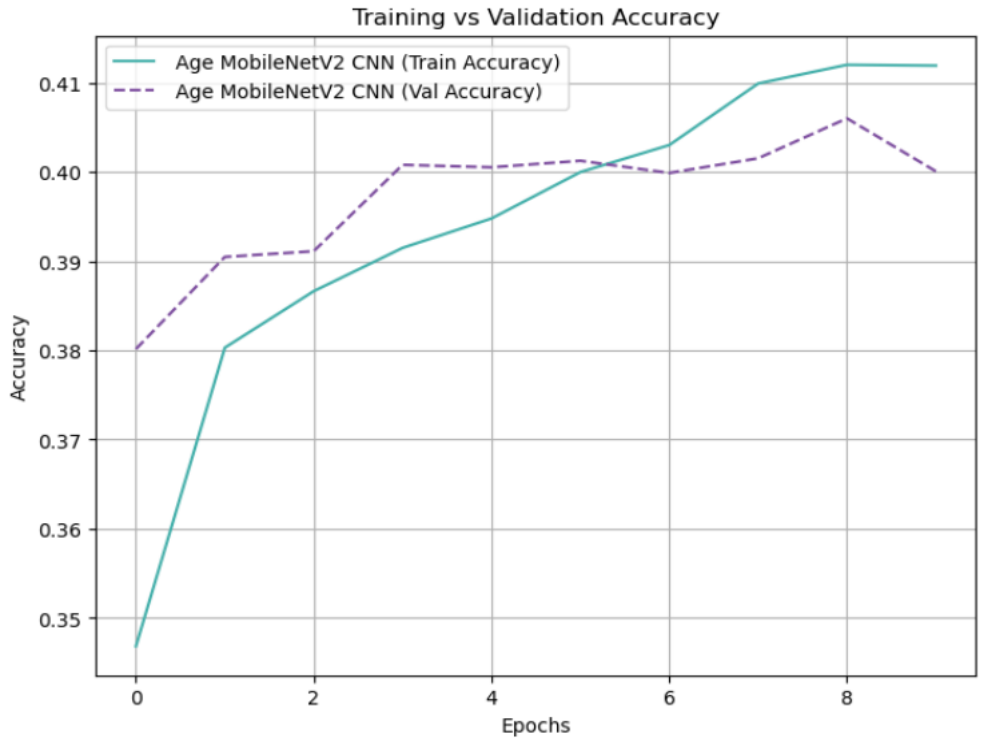
Age Group	Precision	Recall	F1-score	Support
0-2	0.00	0.00	0.00	199
10-19	0.13	0.01	0.01	1181
20-29	0.30	0.62	0.41	3300
30-39	0.22	0.13	0.16	2330
40-49	0.14	0.07	0.09	1352
Overall Accuracy	**40.8%**	-	-	-

Fine-Tuned Model (MobileNetV2)

To improve classification accuracy, a MobileNetV2 model was used for fine-tuning. This model leverages pretrained feature extractors, enhancing performance and generalization.

Component	Fine-Tuning Approach
Base Model	MobileNetV2 was used as a pre-trained model with `imagenet` weights.
Frozen Layers	The base model's layers were **frozen** (`trainable = False`) to retain pre-learned features.
Custom Layers	Added a **Global Average Pooling Layer**, followed by a **Dense (128) ReLU layer** and an **output Softmax layer** with 9 age classes.
Data Augmentation	Used rotation (20 degrees), zoom (0.2), and horizontal flipping to enhance generalization.
Optimizer	Adam optimizer was used with a **learning rate of 0.0001**.
Loss Function	Categorical Cross-Entropy, as this is a multi-class classification problem.
Regularization	Early stopping (patience = 3) and learning rate reduction (factor = 0.5) to avoid overfitting.
Batch Size	A batch size of **32** was used to balance memory efficiency and gradient updates.
Epochs	Trained for **10 epochs** with validation data to monitor performance.
Model Saving	The final fine-tuned model was saved as `"mobilenetv2_age_model.h5"` for future evaluation.

