

XAIface

Measuring and Improving Explainability for AI-based Face Recognition

Explainability Protocol and Methods (v2)

Deliverable number: D4.1

Version: 2.65

Acronym of the project: XAIface

Title of the project: Measuring and Improving Explainability for AI-based Face Recognition

Grant: CHIST-ERA-19-XAI-011

Web site of the project: <https://xaiface.eurecom.fr/>

Deliverable

Editors:	Chiara GALDI and Nelida MIRABET HERRANZ (EURECOM)
Deliverable nature:	Report (R)
Dissemination level: (confidentiality)	Public (PU)
Contractual delivery date:	30th April 2023
Actual delivery date:	30th April 2023
Keywords:	face recognition, explainability, assessment
Author(s): (names and affiliations)	Chiara GALDI, EURECOM; Nelida MIRABET HERRANZ, EURECOM; Martin WINTER, JRS; Yuhang LU, EPFL; Naima BOUSNINA, IT; Fernando PEREIRA, IT; Paulo LOBATO CORREIA, IT; João ASCENSO, IT

Short abstract

Deliverable D4.1 – “Explainability protocol and methods” – reports the consortium work towards the development of explainable techniques for face recognition during the project. The deliverable has been published in two versions, one in month 12 and one in month 24.

The first version (v1) already summarised relevant existing explainability techniques for AI face recognition solutions addressing both the cases of creating transparent models and post-hoc explanation of black-box models, described some plans and evaluation performance protocols, and provided a first analysis of ethical and legal issues.

This is the second version of the deliverable which, in addition to v1 provides more detailed descriptions of the methods initially described in version 1, collects descriptions of additional methods developed and includes performance results obtained during experiments with their first implementations. The chapter addressing ethical and legal issues is extended and provides a more in depth look at various examinations of "interpretability"-requirements according to the AI-Act proposal.

Table of content

Abbreviations	7
Definitions	8
1. Introduction	9
1.1. Motivation	9
2. Related Literature Works for AI Explainability with Potential Application for Face Recognition	11
2.1. General Observations	11
2.2. Shapley-values (SHAP)	12
2.2.1. BreakDown	12
2.3. Local Interpretable Model-agnostic Explanations (LIME)	12
2.4. Saliency Maps for CNNs	13
2.5. Class Activation Maps (CAM), Grad-CAM, and Grad-CAM++	15
2.6. Layer-wise Relevance Propagation (LRP)	15
2.7. Explainable Boosting Machines	16
2.8. Randomised Input Sampling for Explanation (RISE)	17
2.9. Spatial and Feature Activation Diversity Losses for Structured Face Representations	18
3. XAIface Contributions to AI-based Face Recognition Explainability	20
3.1. Explainable Face Recognition by Interpretable, Local Features (LIBF)	20
3.1.1. Introduction	20
3.1.2. Description of the LIBF method	21
Heuristic selection of meaningful patch locations	23
Self-supervised learning of feasible patch embedding	25
3.1.3. Performance Evaluations	26
Evaluation of verification performance using the learned self-supervised embeddings	26
Explanations gained using EBM in the verification scenario	30
3.1.4. Legal and Ethical guidelines concerned by the method	36
3.1.5. Summary, conclusion and outlook	36
3.2. Improving Face Verification Explainability using Spatially-Biased Similarity Metrics and Training Loss Functions	37
3.2.1. Introduction	37
3.2.2. Description	37
3.2.3. Performance evaluation	38
3.3. Similarity-based RISE Algorithm for Explainable Face Recognition System	38
3.3.1. Introduction	38
3.3.2. Description	39
3.3.3. Visual results of generated saliency maps	40
3.3.4. Performance evaluation	42
3.3.5. Legal and ethical guidelines concerned by the S-RISE method	45
3.4. Demographic Information Disentangling	45
3.4.1. Introduction	45

3.4.2. Description	46
3.4.3. Performance Evaluation	47
3.4.4. Legal and Ethical Guidelines Concerned by the Method	47
3.5. Explainable Soft Biometrics Estimation	50
3.5.1. Introduction	50
3.5.2. Description	50
3.5.3. Performance evaluation	51
3.5.4. Legal and ethical guidelines concerned by the method	53
3.6 Face Verification Explainability using Vision Transformer and EBM	54
3.6.1. Introduction	54
3.6.2. Description	54
3.6.3. Performance evaluation	56
4. Addressing Ethical and Legal issues	58
4.1. Explainable Approach Design and Implementation	58
4.2. Scope	58
4.3. Legal Requirements	60
4.3.1. Explainability & transparency	60
4.3.2. Current obligations according to the GDPR	61
General remarks	61
Modalities of transparency	62
Visualisation	63
Information	63
Right to access and specific provisions on automated decision-making	64
Exceptions and opening clauses	66
4.3.3. Future obligations according to the AI-Act	67
General Remarks	67
High-Risk Systems	68
Requirements according to Chapter 2	68
4.4. Compliance Efforts	70
4.4.1. AI-ACT	70
4.4.2. GDPR	71
4.5. Ethical Requirements	73
4.5.1. HLEG-guidelines	73
Transparency & Explainability	73
4.6. Preliminary Conclusion	74
5. Summary	75
References	76

Abbreviations

AI	Artificial Intelligence
CAM	Class Activation Maps
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Networks
DET	Detection Error Tradeoff
DL	Deep Learning
EBM	Explainable Boosting Machine
FAD	Feature Activation Diversity
FMR	False Match Rate
FNMR	False Non-match Rate
FR	Face Recognition
GDPR	General Data Protection Regulation
JPEG	Joint Photographic Experts Group
LIME	Local Interpretable Model-Agnostic Explanations
LMF	Large Magnitude Filtering
LRP	Layer-wise Relevance Propagation
RISE	Randomised Input Sampling for Explanation
RNN	Recurrent Neural Network
SAD	Spatial Activation Diversity
SHAP	SHapley Additive exPlanations
XAI	Explainable Artificial Intelligence

Definitions

ArcFace is a CNN based model for face recognition which learns discriminative features of faces and produces embeddings for input face images. To enhance the discriminative power of softmax loss, a novel supervisor signal called additive angular margin (ArcFace) is used as an additive term in softmax loss.

Automation bias: Over-reliance on automated aids and decision support systems. The same concept can be translated to the fundamental way that AI and automation work, which is mainly based on learning from large sets of data. This type of computation assumes that things won't be radically different in the future. Another aspect that should be considered is the risk of using flawed training data then the learning will be flawed.

Bagging is an ensemble learning technique that helps to improve the performance and accuracy of machine learning algorithms. It is used to deal with bias-variance trade-offs and reduces the variance of a prediction model.

Demographic information: Socio-economic information from a subject such as population, race, income, education and employment.

Explainable AI (XAI) is artificial intelligence that is programmed to describe and understand its purpose, rationale and decision-making process in a way that can be understood by the average (human) person.

Greedy algorithms are any algorithm that follows the problem-solving heuristic of making the locally optimal choice at each stage.

Kernels are filters used to extract the features from the images. More specifically, a kernel is a matrix that moves over the input data, performs the dot product with the sub-region of input data, and gets the output as the matrix of dot products.

MagFace is a category of losses that learns a universal feature embedding whose magnitude can measure the quality of the given face. Under the new loss, the magnitude of the feature embedding monotonically increases if the subject is more likely to be recognized. In addition, MagFace introduces an adaptive mechanism to learn well structured within-class feature distributions by pulling easy samples to class centres while pushing hard samples away. This prevents models from overfitting on noisy low-quality samples and improves face recognition in the wild.

Pooling layers: Used to reduce the dimensions of the feature maps. It reduces the number of parameters to learn and the amount of computation performed in the network. The pooling layer summarises the features present in a region of the feature map generated by a convolution layer.

Post-hoc: Analysis of the results of experimental data.

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Soft biometrics are attributes that are not necessarily unique to an individual, but can be used alone or in conjunction with primary biometric traits for improving performance or explainability in a variety of applications.

1. Introduction

The objective of deliverable D4.1 “Explainability protocol and methods” is to plan, report and track the consortium work towards the development of explainable techniques for face recognition during the project.

This deliverable is published in two versions, one in month 12 and one in month 24 of the XAIface project. The second version, in addition to collecting the description of the methods developed, also includes a more detailed description of the performance assessment of the proposed methods and their links to legal and ethical aspects and issues that they may raise.

The document contains the following sections:

- Section 1: Introduces the general motivation for the document and introduction.
- Section 2: Provides an overview of the state of the art of AI explainability methods, focused on face recognition domain. The reported methods should give a necessary background for the development of new methods. Thus, the methods described are already used - or seem at least promising (potentially applicable) - for Explainable Artificial Intelligence (XAI) face recognition.
- Section 3: Describes the development of new methods and reports on their progress and performance assessment. This section also provides a link between the developed methods and the eventual legal and ethical issues described in Section 4 of this document.
- Section 4: Provides an assessment of potential legal and ethical issues for the face-recognition systems and questions regarding the right to obtain explanations are elaborated and discussed.

1.1. Motivation

Face recognition has become a key technology in our society, frequently used in multiple applications, while creating an impact in terms of privacy. As face recognition solutions based on artificial intelligence (AI) are becoming more and more popular, it is critical to fully understand and explain how these technologies work in order to make them more effective and accepted by society. Thus in XAIface, we focus on the analysis of the influencing factors relevant for the final decision of an AI-based face recognition system as an essential step to understand and improve the underlying processes involved.

Most state-of-the-art work regarding XAI is nowadays working either on the development of transparent machine learning models, or on the application of post-hoc explainability models. While transparency is most often an inherent property of simpler, classical machine-learning models, state-of-the-art (more complex) face recognition methods require “post-hoc” or “post-modelling” explainability techniques like simplification, feature relevance, or class activation maps. Of major interest are also model-agnostic techniques, which can be plugged into any model to extract information from the (black-boxed) prediction process.

Another related aspect is attention, as this will also facilitate the understanding which features of a face contributed to the decision made. Attention is a mechanism allowing the model to focus on certain parts of the input sequence when predicting a certain part of the output sequence, enabling easier learning and of higher quality. Attention can be introduced into different network architectures, such as multi-task networks or recurrent neural networks (RNNs). In face recognition, it has for example been proposed for weighting features in face verification and identification. Also, attention has been used to select face local patches without resorting to face landmarks. A related application of attention is modifying face attributes in generative approaches, which could be a component in adversarial learning. Some of these methods are also reviewed in Section 2.

Another main field this project will contribute to knowledge beyond state of the art in XAI will be the progress in optimising the tradeoff between interpretability and performance. This will be addressed by the development of adjustable explainability components of face recognition pipelines, which may affect some or all components of the pipeline. The plans for this work as well as results obtained so far are outlined and reported in Section 3.

2. Related Literature Works for AI Explainability with Potential Application for Face Recognition

In this section, a brief overview of state-of-the-art methods related to the planned XAIface contributions to AI face recognition explainability is provided. This review will serve as background to understand the developed techniques in XAIface as well as highlighting the contributions of XAIface to the state of the art.

2.1. General Observations

Model explainability, one of the most important problems in machine learning today, refers to the concept of being able to understand a deep learning network. It aims for a better comprehension of why machine learning models make certain decisions and which factors are most relevant for it.

Being able to interpret a machine learning model leads to several benefits because understanding the decision process of a network increases the user's trust in the prediction, helps to debug a model during its development phase and can be relevant in determining whether or not a model is suitable for a real world application. Model interpretability, when successfully performed, can also help and guide in assessing the presence of bias.

In the existing literature, many classifications of explainability models have been used. In this document we distinguish according to the following (most common) XAI grouping:

- **Local vs. global methods:** Local methods investigate a specific set of samples and try to explain the model behaviour on those specific samples. In opposition, global methods give an overall explanation of model behaviour by studying how features collectively affect the final prediction.
- **Model specific vs. model agnostic:** Model specific techniques work under certain, clearly defined model structures, while model agnostic tools work in a more general manner by analysing directly the input-output pair without specific knowledge on the model.
- **White-box vs black-box models:** A white-box model is a technique designed to be explainable per se, thus it does not require any other additional explainability technique on top or besides of it to interpret the decision process. A black-box model is not explainable by itself. Instead of that it requires the adaptation of other techniques in its pipeline in order to make it explainable. The main advantage of black-box models is the fact that one can take an already trained model and train an explanation model on top of it. As a general rule, the explainer is much easier to understand.
- **Gradient-based backpropagation methods vs perturbation-based forward propagation methods:** Gradient-based methods investigate the signal flow through the classifier, investigating how changes in the weights and parameters in the model will influence the output. On the other hand, perturbation-based models change only the input of the model and investigate how the output changes.

2.2. Shapley-values (SHAP)

SHAP (Lundberg and Lee 2017), SHapley Additive exPlanations, is a framework for method interpretability that gives *data features* and *importance values* for specific predictions. SHAP values are based on Shapley-values, which refer in game theory to the average of all the marginal contributions to all possible coalitions. This average is used as a baseline and SHAP shows the impact of each feature by interpreting and measuring the relative impact of a certain value compared to that baseline. In this way, SHAP values allow us to determine any prediction as a sum of the effects or amount of influence of each feature value to the final decision.

Interpreting a model using SHAP has the advantage of providing both local and global model interpretability simultaneously. The contribution of each predictor to the final prediction, positively or negatively, can be measured by looking at the collective SHAP values thus ensuring global interpretability. Furthermore, each individual sample obtains its own set of SHAP values allowing us to assess and compare the relevance and impact on the final prediction of all individual factors. It is also important to mention that the calculation of the SHAP values is computationally expensive. In fact it is exponential in the number of features and this can be especially critical for highly dimensional (face-) features.

Summing up SHAP is a model agnostic technique thus it focuses on analysing the relation between the inputs and the predictions for black-box approaches in a perturbation-based forward propagation manner. Its adaptability to any type of deep learning structure makes it especially interesting for the project since SHAP can help to understand the most significant regions for face recognition and to check if those areas are consistent among different face recognition systems.

2.2.1. BreakDown

BreakDown (Staniak and Biecek 2018) is similar to SHAP. It is also based on the conditional response of a black-box model and it then attributes the response to the input features. The only difference is that breakDown deals with conditionings in a greedy way instead of averaging.

A GitHub implementation of the method can be found at:



<https://github.com/MI2DataLab/pyBreakDown> - python



<https://github.com/pbiecek/breakDown> - R

2.3. Local Interpretable Model-agnostic Explanations (LIME)

LIME (Ribeiro et al. 2016) (Local Interpretable Model-Agnostic Explanations) is one of the most important interpretability techniques developed by University Of Washington researchers to study the decision process inside an algorithm by capturing feature interactions. It is a model agnostic technique that analyses the relation between the input data and the prediction in a perturbation-based forward propagation manner, therefore it can

be used on any model acting as a surrogate model. LIME modifies a test data instance by altering its input values (slightly) and observes the impact on the output. LIME tries to model the local neighbourhood of any prediction by focusing on a small decision surface around the input point. The basic assumption is that even very simple (linear) models are a good approximation of the black-box model under investigation. Those more simple models are often linear/logistic regressions or decision trees. More specifically, LIME defines distance measures in order to compute the distance between the original sample and the altered one. Then, new predictions are computed by passing the altered samples through the black-box model. After that, LIME picks a number “ n ” of features by maximising the maximum likelihood, in order to select the “ n ” features that better describe the deep learning network. Those features are used to fit a simple model to the permuted data with the similarity scores as weights. Finally, by looking into this simple (surrogate) model, it is possible to get insights into the initial black-box model.

The main advantage of LIME is its ability to maintain local fidelity even when the number of dimensions is high. LIME ensures this property by finding a model that approximates the black-box model only locally. It is also able to handle irregular inputs. On the other hand, there is the danger that a bad or insufficient approximation might lead to misleading explanations. In addition, explanations are sometimes unstable and depend on the underlying process of input perturbation. LIME has been considered of interest for XAIface since it can provide quantitatively visual information on the most relevant image regions for face recognition tasks for any type of deep learning structure.

2.4. Saliency Maps for CNNs

Saliency maps are a special visualisation technique providing insight into the decision making process of a neural network. Several methods have been proposed in literature and in principle they identify and highlight (e.g. via heat maps) the regions a network focuses on when providing decisions. Several variants of these techniques exist, for example, Class Activations Maps, which will be described in more detail in section 2.5. Here we focus on one of the earliest approaches regarding saliency map visualisation, namely the “deconvolutional network approach” from (Zeiler and Fergus 2014) which we will describe in more detail in the following.

As the name might presume, this approach uses an inverted convolutional network (deconvolution network) which maps all activations of the individual feature maps (in all layers of the neural network, including the intermediate ones) to the corresponding input patterns in the input pixel space.

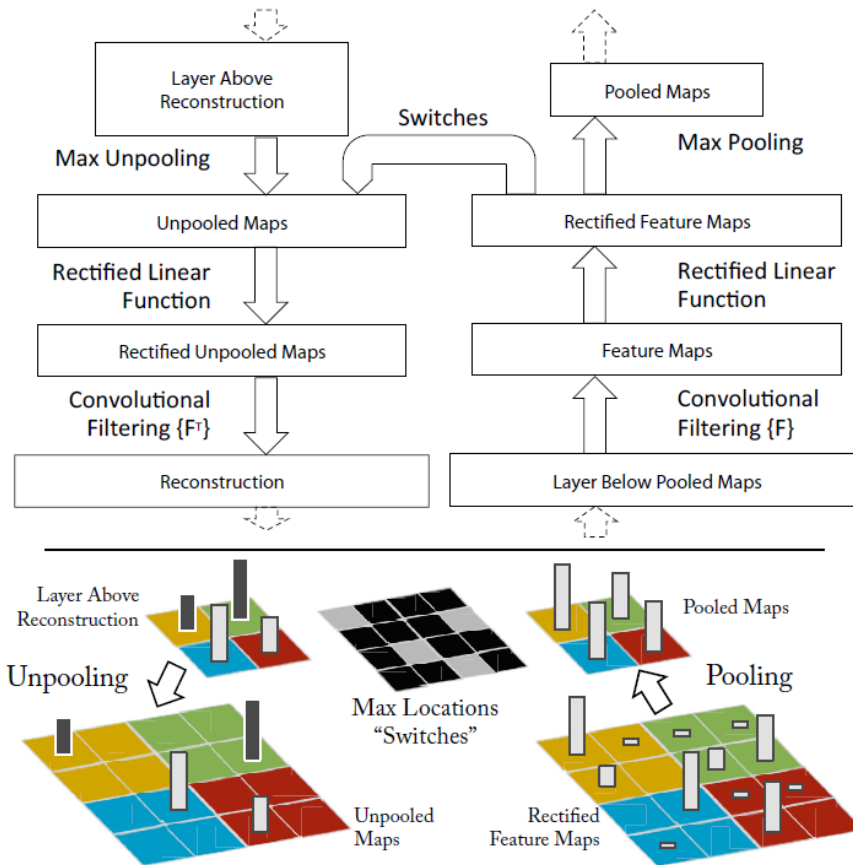



Figure 2.4a: Overview about the saliency map visualisation principle. The network on the left side reconstructs an approximate version of the traditional convolutional neural network illustrated on the right side. The unpooling operation in the bottom of the figure uses so-called “switches” to record the locations of the local max in each pooling region (coloured zones) during traditional operation.

The illustration in Figure 2.4a – taken from (Zeiler and Fergus 2014) – outlines the principle. The traditional image classification pipeline (ImageNet in this example) is shown on the right side and processed from the bottom to the top direction. The deconvolution network (deconv) is depicted on the left side and consists in principle by the same components, but in reverse order. As in the deconv-network the features are mapped to pixels, inverse operations of all convolutional network steps are required. While this is no problem for e.g. “filter” operations by using their transposed kernel versions, the situation is more difficult for non-invertible operations, e.g. max-pooling. In those cases only approximations of an inverse are feasible. In the case of the max-pooling operation for example, the intermediate “recording” (storage) of locations of the maxima within each pooling region in a set of so-called “switch variables” can be used to overcome that issue.

Regarding the categorization in Section 2.1 the “deconvolutional network approach” requires specific interventions (additional layers) within the structure of the neural network to provide a proper explanation. Thus, the approach is definitely a model specific one. In addition, it allows the explanation of each sample per se thus providing a global explanation of the entire model behaviour for a sufficient number of testing samples analysed.

2.5. Class Activation Maps (CAM), Grad-CAM, and Grad-CAM++

Class Activation Maps (Zhou et al. 2016), or CAMs, is a deep learning interpretability method used for CNNs. It is used to indicate the discriminative regions of an image used by a CNN to identify the category of the image. In detail, the CAMs method performs the following steps. First of all, it modifies the network architecture by replacing fully-connected layers in the end by a Global Average Pooling layer and concatenates the averages of the activations of convolutional feature maps that are located before the final output layer and create a feature vector. The weighted sum of the vector is fed to the final softmax loss layer. Finally, the important image regions are identified by projecting back the weight of the output layer to the convolutional feature maps. Publicly available GitHub-implementations can be found at:

 <https://github.com/zhoubolei/CAM> or <https://github.com/frgfm/torch-cam>

The main drawback of CAM is that it requires neural networks to have a specific architecture in the final layer. This is not the case for the Grad-CAM (Selvaraju et al. 2019) method, which is a generalisation of CAM that can produce visual explanations for any CNN architecture. As a gradient-based method, Grad-CAM uses the class-specific gradient information flowing into the final convolutional layer of a CNN, producing a coarse localization map of the important regions in the image. Grad-CAM++ (Chattopadhyay et al. 2018) is an extension of the Grad-CAM method that provides better visual explanations of CNN model predictions.

Proper GitHub-implementations for Grad-CAM and Grad-CAM++:

 <https://github.com/jacobgil/pytorch-grad-cam>

2.6. Layer-wise Relevance Propagation (LRP)

Layer-wise Relevance Propagation (LRP) is a framework that decomposes the prediction of a deep neural network over a sample image down to relevance scores for image regions. It helps to interpret highly complex deep neural networks by back-propagating the predictions. The LRP method supports various data types, e.g. images, text, etc, and various neural network architectures. Moreover, it has been successfully applied to explainable facial expression recognition (Arbabzadah et al. 2016), and document categorization (Arras et al. 2017).

The main idea of Layer-wise relevance propagation is to trace back the contributions of input nodes to the final network prediction. It performs backward propagation using a set of purposely-designed propagation rules from the output, identifying the most relevant neurons within the neural network until returning to the input. In detail, firstly the relevance score of the specified node in the final layer is set as output. The relevance score is then back propagated to the input layer. As a result, the prediction is decomposed into pixel-wise relevance indicating the contribution of a neuron to the final decision.

Figure 2.6a illustrates the general process and results of applying pixel-wise decomposition and LRP method in image classification tasks. A deep learning-based image classifier attempts to extract feature vector representations with deep convolutional neural networks and perform prediction with some classification layers. The LRP method decomposes the classification output into sums of feature and pixel relevance scores and consequently generates a heat map visualising the contributions of every single pixel to the prediction.

