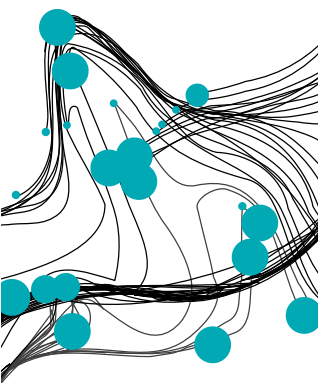




.75288

DMB

DATA MANAGEMENT
AND
BIOMETRICS

EXPLAINABLE FORENSIC FACE RECOGNITION BASED ON FISWG FEATURES

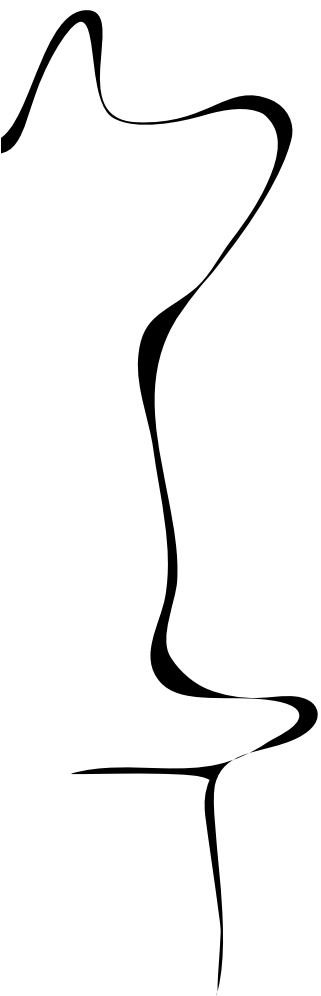
Thimo Veldema

MASTER'S ASSIGNMENT

Committee:
Luuk Spreeuwers
Moritz Hahn

May, 2026

Data Management and Biometrics
EEMathCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands



Explainable Forensic Face Recognition based on FISWG Features

Thimo Veldema, Luuk Spreeuwiers

Submission date: May 19, 2026

Modern face recognition systems function as black boxes, providing high accuracy but lacking the transparency required for forensic and legal utility. This research introduces the Explainable Forensic Face Recognition (EFFR) model, an architecture designed to be explainable by design. The system utilizes a hybrid methodology, combining AI-driven tasks—such as HRNet for landmarking and BiSeNet for face segmentation, with classical computer vision to extract geometric and textural features from the eyebrows, ears, skin, and face proportions. These features are integrated into a hierarchical framework that transforms comparison scores into Likelihood Ratios (LR), offering a measure of evidence strength. While EFFR's Equal Error Rate of 4.2% currently falls below deep learning models like VGG-Face (0.4%), it outperforms old deep learning benchmarks. These results suggest that interpretable architectures can achieve competitive performance while being explainable and transparent for forensic applications.

Face Recognition; Explainable Face Recognition; Forensic Face Recognition; FISWG;

1. Introduction

The rapid advancement of deep learning has revolutionized face recognition, with state-of-the-art (SOTA) models now achieving near-perfect accuracy in ideal conditions. However, in the domain of forensic face recognition, accuracy alone is insufficient. Forensic examiners and legal systems require explainability. Deep learning systems, such as ArcFace or VGG-Face, operate as black boxes, generating high-dimensional embeddings that are powerful but unintelligible to a human jury.

There have been attempts to retroactively make face recognition systems more explainable using post-hoc techniques. However, these methods often merely provide a heatmap of pixel importance, explaining what the model looked at, instead of why it made the decision. These retrospective explanations are inherently limited. Instead it is contended that it is more robust to build an explainable by design face recognition system from the ground up; an intrinsically transparent architecture provides more reliable evidence and verifiable logic than a system explained after the fact.

The fundamental issues with explaining black-box models via post-hoc methods are primarily that these explanations are often unfaithful to the original model, providing a simplified approximation rather than a true reflection of the underlying decision logic. Furthermore, they lack the detail necessary for forensic utility; for instance, while a saliency map might highlight the ear region as a point of similarity, it fails to explain why, such as the specific curvature of the helix or the attachment of the lobule [1]. Consequently, there is a critical need for face recognition systems built upon intrinsically interpretable models.

To create an explainable face recognition system that can be used in a forensic context, the FISWG standard was used. FISWG is an international body dedicated to developing guidelines and best practices for the forensic of human faces. FISWG provides a comprehensive list of descriptors for facial components, including: eyebrows, ears, and skin. By adopting the FISWG standard as the foundation for the Explainable Forensic Face Recognition (EFFR) model, this project ensures that the machine's "reasoning" is directly aligned with forensic standards, and as the FISWG is made for humans, its also directly explainable.

However, translating qualitative human standards into an explainable face recognition system presents several technical challenges. To evaluate the viability of this approach, this paper seeks to address the following research questions:

- How to automatically extract FISWG features that can be used for explainable face recognition?
- How to combine the extracted FISWG features in an explainable face recognition system with good performance?
- What is the performance compared to state of the art face recognition?

2. Related Work

2.1. Explainable Face Recognition

Existing post-hoc explainability methods generally fall into two categories: local perturbations and gradient-based visualizations. Saliency Maps, such as Class Activation Mapping (CAM) [2], highlight specific pixels that the model weighted most heavily during a prediction. Additionally, model-agnostic approaches like LIME [3] attempt to explain individual predictions by repeatedly masking out certain regions of the input to observe changes in the output score.

Lestriandoko et al. [4] demonstrated that certain facial features carry more weight in recognition than others, identifying the eyebrows as the most critical component, followed by the eyes or nose, while the mouth was the least informative. To reach this conclusion, the researchers employed advanced morphing techniques to neutralize specific facial regions without creating digital artifacts, subsequently evaluating the impact across four different deep recognition systems. Their findings further revealed that facial texture contributes more significantly to successful identification than the overall geometric shape of the face.

2.2. Forensic Face Recognition

The field of forensic facial comparison relies heavily on structured morphological analysis. The Facial Identification Scientific Working Group (FISWG) provides the global standard for this discipline, defining a comprehensive framework for human examiners to describe facial characteristics. Central to this framework is the FISWG "Facial Image Comparison Feature List," which identifies 19 distinct facial components, such as the forehead, eyebrows, eyes, nose, and mouth [5].

These components are subdivided into two levels of detail: component characteristics, which are features of the facial components; and characteristic descriptors, which provide detailed descriptors like shape, size, and orientation for the component characteristics. However, while FISWG provides a standardized vocabulary for these features, it does not define the specific comparison process itself or provide a weight for the importance of each feature. Consequently, while it standardizes what an examiner should look at, it does not offer a method for how to compare those features.

In forensic face recognition, the transition from investigative use to court admissible evidence relies on the Likelihood Ratio (LR) framework, which provides the strength of a match. As detailed by Jacquet and Champod [6], because automated systems produce numerical similarity scores rather than direct identities, researchers use Score-based Likelihood Ratios (SLR) to interpret these results. This method calculates the ratio between the probability of a specific score occurring if the images are from the same person, versus the probability if they are from different people. By modeling both within-source variability and between-sources variability, the SLR

offers a standardized, transparent, and empirical way to communicate the weight of facial evidence to a court.

2.3. Extraction of facial features

The focus of this section is on the classical computer vision methods used to automatically extract features, rather than the physical appearance of the facial parts themselves. Classical computer vision methods are emphasized because they are more explainable; unlike deep learning, we know exactly which physical attributes the algorithms are measuring.

2.3.1. Eyebrow

The process of eyebrow recognition typically follows a three steps, beginning with the segmentation of the eyebrow, followed by the extraction of distinctive features, and concluding with a comparison of those features. Regarding segmentation, researchers have employed various methodologies to isolate the region of interest. Garcia-Ramirez et al. [7] utilize a hyperbolic tangent filter applied specifically to the red color channel to enhance image contrast, after which a threshold operator separates the eyebrow and eye regions; this process is refined by applying morphological closing filters, smoothing the result, and finally employing the Otsu thresholding algorithm. Alternatively, Hoang Ngan et al. [8] implement an active shape model approach that defines three distinct regions: foreground, background, and undetermined, using landmarks. In this method, the shortest path from each undetermined pixel to pixels in the known regions is calculated, and the pixel is subsequently classified as either foreground or background based on a combination of edge weight, smoothness, and distance.

Techniques for eyebrow feature extraction are generally categorized into geometric-based, appearance-based, and dimensionality reduction-based approaches. Geometric-based methods focus on the physical shape of the eyebrow. For instance, Dong et al. [9] manually extract the eyebrow to analyze three global shape features: rectangularity, eccentricity, and isoperimetric quotient, alongside local area and critical point features. Similarly, Zeinstra et al. [10] utilize Fourier Shape descriptors in conjunction with FISWG features to describe both the shape and hair distribution, achieving performance levels comparable to those of Dong et al. [9]. In contrast, appearance-based techniques rely on global feature descriptors; Turkoglu et al. [11] specifically utilize LBP, WLD, and HOG as the primary extractors for recognition. Finally, dimensionality reduction techniques leverage mathematical transformations to isolate key data. Jun Bin et al. [12] apply wavelet transformations for feature extraction and utilize a Support Vector Machine (SVM) for the final classification task.

2.3.2. Ear

The ear recognition process generally consists of three primary steps: segmentation, feature extraction, and classification [13]. The first step is segmentation, where the ear is isolated from the rest of the image. This is followed by feature extraction, which prepares the data for the final classification step.

Methods for ear recognition are broadly divided into two categories: geometric-based and appearance-based techniques.

Geometric-based techniques rely on the ear's unique shapes, such as the helix, tragus, antihelix, and concha to extract structural features [13]. This can be achieved using algorithms like Canny edge detection to map the contours of these specific shapes [14]. Once the contours are mapped, features can be extracted. For example, Polin et al. [15] use reference lines to locate specific points along the ear's contour; these points are then converted into a polar coordinate format to construct feature vectors.

Appearance based techniques, on the other hand, derive feature sets from pixel intensities and surface textures. These methods typically utilize a combination of feature descriptors and dimensionality reduction. For instance, Omara et al. [16] employ various feature extractors—including LBP, LPQ, and HOG, coupled with Discriminant Correlation Analysis (DCA) for dimensionality reduction

and a Support Vector Machine (SVM) for classification.

Finally, after the feature extraction phase is complete, the resulting features are categorized. Common algorithms used for this classification step include Neural Networks, SVMs, Normalized Cross-Correlation (NCC), k-Nearest Neighbors (KNN), and minimum distance classifiers [13].

2.4. Mouth

Research specifically focusing on the mouth as a biometric feature remains limited; however, work exists regarding lip print-based identification. For instance, Farrukh et al. [17] compared SURF with deep learning features to identify individuals based on minute lip details, such as ridge patterns. A primary constraint of such methods is the requirement for high resolution imagery, as these discriminative details must be clearly visible for accurate feature extraction.

2.5. Skin

The unique texture and irregularities of human skin have been explored as distinct biometric identifiers. Bianconi et al. [18] demonstrated the effectiveness of skin texture on the forearm, utilizing classical feature extractors such as Local Binary Patterns (LBP), Gray-Level Co-occurrence Matrices (GLCM), and Gabor filters to achieve high-accuracy matching.

Complementing texture-based analysis, Jain et al. [19] focused on soft biometrics by mapping facial marks like moles and freckles. Their approach employs Laplacian-of-Gaussian (LoG) filters and morphological operators to isolate these points. Identification is determined by calculating the spatial distance between detected marks; a match is recorded if the distance falls below a predefined threshold, with the final biometric score derived from the total number of matching pairs.

3. Methodology

The Methodology section is structured from general techniques to more specific applications. First, the General Methods section outlines the core processes utilized across multiple features. This is followed by a detailed breakdown of the Individual Features, each of which is subdivided into specific Descriptive Features. Finally, the chapter concludes by explaining how these features are integrated and how the resulting data is visualized and showcased.

3.1. General Methods

3.1.1. Fourier Shape Descriptors

To describe a shape, Fourier Shape Descriptors (FSD) [20] can be employed. This method represents a contour using a set of coordinates that are treated as a periodic signal, where each individual point is converted into a complex number. To analyze the shape's geometric properties, this signal is then transformed into the frequency domain using the Discrete Fourier Transform (DFT). To normalize for translation, scale, and rotation, the coefficients of these descriptors are adjusted, ensuring the resulting representation is invariant to the shape's position or orientation. Once normalized, two shapes can be compared by calculating the L_2 norm of the difference between their descriptors. This process yields a dissimilarity score, where a smaller value indicates a higher degree of similarity between the shapes.

3.1.2. Local Binary Pattern

The Local Binary Pattern (LBP) serves as a global feature descriptor designed to characterize image textures. The algorithm functions by evaluating a central pixel against its surrounding neighbors on a predefined radius. During this comparison, a binary value is assigned to each neighbor based on its relative intensity: a

neighbor is assigned a 1 if its intensity is greater than or equal to the center pixel, and a 0 otherwise. This process generates a bitmask that is converted into a decimal value. By aggregating these values across the entire image, a feature vector is produced in the form of a histogram. These histograms act as unique textural signatures, allowing for the comparison of different surfaces using statistical measures such as the Kullback-Leibler Divergence.

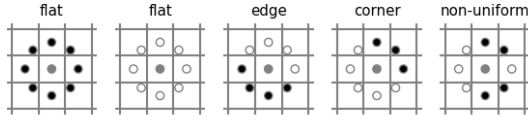


Figure 1. Local Binary Patterns [21]

As illustrated in Figure 1, different bit combinations correspond to specific geometric features. For instance, flat textures are represented by patterns where neighbors are either entirely more intense or entirely less intense than the center. In contrast, an edge is identified when there is a roughly equal split between intensity levels, whereas a corner is defined by a distinct majority of one binary state that does not constitute a flat region. Finally, non-uniform patterns are characterized by having more than two transitions between the binary states of 0 and 1.

To improve the utility of the descriptor, several variations of LBP exists. The most prominent of these is the uniform LBP, which classifies a pattern as uniform only if it contains two or fewer transitions between bits. This approach is used because non-uniform patterns are rarely found in textures and are often just noise; by focusing only on uniform patterns, the descriptor becomes more discriminative and computationally efficient. Furthermore, the LBP can be made rotation invariant by cyclically shifting the binary sequence to its minimum possible value. This ensures that identical local structures yield the same descriptor regardless of their orientation.

3.1.3. Color

To generate color features, the focus was on Hue because it identifies the actual color regardless of how bright or dark the lighting is. Since colors on a color wheel "wrap around" (like 0° and 360° both being red), circular statistics was used. The following attributes were calculated and used as features: The mean Color to find the primary shade, the Standard Deviation to see how consistent or varied the color is, and the Skewness to see if the colors are shifted in a specific direction.

