Tatiana Eng, Matt Kilmer, Orlando Marin

Artificial Intelligence - CSC 481

July 2, 2025

Gender Classification Using Facial Landmark Ratios from the AR Face Database

Team Members

- **Orlando Marin** led the efforts in data cleaning and feature engineering for the project. He implemented the K-Nearest Neighbors (KNN) classifier and contributed an additional model using Random Forest. Orlando also split the dataset into gender-balanced training and test sets based on unique person identifiers to prevent data leakage. He served as the primary author of this report, with input provided by Tatiana and Matt.
- **Tatiana Eng** implemented the Artificial Neural Network (ANN) and Naive Bayes classifiers. She created the ROC curve and confusion matrix visualizations used in the presentation and report. Tatiana, along with Matt, designed the presentation slides, incorporating input from Orlando.
- **Matt Kilmer** implemented and tuned both the Decision Tree and Support Vector Machine (SVM) classifiers. Matt, along with Tatiana, designed the presentation slides, with contributions from Orlando.

Introduction

Facial analysis using computer vision and machine learning has become a popular area of research. In this project, we explore the problem of gender classification using facial landmark data. Our approach involves extracting engineered geometric features from facial points and training several classification models to determine whether an image depicts a male or female face.

Problem Description

The task is to classify a person's gender using facial images from the AR Face Database. Each image contains 22 consistent landmark points (each with x and y coordinates), which represent key facial features such as eyes, eyebrows, nose, and lips. These points allow us to define ratios that describe geometric relationships unique to individuals' faces.

Objective

Our goal is to extract meaningful features from facial landmarks and evaluate the performance of five machine learning classifiers (K-Nearest Neighbor, Naïve Bayes, Artificial Neural Network, Decision Tree, Support Vector Machine), plus one bonus model (Random Forest), for a total of six classifiers, in accurately predicting gender. The models are compared using evaluation metrics including accuracy, precision, recall, AUC, and confusion matrix.

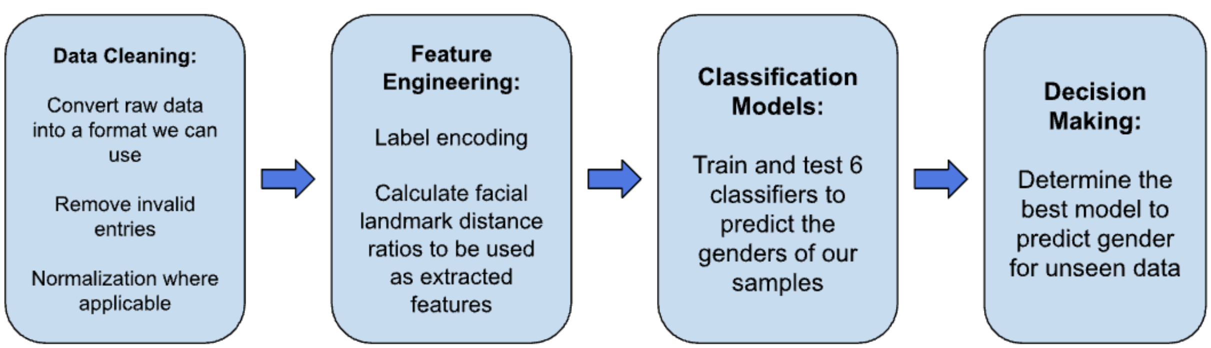
Dataset

We used the publicly available AR Face Database. This dataset includes frontal face images of 136 individuals (76 men and 60 women), each labeled with their gender and a specific facial expression or lighting condition.

Data Source and Format

The image files are labeled using the format "m-xx-yy.pts" or "w-xx-yy.pts", where "m" or "w" denotes gender, "xx" is a person identifier, and "yy" indicates the lighting or expression condition. Each file contains 22 pairs of (x, y) coordinates marking key facial features.

Framework of Overall Procedure



Preprocessing

Data preprocessing steps included removing rows with invalid entries, parsing gender and metadata from file names, and label-encoding the gender as "0" for female and "1" for male. For models sensitive to feature scale (KNN, SVM, ANN), we applied z-score normalization using scikit-learn's StandardScaler to improve convergence and distance-based calculations.

Feature Engineering

We extracted seven ratio-based geometric features from the landmark coordinates, using Euclidean distances as described in the project instructions:

- **Eye Length Ratio**: The length of the longer eye (either between points 0 and 1 or 2 and 3) divided by the distance between points 8 and 13.
- **Eye Distance Ratio**: The distance between the centers of the eyes divided by the distance between points 8 and 13.
- **Nose Ratio**: The distance between points 15 and 16 divided by the distance between points 20 and 21.
- **Lip Size Ratio**: The distance between points 2 and 3 divided by the distance between points 17 and 18.
- **Lip Length Ratio**: The distance between points 2 and 3 divided by the distance between points 20 and 21.
- **Eyebrow Length Ratio**: The longer of the distances between (4,5) or (6,7), divided by the distance between points 8 and 13.