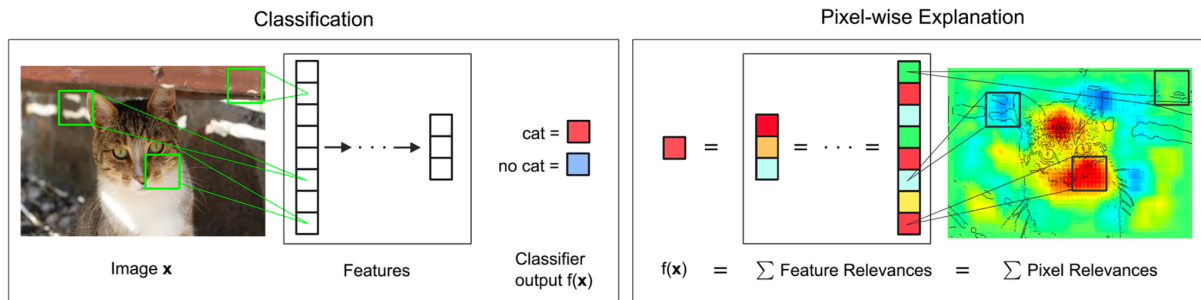


Figure 2.6a. Visualisation of the pixel-wise decomposition process in image classification tasks.

Note that a (non-official) GitHub implementation is available at:

https://github.com/sebastian-lapuschkin/lrp_toolbox.

2.7. Explainable Boosting Machines

Explainable Boosting Machine (EBM) is a glassbox model, designed to have comparable accuracy to state-of-the-art machine learning methods like Random Forest or XGBoost (Chen and Guestrin 2016) while being highly intelligible and explainable (Nori et al. 2019). They are very similar to generalised additive models (Hastie and Tibshirani 1987), but exhibit a number of important improvements besides their ability to uncover the feature contributions to the prediction of a model.

The basic idea behind of the approach is the selection of a simple additive model in the form

$$y = a_0 + a_1 \times f_1(x_1) + \dots + a_k \times f_k(x_k) \quad [1]$$

given y the output of the model and x_1, \dots, x_k the input features. a_n denotes the coefficients and f_n are the (even non-linear) functionals covering more complex correlations between input and output. EBMs learn the feature functions f by using modern machine learning techniques such as bagging and gradient boosting. The boosting procedure is carefully restricted to train on one feature at a time in round-robin fashion using a very low learning rate so that feature order does not matter. Moreover it round-robin cycles through the features to learn the best feature function and enables for easy estimation of the feature's contribution to the final prediction.

So in particular, the method uses very small trees (decision stumps) as simple and feature specific functionals and each tree is trained using only one single feature at a time. So the model created is a simple additive combination of r trees (number of training samples) in the form of

$$y = a_0 + a_1 \times (T_1^{(1)}(x_1) + \dots + T_r^{(1)}(x_1)) + \dots + a_k \times (T_1^{(k)}(x_k) + \dots + T_r^{(k)}(x_k)) \quad [2]$$

and the sum of all trees for a particular feature represent the (highly non linear) functional f . The main advantage of EBMs in contrast to other approaches for explainability is their capability to discover not only importances for individual feature-dimensions, but also for (dual-feature) correlations. Higher order correlations are also possible in principle and first implementations are already integrated in the InterpretML framework (<https://interpret.ml/>). A potential disadvantage of the method is the fact that compared to other approaches the training time is larger, but during inference the approach requires only simple additions and lookups. This enables EBMs to be one of the fastest models to execute at prediction time by coevally using low memory, as the trees can be represented as simple graphs and thus deleted after the training.

EBMs are especially important for the XAIface project, as they are an inherent part of some of the planned contributions described in section 3.1. They are a prototype for a whitebox-model as each step in the decision process can be explicitly tracked through the algorithm and explained. As they allow for the explanation of each single sample they can provide global explanations of the entire model once a sufficient number of testing samples are analysed.

2.8. Randomised Input Sampling for Explanation (RISE)

Unlike white-box approaches that estimate pixel importance using gradients, RISE (Petsiuk et al. 2018) works on black-box models. The RISE algorithm generates a saliency map for any black-box model, indicating how important each pixel of the image is with respect to the network's decision. This is similar to the saliency maps method mentioned before which require access to the internal structure of the model, such as the gradients of the output with respect to the input, intermediate feature maps, or the network's weights. Therefore they are limited to certain types of network architectures or layers.

In contrast to the plain saliency maps for specific networks presented before, RISE provides a more general approach to produce saliency maps for an arbitrary network architecture. Figure 2.8a illustrates the overall workflow of the RISE algorithm. In detail, it firstly generates random binary masks following a uniform distribution. Then each input image is element-wise multiplied with the random masks and the resulting image is subsequently fed to the model for classification. The model produces probability-like scores for the masked images. In the end, a saliency map for the original image is created as a linear combination of the masks using the probability-like scores as coefficients.

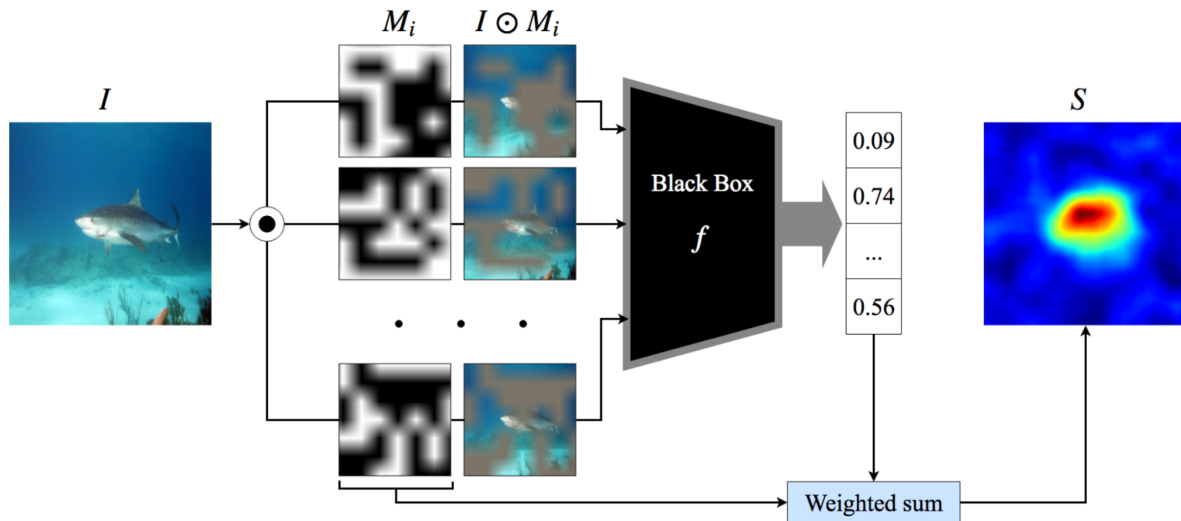


Figure 2.8a: Overview of the RISE algorithm. The input image I is multiplied with some randomly generated masks M_i in an element-wise manner. The black-box model takes masked images as input and produces scores of classes. The score vector serves as weights and linearly combines with the masks to create the saliency map.

RISE is an approach that explains black-box models via estimating the salient regions of input images for the model's predictions. It matches the goal of the XAIface project as it provides a way to interpret a face recognition model by highlighting the important regions of two matching or non-matching faces.

A proper GitHub-implementation can be found at:

 <https://github.com/eclique/RISE>

2.9. Spatial and Feature Activation Diversity Losses for Structured Face Representations

The work presented in (Yin et al. 2019) proposes the usage of two loss functions – spatial activation diversity loss and feature activation diversity loss – to learn more structured face representations. Filters are learned end-to-end from data and constrained to be locally activated with the proposed spatial activation diversity loss. The feature activation diversity loss is introduced to better align filter responses across faces and encourage filters to capture more discriminative visual cues for face recognition, notably when dealing with occluded faces.

By leveraging the face structure, considering part-based representations, the proposed spatial and feature activation diversity losses strive for interpretable representations, which are discriminative and robust to occlusions. The authors claim that the final face representation does not compromise recognition accuracy.

The system architecture proposed by (Yin et al. 2019) for learning meaningful part-based face representations with a deep CNN, using carefully designed losses, is presented in Figure 2.9a. It consists of a Siamese network with two branches sharing weights to learn face representations from two faces: one with synthetic occlusion and one without.

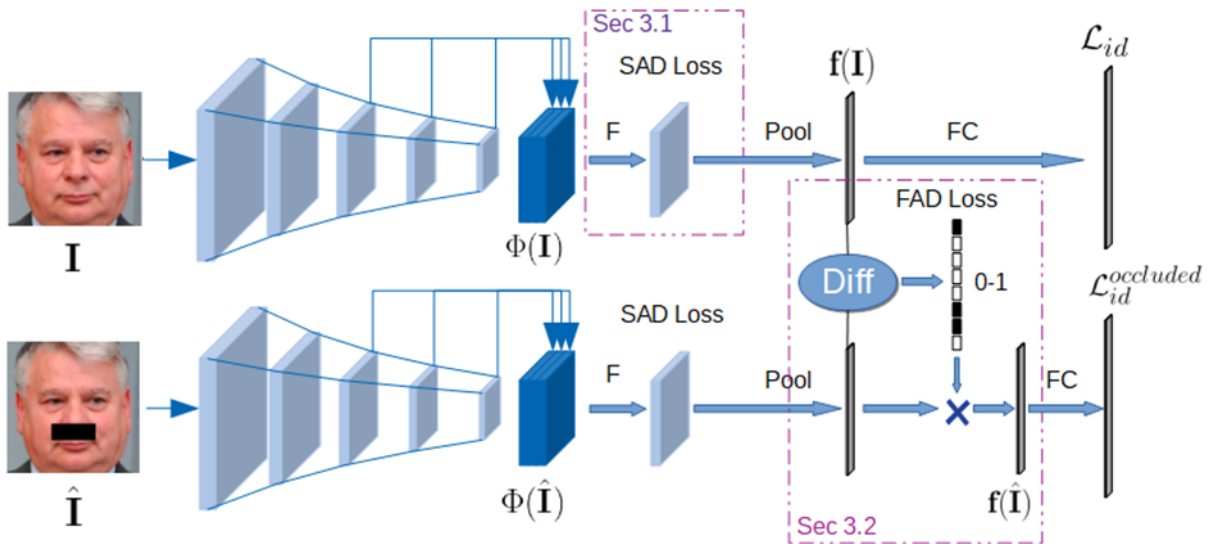


Figure 2.9a: Overall network architecture of the proposed framework. Spatial Activation Diversity (SAD) loss – promotes structured feature responses Feature Activation Diversity (FAD) loss – enforces features to be insensitive to local changes (occlusions).

The Spatial Activation Diversity (SAD) loss encourages the face representation to be structured, with consistent semantic meaning. Softmax loss helps to encode the identity information. The SAD loss learns one set of filters for discriminating the K considered classes and employs the Large Magnitude Filtering (LMF), also proposed in (Yin et al. 2019), which removes small magnitude values, favouring discriminative feature learning. SAD loss improves the spatial spread of peak locations in the feature response maps, and more spreadness indicates higher interpretability.

The Feature Activation Diversity (FAD) loss requires filters to be insensitive to the occluded part, hence more robust to occlusion. The input to the lower network branch is a synthetic occluded version of the top branch input. This way, parts of the face representation sensitive to the occlusion are masked out, and training is performed to identify the input face solely based on the remaining elements. As a result, the filters that respond to the non-occluded parts are trained to capture more discriminative cues for identification. The FAD loss encourages that any local face area only affects a small subset of the filter responses, thus learning part-based face representations, when learning the network model. Therefore, the learned filters are also robust to occlusions. This is achieved by leveraging pairs of face images (one of them with a synthetically occluded region), enforcing the two feature representations to be similar.

The implementation of the interpretable face model using the SAD and FAD loss functions is available at:

 https://github.com/yubangji123/Interpret_FR

2.10 Identifying sources of Harm throughout the Machine Learning Life Cycle

The work presented in (Susesh et al. 2021) aims to prevent and mitigate unexpected system behaviours in real case scenarios. For doing so, the authors state that it is critical to understand when and how harm might have been introduced throughout the model's life cycle. With this purpose, the authors divide the ML life cycle into two streams and identify seven sources of bias, which may cause harm, across the two streams: 1) the data generation stream can contain historical, representational, and measurement bias; and 2) the model building and implementation stream can contain learning, aggregation, evaluation, and deployment bias.

Each type of bias is defined as follows:

1. *Historical bias* replicates biases, like stereotypes, that are present in the world as is or was;
2. *Representation bias* underrepresents a subset of the population in its sample, resulting in poor generalisation for that subset;
3. *Measurement bias* occurs in the process of designing features and labels to use in the prediction problem;
4. Aggregation bias arises when data contains underlying groups that should be treated separately, but that are instead subjected to uniform treatment;
5. *Learning bias* concerns modelling choices and their effect on amplifying performance disparities across samples;
6. *Evaluation bias* is attributed to a benchmark population that is not representative of the user population, and to evaluation metrics that provide an oversimplified view of model performance;
7. *Deployment bias* arises when the application context and usage environment do not match the problem space as it was conceptualised during model development.

3. XAIface Contributions to AI-based Face Recognition Explainability

In this section, we describe the implemented methods for explanations as well as some preliminary and future ideas still under development and investigation by the XAIface consortium. In particular we provide detailed descriptions of the successfully implemented approaches and (basic) performance evaluations of the modules. A link to legal and ethical guidelines concerned by each method is also shortly mentioned.

3.1. Explainable Face Recognition by Interpretable, Local Features (LIBF)

3.1.1. Introduction

Although explaining the contribution of several parts of a face image using standard explanation methods (e.g. Class Activation Maps (CAM), Grad-CAM, Grad-CAM++ or

Saliency Maps) is very useful, the principle disadvantages of such methods already mentioned (e.g. extra layer required, runtime issues, equivalency not guaranteed) are a major drawback. For example, a well known, straightforward approach to explain the end-to-end face-recognition process is the investigation of locally similar perturbations on input images or features following the paradigm of local interpretable model-agnostic explanations (LIME) already described in chapter [2.3](#), but the result, showing how specific pixel perturbations on different locations influence the final results, are not intuitive for the human. In case the recognition process fails, it will not be clear for the user how such single pixel perturbations should be altered or even worse, how single face-feature dimensions can be influenced in order to enable a successful recognition procedure. In other words, it would be rather impossible to take any countermeasures to improve the recognition process by altering the base (face-acquisition) situation with respect to any changes possible through end-users interactions (e.g. by changing the light conditions, acquisition geometry or presentation of different viewpoints).

Similarly, non-interpretability with respect to the original face image presented may occur in the case of application of per se explainable and understandable algorithms (e.g. decision trees) on face features obtained by deep learning. Face features in the latent space are usually non-locality preserving, abstract representations (e.g. 512-dimensional ArcFace features), and the influence of an individual feature dimension has no clear relation to the locality of origin (root cause) in the input face image.

The proposed method in this chapter targets all the issues mentioned above, and it can be seen as a general method for having the ability to map novel, human understandable and traceable, and localised features to influence and importance parameters of the facial input image. We name the method Locally Interpretable Boosted Features (LIBF) and describe the method in more detail below.

3.1.2. Description of the LIBF method

In order to obtain meaningful explanations for the decisions of face recognition and hints on how to eliminate failures through changes made by the end user of a system, we need to solve two problems. The first one is related to the high complexity of classification layers in state-of-the-art, unexplainable deep-learning based face recognition approaches, prohibiting a direct traceability and understanding of decisions taken by a system. Although it is possible to substitute the final classification layers by (inherently) understandable approaches (such as e.g. decision trees), there is a significant loss in classification accuracy. The second one is caused by the highly non-linear operations performed during calculation of deep-learning based face-recognition features. Although it is possible to establish relations between certain input pixels of the input image and the features obtained for classification (by e.g. LIME), there is no guarantee for their spatial proximity and more important, human interpretability. For instance it may be not intuitive, how single pixel perturbations can be altered or even worse, how single face-feature dimensions can be influenced.

To overcome the two issues stated, we propose to combine the aspects of locally interpreting the input data of the face-recognition pipeline with the easy explanation capability of a lightweight classification engine. In particular we use the explainable boosting machine from chapter [2.7](#) together with novel, human-understandable face region

descriptors around local, interpretable (meaningful) and distinctive face image patches and call them Locally Interpretable Boosted Features (LIBF).

LIBFs can be either manually designed or ideally be learned in an unsupervised setup. Thus we can either directly predict the decision of an unexplainable face recognition system (similar to a black-box explanation approach) or estimate at least its probability for success in order to understand the reasons for the unexplainable system's decision. This basic idea is illustrated in see Figure 3.1a.

The left side of Figure 3.1a depicts schematically the workflow of a standard unexplainable reference pipeline defined and to be explained. In particular an input image containing persons (top left) is processed by a face detector (e.g. Retina-Face) to cut out and normalise a face image. On this face-image descriptive features are extracted using non-explainable approaches such as ArcFace or MagFace. Finally a classifier is applied, depending on the face analysis task to solve. This can be e.g. a fully-connected neural network, or also a simple cosine similarity distance with thresholding or a nearest-neighbour classifier.

The right side of Figure 3.1a shows schematically our proposed approach starting with the extraction of patches around face landmarks (e.g. eyes, nose, corners of the mouth) using the framework of e.g. Bulat et al. (Bulat 2017), and several other selected and significant patches relative to them (upper red circle) using a specific heuristic described in more detail later. To describe the content of those local patches we can either use hand-crafted but local features such as blurriness-levels or colour histograms. More promising is to use specialised local features which can be e.g. learned using self-supervised techniques as also described below in more detail.

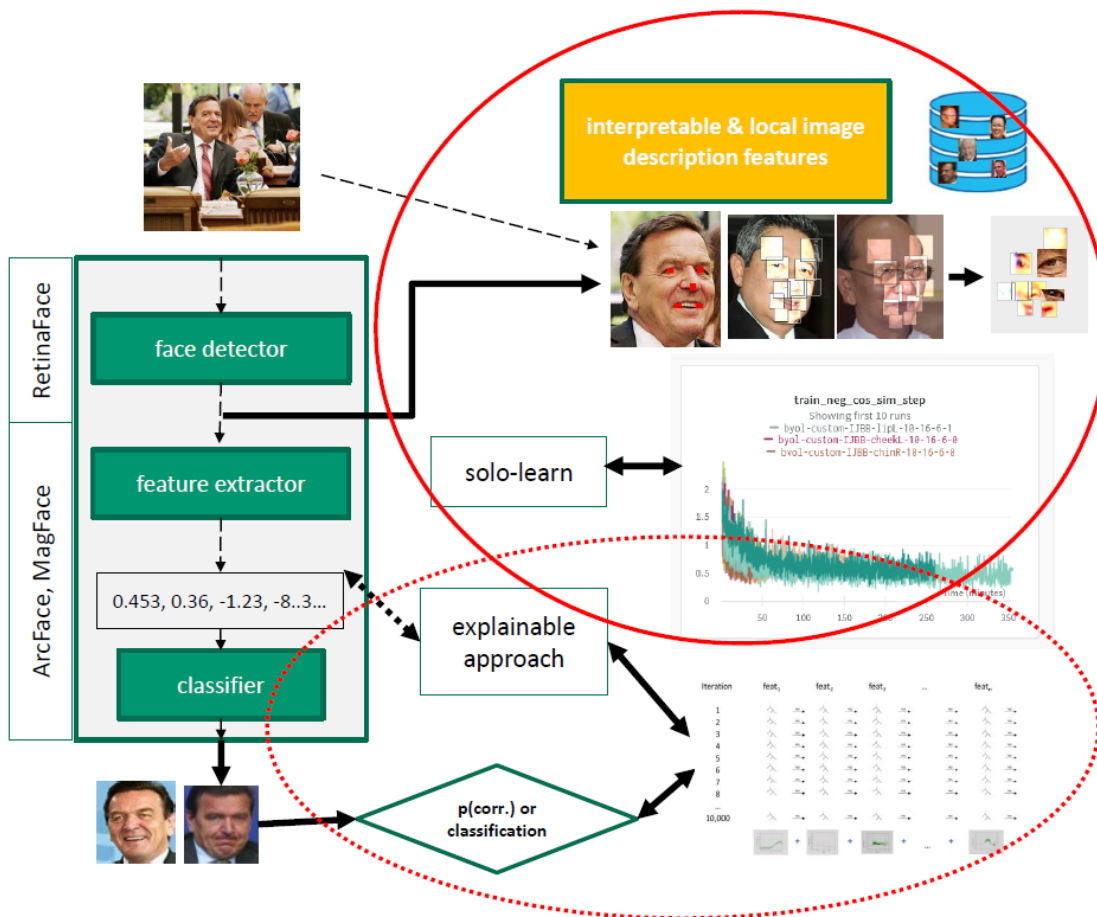


Figure 3.1a: Principle Principle of the proposed LIBF approach. Local, interpretable features, calculated on regions around distinctive points (anchor points) in the face (thus meaningful and understandable for humans) are used to train an inherently explainable approach and estimate probability for correct decisions of the unexplainable pipeline. The dotted arrow indicates a potential extension for loosely coupling our approach to the reference pipeline and avoiding drifting.

Finally, such features can then be used to train an inherently explainable approach (e.g. decision tree or similar approaches like the explainable boosting machine mentioned in Section 2.7) in parallel to the reference pipeline providing similar recognition results or estimating the probability for correct decisions of the native pipeline. Please note that this approach is similar to already presented black-box end-to-end approaches such as RISE (see Section 2.8), but as a main difference to them we will not work on original input data (face images) but on (semi-) local face and image features, which can be easily interpreted by humans. In addition, our approach does not require generating variations of the input image.

Heuristic selection of meaningful patch locations

For human understandability of face recognition processes it is a good idea to relate explanations to distinctive regions in the face which could easily be identified and interpreted by humans. In our approach we select the five natural anchor points, namely eyes, nose and

corners of the mouth as base patches. They could be detected by e.g. the algorithm of Bulat (as mentioned above) or any other method for facial landmark detection.

In order to capture other potentially important parts of the face, we heuristically selected six additional patches - namely (left and right) forehead, cheek and chin. The position and size of the additional patches are determined based on the position of the base patches (such as distance between left eye - right eye or mouth - nose) to be independent of resolution, scale and rotation of the face. Hence in total 11 patches are extracted from a face image. Figure 3.1b shows the principle as well as some examples of extracted face patches.

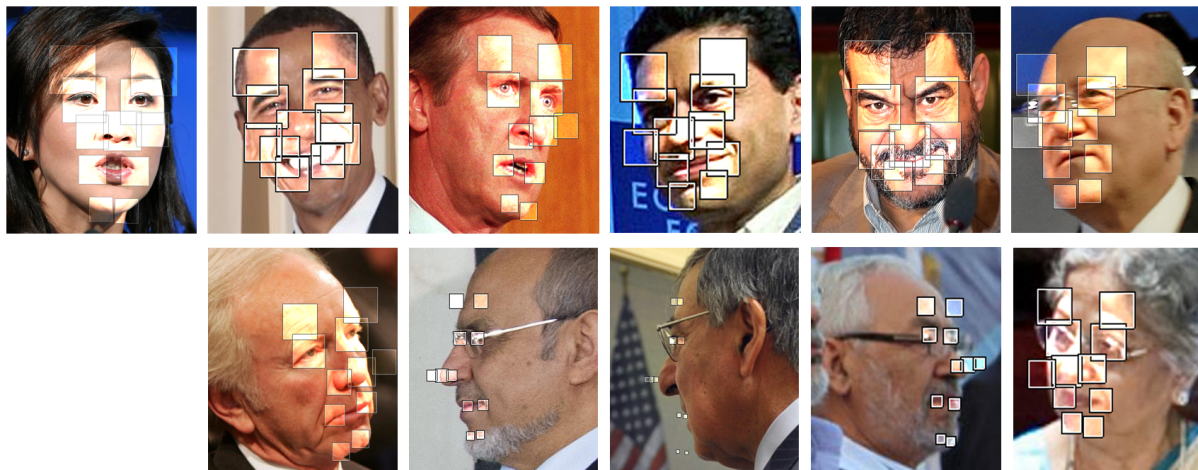
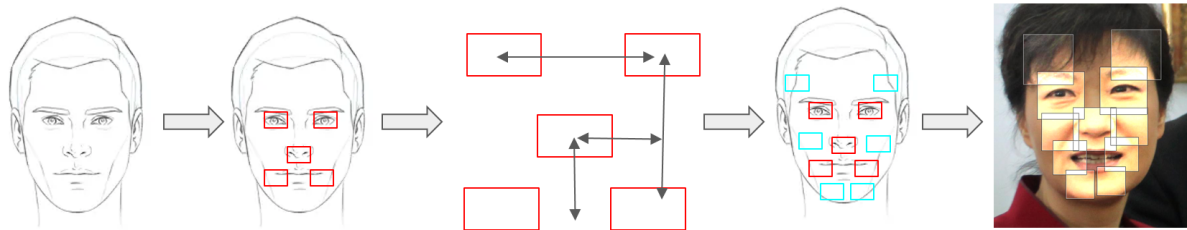


Figure 3.1b: Illustration of the heuristic patch selection principle and some examples of meaningful patches extracted.

