# Fusing Forensic Features and a Face Recognition System on Lookalike Faces

Melissa Tijink s1681028
*University of Twente*
Data Management and Biometrics
Enschede December 15, 2022
m.l.tijink@student.utwente.nl

***Abstract*—Face Recognition Systems are popular and widely used, however their performance on challenging cases can still be improved. One of the challenges are lookalikes, which are subjects who look similar, but have a different identity. In this research the dependence of the score computation of an existing Face Recognition System on face regions will be analyzed. By occluding parts of the face and visualizing the change in score as heatmaps, the cases of mated, (random) non-mated and lookalike pairs can be compared. The heatmaps show that the regions of the eye(brows) and nose are important for mated and lookalike pairs. The next step is to investigate whether the performance of the Face Recognition System can be improved by fusing its output with forensic features. The idea is to force the whole system to pay attention to details in important facial regions. A methodology is presented to automatically retrieve forensic features from an image. Several fusing strategies are compared with a focus on the case of lookalikes. Results show that, although overall performance is not significantly improved, comparable results with a face recognition system are reached and several fused systems show potential on individual lookalike cases.**

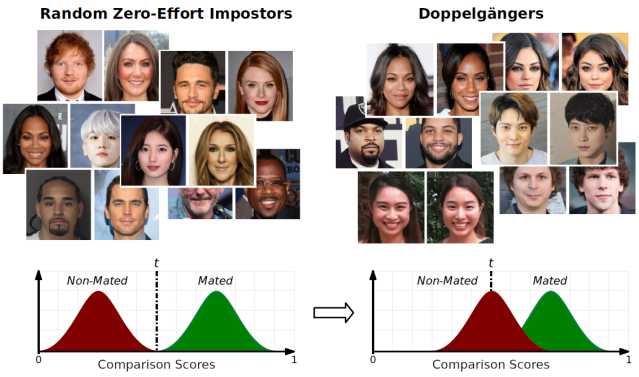



Fig. 1: On the left are random zero-effort impostors shown and their corresponding comparison scores (non-mated). On the right are doppelgangers shown, again with comparison scores (non-mated). The image is copied from [3].

## I. Introduction

In the recent past, the performance of facial recognition has made giant steps, especially with the implementation of deep neural networks [1]. However, when the system is robust against ageing, make-up and other variations within subjects, it may be vulnerable against imposters [2], [3]. An imposter is a subject a different identity as the target, but has a very similar appearance, such as lookalikes.

### A. Lookalikes

There are two types of lookalikes: related and unrelated lookalikes. Related lookalikes can be twins, which pose a serious challenge for face recognition systems [4], [5]. Unrelated lookalikes often share demographic and facial properties with a person [2], [3], [6] and are also known as doppelgängers. The effect of lookalikes on the performance of a Face Recognition System (FRS) can be seen in Figure 1. In this figure random non-mated pairs (which in the context of [3] were called random zero-effort imposters) are shown on the left, and on the right are sets of lookalikes. It is visible that in the situation of the random pairs, the non-mated and mated scores can be separated. On the contrary the non-mated and mated scores of the lookalikes overlap, meaning that high non-mated comparison scores may be falsely matched if they surpass the threshold $t$ [3]. The researchers of [6] have shown that not only face recognition systems struggle with correctly discriminating between lookalikes, but humans too do not reach high accuracy scores. They reported that face shape, eyes, nose and lips play an important role in decision making for humans in the situation of recognizing lookalikes [6].

The researchers of [7] created a new dataset with images that are perceptually similar (not only faces, see example Figure 2). They extract facial features with a deep neural network designed for face recognition and showed that machine-extracted representations perform very poorly in terms of reproducing the matching of image pairs as done by humans. On the contrary, the humans were consistent in choosing image pairs. This indicates that human can point to the (facial) features in which images resemble one another and the machine cannot in this situation [7].

The authors of [8] examined the difference between the performance of forensic facial examiners and deep convolutional neural networks based on a challenging database (lookalike images). Forensic facial examiners are trained in recognizing faces by using standardized procedures, mainly by the Facial Identification Scientific Working Group (FISWG) [8]. Contrary to the comparison with human performance by [6] and [7], which was done with untrained humans, [8] found that only the newest network at that time was able

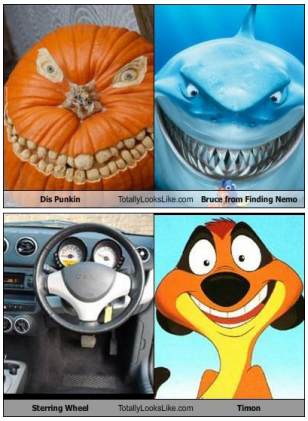



Fig. 2: Two examples from the *Totally-Looks-Like* dataset from [7].

to outperform the trained humans and, more interestingly, when they combined the results of the facial examiners and the algorithms, they could achieve even higher recognition accuracy.

### B. Intentional lookalikes

A unnatural cause of lookalikes is plastic surgery, as a subject changes its facial features. The subject looks less similar to their original face and becomes a "lookalike" of themselves. Furthermore, popular choices in plastic surgery (large lips, celebrity looks, etc.) can cause people that underwent surgery to become lookalikes of each other. Rathgeb et al. recently published a new database with "before" and "after" plastic surgery images [9], which they use to evaluate the performance of face recognition systems. They found that these systems are robust to facial alterations and most plastic surgeries do not cause significant errors. They showed that facial bones corrections have the most impact, followed by eyebrow corrections. This shows that differences between facial structures and eyebrows might be important for differentiating between mated and lookalike subjects.

Another technique to create "unnatural" lookalikes is morphing. This is an image manipulation technique that combines biometric information of two (or more) inputs into one image. This morphed image will match with various probe images from both subjects in the comparison, which shows the flaws in the current facial recognition algorithms [10]–[12]. The researchers of [13] explain that *"in face images, intrinsic feature relations exist between different semantic parsing regions and we find that face forgery algorithms always change such relations"*. Semantic features can be defined as features to which humans can give meaning, are visibly prevalent and can be described in words. They mention that semantic facial regions are crucial in face forgery detection [13] and they propose a rather successful method to detect such forgeries. The researchers of [14] state that *"fake face images are generally of high quality, but they still have tiny defects in the details."*. They explain that these defects are closely related to semantic features, as they appear in regions that contain key information in face images: the eyes, nose and mouth.

### C. Research Questions

A downside of using deep networks for face recognition is that the system is a black box, meaning it is unclear exactly what information is used to determine the outcome [15]–[17]. It was found that the regions around the eyes, eyebrows, nose and mouth are important in the decision making for humans regarding recognizing lookalikes [6], [7]. These regions also seem to play a role in the classification problems in plastic surgery and morphing context, [9], [13], [14]. The focus of this research will be to investigate the differences between mated, (random) non-mated and lookalike pairs concerning these regions. An existing face recognition system will be analyzed to find what information is important for the different types of pairs. This leads to the first research question:

**Research question 1. To which extend does the score calculation of a Face Recognition System depend on facial regions, with an explicit emphasis on examining lookalike sets?**

There are multiple researchers who investigated the influence of facial regions on the output score of the FRS by occluding regions in the face [15]–[17]. The first step of this research question is to evaluate what is known based on literature. An investigation will be conducted on the effect that occlusions of facial regions have on the performance of an FRS with an explicit emphasis on lookalikes compared to mated and (random) non-mated image pairs. The main contributions of this research within this topic are:

- Implementation of different occlusion strategies, based on literature, to analyze the dependence of the score calculation of an FRS on facial regions.
- The analysis of the dependence of the score calculation of an FRS on facial regions on the specific target group of lookalikes compared to mated and (random) non-mated subjects.

As mentioned before, a combination of forensic classification (done by humans) and the one done with a neural network improved the total classification performance [8]. In the second part of this research we are interested in using forensic features to improve the performance of a face recognition system on a dataset containing images of lookalikes. This results in the second research question:

**Research question 2. Can the performance of an existing Face Recognition System on a lookalike dataset be improved by fusing the output with features based on forensic features?**

To ensure the use of *forensic* features, descriptors from the facial image comparison feature list for morphological analysis by the FISWG will be used [18]. This list describes facial features in detail. The goal is to investigate whether a gain in performance can be achieved by fusing an FRS with forensic features. The fusing can be done at feature level, score

level or decision level [19]. However, as most commercial face recognition systems do not allow for feature fusion [19], only the outcomes of the networks, either at score level or decision level fusion, will be used. The results of several fusion strategies will be compared. The main contributions of this research concerning the second research question can be summarized as:

- A methodology to automatically extract features based on the FISWG [18] list from an image.
- A comparison of several fusion strategies where the forensic features are combined with the output of an FRS. Special attention is given to the performance of the fusion strategies on lookalike subjects.

In Section II an overview of related work will be presented. After which the methodology of both research questions will be explained in Section III. The results are presented in Section IV and discussed in Section V. Lastly the conclusions will be given in Section VI.

## II. Related work

### A. Dependence of score calculation of an FRS on facial regions

Visualization of the influence of facial regions on the output score of an FRS is not a new research topic. The researchers of [20] propose a visualization method that gives insight into the differences of lookalikes found by deep convolutional neural networks. By systematically occluding rectangular regions of the input face pair and monitoring the fluctuation of the distance between the score relative to the original score, they show the importance of each region to the similarity of the face pair [20]. In order to visualize the results, they create heatmaps to show the discriminative areas [20], where each heatmap is created from one pair of images.

The researchers of [17] compared six different techniques for computing a network attention map, which identifies influential local features in a network. They show that the network mainly focuses on the area around the eyes and nose, less on the chin, cheeks and mouth, and ignores the background of the image.

In [16] a work is presented about exploring features computed by neurons in the network, in order to visualize the features. They find that the early layers of the used network (VGGFace) correspond to low level features (edge and color), mid leavel features correspond to similar shapes and higher level layers extract high level features which are more complex, such as baldness [16]. They used patch occlusion on specific features (eyes, nose and mouth) and a random patch to occlude parts of the face to highlight and investigate the influence of these facial parts. Important to note is that these high level features are often so complex that a human cannot describe them using a few words [16].

In the research of [21] a method is proposed where different parts of the face are removed and aggregated to measure the contributions of these parts individually and in-collaboration. In an iterative process the most relevant parts of the probe image are removed and similarly the least relevant parts of an image are removed. These adapted images are fed to the FRS and a saliency map can be produced and are presented as a contour map instead of a heatmap.

The researchers of [22] use a triplet loss based system to create an explainable FRS. Triplet loss means that a known mate and non-mate gallery image are simultaneously presented to the system in combination with a probe image. They occlude parts of the probe image using a prior density to determine the locations of the occlusions, which is based on research that the eyes and nose are the most important areas in the face. They believe that *"the triplet of (probe, mate and non-mate) provides a deeper explanation beyond facial class activation maps for the relative importance of facial regions."* - [22]. The masks are filled with a blurred version of the image. After this they propose an inpainting game, where the features of a mated probe are blended with features of another identity to increasingly adapt the image until the mated image is identified as a non-mate. This strategy allows for a quantitative evaluation of the influence of facial features. This is a promising method and they created a benchmark for explainable face recognition.

An overview of the occlusion methods in aforementioned research is given in Table I. In this research a similar strategy as proposed by [20] will be used. This is: occluding different regions of the face and monitoring fluctuations in the score of an FRS. Several occlusion strategies will be used, not only rectangles, but also circles and feature specific occlusions as used by [16], [21]. The researchers of [20] occluded both gallery and probe image, contrary to the approach of [21] where only occlusions were applied to the probe image. The idea of only adapting the probe will be kept, to create more insight into where the exact change in score is originating from. This research will use heatmaps to visualize the score changes, similar to [20]. Instead of deriving heatmaps from a single pair, as done by [20], the heatmaps will be derived from the whole group of pairs (lookalikes, mated and non-mated). This way, differences between groups can be analyzed. Although the researchers of [22] proposed a benchmark for explainable face recognition, it was meant for triplet loss, where three images are presented simultaneously, and the system used in this research will be based on a standard two image gallery-probe input. This means that triplet loss is not suited for this research.

### B. Fusing FRS with forensic features

As mentioned before, there exists a list of standardized facial features created by FISWG [18]. The researchers of [23] showed that the FISWG eyebrow feature set can be considered state-of-the-art in a semi-automatic setting. It was also found that combining the result of an FRS with the analysis of trained humans using the list of FISWG, results in a higher recognition accuracy than the individual performance rates. [8]. In [24] the researchers showed that in the situation of forensic use, they could find FISWG characteristic descriptors that had a moderate to low discriminating power. However,

TABLE I: Overview occlusion methods in literature

| Research | Occlusion method | Occlusion shape | Occlusion fill | Presentation |
|---|---|---|---|---|
| [20] | Patch-by-patch | Rectangular | Mean value whole database | Heatmaps |
| [16] | Per feature / random | Rectangular | Grey | Consistency of difference |
| [17] | Replace pixels | Pixels | Zeros/Mean Color | Heatmap |
| [21] | Patch-by-patch | Circular | Black | Contour map |
| [22] | Based on a non-uniform prior | Circular / Gaussian and features | Blurred version of original | Heatmap |

in most situations a commercial automatic FRS outperformed the characteristic descriptors [24]. The aim of this research is to combine the FRS with features based on the FISWG list. The researchers of [15] describe an ideal forensic model as an open-source automatic feature-based model. However, they also show the downsides of feature based models, for example the influence of optical distortion on feature proportions. As this system is not yet solely determined for forensic use cases, one can use a semi-blackbox neural network (although it is open source, it is hard to pinpoint what features are used by the network, which will be analyzed in the first research question). Based upon the resesarch by [8] we believe that combining the FRS with forensic features such as the features described by FISWG can improve the performance. As the system will not solely be based on facial features, the system might be more robust against the problems as shown by [15]. The features will be automatically retrieved and [23] already showed that semi-automatic retrieved FISWG features can be considered state-of-the-art.

## III. Method

### A. Dependence of score calculation of an FRS on facial regions

In order to analyze the dependence of the score calculation of an FRS on facial regions, parts of images are occluded while keeping track of the change in scores, based on the literature review shown in Section II-A. The focus of this project is on lookalikes, so these are analyzed separately from mated and (random) non-mated pairs. The change in score due to the occlusion is visualized in heatmaps. A more elaborate description of the occlusion methods will be given below. First, a description of the FRS and the database will be provided.

*1) Face Recognition System*

The FRS used in this research is FaceNet, which is an open source convolutional neural network *"that maps face images to a compact Euclidean space"* [25]. The network is ported via the Face Toolbox Keras, available on [26]. It is trained on the VGGFace2 dataset, which consists of 3.3M faces [25], [26]. The score is computed as a cosine distance between the two facial embeddings returned by the network [26].

*2) Dataset*

The HDA-doppelganger database [27] was used as it contains lookalike pairs. This database consists of 200 pairs of lookalikes for male and female individuals (800 images in total). The pairs are divided in original (gallery) and lookalike (probe) images. The subjects are mainly celebrities. By combining the images of different individuals, non-mated pairs are made. In this experiment one non-mated probe per gallery image is used. In order to get mated probe images, the unknown identities of the subjects in the gallery had to be retrieved, which was done by using Google Image Search. The identities of 107 of the 200 female individuals were found and 150 of the 200 male individuals. An example of a gallery image with corresponding probes (lookalike, non-mated and mated) is shown in Figure 3.

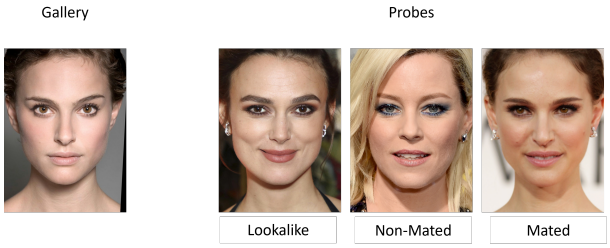



Fig. 3: An example of a gallery image with corresponding probes. The gallery image and the lookalike and non-mated probe images are from the original HDA-doppelganger dataset [27], the mated probe image is added in this research.

Fig. 4: Example of the segmentation of the facial parts. Image is copied from [29].

*3) Preprocessing*

Before the occlusion methods are applied, all faces are aligned using the outer corners of the eyes, that were found using *dlib landmarks* [28]. The images are resized to 768 by 576 pixels. A segmentation map of the face is made using a BiSeNet from the Face Toolbox Keras, available via [26] and ported from [29]. An example of the parsing is shown in Figure 4.

*4) Occlusion methods*

An overview of the whole system is shown in Figure 5 and will be explained below. An overview of the used occlusion methods is shown in Table II. All pixels in an occlusion will be replaced with the average skin color of the subject, which is determined by the segmentation map of the skin. This is

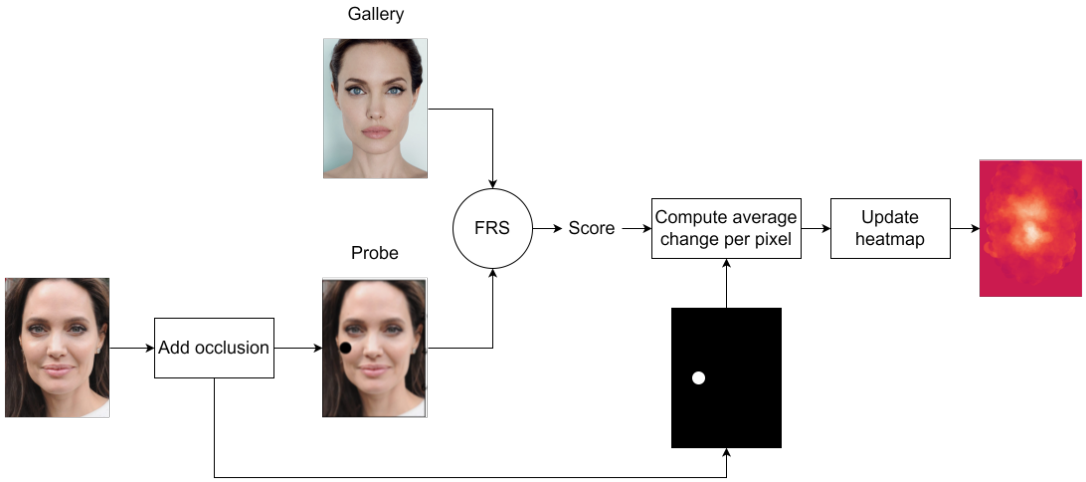



Fig. 5: Summary of applying occlusion.

visible in Figure 4 as the large blue area.

**Rectangles** One of the used occlusion masks in literature is the rectangular mask [20], that is shifted over the image. Since each position requires a new evaluation by the FRS, it was decided not to shift over the whole image, as this is a time consuming procedure. Therefore the mask is placed within the face at random positions. The rectangular masks are placed at 42 different starting positions and are then increased 10 times in size, meaning that for each image pair $42 \cdot 11 + 1 = 463$ FRS evaluations have to be carried out. The sizes are based on the average areas of the features and all approximate areas of {3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 10000, 12500} pixels.

**Circles** Besides a rectangular shape, a circle is used to occlude parts of the face. It is expected that the shape of the occlusion could have an effect on the result, due to the rectangular filters of convolutional neural networks and the organic shapes of facial features. The methodology is the same as for the rectangle, again 42 different starting positions will be used for 11 different sizes of the occlusions. The diameters are chosen in such a way that the areas of the circles are equivalent to those used with the rectangles.

**Features** Similar as is done in [16] whole facial features will be occluded. The left and right eyebrows, left and right eye, nose, upper lip and lower lip will be occluded and the initial occlusion of each feature is based on its segmentation map (see Figure 4). After which the size of the shape is increased using dilation. This is repeated 12 times.

**Pixels** As a baseline experiment random pixels are occluded. 11 different numbers of occluded pixels will be used and the pixels will be initiated $42 * 11$ times. Again, only pixels within the face are changed.