3.1.4. Face Landmarks

Facial landmark detection identifies specific points on a human (as seen in Figure 2) face to map out the location and shape of different features. These points can be used to define a Region of Interest and as features in themselves. For facial landmark detection, High-Resolution Networks (HRNets) [22] pre-trained on the WFLW [23] dataset was used. The WFLW dataset was chosen because it provides the most detailed landmarks (98 points), capturing the entire eyebrow, eyes, lips, nose, and facial outline. WFLW also includes images with occlusion, blur, makeup, and a range of expressions, which makes the model more robust. HRNet was chosen because it maintains a high-resolution throughout the entire process. This prevents the loss of detail common in other models, resulting in more spatially precise landmarks.

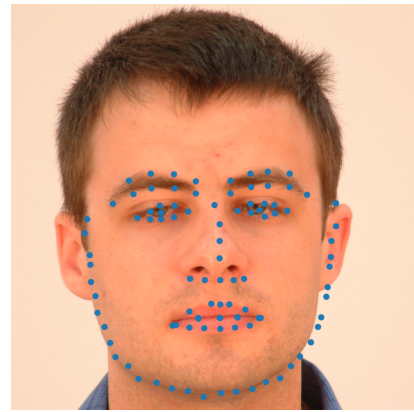


Figure 2. Face Landmarks

3.1.5. Face Segmentation

Face segmentation can be used to extract the exact shapes of facial features on a pixel level, it is a semantic segmentation task that labels every pixel in a face image into facial classes, as seen in Figure 3. For this task, BiSeNet [24] architecture (with a ResNet backbone) pre-trained on the CelebAMask-HQ [25] dataset was used. The CelebAMask-HQ [25] dataset provides detailed masks for 19 different facial classes, including the skin, nose, eyes, eyebrows, mouth, and hair. BiSeNet [24] is originally designed for real-time applications, a feature that was not necessary for this project, however it is currently the most popular and well-supported option for face segmentation. This popularity meant that reliable, high-quality pre-trained models were available, making it a good choice for this implementation.



Figure 3. Face Segmentation

3.1.6. Features to likelihood ratio

All the following facial features share a common workflow(as illustrated in Figure 4): converting feature comparisons into a likelihood ratio. To standardize this process, the features of two faces are first compared using a method appropriate for that specific feature type. For most, this involves calculating the absolute difference, while more complex features, such as LBP or Fourier Shape Descriptors, are compared using their respective similarity metrics. The resulting comparison score is then treated as a feature value itself, resulting in a new feature vector.

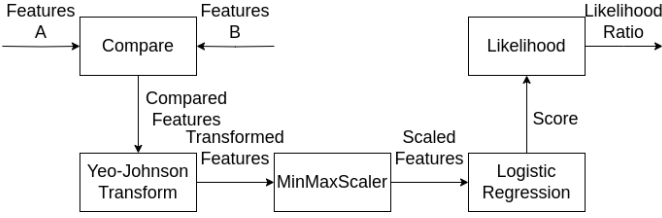


Figure 4. Features to likelihood ratio diagram

The resulting features from the comparison are transformed using the Yeo-Johnson transform [26], this brings the data in a more gaussian like distribution while also reducing the influence of outliers by mapping them closer to the distribution. The features were then transformed to a scale of 0 to 1 to ensure all values are positive. Because all features are now on the same uniform scale, the magnitudes of the logistic regression weights can be directly compared to determine each feature's contribution to the final decision. The Logistic Regression utilizes a positive intercept as a baseline 'match' score, from which weighted feature differences are subtracted. If these differences are significant, the resulting score becomes negative, indicating a non-match. This output serves as a similarity metric, where the magnitude of the score reflects the degree of similarity between the faces. Lastly, the similarity score is changed into a likelihood ratio [27], this is done by using the normalized histograms of the non-matched and matched similarity scores across the entire dataset, the matched density at a specific score is divided by the non-matched density at the same point, giving the likelihood ratio.

$$LR = \frac{P(E|H_p, I)}{P(E|H_d, I)} \quad (1)$$

H_p : Prosecution proposition
 H_d : Defense proposition
 I : Background information

3.1.7. Anomaly Detection

Anomaly detection is used because when two anomalous features match, it provides stronger evidence for a true match because such outliers are statistically rare. For example, if an individual possesses an unusually sharp ear shape that is statistically rare, its presence in both images makes the comparison far less likely to be a coincidental match. Consequently, this shared anomaly provides much stronger evidence for a true match than a common trait would. For the anomaly detection itself, the IsolationForest [28] model was selected. Unlike traditional methods that first define normal data to identify deviations, IsolationForest explicitly targets anomalies by exploiting their inherent susceptibility to isolation, making it an unsupervised model.

3.1.8. Categorization

To improve explainability, features are categorized into distinct classes (e.g., pointy vs. round ears) using Agglomerative Clustering [29] in conjunction with a k-Nearest Neighbors (k-NN) classifier. Agglomerative Clustering was selected for its ability to build a hierarchy of features without an arbitrary initial cluster count. By employing the Ward linkage criterion [30], the model minimizes cluster variance, ensuring that the resulting categories are as cohesive and distinct. Once these baseline clusters are established, a k-NN model (k=5) is used to categorize new data points based on a majority vote of their nearest neighbors, ensuring consistent classification across the dataset.

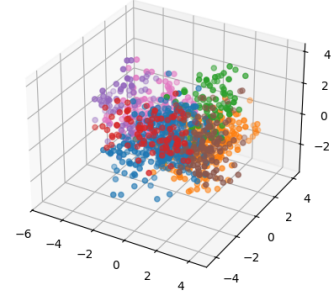


Figure 5. Clustering results visualized in 3D by using PCA

3.2. Features

3.2.1. Ears

The ear recognition pipeline (illustrated in Figure 6) begins with ear extraction using facial segmentation masks that identify individual facial components. The input image is split in the middle, and each half is processed independently. Using the segmentation mask corresponding to the ear region, the image is cropped to isolate the ear. Images below a minimum size threshold are discarded, as reduced dimensions typically indicate occlusion by hair or other obstructions.

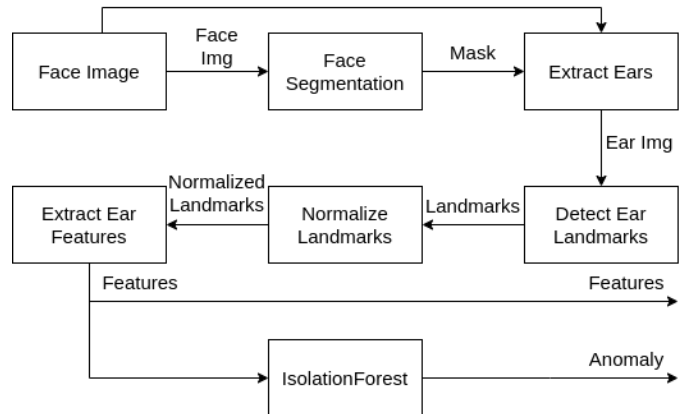


Figure 6. Ear Feature Extraction Diagram

The extracted ear images are subsequently processed through a landmark detection model specifically trained on right ear imagery using an ear landmark dataset [31]. To maintain consistency with the model's training data, left ear images are flipped around the vertical axis. The resulting landmarks can be seen in figure 7.

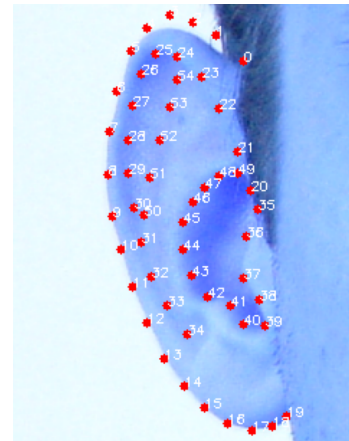


Figure 7. Ear Landmarks

Following landmark extraction, a normalization procedure is applied to ensure spatial invariance. The landmarks are first rotated such that the top and bottom attachment points align with the vertical axis. Subsequently, all landmark coordinates are scaled to fit within a unit bounding box, yielding a standardized representation invariant to orientation and scale as seen in figure 8.

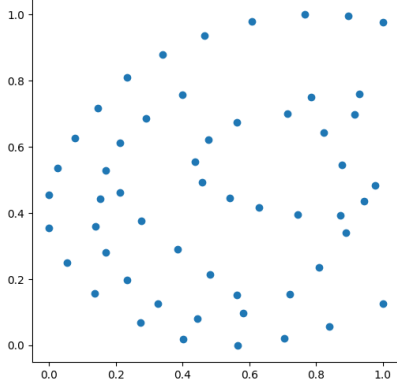


Figure 8. Normalized Ear Landmarks

Feature extraction is performed on the normalized landmarks, targeting the primary anatomical structures of the ear. As illustrated in Figure 7, the landmark model outlines the helix, concha, lobule, tragus, and anti-tragus, which serve as the basis for subsequent feature extraction.

The feature set is inspired by the Facial Identification Scientific Working Group (FISWG) guidelines [5]; however, these guidelines lack specificity regarding quantitative ear measurements. The FISWG recommendations primarily focus on qualitative descriptions of shape and size characteristics for various ear structures, without providing a standardized protocol.

To address this limitation, a novel set of quantitative measurements was developed, analogous to the FISWG's approach for eyebrow characterization. These measurements are depicted in Figure 9. The helix and lobule are characterized by the helix thickness; and the angles at the superior, and inferior attachment points. The tragus and anti-tragus are quantified by their respective indentation depths, while the concha is described by its size.

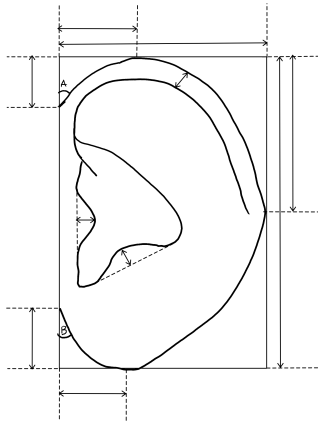


Figure 9. Ear Measured Features

To provide more comprehensive shape descriptors, two additional features are extracted from the contours. The first is a curvature measure that quantifies the degree of curvature variation along an outline. Contours exhibiting sharp angular transitions yield higher curvature scores, whereas smoothly curved outlines produce lower values.

The second feature represents each contour as a second-degree polynomial approximation. A least-squares fitting is used to generate six polynomial coefficients along with a goodness-of-fit metric, all of which are incorporated as features. These shape characterization methods are applied to the contours of the following anatomical structures: the outer ear boundary (helix and lobule combined), the concha, the tragus, the anti-tragus, and the anti-helix.

Curvature

The curvature measure is computed by analyzing the angular changes along the contour. Given a sequence of n points defining a contour, consecutive points are converted into displacement vectors:

$$\mathbf{v}_i = (\Delta x_i, \Delta y_i) = (x_{i+1} - x_i, y_{i+1} - y_i) \quad (2)$$

The angle of each displacement vector is computed using the two-argument arctangent function:

$$\theta_i = \text{atan2}(\Delta y_i, \Delta x_i) \quad (3)$$

This yields a sequence of angles representing the local direction at each point along the contour. The angular differences between consecutive segments are then calculated:

$$\Delta\theta_i = \theta_{i+1} - \theta_i \quad (4)$$

These angular differences quantify the local curvature at each position. Finally, the curvature measure J is defined as the root mean square of these angular changes:

$$J = \sqrt{\sum_{i=1}^{n-2} (\Delta\theta_i)^2} \quad (5)$$

This metric is sensitive to abrupt directional changes: localized sharp corners produce larger values, while gradual, smooth curves yield smaller values.

Parametric Shape

To capture the global shape characteristics of each contour, a parametric polynomial fitting approach is employed. First, the contour coordinates are normalized to the unit interval:

$$\tilde{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \quad \tilde{y} = \frac{y - y_{\min}}{y_{\max} - y_{\min}} \quad (6)$$

The normalized coordinates $\tilde{x}$ and $\tilde{y}$ are then independently fitted to polynomial functions of degree d using a parametric variable t . The polynomial coefficients are obtained by minimizing the sum of squared residuals:

$$\hat{a} = \arg \min_a \sum_{i=1}^n \left(\tilde{x}_i - \sum_{j=0}^d a_j t_i^j \right)^2 \quad (7)$$

$$\hat{b} = \arg \min_b \sum_{i=1}^n \left(\tilde{y}_i - \sum_{j=0}^d b_j t_i^j \right)^2 \quad (8)$$

where t_i represents the parametric position along the contour. This yields two coefficient vectors, $\hat{a}$ and $\hat{b}$, each of length $d + 1$, which collectively describe the shape of the contour.

To quantify the quality of the polynomial approximation, the coefficient of determination (R^2) is computed separately for each coordinate:

$$R_x^2 = 1 - \frac{\sum_{i=1}^n (x_i - \hat{x}_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad R_y^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (9)$$

The overall goodness-of-fit is then computed as the average of the two coordinate-specific values. The final feature vector for each contour comprises the $2(d + 1)$ polynomial coefficients from $\hat{a}$ and $\hat{b}$, along with the R^2 value, providing both a parametric shape representation and a measure of fit quality.

3.2.2. Eyebrows

Eyebrow recognition follows a specific pipeline (as shown in Figure 10). The process starts by masking the eyebrow out of the image using facial landmarks (seen in Figure 11). Features are then extracted using these landmarks to determine shape and size, while the original image data is used to analyze hair density and color. To determine the density distribution, the eyebrow is segmented into four distinct parts, and density features are extracted from each. Finally, these elements are combined into a single feature array for comparison.