Moreover, in the case of extreme rotation around the vertical axis (i.e. side views) some of the patches extracted could be located outside the face or are even invisible. The same is true for images with very low resolution, as we require a minimum patch size of 20x20 pixels for taking an image as a candidate in the training stage. Thus we added a simple heuristic to identify such patches, and store the binary information for the selection of patches to use for the training of the patch embedding and the verification-feature creation. Figure 3.1c shows some examples for the latter case.

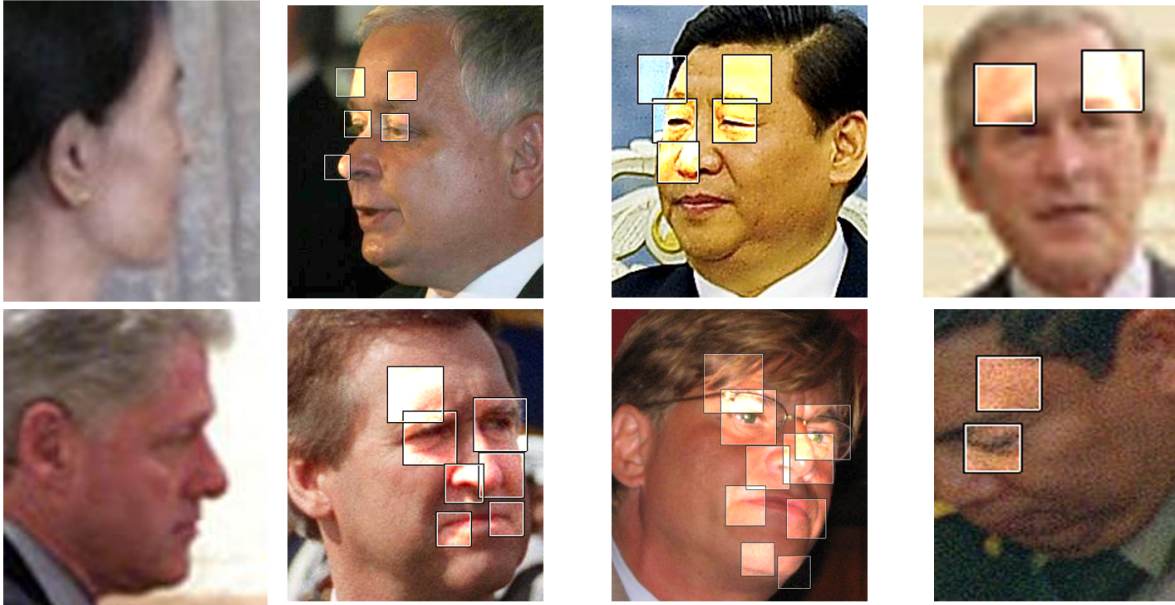


Figure 3.1c: Examples for patches faces with extreme rotation, low resolution and bad allocation causing out of image cropping and align errors.

Self-supervised learning of feasible patch embedding

The features describing the individual face patches should be highly distinctive but also as compact as possible to allow for easy interpretation and understanding. Thus it is not feasible

to use pre-trained image features from other tasks as e.g. object detection. Moreover it is necessary to specifically tune the representation to the nature of the patches themselves and

coevally minimise the number of dimensions. In order to avoid the need for annotated and labelled data, we propose to use a self-supervised technique for generating such embeddings.

In particular we follow the approach of Grill et al. (Grill 2020) , where the authors use a two-stream competitive neural network architecture (Siamese network) to learn an optimal embedding using various augmented views of single images. The network iteratively bootstraps the output of two competitive networks that interact and learn from each other. A stop-gradient module avoids the collapsing of the networks to trivial solutions. The main advantage is that the method achieves state-of-the-art results as proven for e.g. the evaluation protocol of ImageNet – but without using any negative image pairs – which makes the approach best suited for our purposes. In our approach we set the dimensionality of the latent feature

vector (the embedding) to $d = 16$. This selection is inspired by the fact that a dimensionality of face features of 10 already enables useful reconstruction (Celis 2019). Thus we obtain a 16-dimensional embedding for each of the patch types (e.g. left-mouth corner, chin etc.).

3.1.3. Performance Evaluations

Performance evaluations have been continuously conducted during the development of our methods. Thus we make use of the evaluation metrics and protocols already documented on deliverable 2.2. In particular, we use the publicly available IARPA Janus Benchmark-B (IJB-B) and Benchmark-C (IJBC) datasets as main image sources as well as their corresponding verification protocols to allow for fair comparison. Note, that especially the IJB-C dataset is a de facto standard for unconstrained face recognition.

In the following performance evaluations reported we focus on the IARPA Janus Benchmark-C face challenge baseline 1:1 Verification protocol as this is actually one of the most recent and most comprehensive verification protocols. The verification process is a two-stage one and consists of an embedding pre-calculation step generating all the face image patches on individual locations followed by the verification procedure. The first one is the most time-consuming step conducted for each patch separately, and lasts between 7 and 29 hours (per patch-type) depending on the patch-type on an Intel Xeon E5-2640 CPU (2,4GHz).

Evaluation of verification performance using the learned self-supervised embeddings

In order to check the feasibility of the self-supervised embeddings learned, we applied the learned patch representations to a face verification task according to the IJB-x verification protocols. Both the IJB-B and IJB-C verification protocol consist of dedicated images used for the verification task as a total 1:1 evaluation would be too complex and needs too much time.

Moreover the protocol defines a dedicated set of face images from a single person to build a so-called template representation of a face. This is simply done by averaging all the feature vectors from the face image set. For generating the query templates procedure we follow the same pipeline for patch-extraction as described before and obtain a 16-dimensional feature vector for each patch through the embedding learned.

The IJB-x verification protocol then simply uses the cosine-similarity distance between query and template features to calculate a matching score. Following the IJB-x 1:1 Verification protocol we present the results as detection error tradeoff curves (DET) on false-match and non-false-match rates (FMR, NFMR).

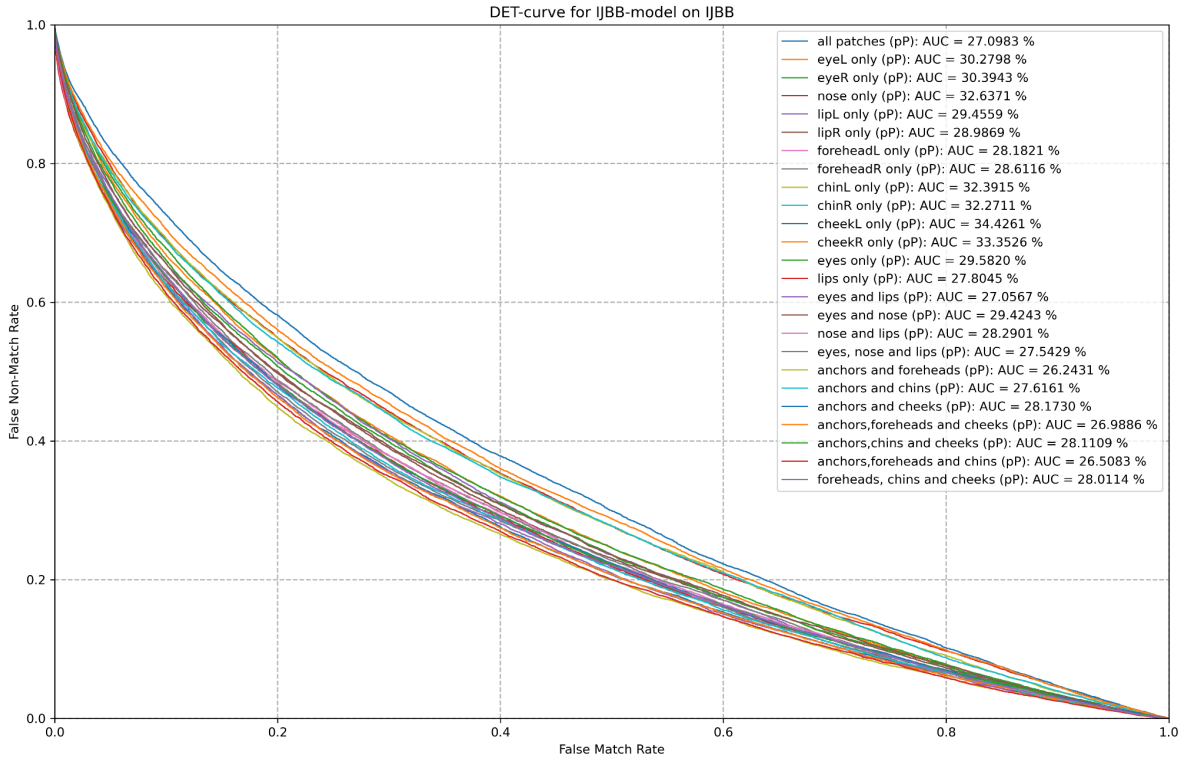


Figure 3.1d: DET curves according to the IJB-B 1:1 Verification protocol using the learned embeddings trained on IJBB for each individual type of face-patches.

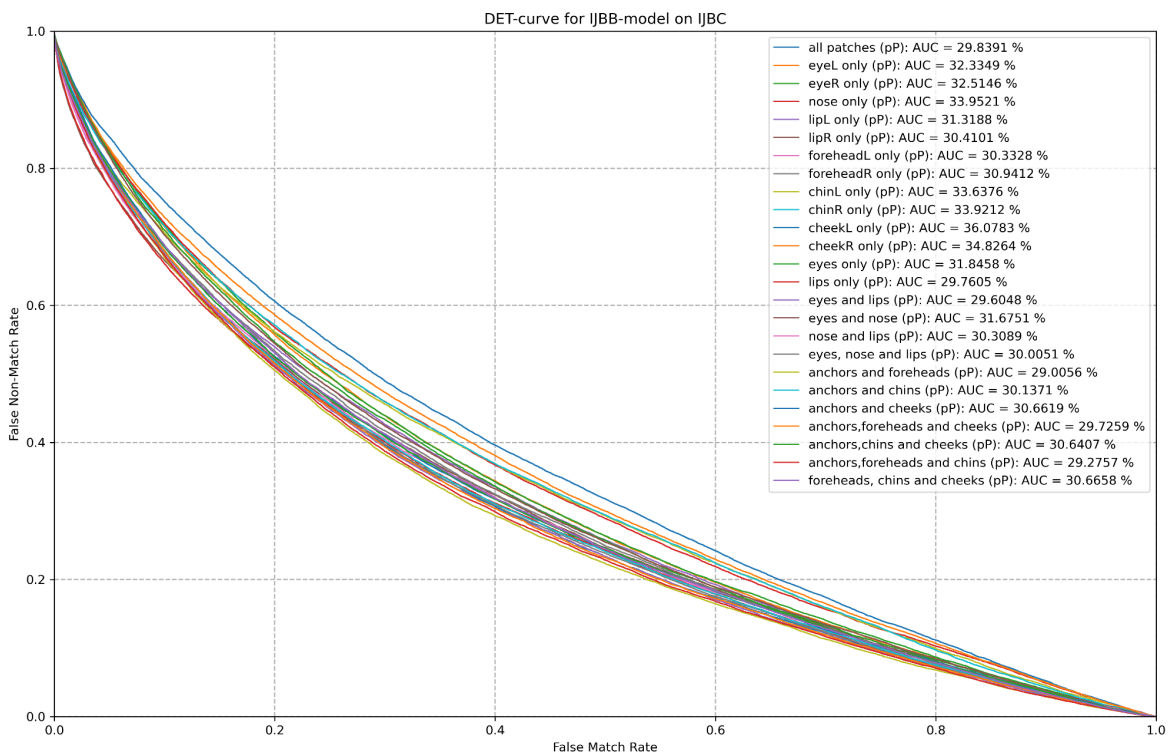


Figure 3.1e: DET curves according to the IJB-C 1:1 Verification protocol using the learned embeddings trained on IJBB for each individual type of face-patches.

The DET curves are calculated for each patch-type separately. The Figures 3.1d and e (for the IJB-B and IJB-C dataset respectively) clearly show, that the performance of the verification approach using a single patch-feature individually is clearly above randomness (which would be a diagonal line) and thus the embeddings calculated are feasible.

In order to check, if there is a dependency of the trained embedding on the verification performance, we also trained the embedded representations on IJBC dataset and obtained similar results as shown in the Figures 3.1f and g.

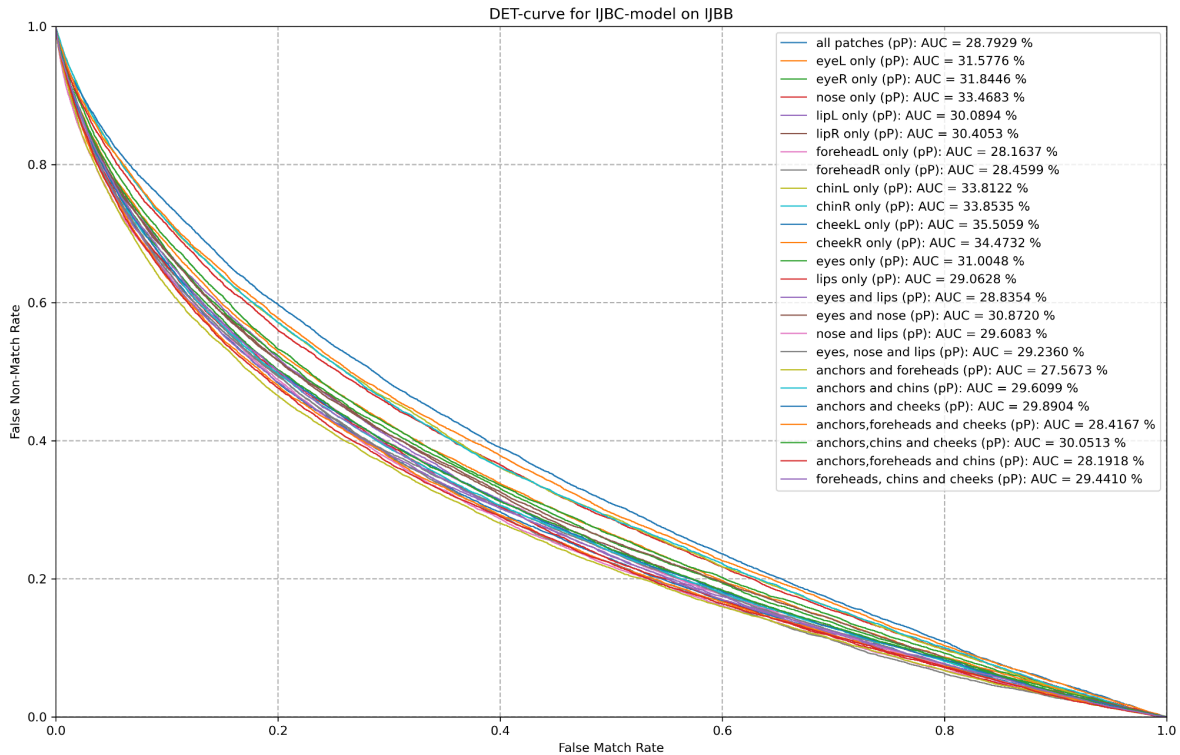


Figure 3.1f: DET curves according to the IJB-B 1:1 Verification protocol using the learned embeddings trained on IJBC for each individual type of face-patches

In order to allow an estimation of the evaluation consistency among the evaluations and gain more insight regarding the importance and performance of the individual patch-representations (patches) we re-arranged the results in a table, colour-coded the individual patches and ranked the patch-locations in decreasing AUC order (note, that the smallest AUC value is best in DET-curves) as shown in Table 3.1

Taking into account the presented results, we draw the following main conclusions:

- The BYOI-Trainings for IJBB and IJBC provide similar results independent from the selection of training/evaluation set (IJBx train / IJBx test).
- The ranking of the patches is consistent over all evaluations. The selection of anchors and forehead patches are best and cheek, chins and eyes are the worst selections regarding verification performance.
- As expected the combination of patches outperform single patch verification alone.

Rank	calcAllROCVals_perPatch_IJBB_IJBB		calcAllROCVals_perPatch_IJBB_IJBC		calcAllROCVals_perPatch_IJBC_IJBB		calcAllROCVals_perPatch_IJBC_IJBC		
	patches used	AUC [%]	patches used	AUC [%]	patches used	AUC [%]	patches used	AUC [%]	
1	anchors and foreheads (pP):	26.24	anchors and foreheads (pP):	29.01	anchors and foreheads (pP):	27.57	anchors and foreheads (pP):	30.02	BEST
2	anchors,foreheads and chins (pP):	26.51	anchors,foreheads and chins (pP):	29.28	foreheadL only (pP):	28.16	foreheadL only (pP):	30.15	
3	anchors,foreheads and cheeks (pP):	26.99	eyes and lips (pP):	29.60	anchors,foreheads and chins (pP):	28.19	anchors,foreheads and chins (pP):	30.68	
4	eyes and lips (pP):	27.06	anchors,foreheads and cheeks (pP):	29.73	anchors,foreheads and cheeks (pP):	28.42	foreheadR only (pP):	30.68	
5	all patches (pP):	27.10	lips only (pP):	29.76	foreheadR only (pP):	28.46	anchors,foreheads and cheeks (pP):	30.83	
6	eyes, nose and lips (pP):	27.54	all patches (pP):	29.84	all patches (pP):	28.79	eyes and lips (pP):	31.00	
7	anchors and chins (pP):	27.62	eyes, nose and lips (pP):	30.01	eyes and lips (pP):	28.84	lips only (pP):	31.11	
8	lips only (pP):	27.80	anchors and chins (pP):	30.14	lips only (pP):	29.06	all patches (pP):	31.24	
9	foreheads, chins and cheeks (pP):	28.01	nose and lips (pP):	30.31	eyes, nose and lips (pP):	29.24	eyes, nose and lips (pP):	31.34	
10	anchors,chins and cheeks (pP):	28.11	foreheadL only (pP):	30.33	foreheads, chins and cheeks (pP):	29.44	nose and lips (pP):	31.61	
11	anchors and cheeks (pP):	28.17	lipR only (pP):	30.41	nose and lips (pP):	29.61	anchors and chins (pP):	31.84	
12	foreheadL only (pP):	28.18	anchors,chins and cheeks (pP):	30.64	anchors and chins (pP):	29.61	foreheads, chins and cheeks (pP):	31.94	
13	nose and lips (pP):	28.29	anchors and cheeks (pP):	30.66	anchors and cheeks (pP):	29.89	lipL only (pP):	31.99	
14	foreheadR only (pP):	28.61	foreheads, chins and cheeks (pP):	30.67	anchors,chins and cheeks (pP):	30.05	anchors and cheeks (pP):	32.02	
15	lipR only (pP):	28.99	foreheadR only (pP):	30.94	lipL only (pP):	30.09	lipR only (pP):	32.04	
16	eyes and nose (pP):	29.42	lipL only (pP):	31.32	lipR only (pP):	30.41	anchors,chins and cheeks (pP):	32.27	
17	lipL only (pP):	29.46	eyes and nose (pP):	31.68	eyes and nose (pP):	30.87	eyes and nose (pP):	32.62	
18	eyes only (pP):	29.58	eyes only (pP):	31.85	eyes only (pP):	31.00	eyes only (pP):	32.68	
19	eyeL only (pP):	30.28	eyeL only (pP):	32.33	eyeL only (pP):	31.58	eyeL only (pP):	33.03	
20	eyeR only (pP):	30.39	eyeR only (pP):	32.51	eyeR only (pP):	31.84	eyeR only (pP):	33.46	
21	chinR only (pP):	32.27	chinL only (pP):	33.64	nose only (pP):	33.47	nose only (pP):	34.71	
22	chinL only (pP):	32.39	chinR only (pP):	33.92	chinL only (pP):	33.81	chinL only (pP):	35.37	
23	nose only (pP):	32.64	nose only (pP):	33.95	chinR only (pP):	33.85	chinR only (pP):	35.61	
24	cheekR only (pP):	33.35	cheekR only (pP):	34.83	cheekR only (pP):	34.47	cheekR only (pP):	35.89	
25	cheekL only (pP):	34.43	cheekL only (pP):	36.08	cheekL only (pP):	35.51	cheekL only (pP):	36.83	WORST

Table 3.1: Ranked and colour-coded AUC values for various embedding types (IJBB and IJBC embeddings) and patch locations.

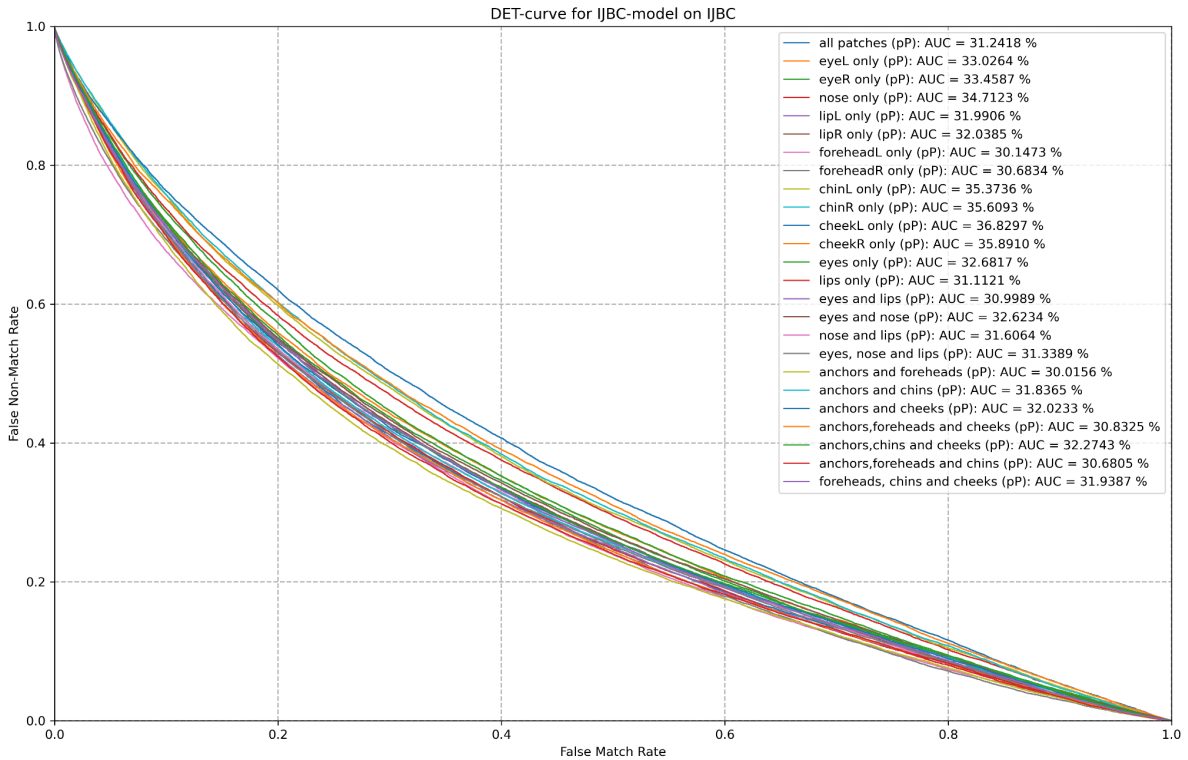


Figure 3.1g: DET curves according to the IJB-C 1:1 Verification protocol using the learned embeddings trained on IJBC for each individual type of face-patches.