After each occlusion is applied, the change in score (compared to the situation without occlusion) is divided by the number of changed pixels (see the mask in Figure 5). The change in score is computed as described in Equation (1). For each instance the moving average of the change per pixel is updated in a heatmap.

$$\text{Change\_per\_pixel} = \frac{\text{score\_occlusion} - \text{score\_original}}{\text{number\_of\_occluded\_pixels}} \quad (1)$$

### *B. Fusing FRS with forensic features*

To fuse the FRS with forensic features several steps have to be taken. Starting with preprocessing the image pair, after which the feature extraction happens. These features are scored by a trained model, which are then fused together with the score of the FRS resulting in a final classification. As mentioned in Section I-C two types of fusion will be applied: score and decision level. An overview of the whole system for score and decision level fusion are shown in Figure 7 and Figure 8 respectively. For training the model and certain aspects of the fusion algorithms, a separate dataset was used. For evaluation of the results a test set is used, which is the HDA Döppelganger database [27]. Some slight alterations to this database are made, which will be explained further on. In this section information on the used datasets will be provided, after which an elaboration on each step shown in Figures 7 and 8 is given.

#### *1) Datasets*

For training a Pinterest face dataset is used, which is publicly available on Kaggle: the Pins Face Recognition dataset [30]. The data is collected from images on Pinterest through a Scrapper-Bot made by Python and Selenium [30]. Hence all individuals in this dataset are celebrities, which is similar to the HDA-doppelganger dataset [27] used for the previous experiment and which will be used as the test set in this experiment. The faces in the training set are preprocessed with dlib [30] similarly as the HDA-doppelganger dataset. 54 female and 50 male identities are used. In total 11012 unique

TABLE II: Overview of the used occlusion methods

| Occlusion shape | Location | Start position remains during size increasement | Number of start positions | Number of size increasements |
|---|---|---|---|---|
| Rectangular | Random in face | Yes | 42 | 11 |
| Circular | Random in face | Yes | 42 | 11 |
| Features | Feature-based | Yes | 7 (features) | 12 |
| Pixels | Random in face | No | 462 | 11 |

Fig. 6: Examples from the Pins Face Recognition dataset, retrieved from [30]

images are left in the dataset and each identity has 106 images on average. Examples of three (original) images of a female and male individual are shown in Figure 6.

For testing the HDA-doppelganger database [27] is used, which was also used in Section III-A2. The dataset is expanded by adding mated images in a similar way as described in Section III-A2[1]. 541 images are added for the female class and 689 images for the male class. Additionally, extra non-mated pairs are created by matching the gallery image to probe images of other identities.

*2) Preprocessing*

Before the features are extracted, the same preprocessing as described in Section III-A3 is applied. The segmentation maps of the eyebrows, eyes, nose, lips and face (blue) will be used for extraction of the FISWG features [18]. In order for easy processing, the contours of the segmented facial parts are used. These contours are retrieved using opencv's FindContours functionality [31].

*3) Feature extraction*

The features that are extracted are based on the FISWG list [18]. The features are selected based on these requirements:

1) **The feature has to be present in a frontal view of the face.** This is important because only frontal view images are available. Some features in the FISWG list [18] require a side or top view of the subject, so these features are dropped.
2) **The feature is quantifiable.** This is used as some of the features involve highly specific descriptions, such as individual hairs in eyebrows. In order to keep the feature retrieval straightforward these features are dropped and only features that are quantifiable (either distance, percentage or other) are used.
3) **The feature is present in the segmentation maps.** This requirement is set as the segmentation maps do not overlap completely with the facial parts used in the FISWG list. For example, no distinction is made between forehead and the cheeks. Also, not all features are always present in the images, such as the ears and the neck. These are left out too.
4) **The feature description is intuitive to the researcher.** As the researcher is not a forensic expert, the descriptions of the features had to be intuitive to some extent, to prevent misconceptions. Very technically described features are left out. An example is this description: *"Visibility of infraorbital furrow (a place where a line or wrinkle may appear parallel to and below the lower eyelid running from near the inner canthus and following cheek bone laterally)"* [18].

The remaining features are retrieved in four steps: the eye and eyebrow features, the mouth features, the nose features and head features. A description of the FISWG feature extraction and lists of all extracted features can be found in Appendix **??**. All these features are extracted per subject. There are seven features left that are computed between two subjects. These features compare the contours of the eyes, eyebrows, nose and lips. A description of the method to extract these features can be found in Appendix A. The features are further processed in two ways:

1) 2D: The FISWG features of both subjects will be concatenated as a two-dimensional feature vector.
2) Diff: The FISWG feature values of one subject are subtracted of the feature values of the other subject, creating a one-dimensional difference feature vector.

*4) Likelihood Ratio*

The next step is to create a model to separate mated and non-mated pairs. The two types of feature vectors of the full training data set are scaled to zero-mean and unit variance. Noise is added to the data, to create a broader representation of values in the training data. The noise is computed by taking one percent of the difference between the maximum and minimum value for a particular feature in the feature vector and multiplying this with random samples from the standard normal distribution. The mean and covariance (matrix) of both mated and non-mated feature data are computed, so the data can be modeled as normal distributions. The contour features

[1]A list of found names of the individuals in the dataset and copyright information on the added images is available upon request.

are only used in the Diff situation, as these features incorporate both subjects.

To score a feature of a pair of subjects the log-likelihood ratio (LLR) is computed. The log-likelihood of the pair under both mated and non-mated distributions can be computed and the ratio between the two log-likelihoods will tell under which distribution the pair is most likely to fit. The equation of the LLR can be found in Equation (2), where $P(X|\text{mated})$ represents the likelihood of sample X under the mated distribution and $P(X|\text{non-mated})$ the likelihood of sample X under the non-mated distribution.

$$\log\left(\frac{P(X|\text{mated})}{P(X|\text{non-mated})}\right) = \log\left(P(X|\text{mated})\right) - \log\left(P(X|\text{non-mated})\right) \tag{2}$$

*5) Feature Fusion*

Now the results of the FRS can be fused with the forensic features. In this project only score level fusion and decision level fusion will be considered, as most commercial systems do not allow feature level fusion [19], as was also mentioned in Section I-C. Although in this project an open source system is used, it was decided to look at these options to offer a fair comparison with commercial systems. Different strategies to fuse the FISWG features with the FRS will be explained in detail. A summary of all different options is shown in Table III and corresponding graphical overviews of the systems are shown in Figure 7 and Figure 8. All strategies are applied to both the "2D" and "Diff" feature extraction methods explained earlier. In both situations the output is the LLR, which is a single value score, so there is no difference for the feature fusion steps between the "2D" and "Diff" process. It is important to mention that the training set will be used to determine the parameters of the fusion models. The features extracted from the original training set are scaled using the scaling parameters found earlier and noise is added again, meaning that this data differs slightly from the original training set. The two different uses of the training set are highlighted in Figures 7 and 8.

*a) Score level fusion*

For these fusion strategies the LLR-scores and the score of the FRS are fused. Four different fusing strategies will be presented. An overview of the system can be found in Figure 7, where optional blocks are shown in color.

**Strategy 1 - Non-scaled** The LLR- and FRS scores are directly fed to a neural network consisting of one neuron. This means that the weighted sum of all scores is taken. The neuron learns the weights of each input and will fuse all features in an optimal way. The single neuron neural networks are implemented using Keras Tensorflow [32] and are trained in 20 epochs with a batch size of 256.

**Strategy 2 - Batch Normalization (BN)** Instead of directly feeding the scores to the neuron, batch normalization is applied to the scores. Batch normalization applies a transformation that maintains the mean output of a batch close to zero and the standard output deviation close to one [33].

**Strategy 3 & 4 - Sum LLR** In strategy 3 independence of all features is assumed, so all log-likelihood ratios can be summed. Again a single neuron network is used to find the weights of the summed LLRs and the FRS score. For strategy 4 a batch normalization layer is applied before the neuron in the classifier and will be referred to as BN/Sum_LLR.

*b) Decision level fusion*

For these fusion strategies the decision (mated or non-mated) per feature are combined to make a final decision. An overview of the two strategies for the fusion step is shown in Figure 8. In order to make the decisions, a threshold is set at 1% False Match Rate (FMR) of the LLR scores per feature and on the score of the FRS. This is the threshold where only 1% of the non-mated is classified as mated match. These thresholds are determined using the training dataset.

**Strategy 5 - Use_decision** All decisions are led to a single neuron neural network, which will determine the weights of each separate decision to come to a final decision.

**Strategy 6 - Voting** A voting system is used on the decisions of the features. It was found that if a pair is assigned a score below 0.2, it is always a mated match and above 1.0, it is always considered non-mated. Whenever the score is between 0.2 and 1.0, the vote that followed from the forensic features is included. The number of necessary LLR of forensic features above the thresholds to get a final mated vote, is varied from 10% to 90% of the total number of forensic features. Only when both systems favour towards mated, a mated vote is given. In all other situations the final vote will give non-mated.

## IV. Results

### A. Research question 1: Dependence of score calculation of an FRS on facial regions

As 4 different methods of occlusion {rectangular, circular, pixels, features}, for two genders {female, male} with three different classes {mated, non-mated, lookalike} and in most cases, for 11 different sizes are applied, over 250 heatmaps are collected. Therefore only a selection of the results is shown. The main results of this research question are shown in Figures 10 and 11. These images show the heatmaps for different types of occlusions and for different types of pairs (mated, non-mated and lookalikes) for female and male subjects. All these occlusions are filled with the average skin color of the subject and only one size of occlusions is visible. In Figure 9 one can see two heatmaps of the score change of mated female pairs with a circular occlusion, one with the minimum area and one with the maximum area. The heatmaps of all optional areas of circular occlusions are shown in the Appendix in Figure 21.

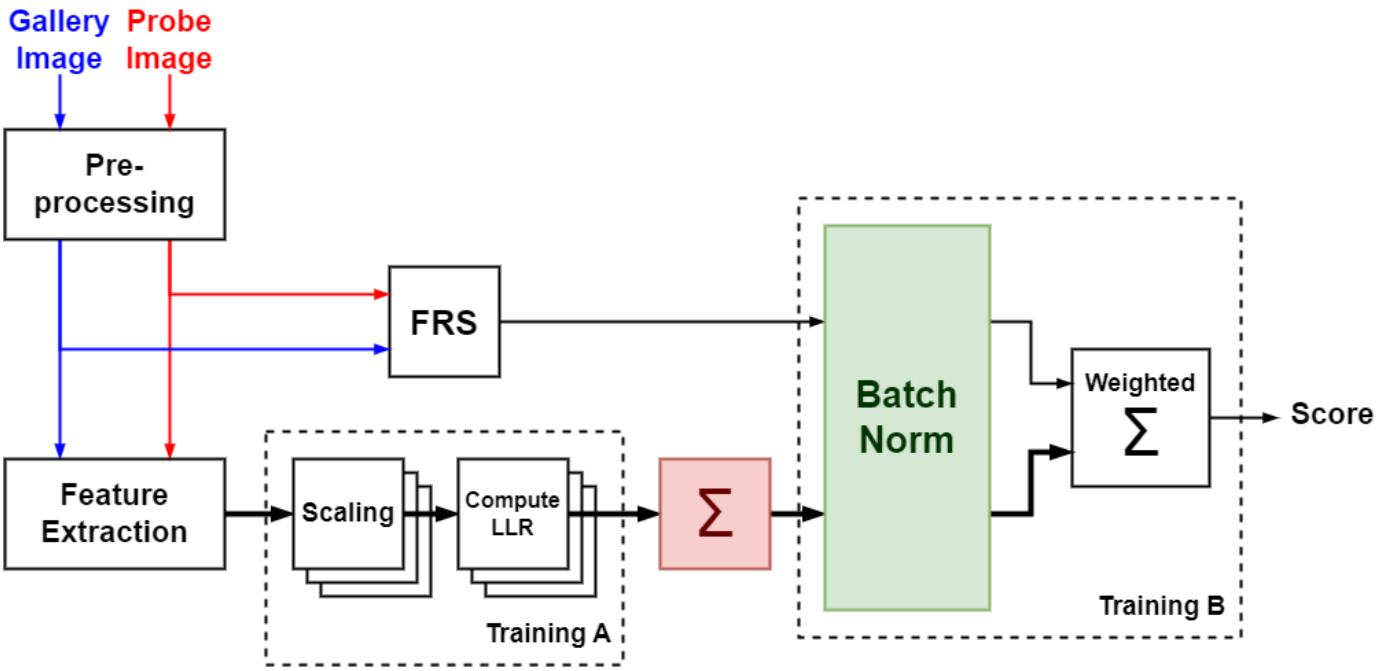



Fig. 7: Overview of the system with score level fusion, showing the optional summation (strategy 3 & 4) and batch normalization (strategy 2 & 4) function blocks in red and green respectively. Highlighted are two subsections of the system that are separately trained (A & B).

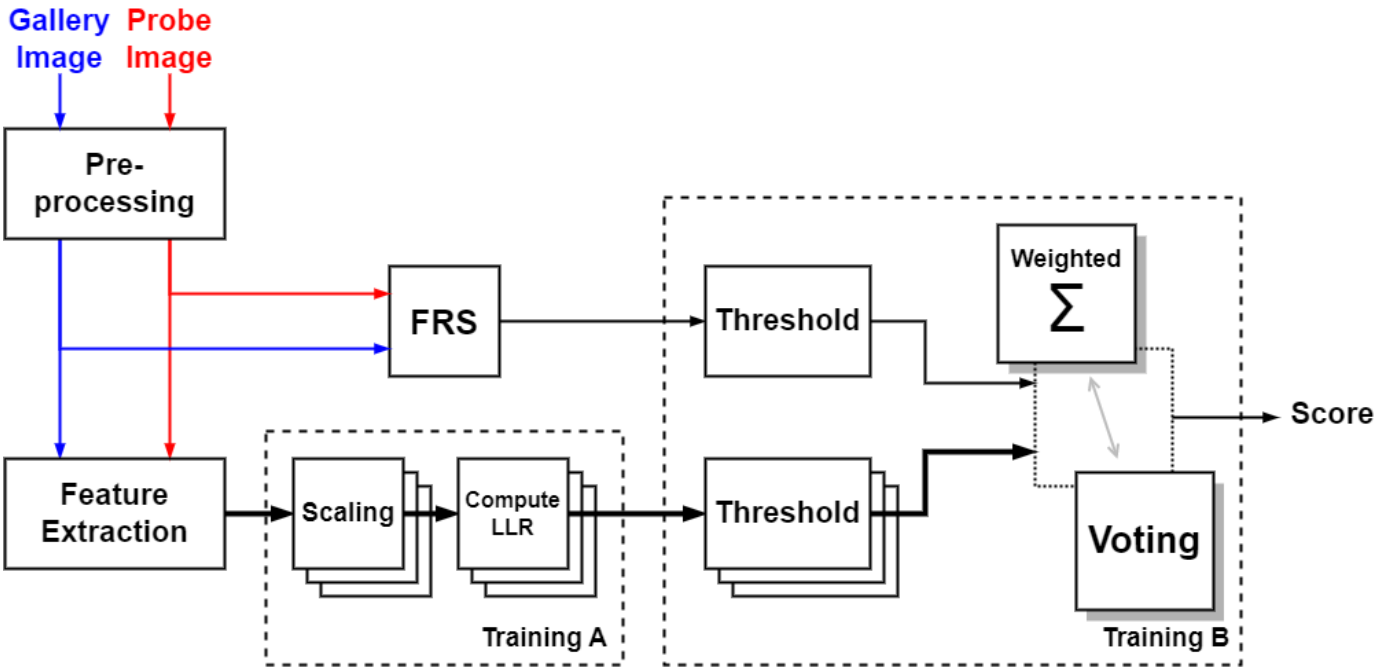


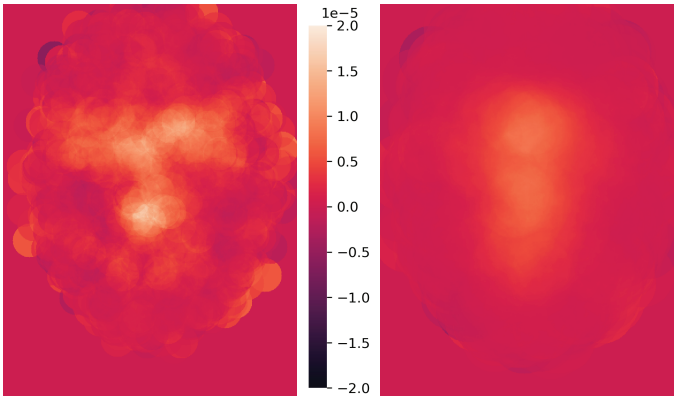

Fig. 8: Overview of the system with decision level fusion, showing the optional fusion strategies *weighted sum* (strategy 5) and *voting* (strategy 6). Highlighted are two subsections of the system that are separately trained (A & B)

TABLE III: Overview of the different options in the fusion strategies.

| Strategy | Name | Sum LLR | Batch Normalization | Threshold LLR | Weighted Sum Fusion | Majority Voting |
|---|---|---|---|---|---|---|
| 1 | Non-scaled | | | | ✓ | |
| 2 | BN | | ✓ | | ✓ | |
| 3 | Sum LLR | ✓ | | | ✓ | |
| 4 | BN/Sum LLR | ✓ | ✓ | | ✓ | |
| 5 | Use_decision | | | ✓ | ✓ | |
| 6 | Majority Voting | | | ✓ | | ✓ |

TABLE IV: Area-under-curve (AUC) and Equal Error Rate (EER) of different fusion strategies.

| | | **FRS** | | **Nonscaled** | | **BN** | | **Sum_LLR** | | **BN + Sum_LLR** | | **Use_decision** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUC | EER | AUC | EER | AUC | EER | AUC | EER | AUC | EER | AUC | EER |
| **Female** | 2D | 0.969 | 0.09 | 0.957 | 0.11 | 0.957 | 0.11 | 0.965 | 0.1 | 0.968 | 0.09 | 0.857 | 0.13 |
| | Diff | | | 0.965 | 0.1 | 0.965 | 0.1 | 0.969 | 0.09 | 0.969 | 0.09 | 0.940 | 0.13 |
| **Male** | 2D | 0.992 | 0.04 | 0.984 | 0.07 | 0.984 | 0.06 | 0.991 | 0.05 | 0.992 | 0.04 | 0.905 | 0.11 |
| | Diff | | | 0.991 | 0.05 | 0.991 | 0.05 | 0.992 | 0.04 | 0.992 | 0.04 | 0.952 | 0.1 |




A) B)

Fig. 9: Heatmaps of the average score change per pixel of mated female pairs with a circular average skin color occlusion. The left heatmap has occlusions with the minimum area and the right heatmap has occlusions with the maximum area.

### *B. Research question 2: Fusing FRS with forensic features*

All different fusion strategies for the 2D and Diff processes are trained and tested as explained in Section III-B5. Table IX shows the top an bottom three (absolute) weights, together with the weight for the FRS score of the single neuron. Table VIII shows the weights of the single neuron for the two different strategies where the LLRs are summed before fusion. For most strategies the final results can be summarized in Receiver-Operator Curves (ROC-curves) which show the False Match Rates (FMR) and True Match Rates (TMR) for varying thresholds on the predicted outcomes of the whole system. This is not possible for the voting strategy (strategy 6), as it only gives two possible outcomes (mated and non-mated). As the curves overlap greatly, the results are summarized as the area under the ROC-curves (Area-under-curve / AUC). The full ROC-curves are presented in Appendix C. The Equal Error Rates (EER) are computed too. The EER is the point where the FMR and the False Non-Match Rate (FNMR) are equal. The AUCs and EERs can be found in Table IV. When computing the AUC and the EER the lookalike pairs are counted as non-mated.

