

# Aligning Characteristic Descriptors with Images for Improved Interpretability

by

Bharat Chandra Yalavarthi

May 15, 2024

A thesis submitted to the  
Faculty of the Graduate School of  
the University at Buffalo, The State University of New York  
in partial fulfilment of the requirements for the  
degree of

Master of Science

Department of Computer Science and Engineering

Copyright by  
Bharat Chandra Yalavarthi  
2024  
All Rights Reserved

# Acknowledgments

I wish to express my sincere gratitude to Dr. Nalini Ratha for his invaluable guidance and support over the past one and a half years. I also thank Dr. Mingchen Gao for her feedback during my thesis defense. I would like to thank Tejas and Sunil for their inputs on this work. Finally, I am grateful to my family for their unwavering support and confidence in me, which empowers me to move forward.

# Abstract

AI advancements have spurred widespread adoption for various applications, but their black box nature makes understanding their predictions challenging. In mission-critical systems, such as face recognition for security/authentication or medical diagnosis, explainability/interpretability is crucial for user trust and effective use. In this work, we address the problem of explainability in deep networks using characteristic descriptors. We propose leveraging these descriptors to explain model decisions, using Vision Language Model such as CLIP to identify the presence of descriptors in an image and using that information to generate textual explanations. A concept bottleneck layer, which computes similarity between image and descriptor embeddings, is baked into the model architecture to provide inherent and faithful explanations. We apply the proposed method for face recognition and chest X-Ray diagnosis. Existing work on explainable face recognition focuses primarily on visual explanations not related to facial characteristic descriptors adopted in the forensic community, while in X-Ray diagnosis prior work is focused on explainability using saliency maps or report generation. By leveraging a curated set of facial descriptions used by forensic examiners for face recognition, and descriptions used by radiologist for X-Ray diagnosis we align the images with these textual representations using the CLIP model. Thus, our model offers a rationale akin to those made by human experts, while also presenting counterfactual instances to elucidate instances of failure. Moreover, our approach achieves comparable recognition/classification performance to that of black-box models on several well-known datasets while providing superior interpretable explanations. Thus, we believe our approach is a serious attempt to make deep learning systems accountable and transparent especially for gaining user trust in the problem domains of face recognition and X-Ray diagnosis.

# Table of Contents

<b>Abstract</b>	<b>iv</b>
<b>Table of Contents</b>	<b>v</b>
<b>List of Tables</b>	<b>vii</b>
<b>List of Figures</b>	<b>viii</b>
<b>Chapter 1:</b>	
<b>Introduction</b>	<b>1</b>
<b>Chapter 2:</b>	
<b>Explainability</b>	<b>3</b>
2.1 Types of Explainability Techniques . . . . .	3
2.1.1 Post-Hoc vs Intrinsic . . . . .	3
2.1.2 Local vs Global . . . . .	4
2.1.3 Model Specific vs Model Agnostic . . . . .	4
2.2 Modalities and Forms of Explanations . . . . .	5
2.3 Advantages of Explainability . . . . .	7
<b>Chapter 3:</b>	
<b>Explainable Face Recognition</b>	<b>10</b>
3.1 Prelude . . . . .	10
3.2 Related Work . . . . .	12

---

3.3	Methodology . . . . .	14
3.4	Experiments and Results . . . . .	19
3.4.1	Ablation Study . . . . .	21
3.4.2	Explanations . . . . .	22
 <b>Chapter 4:</b>		
<b>Explainable Chest X-Ray Diagnosis</b>		<b>27</b>
4.1	Prelude . . . . .	27
4.2	Related Work . . . . .	28
4.3	Methodology . . . . .	29
4.4	Results . . . . .	30
 <b>Chapter 5:</b>		
<b>Conclusion</b>		<b>33</b>
 <b>Bibliography</b>		 <b>34</b>

# List of Tables

3.1	CLIP Variants and Zero-Shot Accuracy on Face Verification.	
	.....	21
3.2	Ablation Study of CLIP Fine-Tuning methods.	
	.....	23
3.3	Ablation Study of Architecture Choices for Fine-Tuning.	
	.....	23
3.4	Ablation Study of our approach with various face recognition training methods.	
	.....	23
3.5	Performance (1:1 Verification Accuracy) on Five Benchmark Face Recognition Datasets.	
	.....	25
3.6	Counterfactual Examples explaining the changes in the predicted concept scores of the probe that would correct the model errors in Challenging Conditions.	
	.....	26
4.1	Comparison of predicted concepts with labels.	
	.....	31
4.2	Performance of black box model vs the proposed explainable model in detecting Pleural Effusion.	
	.....	31

# List of Figures

2.1	Post-Hoc: Analysis is made after the model is trained using input-output relations [4]. . . . .	4
2.2	Intrinsic: A new interpretable layer is added to the model through which explanations are extracted [4]. . . . .	4
2.3	Local vs Global Explanations. . . . .	5
2.4	Example of Natural language justifications for the model’s decision in terms of concepts. . . . .	6
2.5	Example of a saliency map highlighting the important regions for model’s decision [5]. . . . .	6
2.6	Example of logical explanations using boolean expression [6]. . . . .	7
2.7	Example of a Counterfactual Explanation where changing the features makes the model correct its decision [7]. . . . .	8
3.1	Our proposed explainable face recognition system. . . . .	12
3.2	Existing explainable face recognition methods. . . . .	14
3.3	Illustration of some of the facial characteristics descriptor features listed in FISWG Facial Image Comparison Guide. . . . .	15
3.4	An Overview of the Proposed Methodology to extract face embeddings for Explainable Face Recognition. . . . .	15
3.5	Proposed Explainable Face Verification Pipeline. . . . .	16
3.6	Illustration of Group SoftMax. . . . .	18
3.7	Ablation Study of AdaFace margin function parameters $m$ and $h$ . . . . .	24



---

3.8	Justifications (Explanations) provided by our model for its decisions. . . . .	24
4.1	Explainable methods in prior work for chest X-ray diagnosis. . . . .	28
4.2	Proposed explainable chest X-ray diagnosis system. . . . .	29
4.3	An example of extracting characteristic descriptors from radiologist reports.	30
4.4	An Overview of the Proposed Methodology for Explainable Chest X-ray Di- agnosis. . . . .	31
4.5	Examples of the explanations produced by the model for chest X-Ray Diagnosis.	32

# Chapter 1

## Introduction

Deep learning is a specialized field within Artificial Intelligence (AI) that involves processing information through deep artificial neural networks inspired by the structure and function of the human brain. Deep learning algorithms utilize multiple layers of interconnected nodes forming a network, to learn hierarchical representations of data. These networks excel at tasks like image and speech recognition, natural language processing, and autonomous decision-making, revolutionizing fields such as healthcare, security, finance, and transportation with their ability to extract complex patterns from vast amounts of data. Despite their effectiveness and robustness, AI systems face challenges such as bias in data leading to unfair outcomes and the lack of interpretability in models hindering trust and accountability. These issues underscore the critical need for transparency and fairness in AI algorithms, especially in sensitive domains like healthcare and criminal justice. Addressing bias through implementing explainable AI methods can help mitigate these challenges, fostering greater trust and acceptance of AI technologies in society. By ensuring that AI systems are not only accurate but also fair and interpretable, we can maximize their potential benefits while minimizing potential harms.

This thesis introduces methodologies aimed at enhancing the transparency and trustworthiness of deep learning models by furnishing them with explainability and interpretability mechanisms to elucidate their decision-making processes. We specifically focus on the problem domains of face recognition, and chest X-Ray diagnosis. Explainability is vital in

---

face recognition for accountability, transparency, and trust. In face recognition, it helps identify and address biases, ensuring accurate and ethical outcomes. Similarly, in medical diagnosis, it enables patients and healthcare providers to understand and trust diagnostic decisions, improving collaboration and patient care. Explainability or interpretability (used interchangeably in this thesis), refers to an AI model's capacity to clarify its decisions or outputs. It aims to demystify the black-box nature of AI models, offering users insights into the reasoning behind the system's conclusions. By providing transparent insights into the inner workings of deep learning models, these methodologies not only bolster user trust but also facilitate debugging, bias detection, and the mitigation of unintended consequences. Moreover, in domains such as healthcare, finance, access control, and law enforcement, where decisions have significant real-world implications, explainability becomes indispensable for ensuring accountability, fairness, and regulatory compliance. Through the integration of explainability mechanisms, stakeholders can better understand and validate the decisions made by AI systems, fostering greater acceptance and adoption of these technologies. Additionally, by shedding light on how AI models arrive at their predictions or classifications, explainability enables domain experts to collaborate more effectively with AI systems, leveraging their respective strengths to achieve optimal outcomes. Thus, the methodologies introduced in this thesis not only contribute to advancing the field of deep learning but also pave the way for more ethical and responsible deployment of AI in various practical applications.