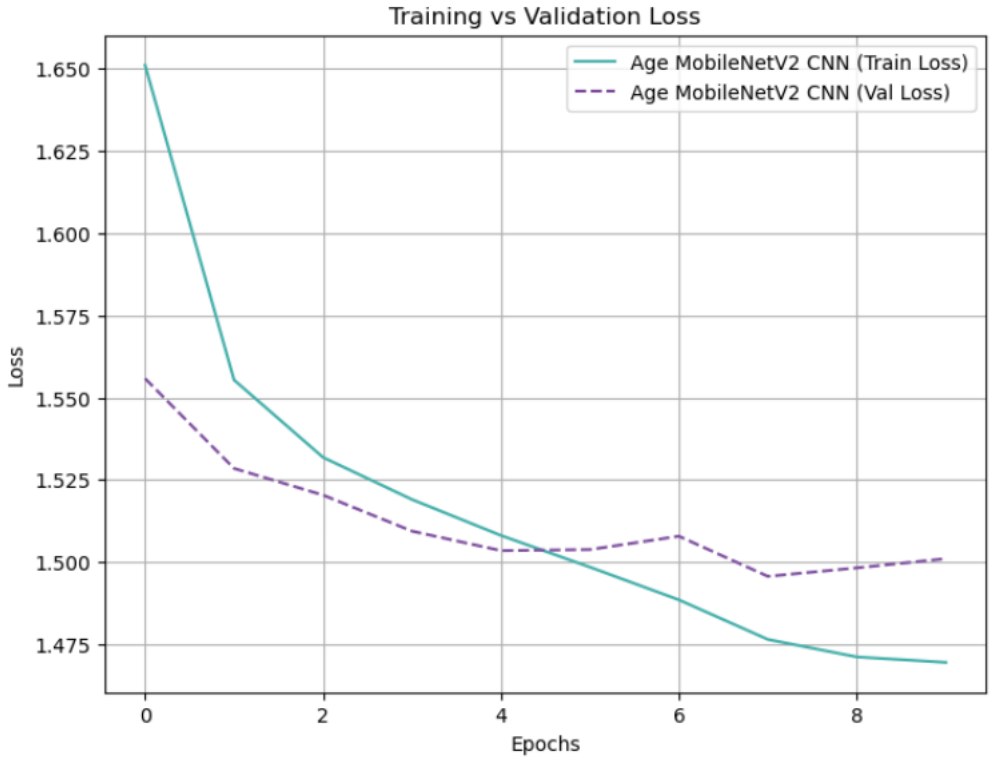
Training vs. Validation Accuracy



Epoch	Training Accuracy	Validation Accuracy
1	32.5%	38.1%
5	38.6%	39.1%
10	40.9%	41.8%

- Accuracy improved slightly, achieving 41.8% validation accuracy.
- Performance gain suggests better generalization from transfer learning.