TABLE V: Confusion matrices (normalized) of the voting strategy on female subjects.

| | **Predicted label 2D** | | **Predicted label Diff** | |
|---|---|---|---|---|
| **True label** | 0.29 | 0.71 | 0.29 | 0.71 |
| | 0.005 | 0.99 | 0.005 | 0.99 |

TABLE VI: Confusion matrices (normalized) of the voting strategy on male subjects.

| | **Predicted label 2D** | | **Predicted label Diff** | |
|---|---|---|---|---|
| **True label** | 0.54 | 0.46 | 0.55 | 0.45 |
| | 0 | 1 | 0 | 1 |

The performance of the voting strategy is shown in Tables V and VI. The shown results have the decision thresholds set at 1% FMR. The confusion matrices remain the same when the system needs 20% or more of the LLR of the forensic features favor a mated match, which is why only one confusion matrix per gender and feature extraction method (2D & Diff) is shown. The confusion matrices of the situation with 10% of necessary features and different decision thresholds are shown in Appendix C.

The performance on the lookalike pairs is evaluated as follows: first the FMR on the mated and non-mated subjects of the test database is set to 1% for all feature fusion strategies (except the voting strategy). The same is done for the FRS without feature fusion. The thresholds are applied to the predictions of lookalikes for all systems. It is important to mention that the lookalikes did not influence the choice of the threshold, as they were left out in this situation. There are 107 usable female lookalikes and 112/114 (Diff & 2D respectively) male lookalikes. The percentages of correctly classified lookalike pairs are shown in Table VII.

Some individual examples are analyzed to compare the

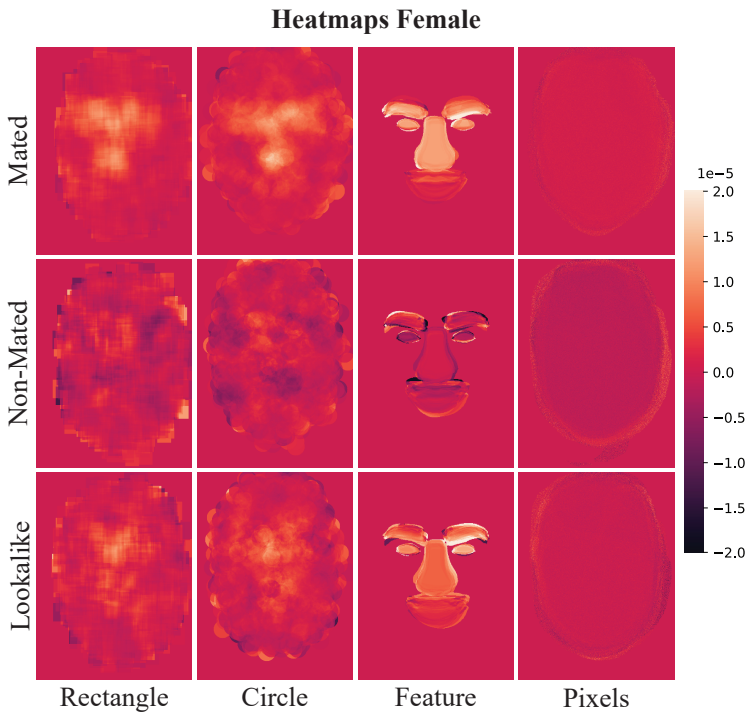



Fig. 10: Heatmaps of the average change in score per pixel for different types of occlusions horizontally (from left to right: rectangular, circular, feature and pixel occlusions) and different pairs vertically (from top to bottom: mated, non-mated and lookalike pairs). All heatmaps are based on female subjects and the occlusions have the average skin color of the subject.

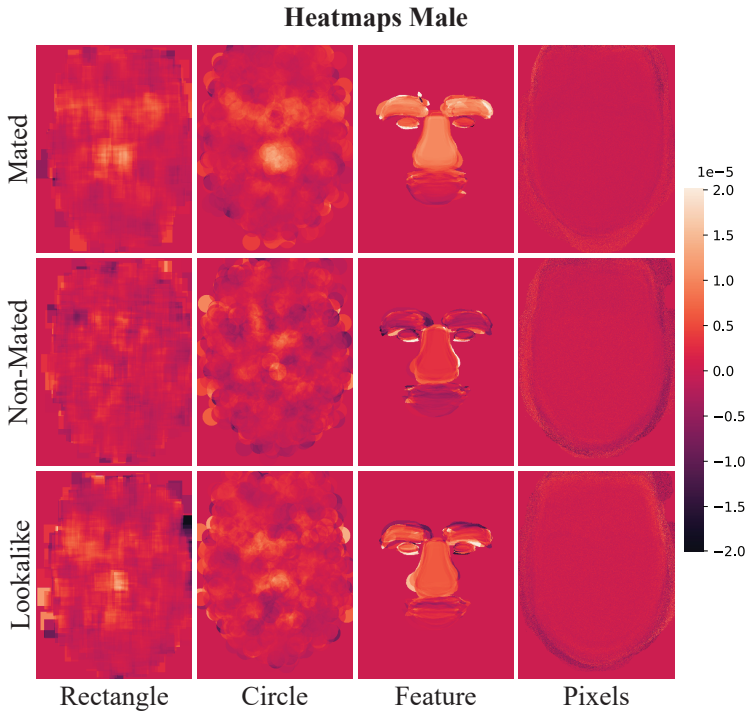



Fig. 11: Heatmaps of the average change in score per pixel for different types of occlusions horizontally (from left to right: rectangular, circular, feature and pixel occlusions) and different pairs vertically (from top to bottom: mated, non-mated and lookalike pairs). All heatmaps are based on male subjects and the occlusions have the average skin color of the subject.

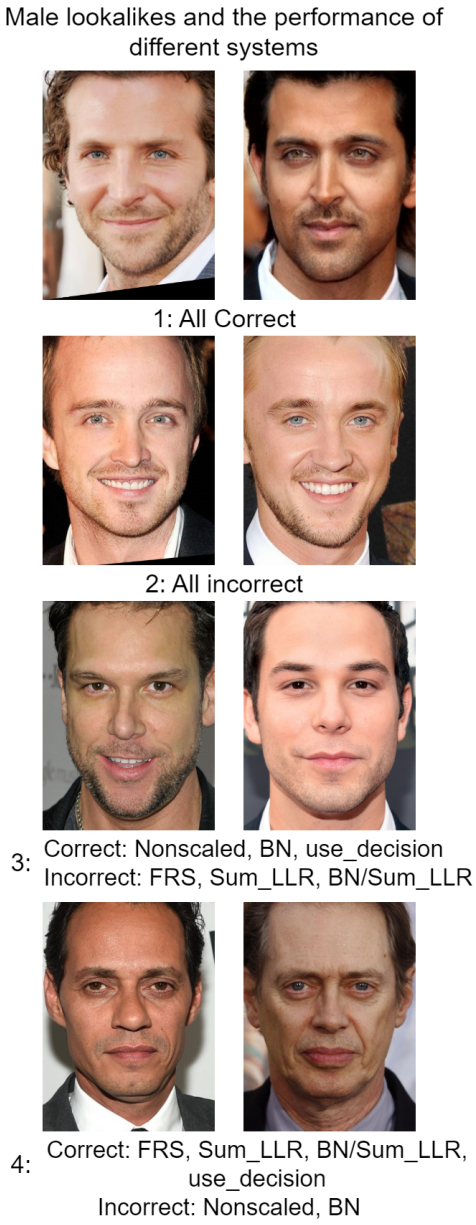



Fig. 12: Examples of male lookalike pairs and the performance of the different (2D) systems on these pairs. All images are from the HDA-döppelganger dataset.

performance of the different strategies (1-5). The images are shown in Figure 12 together with the performance of the different systems on these specific examples. Furthermore, the loglikelihood ratios that followed from the features are shown in a boxplot in Figure 13.

## V. Discussion

### A. *Dependence of score calculation of an FRS on facial regions*

In order to visualize the dependence of the score calculation of an FRS on facial regions, the results of the first research question are presented as heatmaps. The heatmaps show the

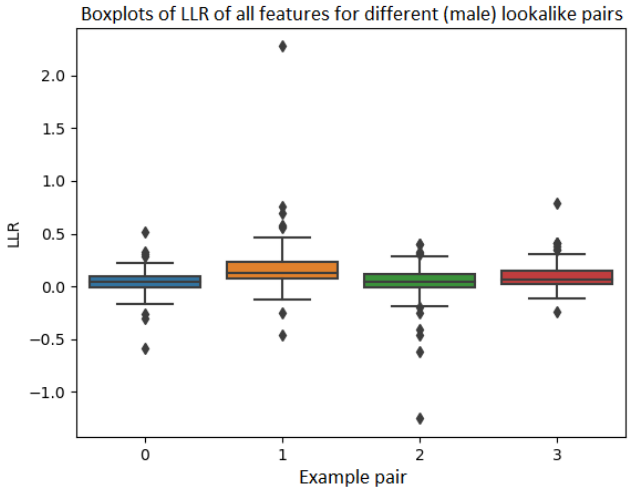



Fig. 13: Boxplots of the loglikelihood ratios of all features. The numbers on the x-axis correspond to the examples in Figure 12.

change in score per pixel (Equation (1)). A positive change means that the occluded image has a higher score than the original image, corresponding with an increase in dissimilarity (cosine distance). Thus, bright areas of the heatmap indicate parts of the image that, when occluded, cause an increase in dissimilarity (and vice versa).

Looking at Figures 10 and 11, a trend is visible that the nose and eye regions are important for the mated pairs for the rectangular, circular and feature occlusions. When these regions are occluded in the probe, it results in a positive change in the score, so the probe and gallery images look less similar. This means that these regions are essential regions for mated subjects. This is similar to the results of [17], who also showed that their evaluated network mainly focuses on the nose and eye areas. This is not clearly visible for the non-mated pairs. It could mean that although a patch in this area is occluded, enough information is still available to determine the pair is non-mated and the score is not influenced greatly. However, when focusing on the non-mated feature occlusions, it appears that the occlusion of contours of the eyebrows, eyes and nose have a negative change in score, meaning that the two subjects look more similar when these parts are covered. This could mean that the contours of facial features are important distinct features, that can be used to distinguish between two subjects. Again, these are the same important regions that were mentioned by [17]. Also [14] mentions that the regions of the eye and nose contain key information for face images. Lastly, looking at the lookalike pairs, it seems like a mix of the heatmaps of the mated and non-mated pairs. This is in line with the definition of a lookalike, namely a non-mated pair which "behaves" like a mated pair. Some parts around the eye and nose region are highlighted, but not as clear as in the case of mated pairs.

A difference between female and male pairs can be found,

TABLE VII: Percentages of correctly classified lookalike pairs per system when the classification thresholds are set at 1% FMR.

| | | FRS | Nonscaled | BN | Sum_LLR | BN/Sum_LLR | Use_decision |
|---|---|---|---|---|---|---|---|
| Female | 2D | 53.27 | 59.81 | 61.68 | 51.14 | 52.33 | 64.49 |
| | DIFF | 53.27 | 56.07 | 53.27 | 53.27 | 53.27 | 75.70 |
| Male | 2D | 67.54 | 71.05 | 65.79 | 67.54 | 69.30 | 66.67 |
| | DIFF | 66.67 | 60.52 | 63.16 | 67.54 | 65.79 | 76.32 |

TABLE VIII: Weights of the single neuron neural networks of both Sum_LLR strategies.

| | | **Sum_LLR** | **BN/Sum_LLR** |
|---|---|---|---|
| **2D** | **Summed Features** | 6.30137 | 0.63693 |
| | **FRS Score** | 12.94781 | 2.39703 |
| **DIFF** | **Summed Features** | 3.6907 | 0.62557 |
| | **FRS Score** | 12.88262 | 1.54494 |

TABLE IX: Top high and low (absolute) weights of the single neuron neural networks of four different fusion strategies. Explanation of the feature names can be found in Appendix B.

| | | **Nonscaled** | | **BN** | | **Use_decision** | |
|---|---|---|---|---|---|---|---|
| **2D** | **High weights** | FRS score | 8.9364 | FRS score | 1.47843 | FRS score | 7.95257 |
| | | nose_rel_inp_eyes | 3.07519 | Nose_Symmetry | 0.48125 | Nose_C | 1.61382 |
| | | Nose_Symmetry | 2.96015 | nose_rel_inp_eyes | 0.43283 | eyebrow_symmetry | 1.40303 |
| | | right_color_iris_G | 1.47614 | Color_Gright | 0.31053 | Nose_Size | 1.1494 |
| | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | **Low weights** | Symmetrylow | 0.01858 | left_eye_pos_rel_face | 0.01183 | Symmetryup | 0.02415 |
| | | symmetry_eyes | 0.01511 | Bleft | 0.0103 | Colorup_B | 0.02354 |
| | | right_diam_iris | 0.01277 | right_diam_iris | 0.00403 | right_diam_iris | 0.0142 |
| **DIFF** | **High weights** | FRS score | 8.88167 | FRS score | 1.40723 | FRS score | 7.95675 |
| | | Eright | 2.41294 | Contourleft | 0.39577 | Contourleft | 1.8342 |
| | | mouth_rel_inp_eyes | 2.1695 | Contourright | 0.32702 | Nose_Contour | 1.71756 |
| | | nose_rel_inp_eyes | 1.97394 | Aup | 0.28252 | Contourright | 1.59296 |
| | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | **Low weights** | Symmetryup | 0.00997 | distance_outer_eyes | 0.0013 | right_color_iris_G | 0.00694 |
| | | left_eye_contour | 0.00344 | right_eye_contour | 0.00093 | Colorup_G | 0.0024 |
| | | nose_rel_inner_eyes | 0.00334 | left_eye_contour | 0.00032 | Nose_Symmetry | 0.00114 |

when looking at the feature-based occlusions. In case of the female lookalike pairs it seems like occluding the features results in more dissimilar scores, especially around the eyebrow regions. This could mean that similar eyebrows are important features to explain why certain female pairs are considered lookalikes. Male non-mated and lookalike subjects show negative score changes around the eyebrow contours, meaning that when these features are covered they are considered more similar. This could indicate that the eyebrow shapes are unique features for non-mated male subjects. The researchers of [9] also showed that eyebrow corrections have an impact on recognition rates, meaning that these features contain important information.

The occlusions with pixels (visible in Figures 10 and 11 barely show any highlighted facial parts. Only the overall shape of the face is visible, which can be be explained by the fact that these pixels are not reached by all samples. This means that occluding random pixels does not change the score within the face and occluding rectangular, circular and feature shapes does. It indicates that the FRS does consider larger shapes in the face as small (pixel) changes do not influence the score remarkably.

Looking at Figure 9 it is visible that the area of focus remains in the center of the face around the nose when the size of the occlusion is changed. However, when the occlusions are small, one can see the shape of the eyes more clearly highlighted, as when the occlusions are large that focus seems to disappear. This is probably due to the fact that the average change per pixel is considered, resulting in less resolution in the heatmap of larger occlusions. The important areas remain highlighted when the size is increased.

*a) Limitations & Future research*

Although these experiments give insight in the focus areas of an FRS that are in line with literature, there are still limitations. The occlusions are not completely uniformly spread over the faces. This is because the positions are randomly determined and a limited number of occlusions per subject are placed. For a full analysis the occlusions should be placed systematically in the face to ensure full coverage, however this comes at a higher computational cost. Furthermore, only one open source FRS is used to compare the different types of pairs. For a more complete analysis, ideally several com-

mercial and open source systems should be compared to see whether they give similar results.

It is clear that the nose and eye(brow) regions are important to determine whether a pair is mated and that these areas seem to be less important for (random) non-mated subjects. However, as lookalikes behave similarly as mated pairs concerning these areas, it might be useful to focus on details in these areas. The features might seem similar, but as the lookalikes are non-mated, there will be differences. And for (random) non-mated pairs contours of the eyebrows and eyes seem important, instead of the whole areas. This is also mentioned by [14], to recognize morphed images they focus on details in the regions of the eyes and nose. This idea is transferred to this research, which leads to the results of the second research question.

### *B. RQ 2 - Fusion of FRS with forensic features*

The goal of this research question was to improve the performance of an FRS by fusing the result with forensic features. The performances of the different fusion strategies are summarized in Table IV. It is visible that several strategies reach similar performance as the original system (only FRS). This holds for both the 2D and Diff implementation. Only the Use_decision implementation seems to perform less than the other systems, which will be discussed later. The EER and AUC that are reached by our systems are comparable with the EER and AUC of the FRS. However, they never perform better.

This changes when we focus on the lookalikes, in Table VII the percentages of correctly classified lookalike pairs per system are shown. It is important to notice that the classification threshold of each system at 1% FMR is based on only the mated and (random) non-mated pairs. The Nonscaled and Use_decision systems seem to outperform the FRS (and other systems) on classifying the lookalikes. However, when the outcomes of the fusion methods are compared to the one of the FRS using the McNemar's test [34], which is *"a statistical test for comparing proportions from two dependent populations"* [35], none of the fusing methods show a significant difference with the original FRS. The p-values of each method can be found in the Appendix (Table XV). The Use_decision implementation seems to perform best on the lookalikes, however this strategy gave the highest EER and lowest AUC (Table IV). This can be explained by the choice of the thresholds on the likelihood ratios. They were set at 1% FMR, meaning that the system will be rather conservative to assign a mated match (as the FMR is to be kept low). More true mated pairs will be considered as non-mated, resulting in a higher FNMR, which could explain the lower performance in the AUC and EER. This also explains the higher percentage correctly classified lookalikes (correct means non-mated), because the system will decide in favor of non-mated in general.

The same holds for the voting system. Tables V and VI show that for both female and male subjects a high True Non-Match Rate (TNMR) is reached, meaning that the system is in almost all situations correct when appointing a non-mated pair, even in the situations of lookalikes. However, this comes at a cost, as the True Match Rate is very low, 29% in the case of female subjects. This means the system performs well by assigning the label non-mated to nearly all pairs, which makes the system unsuitable for real life situations. This is why the results of the voting system are not further considered. Changing the threshold to a higher FMR (visible in the Appendix section C) does result in better True Match Rates, however it comes with lower TNMRates. More experiments with varying thresholds have to be done to find the best settings, as well as the design a good comparison system with the original system (only FRS), to improve this fusion strategy.

The similarity in performance of Sum_LLR and BN/-Sum_LLR can be explained as only one other value (sum of LLR) is fused with the output of the FRS. These systems will mainly follow the score of the FRS, which is also shown in Table VIII where the weight of the score of the FRS has a higher order of magnitude than the weights of the fused scores. The same holds for the BN fusion method. In the situation of Nonscaled and Use_decision the top weights have the same order of magnitude. And although they are visibly lower than the weight of the FRS score, the features are more important than for the strategies previously mentioned. It can be assumed that the whole system therefore pays more attention to the forensic features.

Looking at the specific features that are in the top three of highest and lowest weights in Table IX one can see that in the case of 2D feature processing the nose-related features seem to be important. The Nonscaled, BN and Use_decision strategies all have several nose-related features in their top three. Most also have some eye(brow) related features, such as the eyebrow symmetry, in their top three. This corresponds with the findings in the first research question that the nose and eye region are important for determining whether a pair is mated. The diameter of the iris appears to be the least important feature. This could be caused by the fact that the resolution of the image is small compared to the feature size, and even smaller compared to subject-to-subject variations. The large quantization error in combination with a small signal makes that little information can be extracted from this feature. This could be different when the images have a higher resolution. In the case of Diff feature processing, two strategies show a strong favour for the contour features. In research question 1 it was found that covering the contours of features resulted in more similar scores. Meaning that contours are unique features of subjects, which could explain that these specific features are important to make a distinction between lookalikes. The Nonscaled strategy has no contour features in its top three. It could be that other features show more extreme values, which are highlighted by the Nonscaled strategy.