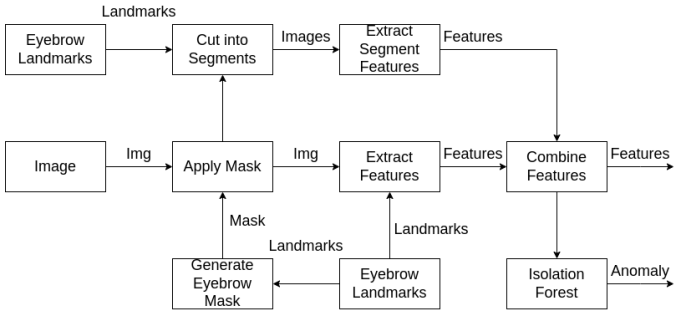


Figure 10. Eyebrow Feature Extraction

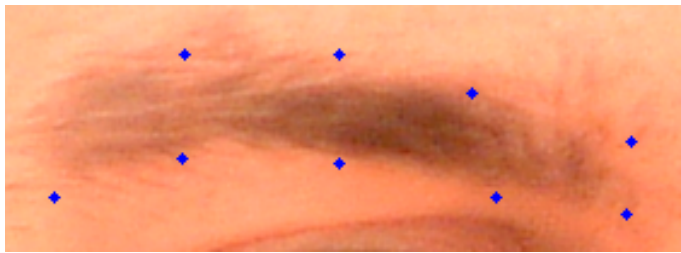


Figure 11. Eyebrow Landmarks

Shape

The shape of the eyebrow is described using Fourier shape features [20]. These features can be adjusted to make them scale, rotation, and translation invariant. Fourier shape was chosen because Chris Zeinstra et al [10]. showed that it performed quite well. However the eyebrow landmarks are not as dense and therefore don't describe the shape as well, this could impact performance, still fourier shape was chosen as its quite flexible, therefore the points from the landmarks can easily be replaced with more detailed landmarks without needing to adjust or replace the fourier shape.

The FISWG [5] also describes some features that are related to the shape of the eyebrow but also the position relative to the eye. These features are shown in figure 12. These features can easily be extracted by using the landmarks. The landmarks are also used to extract the width and height of the eyebrow. All the horizontal features are divided by the distance between the eyes, and all the vertical features are divided by the height of the nose, this is done to make sure the features are in the same scale regardless of the scale of the image itself.

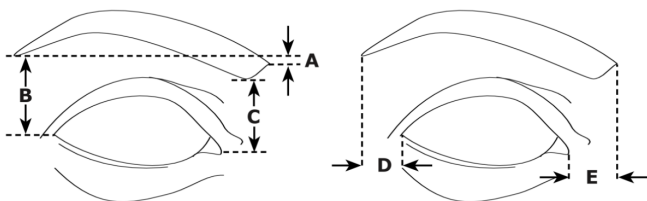


Figure 12. Eyebrow FISWG Shape Features [5]

Hair Density

The hair density is calculated using Local Binary Patterns (LBP) with rotation-invariant uniform patterns. The LBP is applied to a grayscale image, resulting in a histogram of 26 bins, where each bin represents a normalized count of the detected patterns. To determine the hair density, bins 4 through 23, which represent edges and corners, are summed into a single feature value. In this context, these edges and corners represent the individual hairs, while the remaining bins represent flat areas of skin. The final sum serves as the measure of total hair density.

Texture

To capture eyebrow texture, Local Binary Patterns (LBP) with rotation-variant uniform patterns are used. This generates a histogram; however, unlike the hair density feature, these histograms are compared directly via Kullback-Leibler divergence to produce a dissimilarity score.

Color

For the color features the focus was on the hue, the eyebrow picture is converted into a HSV format, the hues are then used to calculate the mean hue of the picture and the standard deviation of the hue distribution, these are then used as the 2 features for the color.

3.2.3. Skin

Skin recognition follows the pipeline shown in Figure 13. First the skin is extracted using face segmentation, only the skin is masked, this mask is then eroded using a cross kernel for 18 iterations, in the end the mask is applied to the image, creating an image with only skin as seen in figure 14b. 50 patches from this image are extracted, for each patch the features are extracted, all the features from the patches are then averaged into 1 feature set, these features can then be used for comparisons.

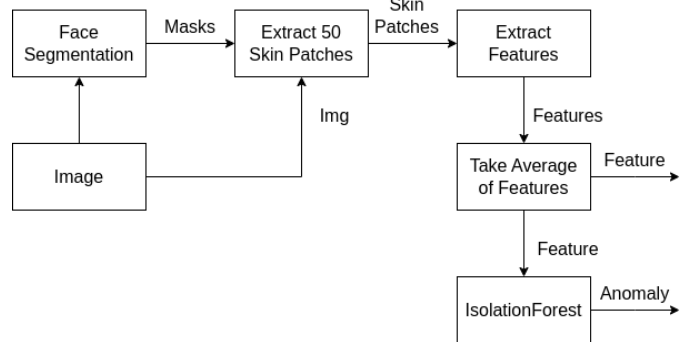


Figure 13. Skin Feature Extraction Diagram



(a). Mask

(b). Mask Applied to Image

Figure 14. Skin Segmentation

Patches

For the 500×500 pixel image, the skin mask (Figure 14a) is first eroded using a 30×30 kernel of ones to ensure that any 30×30 patch extracted remains entirely within the valid skin area. Fifty pixel locations are then chosen at random from this eroded mask to produce a set of 50 patches. Because these locations are sampled randomly, the specific patches vary with each execution; however, averaging the features across these samples allows the result to converge toward the true average of the total skin features. If the number of samples is too low, the result fails to converge to the overall skin average and instead reflects the specific characteristics of those few random patches. While increasing the sample size improves this convergence, 50 patches was found to be the optimal point where the features accurately represent the skin average without wasting computational power on negligible performance gains.

Texture

For the texture, LBP (Local Binary Patterns) and GLCM (Gray-Level Co-Occurrence Matrix) are used. LPB is explained in section 3.1.2, The uniform rotation invariant version of the LBP is used, these describe local patterns, to detect patterns at different sizes, the LBP is used for 3 different radii, 1,3, and 5 pixels.

GLCM counts the re occurring pixel combinations using a specific displacement. This results in a matrix where (i, j) i is the pixel value of the first pixel, j the pixel value of the second pixel and the value is the count of how many time this combination has occurred. This is done for different angles and distances, to make the feature rotation invariant the matrices of the different angles are summed into 1 matrix. All the matrices for all the skin patches are summed up into one matrix. Due to the matrix being sparse, Haralick features are extracted from the matrix, these Haralick features are then used for comparisons.

Color

To extract the color features, the focus was on the Hue, the skin patches are converted into a HSV format, the hues are then used to calculate the mean hue of the picture and the standard deviation of the hue distribution, these features are then averaged to give the average color features for the skin.

3.2.4. Mouth

The FISWG features chosen to be implemented for the mouth are the shape of the upper lip, lower lip, mouth, and lip fissure; and the fullness of the upper lip, lower lip.

Shape

To describe the shape of the lips, Fourier Shape Descriptors (FSD) are used. These were selected because they are versatile, capable of describing complex contours, and provide a direct similarity score when comparing different outlines. A key advantage of FSDs is that they are invariant to translation, scale, and rotation, which ensures that the comparison remains accurate even if the face is positioned at a different distance or angle in each image. However the disadvantage of FSDs is that they are not explainable.

Using the facial mouth landmarks (as seen in Figure 15), the system generates four distinct sets of Fourier descriptors to capture the full geometry of the mouth. This includes separate descriptors for the outer boundary of the upper lip, the outer boundary of the lower lip, and for the full mouth outline as a single unit. By generating these descriptors for the upper and lower lips independently, the system can compare specific parts of the mouth's geometry. These various descriptors are then compared to those of another face to produce a similarity score, which are used as comparison features.

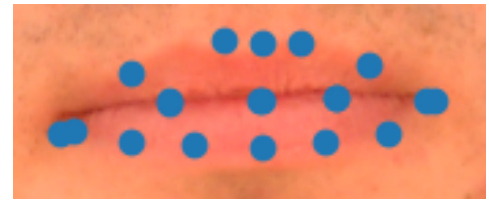


Figure 15. Mouth Landmarks

Lip Fissure

To localize the lip fissure, the Region of Interest (ROI) is first extracted using facial landmarks and converted to grayscale. Given that the fissure represents the region of lowest intensity, the ROI is processed on a column-wise basis. Each column is smoothed using a 1D Gaussian filter to reduce noise, resulting in a distinct signal dip at the fissure's vertical position. The fissure's coordinates are identified by locating the minimum intensity value within each filtered column. Finally, an intensity threshold is applied to these minimum values to define the beginning and end points of the lip fissure. These points are then used to generate a FSD, which resulting similarity score can be used as a comparison feature.

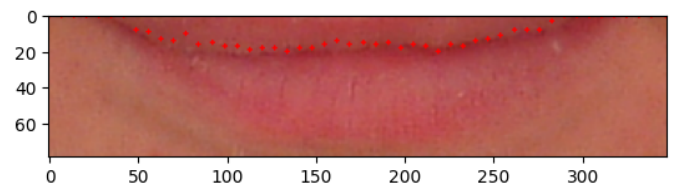


Figure 16. Lip Fissure

Fullness

While the shape of the lips is a complex feature that is difficult to quantify, lip fullness can be represented as a numerical value. This makes the feature highly explainable: a lower value indicates thinner lips, while a higher value represents fuller lips. The Upper Lip Fullness and Lower Lip Fullness are calculated as separate features. Although fullness is technically a measure of 3D volume, this study focuses on 2D frontal images. Therefore, the mean thickness, measured vertically at three distinct points across the lip from the facial landmarks, is used as a reliable proxy for total fullness.

3.2.5. Face Composition

A portion of the FISWG guidelines [5] focuses on the relative composition of the face, specifically how the position of one feature relate to another. This includes measuring the approximate width of the nose and mouth relative to eye distances, such as the interpupillary distance, as well as the width of the nose in relation to the mouth. Furthermore, the vertical layout of the face is captured by calculating the distance from the nose to the upper lip and from the chin to the lower lip, both relative to the overall face length. The positioning of the eyes and ears relative to other major landmarks are also included.

While most of these features are automatically extracted using the WFLW [23] facial landmarks, the ears are not represented in landmark sets. To account for this, the face segmentation masks are utilized to determine the location of the ears. Once all measurements are collected, they are compiled into a feature vector. Because these features are expressed as ratios, making them scale-invariant, they can be compared between two different images by calculating the absolute difference between their respective vectors.

3.2.6. Facial Marks

First marks are detected using YOLO [32], for each mark YOLO gives a bounding box, the box is used to extract an image of the mark, and the coordinates are used to find the location, however

due to the possibility of the face not being in the exact same position for each picture, the coordinates are brought into the coordinate system of the face itself. Then these coordinates and the features extracted from the patch are used to match and compare the marks, if the location is too different then the comparison is not done, if the location of both marks are within a certain range, then the marks features are compared, if the marks are a match then this is positive evidence that the faces belong to the same person, however if it is not a match, the system ignores it.

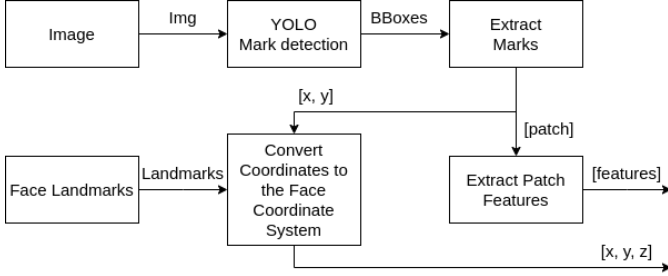


Figure 17. Facial Marks Extraction Diagram

YOLO

An YOLO 11M model is trained on a facial marks dataset [33] to detect facial marks. It is trained on a high resolution (1280x1280) so that it can detect small facial marks as seen in Figure 18.



Figure 18. Marks detected by YOLO

Face Frontalization

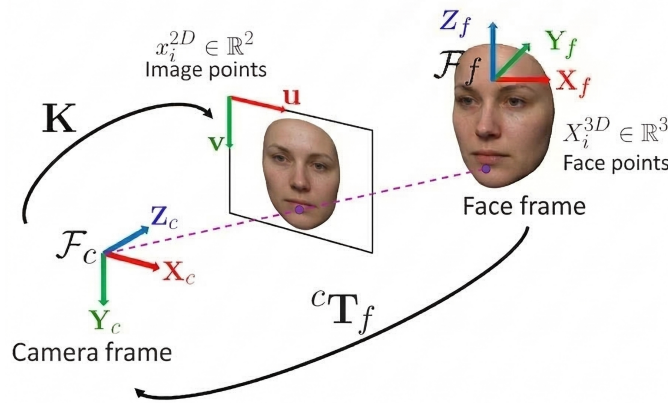


Figure 19. perspective projection model edited from [34]

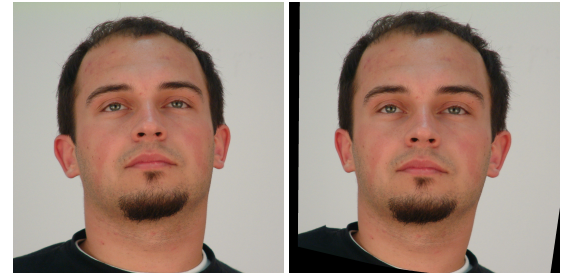
Facial landmarks and the corresponding 3D face coordinates are used to find cT_f using perspective-n-point pose computation in OpenCV [34]. This matrix is used to map points in the face coordinate system to the camera coordinate system. Then using the camera intrinsic parameters matrix the points in the camera coordinate system can be mapped to the image.

$$X = \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K[I|0]^c T_f \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (10)$$

This can also be used in reverse, to go from a point in the image to the corresponding 3D coordinates in the face frame, however the depth is missing, therefore this will result into a line instead of a point.

$$\begin{bmatrix} x \\ y \\ f(x, y) \\ 1 \end{bmatrix} = {}^fT_c \begin{bmatrix} s * K^{-1} X \\ 1 \end{bmatrix} \quad (11)$$

This equation can be solved if we know the z value, which we can get by estimating the z value using the form of the face, the form is approximated using a cylindrical shape fitted on the 3D face landmarks. Then the equation above combined with the equation for a cylindrical shape is solved. The result is the x, y and z coordinates in the face coordinate system of the facial mark.



(a). Input (b). Frontalized Output

Figure 20. Frontalization

Mark Comparison

To compare the facial marks on 2 pictures, first the distance between all the marks in the face frame are used, they are then matched based on the minimum distance, if the minimum distance between 2 marks is above a certain threshold then they are deemed different marks, if the distance is below the threshold, certain features are used to match the marks. Which are the color and texture, the texture feature is done by using LBP with a radius of 2, the color is the mean, standard deviation and skew of the Hue distribution of the mark. Then the features are compared to each other and the resulting compared features are put into a logistic regression model. Which classifies the mark as either a match or a non-match, only in the case of a match does the system treat the mark as evidence for a match. The system ignores it completely in the case of a non-match.

3.3. Combination of Features

To determine whether a face comparison result is a match or a non-match, the data from individual components, such as the ears, eyebrows, and mouth, must be combined into a single conclusion. While there are several more complex and higher-performing methods to achieve this, such as concatenating all data into a single feature vector for a machine learning model or implementing a binary voting system, this study utilizes the multiplication of Likelihood Ratios (LR). This approach was chosen specifically for its simplicity and the high explainability of the final decision rather than for its raw performance.