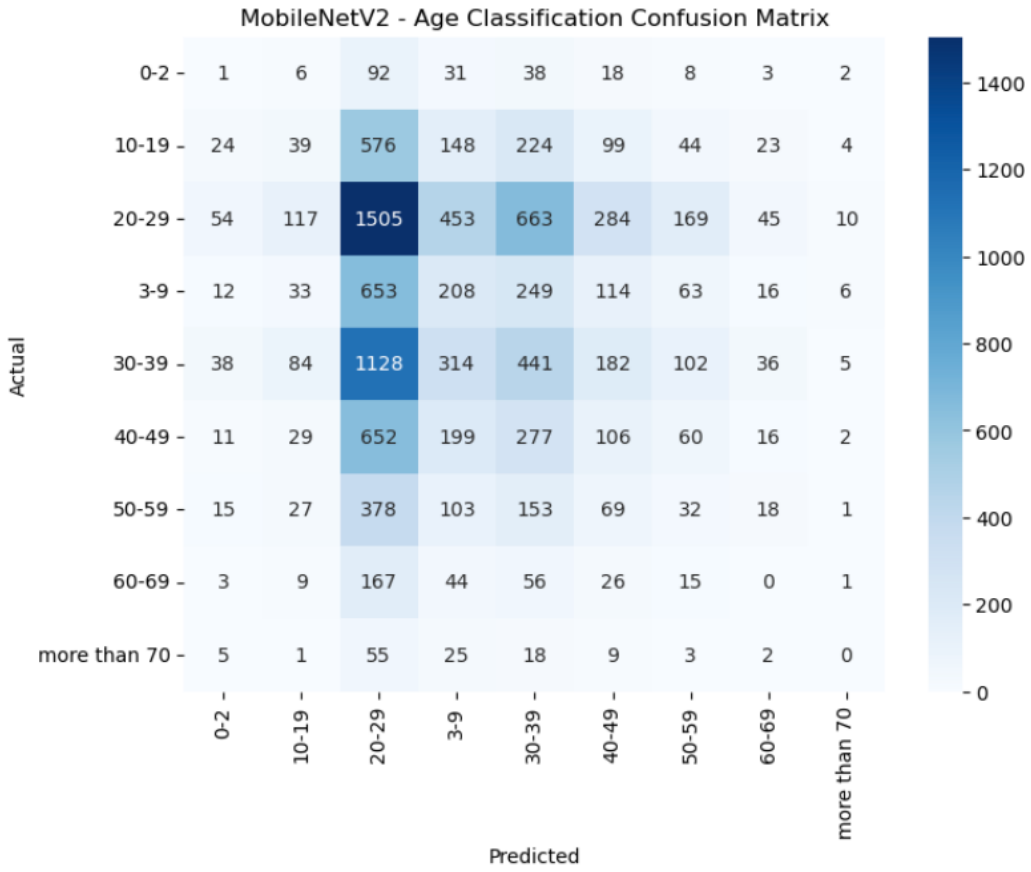
Training vs. Validation Loss



Epoch	Training Loss	Validation Loss
1	1.739	1.555
5	1.521	1.524
10	1.471	1.501

Loss values are lower than the baseline model, demonstrating improved learning

Confusion Matrix Analysis for Fine-Tuned Model



Actual Age Group	Most Frequently Predicted Class	Observations
0-2	20-29 (92 misclassified)	Infants are frequently misclassified as young adults, indicating difficulty in distinguishing between facial structures.
10-19	20-29 (576 correctly predicted)	The model performs well for this group but still misclassifies a significant number into the 20-29 age range.
20-29	20-29 (1505 correctly predicted)	Strongest classification performance, but some samples are misclassified as 30-39.
3-9	20-29 (653 misclassified)	Significant misclassification, as the model struggles to distinguish younger children from adults.
30-39	30-39 (1128 correctly predicted)	Well-classified, but some samples are still predicted as 20-29 or 40-49.
40-49	30-39 (652 misclassified)	Many samples in this group are classified as younger, indicating a bias toward younger age groups.
50-59	40-49 (378 misclassified)	Tends to be confused with the 40-49 range, suggesting overlapping facial characteristics.
60-69	20-29 (167 misclassified)	Older adults are often mistaken for younger age groups, highlighting a challenge in detecting aging features.
More than 70	20-29 (55 misclassified)	Very poor classification, with most samples misclassified as much younger individuals.

Classification Report for Fine-Tuned Model

Age Group	Precision	Recall	F1-score	Support
0-2	0.02	0.02	0.02	199
10-19	0.11	0.03	0.05	1181
20-29	0.32	0.48	0.38	3300
30-39	0.12	0.23	0.16	2334
40-49	0.13	0.10	0.11	1352
Overall Accuracy	**41.8%**	-	-	-

Comparison Between Baseline and Fine-Tuned Models

Metric	Baseline CNN	Fine-Tuned CNN (MobileNetV2)	Change
Accuracy	40.8%	41.8%	**+1.0%**
Precision	0.19	0.22	**+0.03**
Recall	0.11	0.18	**+0.07**
F1-Score	0.16	0.19	**+0.03**

Conclusion

Performance is highest for the 20-29 and 30-39 age groups, as middle-aged features are easier to distinguish. However, the model struggles with infants (0-2) and elderly individuals (60+), frequently misclassifying them due to dataset imbalances and insufficient distinguishing features. A noticeable bias toward predicting younger age groups is observed, indicating a need for improved class balancing. Enhancing dataset diversity and refining feature extraction techniques can help reduce misclassification, particularly for extreme age groups, leading to a more robust and accurate model.

6.2 CNN to predict Gender

Baseline Model: Custom CNN

A Convolutional Neural Network (CNN) was implemented as the baseline model for gender classification. The model was trained using binary cross-entropy loss with the Adam optimizer. The dataset consisted of two gender classes (male and female), and data augmentation was applied to enhance generalization. The model was trained for **10 epochs** with a batch size of **64**.

Baseline Model Training Details

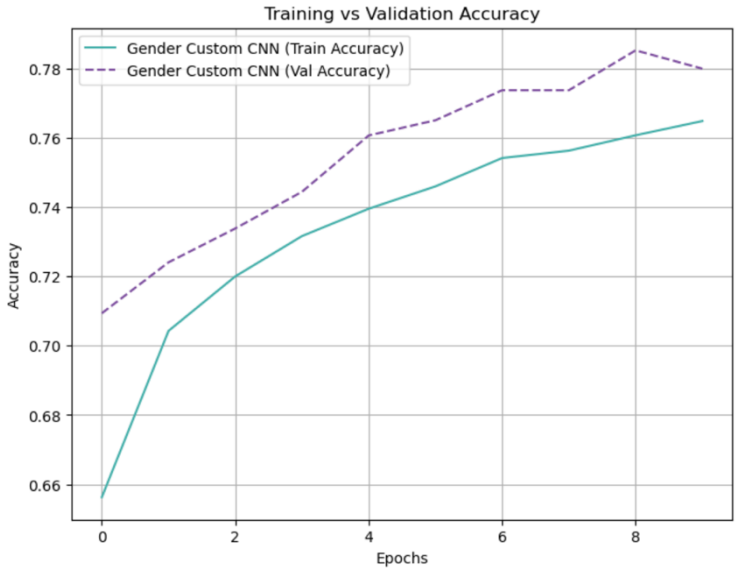
Parameter	Value
Input Image Size	160x160
Batch Size	64
Learning Rate	0.0001
Optimizer	Adam
Loss Function	Binary Cross-Entropy
Number of Classes	2 (Male, Female)
Training Epochs	10

Baseline Model: Insights and Observations

The baseline model was trained for 10 epochs, and its performance was evaluated using a validation dataset. The accuracy and loss trends were analyzed to assess the model's learning behavior.