The individual lookalike examples in Figures 12 and 13 show that when a feature has an "extreme" postive LLR in comparison to the other values, all fusing strategies favour towards an incorrect mated match (example 2), as well as the FRS without fusion. Similarly, when there are no real extreme LLR values, the fusion strategies correspond with the

FRS, in this case a correct label is chosen. In the situation where only the Nonscaled, BN and Use_decision strategies are correct (example 3), one can see that the boxplot falls mainly on the positive side of LLR values. However, several extreme negative features are visible, which could explain why the strategies that make use of the individual features do give a correct prediction and the systems that sum all features do not, as the extreme values are outweighted by the overall positive LLRs. Furthermore, the FRS was wrong in this situation, so there are situations where some of the fused strategies clearly make use of the features to end up with a correct prediction. However, in the last example the FRS and strategies that sum the LLR do give a correct prediction. It is visible that the lower quartile of the data is very dense, meaning that the overall sum will remain low and thus a correct prediction follows. The other systems will probably pay too much attention to the positive extreme values and therefore give a wrong prediction.

*a) Limitations & Future research*

The feature extraction of all forensic features is based on the segmentation tool provided by [26]. Another segmentation tool could be used to evaluate the dependence of the feature extraction on the performance of the segmentation tool. Furthermore, it was found that some features probably suffer from quantization noise, in future research images with higher resolution should be used to investigate whether certain features perform differently. The features are modeled as normal distributions, however, it could be that the true densities are not normal distributions. And as *"the likelihood ratio test is optimal only when the underlying densities can be estimated very accurately"* [19], it could be that the LLR is not optimal in this situation. One could investigate the use of other density estimation techniques, such as kernel density estimation. In this research this option was explored shortly, however as it comes with higher computational costs (as the density estimation is more accurate), it was not used in further experiments and results were not collected. The evaluation on the lookalikes was only possible on a small dataset, around 200 valid samples of male and female pairs in total. In future research this dataset could be increased, for example by combining more images from each identity. Lastly, in the fusion strategies all features are combined directly. One could also consider fusing features of facial parts (such as eyes, nose, etc.) first, and then fuse the facial parts with the FRS score.

## VI. Conclusions

In this work two research questions were investigated with the goal to improve the performance of a Face Recognition System on challenging lookalike faces by focusing on forensic features.

The first question focused on the existing system: *To which extend does the score calculation of a Face Recognition System, with an explicit emphasis on lookalike sets, depend on facial regions?* Based on literature a system was developed that increasingly occluded parts of the face to visualize the change in score of a Face Recognition System (FRS) in a heatmap. Several occlusion strategies have been applied and it was shown that a FRS does look at larger shapes, as changing random pixels did not show a remarkable change in score. It was found that the areas of the eye(brows) and nose seem to be rather important for mated pairs, as occluding these areas results in a visible change in score. However, these areas seem of less importance for non-mated pairs. Special attention was given to the case of lookalike pairs, and it was found is that these areas also seem important for lookalikes, but less than for mated pairs. Meaning that these could be crucial areas for distinguishing between lookalikes. When the system would be forced to pay more attention to these areas, it could learn how to differentiate better between the subjects.

This leads us to the second research question: *Can the performance of an existing Face Recognition System be improved by fusing the output with forensic features?* In order to do so features from the forensic feature list of FISWG [18] were selected and automatically retrieved from subjects. These features were modelled as normal distributions for mated and non-mated pairs. For each pair of subjects the log-likelihood ratios are computed for all features, which are then fed to a fusion system where these ratios are combined with the score of the FRS. Several fusion strategies were explored and both score and decision level fusion were applied. It was found that the overall system was not improved by any of the strategies, however comparable performance is reached. The important features were found to be features concerning the eyebrows, nose and contours of facial features, which is in line with the results found in the first research question. When examining the performance of the systems on lookalikes, it was found that the system did not significantly improve the overall performance, but on individual examples some systems outperformed the FRS. These systems used a score based fusion and incorporated the separate features and therefore showed that incorporating forensic features has potential for better performance on lookalikes.

Summarizing, in this research it was investigated whether the performance of an existing face recognition system could be improved by incorporating forensic features. Although the overall performance was not improved, on individual examples several new systems showed improvement. This shows potential for future research, where the experiments can, for example, be expanded by larger databases on lookalikes and different density estimation techniques can be applied to further improve performance.

## VII. Acknowledgements

I would like to thank my supervisors for our weekly interesting discussions and their overall support during the process. I would also like to show my appreciation to my peers for assisting me in expanding the HDA-doppelganger database [27] with genuine images.

## References


[1] Z. Akhtar and A. Rattani, "A face in any form: New challenges and opportunities for face recognition technology," *Computer*, vol. 50, no. 4, pp. 80–90, 2017. DOI: 10.1109/MC.2017.119.

[2] C. Rathgeb, D. Fischer, P. Drozdowski, and C. Busch, *Reliable detection of doppelgängers based on deep face representations*, 2022. arXiv: 2201.08831 [cs.CV].

[3] C. Rathgeb, P. Drozdowski, M. Obel, *et al.*, "Impact of doppelgängers on face recognition: Database and evaluation," in *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2021, pp. 1–4. DOI: 10.1109/BIOSIG52210.2021.9548306.

[4] J. McCauley, S. Soleymani, B. Williams, J. Dando, N. Nasrabadi, and J. Dawson, "Identical twins as a facial similarity benchmark for human facial recognition," in *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2021, pp. 1–5. DOI: 10.1109/BIOSIG52210.2021.9548299.

[5] J. R. Paone, P. J. Flynn, P. J. Philips, *et al.*, "Double trouble: Differentiating identical twins by face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 2, pp. 285–295, 2014. DOI: 10.1109/TIFS.2013.2296373.

[6] H. Lamba, A. Sarkar, M. Vatsa, R. Singh, and A. Noore, "Face recognition for look-alikes: A preliminary study," in *2011 International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1–6. DOI: 10.1109/IJCB.2011.6117520.

[7] A. Rosenfeld, M. D. Solbach, and J. K. Tsotsos, *Totally looks like - how humans compare, compared to machines*, 2018. arXiv: 1803.01485 [cs.CV].

[8] P. J. Phillips, A. N. Yates, Y. Hu, *et al.*, "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms," *Proceedings of the National Academy of Sciences*, vol. 115, no. 24, pp. 6171–6176, 2018, ISSN: 0027-8424. DOI: 10.1073/pnas.1721355115. eprint: https://www.pnas.org/content/115/24/6171.full.pdf. [Online]. Available: https://www.pnas.org/content/115/24/6171.

[9] C. Rathgeb, D. Dogan, F. Stockhardt, M. De Marsico, and C. Busch, "Plastic surgery: An obstacle for deep face recognition?" In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 3510–3517. DOI: 10.1109/CVPRW50498.2020.00411.

[10] A. Röttcher, U. Scherhag, and C. Busch, "Finding the suitable doppelgänger for a face morphing attack," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 2020, pp. 1–7. DOI: 10.1109/IJCB48548.2020.9304878.

[11] E.-V. Pikoulis, Z.-M. Ioannou, M. Paschou, and E. Sakkopoulos, "Face morphing, a modern threat to border security: Recent advances and open challenges," *Applied Sciences*, vol. 11, no. 7, 2021, ISSN: 2076-3417. DOI: 10.3390/app11073207. [Online]. Available: https://www.mdpi.com/2076-3417/11/7/3207.

[12] I. Batskos, F. F. de Wit, L. J. Spreeuwers, and R. J. Veldhuis, "Preventing face morphing attacks by using legacy face images," *IET Biometrics*, vol. 10, no. 4, pp. 430–440, 2021. DOI: https://doi.org/10.1049/bme2.12047. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/bme2.12047. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/bme2.12047.

[13] Y. Xu, G. Jia, H. Huang, J. Duan, and R. He, "Visual-semantic transformer for face forgery detection," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, 2021, pp. 1–7. DOI: 10.1109/IJCB52358.2021.9484407.

[14] Z. Chen and H. Yang, "Attentive semantic exploring for manipulated face detection," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021. DOI: 10.1109/icassp39728.2021.9414225. [Online]. Available: http://dx.doi.org/10.1109/ICASSP39728.2021.9414225.

[15] M. Jacquet and C. Champod, "Automated face recognition in forensic science: Review and perspectives," *Forensic Science International*, vol. 307, p. 110 124, 2020, ISSN: 0379-0738. DOI: https://doi.org/10.1016/j.forsciint.2019.110124. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0379073819305365.

[16] Y. Zhong and W. Deng, "Exploring features and attributes in deep face recognition using visualization techniques," in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, 2019, pp. 1–8. DOI: 10.1109/FG.2019.8756546.

[17] G. Castanon and J. Byrne, "Visualizing and quantifying discriminative features for face recognition," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 2018, pp. 16–23. DOI: 10.1109/FG.2018.00013.

[18] Facial-Identification-Scientific-Working-Group, *Facial Comparison Overview and Methodology Guidelines*, 2nd ed. FISWG, 2019. [Online]. Available: https://fiswg.org/documents.html.

[19] A. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of multibiometrics*, 1st ed. Springer, 2006.

[20] Y. Zhong and W. Deng, "Deep difference analysis in similar-looking face recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3353–3358. DOI: 10.1109/ICPR.2018.8545449.

[21] D. Mery and B. Morris, "On black-box explanation for face verification," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1194–1203. DOI: 10.1109/WACV51458.2022.00126.

[22] J. R. Williford, B. B. May, and J. Byrne, *Explainable face recognition*, 2020. DOI: 10.48550/ARXIV.2008.

00916. [Online]. Available: https://arxiv.org/abs/2008.00916.

[23] C. Zeinstra, R. Veldhuis, and L. Spreeuwers, "Towards the automation of forensic facial individualisation: Comparing forensic to non forensic eyebrow features," English, in *35rd WIC Symposium on Information Theory in the Benelux and The 4th WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux*, B. Skoric and T. Ignatenko, Eds., http://www.win.tue.nl/sitb2014/proceedings.pdf; 35th WIC Symposium on Information Theory in the Benelux 2014 ; Conference date: 12-05-2014 Through 13-05-2014, Netherlands: Werkgemeenschap voor Informatie- en Communicatietheorie (WIC), May 2014, pp. 73–80, ISBN: 978-90-365-3383-6.

[24] C. Zeinstra, R. Veldhuis, and L. Spreeuwers, "Discriminating power of fiswg characteristic descriptors under different forensic use cases," in *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2016, pp. 1–7. DOI: 10.1109/BIOSIG.2016.7736919.

[25] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2015. DOI: 10.1109/cvpr.2015.7298682. [Online]. Available: https://doi.org/10.1109%5C%2Fcvpr.2015.7298682.

[26] Shaoanlu, *Github - shaoanlu/face$_t$oolbox$_k$eras: A collection of deep learning frameworks ported to keras for face analysis.* 2019. [Online]. Available: https://github.com/shaoanlu/face_toolbox_keras.

[27] C. Rathgeb, P. Drozdowski, M. Obel, *et al.*, "Impact of doppelgngers on face recognition: Database and evaluation," *20th International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2021.

[28] "Dlib c++ library." (May 8, 2022), [Online]. Available: http://dlib.net/.

[29] Zllrunning, *Github - zllrunning/face-parsing.pytorch: Using modified bisenet for face parsing in pytorch*, 2019. [Online]. Available: https : / / github . com / zllrunning/face-parsing.PyTorch.

[30] Burak, *Pins face recognition*, 2022. [Online]. Available: https://www.kaggle.com/datasets/hereisburak/pins-face-recognition.

[31] 2022. [Online]. Available: https://docs.opencv.org/3.4/d4/d73/tutorial_py_contours_begin.html.

[32] "Keras tensorflow." (), [Online]. Available: https://keras.io/about/ (visited on 11/03/2022).

[33] "Tensorflow batch normalization." (Oct. 26, 2022), [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/layers/BatchNormalization (visited on 10/29/2022).

[34] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1, 1947, ISSN: 1860-0980. DOI: 10 . 1007 / BF02295996. [Online]. Available: https : / / doi . org / 10 . 1007 / BF02295996.

[35] C. De-Yu. "Mcnemar's test, with python." (Feb. 2022), [Online]. Available: https : / / towardsdatascience . com / mcnemars-test-with-python-e1bab328d15c (visited on 12/05/2022).

[36] S. Park, *Github - swook/gazeml: Gaze estimation using deep learning, a tensorflow-based framework.* 2019. [Online]. Available: https://github.com/swook/GazeML.

[37] Saigolp. "Eyes vector," Deviant Art. (Sep. 17, 2015), [Online]. Available: https : / / www . deviantart . com / saigolp / art / Eyes - Vector - 560869217 (visited on 10/21/2022).

[38] "Nose PNG," PNG ALL. (Jan. 23, 2020), [Online]. Available: https://www.pngall.com/nose-png/download/39923 (visited on 10/21/2022).

[39] 2022. [Online]. Available: https://docs.opencv.org/3.4/d8/d23/classcv_1_1Moments.html.

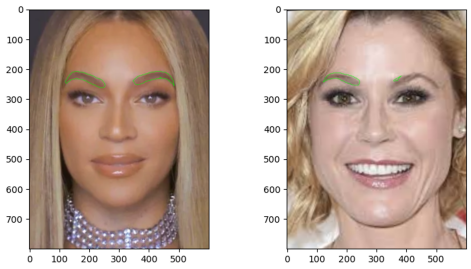


Fig. 14: Examples of issues with eyebrow segmentation that can be solved. Images are from [27].

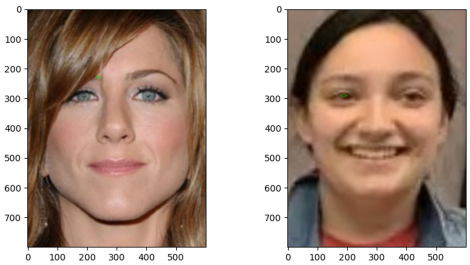


Fig. 15: Examples of issues with eyebrow segmentation that cannot be solved. Images are from [27].

## APPENDIX A
## DESCRIPTION OF FISWG FEATURE EXTRACTION

*a) Eye and eyebrow features*

These features were most challenging to retrieve, as the segmentation process was not always perfect. First of all, sometimes no eyebrow or eye could be found due to occlusions. It also happened that only a part of the eyebrow or eye was recognized. Examples are shown in Figure 15, the segmented parts are shown as a green contour. In both examples the segmented parts are too small. Another problem was that both eyebrows were classified as the right or left eyebrow and the other was not found, the eyebrows were swapped or an extra small contour was found additional to a correct contour, see the examples in Figure 14. These situations could be solved. The same holds for the eyes. After checking for these mistakes and if possible, compensating for them, the features are retrieved. An overview of the eyebrow features is shown in Appendix Table X. A graphical overview is given for the position related features (feature 1-5 and 10-14) in Figure 16. It is important to notice that the x- and y-coordinates start with [0,0] in the upperleft corner, see the axes in Figure 14. An overview of the eye features is shown in Appendix Table XI and several position related features are shown in Figure 17. Some of the eye features use iris landmarks, these are retrieved by an iris detector. This detector is part of the used Facetoolbox, provided by [26]. It is originally an Tensorflow based network that was meant for eye region landmark based gaze estimation [36]. The network estimates the position of the iris and returns 8 landmarks of the contour. An example is shown in Figure 18, where the iris landmarks are shown in blue.

*b) Nose features*

For this feature the only check is whether it is present and whether the size of the found shape is large enough to be considered a nose. Due to, for example, occlusions some noses were not correctly segmented and then the features become meaningless, so these situations are filtered. For most features of the nose in the FISWG list, the profile view is necessary [18], which results in a rather small list of features compared to the eyebrow and eye features. An overview of the features can be found in Appendix Table XII, the position related features are shown in Figure 19.

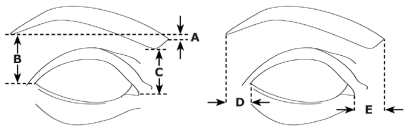



Fig. 16: The position related features of the eyebrow, image is copied from [18] page 11.

*c) Mouth features*

The segmentation parts of the mouth are divided into two shapes: the upper lip and the lower lip. Again, a check for both parts whether they are present and large enough is done. Problems occur for example when the subject has a beard that occludes parts of the lip. An overview of the mouth features is shown in Appendix Table XIII, the position related features are shown in Figure 20.

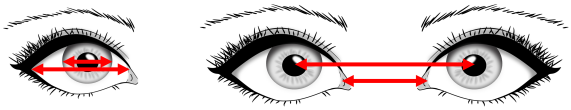


Fig. 17: The position related features of a single eye (left) and both eyes (right). The description of the features can be found in Table XI. The image is copied from [37]; arrows are added.

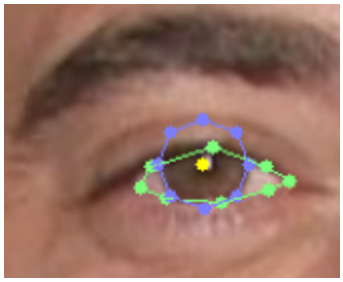


Fig. 18: Example of the eye region landmarks following from the iris detector provided by [26].

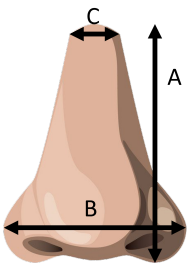



Fig. 19: The position related features of the nose. This image is copied from [38] - licensed under CC-BY-NC; arrows are added.

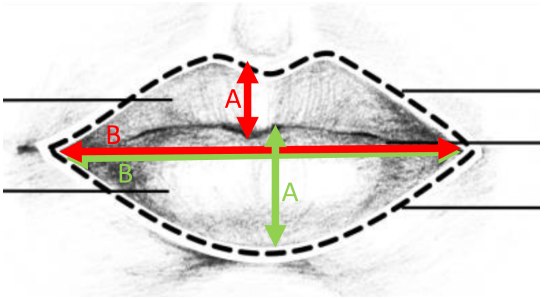



Fig. 20: The position related features of the mouth. Red visualizes the features for the upper lip, the green arrows represent the features for the lower lip. Image is copied from [18]; arrows are added.

*d) Head features*

In the FISWG list there are several features related to facial proportions, which are categorized as head features in this research. The face shape is retrieved by the segmentation map of the skin. It has to be noted that this is without the hair, so it is not the full skull shape, but chin to hairline. The overview of features can be found in Appendix Table XIV. Most features use features that can already be found in the other feature overviews.

*e) Symmetry features*

In Tables X to XIV (found in Appendix) several features related to symmetry have been mentioned. The symmetry between left and right eyebrows and eyes (two shapes) are computed and the symmetry of the shapes of the nose and lips (within shape) are computed.

*2 shape symmetry*

In the situation with two shapes, the first step is to overlay the shapes. This is done by determining the centroid of the shapes. Using the opencv library one can find the moments of a contour [39]. In order to find the centroid one can use Equation (3), where $\{\bar{x}, \bar{y}\}$ are the coordinates of the centroid. The next step is to translate one of the two shapes to $\{0, 0\}$ by subtracting the centroid of all coordinates. Then all x-coordinates are multiplied by $-1$ to mirror the shape, after which another translation is done, to overlay the two shapes. This is done by adding the centroid of the second shape to all the coordinates of the first (mirrored) shape. Lastly, the opencv function cv2.matchShapes which uses Hu moments to compare two shapes.

$$\{\bar{x}, \bar{y}\} = \left\{ \frac{M_{10}}{M_{00}}, \frac{M_{01}}{M_{00}} \right\} \tag{3}$$

*1 shape symmetry*

In case there is one shape, first its centroid (see Equation (3) is found and the shape is translated to $\{0, 0\}$. Then the shape is divided in a left part and right part, by looking at all positive and negative x-coordinates. One of the parts is mirrored and cv2.matchShapes is used to quantify the symmetry.