# Chapter2

## Explainability

### 2.1 Types of Explainability Techniques

#### 2.1.1 Post-Hoc vs Intrinsic

**Post-Hoc Explainability** techniques focus on providing insights into the model's decisions after they have been trained. These methods, such as LIME (Local Interpretable Model-agnostic Explanations) [1] or SHAP (SHapley Additive exPlanations) [2], analyze individual predictions to uncover the underlying logic of complex models. These methods are generally model-agnostic and are useful for post-hoc analysis of a trained model.

**Intrinsic Explainability:** Unlike post-hoc methods, which analyze models after training, intrinsic explainability is built into the model architecture itself, making it easier to interpret and trust. For deep networks, it usually involves adding a new interpretable layer in the network to extract the explanations.

Intrinsic explainability is preferred over post-hoc methods [3] as explanations are directly aligned with the model's internal logic, enhancing trust and reducing the risk of misinterpretation or bias introduced by post-hoc techniques. This alignment with the model's internal logic allows for a more transparent understanding of the model's decision-making process. Figures 2.1 2.2 illustrates the difference between post-hoc and intrinsic explainability methods.

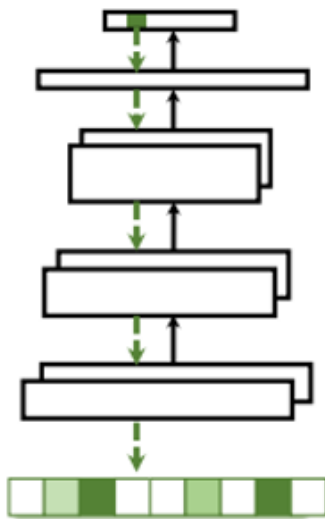


Figure 2.1: Post-Hoc: Analysis is made after the model is trained using input-output relations [4].

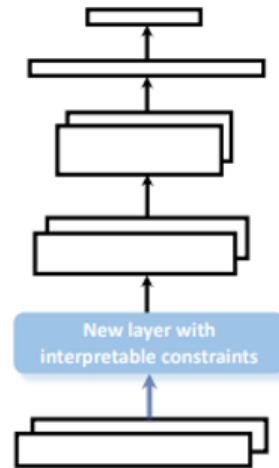


Figure 2.2: Intrinsic: A new interpretable layer is added to the model through which explanations are extracted [4].

### 2.1.2 Local vs Global

**Local explanations** provide insights into individual predictions made by machine learning models, offering context-specific reasoning for a particular instance. They aid in understanding model behavior at a granular level.

**Global explanations** provide an overview of a model's behavior across its entire input space, offering insights into its overall decision-making process. Unlike local explanations, which focus on individual predictions, global explanations highlight broader patterns and trends, aiding in understanding model behavior at a higher level.

### 2.1.3 Model Specific vs Model Agnostic

**Model Specific** techniques are tailored to the unique characteristics and architectures of particular models leveraging internal model structures and parameters to provide precise explanations directly aligned with the model's behavior.

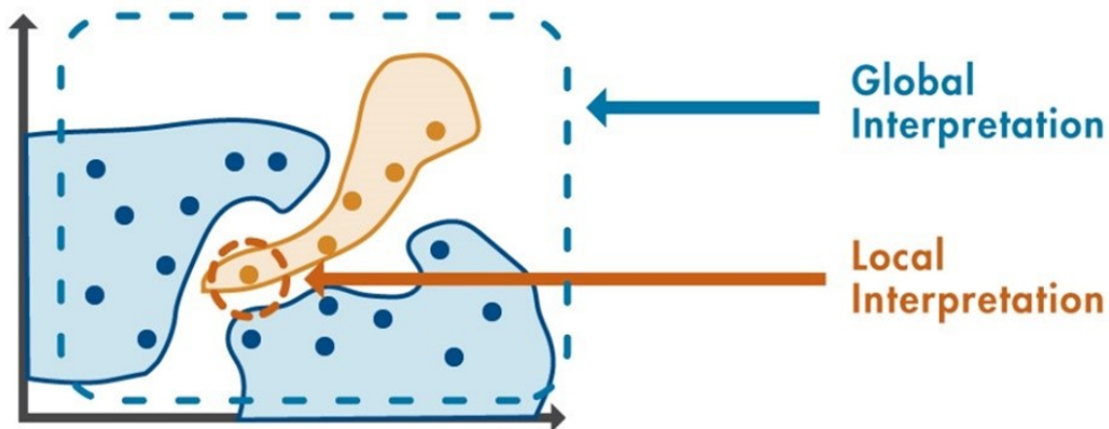


Figure 2.3: Local vs Global Explanations.

**Model Agnostic** techniques offer the flexibility to explain the behavior of various models without relying on specific model architectures by focusing on the model’s input-output relationship rather than its internal workings.

## 2.2 Modalities and Forms of Explanations

Modalities and Forms of Explanations encompass the diverse ways in which explanations are presented and conveyed. Modalities refer to the different mediums through which explanations are communicated, such as textual, visual, or interactive interfaces. Meanwhile, Forms of explanations encompass various styles or structures of explanations, including feature importance rankings, natural language justifications, or counterfactual examples etc.

**Textual explanations** like natural language justifications provide explanations in human-readable text, allowing users to understand the reasoning behind the model’s decisions in familiar language.

**Saliency Maps** highlight the most relevant features or regions in input data that influence a model’s decision-making process, providing visual insights into the model’s attention and reasoning.



Predicted: Match  
Actual: Match

**Similar Concepts in Reference and Probe (Top-5)**

- 1) Forehead height: Long
- 2) Face of a Male
- 3) Hair Color: Blonde
- 4) Brow ridges are Subtle
- 5) Forehead wrinkles are Present

Figure 2.4: Example of Natural language justifications for the model's decision in terms of concepts.



Figure 2.5: Example of a saliency map highlighting the important regions for model's decision [5].

**Logical Explanations** utilize Boolean expressions to articulate the decision-making process of the model, condensing the learned rules into Boolean logic to provide a clear and structured rationale for the model’s outputs.

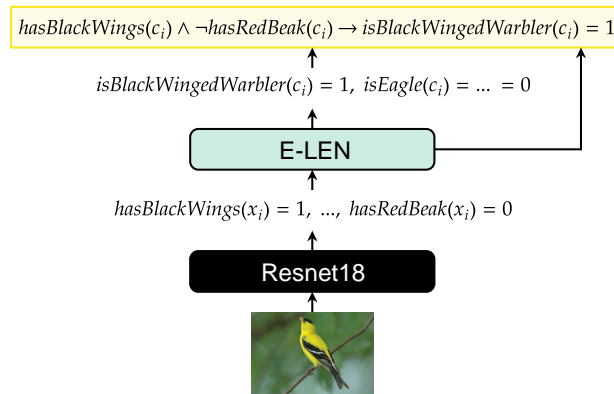


Figure 2.6: Example of logical explanations using boolean expression [6].

**Counterfactual Explanations** explore alternative scenarios by identifying minimal changes to embedding space that would result in different model outputs, offering insights into the model’s decision boundaries and potential biases. By highlighting these counterfactual instances, users gain a deeper understanding of how the model operates and can assess its robustness and bias.

## 2.3 Advantages of Explainability

The opacity of deep learning models poses a substantial hurdle in various sectors, especially those where trust and understanding are crucial, such as security and healthcare. Without transparency, these models operate as enigmatic black boxes, leaving users in the dark about how and why they arrive at certain conclusions. This lack of insight can hinder not only comprehension but also trust in AI-driven decisions, potentially impeding the adoption of these technologies in critical areas.