- **Aggressive Ratio**: The distance between points 10 and 19 divided by the distance between points 20 and 21.

Train-Test Split

We used an 80/20 train-test split based on a gender-balanced, person-wise approach. Each gender group was independently split by a unique person ID, ensuring that no individual's images appeared in both training and test sets. This method preserved equal male/female representation across both sets while preventing data leakage.

Classifier Implementation and Tuning

We tested six different classifiers. For K-Nearest Neighbors, we evaluated values of k from 1 to 50 and found the best performance at k = 18. Random Forest was tuned across multiple tree counts and depths, with the best setup including 30 trees and a max depth of 15. The Artificial Neural Network achieved its best performance with one hidden layer of 100 neurons, a learning rate of 0.01, and a tolerance of 0.01. Naive Bayes was used without tuning. For Decision Tree, we used grid search with class weighting to find the best values for maximum depth of each tree, minimum number of samples needed to split a node, and the minimum number of samples needed in a leaf. Support Vector Machine performed best with a radial basis function (RBF) kernel and a regularization value (C) of 1.

Evaluation Metrics

We evaluated each classifier using accuracy, precision, recall, area under the ROC curve (AUC), and the confusion matrix. Accuracy measures overall correctness. Precision measures how often positive predictions are correct. Recall indicates how well the model identifies true positives. AUC reflects the model's ability to separate the two classes. The confusion matrix breaks down model predictions into true positives, true negatives, false positives, and false negatives.

Results and Analysis

The best-performing model was Random Forest, with 77.27% accuracy, 76.00% precision, 89.06% recall, and an AUC of 0.7362.

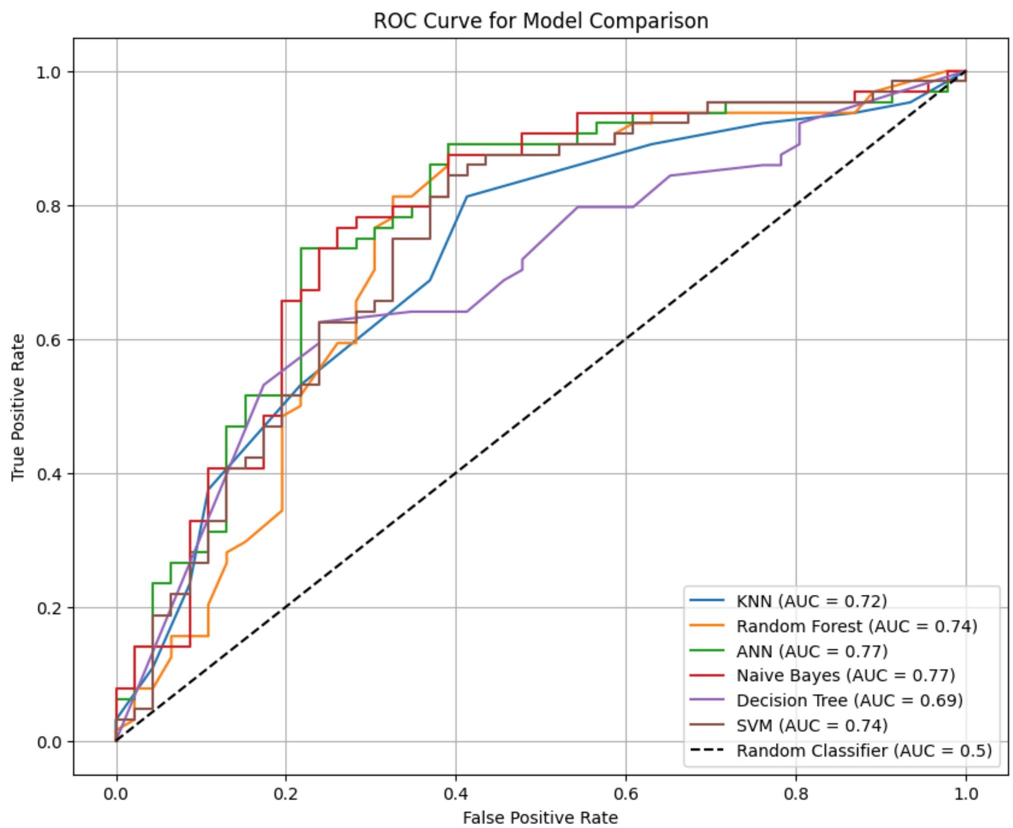
The Artificial Neural Network was a close second, which achieved an accuracy of 76.36%, precision of 75.68%, recall of 87.50%, and an AUC of 0.7741.

KNN achieved 71.82% accuracy, 73.24% precision, 81.25% recall, and an AUC of 0.7215.

Naive Bayes also performed competitively with 72.73% accuracy, 75.00% precision, 79.69% recall, and an AUC of 0.7738.

SVM showed strong recall at 89.06% and an AUC of 0.7446 but had lower precision (69.51%) and accuracy (70.91%).