*f) Contour features*

All previously mentioned features are computed per subject. There are seven features left that are computed between two subjects. The face parts that are used are expressed in segmentation maps and later on as contours (see Section III-B2). The contours are expressed in coordinates. Using these contours one can compare two areas. First the two shapes are overlaid using the centroids (see Equation (3), where $\{\bar{x}, \bar{y}\}$ are the coordinates of the centroid). Then the contours are filled and a logical XOR operation is performed on the two filled shapes. Then all pixels are counted to compute the unique areas (the non-overlapping parts of the contours). Which is the difference of the two shapes between two subjects. This is done for the contours of the left and right eyebrow, left and right eye, nose, upperlip and lowerlip.

## APPENDIX B
## FEATURE EXTRACTION TABLES

TABLE X: Overview of eyebrow features

| | Facepart | Feature | How is it retreived |
|---|---|---|---|
| 1 | Eyebrow | Aleft | Upperright y-coordinate of the eyebrow minus the lowest y-coordinate of the left side of the eyebrow. (First all extreme x-coordinates are found and then extreme y-coordinates among these sets are found.) |
| 2 | Eyebrow | Bleft | Upperleft y-coordinate of the eye (left eyecorner) minus the lowest y-coordinate of the left side of the eyebrow. (First all extreme x-coordinates are found and then extreme y-coordinates among these sets are found. ) |
| 3 | Eyebrow | Cleft | Lowerright y-coordinate of the eye (right eyecorner) minus lowerleft y-coordinate of the eyebrow. (First all extreme x-coordinates are found and then extreme y-coordinates among these sets are found.) |
| 4 | Eyebrow | Dleft | Lowerleft x-coordinate of the eyebrow minus upperleft x-coordinate of the eye (left eyecorner) |
| 5 | Eyebrow | Eleft | Lowerright x-coordinate of the eye minus the most left coordinate of the eye. |
| 6 | Eyebrow | Sizeleft | Number of pixels in segmentation map |
| 7 | Eyebrow | Color_Rleft | Average color in RGB, R value |
| 8 | Eyebrow | Color_Bleft | Average color in RGB, B value |
| 9 | Eyebrow | Color_Gleft | Average color in RGB, G value |
| 10 | Eyebrow | Aright | see Aleft, but for other eyebrow |
| 11 | Eyebrow | Bright | see Bleft, but for other eyebrow |
| 12 | Eyebrow | Cright | see Cleft, but for other eyebrow |
| 13 | Eyebrow | Dright | see Dleft, but for other eyebrow |
| 14 | Eyebrow | Eright | see Eleft, but for other eyebrow |
| 15 | Eyebrow | Sizeright | see Sizeleft, but for other eyebrow |
| 16 | Eyebrow | Color_Rright | see Color_Rleft, but for other eyebrow |
| 17 | Eyebrow | Color_Bright | see Color_Bleft, but for other eyebrow |
| 18 | Eyebrow | Color_Gright | see Color_Gleft, but for other eyebrow |
| 19 | Eyebrow | eyebrow_symmetry | see symmetry explanation in **??** A-0e |

TABLE XI: Overview of eye features

| | Facepart | Feature | How is it retreived |
|---|---|---|---|
| 1 | Eye | left_size_iris | Take the landmarks of the iris and create a cv2 contour; take the number of pixels within the contour as the size. |
| 2 | Eye | left_size_iris_circle | Take the landmarks of the iris retreived by the eyerecognizer and compute the average diameter of the iris. Using the diameter, compute the area of the iris as a circle. |
| 3 | Eye | left_color_iris_R | Take the landmarks of the iris and create a cv2 contour; take R value of the average RGB coded color. |
| 4 | Eye | left_color_iris_G | Take the landmarks of the iris and create a cv2 contour; take G value of the average RGB coded color. |
| 5 | Eye | left_color_iris_B | Take the landmarks of the iris and create a cv2 contour; take B value of the average RGB coded color. |
| 6 | Eye | left_diam_iris | Take the landmarks of the iris retreived by the eyerecognizer and compute the average diameter of the iris. Divide by the horizontal distance of the eye opening (most left and most right x-coordinates). |

**Table XI continued from previous page**

| | Facepart | Feature | How is it retreived |
|---|---|---|---|
| 7 | Eye | left_visib_iris | Take the contour of the iris and the contour of the eye as masks and compute the overlap of the masks. Then find the intersections of the overlap and the eyemask (xor operation), which are the non visible parts of the iris. The visible part of the iris is the difference of the total size of the iris and the non visible part of the iris. Divide this difference by the total size of the iris to quantize the visibility as a percentage. |
| 8 | Eye | left_size_eye | Number of pixels in segmentation map of this eye. |
| 9 | Eye | right_size_iris | See left_size_iris |
| 10 | Eye | right_size_iris_circle | See left_size_iris_circle |
| 11 | Eye | right_color_iris_R | See left_color_iris_R |
| 12 | Eye | right_color_iris_G | See left_color_iris_G |
| 13 | Eye | right_color_iris_B | See left_color_iris_B |
| 14 | Eye | right_diam_iris | See left_diam_iris |
| 15 | Eye | right_visib_iris | See left_visib_iris |
| 16 | Eye | right_size_eye | See left_size_eye |
| 17 | Eye | distance_inner_eyes | Most left x-coordinate of the left eye minus the most right x-coordinate of the right eye (intercanthal distance). |
| 18 | Eye | distance_outer_eyes | Most right x-coordinate of the left eye minus the most left x-coordinate of the right eye |
| 19 | Eye | distance_pupils | Compute pupil centers of both eyes as the average x- and y-coordinates of the iris landmarks. Take the difference of the x-coordinates of the left and right eye. |
| 20 | Eye | offset | Difference of y-coordinates of pupil centers (see distance_pupils) of left and right eye. |
| 21 | Eye | symmetry_eyes | See symmetry explanation in **??** A-0e. |

TABLE XII: Overview of nose features

| | Facepart | Feature | How is it retreived |
|---|---|---|---|
| 1 | Nose | Nose_A | Difference between minimum and maximum y-coordinate of the nose contour. |
| 2 | Nose | Nose_B | Difference between minimum and maximum x-coordinate of the nose contour. |
| 3 | Nose | Nose_C | Look in the lower ten percent of the y-coordinates of the nose. Find the minimum and maximum x-coordinates to compute the width of the nasal bridge. |
| 4 | Nose | Nose_Size | Number of pixels in segmentation map. |
| 5 | Nose | Nose_Color_R | Average color in RGB format, R value. |
| 6 | Nose | Nose_Color_B | Average color in RGB format, B value. |
| 7 | Nose | Nose_Color_G | Average color in RGB format, G value. |
| 8 | Nose | Nose_Symmetry | See symmetry explanation in **??** A-0e. |

TABLE XIII: Overview of mouth features

| | Facepart | Feature | How is it retreived |
|---|---|---|---|
| 1 | Mouth | Aup | Take the middle of the lip as the average of the most left and most right x-coordinates. Search in the region of \pm 10 coordinates around the middle of the contour for the maximum and minimum y-coordinate. Take the difference of these y-coordinates to find the height of the lip. |
| 2 | Mouth | Bup | Take the difference of the most left and most right coordinates of the contour. |
| 3 | Mouth | Sizeup | The number of pixels in the contour. |
| 4 | Mouth | Colorup_R | Average color in RGB format, R value. |
| 5 | Mouth | Colorup_B | Average color in RGB format, B value. |
| 6 | Mouth | Colorup_G | Average color in RGB format, G value. |
| 7 | Mouth | Symmetryup | See symmetry features. |
| 8 | Mouth | Alow | See Aup. |
| 9 | Mouth | Blow | See Bup. |
| 10 | Mouth | Sizelow | See Sizeup. |
| 11 | Mouth | Colorlow_R | See Colorup_R. |
| 12 | Mouth | Colorlow_G | See Colorup_G. |
| 13 | Mouth | Colorlow_B | See Colorup_B. |
| 14 | Mouth | Symmetrylow | See symmetry features in **??** A-0e. |

TABLE XIV: Overview of head features

| | Facepart | Feature (short) | Feature (long) | How is it retreived |
|---|---|---|---|---|
| 1 | Head | nose_rel_inp_eyes | Nose width relative to innerpupillary distance eyes | Nose_B divided by distance_pupils |
| 2 | Head | nose_rel_outer_eyes | Nose width relative to width outer corners eyes | Nose_B divided by distance_outer_eyes |
| 3 | Head | nose_rel_inner_eyes | Nose width relative to width inner corners eyes | Nose_B divided by distance_inner_eyes |
| 4 | Head | nose_rel_left_eye | Nose width relative to left eye width | Nose_B divided by difference of most left and most right x-coordinates of left eye |
| 5 | Head | nose_rel_right_eye | Nose width relative to right eye width | Nose_B divided by difference of most left and most right x-coordinates of right eye. |
| 6 | Head | mouth_rel_inp_eyes | Mouth width relative to interpupillary distance eyes | Bup divided by distance_pupils |
| 7 | Head | mouth_rel_outer_eyes | Mouth width relative to width outer corners eyes | Bup divided by distance_outer_eyes |
| 8 | Head | mouth_rel_inner_eyes | Mouth width relative to width inner corners eyes | Bup divided by distance_inner_eyes |
| 9 | Head | mouth_rel_left_eye | Mouth width relative to left eye width | Bup divided by difference of most left and most right x-coordinates of left eye |
| 10 | Head | mouth_rel_right_eye | Mouth width relative to right eye width | Bup divided by difference of most left and most right x-coordinates of right eye. |
| 11 | Head | nose_rel_mouth | Nose width relative to mouth width | Nose_B divided by Bup |
| 12 | Head | nose_upperlip_rel_face | Distance nose to upperlip relative to facelength | Take the difference between highest y-coordinate of the nose and the lowest y-coordinate of the upperlip. Divide by the difference of the highest and lowest y-coordinates of the face (length of the face). |
| 13 | Head | chin_lowerlip_rel_face | Distance chin to lowerlip relative to facelength | Take the difference between the highest y-coordinate of the nose and the highest y-coordinate of the face and divide this by the length of the face. |
| 14 | Head | left_eye_pos_rel_face | Left eye postition relative to facelength | Use cv2.moments to find the center of the eye (see symmetry features) and take the difference between this center and the lowest point of the face. Divide this by the length of the face. |
| 15 | Head | right_eye_pos_rel_face | Right eye position relative to facelength | See left_eye_pos_rel_face. |
| 16 | Head | face_length | Facelength. | Difference between lowest and highest y-coordinates of the face. |

TABLE XV: Percentages of correctly classified lookalike pairs per system and p-value of the McNemar's test [34] of the system compared to the FRS.

| | | **FRS** | **Nonscaled** | | **BN** | | **Sum_LLR** | | **BN/Sum_LLR** | | **Use_decision** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | % correct | % correct | p-value | % correct | p-value | % correct | p-value | % correct | p-value | % correct | p-value |
| Female | 2D | 53.27 | 59.81 | 0.1213 | 61.68 | 0.0523 | 51.14 | 0.4795 | 52.33 | 1.0 | 64.49 | 6.834e-09 |
| | Diff | 53.27 | 56.07 | 0.4497 | 53.27 | 0.7237 | 53.27 | n/a | 53.27 | n/a | 75.70 | 7.998e-05 |
| Male | 2D | 67.54 | 71.05 | 0.5224 | 65.79 | 0.8231 | 67.54 | 0.6831 | 69.30 | 0.6171 | 66.67 | 0.001723 |
| | Diff | 66.67 | 60.52 | 0.1456 | 63.16 | 0.4227 | 67.54 | 1.0 | 65.79 | 1.0 | 76.32 | 0.07249 |

## APPENDIX C
## EXTRA RESULTS

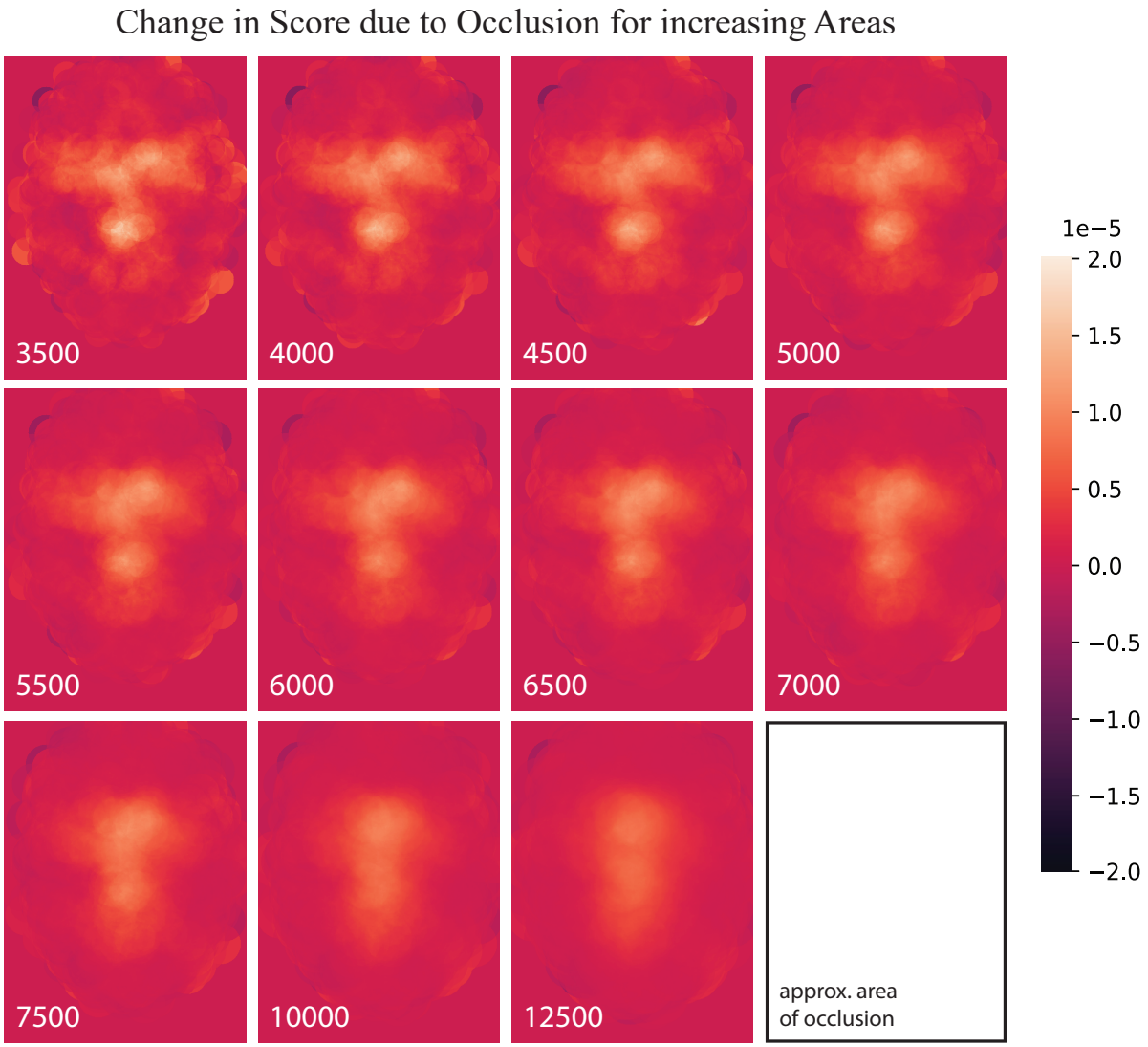



Fig. 21: Heatmaps of the score change of genuine female pairs with a circular average skin color occlusion. The eleven heatmaps all have different areas of occlusion which are shown in the lower left corner of each heatmap.

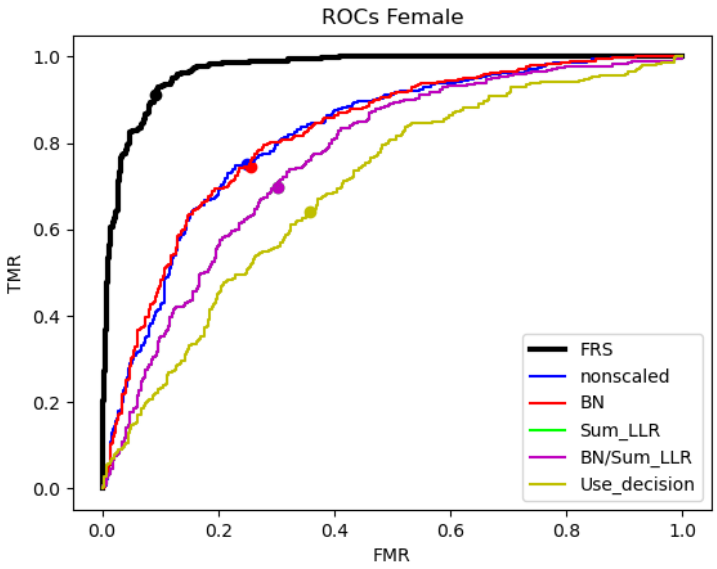



Fig. 22: ROC-curves of different '2D' feature fusion strategies without the score of the FRS on female subjects. The black ROC-curve shows the result of using only the FRS. The dots show the points of the EER.

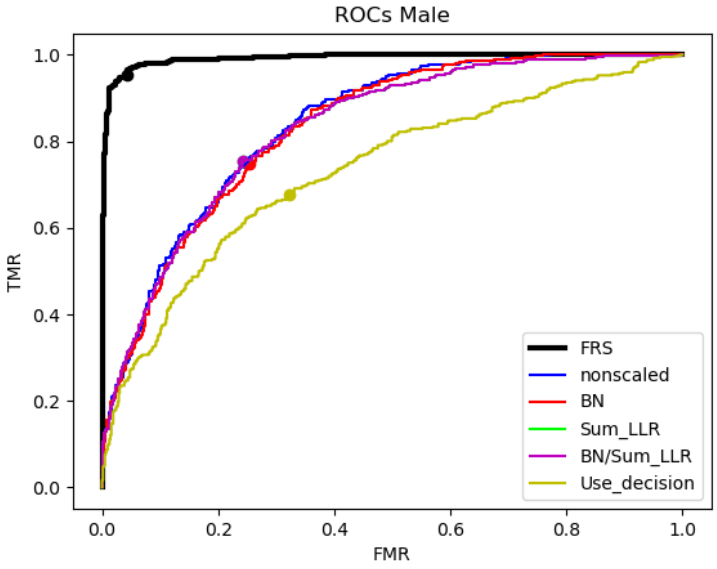



Fig. 23: ROC-curves of different '2D' feature fusion strategies without the score of the FRS on male subjects. The black ROC-curve shows the result of using only the FRS. The dots show the points of the EER.

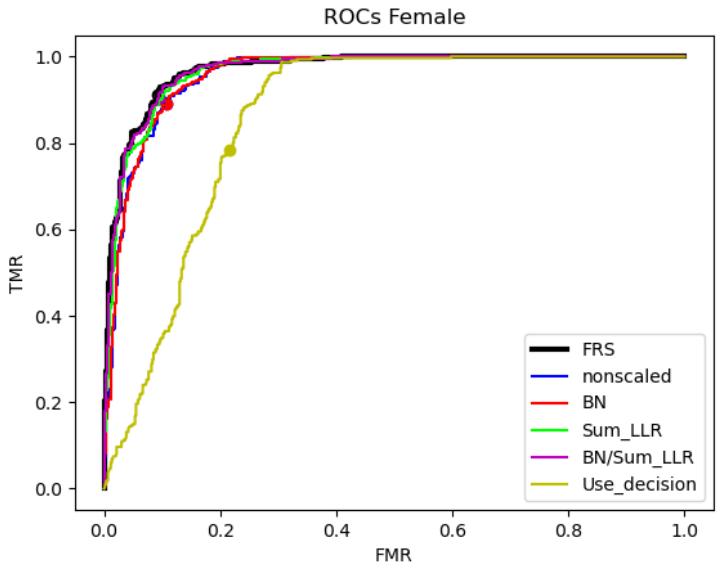



Fig. 24: ROC-curves of different '2D' feature fusion strategies using the score of the FRS on female subjects. The black ROC-curve shows the result of using only the FRS. The dots show the points of the EER.