To overcome this challenge, explanations emerge as indispensable tools. They serve multifaceted purposes, playing a pivotal role in bridging the gap between users and machine



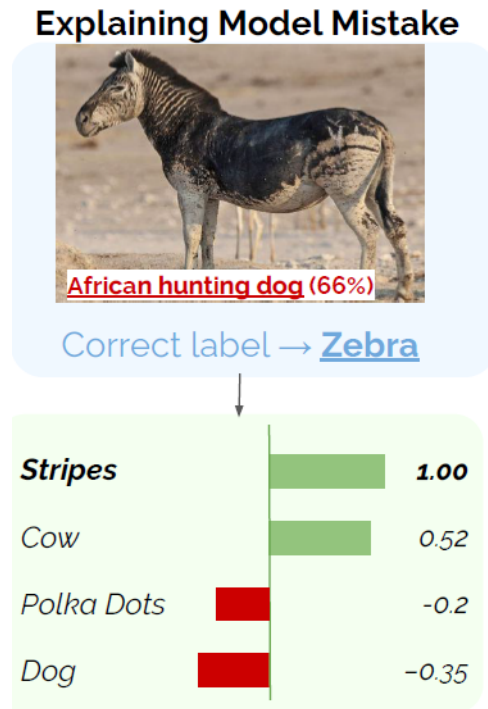


Figure 2.7: Example of a Counterfactual Explanation where changing the features makes the model correct its decision [7].

intelligence. Firstly, explanations provide clarity, offering insights into the inner workings of AI systems, thereby enhancing users' understanding and confidence in the technology. By shedding light on the decision-making process, explanations empower users to assess the reliability of AI outputs and make informed judgments based on solid reasoning.

Moreover, explanations serve as invaluable aids in debugging models and identifying potential biases. By dissecting the rationale behind AI decisions, researchers and developers can pinpoint errors, refine algorithms, and enhance the overall performance of deep learning systems. Additionally, explanations play a crucial role in mitigating biases, as they enable stakeholders to scrutinize and rectify any prejudicial tendencies encoded within the models. This proactive approach not only fosters fairness in decision-making processes but also promotes inclusiveness and equity in AI applications across diverse populations.

Furthermore, explanations are instrumental in upholding the safety and reliability of AI systems. By providing insights into the factors influencing decision-making, explanations

---

enable continuous monitoring and validation of model behavior. This proactive stance allows for early detection of anomalies or deviations from expected norms, facilitating timely interventions to prevent potential harm or misuse of AI technologies.

In essence, explanations serve as a linchpin for building trust, ensuring fairness, and bolstering the reliability of deep learning models. As AI continues to permeate various aspects of society, the imperative for transparent and interpretable AI becomes ever more pronounced. By embracing explanations as integral components of AI development and deployment, stakeholders can harness the full potential of these technologies while safeguarding against potential risks and pitfalls.

# Chapter 3

## Explainable Face Recognition

### 3.1 Prelude

Deep Learning models have truly transformed the landscape of face recognition, ushering in a new era of unprecedented accuracy, efficiency, and scalability. Yet, amidst these remarkable advancements lies a crucial deficiency: a lack of transparency. Users find themselves navigating in the shadows, unaware of the inner workings and decision-making processes of these sophisticated algorithms.

This opacity poses a significant challenge, particularly in domains where accountability and transparency are paramount, such as security applications [8]. Imagine a scenario where a face recognition system misidentifies an individual, potentially leading to wrongful accusations or security breaches. Without a clear understanding of how and why these decisions are made, addressing biases or rectifying failures becomes an uphill battle.

Legal contexts further underscore the necessity for transparent face recognition systems. In courtrooms, where the stakes are high, identity decisions must be supported by justifiable reasoning. Traditionally, facial forensic examiners have provided this critical analysis, offering insights into the validity and reliability of identification evidence. However, as technology advances, there's a growing gap between the traditional methods of human expertise and the opaque operations of Deep Learning models [9].

Numerous investigations into commercial face recognition systems have shed light on the inherent biases lurking within them [10]. These biases, often reflecting societal prejudices or flawed data inputs, can have profound real-world consequences. Take, for instance, the troubling case of a woman wrongfully arrested due to a flawed face recognition [11] match—a stark reminder of the urgent need for accountability and fairness in algorithmic decision-making.

The incorporation of interpretability or explainability mechanisms holds promise in mitigating such challenges by facilitating effective debugging processes and illuminating the underlying biases ingrained within the model.

Although there are several papers on explainable face recognition [8, 12, 13, 14] they focus on indistinct visual explanations which are less interpretable than textual explanations. Moreover, the existing method [5] that offers textual explanations necessitates the availability of labeled face attributes. Moreover, unlike our approach, the explanations provided by [5] is not similar to that of a forensic examiner. This requirement presents practical limitations and constrains the scope of explanation solely to labeled face attributes. Visual explanations can lack the nuanced detail and context often provided in textual explanations and may be subject to misinterpretation [9]. Our approach adopts a novel strategy that leverages precisely defined textual concepts to mimic the explanatory prowess of human experts. By employing a self-explainable architecture, we strive to ensure the fidelity and reliability of our explanations, thereby bridging the gap between machine-driven recognition and human-understandable justifications. This integration of textual elements not only enhances interpretability but also enriches the explanatory power of facial recognition systems, ultimately advancing their utility across various domains.

We make the following contributions in this paper:

- We propose a generic explainable face feature extractor that can be used with any of the existing State Of The Art (SOTA) face recognition techniques like AdaFace, ArcFace, etc.

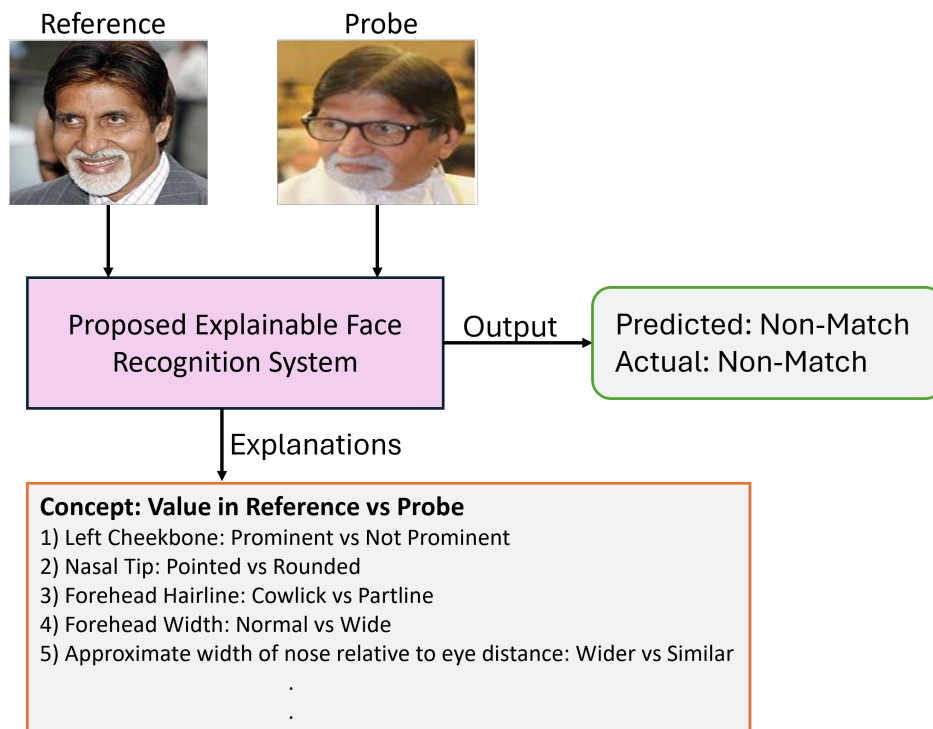


Figure 3.1: Our proposed explainable face recognition system.

- We provide coherent and user-friendly textual justification for face-matching based on concepts used by professional forensic experts with the faithfulness of the explanations as an inherent characteristic.
- We analyze the failure cases in challenging conditions for face recognition using counterfactual examples produced by our approach.