To prevent a single feature from disproportionately influencing the outcome, the individual likelihood ratios are clipped at pre-defined upper and lower bounds of 1000 and 1/1000. This ensures that a single outlier, cannot dominate the decision making process. By multiplying these clipped ratios, the system produces a final, likelihood ratio that reflects the cumulative strength of the evidence across all facial regions.

3.4. Report Showcase

A primary objective of this research is to make the system explainable, this is done by providing an comparison report. When two faces are compared, the system generates a structured breakdown of individual facial features, such as the eyebrows, ears, and mouth, presented in a side-by-side format with Person A on the left and Person B on the right.

To make the underlying machine learning logic transparent, each feature comparison is visualized using a diverging bar chart (Figure 21). In this visualization, bars extending to the left signify dissimilarity, indicating a likely non-match for that specific trait. Conversely, bars extending to the right represent high similarity, supporting a match.

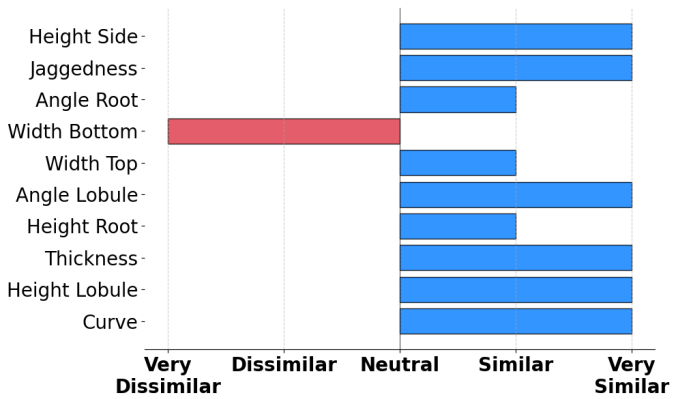


Figure 21. Explanation of the comparison between the helix of 2 ears

A five-point scale was implemented to ensure the results are interpretable. To further enhance clarity, features are automatically sorted by their weighted importance (from most significant to least significant). This approach allows a user to understand the reasoning behind a “Match” or “Non-Match” verdict at a glance, without requiring prior technical knowledge of the algorithm.

The machine learning comparison model outputs a raw similarity score for each facial trait. To translate this into a forensic context, each score is converted into a Likelihood Ratio (LR) based on the statistical distribution of known matches and non-matches. This process is visualized using a histogram (Figure 22), which displays two normalized distributions: one representing the scores of “same-source” comparisons (matches) and the other representing “different-source” comparisons (non-matches).

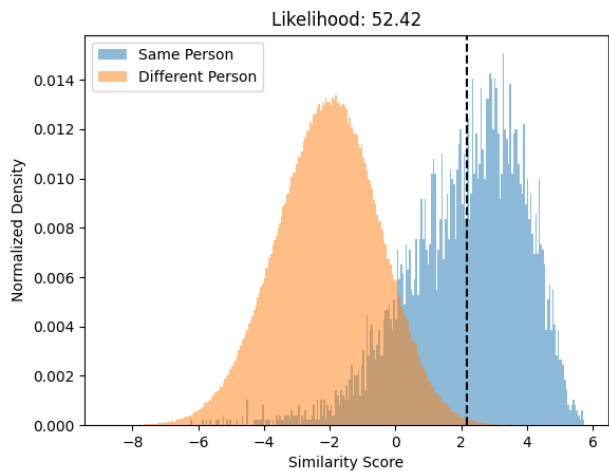


Figure 22. Histogram of similarity scores of matched and non-matched ear comparisons

During a comparison, a vertical line marks the specific similarity score of the feature being analyzed. The point at which this line

intersects both distributions determines the likelihood ratio, calculated as the height of the match distribution divided by the height of the non-match distribution. This resulting ratio is displayed in the figure title, providing a measure of how much a specific feature (such as the ear) supports the overall conclusion.

Ear Person A	Attributes	Ear Person B
True	Match	True
112.48	Likelihood	112.48
Class 3	Class	Class 8
True	Anomaly	True

Table 1. Comparison of Attributes between Subject A and Subject B

Finally a summary Table is presented, it serves as the concluding component of the feature comparison. This table shows several metrics, beginning with a Match Status that provides a binary of whether the system considers the features to be a match. The Likelihood Ratio (LR) is included to show the weight of evidence derived from the distributions. Furthermore, the table specifies the Feature Classification, and it includes an Anomaly Detection flag, indicating if the feature is an outlier.

4. Experimental Setup

4.1. Dataset

The PUT Face Database [35] was selected for this research due to its high resolution images captured under partially controlled conditions. The dataset consists of approximately 10,000 images representing 100 unique subjects. Since the primary objective of this project is the development of an explainable forensic framework rather than achieving state-of-the-art accuracy on unconstrained data, the controlled nature of the PUT dataset is advantageous. By utilizing images with consistent lighting and high clarity, the focus remains on the transparency and interpretability of the recognition process.

4.2. Preprocessing

To preprocess the dataset, faces are first detected using an OpenCV Haar Cascade and then cropped. While Haar Cascades are considered an older detection method and more modern, higher-performing alternatives are available, this approach remains effective for this specific dataset due to the high clarity of the images. Following detection, a landmark detector is applied to the image; these landmarks are used to rotate the face into an upright position based on the alignment of the eyes. Additionally, the landmarks serve as a filter to remove faces turned too far to the left or right. This is achieved by comparing the mean x-position of the head’s outline landmarks to the mean x-position of internal features, such as the eyes and eyebrows. If the horizontal distance between these two means exceeds a predefined threshold, the image is discarded. This process results in a standardized dataset where all faces are maintained in a consistent, near-frontal, and upright position.

4.3. Data Split

For the evaluation of the system, the dataset was partitioned using an 80-20 split based on subjects rather than individual images. This means that 80% of the unique individuals were used for training, while the remaining 20% were reserved for testing. By ensuring that no images of a test subject appear in the training phase, the system is forced to demonstrate its ability to generalize to unseen faces.

5. Results

5.1. Features Results

The eyebrow was the most discriminative feature in the study, outperforming all other individual components. The ear feature followed quite closely, then the skin and face composition features followed with comparable results, though face proportions performed slightly better. The mouth feature yielded the lowest performance by a significant margin.

Biometric Feature	EER	AUC
EyeBrow	0.0923	0.9691
Skin	0.1666	0.9148
Face Proportions	0.1576	0.9179
Ear	0.1103	0.9478
Mouth	0.3283	0.7446

Table 2. Performance Comparison of Biometric Features

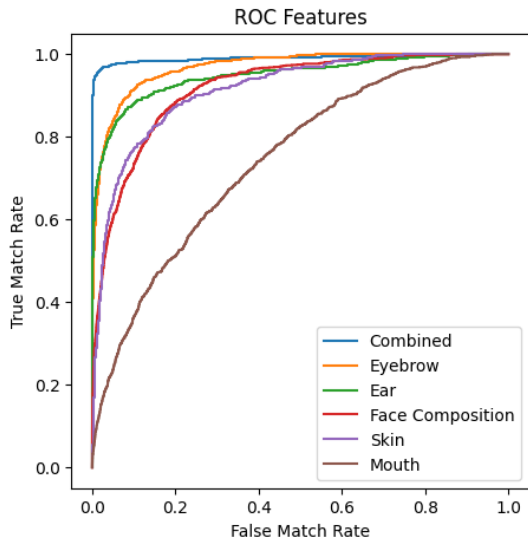


Figure 23. ROC biometric features

5.2. Combined Results

To evaluate the performance of the proposed Explainable Forensic Face Recognition (EFFR) system, it was benchmarked against several established face recognition models: ArcFace [36], VGG [37], Dlib [38], OpenFace [39], and DeepID [40]. All systems were evaluated using the DeepFace library [41], cosine similarity was used as the distance metric, and all face recognition models were tested on the exact same test set.

The results show that the proposed EFFR system with facial marks provides only a marginal improvement over the version without marks, evidenced by a slight 0.1% difference in Equal Error Rate (EER) and a 1% increase in True Positive Rate (TPR) at a False Positive Rate (FPR) of zero. ArcFace and VGG achieved the highest performance, producing nearly perfect scores on this dataset, while Dlib followed closely with slightly lower accuracy. In contrast, EFFR performed significantly worse than these state-of-the-art models, with an EER more than three times higher than that of Dlib. Finally, OpenFace and DeepID demonstrated the lowest efficacy among all evaluated systems; notably, OpenFace was outperformed not only by the individual eyebrow feature of the EFFR system but also by the skin and face composition features.

Model	EER	AUC
ArcFace [36]	0.0060	0.9998
Dlib [38]	0.0117	0.9993
VGG-Face [37]	0.0037	0.9999
EFFR with Marks	0.0302	0.9906
EFFR without Marks	0.0311	0.9905
DeepID [40]	0.1608	0.9099
OpenFace [39]	0.1936	0.8847

Table 3. Performance Comparison of Face Recognition Models

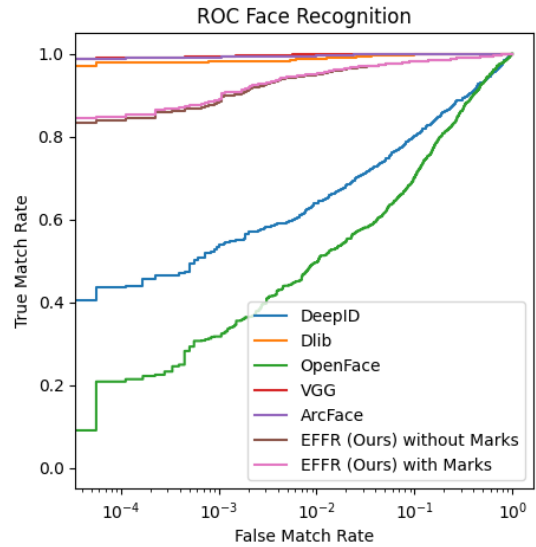


Figure 24. ROC Face Recognition Models

6. Discussion

6.1. Features

The eyebrow is the best performing feature in this set, which makes sense as it is known to be one of the most discriminative features in face recognition generally. Next are the skin and face proportion features. These performed quite equally, with skin features being slightly better. This is also expected, as while these features are not highly distinctive and can look similar across different people, they still provide significant value. Given the testing set of 20 people, it is clear that these features are distinctive enough to contribute effectively to recognition.

The ear, however, is a highly unique structure but performs a bit worse than the eyebrow, though it still performs quite well. The lower performance relative to the eyebrow is likely due to the use of frontal images. Ear recognition is typically most effective from a side profile where the ear is fully visible. In frontal images, some subjects have clearly visible ears, while others have less protruding ears where the inner structures are partially or entirely hidden. This inconsistency leads to less accurate landmarks, or a total lack of usable landmarks, which directly degrades the resulting features and the final recognition results.

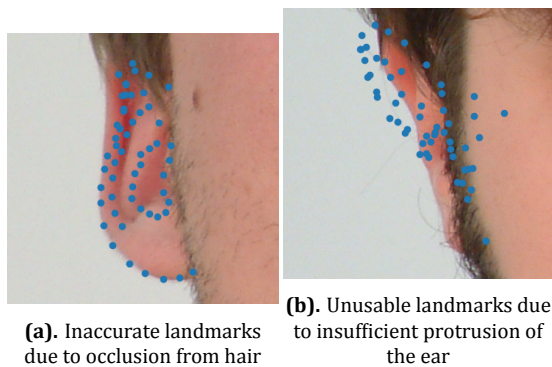


Figure 25. Poor Performing Ear Landmarks

Furthermore, the ear and face proportion features are the most explainable in this study. Unlike the other sets, they do not rely on global descriptors like Local Binary Patterns (LBP) or Fourier Shape Descriptors. While these mathematical descriptors are less explainable to a human observer, they contain more dense information. This trade-off between high information content and human interpretability likely explains why the more explainable ear features did not perform as well as the more complex feature sets.

I suspect the eyebrow shape features could be further improved. Currently, only 9 landmarks are used to define the eyebrow's shape; while this is sufficient to extract the FISWG features shown in Figure 12, it lacks the detail necessary to describe the full, distinctive contour of the eyebrow.

While face segmentation masks were considered for eyebrow segmentation, landmarks were ultimately chosen because they integrate more effectively with Fourier shape descriptors. Furthermore, landmarks offer superior robustness, as the model remains stable enough to estimate positions even under partial occlusion. Looking ahead, eyebrow recognition performance could be further enhanced while keeping the robustness by training landmark detection models with more eyebrow landmarks.

The face composition features outperformed expectations. This set consists of 21 ratios, all of which are scale-invariant and highly explainable, as they represent simple distances between facial parts. Despite their simplicity, they perform remarkably well; this is likely due to the stable nature of these features, as they do not change significantly with different facial expressions.

Because the standard facial landmarks did not include the ears, face segmentation masks were used to fill this gap. However, since these masks are less robust against occlusion, the predicted location of the ears can be inaccurate if hair or other objects obstruct the view. To further improve the performance of face composition recognition, a more robust method for locating the ears, specifically one that handles occlusion more effectively should be utilized.

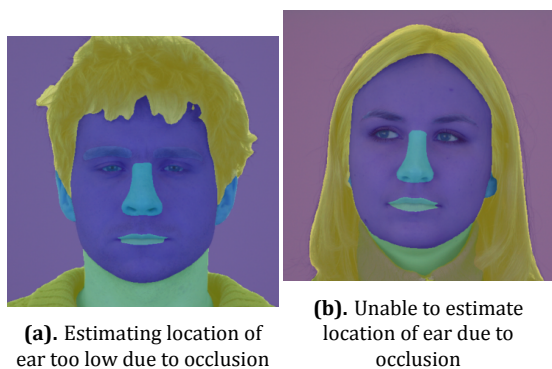


Figure 26. Face Masks

Skin recognition performs well but remains one of the least explainable feature sets in this study. This is because it relies on two global feature extractors, Local Binary Patterns (LBP) and the Gray-Level Co-occurrence Matrix (GLCM), to describe skin texture. While

these descriptors are highly effective, they primarily output a numerical similarity score based on texture patterns, offering little insight into which specific visual markers are being compared. The only truly explainable component is color; however, this was also found to be the least distinctive feature, as it lacks the complexity needed to reliably differentiate between individuals.

The mouth was the least effective feature, likely because of the limited sub-features, just the shape of various components and their color was used. Furthermore, the mouth is quite inconsistent; its shape fluctuates based on expression, which can drastically reduce performance. This restricted feature set was primarily the result of time constraints and the initial assessment that the mouth was a less promising area of focus due to the inconsistency.