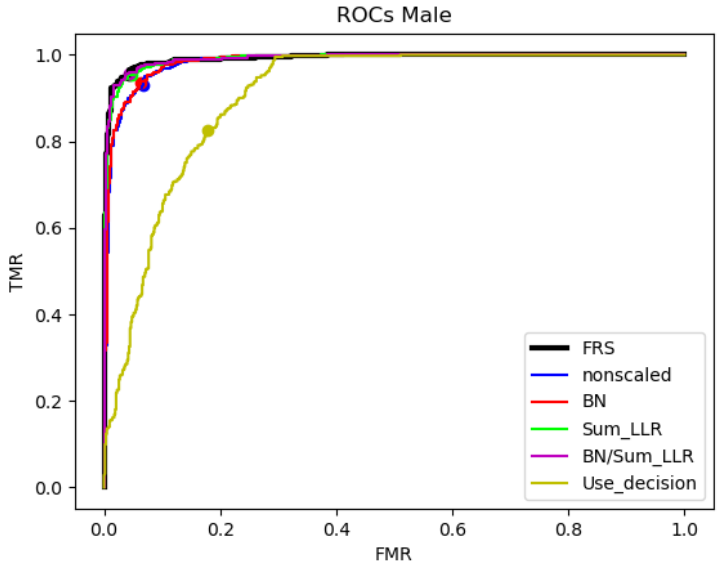



Fig. 25: ROC-curves of different '2D' feature fusion strategies using the score of the FRS on male subjects. The black ROC-curve shows the result of using only the FRS. The dots show the points of the EER.

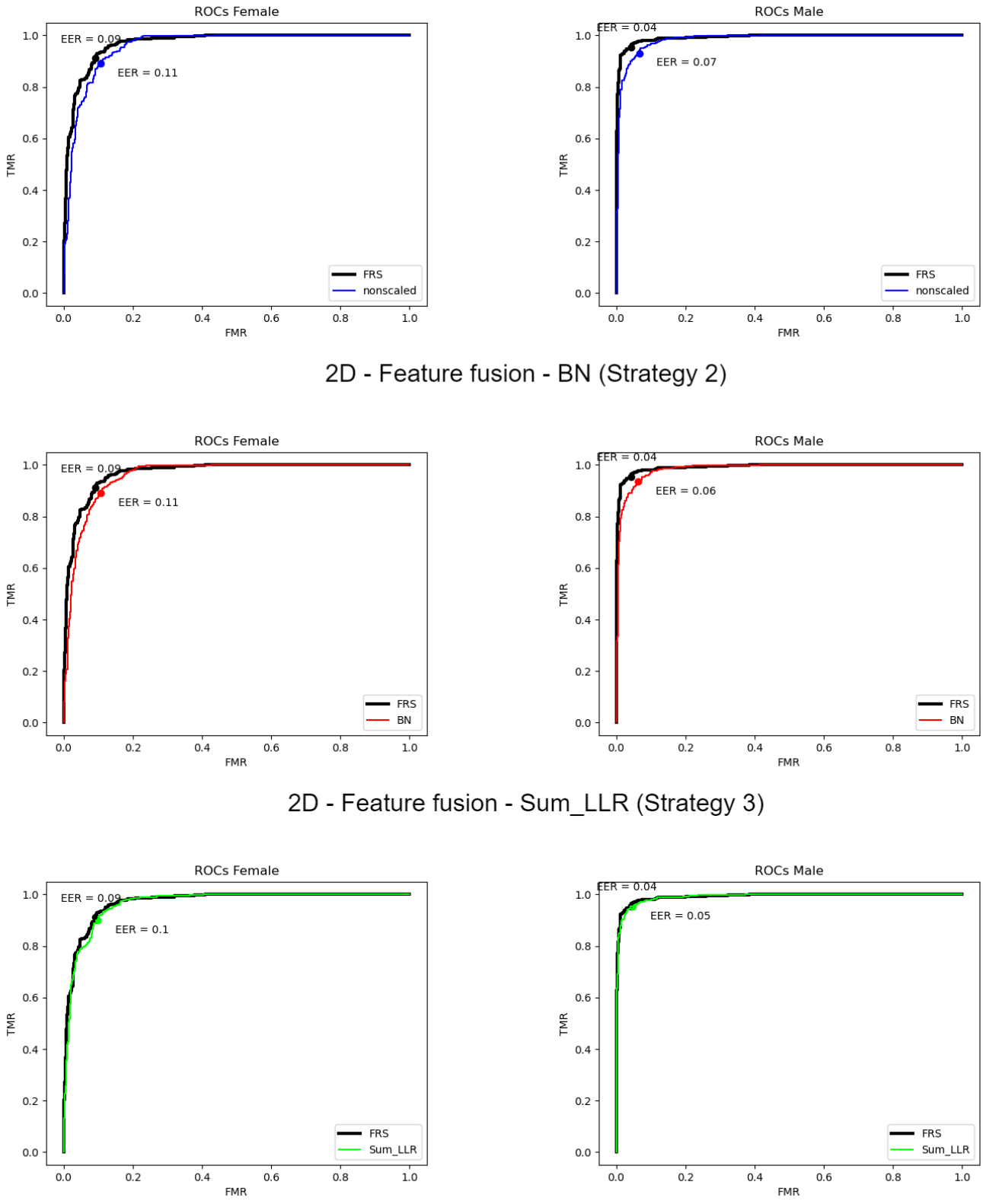



Fig. 26: ROC-curves of '2D' strategy 1, 2 and 3 together with the ROC-curve of the FRS.

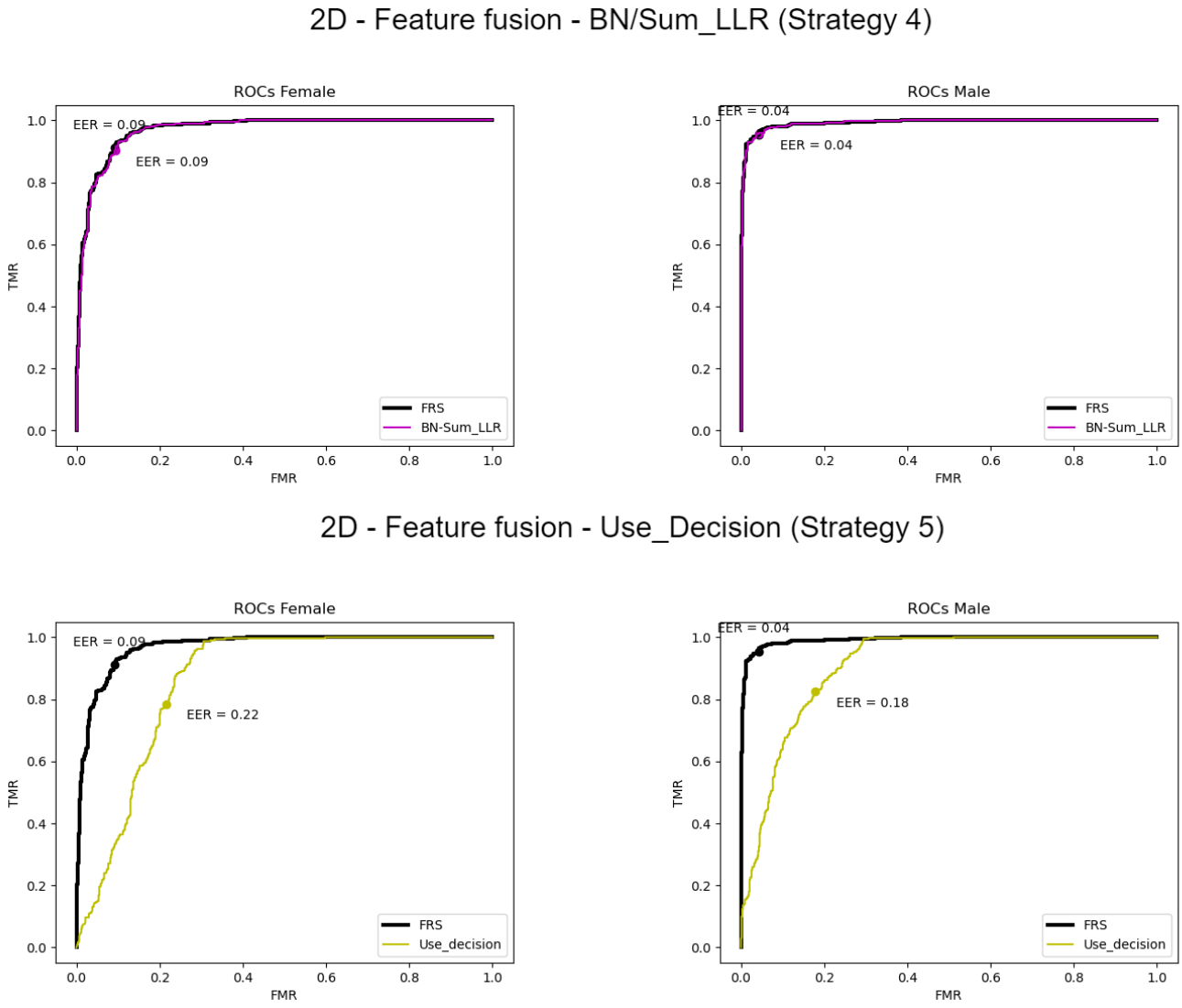



Fig. 27: ROC-curves of '2D' strategy 4 and 5 together with the ROC-curve of the FRS.

TABLE XVI: Top High and Low LLR of the forensic features for different examples of lookalike pairs. The examples can be found in Figure 12

| | **All correct** | | **All incorrect** | | **Correct: Non-scaled; BN; use_decision Incorrect: FRS, Sum_LLR, BN/Sum_LLR** | | **Correct: FRS, Sum_LLR, BN/Sum_LLR, use_decision Incorrect: Non-scaled, BN** | |
|---|---|---|---|---|---|---|---|---|
| | *FRS score* | *0.597520202* | *FRS score* | *0.393041432* | *FRS score* | *0.42572993* | *FRS score* | *0.489703774* |
| **High LLR** | Bright | 0.515560502 | Sizeright | 2.275894344 | nose_rel_inner_eyes | 0.405804894 | Bright | 0.793499295 |
| | Cright | 0.32815138 | left_color_iris_G | 0.753499706 | Bright | 0.399243287 | Aright | 0.411233416 |
| | Symmetryup | 0.306696337 | left_color_iris_B | 0.692576374 | mouth_rel_inner_eyes | 0.329502466 | Bleft | 0.407136853 |
| **Low LLR** | Aleft | -0.26339648 | Symmetrylow | -0.12539211 | Cright | -0.45520606 | Nose_Color_R | -0.04979515 |
| | Aup | -0.30474284 | Alow | -0.25365471 | Aleft | -0.61675216 | Dright | -0.07630265 |
| | eyebrow_symmetry | -0.58620728 | eyebrow_symmetry | -0.45903589 | Aright | -1.25197852 | eyebrow$_{symmetry}$ | -0.23745861 |

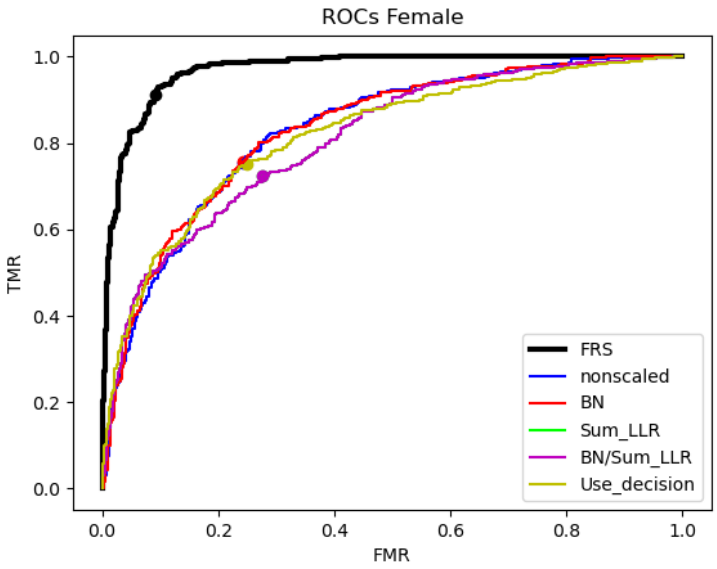



Fig. 28: ROC-curves of different 'DIFF' feature fusion strategies without the score of the FRS on female subjects. The black ROC-curve shows the result of using only the FRS. The dots show the points of the EER.

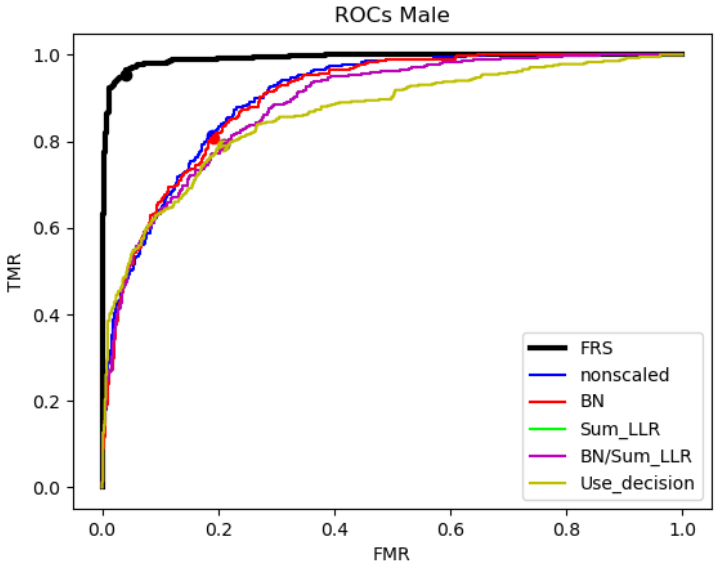



Fig. 29: ROC-curves of different 'DIFF' feature fusion strategies without the score of the FRS on male subjects. The black ROC-curve shows the result of using only the FRS. The dots show the points of the EER.

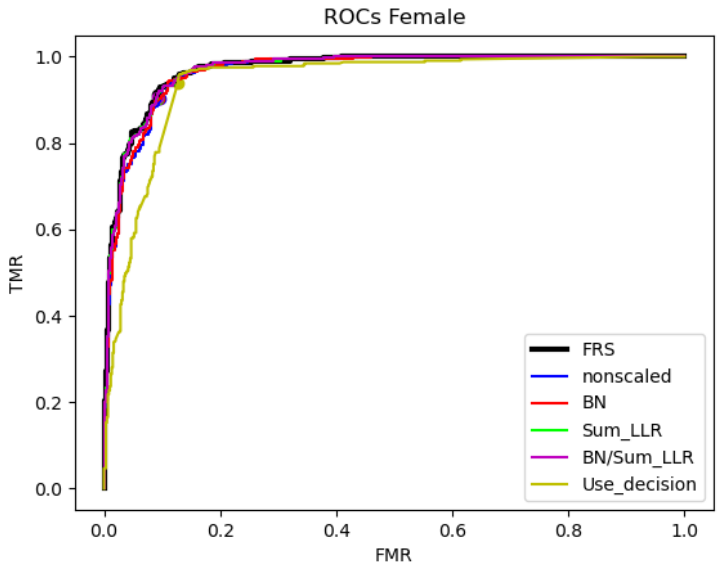



Fig. 30: ROC-curves of different 'DIFF' feature fusion strategies using the score of the FRS on female subjects. The black ROC-curve shows the result of using only the FRS. The dots show the points of the EER.

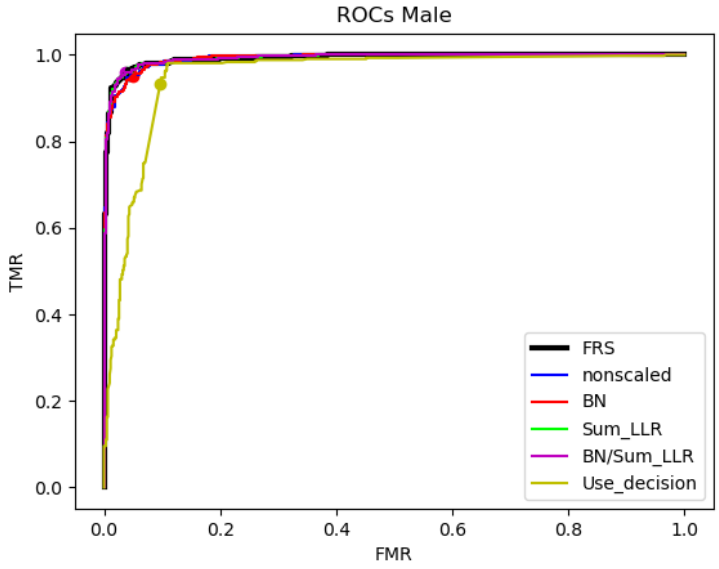



Fig. 31: ROC-curves of different 'DIFF' feature fusion strategies using the score of the FRS on male subjects. The black ROC-curve shows the result of using only the FRS. The dots show the points of the EER.

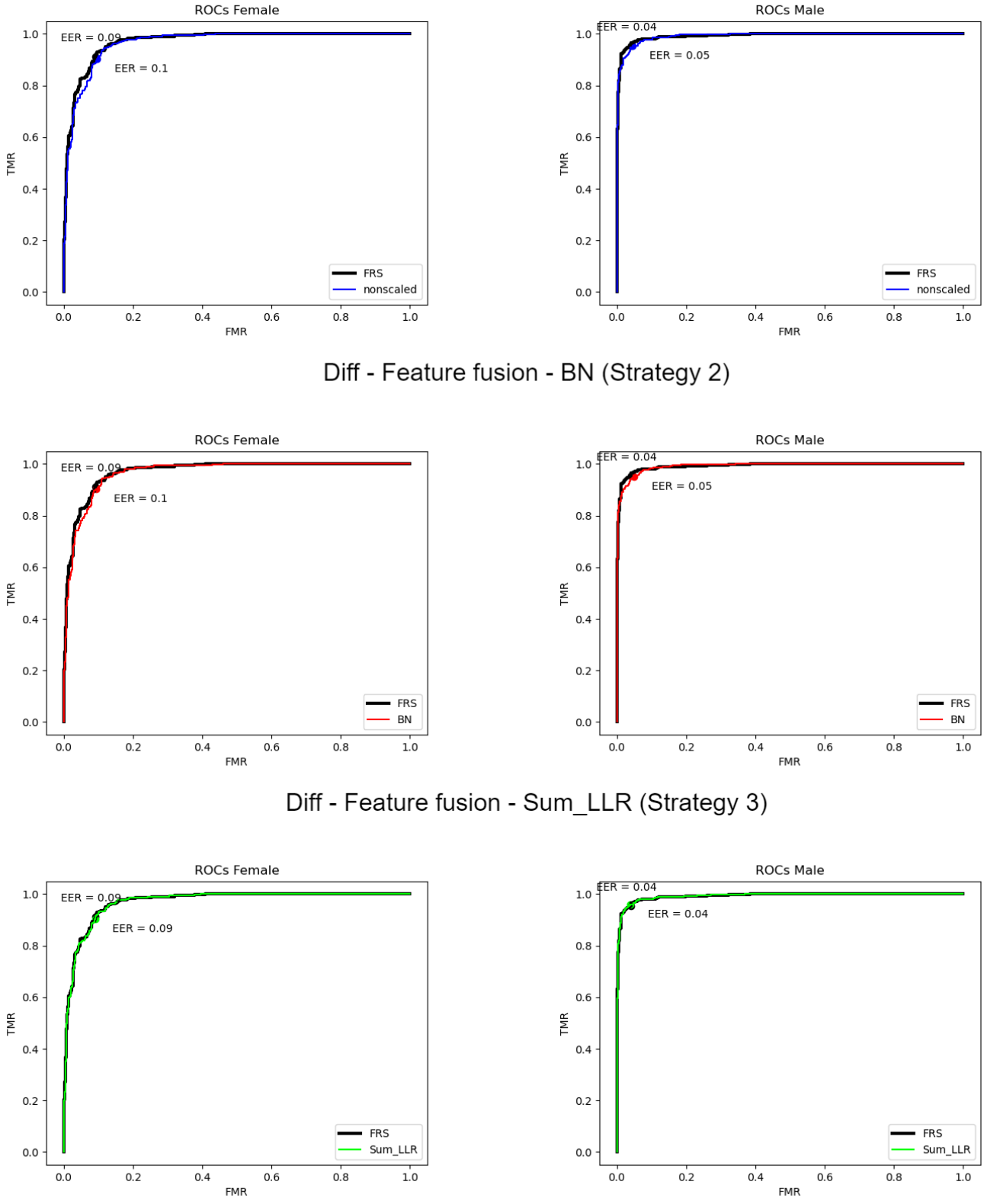



Fig. 32: ROC-curves of 'Diff' strategy 1, 2 and 3 together with the ROC-curve of the FRS.