## 3.2 Related Work

The prior work related to our proposed approach can be broadly divided into two categories i) Explainable Face Recognition and ii) Vision Language Models (VLMs) for Explainability

Explainability methods have two main approaches i) Self-explainable models where explainability is embedded in the network architecture, ii) Post-hoc methods that try to explain decisions made by regular black box models using various approaches. Prior literature [15,

16, 17] has pointed out that self-explainable models generate more faithful explanations than post hoc methods.

Most of the prior work on explainable face recognition focused on generating visual explanations in the form of saliency maps depicting regions of the face the model focused on while making a decision. [18, 13, 19] identify these important regions through forms of occlusion or perturbation and how they affect the face-matching decision. Another approach in creating these saliency maps [8, 20, 14] [21] is identifying regions of the face pairs that are similar and dissimilar or identifying regions that lead to an imposter decision. [12] provides both patch-wise similarity of face images and attention weights indicating the importance of each patch in making a matching decision. Unlike the above works [5] provides both textual and visual explanations. It works by training separate networks to identify face attributes from images and finding important attributes for each match based on counterfactual examples. Unlike our approach, this requires labeled attributes for faces, and explanations are limited in precision and coverage by the available labels. Visual explanations have several disadvantages, they are not precise, fine-grained, and may be subject to interpretation. In human communications, an explanation response is usually in a textual medium either in spoken or written form as this can provide clear and concise explanations in most cases [9]. Observing these limitations, our method provides explanations through textual descriptions of facial features used by face forensic experts with a self-explainable architecture to ensure faithfulness. Explanations provided by some of the existing methods are shown in Figure 4.1

Pre-trained Vision Language Models (VLMs) like CLIP [22], and ALIGN [23] have shown good performance in image classification across several domains in zero-shot and fine-tuning settings [24, 25]. In a zero-shot setting, the similarity between a textual description of the class labels and an image is used for classification. Various fine-tuning techniques also exist including training a linear probe [25], tuning the image encoder or textual embeddings, weight ensembling of zero-shot and fine-tuned versions [24]. Several works [26, 27, 28, 29]

have also focused on improving the quality of textual descriptions using Large Language Models (LLMs) like GPT-4 to extract better image classification performance.

Recently there have been several works where VLM’s like CLIP were used to design explainable image classification models. [30, 31, 32] uses the concept bottleneck layer formed by aligning textual concepts and images for explainable image classification. Candidate concepts are usually generated by prompting the LLMs and undergo a selection process. We extend the usage of VLM’s for explainability to face recognition which has its challenges including meticulous description of facial concepts, intra-class variability, inter-class similarity, and low threshold for error due to its usage in critical applications.

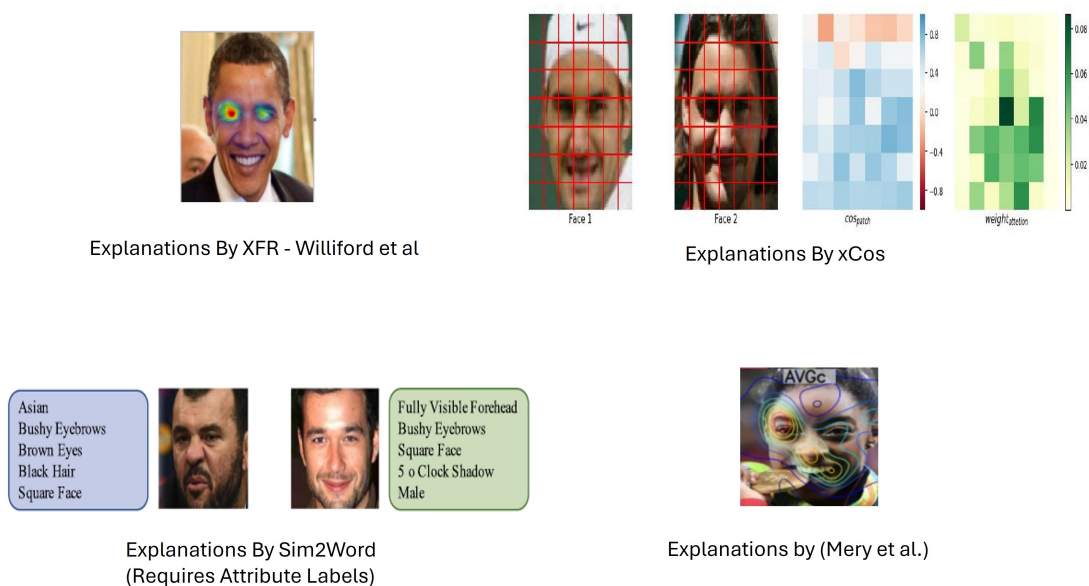


Figure 3.2: Existing explainable face recognition methods.

### 3.3 Methodology

Our proposed novel explainable face recognition methodology is based on definitive textual concepts or characteristic descriptors (used interchangeably in this report). These concepts are derived from the facial features standard published by Facial Identification Scientific Working Group (FISWG) [33] for morphological analysis in face comparison. This FISWG

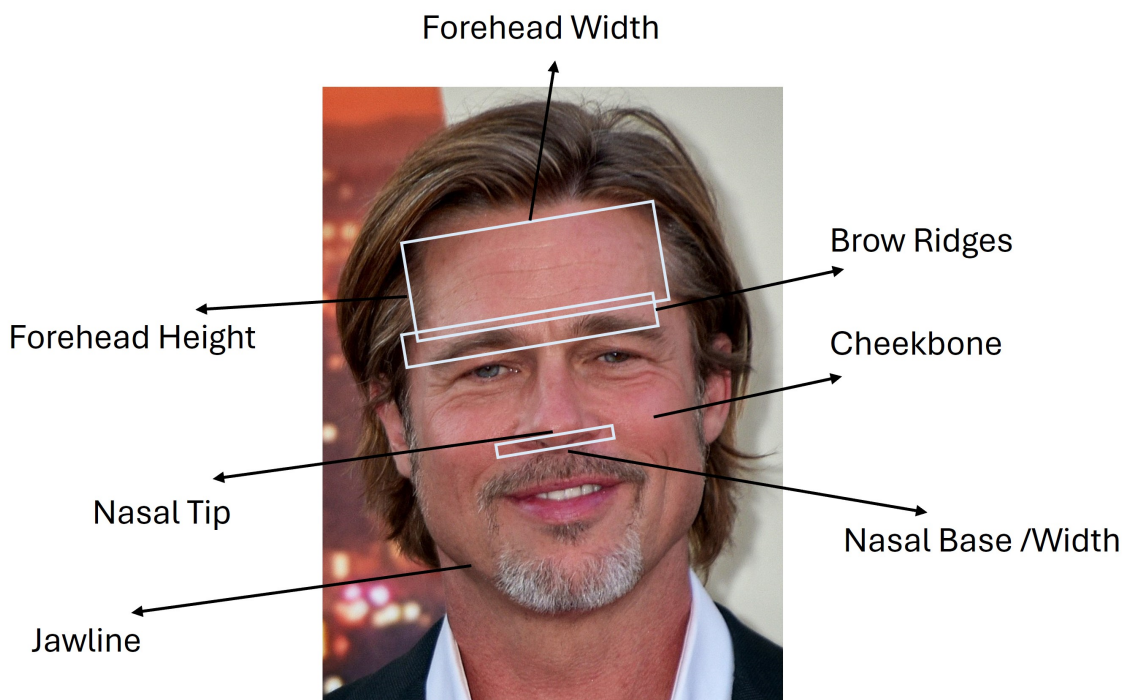


Figure 3.3: Illustration of some of the facial characteristics descriptor features listed in FISWG Facial Image Comparison Guide.