Explanations gained using EBM in the verification scenario

In this paragraph we present first insights from applying the EBM to the features obtained through learning the embeddings described in the previous section. In particular, we use the implementation of EBM contained in Microsoft's InterpretML framework (see <https://github.com/interpretml/interpret>) and follow the feature extraction scheme described above. In addition to the 16-dimensional feature vector obtained by applying the learned embedding to the face patches, we use the binary information on visibility of face patches (caused by extreme vertical rotation) as an additional feature. Thus we end up with an 11x17 dimensional description for faces and a final normalisation step is applied to query and template features. We empirically found that per-patch normalisation of the feature vectors (normalisation for each group of 17 features for one single patch) provided best results.

In order to train the EBM for a verification task, we have provided both the query and the template LIBF at the same time. Thus we encode the comparison of the 187-dimensional LIBFs for query (L_q) and template features (L_t) directly in the feature representation. Such a verification feature representation (V) can then be directly used to estimate the label (match or no match) encoded as [0, 1]. We investigated two different encodings, namely

- a simple concatenation or stacking of the query and template LIBFs ($V_c = L_{q,1}, \dots, L_{q,187}, L_{t,1}, \dots, L_{t,187}$) resulting in a 374 dimensional verification-feature and
- a verification-feature obtained through multiplying the corresponding query and template LIBF dimensions ($V_m = L_{q,1} \cdot L_{t,1}, \dots, L_{q,187} \cdot L_{t,187}$).

For the training it is furthermore important to use a dataset with an approximately equal number of matching and nonmatching template pairs from the IJB-B dataset. Thus we reduce the number of pairs to ~10k samples each, coevally taking care of an almost balanced distribution of face-samples per person. The training procedure itself is rather fast using 90% of the 20 cores on an Intel Xeon E5-2640 CPU (2,4GHz).

To gain first global insights on the EBM-classifier trained to mimic a face verification task we analyse the visualisations of the mean absolute importance scores for various features and feature-pairs. Figure 3.1*h* depicts the 15 most important contributions to the decision of the classifier made for both verification feature encodings. There is obvious prioritisation for several single feature dimensions and pairwise correlation depending on the type.

Overall Importance: Mean Absolute Score

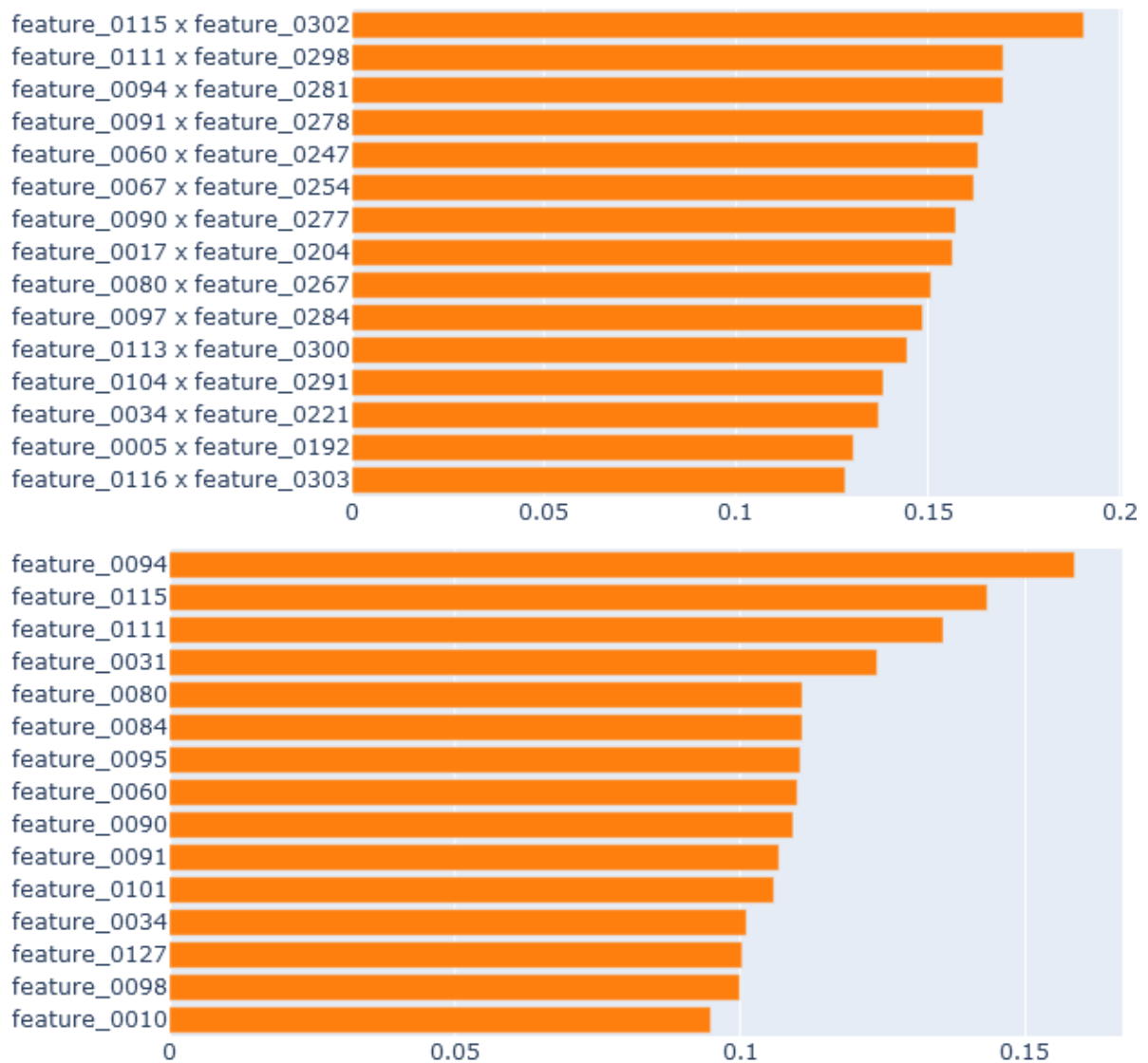


Figure 3.1h: Global explanation for the stacked verification features (V_c) on top and multiplied verification features (V_m) on bottom.

As expected, the EBM correctly identifies dual-feature correlations (we can find only feature-pairs among the entries) as most important contributions in the case of the stacked (V_c) verification feature type (top). Even if displaying all the EBM selected feature dimensions as shown in Figure 3.1i only dual-feature correlations contribute to the overall verification procedure. The opposite is true for the multiplied (V_m) verification features types (bottom).

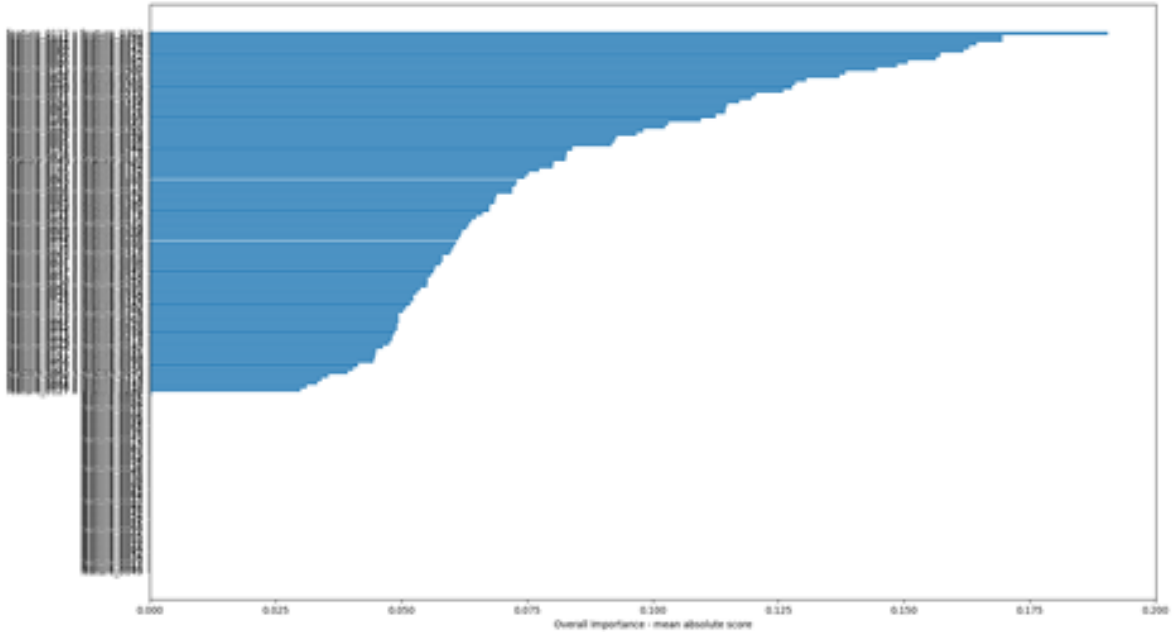
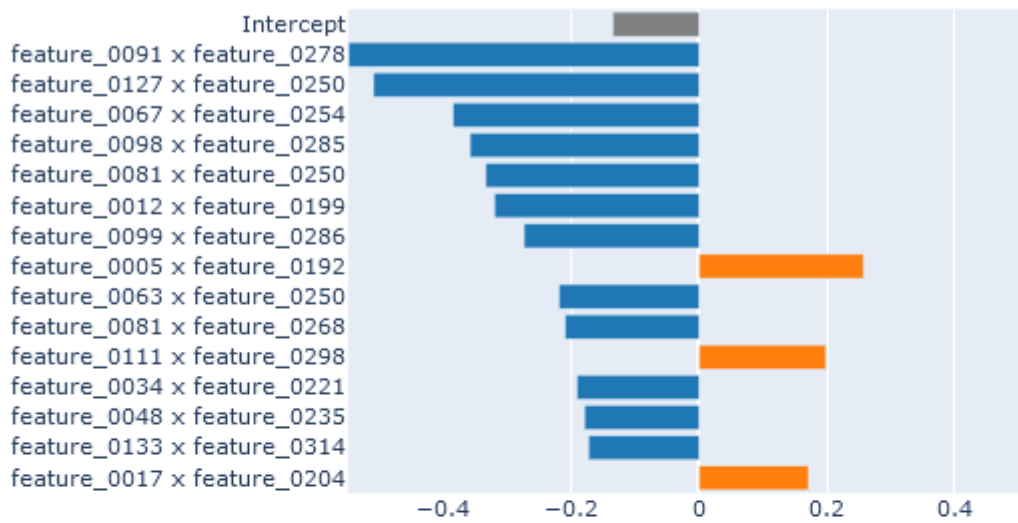


Figure 3.1i: All global explanations for the stacked verification features (V_c) as identified by the EBM. Although the image is not sharp due to limited resolution, the selection of dual-feature correlations contributing to the verification process is clearly visible.

Besides the global explanation of the entire classifier trained with the EBM it is also interesting how individual features of an arbitrary instance contribute to the verification output. To provide initial insights on that, we exemplarily show some results in Figures 3.1j and k for the stacked (V_c) and multiplied (V_m) verification feature encodings, respectively.

Predicted (0): 0.946 | Actual (0): 0.946



Predicted (1): 0.923 | Actual (1): 0.923

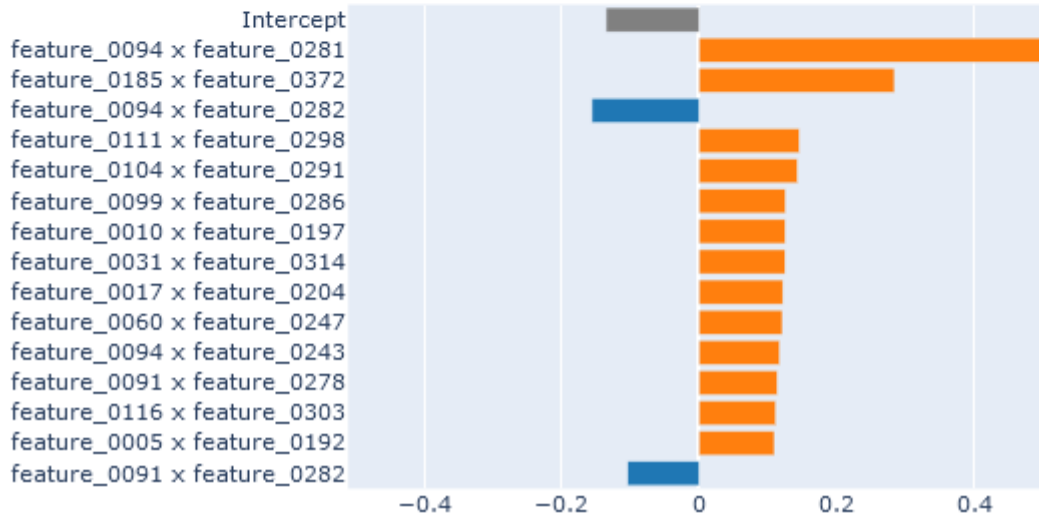
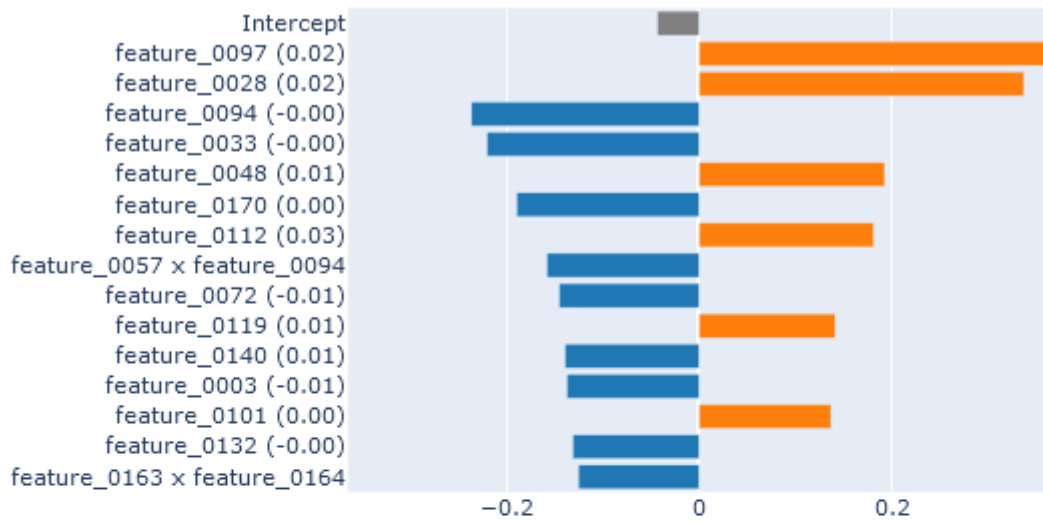


Figure 3.1j: Local explanation examples of non-matching, verification features (V_c) on top and matching verification features on bottom.

In Figure 3.1j one can see that for the stacked verification features (V_c) only feature combinations provide most important contributions to the final matching decision. In the case of features from non-matching faces (Figure 3.1j top) the majority of correlated feature pairs vote for the desired output, while the opposite is true for the matching ones in the graphic below.

For the multiplied verification features in Figure 3.1k a similar behaviour is observed, but mostly for single features. It is interesting to see that also some feature combinations provide contributions. This might indicate some hidden correlation between (neighbouring) face image patches extracted.

Predicted (0): 0.683 | Actual (0): 0.683



Predicted (1): 0.863 | Actual (1): 0.863

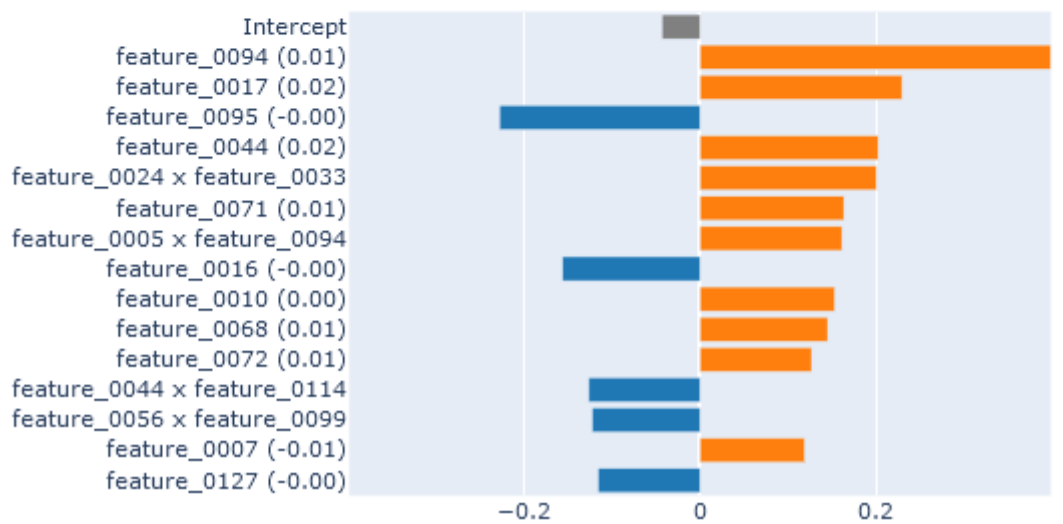


Figure 3.1k: Local explanation examples of non-matching (above) and matching (below) multiplied verification features (V_m).

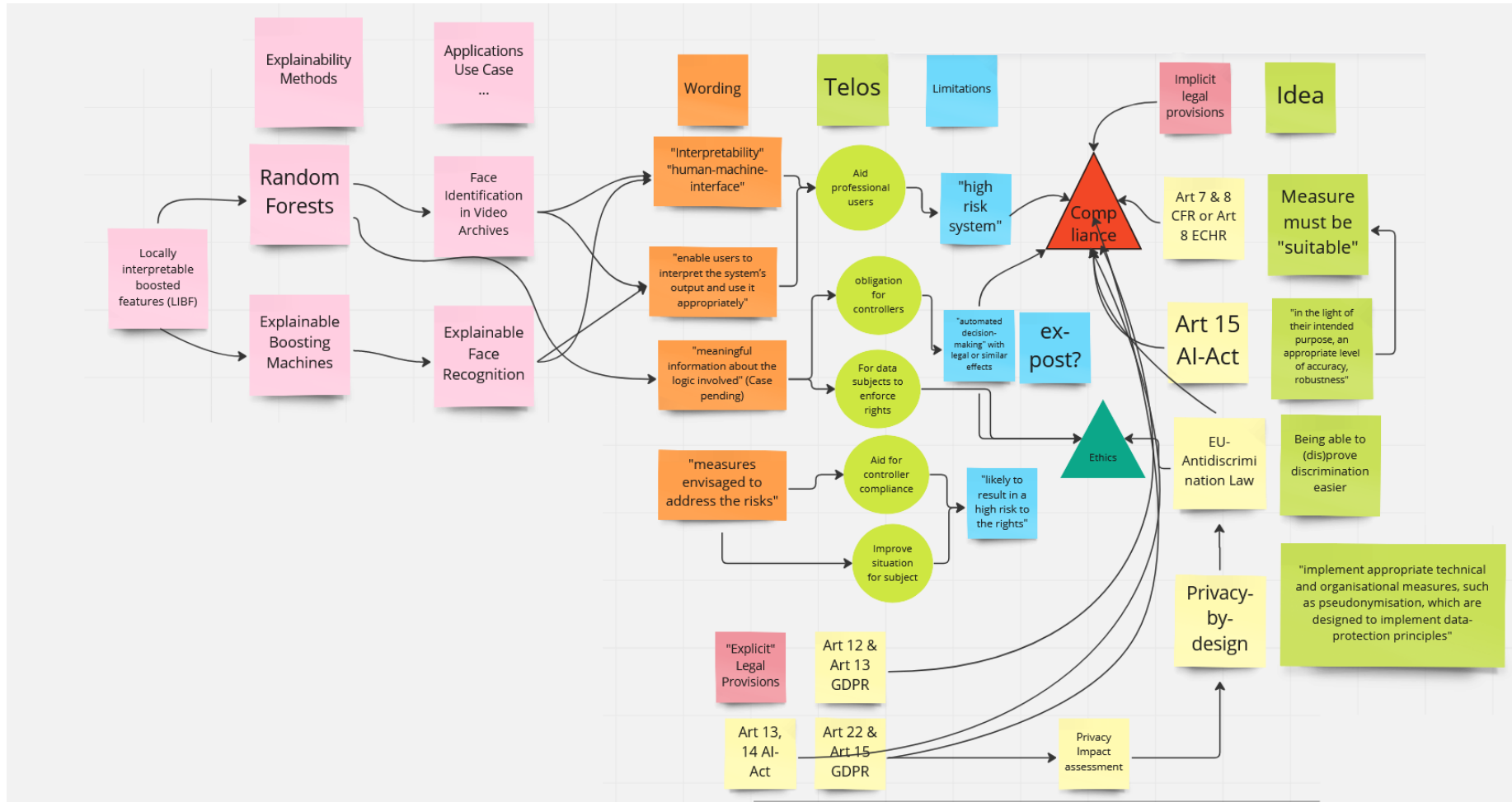


Figure 3.11: Locally Interpretable Boosted Feature (LIBF) in the context of legal and ethical requirements concerned.

3.1.4. Legal and Ethical guidelines concerned by the method

The proposed Locally Interpretable Boosted Features (LIBF) method and the explainable boosting machine subsequently used tackle several legal and ethical obligations required by the in depth-analysis from deliverables 3.3 and 3.4. To identify and highlight these dependencies we make use of Figure 3.1I and identify the following wordings as most relevant and applicable for our method.

- “interpretability”,
- “human-machine interface”,
- “enable users to interpret the system’s output and use it appropriately” and
- “meaningful information about the logic involved” (Case pending)”

Hence we can deduce from the dependencies in Figure 3.1I, that LIBFs used as an aid for professional users can be treated as a “high risk system” causing compliance with several articles of the CFR and ECHR as well as articles 13-15 of the AI-Act and article 12, 13, 15 and 22 of the GDPR.

In addition our method can be linked against the explicit compliance efforts regarding AI-Act and GDPR elaborated and reported in chapter 4.4.1 and 4.4.2 respectively. Thus our method is especially relevant for the numbered topics

- A-13-1, A-15-1, A-13-2 and A14-3 for AI-Act related recommendations - and
- G-15-1, G-14-1, G-14-2 for GDPR related recommendations.

3.1.5. Summary, conclusion and outlook

As a conclusion of this chapter we state that we proposed a novel, general black-box explanation method for face recognition tasks and Locally Interpretable Boosted Features (LIBF). The method provides a different way for explaining most important regions in face images for the recognition task, but taking into account human understandability and thus providing the ability to apply countermeasures in the case of failures. To prove the feasibility of unlabelled, self-supervised face embeddings learned for specific regions, we verified that the approach already works in principle. Finally, we have shown first results regarding the suitability of a recent high performance explanation technique termed explainable boosting machine for uncovering local feature importance using our proposed features.