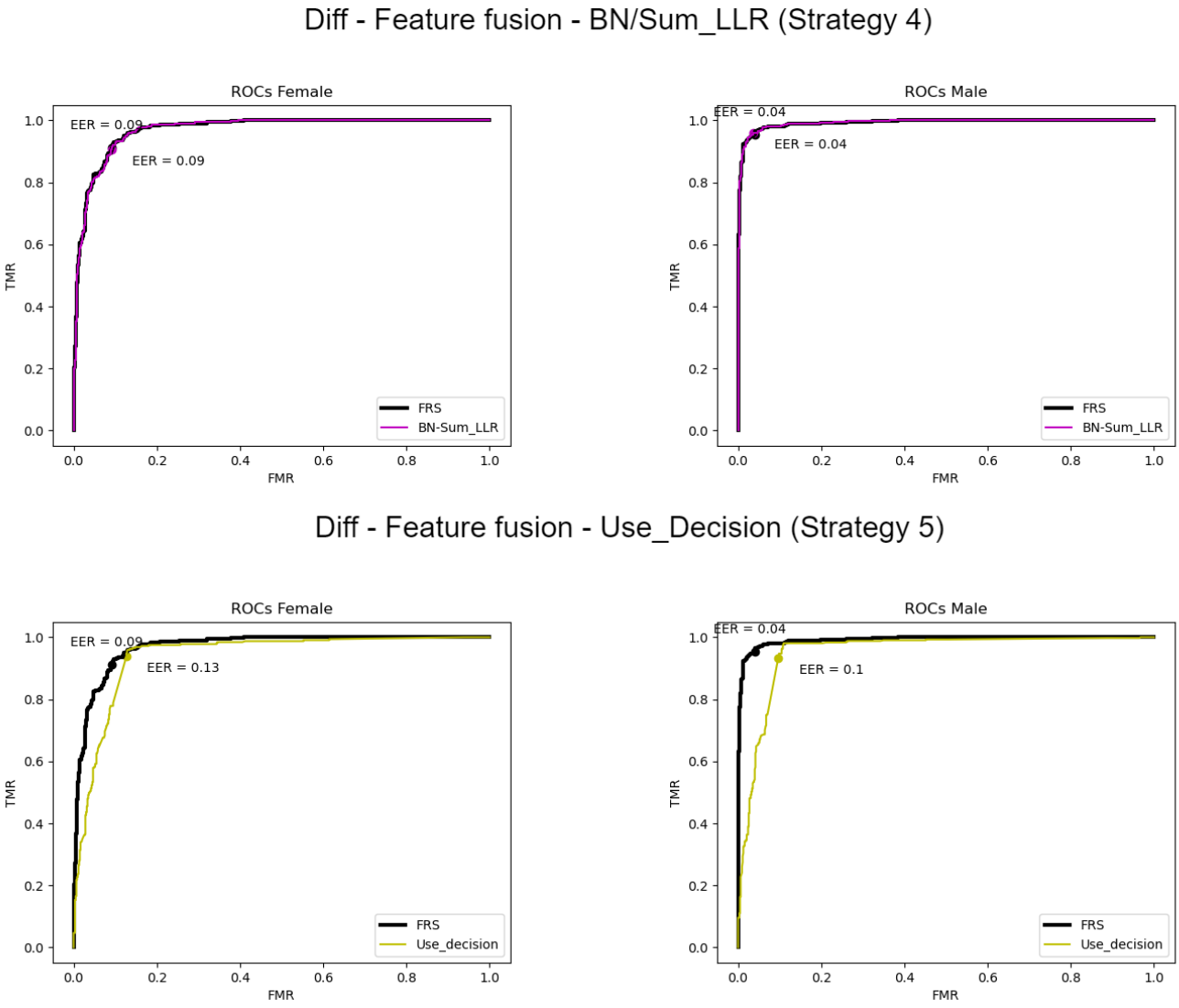



Fig. 33: ROC-curves of 'Diff' strategy 4 and 5 together with the ROC-curve of the FRS.

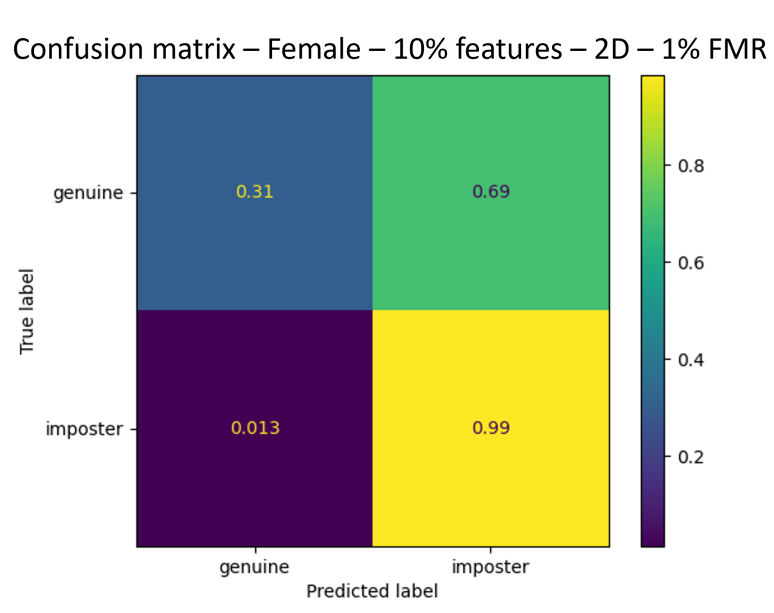



Fig. 34: Confusion matrix of the 2D voting strategy on female subjects. The thresholds are set at 1% FMR and 10% of the forensic features have to be in favour of a match for a mated vote from the voting system.

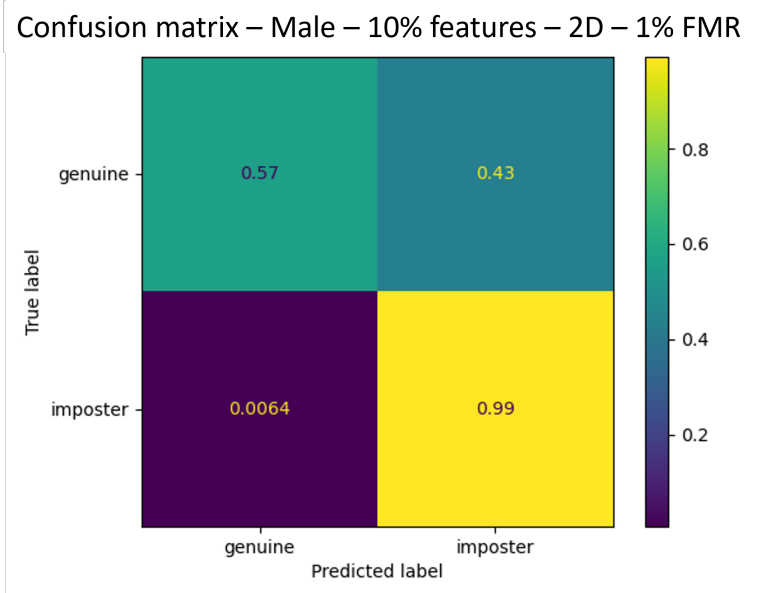



Fig. 35: Confusion matrix of the 2D voting strategy on male subjects. The thresholds are set at 1% FMR and 10% of the forensic features have to be in favour of a match for a mated vote from the voting system.

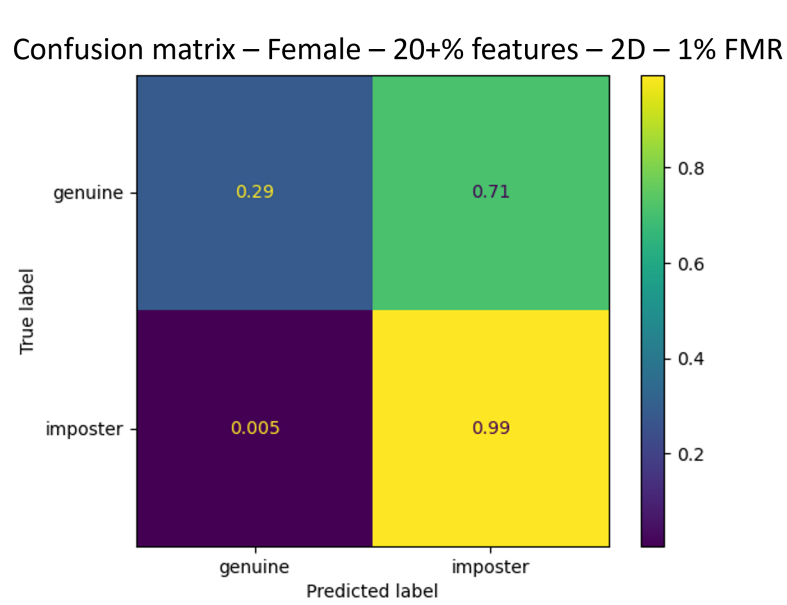



Fig. 36: Confusion matrix of the 2D voting strategy on female subjects. The thresholds are set at 1% FMR and 20-90% of the forensic features have to be in favour of a match for a mated vote from the voting system.

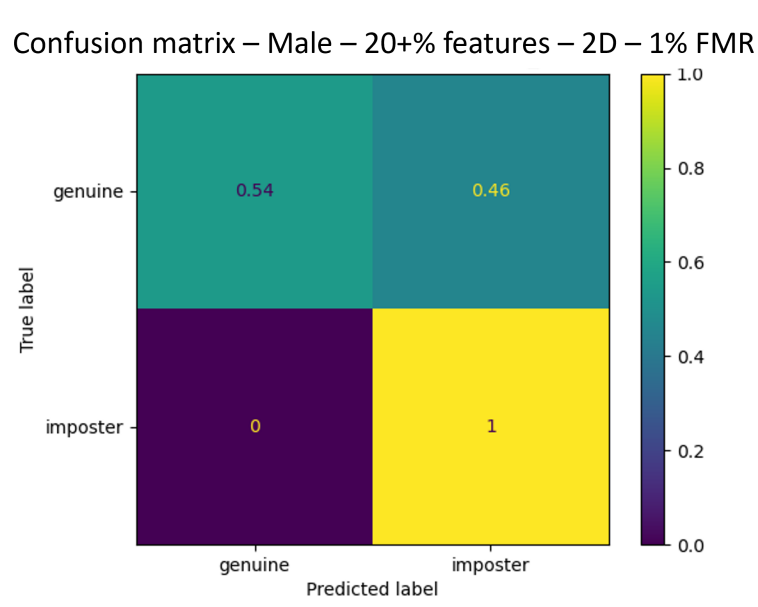



Fig. 37: Confusion matrix of the 2D voting strategy on male subjects. The thresholds are set at 1% FMR and 20-90% of the forensic features have to be in favour of a match for a mated vote from the voting system.

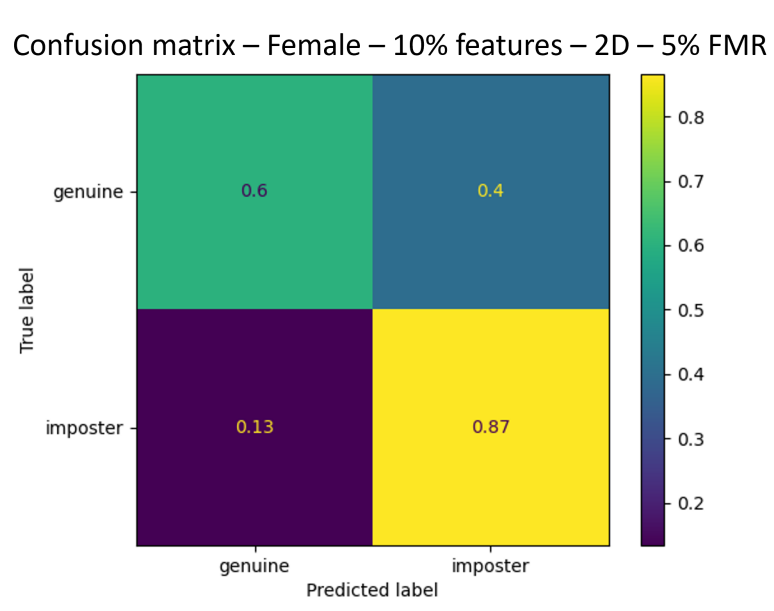



Fig. 38: Confusion matrix of the 2D voting strategy on female subjects. The thresholds are set at 5% FMR and 10% of the forensic features have to be in favour of a match for a mated vote from the voting system.

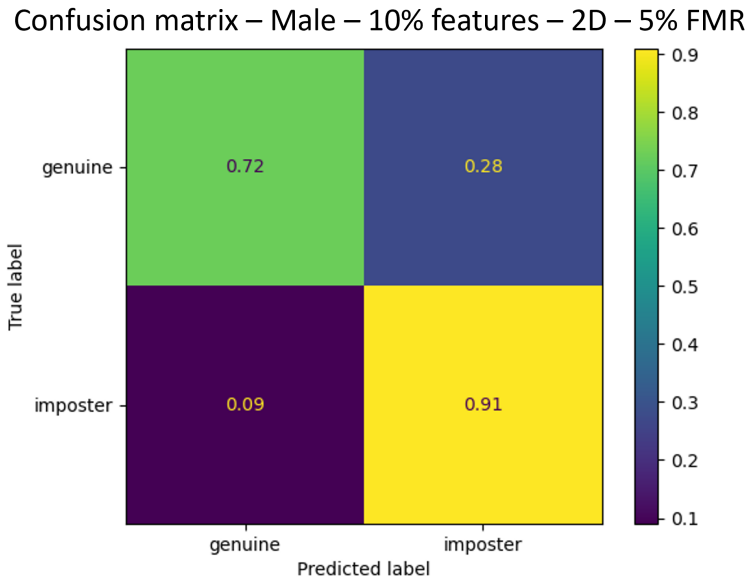



Fig. 39: Confusion matrix of the 2D voting strategy on male subjects. The thresholds are set at 5% FMR and 10% of the forensic features have to be in favour of a match for a mated vote from the voting system.

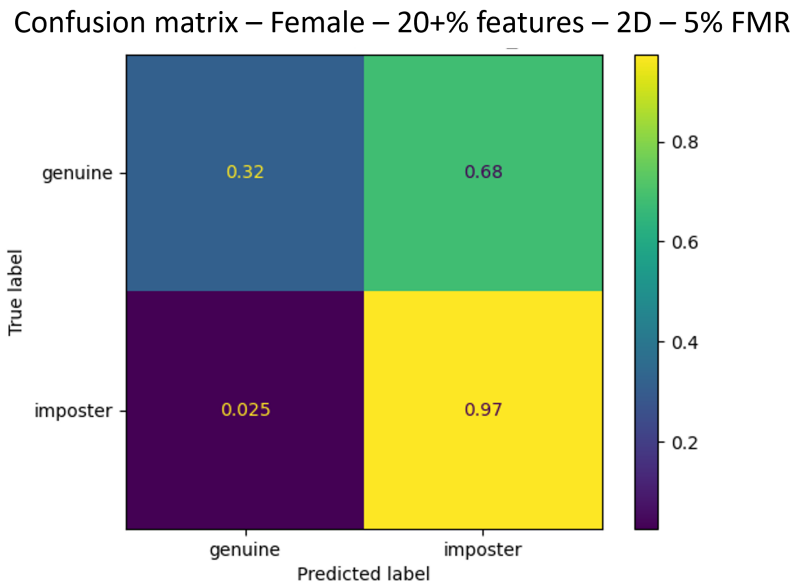



Fig. 40: Confusion matrix of the 2D voting strategy on female subjects. The thresholds are set at 5% FMR and 20-90% of the forensic features have to be in favour of a match for a mated vote from the voting system.

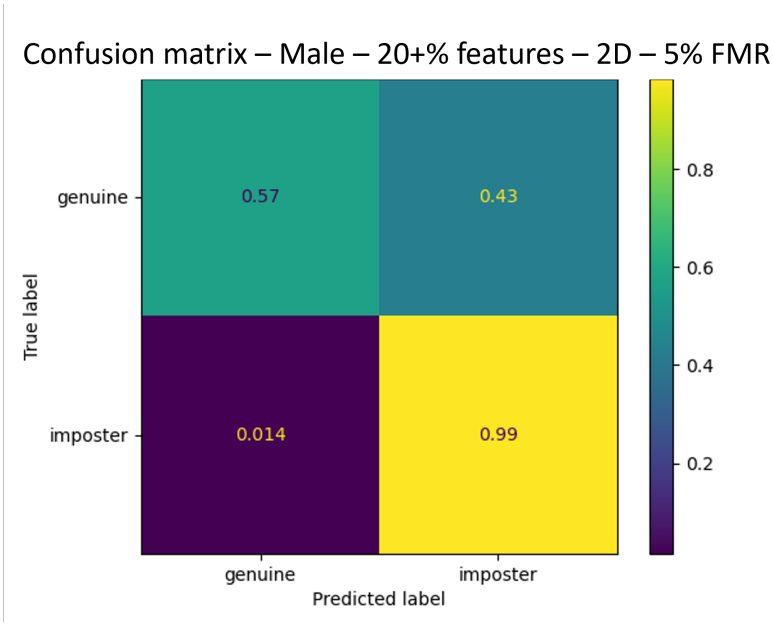



Fig. 41: Confusion matrix of the 2D voting strategy on male subjects. The thresholds are set at 5% FMR and 20-90% of the forensic features have to be in favour of a match for a mated vote from the voting system.

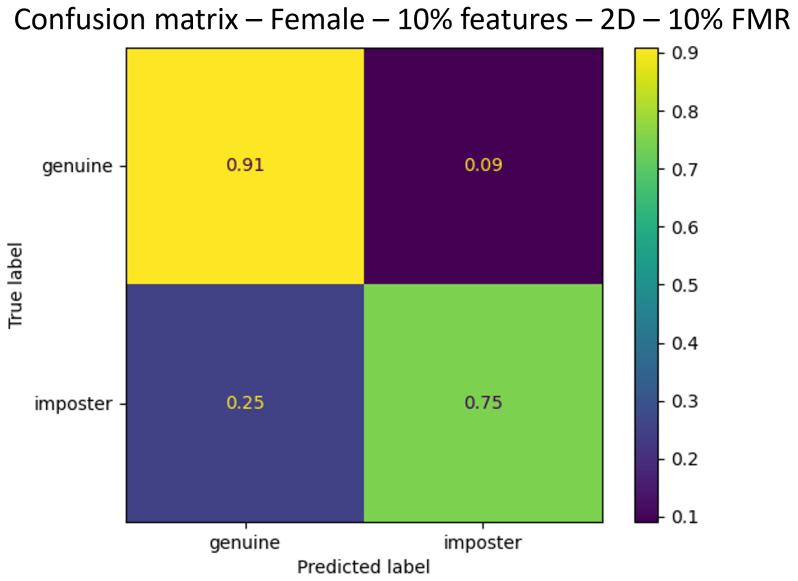



Fig. 42: Confusion matrix of the 2D voting strategy on female subjects. The thresholds are set at 10% FMR and 10% of the forensic features have to be in favour of a match for a mated vote from the voting system.