6.2. Face Recognition

While EFFR falls short of state-of-the-art (SOTA) models like ArcFace, VGG-Face, and Dlib, it outperforms old deep learning benchmarks such as DeepID and OpenFace. Unlike these high-performing black-box models, EFFR prioritizes transparency; however, this explainability comes at the cost of raw performance, as individual handcrafted features cannot match the information density of deep neural embeddings. Deep learning models extract vastly more intricate facial details through training, whereas EFFR relies on handcrafted features that, due to time constraints, represent only a subset of possible facial descriptors.

Another current disadvantage of EFFR is its lack of robustness, particularly regarding pose changes which significantly impact recognition scores. EFFR utilizes a hybrid of classical computer vision and AI; while the AI components (such as facial landmarking) are highly robust, the classical methods often struggle with environmental variability. A promising future approach would involve using deep learning to extract explainable features directly, or using them as a foundation for more robust handcrafted descriptors. The decent performance already achieved suggests that face recognition can indeed be engineered to be both effective and interpretable.



Figure 27. False Negative due to extreme pose change



Figure 28. False Positive due to matching face proportion and matching right ear (left on image)

As expected, EFFR utilizing facial marks offers only a marginal improvement. This limited gain stems from the model's focus on providing positive evidence for matches rather than accounting for negative evidence. Furthermore, the YOLO model lacks the reliability needed to consistently detect marks across different images. The frontalization process introduces inaccuracies by utilizing a generic cylindrical shape as a proxy for the head instead of a precise 3D reconstruction. This lack of anatomical detail increases the margin of error when comparing landmark locations, necessitating a larger threshold that ultimately leads to more false positives. Furthermore, because frontalization requires the head's rotation, relying solely on facial landmarks makes the predicted rotation inaccurate. The frontalization process is effective at correcting roll and pitch, but struggles with yaw. Even when a face is already largely frontalized in the yaw plane, the method still attempts a correction, leading to over-correction. Additionally, it frequently miscalculates the correct rotation direction required to frontalize the yaw.

6.3. FISWG

A primary objective of this project was the automated extraction of FISWG features. While these features are ideal for explainability, having been designed specifically for human interpretation, their human-centric nature makes them exceptionally difficult for computer vision systems to extract without sacrificing that very clarity.

The FISWG standard often relies on qualitative rather than quantitative descriptions, leading to inherent vagueness that complicates the transition to measurable data. In this study, an attempt was made to map these features into discrete categories. While this categorization was successful, it resulted in significant information compression, stripping away the detail necessary for high-precision comparisons. Although this approach was only applied to high-level features like the ears and eyebrows, applying a similar categorical framework to more granular, low-level structures, such as the specific shape of the ear helix, could mitigate this information loss and improve comparative utility.

Ultimately, to develop a higher-performing system, a new standard must be established, one designed with both human interpretability and machine compatibility in mind. Such a standard should prioritize quantitative descriptions, ensuring that facial features are both measurable and human-explainable.

7. Conclusion

This research addressed the extraction and integration of FISWG features to develop an explainable face recognition system. To answer the first research question regarding automated extraction, this study demonstrates that FISWG features can be effectively recovered through a hybrid methodology. By combining AI-driven tasks such as facial landmarking and parsing—with classical computer vision techniques for geometric and textural extraction, the system successfully generates interpretable descriptors for eyebrows, skin, and face proportions.

Regarding the combination of these features, the EFFR system utilizes a three-level hierarchical approach. At the foundational level, lower-level descriptors are aggregated using an interpretable Logistic Regression model to synthesize them into high-level anatomical units, such as the ears and eyebrows. These high-level features are then transformed into Likelihood Ratios (LR) to represent the statistical strength of the evidence. Finally, these individual ratios are combined via multiplication to produce a single, unified likelihood ratio that serves as the final, transparent recognition score.

Finally, in terms of comparative performance, the EFFR system currently falls below modern state-of-the-art black-box models. The system achieved an Equal Error Rate (EER) of 3.0%, compared to the 0.4% EER reported by VGG-Face. While less performant than deep learning benchmarks, these results suggest that an explainable architecture has the potential to achieve competitive performance, offering a level of transparency and forensic utility that current state-of-the-art models lack.

8. Future Work

Future research should prioritize expanding the feature set to include the nose and eyes while implementing improved extraction methods, such as high-density landmarks for eyebrows and occlusion robust models for ear landmarks. To improve accuracy across all features, a more sophisticated frontalization process should be used. Additionally, the system's reliance on unexplainable texture descriptors highlights the need for a more explainable approach for texture analysis and a more reliable, robust implementation for facial mark detection is needed.

The project's evaluation should also be extended to larger, more challenging datasets to validate performance under diverse conditions. Ultimately, a new quantitative standard must be established, one designed specifically to bridge the gap between human-centric FISWG descriptions and machine-compatible data to ensure that high performant and explainable face recognition.

9. AI

This document was edited and refined with the help of AI tools, such as ChatGPT, for clarity, grammar, and phrasing. All technical content, ideas, and structure were created by me. I have thoroughly reviewed the entire text and take full responsibility for the content.

References

- [1] C. Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019, ISSN: 2522-5839. doi: 10.1038/s42256-019-0048-x.
- [2] R. Shadman, D. Hou, F. Hussain and M. G. S. Murshed, 'Explainable face recognition via improved localization', *Electronics*, vol. 14, no. 14, 2025, ISSN: 2079-9292. doi: 10.3390/electronics14142745. [Online]. Available: <https://www.mdpi.com/2079-9292/14/14/2745>.
- [3] A. Rajpal, K. Sehra, R. Bagri and P. Sikka, 'Xai-fr: Explainable ai-based face recognition using deep neural networks', *Wireless Personal Communications*, vol. 129, no. 1, pp. 663–680, Dec. 2022, ISSN: 1572-834X. doi: 10.1007/s11277-022-10127-z.
- [4] N. H. Lestriandoko, R. Veldhuis and L. Spreeuwers, 'The contribution of different face parts to deep face recognition', English, *Frontiers in Computer Science*, vol. 4, pp. 1–16, Aug. 2022, ISSN: 2624-9898. doi: 10.3389/fcomp.2022.958629.
- [5] Facial Identification Scientific Working Group (FISWG), *Facial image comparison feature list for morphological analysis*, <https://www.fiswg.org>, version 2.0, 2022.
- [6] M. Jacquet and C. Champod, 'Automated face recognition in forensic science: Review and perspectives', *Forensic Science International*, vol. 307, p. 110 124, Feb. 2020, ISSN: 0379-0738. doi: 10.1016/j.forsciint.2019.110124.
- [7] J. García-Ramírez, J. A. Olvera-López, I. Olmos-Pineda and M. Martín-Ortiz, 'Mouth and eyebrow segmentation for emotion recognition using interpolated polynomials', *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 5, pp. 3119–3131, 2018. doi: 10.3233/JIFS-169496.
- [8] T. H. N. Le, U. Prabhu and M. Savvides, 'A novel eyebrow segmentation and eyebrow shape-based identification', in *IEEE International Joint Conference on Biometrics*, 2014, pp. 1–8. doi: 10.1109/BTAS.2014.6996262.
- [9] Y. Dong and D. L. Woodard, 'Eyebrow shape-based features for biometric recognition and gender classification: A feasibility study', in *2011 International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1–8. doi: 10.1109/IJCB.2011.6117511.
- [10] C. Zeinstra, 'Towards the automation of forensic facial individualisation: Comparing forensic to non forensic eyebrow features', Ph.D. dissertation, University Library/University of Twente, 2017, ISBN: 9789036543750. doi: 10.3990/1.9789036543750.
- [11] M. O. Turkoglu and T. Arican, 'Texture-based eyebrow recognition', English, in *2017 International Conference of the Biometrics Special Interest Group, BIOSIG 2017*, C. Busch, A. Uhl, A. Bromme, A. Dantcheva and C. Rathgeb, Eds., 2017 International Conference of the Biometrics Special Interest Group, BIOSIG 2017, BIOSIG ; Conference date: 20-09-2017 Through 22-09-2017, United States: IEEE, Sep. 2017, ISBN: 978-1-5386-0396-3. doi: 10.23919/BIOSIG.2017.8053522.
- [12] C. Jun-Bin, Y. Haitao and D. Lili, 'Eyebrows identity authentication based on wavelet transform and support vector machines', *Physics Procedia*, vol. 25, pp. 1337–1341, Dec. 2012. doi: 10.1016/j.phpro.2012.03.241.

- [13] M. Alva, A. Srinivasaraghavan and K. Sonawane, 'A review on techniques for ear biometrics', in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2019, pp. 1–6. doi: 10.1109/ICECCT.2019.8869450.
- [14] M. Chowdhury, R. Islam and J. Gao, 'Robust ear biometric recognition using neural network', in *2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2017, pp. 1855–1859. doi: 10.1109/ICIEA.2017.8283140.
- [15] M. Polin, A. kabir and M. Sadi, '2d human-ear recognition using geometric features', Dec. 2012, pp. 9–12, ISBN: 978-1-4673-1434-3. doi: 10.1109/ICECE.2012.6471471.
- [16] I. Omara, X. Li, G. Xiao, K. Adil and W. Zuo, 'Discriminative local feature fusion for ear recognition problem', in *Proceedings of the 2018 8th International Conference on Bioscience, Biochemistry and Bioinformatics*, ser. ICBBB '18, Tokyo, Japan: Association for Computing Machinery, 2018, pp. 139–145, ISBN: 9781450353410. doi: 10.1145/3180382.3180409.
- [17] W. Farrukh and D. van der Haar, 'Lip print-based identification using traditional and deep learning', *IET Biometrics*, vol. 12, no. 1, pp. 1–12, May 2022, ISSN: 2047-4938. doi: 10.1049/bme2.12073.
- [18] F. Bianconi, E. Chirikina, F. Smeraldi, C. Bontozoglou and P. Xiao, 'Personal identification based on skin texture features from the forearm and multi-modal imaging', *Skin Research and Technology*, vol. 23, no. 3, pp. 392–398, Nov. 2016, ISSN: 1600-0846. doi: 10.1111/srt.12348.
- [19] A. K. Jain and U. Park, 'Facial marks: Soft biometric for face recognition', in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 37–40. doi: 10.1109/ICIP.2009.5413921.
- [20] W. Burger and M. J. Burge, 'Fourier shape descriptors', in *Principles of Digital Image Processing: Advanced Methods*. London: Springer London, 2013, pp. 169–227, ISBN: 978-1-84882-919-0. doi: 10.1007/978-1-84882-919-0_6.
- [21] scikit-image developers, *Local binary pattern for texture classification*, https://scikit-image.org/docs/stable/auto_examples/features_detection/plot_local_binary_pattern.html, Accessed: 2024-03-29, 2024.
- [22] J. Wang et al., 'Deep high-resolution representation learning for visual recognition', *CoRR*, vol. abs/1908.07919, 2019. arXiv: 1908.07919. [Online]. Available: <http://arxiv.org/abs/1908.07919>.
- [23] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai and Q. Zhou, 'Look at boundary: A boundary-aware face alignment algorithm', in *CVPR*, 2018.
- [24] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu and N. Sang, *Bisenet: Bilateral segmentation network for real-time semantic segmentation*, 2018. arXiv: 1808.00897 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/1808.00897>.
- [25] C.-H. Lee, Z. Liu, L. Wu and P. Luo, 'Maskgan: Towards diverse and interactive facial image manipulation', in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [26] I.-K. Yeo and R. A. Johnson, 'A new family of power transformations to improve normality or symmetry', *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000, ISSN: 00063444, 14643510. Accessed: 9th Apr. 2026. [Online]. Available: <http://www.jstor.org/stable/2673623>.
- [27] D. Ramos, R. P. Krish, J. Fierrez and D. Meuwly, 'From biometric scores to forensic likelihood ratios', in *Handbook of Biometrics for Forensic Science*, M. Tistarelli and C. Champod, Eds. Cham: Springer International Publishing, 2017, pp. 305–327, ISBN: 978-3-319-50673-9. doi: 10.1007/978-3-319-50673-9_14.
- [28] F. T. Liu, K. M. Ting and Z.-H. Zhou, 'Isolation forest', in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422. doi: 10.1109/ICDM.2008.17.
- [29] F. Pedregosa et al., 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [30] J. H. Ward, 'Hierarchical grouping to optimize an objective function', *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.
- [31] Y. Zhou and S. Zafeiriou, 'Deformable models of ears in-the-wild for alignment and recognition', in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, IEEE, Washington, DC, USA, 2017, pp. 626–633.
- [32] G. Jocher and J. Qiu, *Ultralytics yolo11*, version 11.0.0, 2024. [Online]. Available: <https://github.com/ultralytics/ultralytics>.
- [33] M. Momeni, *Acne / computer vision*, <https://www.kaggle.com/datasets/imtkaggleteam/acne-computer-vision>, 2023.
- [34] *OpenCV: solvePnP function reference*, 4.13.0, Available at https://docs.opencv.org/4.13.0/d5/d1f/calib3d_solvePnP.html, OpenCV Team, 2025.
- [35] A. Kasiński, A. Florek and A. Schmidt, 'The put face database', *Image Processing and Communications*, vol. 13, pp. 59–64, Jan. 2008.
- [36] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia and S. Zafeiriou, 'Arcface: Additive angular margin loss for deep face recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022, ISSN: 1939-3539. doi: 10.1109/tpami.2021.3087709.
- [37] O. Parkhi, A. Vedaldi and A. Zisserman, 'Deep face recognition', in *BMVC 2015- Proceedings of the British Machine Vision Conference 2015*, British Machine Vision Association, 2015.
- [38] D. E. King, 'Dlib-ml: A machine learning toolkit', *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.
- [39] B. Amos, B. Ludwiczuk, M. Satyanarayanan et al., 'Openface: A general-purpose face recognition library with mobile applications', *CMU School of Computer Science*, vol. 6, no. 2, p. 20, 2016.
- [40] Y. Sun, X. Wang and X. Tang, 'Deep learning face representation from predicting 10,000 classes', in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898. doi: 10.1109/CVPR.2014.244.
- [41] S. I. Serengil and A. Ozpinar, 'Lightface: A hybrid deep face recognition framework', in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, IEEE, 2020, pp. 23–27. doi: 10.1109/ASYU50717.2020.9259802. [Online]. Available: <https://ieeexplore.ieee.org/document/9259802>.