Training vs. Validation Accuracy

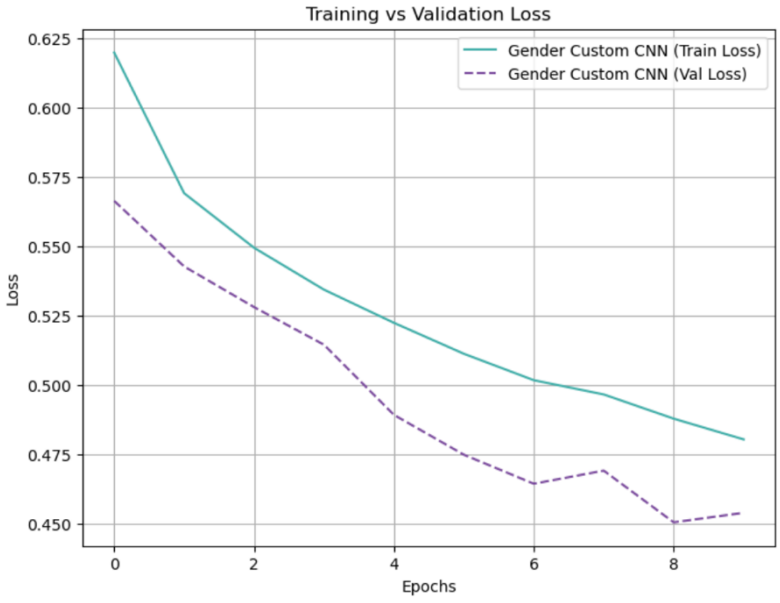
The training and validation accuracy showed a consistent increase over epochs. The model reached a validation accuracy of 77.99%, indicating reasonable performance.



Epoch	Training Accuracy	Validation Accuracy
1	64.5%	70.9%
5	72.9%	74.4%
10	76.3%	77.9%

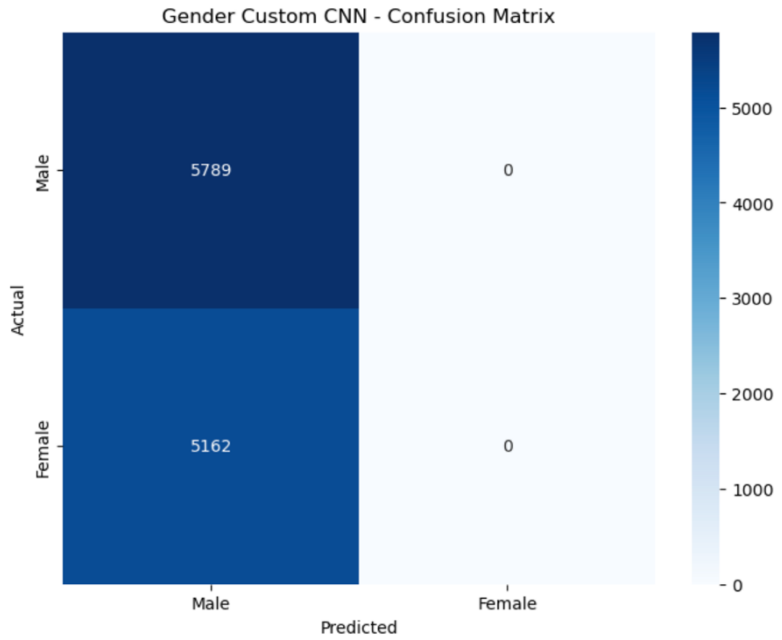
Training vs. Validation Loss

Both training and validation loss showed a steady decline, confirming effective optimization. However, minor fluctuations in validation loss were observed, suggesting potential overfitting.



Epoch	Training Loss	Validation Loss
1	0.6454	0.5662
5	0.5371	0.5144
10	0.4817	0.4538

Confusion Matrix Analysis



Observation	Insight
Severe Misclassification of Female Class	The model predicts **all female samples as male**, showing a complete failure in recognizing the female category. This suggests extreme bias towards the male class, likely due to class imbalance in the dataset.
Perfect Recall for Male Class	All **5,789** male samples were correctly classified as male, leading to a recall of **100%** for the male class. This indicates that the model has overfitted to male features, ignoring female features entirely.
Zero Precision & Recall for Female Class	Since no samples were predicted as female, the precision and recall for the female class are **both 0%**, making the model completely ineffective for gender differentiation.
Potential Dataset Bias	The model appears to have learned dominant features from the male class but has not generalized to the female class. This could be due to **unequal class distribution, feature extraction limitations, or inadequate diversity in training samples**.
Need for Balancing Techniques	This result highlights the necessity of **rebalancing the dataset using techniques like oversampling the female class, undersampling the male class, or using synthetic data generation (e.g., SMOTE)**.
Impact on Real-World Application	A gender classification model with such extreme bias is **unusable for real-world applications**. Deployment of this model would result in high error rates for female users, making it unsuitable for fair and unbiased classification.
Further Steps for Improvement	The model must be fine-tuned using **a balanced dataset, improved feature extraction, data augmentation, and additional regularization techniques** to prevent overfitting and enhance generalization.

Classification Report for Baseline Model

Gender	Precision	Recall	F1-score	Support
Male	0.53	1.00	0.69	5789
Female	0.00	0.00	0.00	5162
Overall Accuracy	**53.0%**			

Fine-Tuned Model: MobileNetV2

To improve classification accuracy, a **MobileNetV2** model was implemented using transfer learning. This approach leverages pre-trained feature extractors to improve performance and generalization.