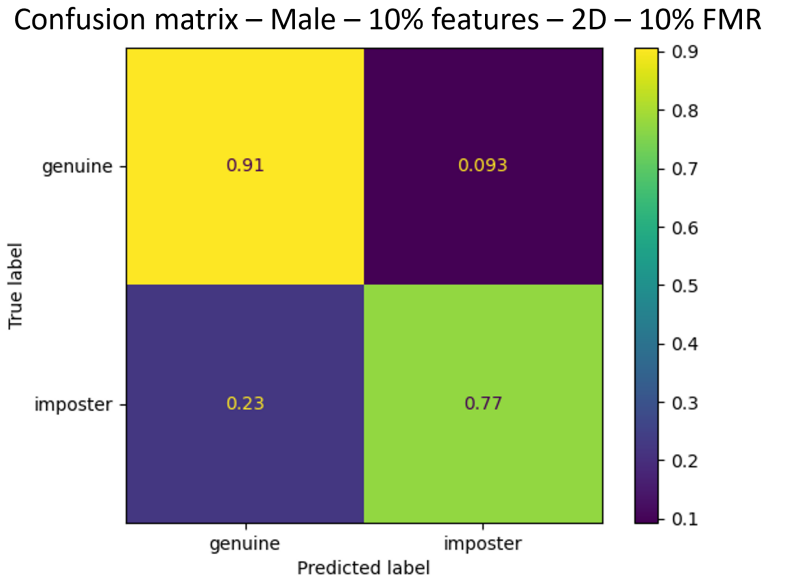



Fig. 43: Confusion matrix of the 2D voting strategy on male subjects. The thresholds are set at 10% FMR and 10% of the forensic features have to be in favour of a match for a mated vote from the voting system.

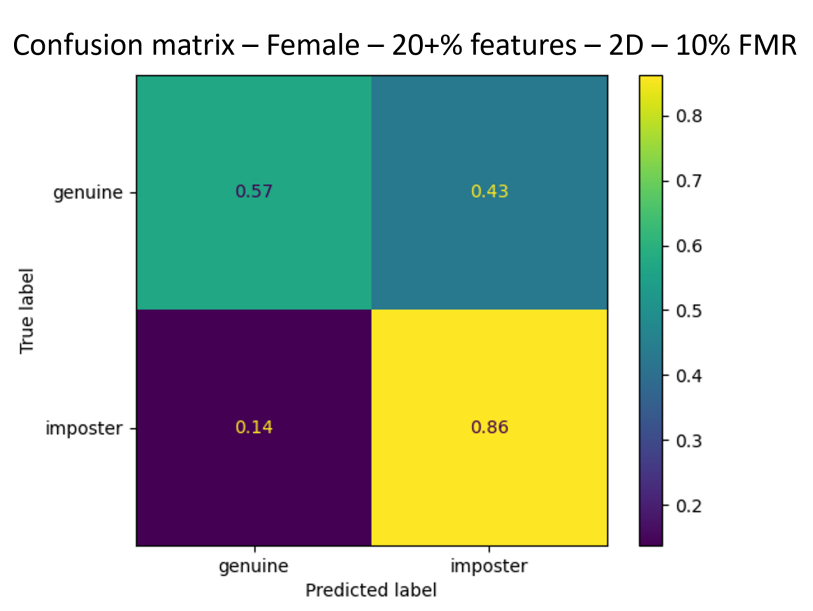



Fig. 44: Confusion matrix of the 2D voting strategy on female subjects. The thresholds are set at 10% FMR and 20-90% of the forensic features have to be in favour of a match for a mated vote from the voting system.

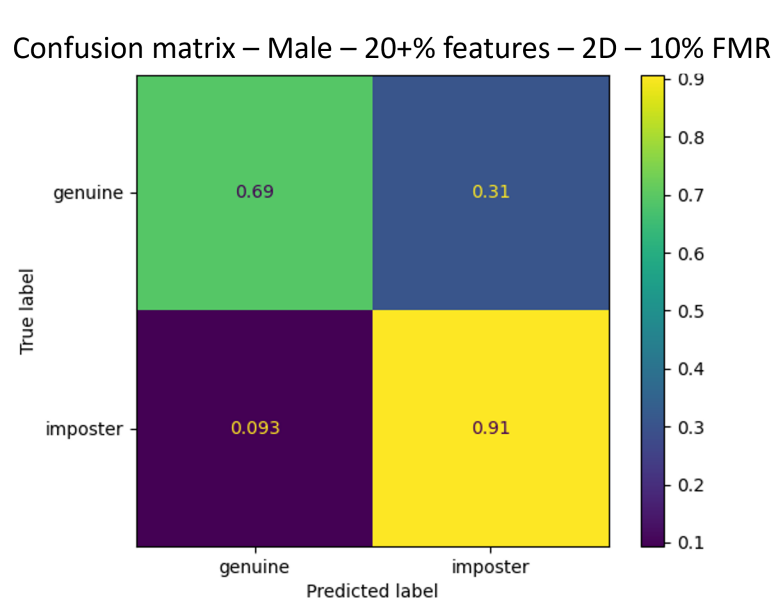



Fig. 45: Confusion matrix of the 2D voting strategy on male subjects. The thresholds are set at 10% FMR and 20-90% of the forensic features have to be in favour of a match for a mated vote from the voting system.

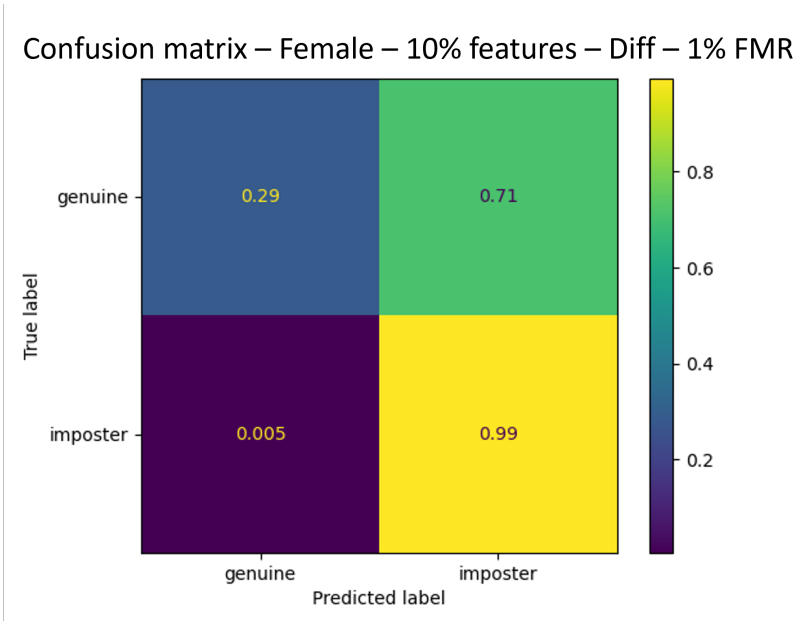



Fig. 46: Confusion matrix of the Diff voting strategy on female subjects. The thresholds are set at 1% FMR and 10% of the forensic features have to be in favour of a match for a mated vote from the voting system.

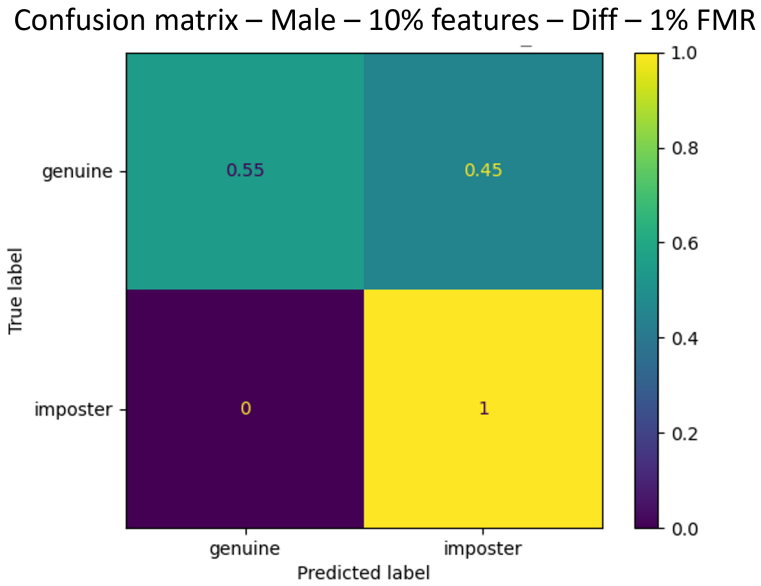



Fig. 47: Confusion matrix of the Diff voting strategy on male subjects. The thresholds are set at 1% FMR and 10% of the forensic features have to be in favour of a match for a mated vote from the voting system.

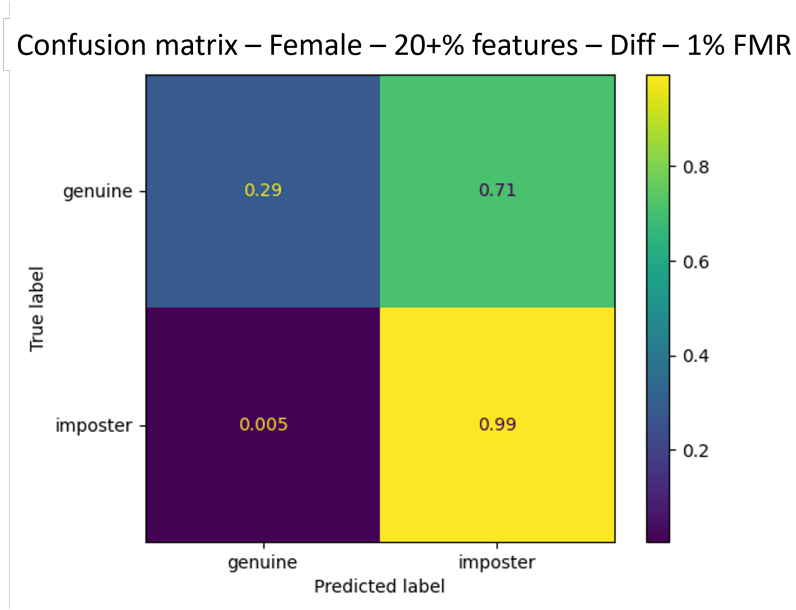



Fig. 48: Confusion matrix of the Diff voting strategy on female subjects. The thresholds are set at 1% FMR and 20-90% of the forensic features have to be in favour of a match for a mated vote from the voting system.

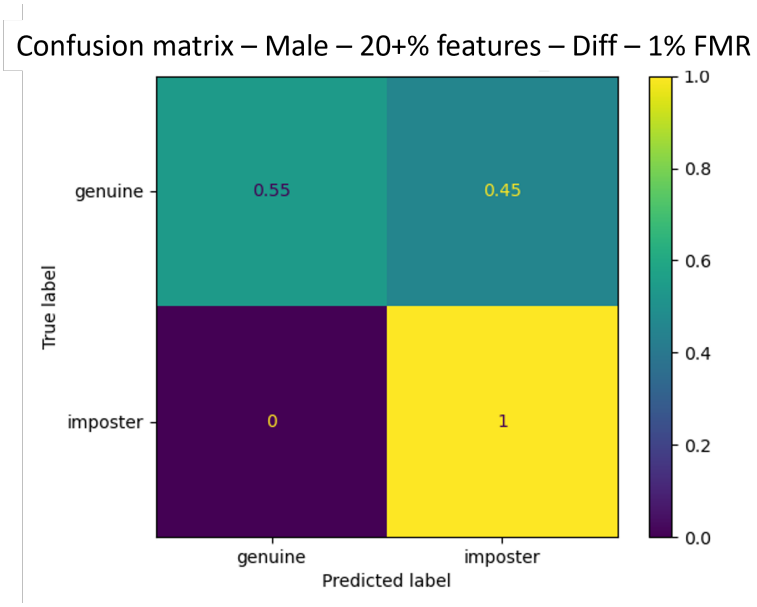



Fig. 49: Confusion matrix of the Diff voting strategy on male subjects. The thresholds are set at 1% FMR and 20-90% of the forensic features have to be in favour of a match for a mated vote from the voting system.

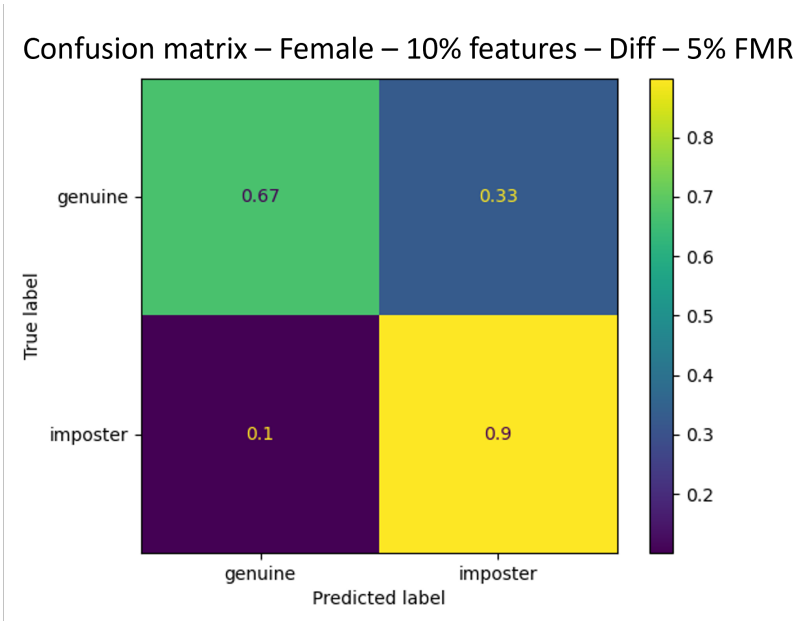



Fig. 50: Confusion matrix of the Diff voting strategy on female subjects. The thresholds are set at 5% FMR and 10% of the forensic features have to be in favour of a match for a mated vote from the voting system.

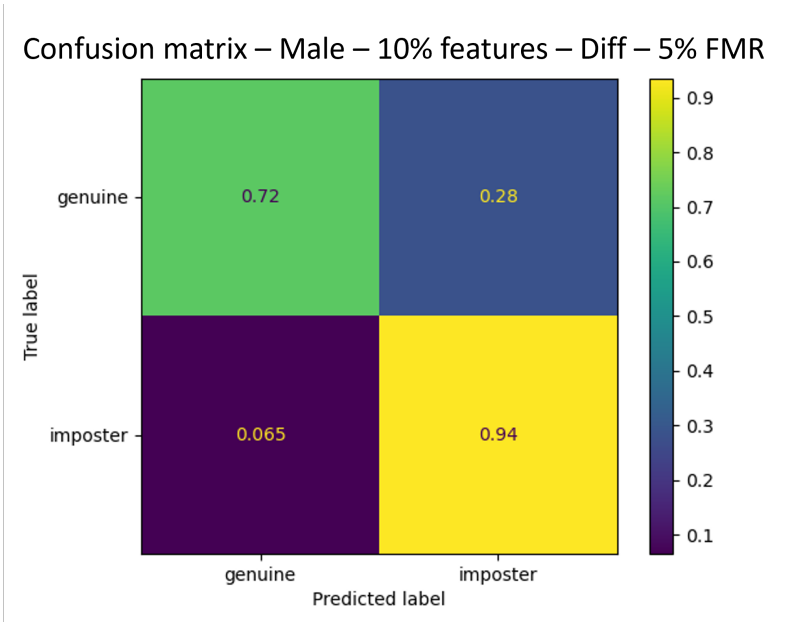



Fig. 51: Confusion matrix of the Diff voting strategy on female subjects. The thresholds are set at 5% FMR and 10% of the forensic features have to be in favour of a match for a mated vote from the voting system.

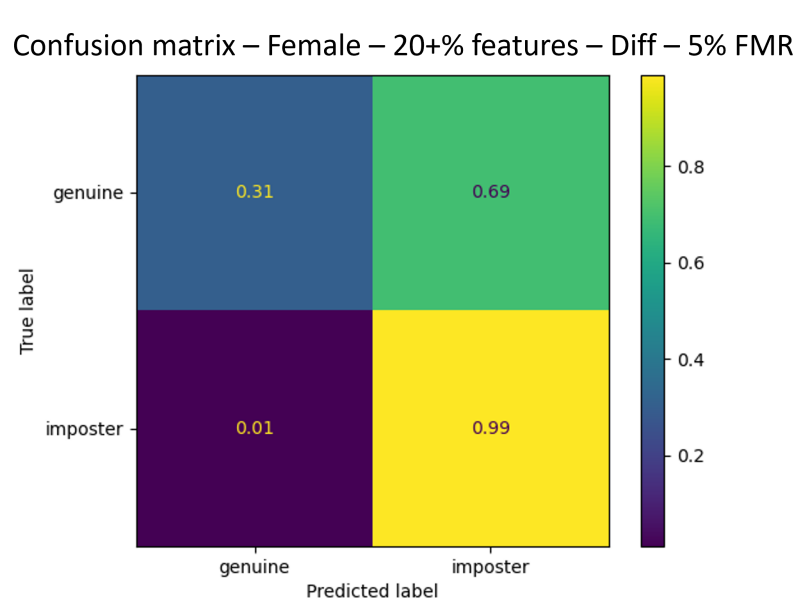



Fig. 52: Confusion matrix of the Diff voting strategy on female subjects. The thresholds are set at 5% FMR and 20-90% of the forensic features have to be in favour of a match for a mated vote from the voting system.

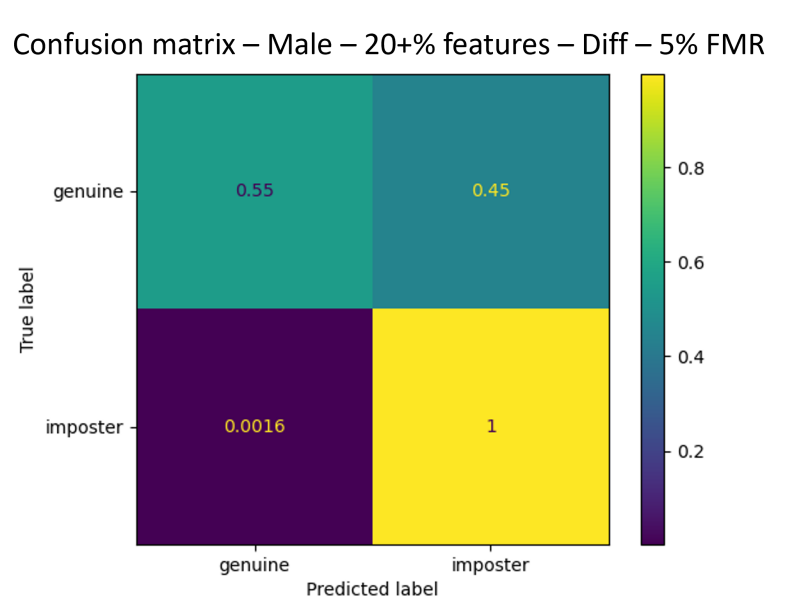



Fig. 53: Confusion matrix of the Diff voting strategy on female subjects. The thresholds are set at 5% FMR and 20-90% of the forensic features have to be in favour of a match for a mated vote from the voting system.