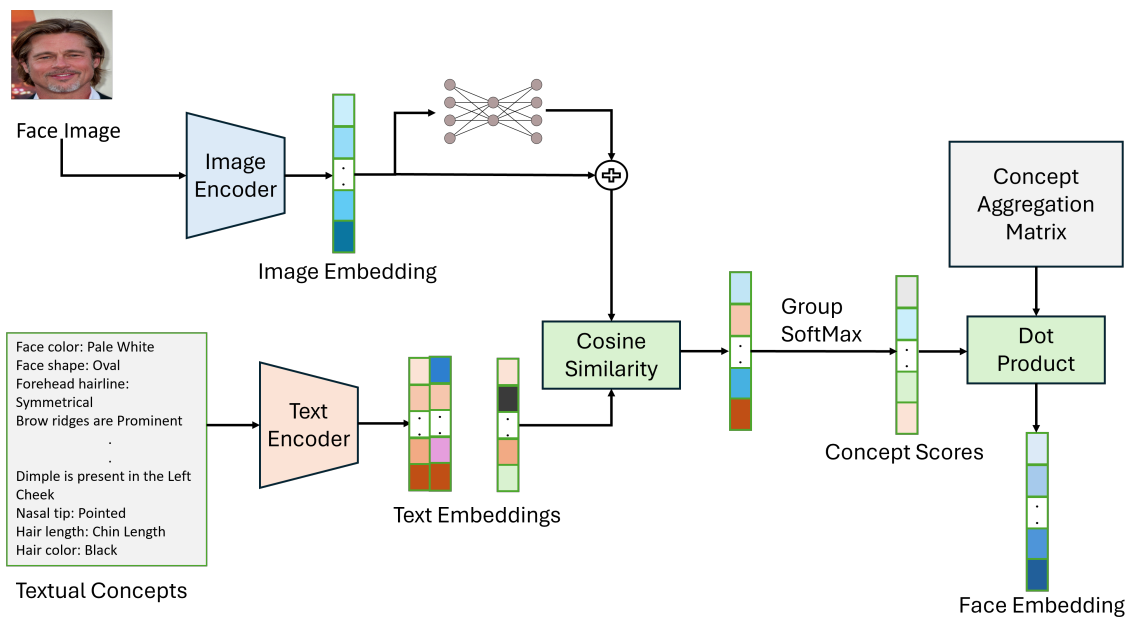
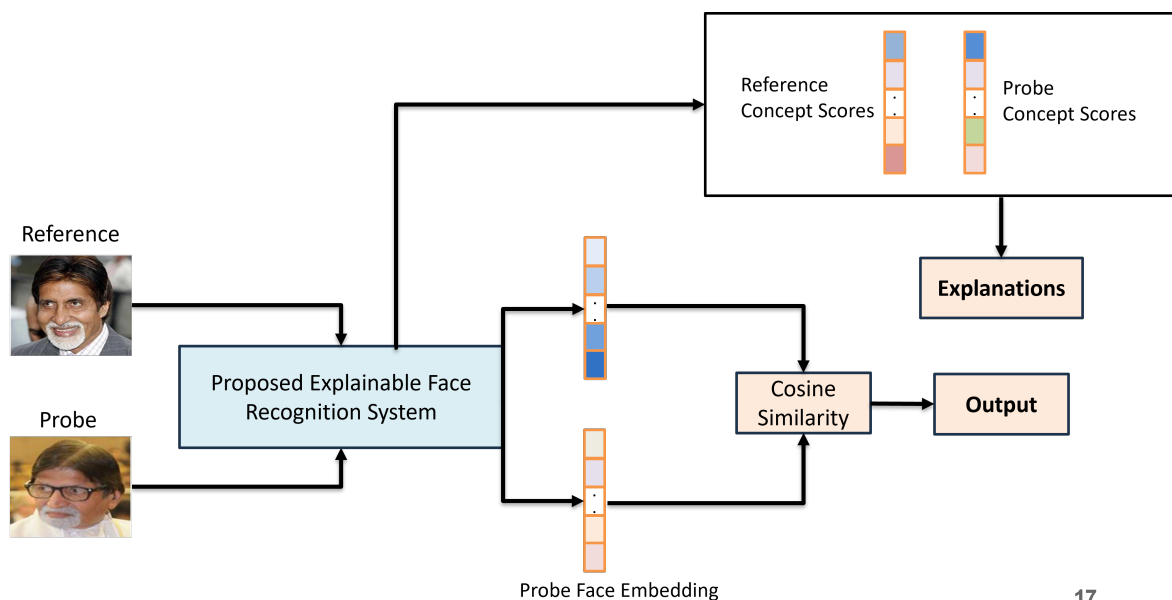


Figure 3.4: An Overview of the Proposed Methodology to extract face embeddings for Explainable Face Recognition.





17

Figure 3.5: Proposed Explainable Face Verification Pipeline.

guide listing the facial features to be used for face comparison and morphological analysis is commonly followed by forensic experts. It has detailed characteristic descriptors for each of the nineteen components in human faces. Figure 3.3 shows some examples of the kind of characteristic descriptors presented in [33]. Based on the face component characteristics described in [33] we handcraft a set of textual concepts denoted by  $C = \{c_1, c_2, \dots, c_n\}$  based on their adaptability with CLIP. These concepts are divided into subgroups based on the face components they are representing.

Inspired by the concept bottleneck models (CBM) [34] to design inherently explainable models we use a bottleneck layer in our proposed methodology. To overcome the requirement of labeled face concepts (attributes) to train CBM models we use CLIP [22] [35] to identify the textual concepts in the facial images without requiring labels. The similarity score between the textual and image embeddings produced by CLIP gives us the concept scores that form the bottleneck layer in our architecture. The concept scores of each image reveal the extent of a concept’s presence. Linearly transformed concept scores give the face embedding of the image which can be used for matching faces based on L2 distance. We used AdaFace’s [36] adaptive margin technique for fine-tuning CLIP image encoder for face recognition.

Let  $X \in R^{H*W*D}$  denote a face image,  $y$  its identity. We denote the image encoder module of the CLIP as  $E_i$  and the text encoder module as  $E_t$ . The dot product between the embeddings of  $E_i$  and  $E_t$  shows the match between the image and the text modalities. To fine-tune CLIP we use a skip connection to extract image embeddings as first detailed in [25]. Image embedding  $I \in R^d$  is extracted using the following equation:

$$I = \alpha * E_i(X) + (1 - \alpha) * F_i(E_i(X)) \quad (3.1)$$

where  $F_i$  is a two-layer network with ReLU activation which down-samples and up-samples the embedding back to its original size.

The tokenized concepts are fed into  $E_t$  to get the text embeddings  $T \in R^{N*d}$ . We compute the cosine similarity between  $I$  and  $T$  to get concept scores  $S \in R^N$  which represents the presence of concepts in the face image. To better represent these concept scores and the dependencies within a concept subgroup we apply SoftMax independently within each subgroup (denoted as Group SoftMax) of the concept set to obtain  $S_{sm}$ . We then transform the  $S_{sm}$  using a learned concept transformation matrix  $W \in R^{N*m}$  to get the final face embedding  $X_{emb} \in R^m$  used for matching. Since there are groups of these characteristic descriptors each defining a characteristic of the face, different groups can be fairly considered to be independent of each other. Directly applying SoftMax on the concept scores might not give a good representation, hence we tweak the regular SoftMax to address this issue using Group SoftMax. Figure 3.6 shows the illustration of how Group SoftMax works. Group SoftMax leads to a better representation of the embedding by highlighting the most activated concept within the subgroup aiding in improved accuracy and explainability as shown in the ablation study 3.3. Figure 3.4 shows the architecture of the proposed explainable model. Given a reference and probe face images, we get the face embeddings of these by passing it through the proposed architecture and compute if its a match or not based on their cosine similarity.

**Fine-Tuning:** The base CLIP model was fine-tuned to better recognize the face de-

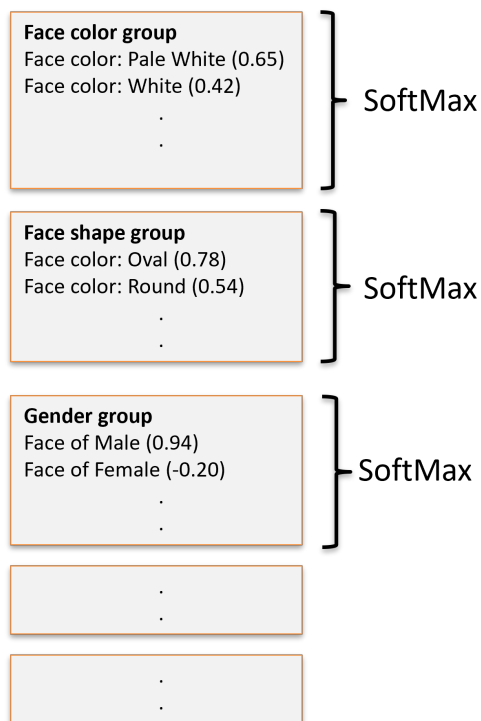


Figure 3.6: Illustration of Group SoftMax.

scriptions in the face images. As labeled data for face concepts is not available we can't use contrastive loss like the original CLIP model to adapt it to face recognition. As expected, experiments with end-to-end fine-tuning of either image encoder or text encoder have been shown to disrupt the capability of CLIP to align text and images. So we refrained from tuning the text encoder, and attention modules of the image encoder and only tuned the last fully connected layer of the image encoder. Additionally, we adopted the idea from [25] to add a fully connected layer on top of the image encoder and form a residual connection between their respective outputs (denoted as Adaptive FT). We used the quality adaptive margin loss function proposed in AdaFace [36] for fine-tuning the model.