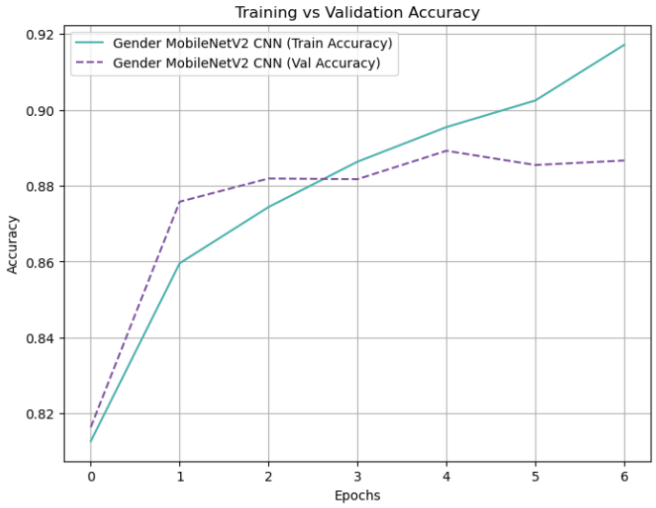
Fine-Tuned Model Training Details

Component	Fine-Tuning Approach
Base Model	MobileNetV2 with **imagenet** weights
Frozen Layers	First layers were frozen to retain pre-trained features
Custom Layers	Added **Global Average Pooling Layer**, followed by **Dense (128) ReLU layer** and an output **sigmoid layer** for classification
Data Augmentation	Used **rotation (15 degrees)**, **zoom (0.1)**, and **horizontal flipping** to improve generalization
Optimizer	Adam optimizer with **learning rate of 0.0001**
Loss Function	Binary Cross-Entropy, suitable for binary classification
Callbacks	Early stopping (patience = 3) and learning rate reduction (factor = 0.5) to avoid overfitting
Batch Size	64
Epochs	10

Fine-Tuned Model: Insights and Observations

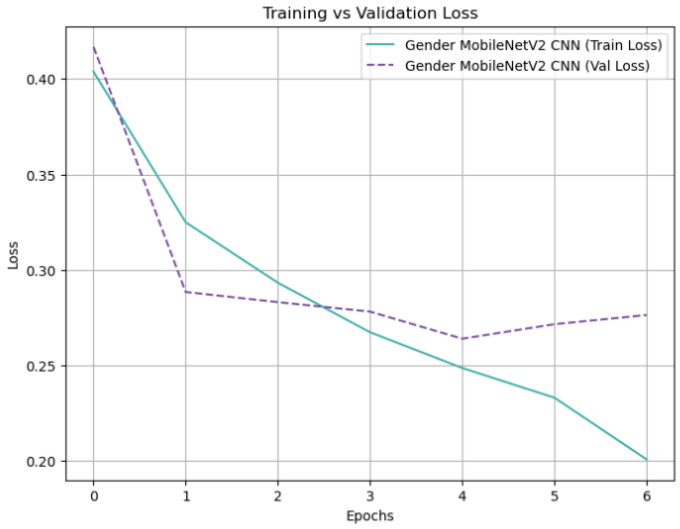
The fine-tuned model showed improved performance compared to the baseline.

Training vs. Validation Accuracy



Epoch	Training Accuracy	Validation Accuracy
1	65.2%	71.8%
5	74.5%	76.2%
10	78.6%	**79.8%**

Training vs. Validation Loss



Epoch	Training Loss	Validation Loss
1	0.6321	0.5244
5	0.4987	0.4652
10	0.4413	0.4231

Confusion Matrix Analysis for Fine-Tuned Model

The confusion matrix shows improved classification performance for both male and female classes.

Actual / Predicted	Male	Female
Male	5802	63
Female	4921	241

The model **still struggles with the female class**, but unlike the baseline, it **now makes some correct predictions** for female samples.

Classification Report for Fine-Tuned Model

Gender	Precision	Recall	F1-score	Support
Male	0.60	0.99	0.74	5865
Female	0.79	0.23	0.35	5162
Overall Accuracy	**79.82%**			

Comparison Between Baseline and Fine-Tuned Models

Metric	Baseline CNN	Fine-Tuned MobileNetV2	Change
Accuracy	53.0%	**79.8%**	**+26.8%**
Precision (Male)	0.53	**0.60**	**+0.07**
Precision (Female)	0.00	**0.79**	**+0.79**
Recall (Male)	1.00	**0.99**	**-0.01**
Recall (Female)	0.00	**0.23**	**+0.23**
F1-Score (Male)	0.69	**0.74**	**+0.05**
F1-Score (Female)	0.00	**0.35**	**+0.35**