The Decision Tree had the lowest performance with 63.64% accuracy, 67.65% precision, 71.88% recall, and an AUC of 0.6929.

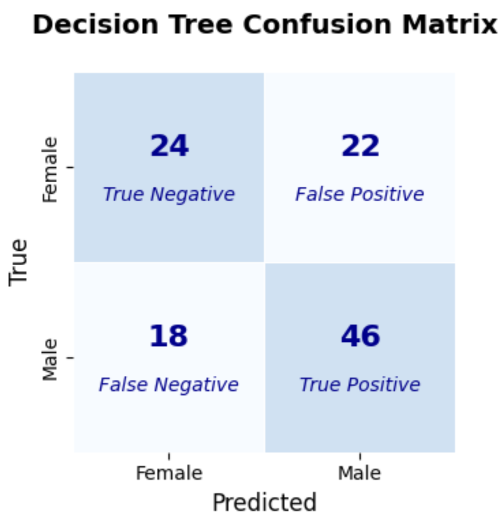


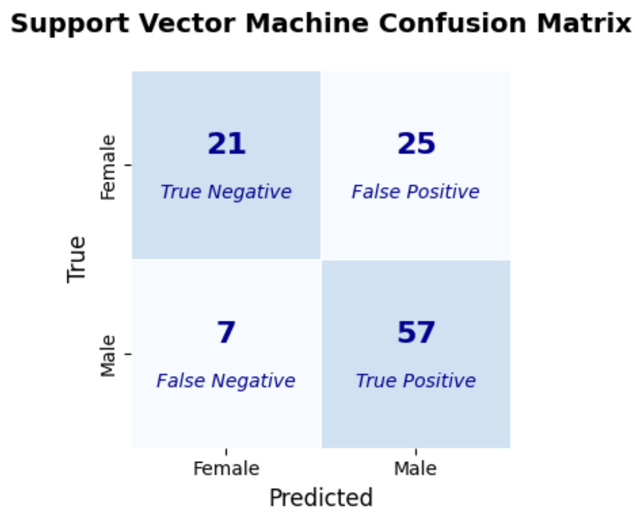
Results: Accuracy, Precision, and Recall Table

Random Forest produced the highest accuracy and precision and tied with Support Vector Machine for the highest recall.

Model	Accuracy	Precision	Recall
KNN	71.82%	73.24%	81.25%
Random Forest	77.27%	76%	89.06%
ANN	76.36%	75.68%	87.50%
Naive Bayes	72.73%	75%	79.69%
Decision Tree	63.64%	67.65%	71.88%
SVM	70.91%	69.51%	89.06%


KNN Confusion Matrix
27
True Negative
19
False Positive
12
False Negative
52
True Positive
Female
Male
True
Female
Male
Predicted
Random Forest Confusion Matrix
28
True Negative
18
False Positive
7
False Negative
57
True Positive
Female
Male
True
Female
Male
Predicted
ANN Confusion Matrix
28
True Negative
18
False Positive
8
False Negative
56
True Positive
Female
Male
True
Female
Male
Predicted
Naive Bayes Confusion Matrix
29
True Negative
17
False Positive
13
False Negative
51
True Positive
Female
Male
True
Female
Male
Predicted





Conclusion

Our findings demonstrate that facial landmark ratios can be powerful features for gender classification tasks. The Random Forest achieved the most balanced and accurate results, followed closely by the Artificial Neural Network model. Both models proved effective at capturing the non-linear geometric relationships inherent in facial structure.

KNN, Naive Bayes, and SVM also delivered reliable results, especially in terms of recall. Although the Decision Tree model lagged behind in overall performance, it still achieved moderate recall, suggesting some utility in contexts where recall is more critical than precision.

The consistent results across models validate the effectiveness of the feature engineering approach and person-wise data split strategy used in this study.

Limitations

One limitation of this study is the class imbalance present in the dataset, with 301 male images compared to only 208 female images. Although we used stratified and balanced splitting techniques to address this, the imbalance may still influence model performance.

Another limitation is the restricted number of facial landmarks, only 22 points per image, which limits the amount of spatial detail the model can learn.

Finally, the dataset is lacking in categorical data such as the subject's age group or skin tone, which could enhance the model's ability to generalize across a broader population.

Future Work

Future studies can address these limitations by using a more gender-balanced dataset and including additional landmark points to capture more nuanced facial features. Expanding the feature set to include symmetry measures, facial angles, or texture-based descriptors may also enhance model accuracy. Furthermore, incorporating demographic data such as age group or skin tone could provide valuable context to the model, improving both accuracy and fairness in real-world applications.

Appendix

- Values of seven extracted features:
  https://github.com/orlandojmarin/gender-identification-from-face-images/blob/main/Extracted%20Feature%20Values.csv
- Source code:
  https://github.com/orlandojmarin/gender-identification-from-face-images/blob/main/Code_Gender_Identification_from_Face_Images.ipynb

---

References

- AR Face Database
- https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/
- https://www.geeksforgeeks.org/machine-learning/naive-bayes-classifiers/