**Counterfactual Explanation:** Explanation in the form of counterfactual examples is a useful tool for understanding the mistakes made by the model, debugging, and mitigating bias [7]. To generate counterfactual explanations we modify the concept scores  $S_{sm}$  to affect a change in the face embedding  $X_{emb}$ . We modify the concept scores of both probe and reference images to achieve a match for false negative or a non-match for false positive and

identify the concepts whose change in scores leads to the model correcting its decision. The algorithm to generate counterfactuals is shown given in algorithm 1 where  $C_r, C_p$  are the concept scores of reference, and probe images respectively while  $E_r$ , and  $E_p$  are their face embeddings. *thresh* is the threshold used to determine a match, and *match* is a boolean denoting the model decision. The algorithm takes these inputs for the errors made by the model and optimizes the concept scores of the probe until the prediction is corrected. We use elastic net regularization while optimizing concept scores to limit the number of concepts used in a counterfactual example for human interpretability.

### 3.4 Experiments and Results

We evaluate our approach on several standard face recognition datasets on performance, generate justifications for predictions made by the model, and counterfactual explanations for incorrect predictions.

**Datasets:** MS1MV2 [37] dataset containing 5.8M images with 85K identities was used for fine-tuning our model. Standard face recognition datasets including LFW [38], CFP-FP [39], AgeDB [40], CPLFW [41], CALFW [42] are used for validation. In all the ablation studies we report the average accuracy on these five datasets. In addition to the standard LFW dataset, we used more challenging datasets like AgeDB, and CALFW where there is the age difference between the faces being matched, and also CPLFW, and CFP-FP where the face poses are different.

Table 3.5 shows the 1:1 verification performance of our proposed approach on these five datasets.

**CLIP Model:** As there are multiple variants of CLIP differing in the architecture and training dataset we evaluate various CLIP variants based on their Zero-shot performance in 1:1 verification task. We report the average accuracy on the five validation datasets of each variant in table 3.1. Based on these results we chose ViT-L/14 trained on LAION-2B

---

**Algorithm 1** Generate Counterfactual Examples

---

**function** GEN\_CF( $C_r, C_p, E_r, E_p, thresh, match$ ) $C1 \leftarrow C_r$  $C2 \leftarrow C_p$  $dist \leftarrow \|E_r - E_p\|_2$ **if**  $match == \text{True}$  **then**  **while**  $dist > thresh$  **do**     $E_p \leftarrow \text{Linear}(C_p)$     ▷ *Linear* - Linear layer     $dist = \|E_r - E_p\|_2$     **if**  $dist > thresh$  **then**

break

**end if**     $L \leftarrow -MSE(E_r, E_p) + \|C2 - C_p\|_1 + 0.1 * \|C2 - C_p\|_2$      $C_p \leftarrow SGD(L, C_p)$     ▷ Optimizes  $C_p$  based on  $L$   **end while****end if****if**  $match == \text{False}$  **then**  **while**  $dist < thresh$  **do**     $E_p \leftarrow \text{Linear}(C_p)$      $dist = \|E_r - E_p\|_2$     **if**  $dist < thresh$  **then**

break

**end if**     $L \leftarrow MSE(E_r, E_p) + \|C2 - C_p\|_1 + 0.1 * \|C2 - C_p\|_2$      $C_p \leftarrow SGD(L, C_p)$   **end while****end if**  **return**  $C_p$ **end function**

---

dataset variant for all further experiments. Interestingly, it outperformed other variants with larger parameters such as ViT-H-14-quickgelu and ViT-H-14-378-quickgelu, and also a variant with the same architecture trained on a different dataset.

**Training Settings:** We take the cropped and aligned MS1MV2 images and resize them to the shape of 224 x 224 to make it suitable for CLIP’s image encoder. We fine-tune the model for 5 epochs using AdamW optimizer with a learning rate of 0.0003 as in [24]. We use the same hyper-parameter values for margin  $m$  and image quality indicator concentration  $h$  as in Adaface [36].

Table 3.1: CLIP Variants and Zero-Shot Accuracy on Face Verification.

<b>Variant</b>	<b>Training Dataset</b>	<b>Accuracy</b>
ViT-B/16	DataComp-1B	68.72%
ViT-L/14	OpenAI’s WIT	71.21%
ViT-H-14-378-quickgelu	dfn5b	72.29%
ViT-H-14-quickgelu	dfn5b	72.87%
<b>ViT-L/14</b>	<b>LAION-2B</b>	<b>74.11%</b>

### 3.4.1 Ablation Study

**Fine-Tuning:** For the sake of explainability, it is important to fine-tune the CLIP backbone without affecting its text-image alignment capability as faithfulness and validity of the explanations are directly dependent on it. Prior works have shown strong performance of CLIP in zero-shot image classification settings owing to its strong capability to align text and images. We use the zero-shot performance of our fine-tuned CLIP on CIFAR-10 as a proxy for its alignment capability to find the right level of fine-tuning. Our goal in this experiment is to fine-tune to improve the face recognition accuracy without affecting the alignment capability. As shown in table (Table: 3.2) fine-tuning the entire image encoder part of CLIP has led to the best face recognition accuracy of all the experimented cases, but as evident from the alignment proxy accuracy it has lost its ability to align text and images

rendering it unusable for explanations. Weight ensemble fine-tuning which was proposed in [24] interpolates the weights of zero-shot and end-to-end fine-tuned models to bring the generality of zero-shot model to fine-tuned one. Although this provided improvement in the alignment but was not able to bring it to the levels of zero-shot model. Further experiments have proved that fine-tuning using Adaptive-FT as proposed in [25] or fine-tuning the last Multi-Layer Perceptron (MLP) of the CLIP image encoder showed improvements in face recognition while not losing significantly on the alignment. Moreover, combining both these techniques was shown to be the best balance between face recognition performance and text-image alignment.

**Face Recognition Training:** We have experimented with existing state-of-the-art methods of AdaFace [36], ArcFace [37], and CosFace [43], for training face recognition models to find the best suitable one to fine-tune our model whose results are shown in Table 3.4. Figure 3.7 shows the ablation study for finding the best AdaFace parameter’s margin  $m$ , and image quality indicator concentration  $h$ . Varying slightly from the AdaFace’s original parameter values we find that for our setting  $m=0.5$ , and  $h=0.33$  was the best performing combination.

### 3.4.2 Explanations

We show the ability of our approach to produce faithful explanations justifying its face matching decisions. Next, we also show the utility of our approach to debug the failure cases in challenging conditions induced due to factors like a)Pose, b)Occlusion, c)Illumination, d)Image Quality, e)Age, f)Expression. Through counterfactual explanations we try to get a deeper understanding of the errors made by the face recognition model in these challenging conditions. Figure 3.8 shows the examples of the model justifying its decision based on important concepts. The top-5 closest concepts of reference and probe which are commonly activated in each concept group are shown for matching cases, while the top-5 concept groups where the activated concepts differ the most between the reference and probe are shown for

non-match cases.

Table 3.2: Ablation Study of CLIP Fine-Tuning methods.