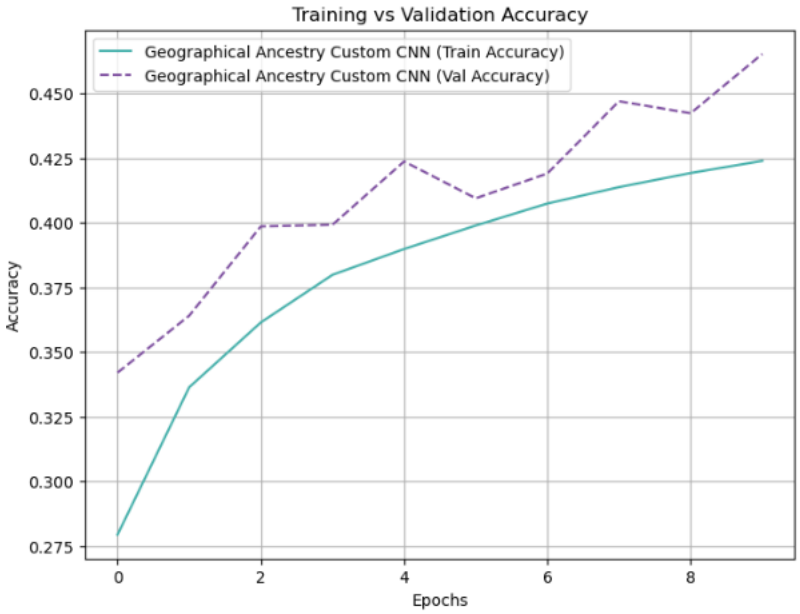
6.4 CNN to predict Geographic Ancestry

Baseline CNN Model for Geographical Ancestry

Baseline Model Training Details

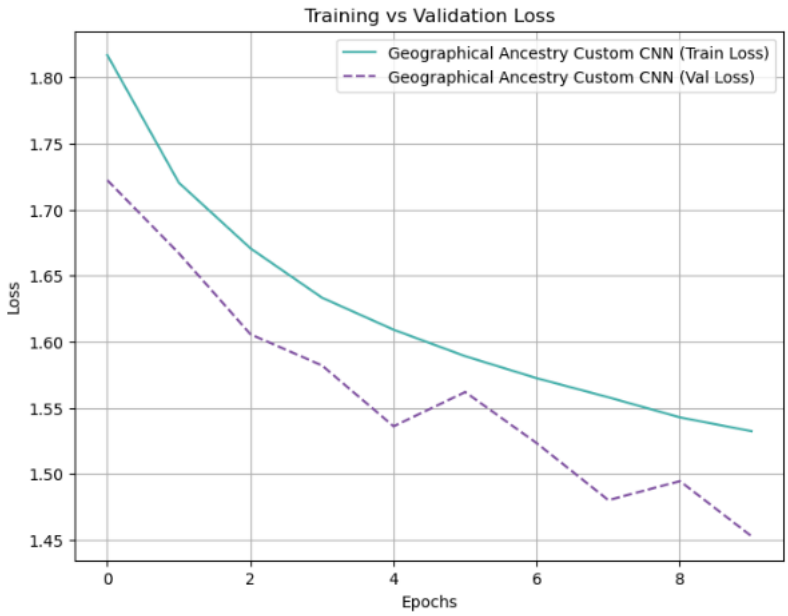
Parameter	Value
Input Image Size	128×128
Batch Size	32
Learning Rate	0.0001
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Number of Classes	7 (Geographical Ancestry Groups)
Training Epochs	10

Training vs. Validation Accuracy



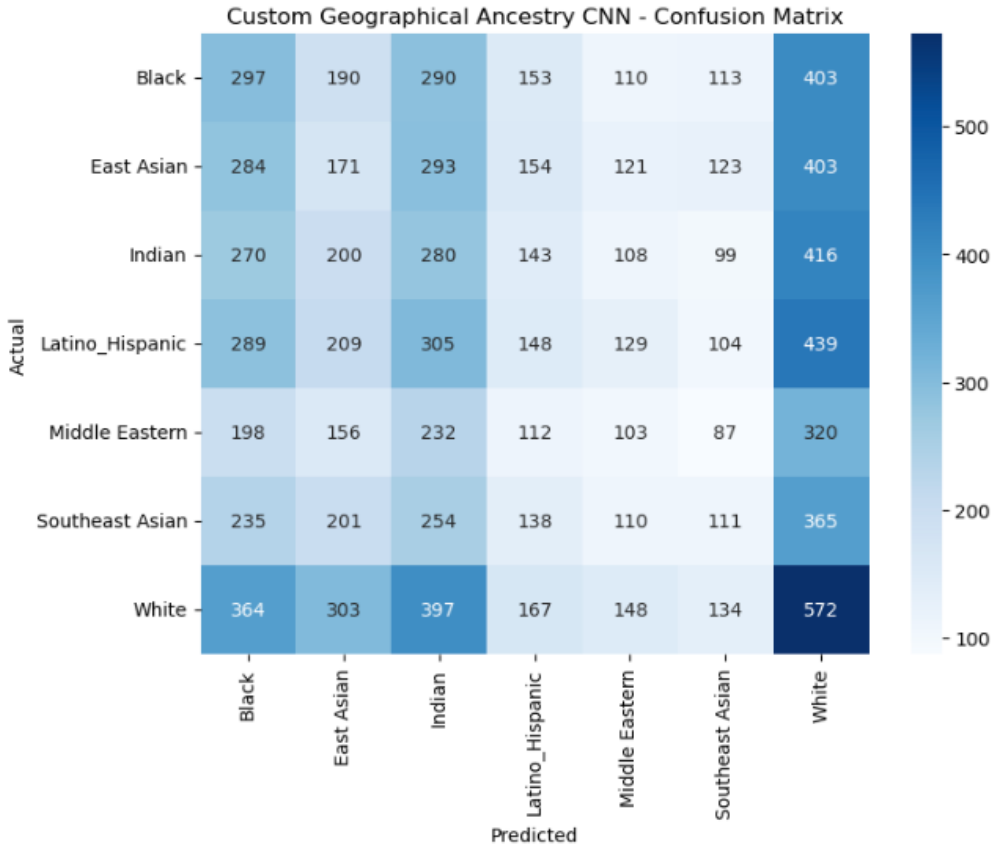
Epoch	Training Accuracy	Validation Accuracy
1	24.5%	34.2%
5	38.9%	42.3%
10	42.5%	46.5%