Future work will be the application of our approach to a face-identification scenario and comparison to other state-of-the art face recognition pipelines such as ArcFace or MagFace. Moreover we will apply the approach to the use case of “Video-Archive content search and documentation”. Therefore we plan to optimise the parameters and provide more in-depth results working on this much more demanding dataset. The potential reduction of embedded layer dimensions is also a hot topic of interest we plan to investigate in upcoming research.

Finally, and as a closing remark for this chapter, we want to mention that the method proposed can also be seen as a generic approach to address explainability tasks for other domains. Thus it will help to avoid (classical) pitfalls such as uncovering the presence of non-task-related image features (e.g. specific background for a certain class) feigning perfect recognition/classification performance.

3.2. Improving Face Verification Explainability using Spatially-Biassed Similarity Metrics and Training Loss Functions

3.2.1. Introduction

Convolutional neural networks (CNNs) have achieved top performance for many computer vision tasks, notably face recognition. However, there is still a lack of effective processes to explain the complex decision-making process of deep learning (DL)-based solutions, especially due to their large, non-linear components. This situation is largely known as the “black box” approach, basically recognizing that the performance may be excellent but there is no clear idea how the final result is obtained. This limits the application of this technology, especially when the decisions have serious implications in real-world applications, e.g. access to critical infra-structures, gender or race discrimination, etc.; in fact, unexplainable false-positive results may lead to serious security and privacy issues. In this context, it is crucial to improve the transparency of the decision-making process for DL-based face recognition solutions which are the target of the XAIface project. Thus, with the help of explainable AI, it is possible to understand and trust more the DL-based models results.

3.2.2. Description

It is well-known that DL-based face recognition results are largely impacted both by the loss function at training time as well as the features/description similarity metric at decision time. It is also intuitive that the similarity between two faces is not equally determined by all parts of the corresponding faces, notably face landmarks may be more relevant.

In the selected XAIface reference face recognition pipelines, notably based on ArcFace and MagFace, two images are said to be from the same person simply if the adopted similarity metric is larger than a specific threshold value. However, this standard procedure does not help in interpreting and explaining to humans the relevance and impact of the high-dimensional features involved in the recognition process.

In this context, the plan for this novel explainable DL-based face recognition solution is to include new similarity metrics and training loss functions that consider the different impact of the various spatial regions in a face image, thus offering a local, spatially-biassed approach. Naturally, the face landmarks may have a special impact in face recognition, which is not even always the same among them, e.g. eyes, nose and mouth. The target is that this novel solution provides quantitative and qualitative reasons to explain why two face images are

from the same person or not. For example, if two face images are associated with the same person, the proposed solution may identify which parts of the face were more impactful and representative, e.g. by providing local similarity values. In this spatially-biased process, it is also possible to include a spatial attention model, which should replicate the attention that humans dedicate to different regions when they recognize faces.

However, explainability capabilities, notably involving new similarity metrics and training loss functions, should not come at an unreasonable price in terms of recognition performance, notably for the verification protocol. Therefore, a novel explainable DL-based face recognition solution must offer an appropriate balance between verification accuracy and human interpretability, e.g. through some meaningful spatial maps which offer spatial explanations for the final decisions. By using the local similarities with different weights, the proposed, improved solution has the potential to become more robust than the original one for partially occluded faces.

Since it is common to perform face recognition on images which have suffered image compression, the recognition performance of the proposed explainable DL-based face recognition solution will also be studied for decoded images, not only using the so-called conventional image codecs, e.g. JPEG, JPEG 2000, JPEG XL, but also the recently emerged DL-based image codecs, notably those presented in the context of the JPEG AI project. In this context, it will also be possible to explain and assess the impact of coding the various face landmarks with different qualities after face and landmarks detection, if image encoders including that facility are available.

3.2.3. Performance evaluation

The novel similarity metrics and training loss functions will be developed and assessed using the two XAIface reference face recognition pipelines as baseline solutions and anchors (notably ArcFace and MagFace) under the recommended verification protocol.

3.3. Similarity-based RISE Algorithm for Explainable Face Recognition System

3.3.1. Introduction

Face recognition has been a crucial task in recent years and the recognition systems have been widely deployed to various applications, such as smartphones or surveillance. With the development of deep convolutional neural networks (DCNN) and an increasingly large amount of available dataset, the methods developed have shown remarkable results on face recognition tasks with DCNN-based approaches. However, beyond the exceptional performance, deep face recognition models trained with large scale datasets were usually treated as "black-boxes", because the model designers can only have an understanding of the dataset or loss functions for training, but very limited understanding of the learned model itself. The deep learning-based recognition systems need more explainability and transparency so that people can truly trust the results they predict or understand the possible failures from them. Visual saliency algorithms have been widely employed to explain the decisions of deep learning systems acting on vision tasks. This section introduces a new

model-agnostic explanation method based on visual saliency map for face recognition models.

3.3.2. Description

Explainable face recognition is a problem of explaining the matches returned by a recognition system in order to provide insight into why a probe face image can be matched with one identity over another. A saliency map-based explainable face recognition method should be able to generate heat maps for critical regions of the input image and interpret the decision of the recognition model. Under this context, we propose a new model-agnostic explanation method for face recognition systems.

A popular group of methods that produce saliency explanations without accessing the intrinsic network architecture mainly performs random perturbation on input image, e.g. noise, occlusion, etc, and determines the importance region by observing the impact of such perturbation on the final output predictions. Despite being useful in principle, existing explainability tools for other image understanding tasks cannot be directly applied to the face recognition task. For example, the Randomized Input Sampling for Explanation (RISE) method explains a classification model by leveraging the categorical output probability of the classifier as the weight to aggregate the final saliency map. However, the decision-making process of a face recognition system mainly involves deep face representation extraction and similarity calculation between at least two images. To address this issue, we propose a Similarity-based RISE algorithm (S-RISE) that leverages the similarity score as weights for the masks and provides explanation saliency maps without accessing the internal architecture or gradients of the face recognition system.

Figure 3.3a depicts an overview of the proposed S-RISE algorithm. In general, given a pair of images $\{img_A, img_B\}$, a mask generator will first randomly produce a fixed number of masks. For each mask M_i , it will be applied to the input image, e.g. img_A . The masked img_A and unmasked img_B are then fed into the face recognition model respectively to capture the deep face representation. Afterward, the cosine similarity is computed as the weight of the corresponding mask. After iterating all the masks, the final saliency map S_A for img_A is the weighted combination of the generated masks.

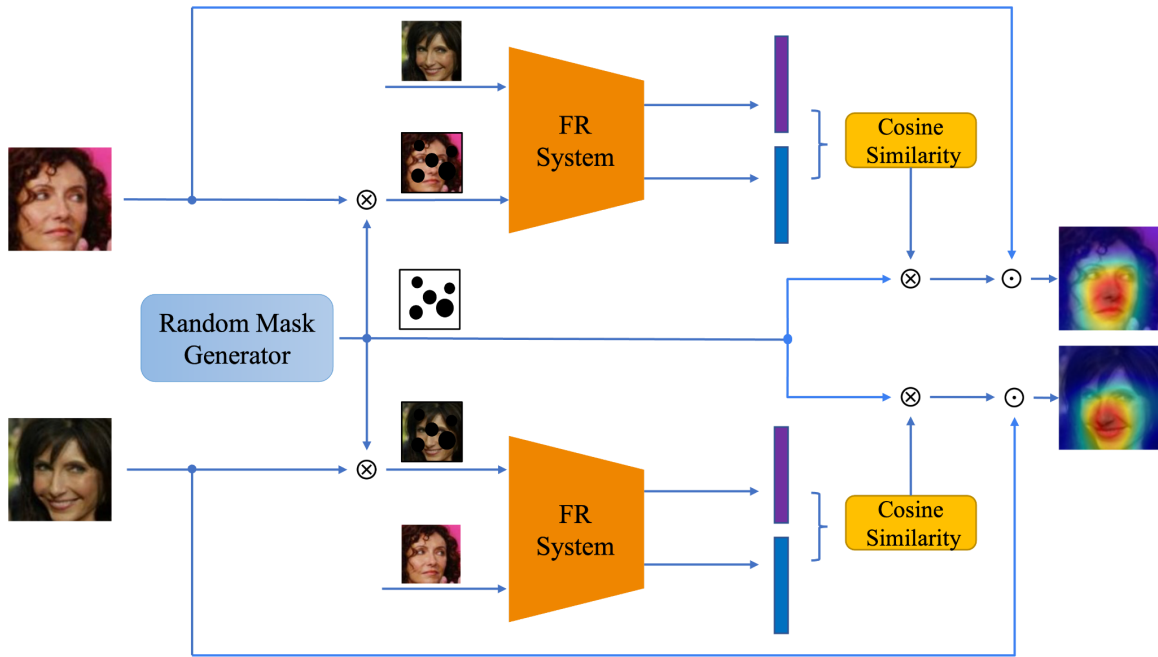


Figure 3.3a: Workflow of the S-RISE explanation method

We interpret the decision-making process of a face recognition model by inputting a triple of probe, mate, and non-mate images. In this context, a well-developed explanation method first should be able to generate saliency maps that highlight the critical regions that are similar to the FR model. Secondly, it should explain why the FR model believes the probe-mate pair is a better matching than the probe-nonmate pair. In practice, the explanation method is applied to the ArcFace face recognition system (one of the reference pipelines to work on agreed in this project). The saliency maps for the probe-mate pair and probe-nonmate pair are calculated separately according to the steps above and re-weighted according to the similarity differences between them.

3.3.3. Visual results of generated saliency maps

This section presents the visual results of the saliency map generated by our proposed S-RISE algorithm. As shown in Figure 3.3b, the produced saliency map properly highlights the regions between the matching pairs that the FR model believes are very similar. As for the probe and nonmate pairs, the heatmaps also represent similar regions but they are much shallower, indicating low similarities between them, which explains why the model believes they are not from the same subject.

To further validate the effectiveness of the proposed explainability model, an additional test has been done with self-occluded faces. Some studies have shown that the current deep face recognition model is capable of identifying partially occluded faces despite lower confidence. In this case, an ideal explanation method should provide low saliency values for occluded pixels while high values for other similar regions. In this experiment, the non-matching face in a triplet is replaced by an occluded image from the same subject as the probe face. As presented in Figure 3.3c, the S-RISE algorithm explains that the FR model manages to verify them through the eye regions when images are occluded by facial masks, and through mouth and nose areas when masked by sunglasses.

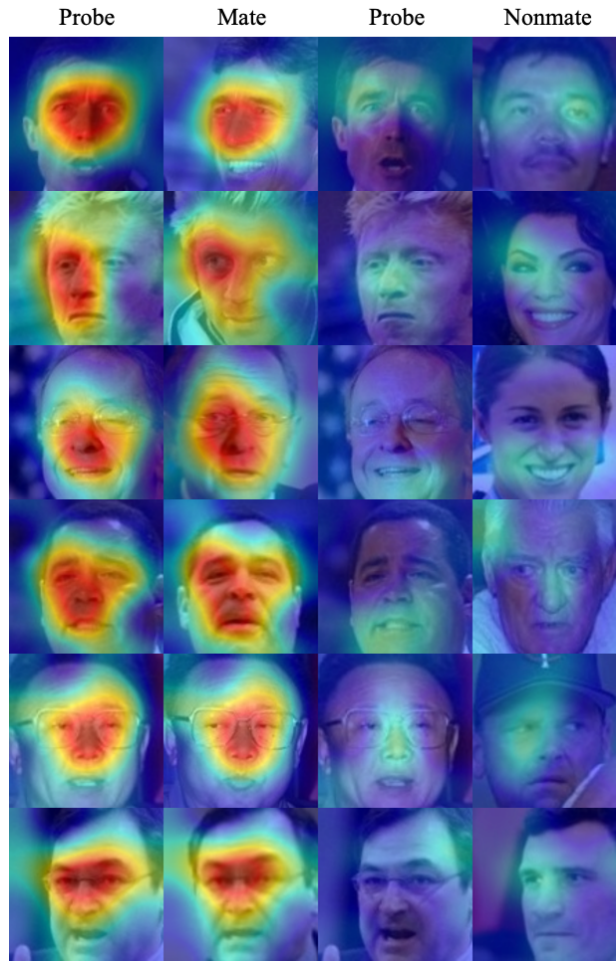


Figure 3.3b: Saliency map explanations for the FR model's prediction on the matching (left) and non-matching (right) image pairs.

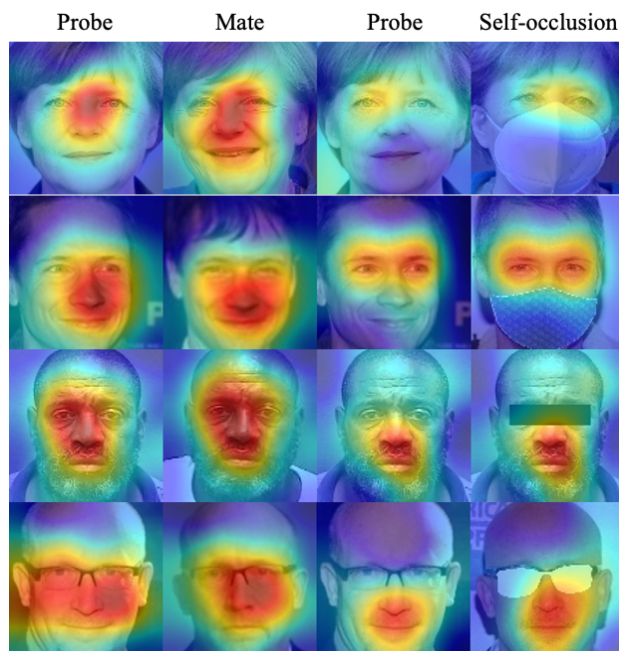


Figure 3.3c: Saliency map explanations for the predictions of the FR model on partially-occluded faces

3.3.4. Performance evaluation

This section presents a new evaluation approach to systematically validate the reliability and measure the performance of the general saliency-based EXplainability Face Recognition (XFR) method. The conventional "Deletion" and "Insertion" evaluation metrics are adapted and improved to better fit the explainable face recognition framework. The main insight is that the explanation saliency map is expected to precisely highlight the most important regions of the face with the smallest number of pixels for the face recognition model to make a correct decision.

In general, the Deletion and Insertion metrics measure how fast the similarity between two faces drops/rises to a threshold value after removing/adding saliency pixels from them. More specifically, the deletion process starts with original images, and the pixels with the highest saliency values are sequentially removed and replaced with a constant value. After removing each pixel, the similarity score is recalculated until it is lower than a predefined threshold. On the contrary, the insertion process starts with the constant value, and the most critical pixels in the image sorted by the saliency map are added to the plain image. The similarity score is recalculated each time after adding one pixel until it is larger than the threshold. The number of pixels deleted from or added to the image is accumulated until the recognition model changes the decision. Overall, the deletion and insertion metrics are defined as $\frac{\#Removed\ Pixels}{\#All\ Pixels}$ and $\frac{\#Added\ Pixels}{\#All\ Pixels}$. In practice, removing pixels from an image alters the original distribution and can eventually affect recognition results. Hence, the constant value above is set as the mean value of the specific image.

Table 3.3: Quantitative evaluation of saliency maps using proposed Deletion and Insertion metrics

Methods	Iterations	Deletion	Insertion	Average
S-RISE	10	0.4466	0.3582	0.4024
	100	0.2617	0.1983	0.2300
	500	0.2071	0.1459	0.1765
	1000	0.2077	0.1384	0.1731

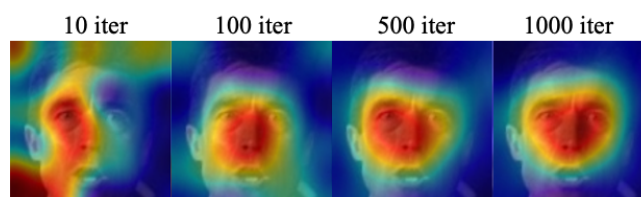


Figure 3.3d: Saliency map generated by S-RISE algorithm with different iteration configurations

Table 3.3 shows the quantitative evaluation for the proposed S-RISE algorithm under different configurations in terms of iterations. The metrics show that a small number of iterations results in poor explanation performance. On the other hand, the metrics gradually converge at around 1000 iterations, which corresponds to stable and accurate saliency maps. Figure 3.3d further validates the conclusion drawn by quantitative evaluation results.

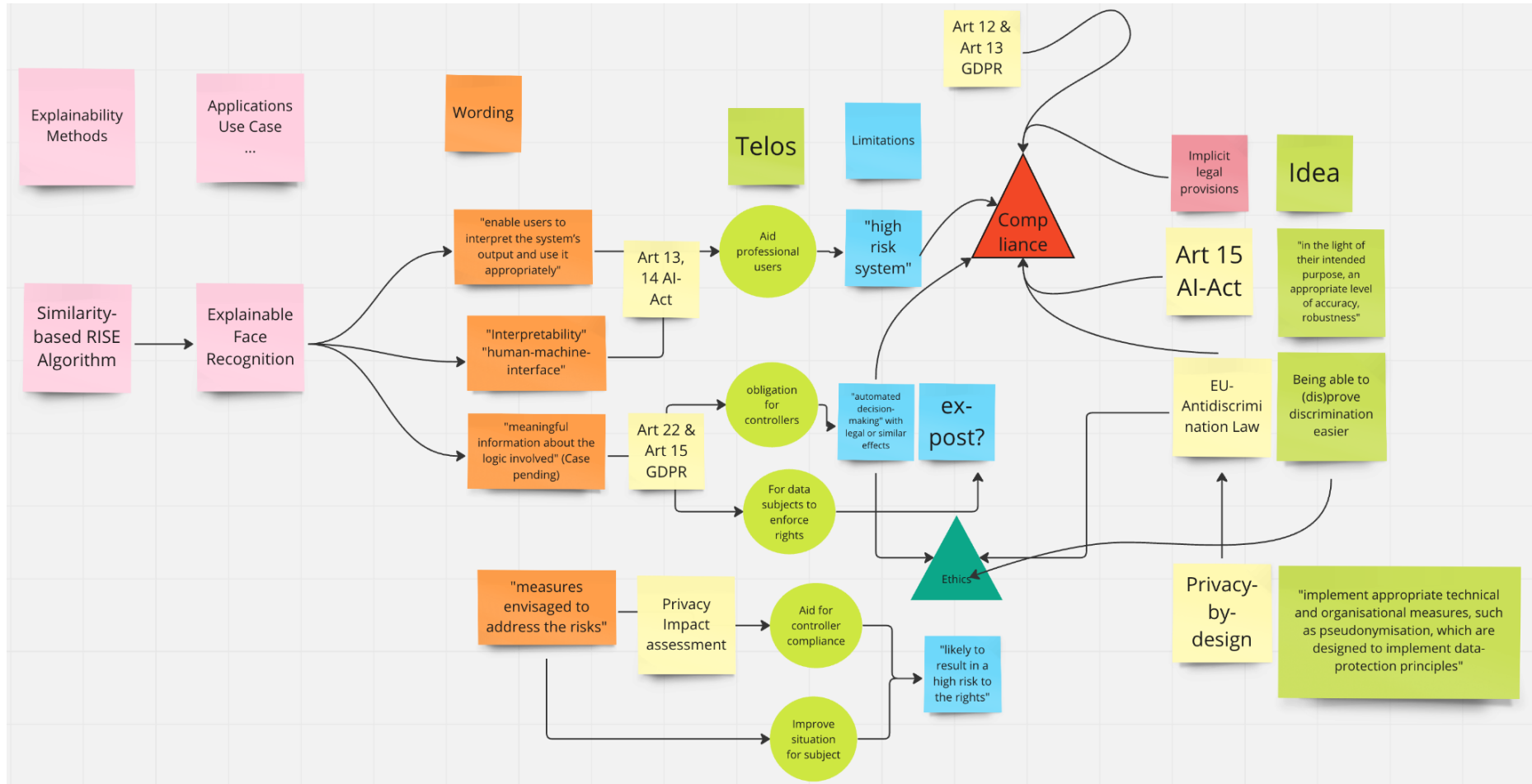


Figure 3.3e: Similarity-based RISE algorithm in the context of legal and ethical requirements concerned.

3.3.5. Legal and ethical guidelines concerned by the S-RISE method

The proposed Similarity-based RISE method tackles several legal and ethical obligations required by the in depth-analysis from Deliverable 3.3 and 3.4. To identify and highlight these dependencies, we make use of the overview graphics shown in Figure 3.3e and identify the following wordings as the most relevant and applicable for our method.

- Interpretability
- Human-machine interface
- Enable users to interpret the system's output and use it appropriately
- Meaningful information about the logic involved

Hence, we can deduce from the dependencies in Figure 3.3e that the S-RISE method used as an aid for professional users can be treated as a “high risk system” causing compliance with several articles of the CFR and ECHR as well as articles 13-15 of the AI-Act and article 12, 13, 15 and 22 of the GDPR.

In addition, our method can be linked against the explicit compliance efforts regarding AI-Act and GDPR elaborated and reported in chapter 4.4.1 and 4.4.2 respectively. Thus our method is especially relevant for the following numbered topics.

- A-13-1, A-15-1, A-13-2 and A14-3 for AI-Act related recommendations
- G-15-1, G-14-1, G-14-2 for GDPR related recommendations.

3.4. Demographic Information Disentangling

3.4.1. Introduction

Face recognition (FR) is currently predominantly based on convolutional neural networks (CNNs). CNNs are used to extract face representations that can be used for several tasks, including identity recognition. Although CNNs are intended to generate representations encoding only the identity information, recent studies have shown that information about soft-biometric traits, including gender, age, and other demographics, is also encoded in these representations (Dantcheva et al. 2015) (Dhar et al. 2020) (Nagpal et al. 2019) (Parde et al. 2017). Soft biometrics are attributes that are not necessarily unique to an individual but can be used alone or in conjunction with primary biometric traits for a variety of applications. However, since these attributes can be extracted from the computed face representations and may potentially be misused, they represent a considerable privacy risk.

3.4.2. Description

For the so-called black-box FR systems, it is not clear how face features such as gender, age, and ethnicity are encoded in the overall face description, namely the face template. Several works have addressed the problem of masking or altering the face representation in order to conceal the soft-biometric traits (gender, ethnicity, age). There are two approaches: image-level techniques operate by suppressing the soft-biometric information in face images so that machine-learning based models, such as CNN, fail to infer it (Mirjalili et al. 2019) (Rozsa et al. 2019). However, they rely on pre-trained classifiers to learn the perturbation to be applied to the image and thus do not generalise well for other classifiers. Template-level approaches try to suppress the soft-biometric information from the more compact face representation encoded in face templates. Existing techniques from this group were shown to generalise well to arbitrary attribute classifiers (Terhörst et al. 2019).

In XAIface, instead of masking or altering the gender, age, or ethnicity information, we plan to develop techniques for disentangling demographic information from the overall face representation in order to understand the impact of such traits on face recognition but also to develop demographic-free face recognition. The latter will also indirectly address fairness and non-discrimination issues by following the idea of de-biasing during the training, as the only information used by the FR pipeline will be related to identity and not to other soft biometric traits.