<b>Fine-Tuned Model</b>	<b>Face Recognition Accuracy</b>	<b>Alignment Proxy Accuracy</b>
CLIP Zero-Shot (No FT)	74.11%	96.91%
Entire Image Encoder FT	94.80%	12.59%
Weight Ensemble FT [24]	76.72%	36.16%
Adaptive FT [25]	83.50%	96.82%
Image Encoder MLP FT	83.72%	93.59%
<b>Image Encoder MLP + Adaptive FT</b>	<b>89.50%</b>	<b>94.27%</b>

Table 3.3: Ablation Study of Architecture Choices for Fine-Tuning.

<b>Fine-Tuned Model</b>	<b>Face Recognition Accuracy</b>
CLIP Zero-Shot (No FT)	74.11%
CLIP Zero-Shot (No FT) + Group SoftMax	76.91%
CLIP Zero-Shot (No FT) + Group SoftMax + Linear Layer	76.91%
CLIP Image-Encoder MLP FT + Group SoftMax + Linear Layer	83.72%
<b>CLIP Image-Encoder MLP FT + Group SoftMax + Linear Layer + Adaptive FT</b>	<b>89.50%</b>

Table 3.4: Ablation Study of our approach with various face recognition training methods.

<b>Method</b>	<b>Accuracy</b>
CosFace (m=0.35)	80.14%
ArcFace (m=0.50)	83.24%
<b>AdaFace (m=0.5)</b>	<b>89.50%</b>

**Counterfactual Example Analysis:** To show the debugging capabilities of our approach, we analyze the model errors on the evaluation datasets using the counterfactual examples. By analyzing these examples against the images we can understand the underlying biases, and spurious correlations, and precisely point out where the model is going wrong. We try to identify those concepts in counterfactual examples which required changes in their concept scores the most number of times as these concept scores were either underestimated



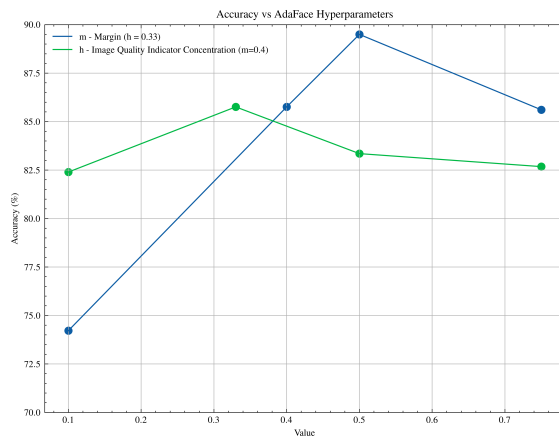


Figure 3.7: Ablation Study of AdaFace margin function parameters  $m$  and  $h$ .

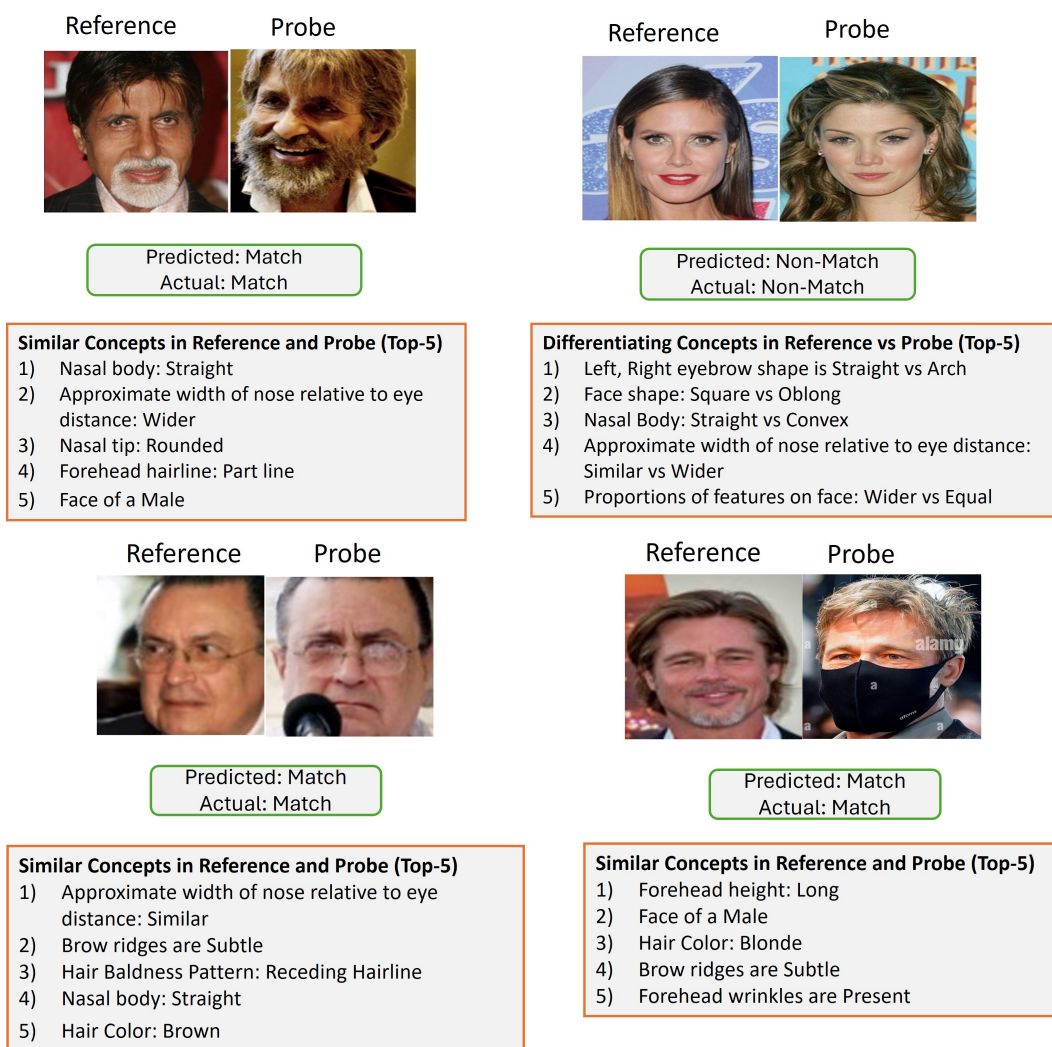


Figure 3.8: Justifications (Explanations) provided by our model for its decisions.



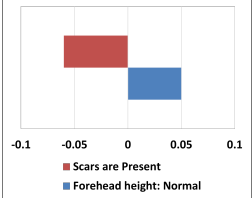


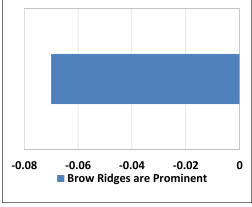
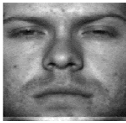
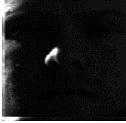
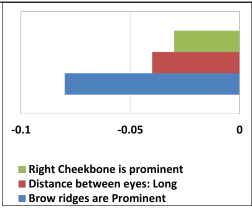
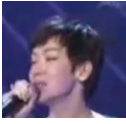

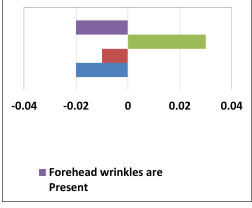


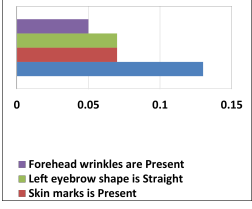


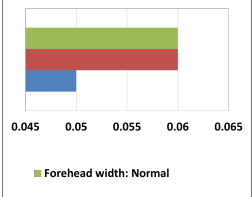
Table 3.5: Performance (1:1 Verification Accuracy) on Five Benchmark Face Recognition Datasets.

Dataset	Accuracy (Proposed Explainable Approach)	Accuracy (Black Box - SOTA)
LFW [38]	98.38%	99.82%
CFP-FP [39]	90.88%	98.49%
CALFW [42]	89.36%	96.08%
AgeDB [40]	81.18%	98.05
CPLFW [41]	87.71%	93.13%
Average	89.50%	97.11%

or overestimated by the model. For example, in the AgeDB [40] dataset where there are more images of older people we identify that the model overestimates the presence of scars (in 13% of the samples wrongly predicted), piercing (in 12%), and skin marks (in 10%) on the face. Inspection of these images suggests that this is caused by the presence of wrinkles in old people’s faces, where the model is wrongly correlating wrinkles to the presence of scars, or piercing, or skin marks.