10. Appendix A Example Report

Final Results

Face A



Face B



Feature	Likelihood
Left Ear	0.11
Right Ear	44.75
Left Eyebrow	73.82
Right Eyebrow	73.82
Face Composition	21.46
Skin	3.39
Marks	11.75
Combined	22846591.37

Face Composition

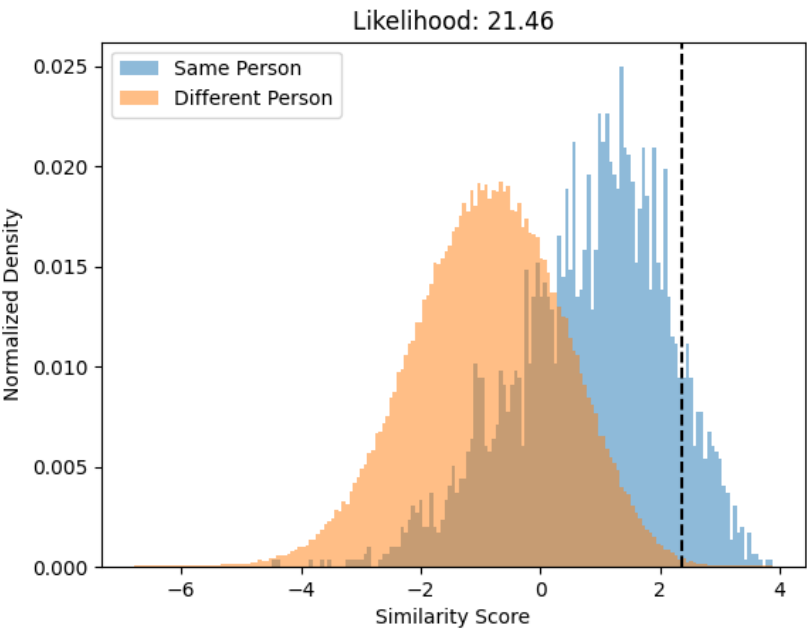
Face A



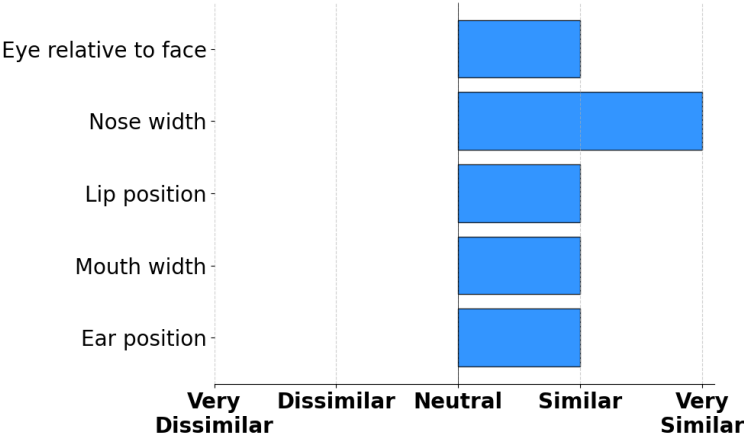
Face B



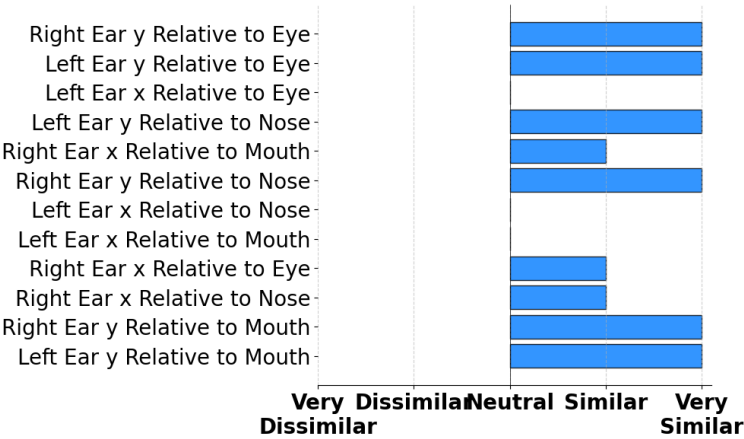
Person A	Attributes	Person B
21.46	Likelihood	21.46
Class 1	Class	Class 1
False	Anomaly	False



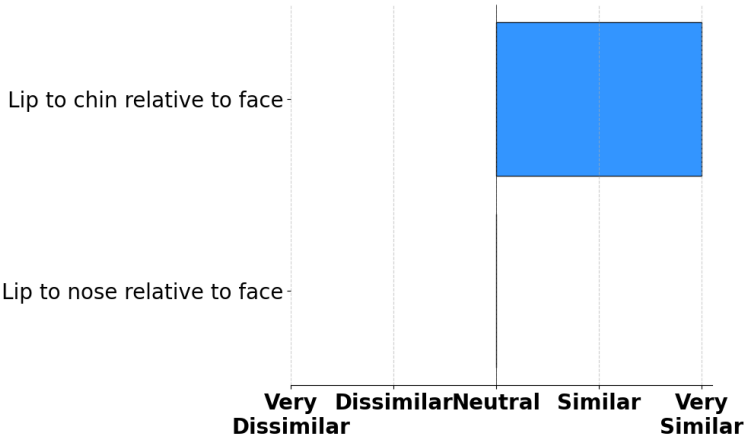
Face



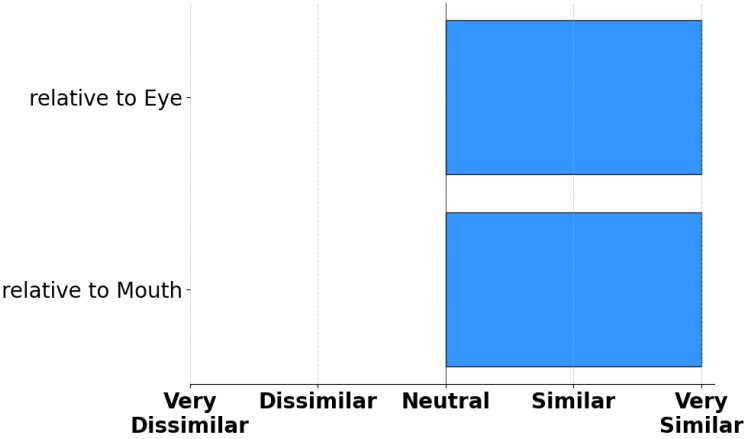
Ear position



Lip position



Nose width



Left Ears

Original Images

Person A



Person B



Typical Class 3 ear



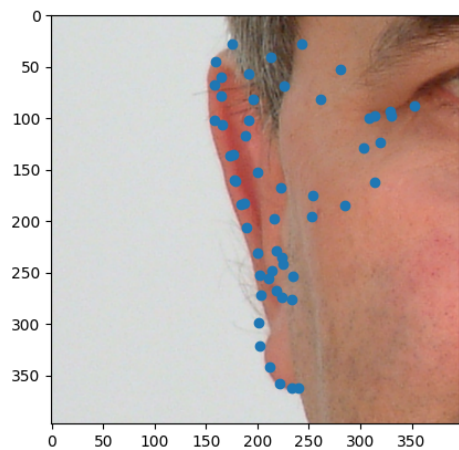
Typical Class 3



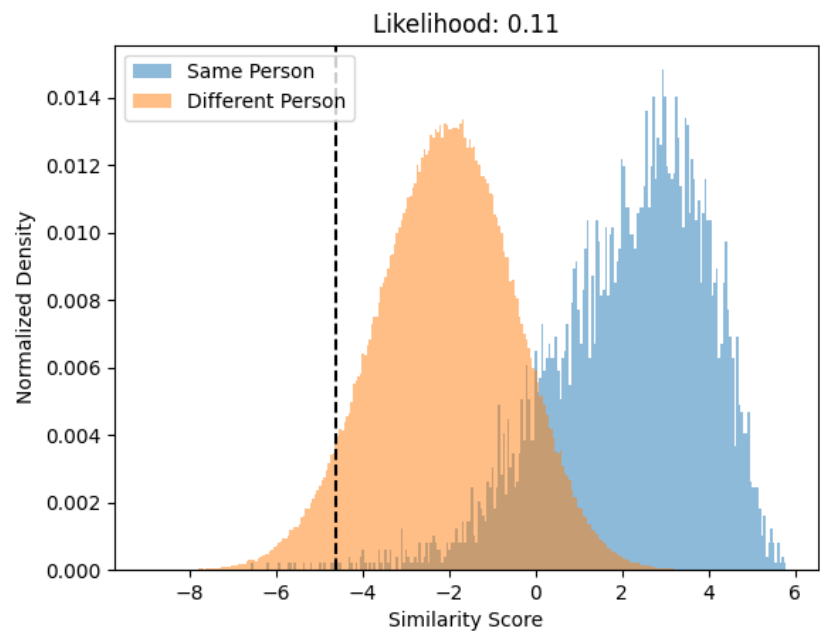
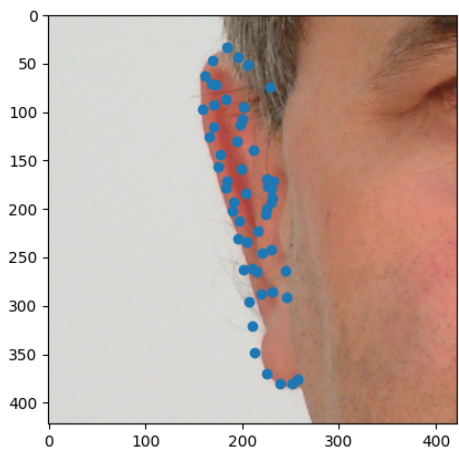
Person A	Attributes	Person B
0.11	Likelihood	0.11
Class 3	Class	Class 3
True	Anomaly	True