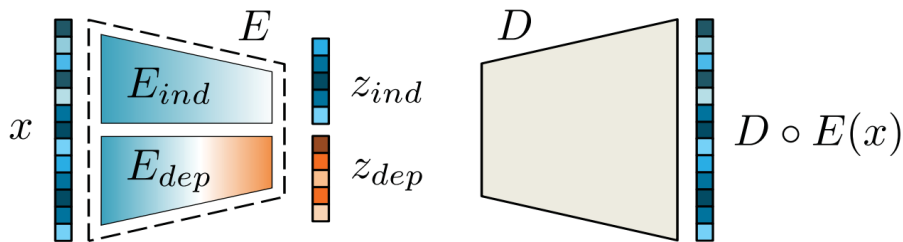


Figure 3.4a: Overall architecture of the PFRNet autoencoder.

An autoencoder will be used to reduce the dimensionality of a facial biometric template (feature vector) and at the same time to separate identity information from demographic information. To do this, specific loss functions will be used to drive the separation of the information into two smaller vectors. This work is currently under development and will build upon the method presented in (Bortolato et al. 2020), where an autoencoder, namely PFRnet, separates the information contained in a FR feature vector into two sub-vectors that encode the identity representation and the demographic information. The architecture of PFRNet is illustrated in Figure 3.4a. PFRNet consists of a two-path encoder E and a single-path decoder D . E comprises two separate encoders E_{ind} and E_{dep} . The first encoder E_{ind} maps the face representation x (commonly generated by a CNN face recognition model, such as FaceNet, VGGFace, or VGGFace2) into a latent vector z_{ind} , which preserves identity cues, while greatly reducing the amount of information related to selected demographic attributes, such as gender. Thus, the latent representations z_{ind} can ideally be used in biometric systems for identity recognition without privacy-related concerns regarding the misuse of demographic information. The second encoder E_{dep} maps the original face representation x into a latent vector z_{dep} that encodes demographic attributes only. The

complete latent representation of x generated by the encoder E is a concatenation of the latent representations z_{ind} and z_{dep} , i.e., $z = z_{ind} \oplus z_{dep}$. This latent representation can in principle be reconstructed by the decoder function $D : z \rightarrow x$. In their paper, Bortolato et al. train PFRNet to suppress gender information, but in general the same concept can also be extended to other demographic attributes.

The properties of the subvectors z_{ind} and z_{dep} discussed above follow from the learning objective devised for PFRNet. The model makes use of three different loss functions. The first ensures a good reconstruction of the input data (reconstruction loss), the second suppresses gender information in z_{ind} and the third loss forces the distributions of z_{dep} for male and female subjects to be as different as possible. The second and third loss functions require attribute labels for the training data, while the first relies on self-supervision.

In XAIface the FR feature vectors will be extracted by the selected FR pipelines (ArcFace and MagFace) and ways to improve the soft-biometric disentanglement as well as to extract other types of demographic beyond gender will be investigated.

3.4.3. Performance Evaluation

Performance evaluation will be carried out by analysing and measuring the impact of demographics on face recognition in terms of loss/gain in performance, both on the whole dataset and on subsets of specific classes. Face recognition performances will be assessed using the metrics recommended in the ISO/IEC 19795-1, that is false match rate (FMR), false non-match rate (FNMR) for verification (one-to-one), and false-negative identification-error rate and false-positive identification-error rate for identification (one-to-many), as well as with graphical illustrations through the detection-error tradeoff (DET) plot.

Along with classical performance metrics, ad-hoc metrics will be adopted for specific tasks. For example, for assessing the efficiency of the disentangling method PRFNet (Bortolato et al. 2020), the authors propose the privacy-gain identity-loss coefficient (PIC). This metrics combines the performances for the following two tasks:

- for gender recognition, the fraction of incorrectly classified images (fic), and
- for face recognition, the equal error rates (eer), computed in face verification experiments.

$$PIC = \frac{fic' - fic}{fic} - \frac{eer' - eer}{eer}$$

Where fic' and eer' are computed from the attribute suppressed representations, whereas the errors fic and eer are computed from the original (unmodified) face feature vectors. Positive PIC values imply that the privacy gain is higher than the potential loss in face recognition performance and higher values indicate better performances.

3.4.4. Legal and Ethical Guidelines Concerned by the Method

Demographic information disentangling is concerned with several legal and ethical obligations required by the in depth-analysis from Deliverables 3.3 and 3.4. Such

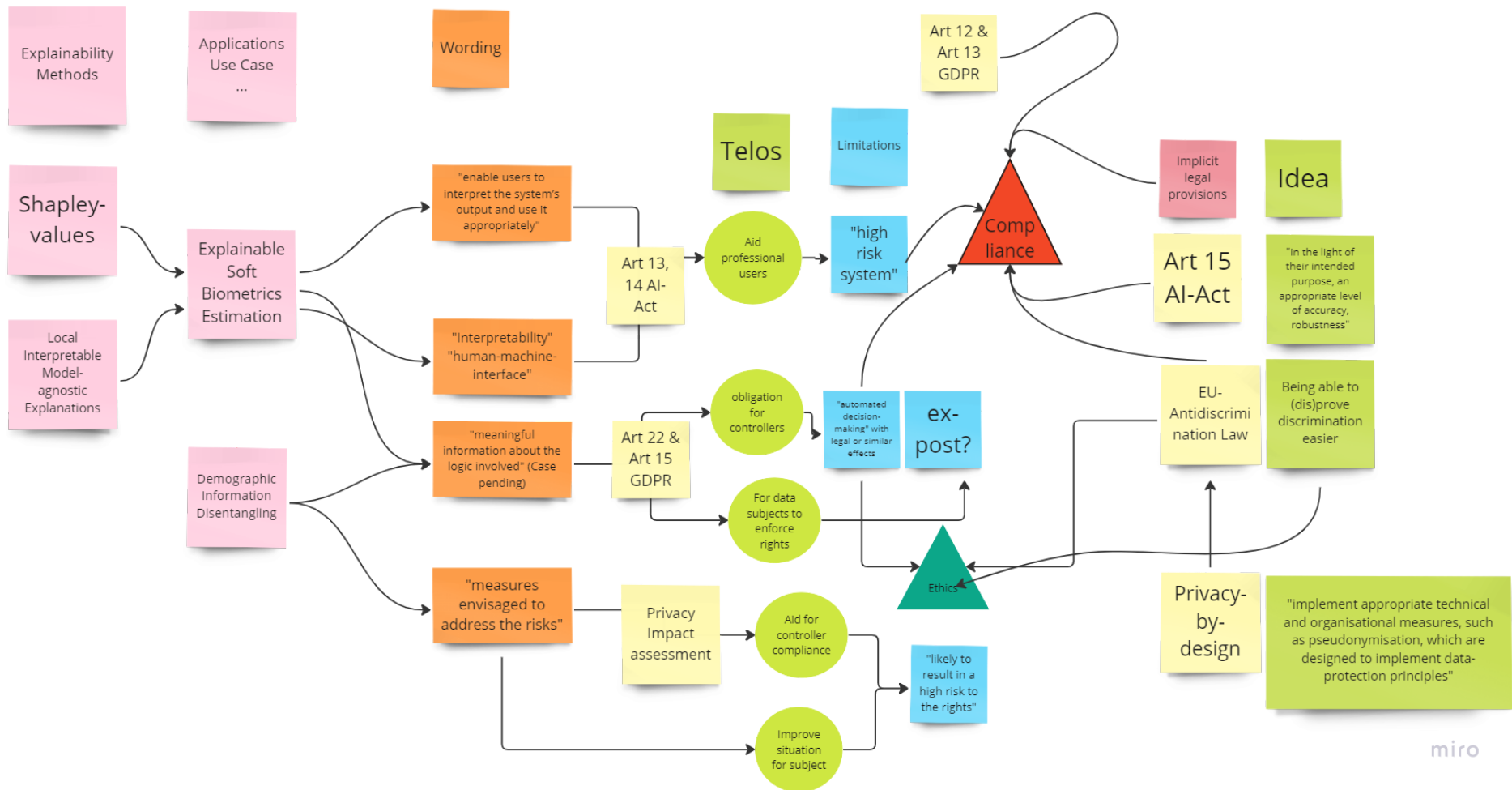
dependencies are illustrated in the graph shown in Figure 3.4b and identify the following wordings as the most relevant and applicable for our method to be developed:

- “Measures envisaged to address the risks”: as this approach aims at separating the identity information from other demographic information so that only the strictly needed information for a specific task is used, preventing the revealing of additional unnecessary information and in a way protect the user's privacy as much as possible while still allowing the recognition system to work efficiently.
- “Meaningful information about the logic involved”: this approach will also contribute in making the system’s logic more understandable as it will allow the user to better understand what part of the information is used.

We can thus conclude that from the dependencies in Figure 3.4b, the demographic information disentangling method is interested by articles 15 and 22 of the GDPR.

Finally, this method is especially relevant for the following numbered topics:

- A-13-2, A-14-1, and A-14-4 for AI-Act related recommendations;
- G-15-1 for GDPR related recommendations.



miro

Figure 3.4b: Demographic Information Disentangling and Explainable Soft Biometrics Estimation in the context of legal and ethical requirements concerned.

3.5. Explainable Soft Biometrics Estimation

3.5.1. Introduction

Human face images encode different types of biometric information. Soft biometrics such as gender, height, and weight do not have the capacity to differentiate between two different identities, however they can be useful to improve the quality of different systems. Among those, weight is also an indicator of both physical appearance and health conditions and unlike gender and height, body weight changes during a person's adult life and needs to be periodically measured. Conventional weight measurement techniques require the cooperation of the subject to be measured, which might not be possible during medical emergencies, video surveillance for criminal pursuit or due to different patient disabilities. When non-cooperative scenarios occur, visual estimation of the weight of the patient by a health professional is preferred but such estimations might not always be accurate.

Furthermore, face images are a rich source of personal and sensitive data that can be used to support a wide range of applications spanning from biometric recognition to user profiling. It is therefore essential that face images are adequately protected so that they cannot be misused ensuring their use exclusively for the target application. Individuals should be able to have access to privacy preserving systems that allow them to select the attributes to be kept and the attributes to be protected or suppressed from their face. Being able to understand which face image factors are important for a weight estimation model in opposition to a face recognition system, is a first step to performing weight anonymization, avoiding targeted advertisement in social networks that might lead to large-scale eating disorders.

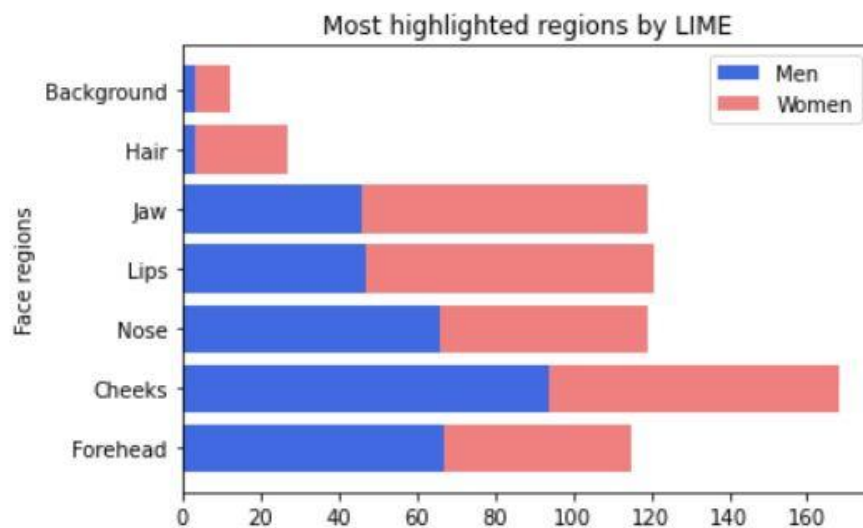
3.5.2. Description

Nowadays, deep learning technologies provide new solutions to obtain end-to-end learning models that gain knowledge and insights from complex, high-dimensional biomedical data. However, the general user might be still sceptical when facing black-box approaches in applications where model interpretability is a concern. Understanding the decision making of a predictor is crucial when actions are taken in the medical domain. Assessing trust in a model cannot be achieved only through accuracy metrics, a trustworthy model should be evaluated using interpretability techniques.

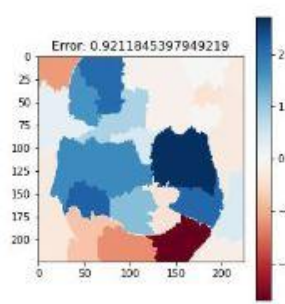
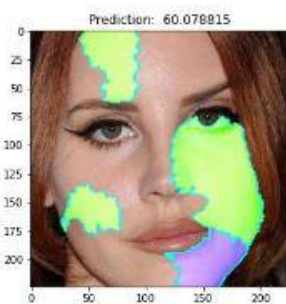
In this context, we aim to increase the trust of the user in soft biometrics estimation models, especially for one of the most discriminative as it is the weight. We intend to do it by complementing the prediction delivered by the deep learning network with a visualisation of the most contributive facial regions that lead to it. To this end, we explore two model-agnostic explainability techniques, SHAP and LIME. The interpretability techniques do not assess the validity of the result, instead, they give complementary information on which images areas were most significant for the prediction.

As a result of the regions highlighted by the interpretability approaches, we focus on assessing the impact of different occlusion factors on the weight estimation model. We explore the impact that hairstyle, and more particularly facial hair, has in the final weight prediction. We also evaluate different face detection and cropping techniques to assess whether different detection methods will consider more of the most meaningful areas highlighted by the interpretability approaches. As suggested by the explainability techniques, the occlusion or omission of those areas will lead to less accurate predictions, causing a misestimation of the soft biometric (in this case the weight) thus misleading the face recognition system. After those experiments, we are able to affirm that the optimal image pre-processing of image faces for weight estimation must take into account factors such as hair occlusions and the shape and size of face cropping bounding boxes in a different manner as usually done for face recognition tasks. Finally, we show how the most relevant face regions for estimating the weight differ by gender thus validating a gender-based model selection.

3.5.3. Performance evaluation



(a) Count of the most contributive regions after applying LIME.



(b) LIME example

(c) SHAP example

Figure 3.5a: Explainability approaches applied to the VIP_attribute dataset.

In Figure 3.5a we present representations of the output of LIME (b) and SHAP (c) when those explainability techniques are applied to the face images.

In (b), the highlighted green areas represent the image regions contributing to an increase of the weight and opposite to them, the purple ones constitute the portions decreasing the weight value. In (c), the red dots represent meaningful players for the game outcome. Nearly all the highlighted pixels for both methods lay in the face skin areas of the image, excluding regions such as background or eye pupil, reinforcing our model trust since meaningful parts of the image were taken into account.

We made a count of the most returned face areas by LIME across the test set and presented them in Figure 3.5a (a) for the male and female subjects. We observed how in all cases, the cheek area was the most highlighted feature. We also noticed how different areas were highlighted for males and females, as is the case of the jaw for women confirming that the model focuses on different face regions depending on gender.

Finally, we also counted the times that a non facial region was highlighted. The background was not often considered relevant information while the hair areas were wrongly taken into account in more cases.

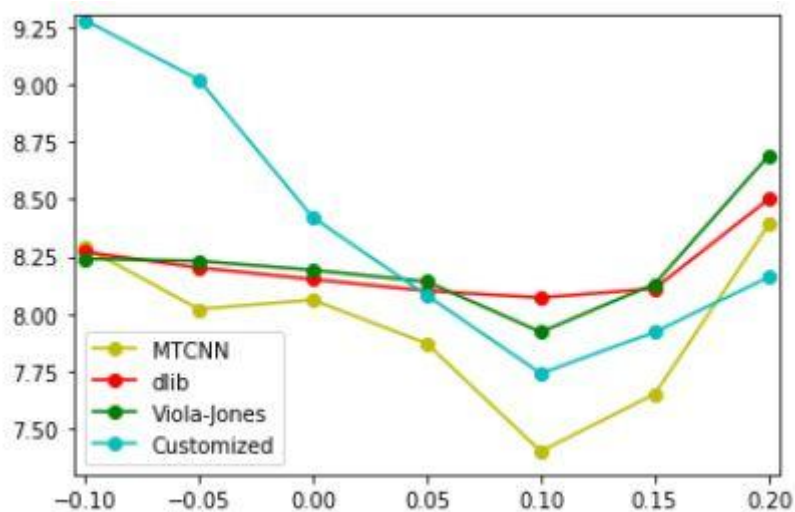


Figure 3.5b: MAE in kg of the VIP_attribute test set for various face detectors and cropping margins.

While assessing trust in the predictions through LIME, we discover that significant regions for our weight estimation model such as the face contour (forehead, cheeks and jaw) are usually excluded from face cropping algorithms since eyes contain the most meaningful information for face recognition. We evaluated whether different croppings, specially the ones considering larger face areas, will lead to a more accurate prediction as suggested by the explainability approaches. In our experiment, we have trained and tested our network for 4 different face cropping methods: Viola-Jones, Multi-Cascade CNN (MTCNN), the *dlib* python package and a customised cropping. We defined our face cropping by considering the highest, lowest, and furthest at the left and furthest at the right facial landmarks computed by the *dlib* landmark detector. The results presented in Figure 3.5b, represent the Mean Absolute Value (MAE) (y-axis) for different cropping margins (x-axis). It also highlights

the benefit of an increased margin, specifically of 0.1 (10% increase of the original bounding box), specially for the rectangular face croppings (MTCNN and customised) whose output is more adapted to the face shape. Nevertheless, large croppings include a high amount of hair and background regions increasing the network's MAE.

Hairstyle	# of subjects	MAE	MAPE
Bold - Short Bold	11	11.32	12.98
Short	96	8.73	11.12
Medium	26	5.47	8.00
Long - Long volume	72	6.12	8.92
Facial hair			
Clean	136	7.40	10.17
Moustache - Goatee	11	7.47	8.84
Beard	47	8.30	10.16
Fringe	11	6.10	9.28

Table 3.5a: MAE in kg of the VIP_attribute test set for various face detectors and cropping margins.

Another occlusion factor when considering a face image is hair. LIME highlighted in some cases hair areas as relevant as those can be present above the forehead and cheeks. We extended the annotation of the VIP_attribute database by adding for every subject annotations of their hairstyle, presence and type of facial hair and presence of glasses thus making possible further studies of those categories. Table 3.5a presents the MAE and Mean Absolute Percentage Error (MAPE) per category. Some categories such as "Bold-Short bold" are underrepresented so the high values can be due to the presence of outliers. But in the case of facial hair, we can observe how the presence of fringe as an occlusion is not as meaningful for the prediction as the presence of a beard which increases the error by more than 2 kg on average.

3.5.4. Legal and ethical guidelines concerned by the method

Explainable Soft Biometrics Estimation is concerned with several legal and ethical obligations required by the in depth-analysis from Deliverables 3.3 and 3.4. Such dependencies are illustrated in the graph shown in Figure 3.4b and identify the following wordings as the most relevant and applicable for our method:

- “enable users to interpret the system’s output and use it appropriately”;
- “interpretability”;
- “human-machine interface”;

- “meaningful information about the logic involved”.

As illustrated by the dependencies in Figure 3.4b, this method provides sensible information to be used as an aid for professional users and as such must be considered a “high risk system” causing compliance with several articles of the CFR and ECHR as well as articles 13-15 of the AI-Act and article 12, 13, 15 and 22 of the GDPR.

Finally, this method is especially relevant for the following numbered topics:

- A-13-1, A-15-1, A-13-2 and A14-3 for AI-Act related recommendations;
- G-15-1, G-14-1, G-14-2 for GDPR related recommendations.

3.6 Face Verification Explainability using Vision Transformer and EBM

3.6.1. Introduction

Modern deep-learning solutions often rely on the usage of transformers, also for computer vision tasks. In particular, the Vision Transformer (ViT) attains excellent results when compared to state-of-the-art convolutional networks, while requiring fewer computational resources for training (Dosovitskiy et al. 2021). The ViT splits an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard transformer encoder (Vaswani et al. 2017). The resulting embeddings can then be used for classification. Additionally, the ViT includes a series of attention heads, which already provide a way to visualise the most relevant input image regions that contribute to compute the embeddings describing the input image. In other words, the basic ViT already provides some degree of explainability.

ViT has been used for face recognition with simple modifications such as modifying the tokens generation method to consider overlapping image patches, for better description of the inter-patch information (Zhong et al. 2021). Using the Attention Rollout tool (Abnar et al. 2020), it is possible to create a heat map highlighting the face regions that contribute most to creating the embeddings that led to the face recognition decision. Attention rollout tracks the information propagated from the input layer to the embeddings in the higher layers.

Further developments in terms of explainability when using the ViT have been proposed, to go beyond attention visualisation. The method proposed in (Chefer et al. 2021) assigns local relevance scores and propagates them through the layers, handling the problems caused by the usage of non-positive activation functions, or the frequent use of skip connections in transformer models.

The work reported here started by exploring face verification explainability using vision transformers, and then, inspired by the work presented in Section 3.1, combining the usage of the ViT with the usage of explainable boosting machines (EBM), described in Section 2.7, to further increase the explainability. The possibility of applying multiple ViTs to different image patches, in combination with EBM, will also be explored.

3.6.2. Description

As described above, the idea behind the proposed explainability solution is combining the intrinsic explainability characteristics of Vision Transformers (ViT) and Explainable Boosting Machine (EBM), to create a face verification ante-hoc explainability tool.

The overall architecture of the proposed tool is shown in Figure 3.6a. In terms of face recognition, it can include a single ViT taking as input the complete face image, or include multiple ViT modules, each to process one of the input face patches. The resulting embeddings created by the ViT(s) are then passed to the EBM for classification.

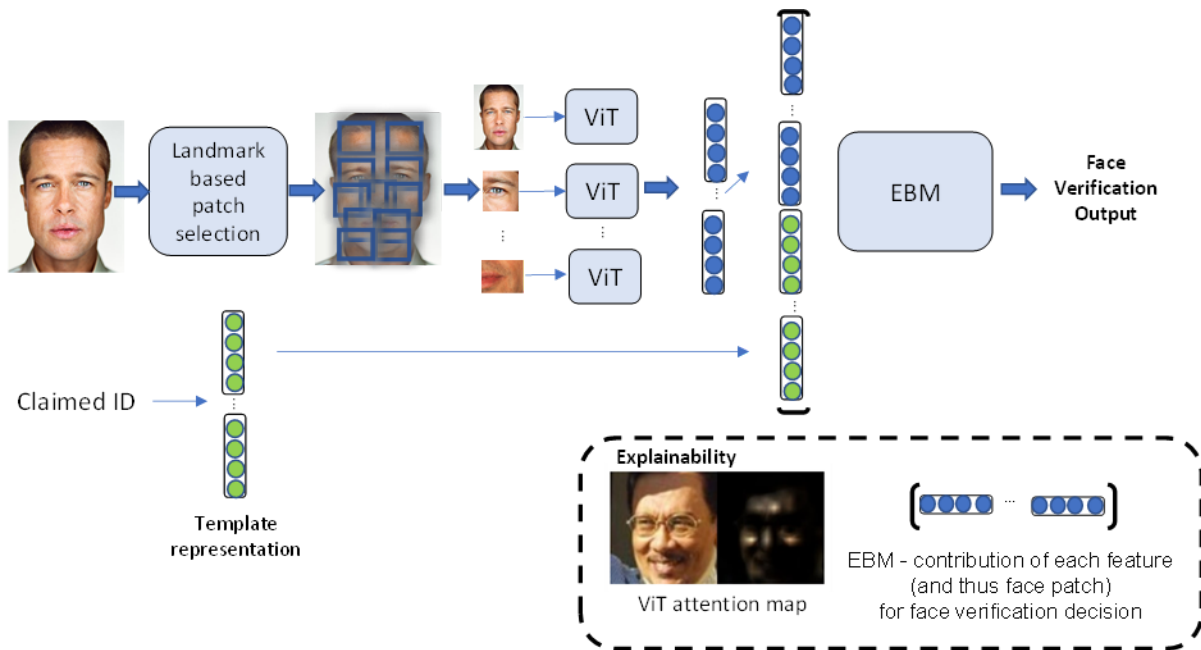


Figure 3.6a: Overall architecture of the proposed multiple ViT + EBM explainable face verification solution.