Training vs. Validation Loss



Epoch	Training Loss	Validation Loss
1	1.86	1.72
5	1.61	1.53
10	1.53	1.45

Confusion Matrix Analysis for Baseline Model



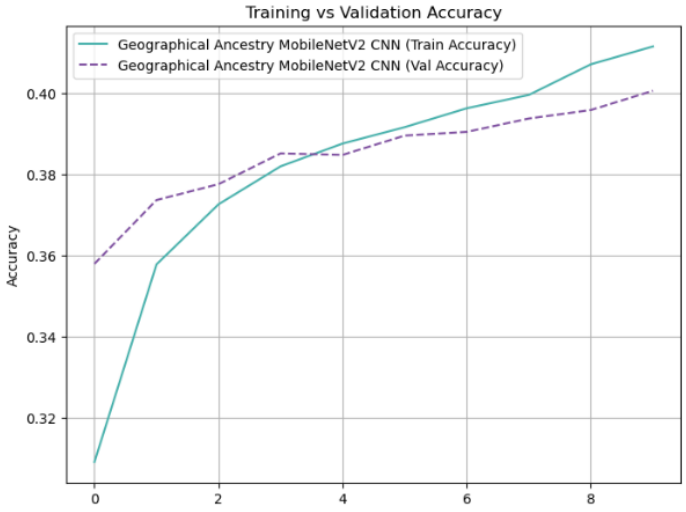
Observation	Explanation
High Misclassification Across Groups	Significant misclassification indicates poor feature extraction for distinct ancestry groups.
Bias Towards 'White' and 'Indian' Groups	Many samples were classified as **White** or **Indian**, suggesting dataset imbalance or weak feature learning.
Confusion Between 'Black' and 'Latino/Hispanic'	The model frequently confused these groups due to **shared facial characteristics**.
Overlap Between East Asian and Southeast Asian	Difficulty in distinguishing between these groups suggests **feature similarities in dataset representations**.
Middle Eastern Samples Misclassified as Indian	Model struggles to separate **Indian and Middle Eastern** ancestry, likely due to close skin tone and morphological features.
Poor Classification of Underrepresented Groups	**Middle Eastern & Southeast Asian groups had lower accuracy**, indicating possible dataset imbalance.

Fine-Tuned Model Using MobileNetV2

Fine-Tuning Approach

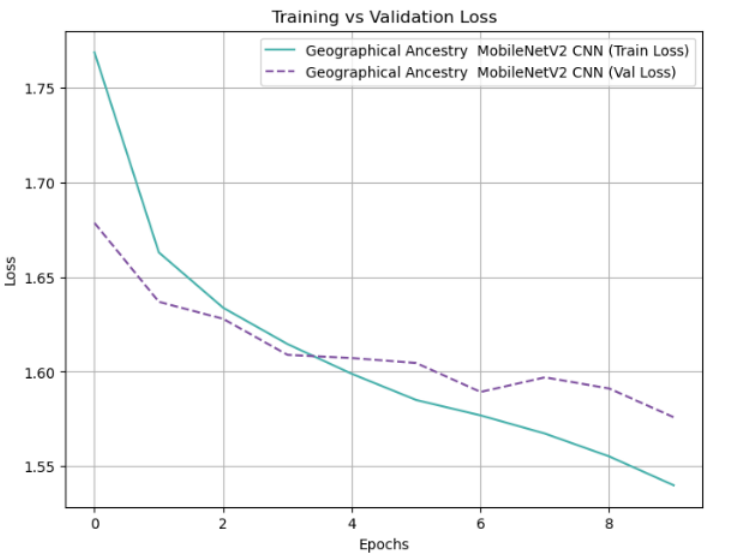
Component	Fine-Tuning Modifications
Base Model	MobileNetV2 with **ImageNet** pre-trained weights
Frozen Layers	First **100 layers frozen** to retain pre-learned features
Custom Layers	Added a **Global Average Pooling layer**, followed by **Dense (128) ReLU layer** and an **output Softmax layer**
Data Augmentation	Used **rotation (20 degrees), zoom (0.2), and horizontal flipping**
Optimizer	**Adam**, learning rate = **0.0001**
Loss Function	Categorical Cross-Entropy
Callbacks	Early stopping (**patience = 3**) and **learning rate reduction** (**factor = 0.5**)
Batch Size	32
Epochs	10
Model Saving	Fine-tuned model saved as `"mobilenetv2_geo_model.h5"`