Landmarks

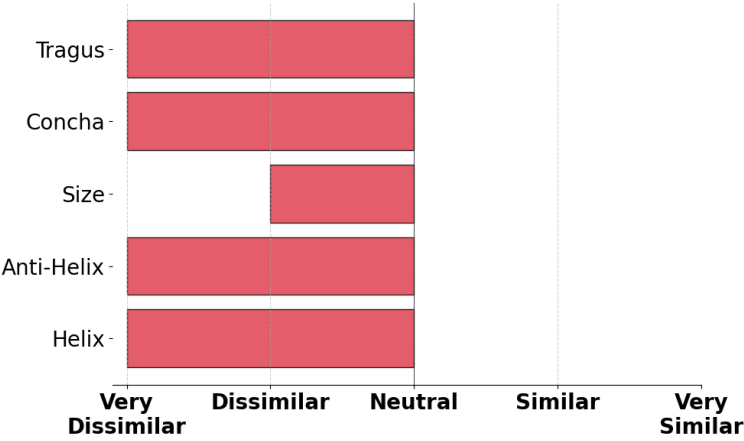
Person A



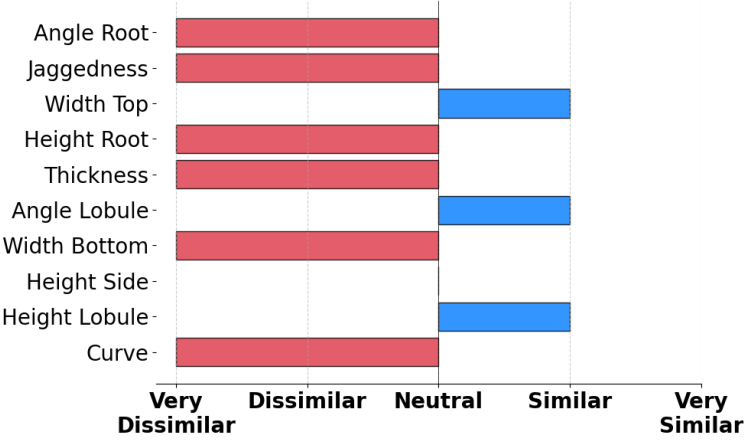
Person B



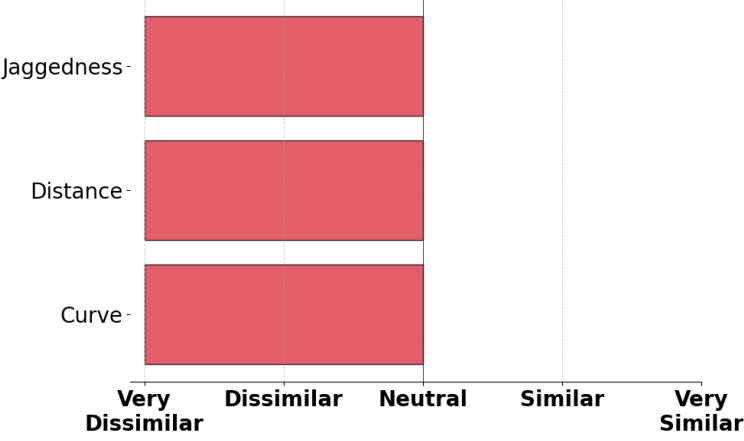
Total Ear



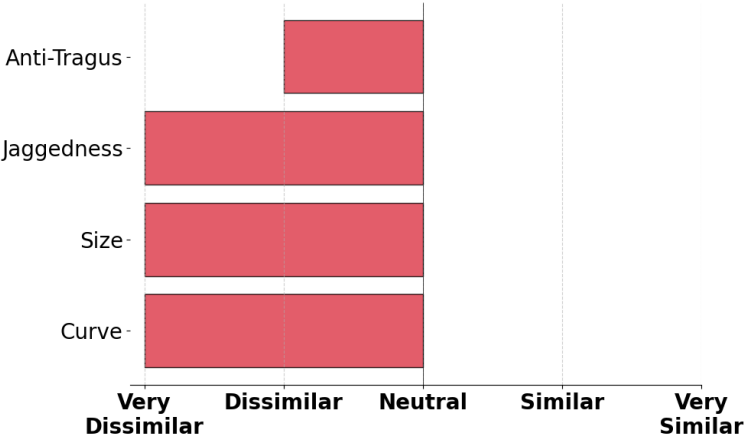
Helix



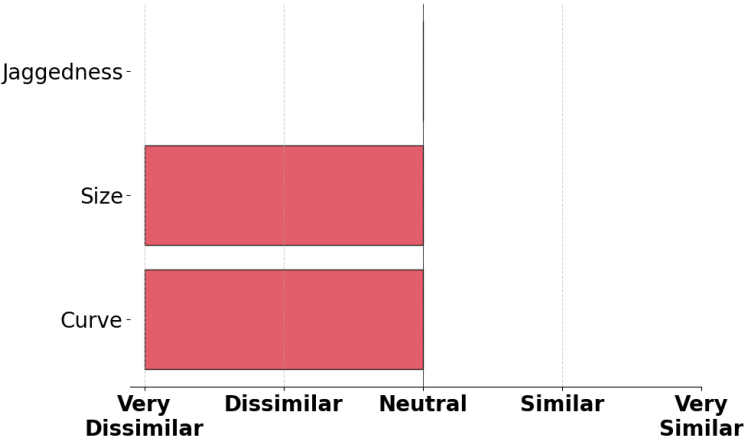
Anti-Helix



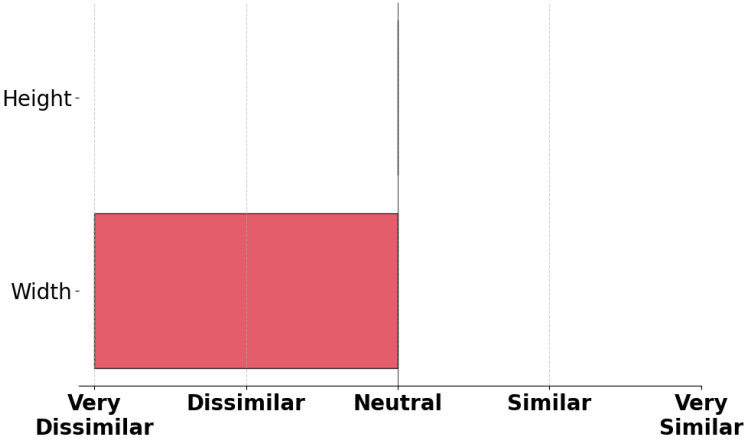
Concha



Tragus



Size



Right Ears

Original Images

Person A



Person B



Typical Class 3 ear



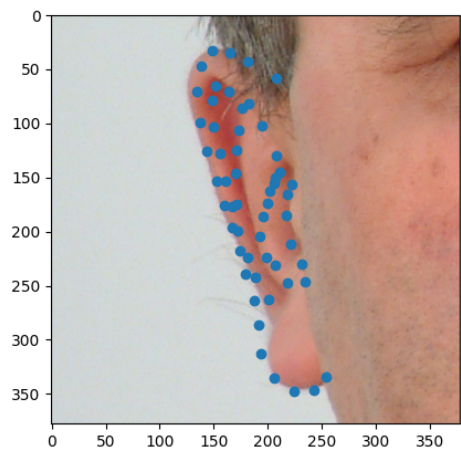
Typical Class 3



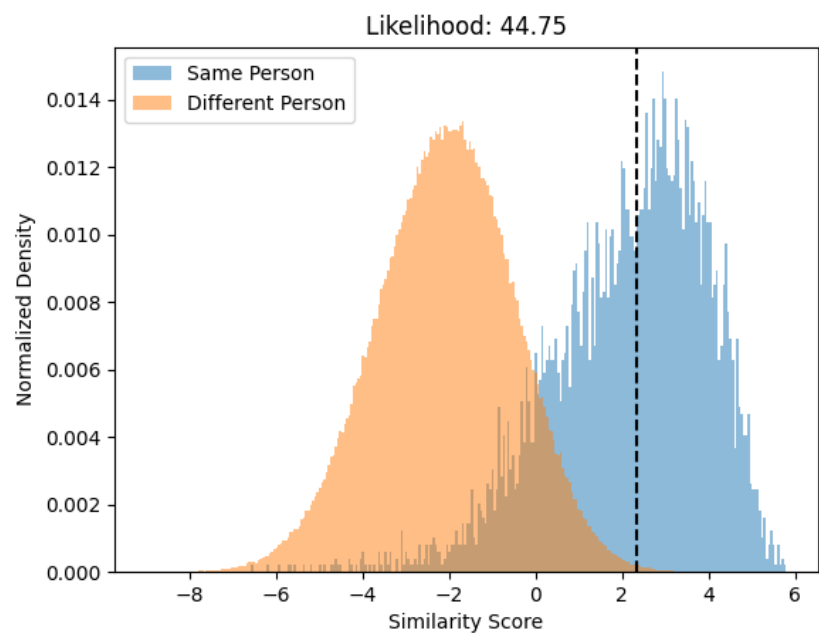
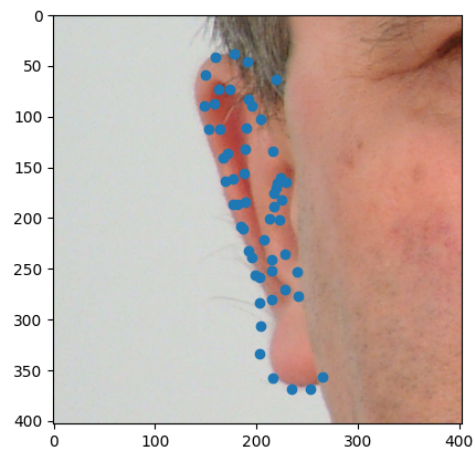
Person A	Attributes	Person B
44.75	Likelihood	44.75
Class 3	Class	Class 3
False	Anomaly	False

Landmarks

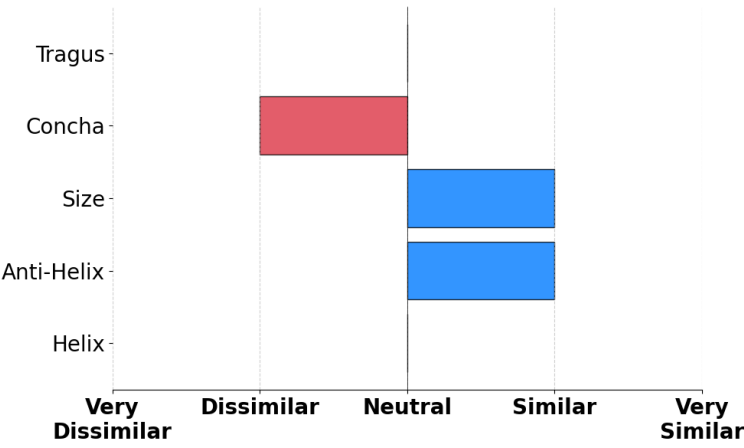
Person A



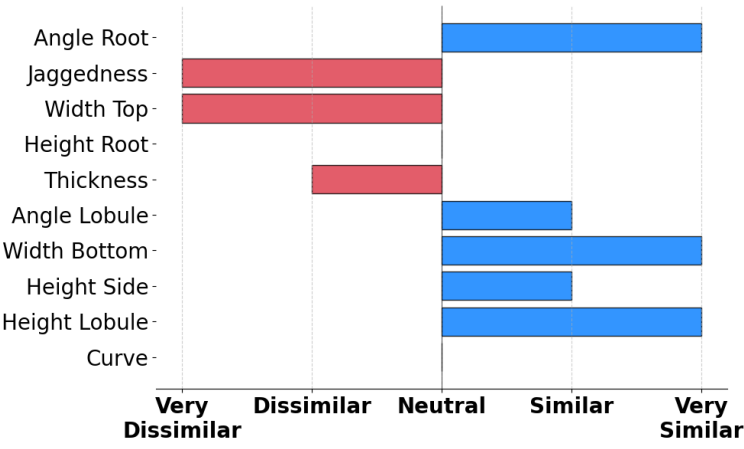
Person B



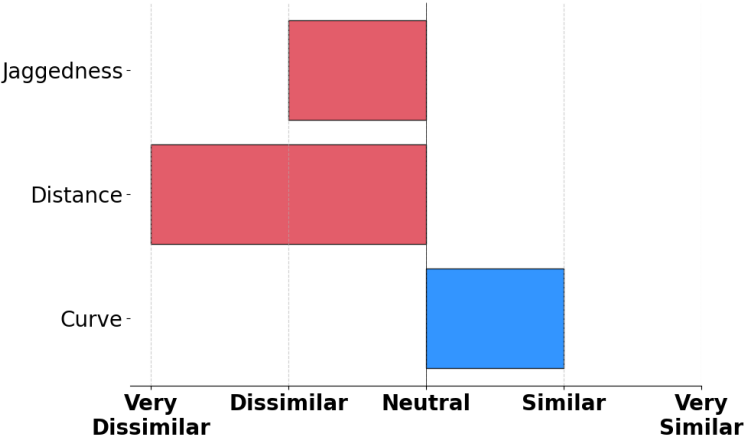
Total Ear



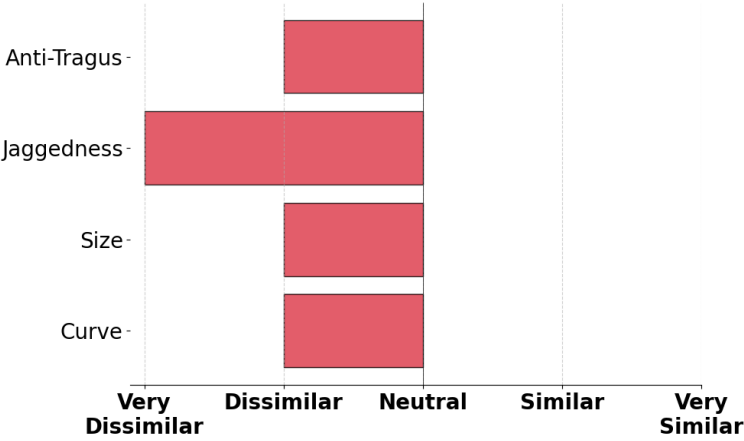
Helix



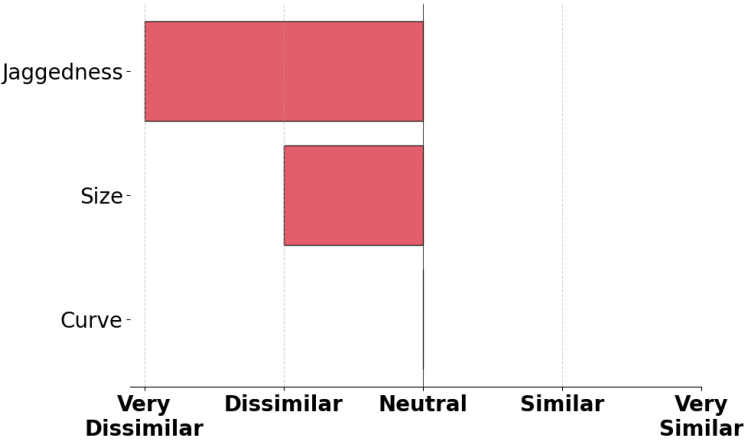
Anti-Helix



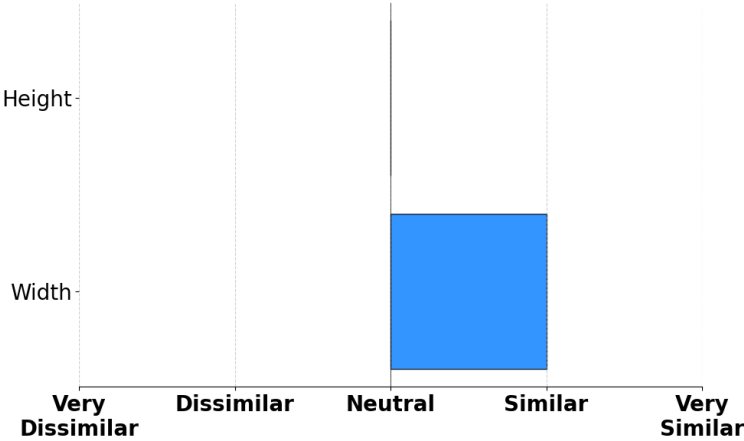
Concha



Tragus

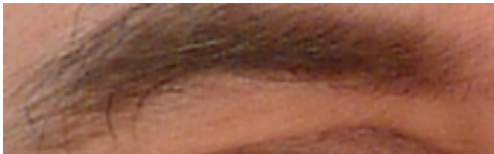


Size



Left Eyebrows

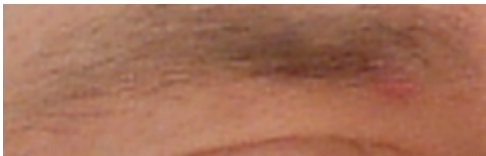
Person A



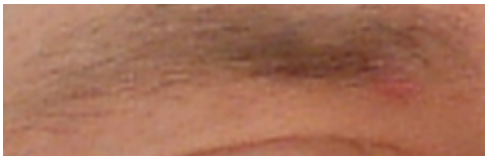
Person B



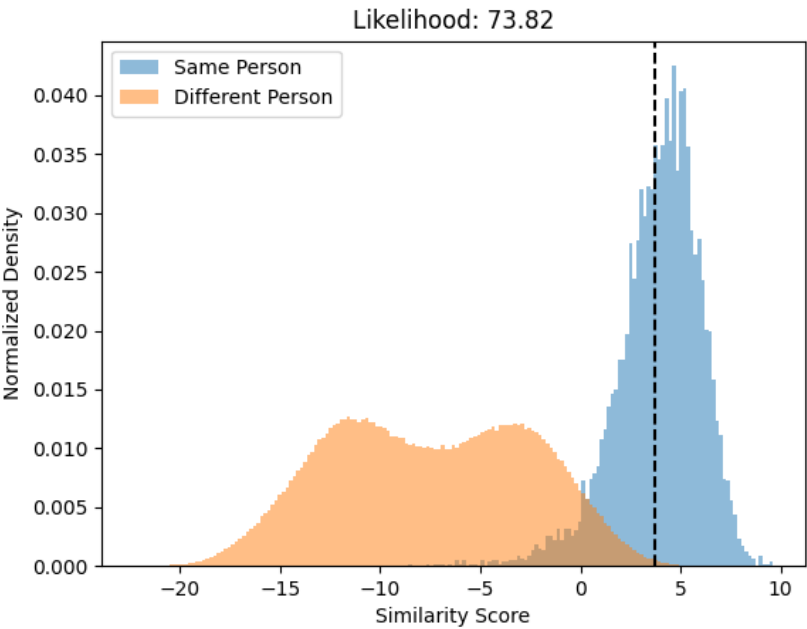
Class 5



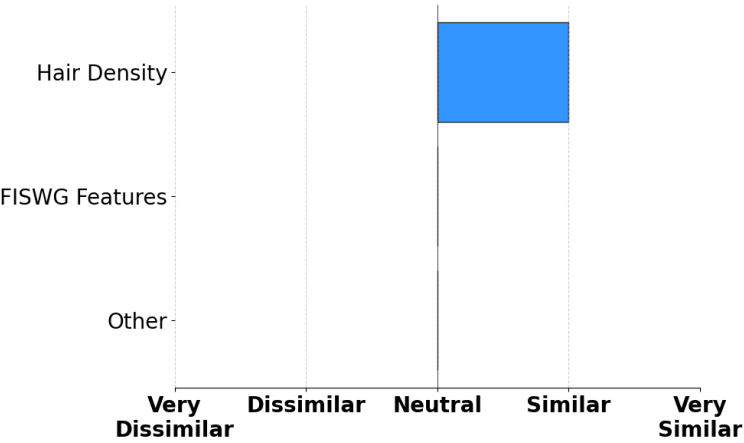
Class 5



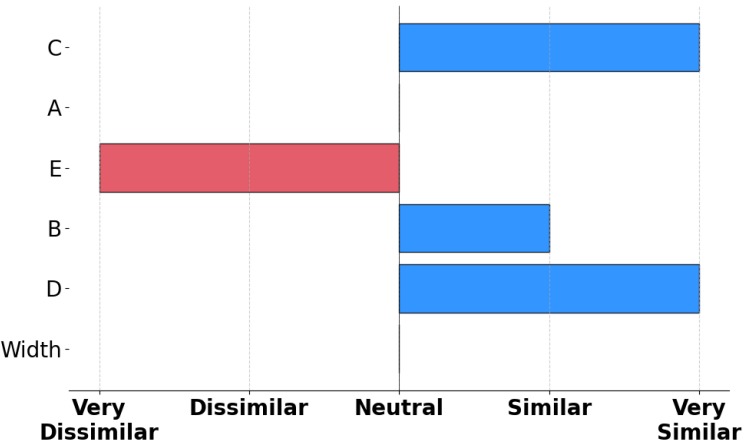
Person A	Attributes	Person B
73.82	Likelihood	73.82
Class 5	Class	Class 5
False	Anomaly	False