In terms of explainability, the goal is to combine the power of both explainability tools, i.e. ViT and EBM. For the ViT, the heat maps generated using both the rollout tool (Abnar et al. 2020) and the interpretability based on propagation tool (Chefer et al. 2021) will be evaluated. Then, since EBM produces relevance scores for each input feature, this will allow to gain further insights into the complete recognition model's behaviour. Notice that having access to the EBM feature relevance scores, this information can be propagated for each face embedding through the ViT, notably when considering the multiple ViT version, as illustrated in Figure 3.6b.

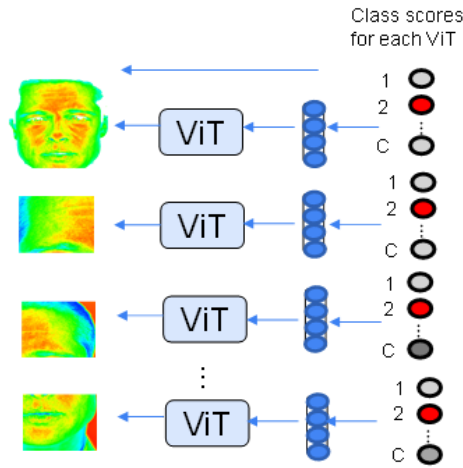


Figure 3.6b: Illustration of explainability architecture with the proposed multiple ViT + EBM face verification solution.

3.6.3. Performance evaluation

Since this work, inspired by the tool reported in Section 3.1, has started later, only preliminary results are available at the time of writing.

The initial results regard the explainability power of the selected visualisation tools that can be used to highlight the regions in the input face image that mostly contribute to the embeddings generated by the ViT. The examples included in Figure 3.6c show heat maps highlighting the most relevant regions contributing to the ViT embeddings for a true positive face verification attempt, using the interpretability based on the propagation tool (denoted as Chefer tool in the image), and the rollout tool. An example for a true negative face verification attempt is included in Figure 3.6d.

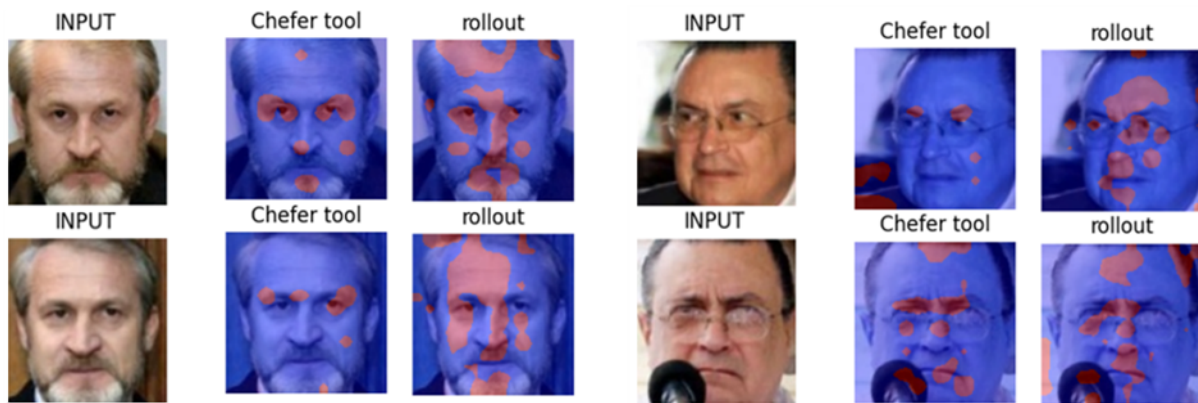


Figure 3.6c: Heat maps highlighting the most relevant regions contributing to the ViT embeddings computed for a true positive face verification attempt.



Figure 3.6d: Heat maps highlighting the most relevant regions contributing to the ViT embeddings computed for a true negative face verification attempt.

3.6.4. Legal and ethical guidelines concerned by the method

The proposed method for face verification explainability, which in a first step is based on using the vision Transformer (ViT) and subsequently will be combined with the explainable boosting machines (EBM) tackle several legal and ethical obligations, required by the in depth-analysis from deliverables 3.3 and 3.4. To identify and highlight these dependencies we make use of the wordings:

- “interpretability”,
- “human-machine interface”,
- “enable users to interpret the system’s output and use it appropriately” and
- “meaningful information about the logic involved”

as most relevant and applicable for our method. Hence we assume that the face verification explainability method using the ViT can be used as an aid for professional users, as this face recognition systems can be treated as “high risk systems”, causing compliance with several articles of the CFR and ECHR as well as articles 13-15 of the AI-Act and article 12, 13, 15 and 22 of the GDPR.

In addition the proposed method for face verification explainability can be linked against the explicit compliance efforts regarding AI-Act and GDPR elaborated and reported in chapter 4.4.1 and 4.4.2 respectively, being especially relevant for:

- AI-Act related recommendations: A-13-1, A-15-1, A-13-2 and A14-3;
- GDPR related recommendations: G-15-1, G-14-1, G-14-2.

4. Addressing Ethical and Legal issues

4.1. Explainable Approach Design and Implementation

According to the project proposal, the aim of WP 4 is to research “an approach to explaining decisions of face recognition systems, approach for explaining decisions of face recognition systems, building on information contributed from the different influencing factors in the form of metrics developed in WP2 and data/metadata obtained in WP3. The WP aims to design a protocol that is applicable to different state-of-the-art face detection and recognition pipelines that include AI components.”

The work package is separated into 3 tasks:

<p>T4.1</p>	<p>Explainability protocol and methods (M06 – M24: 1; 2, 3, 4, 5) This task researches and develops components of a face recognition pipeline producing explanatory information, addressing both the cases of creating transparent models and post-hoc explanation of black-box models. It also develops a protocol for implementing them in face recognition systems relying on AI components.</p>
<p>T4.2</p>	<p>Implementation of explainability approach (M06 – M24: 2; 1, 4, 5) The protocols and methods developed in T4.1 will be prototypically implemented for at least two different state of the art face recognition pipelines. As these pipelines will differ in which components are AI-based (e.g., face detection, feature extraction) and for which transparent models are used (e.g., for matching/classification), as well as in the network architecture (e.g., relying on common backbones vs. multitask networks for the specific problem), the implementation of the protocol will be specific for each of the methods. It is also expected that the trade-off between interpretability and performance will be different for each of the pipelines.</p>
<p>T4.3</p>	<p>Communicating explanations to users (M12 – M36: 1; 2, 3, 4, 5) This task analyses how information produced in T4.2 can be visualised or verbalised to be understandable to users. It will specify approaches to be integrated into user interfaces to summarise explanations for non-expert users and enable them to give feedback that is valuable for further training.</p>

Please note, that although the document focuses on T4.1 and T4.2 in particular, the analysis of ethical and legal aspects is also valid for the part regarding communication to end users, which will be described in another deliverable (D4.2). In order to avoid duplication of content and avoid redundancy, we perform the entire analysis in this document and will provide only a link (reference) and eventually some update notes there.

4.2. Scope

According to task 4.1, components shall be developed, which would produce additional **explanatory information**. Furthermore, the models should be **transparent**. The analysis in task 4.3 explores how such information can be visualised or verbalised. The aim is to deliver understandable information to users and summarise such information for non-expert users.

The major topics which must therefore be addressed within this work package are explainability and transparency obligations. These issues must be analysed from a legal and an ethical standpoint. While the legal analysis deals with the status quo of regulation as well as foreseeable changes in the law, the ethical perspective handles moral issues that may go beyond current regulation.

It should be noted that explainability and transparency requirements differ depending on the specific context in which a system is used. Recent EU-legislation has chosen a risk-based approach to regulation. As a result, the same system may have to adhere to different standards depending on the level of risk involved. Additionally, national legislation may add to those requirements. The starting point of the entire analysis are base requirements, which all or at least most systems must fulfil.

The two main questions, that shall be answered are the following:

- A. Is there a legal obligation to be transparent and to provide explanations for decisions of facial recognition systems?**
- B. Is there a legal right to obtain explanations of facial recognition systems?**

In a preliminary meeting of UNIVIE, the following additional subpoints were identified:

Legal Perspective:

- a) What is "explainability" and how is it linked to transparency?
- b) What is "transparency"?
- c) What transparency obligations does a controller have?
- d) To whom does the controller have an obligation to transparency?
- e) What rights relating to transparency does a data subject have?
- f) What information needs to be communicated to data subjects?
- g) How must the information be conveyed to data subjects?
- h) Must an AI-based face recognition system be explainable?
- i) If so, what information must be conveyed to data subjects?
- j) If so, how must the information be conveyed to data subjects?
- k) What additional requirements will (probably) be added by future EU-regulation?

Ethical Perspective:

- a) What are the ethical requirements concerning transparency and explainability according to the High-Level Expert Group on AI and similar authorities?

4.3. Legal Requirements

4.3.1. Explainability & transparency

For the legal analysis, the terms “explainability” and “transparency” must first be explored. Although classification varies depending on the model and taxonomy, transparency, alongside interpretability, can be seen as a subcategory of explainability.¹ Contrary to this view, the High-Level Expert Group on Artificial Intelligence categorised explainability as a subset of transparency.² In a study prepared for the members and staff of the European Parliament, “algorithmic transparency” was defined as follows:

“Depending on the type and use of an algorithmic decision system, the desire for algorithmic transparency may refer to one, or more of the following aspects: code, logic, model, goals (e.g. optimisation targets), decision variables, or some other aspect that is considered to provide insight into the way the algorithm performs. Algorithmic system transparency can be global, seeking insight into the system behaviour for any kind of input, or local, seeking to explain a specific input - output relationship.”³

The study describes transparency as a prerequisite for accountability.⁴ In order to ensure accountability, a certain amount of information must be provided. The authors of the document identified seven potential areas of transparency for machine learning systems:⁴

1. **Data:** refers to raw data, sources, pre-processing and collection methods;
2. **Algorithms:** refers to testing outputs against inputs;
3. **Goals:** refers to relative priorities of respective goals;
4. **Outcomes:** refers to the outcome of the actual deployment of a system;
5. **Compliance:** refers to the reports on overall compliance of an operator or manufacturer;
6. **Influence:** refers to revealing own interests or third party interests;
7. **Usage:** refers to what personal data is used.

Furthermore, transparency can be differentiated based on the addressees. A system can be transparent vis-à-vis everyone, authorities, third-party analysts, researchers⁵ or data subjects.

The term “explainability” still eludes legal scripture. Attempts to define the term have been undertaken⁶, but no definition has been agreed upon. However, the key points are captured by a similar debate on the “right to explanation”. *Wachter et alia* concisely defined the

¹ Waltl/Vogl, Explainable artificial intelligence – the new frontier in legal informatics, in Schweighofer/Kummer/Saarenpää/Schafer (Eds.) Data Protection/Legal Tech – Proceedings of the 21st International Legal Informatics Symposium IRIS 2018 (2018) 118.

² High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (2019) 18.

³ EPRS, study: A governance framework for algorithmic accountability and transparency (2019) 4.

⁴ EPRS, study: A governance framework for algorithmic accountability and transparency (2019) 5.

⁵ EPRS, study: A governance framework for algorithmic accountability and transparency (2019) 6.

⁶ Waltl/Vogl, Explainable artificial intelligence – the new frontier in legal informatics, in Schweighofer/Kummer/Saarenpää/Schafer (Eds.) Data Protection/Legal Tech – Proceedings of the 21st International Legal Informatics Symposium IRIS 2018 (2018) 117.

potential points an explanation to an automated decision would have to include⁷. According to the authors, the information can be separated into two categories: system functionality and specific decisions. While system functionality includes information about the logic of the system, significance, envisaged consequences, decision trees, pre-defined models, criteria and classification structure, the second category would include more information about the specific decision such as the rationale, the reasons, individual circumstances (weighting of features, profile groups etc.).

Wachter et alia further elaborated that explanations can be distinguished by their timing. An explanation can be ex ante and would therefore be limited to system functionality. Alternatively, an explanation can be provided ex post and could incorporate both categories. This structure has since guided the legal debate.⁸ The definition in the field of ethics will be addressed in the chapter on ethical requirements.

4.3.2. Current obligations according to the GDPR

General remarks

A key regulation that instituted obligations to transparency was the GDPR⁹. With the GDPR, a directive on the processing of data in the context of criminal law was created, which also contains provisions on transparency.¹⁰ Since this context is not a basic use case, the directive will only be mentioned for the sake of completeness.

The term “transparency” is not defined in the GDPR itself.¹¹ Recital 39 of the GDPR establishes that transparency is one of the principles of the GDPR and asserts that “it should be transparent to natural persons that personal data concerning them are collected, used, consulted or otherwise processed and to what extent the personal data are or will be processed”.¹² **From the excerpt one can already derive that transparency according to the GDPR is mostly an obligation vis-à-vis the data subjects.**

⁷ Wachter/Mittelstadt/Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law* (2017) Vol. 7/2, 76.

⁸ For further references: Kim/Routledge, Why a Right to an Explanation of Algorithmic Decision-Making Should Exist: A Trust-Based Approach (2020), available at: <http://dx.doi.org/10.2139/ssrn.3716519>.

⁹ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (referred to as GDPR).

¹⁰ Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA.

¹¹ Art 29 Working Party, Guidelines on transparency under Regulation 2016/679 – WP 260 rev.01 (2018) 6.

¹² Rec 39 GDPR.

Hence, the main objective of the provisions is to ensure accountability of data controllers and to empower them to exercise control over their data¹³.

According to the material scope, the obligations in the GDPR apply to the wholly or partly automated processing of personal data¹⁴. The creation and use of a face recognition system usually falls within the scope. The problem of which criteria must be fulfilled for data to constitute personal data will be addressed in another deliverable. It must however be noted that some relevant exemptions exist for the processing of certain competent authorities (criminal law) or organisations of the European Union¹⁵. This is due to specific regulations in these sectors.

The main addressee of the provisions of the GDPR is the “controller”. According to Art 4 (7) GDPR a “controller” means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law. Within the lifecycle of a face recognition system a typical controller would be the developer with regard to the data (images) used to train the models. Afterwards, the system may be used by an organisation, who may be considered the controller concerning the data used in the productive stage. Hence, these controllers need to fulfil their obligations towards their respective data subjects.

Art 5 GDPR established the key principles relating to processing of personal data. According to the article, personal data shall be “processed lawfully, fairly and in a transparent manner in relation to the data subject (‘lawfulness, fairness and transparency’)”.¹⁶ Art 5 par 2 GDPR highlights that it is not only the obligation of a controller to ensure compliance with the transparency obligation, but also to be able to demonstrate compliance.

Modalities of transparency

Section 1 of Chapter III of the GDPR specifically deals with transparency and modalities. Article 12 is titled “transparent information, communication and modalities for the exercise of the rights of the data subject.” The actual information that must be provided to data subjects is enlisted in Art 13 & 14 GDPR. According to paragraph 1 of Art 12 GDPR, this information must be provided “in a **concise, transparent, intelligible and easily accessible form**, using clear and plain language, in particular for any information addressed specifically to a child.”

The Art 29 Working Party states that “intelligible” means that an average member of the intended audience should be able to understand the information.¹⁷ If there is uncertainty of the intelligibility, it should be tested (p.ex.: readability testing). The standard applied to

¹³ Art 29 Working Party, Guidelines on transparency under Regulation 2016/679 – WP 260 rev.01 (2018) 5.

¹⁴ Art 2 GDPR.

¹⁵ Art 2 par 2 lit d, par 3 GDPR.

¹⁶ Art 5 par 1 lit a GDPR.

¹⁷ Art 29 Working Party, Guidelines on transparency under Regulation 2016/679 – WP 260 rev.01 (2018) 7.

controllers in relation to linguistic presentation is high.¹⁸ “Easily accessible” means that it should be obvious where the information is provided. It should not be the task of the data subject to seek out the relevant information.¹⁹ Common examples would be pop-ups and prominently displayed privacy notices.

Furthermore, the information must (generally) be either provided **in writing or other (electronic) means**. The complete information should be provided within one document. However, a layered approach is recommended.²⁰ Even though the obligations may result in some workload for the controller, the information must always be provided free of charge. The data subject must be informed at the time, when the data is obtained.²¹

Visualisation

The information may also be provided **in combination with so-called standardised icons**. The goal would be to aid visibility, legibility and to provide a meaningful overview over the processing. Such icons must be machine-readable, if they are presented electronically.²² According to the Art 29 Working Party, a variety of visualisation tools may be used: icons, certification mechanisms, data protection seals and marks.²³ However, visualisation tools may only be used in combination with and not as total replacement for language.²⁴ The GDPR requires the controller to make use of standardised icons to increase the utility. Since no decision could be reached on the matter, no annex with standardised icons was delivered with the GDPR. Rather, it is now the task of the European Data Protection Board and the European Commission to create a draft.²⁵ Recital 58 of the GDPR highlights that visualisation may be used in addition to language. However, for the specific information according to Art 13 & 14 GDPR only standardised icons may be used.²⁶

Information

Art 13 & 14 GDPR specify the information that must be provided to the data subjects. Both articles contain similar lists but differ in the scope. Art 13 GDPR is applicable if the data were obtained directly from the data subject. Otherwise, Art 14 GDPR applies. A typical scenario where Art 14 GDPR applies instead of Art 13 GDPR, would be the training of models with data from an available online database.

¹⁸ Schrey in Rücker/Kugler, New European General Data Protection Regulation (2018) 128.

¹⁹ Art 29 Working Party, Guidelines on transparency under Regulation 2016/679 – WP 260 rev.01 (2018) 7.

²⁰ For more information: Art 29 Working Party, Guidelines on transparency under Regulation 2016/679 – WP 260 rev.01 (2018) 11.

²¹ Arg.: Art 13 par 1 GDPR.

²² Art 12 par 7 GDPR.

²³ Art 29 Working Party, Guidelines on transparency under Regulation 2016/679 – WP 260 rev.01 (2018) 25.

²⁴ Franck in Gola, DS-GVO2 (2018) Art 12 cif 47.

²⁵ Franck in Gola, DS-GVO2 (2018) Art 12 cif 49.

²⁶ Veil in Gierschmann/Schlender/Stentzel/Veil (Eds.), Datenschutzgrundverordnung (2018) Art 12 cif 28.

The list of information is extensive, so only a portion of particularly relevant information will be listed below. For a practical overview in table form the authors refer to the annex of the [Guidelines on transparency under Regulation 2016/679](#) of the Article 29 Working party.

The controller must provide information about themselves and potentially their representative and data protection officer.²⁷ The data subject must be informed about the purposes of and legal basis for processing²⁸, storage time²⁹ and the individual rights of the data subject.³⁰

Furthermore, the controller must inform about the existence of automated decision-making, including modalities.³¹ Due to the importance of this specific topic and its inextricable link to explainability, it will be discussed in a separate paragraph below. Art 14 GDPR contains a slightly more extensive list. A controller must inform about the categories of personal data involved.³² Additionally, the source of the data must be revealed.³³

In addition to these obligations, required information may also include inter party communication on the exercise of the subjects' rights (Articles 15-22 GDPR) and communications in relation to data breaches (Article 34 GDPR).

Right to access and specific provisions on automated decision-making

As a counterpart to the obligation of the controller, a data subject has the right to obtain information about the processing activities. Such information includes, for example, the purposes of the processing³⁴ and the categories of personal data concerned.³⁵ Art 15 par 1 lit h also grants the right to data subject to be informed about the existence of automated decision-making, including profiling. The specific provisions on automated individual decision-making in the GDPR can be seen as the connecting piece between transparency, explainability. Art 22 GDPR primarily contains the right of any data subject to not be subject to a decision solely based on automated processing, including profiling. The data subject does not need to actively invoke the right for it to apply.³⁶ The right is restricted to cases, where the automated processing produces legal or similar effects for the data subject. The right does not apply if explicit consent is provided^[38], the decision is necessary for a contract³⁷ or the decision is authorised by the law of the controller and suitable safeguards for the data subject are provided.³⁸ In case one of the exceptions applies, the controller must inform the data subject about: “[...] the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about **the logic involved**, as well as the **significance and the envisaged**

²⁷ Art 13 par 1 lit a, b GDPR.

²⁸ Art 13 par 1 lit c GDPR.

²⁹ Art 13 par 2 lit a GDPR.

³⁰ Art 13 par 2 lit b, c, d GDPR.

³¹ Art 13 par 2 lit f GDPR.

³² Art 14 par 1 lit d GDPR.

³³ Art 14 par 2 lit f GDPR.

³⁴ Art 15 par 1 lit a GDPR.

³⁵ Art 15 par 1 lit b GDPR.

³⁶ Art 29 Working Group, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 – WP 251 rev.01 (2018) 19

³⁷ Art 22 par 2 lit c GDPR.

³⁸ Art 22 par 2 lit b GDPR.

consequences of such processing for the data subject.”³⁹ Hence, the question arises, what the exact telos of the norm is.

Scope

The scope of the provision is limited to specific scenarios. Despite the fact that profiling is mentioned in the title of Art 22 GDPR, it is not a necessary element for the application. The concept of profiling can be neglected for the purposes of this document, since facial recognition systems will rarely fall under the definition due to the lack of evaluation of personal aspects of the data subject.⁴⁰ The concept of “automated decision-making”, however, cannot be dismissed as quickly. Nevertheless, both concepts have a significant overlap.⁴¹ Art 22 GDPR only applies in cases of solely automated decision-making, which “[...] is the ability to make decisions by technological means without human involvement”.⁴² If a human is involved in the decision-making process, Art 22 GDPR usually does not apply, insofar as the contribution to the decision by the human is significant enough and not just formal.⁴³ Additionally the automated decision-making must have legal or similar effects for the data subject. Examples of legal effects mentioned by the Art 29 Working Group include refused admission to a country or denial of a particular social benefit granted by law. Also a person may be similarly affected if the decision leads to exclusion or discrimination.

In the case of a facial recognition system, the application of Art 22 GDPR depends on the context it is used in/for. The argument could be made that at least in case of a classification system as referred to in in-depth analysis of the European Parliamentary Research Service⁴⁴, a decision with potential legal or similar effects for the person is reached fully automatically. Other examples found in the study would be the use for crime prevention at train stations or crime investigations at the 2017 G 20 summit in Germany⁴⁵. Even though it is questionable, if one would have to include potential false positives, such as wrongfully being identified as a suspect, the authors would argue for such an interpretation. After all, the objective of the provision is to address the risk of automated processing. Even if the processing does not strictly fall within the scope of Art 22 GDPR, the Art 29 Working Group recommends fulfilling the additional transparency obligations.⁴⁶

Right to Explanation

As mentioned above, a controller employing automated decision-making must provide additional information to their data subjects:

- a) The fact that the controller is employing the technology

³⁹ Art 13 par 2 lit f, Art 14 par 2 lit g GDPR.

⁴⁰ Compare to Art 4 cif 4 GDPR.

⁴¹ Art 29 Working Group, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 – WP 251 rev.01 (2018) 8.

⁴² Ibid.

⁴³ Art 29 Working Group, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 – WP 251 rev.01 (2018) 21.