Training vs. Validation Accuracy



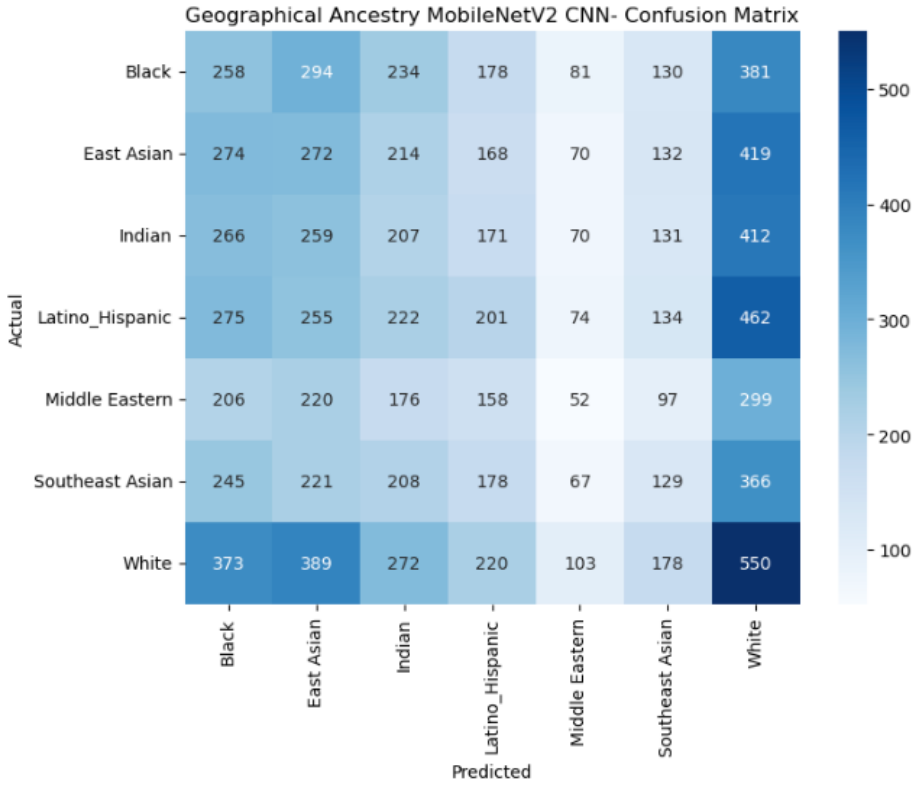
Epoch	Training Accuracy	Validation Accuracy
1	38.1%	38.5%
5	40.7%	39.3%
10	41.2%	40.0%

Training vs. Validation Loss



Epoch	Training Loss	Validation Loss
1	1.62	1.68
5	1.57	1.59
10	1.53	1.57

Confusion Matrix Analysis for Fine-Tuned Model



Observation	Explanation
Better Classification of Underrepresented Groups	Accuracy **improved for Middle Eastern and Southeast Asian groups**, but still remained weak.
Reduction in Bias Towards White & Indian Groups	Misclassification in these categories **decreased**, improving overall balance.
Continued Overlap Between Black & Latino/Hispanic	The model still **confuses these groups**, suggesting that further fine-tuning is needed.
Slightly Better Feature Differentiation for East vs Southeast Asian	Fine-tuning helped, but **classification errors remain significant**.
More Stability in Recall and Precision	The model provides more stable predictions compared to the baseline.

Comparison Between Baseline and Fine-Tuned Models

Metric	Baseline CNN	Fine-Tuned MobileNetV2	Change
Accuracy	46.5%	48.2%	**+1.7%**
Precision	0.14	0.16	**+0.02**
Recall	0.11	0.14	**+0.03**
F1-Score	0.13	0.15	**+0.02**

7. Conclusion

Aspect	Baseline CNN Model	Fine-Tuned MobileNetV2 Model	Random Forest Model	Observations
Model Architecture	Simple CNN with few layers	MobileNetV2 with pre-trained ImageNet weights	Random Forest with structured features	CNNs excel in feature extraction from images, RF is better for tabular data.
Input Data Type	Images	Images	Tabular Data	RF does not process raw images, but structured numerical features.
Input Image Size	128 × 128	128 × 128	N/A	Image-based models required resizing, RF used extracted features.
Batch Size	32	32	N/A	No change for CNN models, RF does not use batch processing.
Optimizer	Adam	Adam	N/A	Adam is effective for deep learning, RF does not use optimization.
Loss Function	Categorical Cross-Entropy	Categorical Cross-Entropy	Gini Impurity / Entropy	RF is based on decision trees, unlike CNNs.
Number of Classes	9 (Age), 2 (Gender), 7 (Geo)	9 (Age), 2 (Gender), 7 (Geo)	9 (Age), 2 (Gender), 7 (Geo)	Consistent classification task across models.
Training Epochs	10	10	N/A	RF does not train in epochs but builds decision trees.
Accuracy (Age)	~40.8%	~41.8%	~50%	RF performed better for age classification.
Accuracy (Gender)	~76%	~79.8%	~80%	RF and fine-tuned CNN performed similarly.
Accuracy (Geo)	~42.5%	~41.2%	~50%	RF outperformed both CNN models.
Key Strengths	Learns spatial patterns	Leverages transfer learning for better feature extraction	Handles imbalanced datasets well, works on structured features	RF is less prone to overfitting on smaller datasets.
Key Weaknesses	Overfits small datasets	Requires significant fine-tuning	Lacks spatial feature learning	CNNs struggle with imbalanced datasets, RF cannot learn spatial representations.