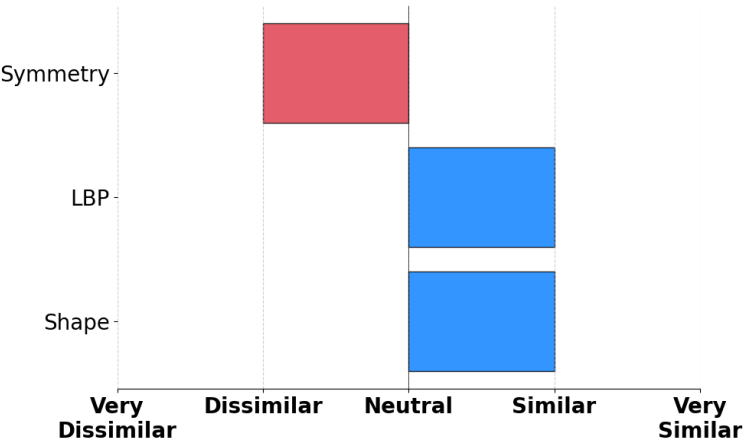
Eyebrow



FISWG



Other

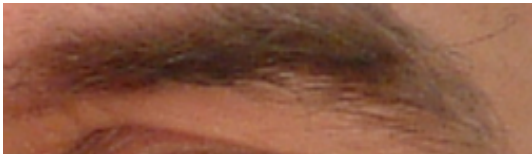


Right Eyebrows

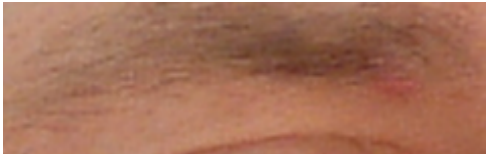
Person A



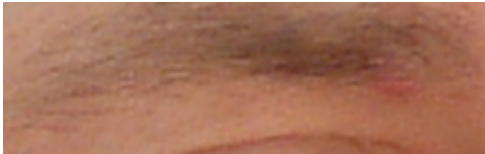
Person B



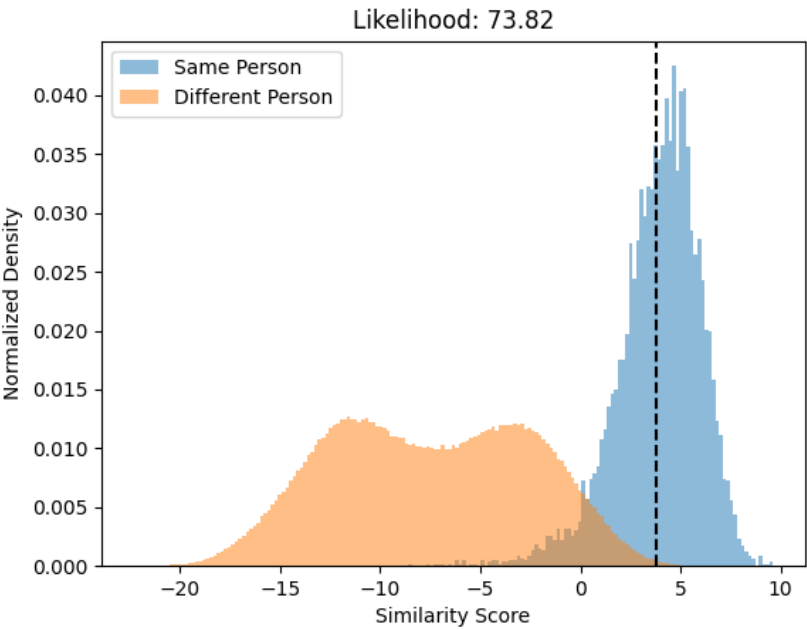
Class 5



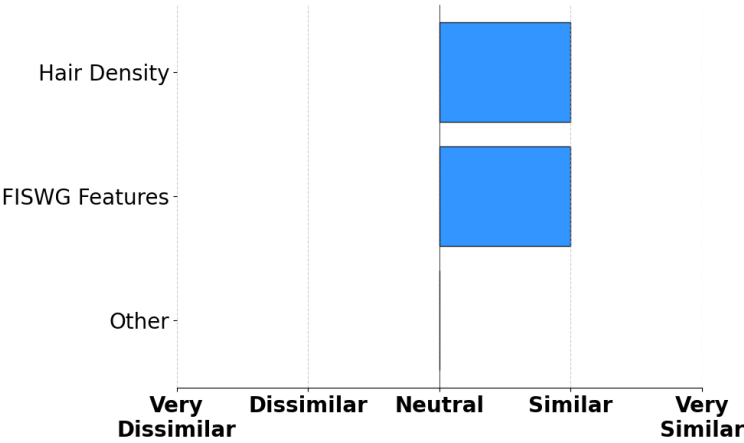
Class 5



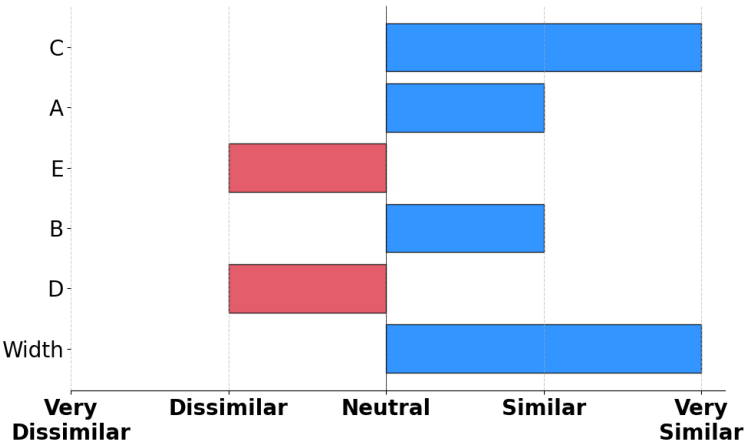
Person A	Attributes	Person B
73.82	Likelihood	73.82
Class 5	Class	Class 5
False	Anomaly	False



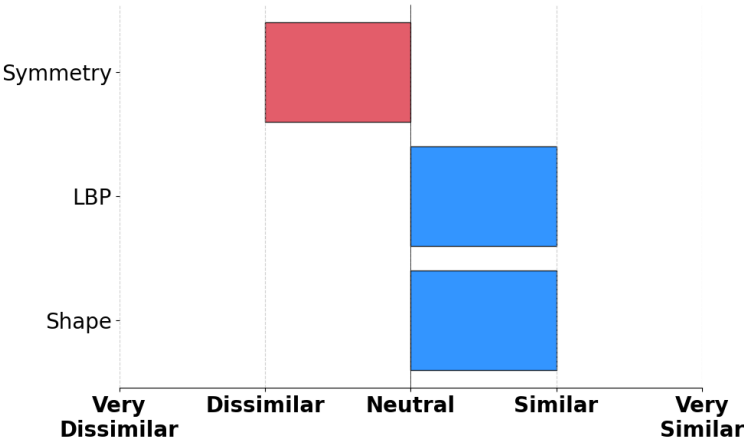
Eyebrow



FISWG

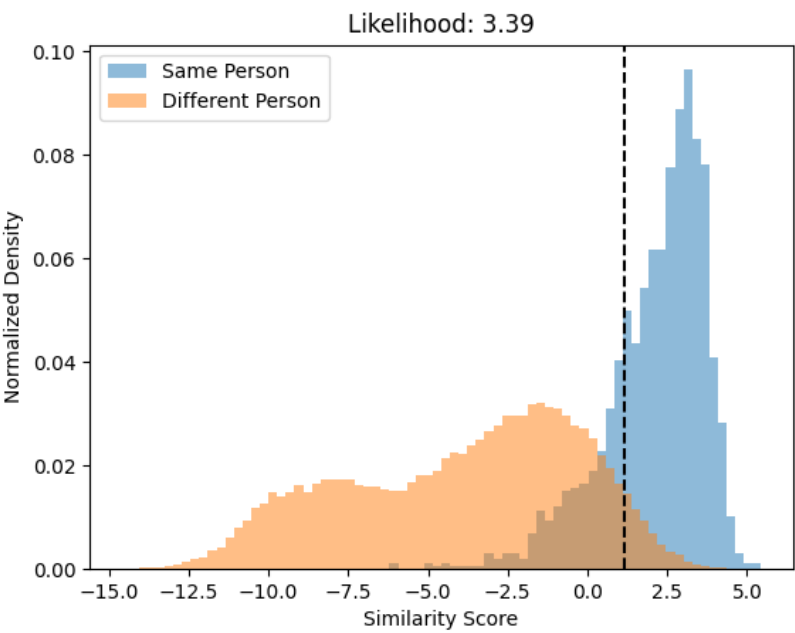


Other

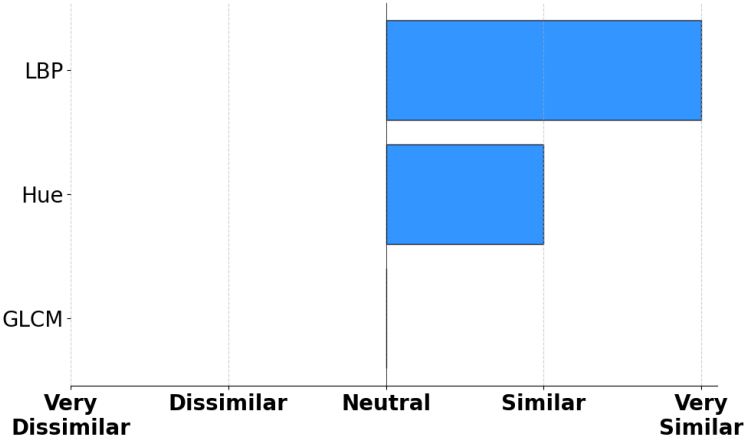


Skin

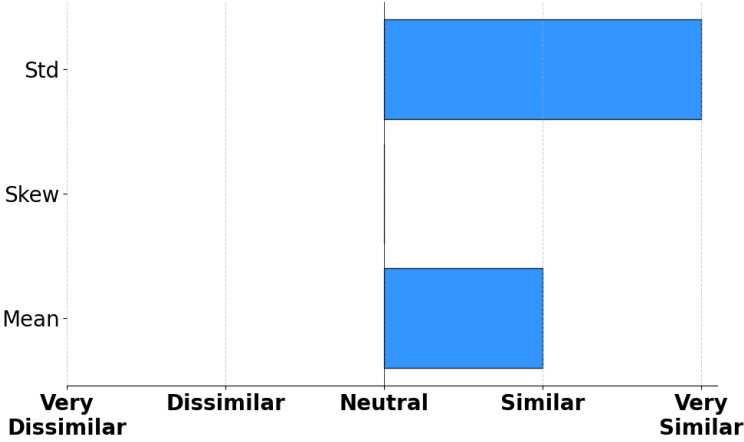
Person A	Attributes	Person B
3.39	Likelihood	3.39
Class 2	Class	Class 1
False	Anomaly	False



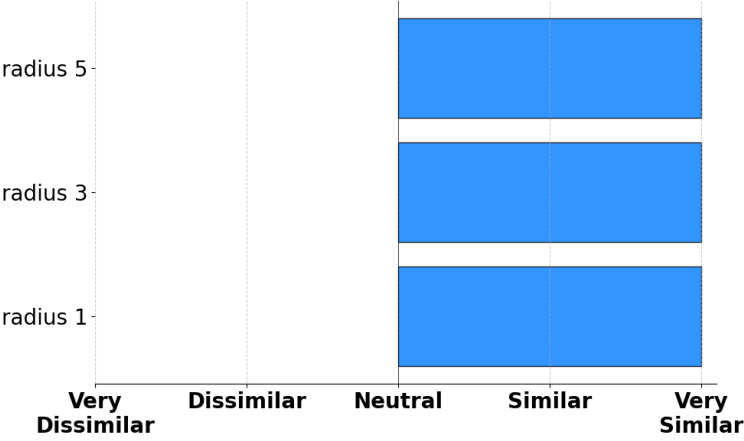
skin



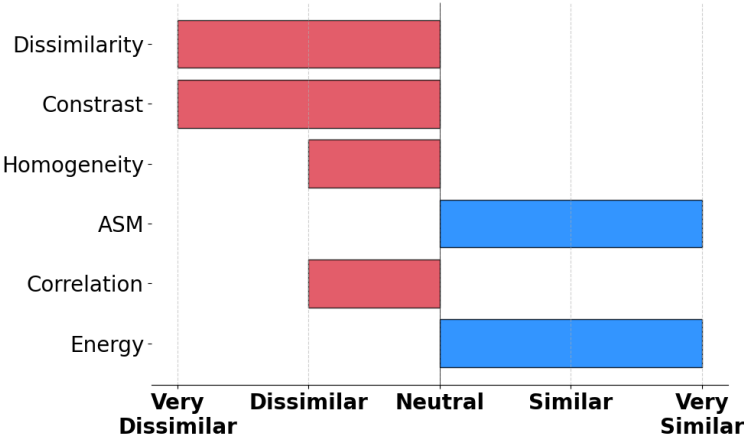
Hue



LBP



GLCM



Mark 0