⁴⁴ Madiega/Mildebrath, Regulation facial recognition in the EU (2021) 2.

⁴⁵ Madiega/Mildebrath, Regulation facial recognition in the EU (2021) 37.

⁴⁶ Art 29 Working Group, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 – WP 251 rev.01 (2018) 25.

- b) Meaningful information about the logic involved
- c) Explanation of the significance and envisages consequences of the processing.

According to the guidelines on automated individual decision-making, it is the obligation of the controller to find simple ways to explain the “rationale behind, or the criteria relied on in reaching the decision”⁴⁷. Understanding the process can be particularly difficult if machine-learning is involved.⁴⁷

The language of Art 22 GDPR and Art 13 & 14 GDPR appears somewhat ambiguous even after the clarification of the Art 29 Working Group. This fact was noticed even before the GDPR came into force by various authors, who published essays questioning the right under the lens of a “right to explanation”.⁴⁸ It is not unreasonable to prima facie detect a right to explanation in those articles, especially if one considers the respective Recital 71: “In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to **obtain an explanation of the decision reached after such assessment** and to challenge the decision.” However, Recital 71 is not legally binding and the right to explanation is not explicitly mentioned in the safeguards of Art 22 par 3 GDPR.⁴⁹

According to the elaborations of *Wachter, Mittelstadt* and *Floridi*, a right to explanation would have to include not only information on system functionality such as the logic, significance (ex-ante) and consequences, but also information on the specific decision (ex-post) such as the reasons, weighting of features and individual circumstances.⁵⁰ Within their paper, the trio comes to the conclusion that, since no ex-post explanation of the specific decision must be provided, no complete right to explanation exists.⁵¹

For the purposes of this analysis, it must therefore be concluded that no right to explanation currently exists. However, controllers have certain obligations to be transparent about automated decision-making towards their data subjects.

Exceptions and opening clauses

Art 11 GDPR

⁴⁷ Ibid.

⁴⁸ Wachter/Mittelstadt/Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law* (2017) Vol. 7/2, 76; Mendoza/Bygrave, The Right not to be Subject to Automated Decisions based on Profiling, University of Oslo faculty of Law Legal Studies Research Paper Series No. 2017-20.

⁴⁹ See also Wachter/Mittelstadt/Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law* (2017) Vol. 7/2, 76 (80).

⁵⁰ Wachter/Mittelstadt/Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law* (2017) Vol. 7/2, 76 (78).

⁵¹ Wachter/Mittelstadt/Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law* (2017) Vol. 7/2, 76 (82).

Art 11 par 1 GDPR states: “If the purposes for which a controller processes personal data do not or do no longer require the identification of a data subject by the controller, the controller shall not be obliged to maintain, acquire or process additional information in order to identify the data subject for the sole purpose of complying with this Regulation.” This exemption may apply, for example, if a developer uses a database for the creation of data derived from images without knowledge of and interest in the identity of the data subjects. In such a case, no further data must be collected to identify the data subjects and to fulfil the obligations to transparency. However, if a data subject provides additional information about their identity in accordance with Art 11 par 2 GDPR, a controller must still comply with the relevant provisions. It should be noted that Art 11 GDPR does not apply if national law applies due to the use of an opening clause.⁵²

Art 89 GDPR

GDPR was the full harmonisation of data protection law, national legislators retained the right to regulate specific areas of data protection law. Art 89 GDPR contains an opening clause for research purposes and is therefore of particular interest. National legislation may contain specific provisions on transparency.⁵³

4.3.3. Future obligations according to the AI-Act

General Remarks

In 2021 the European Commission proposed the Artificial Intelligence Act (AI-Act)⁵⁴ in an effort to introduce harmonised rules on artificial intelligence systems. The act marks a significant new milestone in the process of regulation of new technologies. The main objective is to ensure that AI systems placed on the EU-market are safe and that fundamental rights and EU-values are respected. The legal certainty provided by the act is supposed to foster innovation and investment. AI made in Europe shall be safe, lawful and trustworthy. In essence, the proposal is a product regulation. It includes design requirements for AI systems as well as obligations for import and usage of such systems. Specifically, Art 13 & 14 AI-Act contain provisions on transparency and human oversight, which may increase the required level of interpretability of AI-systems. Interpretability, as mentioned above, can be seen as one aspect of explainability.⁵⁵ However, these provisions only apply to so-called “high risk systems”. It should be noted that the legislative process is extensive and changes to the text of the regulation are expected. In the current version⁵⁶, facial recognition systems, as systems that would potentially use biometric data⁵⁷, even occupy a

⁵² Kampert in Sydow, Europäische Datenschutzgrundverordnung 2 (2018) Art 11 cif 13.

⁵³ Details are provided in Deliverable 3.1.

⁵⁴ Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts (referred to as AI-Act).

⁵⁵ Waltl/Vogl, Explainable artificial intelligence – the new frontier in legal informatics, in Schweighofer/Kummer/Saarenpää/Schafer (Eds.) Data Protection/Legal Tech – Proceedings of the 21st International Legal Informatics Symposium IRIS 2018 (2018) 118.

⁵⁶ Status: 21/04/2021.

⁵⁷ Art 3 cif 33.

special position within the text. Naturally, the application of the specific provisions is dependent on factors such as intent and data used and cannot be decided categorically.

High-Risk Systems

The classification rules for high-risk AI systems are laid out in Art 6 of the proposed AI-Act. In essence, an AI system must be regarded as “high-risk” if it is either a safety component or a product which is covered by Union harmonisation legislation or is required to undergo a third-party conformity assessment. This, for example, includes sector specific regulation on machinery or medical devices.⁵⁸ Additionally, every system that is referred to in Annex III is a high-risk system. Annex III enlists specific areas, which would be deemed “high-risk” per se. The first area is of particular relevance for the project XAIface: “(a) AI systems intended to be used for the ‘real-time’ and ‘post’ remote biometric identification of natural persons;”⁵⁹

Nevertheless, not all facial recognition systems may fall within the provision. If they are used for other purposes than identification and Art 6 par 1 does not apply, the system may still be considered high-risk if it is used for specific intents in areas such as:

- Management and operation of critical infrastructure
- Education and vocational training
- Employment, workers management and access to self-employment
- Access to and enjoyment of essential private services and public services and benefits
- Law enforcement
- Migration, asylum and border control management
- Administration of justice and democratic processes;⁶⁰

These categories can be amended by the European Commission even after the regulation is in force.⁶¹ For the current version of the proposal including the annex, it can be assumed that a significant portion of facial recognition systems would be classified as “high-risk” and would therefore have to comply with the obligations set out in Chapter 2.

Requirements according to Chapter 2

Art 13 AI-Act establishes new requirements for transparency and provision of information to users. Users as referred to in this article are not necessarily the data subjects or end users, but rather professional users of an AI-system.⁶² Art 13 par 1 AI-Act requires high-risk AI systems to “[...] **be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system’s output and use it appropriately**”⁶³ Therefore, transparency and interpretability must be “by design” rather than being established post-hoc.

⁵⁸ See Annex II to AI-Act.

⁵⁹ § 1 lit a Annex III AI-Act.

⁶⁰ § 2 – 8 Annex III AI-Act.

⁶¹ Art 7 AI-Act.

⁶² Art 3 cif 4 AI-Act.

⁶³ Art 13 par 1 AI-Act.

Instructions

Additionally, a document with instructions must be provided to the users of the system.⁶⁴ As modalities Art 13 par 2 AI-Act requires that the information provided be in a digital or other format, concise, complete, correct, clear, relevant, accessible, and comprehensible to users. Similar to the respective provisions on transparency in the GDPR, it can be expected that the demanded standard for readability will be high.

The instructions mainly include information about the provider⁶⁵ and the system. The provider must list the characteristics, capabilities and limitations of performances of the system.⁶⁶ The system information includes its intended purpose, accuracy levels, robustness, cybersecurity, foreseeable misuse and risks, performance as regards the persons on which the system will be used and specifications for input data as well as relevant information on training, validation and testing. Importantly, the instructions must include so-called “**human oversight measures**” as referred to in Art 14 AI-Act. These measures include “[...] technical measures put in place **to facilitate the interpretation of the outputs of AI systems** by the users”.⁶⁷

Human Oversight

Art 14 par 1 of the AI-Act requires high-risk AI-systems to “[...] be designed and developed in such a way, including with appropriate **human-machine interface tools**, that they can be effectively overseen by natural persons during the period in which the AI system is in use.” The aim of the measures is to minimise risks and foreseeable misuse.⁶⁸

Whenever feasible, the measures shall be already built into the system by the provider before the product is placed on the market or put into service.⁶⁹ Otherwise the provider must identify the measures while they will be implemented by the user.

Art 14 par 4 then sets out the specific objectives such measures must enable a human overseer to achieve:

“(a) fully understand the capacities and limitations of the high-risk AI system and be able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible;

(b) remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (**‘automation bias’**), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons; (c) be able to **correctly interpret the high-risk AI system’s output**, taking into account in particular the characteristics of the system **and the interpretation tools and methods** available;

⁶⁴ Art 13 par 2 AI-Act.

⁶⁵ Art 13 par 3 lit a AI-Act.

⁶⁶ Art 13 par 3 lit b AI-Act.

⁶⁷ Art 13 par 3 lit d AI-Act.

⁶⁸ Art 14 par 2 AI-Act.

⁶⁹ Art 14 par 3 lit a AI-Act.

(d) be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system;

(e) be able to intervene in the operation of the high-risk AI system or interrupt the system through a “stop” button or a similar procedure.”

Specifically for AI systems intended to be used for remote biometric identification of natural persons, the measures must ensure that no action or decision is taken as a result of the systems identification without **verification and confirmation by at least two natural persons.**⁷⁰

For the purposes of this analysis, it must therefore be concluded that the new proposal of the AI-Act will significantly increase the requirements of interpretability and transparency of high-risk AI-systems by introducing new design and human oversight obligations. These obligations will require providers to technically enable the correct interpretation of outputs of AI systems. Such measures must include human-machine interfaces that will allow for better interpretability.

4.4. Compliance Efforts

UNIVIE has elaborated in great detail on the legal interpretation of the AI Act as well as the GDPR within the present, but also other deliverables. Great detail, however, also bears the problem of high information density. In this chapter, all basic requirements by the current draft of the AI Act as well as the GDPR will be reiterated, including the exact wording of the draft and the regulation, so it may be cross-referenced to specific solutions or proposed answers by D4.1.

4.4.1. AI-ACT⁷¹

Aid professional users (Art 13)	High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system’s output and use it appropriately. An appropriate type and degree of transparency shall be ensured, with a view to achieving compliance with the relevant obligations of the user and of the provider set out in Chapter 3 of this Title.	A-13-1
Aid professional users: "Interpretability" "human-machine-interface" (Art 15)	High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use.	A-15-1

⁷⁰ Art 14 par 5 AI-Act.

⁷¹ As per EC draft COM(2021) 206 final, 2021/0106(COD), see <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN>; for the wordings in the statement of the Council please refer to https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=consil%3AST_15698_2022_INIT.

Information to be made transparent according to Art 13 (3) “transparency”:	<ul style="list-style-type: none"> - Intended purpose; - Accuracy, robustness and cybersecurity level; - Reasonably foreseeable misuse causing risks to health, safety and fundamental rights; - Performance “as regards the persons or groups of persons on which the system is intended to be used”; - Information about training, validation, testing data sets and input data; - Human oversight measures. 	A-13-2
Aid professional users (Art 14)	[Users should be able to] fully understand the capacities and limitations of the high-risk AI system and be able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible.	A-14-1
Aid professional users (Art 14)	[Users should] remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (‘automation bias’), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons.	A-14-2
Interpretability (Art 14)	[Users should] be able to correctly interpret the high-risk AI system’s output, taking into account, in particular, the characteristics of the system and the interpretation tools and methods available;	A-14-3
Aid professional users (Art 14)	[Users should] be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system;	A-14-4
Aid professional users (Art 14)	[Users should] be able to intervene in the operation of the high-risk AI system or interrupt the system through a “stop” button or a similar procedure.	A-14-5
Accuracy/Robustness (Art 15)	The levels of accuracy and the relevant accuracy metrics of high-risk AI systems shall be declared in the accompanying instructions of use.	A-15-2
Accuracy/Robustness (Art 15)	High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way to ensure that possibly biased outputs due to outputs used as an input for future operations (‘feedback loops’) are duly addressed with appropriate mitigation measures.	A-15-3

4.4.2. GDPR⁷²

⁷² Art 22 as well as the rest of the official consolidated text of the GDPR can be accessed under <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679#d1e2838-1-1>.

Transparency (Art 15)	[The controller shall provide] meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.	G-15-1
Right to access (Art 15)	The controller shall provide a copy of the personal data undergoing processing.	G-15-2
Right to access (Art 15)	The right to obtain a copy referred to in paragraph 3 [Art 15] shall not adversely affect the rights and freedoms of others.	G-15-3
Transparency (Art 14)	[The controller shall provide meaningful information about] the existence of automated decision-making, including profiling, as well as the	G-14-1
Transparency (Art 14)	[The controller shall provide] meaningful information about the logic involved.	G-14-2
Transparency (Art 14)	[The controller shall provide meaningful information about the] significance and the envisaged consequences of such processing for the data subject.	G-14-3
Intervention of automated decision-making (Art 22)	The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her. [This shall not apply where it] is (a) necessary for entering into, or performance of, a contract between the data subject and a data controller; (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or (c) is based on the data subject's explicit consent.	G-22-1
Differentiation between "normal" and special categories of personal data (Art 22)	Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.	G-22-2

4.5. Ethical Requirements

4.5.1. HLEG-guidelines

The Ethics Guidelines for Trustworthy AI⁷³ are a central reference document for issues on AI. They promote trustworthy AI, which must be lawful, ethical, and robust.⁷⁴ The document, however, by its own definition does not provide guidance on lawful AI. Hence, these requirements were established in the chapter on legal requirements. The following section analyses the key requirements on transparency and explainability of the guidelines.

Transparency & Explainability

According to the guidelines, transparency of AI can be structured into three components: **traceability, explainability and communication**.⁷⁵ “Traceability” dictates that data sets and processes that yield a decision must be documented to increase transparency. Likewise, decisions of the AI systems should be traced to identify errors. “Explainability” is the ability to not only explain the technical process, but also the related human decisions. A human must be able to understand and trace the decisions. The guidelines assert that the person or persons concerned should have **a right to explanation**, whenever the AI system has a significant impact on their lives.⁷³ The explanation should be provided in a timely manner and suitable for the receiver. Furthermore, the guidelines demand transparency of the influences of AI-systems on organisation decision-making processes. “Communication” means that AI systems must not represent themselves as humans, but be transparent of their identity. Capabilities and limitations should be communicated appropriately to the parties involved.

The guidelines go into more detail on explanation methods in the section on technical methods. To be considered trustworthy, it must be understandable why a system behaved in a given manner or why it gave a certain interpretation.⁷⁶ The document states that “[...] to address this issue to better understand the system’s underlying mechanisms and find solutions” is still an open challenge for some AI systems and that XAI research is vital.⁷⁷ For the concrete assessment of explainability, the authors refer to the relevant section in the assessment list in the guidelines.⁷⁸

For the purposes of this analysis, it must therefore be concluded that a right to explanation is ethically required, whenever an AI system has a significant impact on a person or group of persons.

⁷³ High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (2019).

⁷⁴ High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (2019) 2.

⁷⁵ High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (2019) 18.

⁷⁶ High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (2019) 21.

⁷⁷ High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (2019) 21, 22.

⁷⁸ High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (2019) 29.

4.6. Preliminary Conclusion

The aim of this Section is to determine whether obligations to be transparent and to provide explanations for decisions of facial recognition systems exist. In turn, it was questioned whether there is a legal right to obtain explanations of facial recognition systems.

Having analysed the current state of EU-legislation, the authors reached the following conclusion:

- The GDPR contains specific provisions that oblige data controllers to be transparent about the data processing towards data subjects.
- Art 22 GDPR in conjunction with Art 13-15 GDPR only requires the controller to inform about the existence of automated decision-making, including profiling, and meaningful information about the logic involved, the significance and the envisaged consequences of the processing. Exceptions may apply.
- The articles do not provide for an obligation to ex-post explain specific decisions and therefore also do not contain a right to explanation in line with the current legal definition of such a right.

With regard to future EU-legislation, the authors reached the following conclusion:

- The proposal for the AI-Act in its current form will increase the level of transparency of high-risk AI systems, which will include a significant portion of facial recognition systems.
- The proposal for the AI-Act requires high-risk AI-systems to be designed in such a way, that their outputs are interpretable.
- Various technical tools for interpretability and human-machine interfaces will have to be provided.
- The document is still subject to change and results of the analysis are therefore only preliminary.

With regard to the Ethics Guidelines on AI, the authors reached the following conclusion:

- A right to explanation is ethically required, whenever an AI system has a significant impact on a person or group of persons.
- Providing solutions to make certain AI systems explainable is still an open challenge.

So, as a final remark, concerning the project XAIface, it must therefore be concluded that the efforts fall in line with the direction of future legislative projects. Even though not all methods may currently be legally required, there is an ethical obligation to ensure explainability. The concrete methods to allow for interpretability are not specified in the proposal of the AI-Act. Rather, they must be implemented or at least provided for by the provider of an AI system. The researched methods may be used to achieve compliance with future legislation. As insinuated in Art 13 and 14 of the proposal of the AI-Act, however, the choice of methods and interpretation tools will depend on the use case.

5. Summary

In this deliverable, the overview of the state of the art on AI explainability methods, focused on potential application for face recognition, is provided. This overview serves to give a necessary background for the development of new face explainability methods in the frame of the project.

In the overview (see Section 2), different categories of AI-explainability methods have been identified, including local and global methods, model specific and agnostic methods, white-box and black-box approaches, and gradient-based backpropagation as well as perturbation-based forward propagation algorithms. For each method, along with a brief description, we report which category it belongs to and what its possible use in XAIface might be. In addition, where possible, links to the implementation of the method are given. Many of the AI explainability methods in Section 2 are model-agnostic and applicable to black-box models. This allows a lot of flexibility in choosing the architecture of the face recognition model. However, a drawback of these methods is that their output can be very generic. Thus, one of the objectives of XAIface is to map novel, human understandable and reasonable (local) features.

Section 3 describes the methods developed in detail and reports basic performance evaluations of the new methods and modules for AI face explainability in XAIface. Furthermore, for each method there is a short paragraph describing the link to the preliminary assessment of potential legal and ethical issues described in more detail in Section 4.

Finally we want to mention again, that the main goals for the methods to be developed in the project are to: (i) develop AI explainability methods specific for face recognition; (ii) to provide meaningful and reasonable feedback for the end-users, (iii) to disentangle demographic information from identity to protect people's privacy; (iv) to understand what other information can be extracted from the face (soft biometrics) apart from identity.

As a conclusion of the legal and ethical requirements it turned out, that while there is an ethical requirement to make AI systems explainable, the current EU-legislation does not provide for a general right to explanation of or an obligation to explain decisions of facial recognition systems ex post. Depending on the context, data subjects may have a right to obtain ex-ante information such as the logic involved. Future legislation may change the status quo insofar as outputs of high-risk AI-systems must be interpretable.

References

- Abnar, S. and Zuidema, W., “Quantifying attention flow in transformers,” arXiv preprint arXiv:2005.00928, 2020
- Arbabzadah, Farhad, et al. “Identifying individual facial expressions by deconstructing a neural network.” *German Conference on Pattern Recognition, Springer*, 2016, pp. 344-354.
- Arras, Leila, et al. ““ What is relevant in a text document?”: An interpretable machine learning approach.” *PloS one*, vol. 12, no. 8, 2017.
- Auret, Lidia, and Chris Aldrich. “Interpretation of nonlinear relationships between process variables by use of random forests.” *Minerals Engineering*, vol. 35, 2012, pp. 27-42.
- Bach, Sebastian, et al. “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation.” *PloS One*, vol. 10, no. 7, 2015.
- Bulat A. and Tzimiropoulos G. , “How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks),” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1021–1030.
- Chattopadhyay, Aditya, et al. “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks.” *2018 IEEE winter conference on applications of computer vision (WACV)*, 2018, pp. 839-847.
- (Celis 2019) Celis, D. and Rao, M.; “Learning facial recognition biases through VAE latent representations;” in 1st International Workshop on Fairness; Accountability; and Transparency in MultiMedia; 2019; pp. 26–32.
- Chefer, H.; Gur, S. and Wolf, L.; “Transformer Interpretability Beyond Attention Visualization”; arXiv, 2021. doi: 10.48550/arxiv.2012.09838

- Chen, Tianqi, and Carlos Guestrin. "gboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794.
- Dantcheva, Antitza, et al. "What else does your biometric data reveal? A survey on soft biometrics." *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, 2015, pp. 441-467.
- Deng, Houtao. "Interpreting tree ensembles with inTrees." *International Journal of Data Science and Analytics*, vol. 7, 2019, pp. 277–287.
- Dhar, Prithviraj, et al. "How are attributes expressed in face DCNNs?" *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2020, pp. 85-92.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2021, doi: 10.48550/arxiv.2010.11929
- Grill, J. B.; Strub, F; Althé, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; Piot B.; Kavukcuoglu K.; Munos R. and Valko, M. ; "Bootstrap your own latent: A new approach to self-supervised Learning;" 2020; to do. [Online]. Available: <http://arxiv.org/abs/2006.07733>.
- Hastie, Trevor, and Robert Tibshirani. "Generalized additive models: some applications." *Journal of the American Statistical Association*, vol. 82, no. 398, 1987, pp. 371-386.
- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Proceedings of the 31st international conference on neural information processing systems*, 2017, pp. 4768-4777.
- Mirjalili, Vahid, et al. "Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers." *IEEE Access* 7, 2019, pp. 99735-99745.

- Nagpal, Shruti, et al. "Deep learning for face recognition: Pride or prejudiced?" *arXiv preprint arXiv:1904.01219*, 2019.
- Nori, Harsha, et al. "Interpretml: A unified framework for machine learning interpretability." *arXiv preprint arXiv:1909.09223*, 2019.
- Parde, Connor J., et al. "Face and image representation in deep cnn features." *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2017, pp. 673-680.
- Petsiuk, Vitali, et al. "Rise: Randomized input sampling for explanation of black-box models." *British Machine Vision Conference*, 2018.
- Ribeiro, Marco Tulio, et al. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144.
- Rozsa, Andras, et al. "Facial attributes: Accuracy and adversarial robustness." *Pattern Recognition Letters*, vol. 124, 2019, pp. 100-108.
- Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618-626.
- Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 618-626.
- Staniak, Mateusz, and Przemysław Biecek. "Explanations of Model Predictions with live and breakDown Packages." *The R Journal*, vol. 10, no. 2, 2018, pp. 395-409.
- Susesh, Harini; Gutttag, John. A framework for understanding sources of harm throughout the machine learning life cycle. En Equity and access in algorithms, mechanisms, and optimization. 2021. p. 1-9.

- Terhörst, Philipp, et al. "Suppressing gender and age in face templates using incremental variable elimination." *2019 IEEE International Conference on Biometrics (ICB)*, 2019, pp. 1-8.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L. and Polosukhin, I.; "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008
- Williford, Jonathan R., et al. "Explainable face recognition, Springer." 2020, pp. 248-263.
- Yin, Bangjie, et al. "Towards interpretable face recognition." *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9348-9357.
- Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *European Conference on Computer Vision, Springer*, 2014.
- Zhong Y.; and Deng, W.; "Face Transformer for Recognition", 2021, doi: 10.48550/arXiv.2103.14803
- Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921-2929.