Another such analysis on the CFP-FP dataset[39] which has a mix of frontal and profile views of the face showed that concepts related to brow-ridges, and cheekbones were changed the most in counterfactual examples. This suggests the ability of the model to better detect concepts related to the prominence of brow-ridges and cheekbones in profile view and not as good in frontal view. Similarly for CPLFW dataset [41] which has variation in face poses model found forehead wrinkles and the size of nasal base better in frontal poses than in other poses.

Table 3.6: Counterfactual Examples explaining the changes in the predicted concept scores of the probe that would correct the model errors in Challenging Conditions.

Condition	Model Decision	Reference	Probe	Incorrectly Predicted Concepts (Correction Required to get correct match decisions)
Pose	Non-Match			
Occlusion	Non-Match			
Illumination	Non-Match			
Image Quality	Non-Match			
Age Difference	Match			
Expression	Non-Match			

# Chapter 4

## Explainable Chest X-Ray Diagnosis

### 4.1 Prelude

Deep learning models have improved the effectiveness of automated chest X-ray diagnosis, offering significant improvements in accuracy, and scalability. However, they often operate as opaque black boxes, lacking transparency in their decision-making processes. This lack of transparency poses a critical challenge in medical contexts where accountability and understanding of model decisions are essential for ensuring patient safety and trust in the healthcare system. Incorporating interpretability or explainability mechanisms holds promise in addressing these challenges by enabling effective debugging processes and uncovering underlying biases within the model, ultimately enhancing the reliability and accountability of AI chest X-ray diagnostic systems.

Several prior works dealt with explaining the chest X-ray diagnosis made by deep learning models, none have explored using fine-grained, and atomic characteristic descriptors for providing explanations. The prior work can be mainly categorized into the explainable X-ray report generation [44] [45], saliency maps for the X-ray images to depict the important regions for the diagnosis [46], and explaining the alignment of image and text [47]. Contrary to these in our approach, we use characteristic descriptors to provide textual, and atomic explanations similar to a human expert.

## 4.2 Related Work

Saliency maps was a common method used by many prior works to explain the decisions made by a model. [48] [49] uses saliency maps for explanation while also using features extracted from the report to improve the disease classification performance. [46] proposes a method to provide explanations by modeling the gaze of a radiologist when he is examining a chest X-Ray. On the other hand, [44] provides explainable report generation capabilities by linking the regions of the image to the parts of the report. [50] uses longitudinal representations to provide interpretable and controllable report generation. [51] provides a novel methodology to use the explanations produced by the model to improve the classification performance. Prior work also addresses the problem of explaining the alignment of images and text modalities in vision language models like [47].

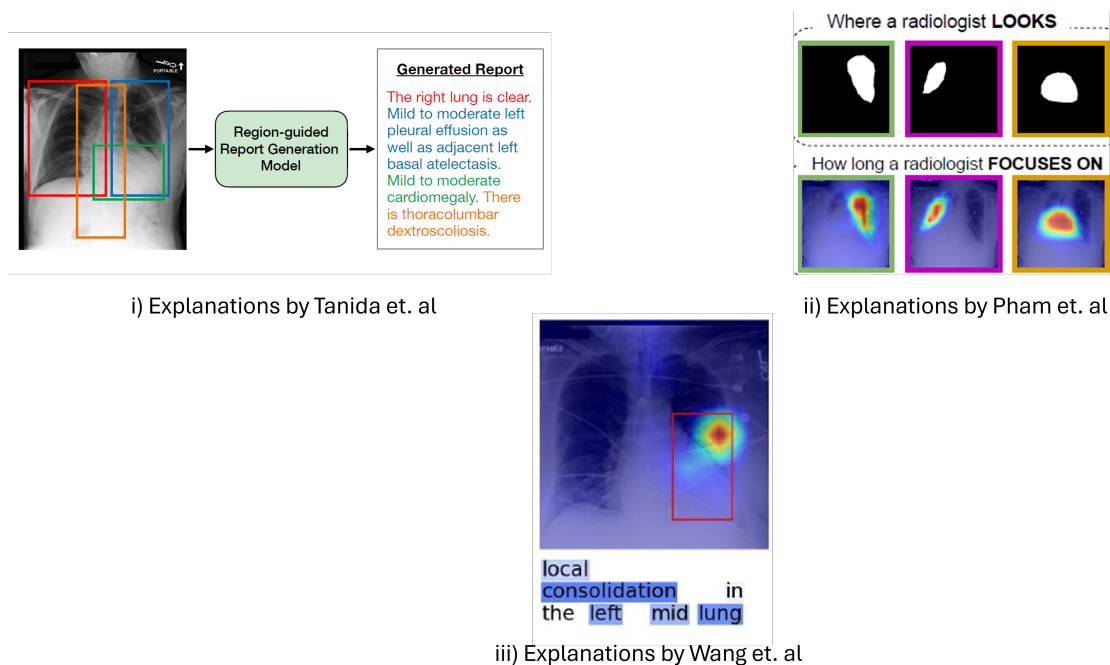


Figure 4.1: Explainable methods in prior work for chest X-ray diagnosis.

### 4.3 Methodology

Our proposed methodology provides the diagnosis result along with explanations through the characteristic descriptors as illustrated in 4.2

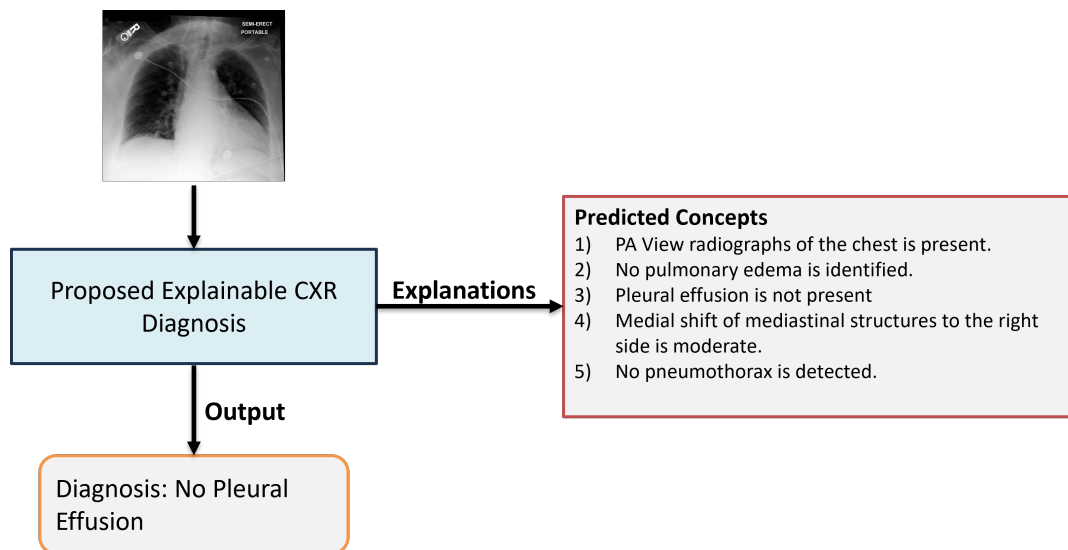


Figure 4.2: Proposed explainable chest X-ray diagnosis system.

**Dataset:** We use MIMIC-CXR dataset which has around 300K chest X-rays, corresponding radiologist reports, and disease labels. For this work, we classify the presence of Pleural Effusion from X-ray images using a balanced set of around 20K data samples. We make an 85% train and 15% test split from the chosen subset of the MIMIC-CXR data.

**Architecture:** We use a CLIP model pre-trained on Chest X-Ray images and reports which is proposed by [52]. This model was trained using contrastive loss similar to the original CLIP model but specialized for chest X-ray data. The convolution and the transformer blocks of the CLIP image encoder and the text encoder are frozen while only the last projection layer of the image encoder is fine-tuned.

As we can extract the concept labels from radiologist reports, we can supervise the explanation generation unlike in face recognition where there was no supervision. Given a corpus of radiology reports, we extract the atomic, fine-grained characteristic descriptors from them using the Mistral 7B language model. We prompt it to disentangle the descriptors

to separate sentences from the report as shown in fig 4.5.

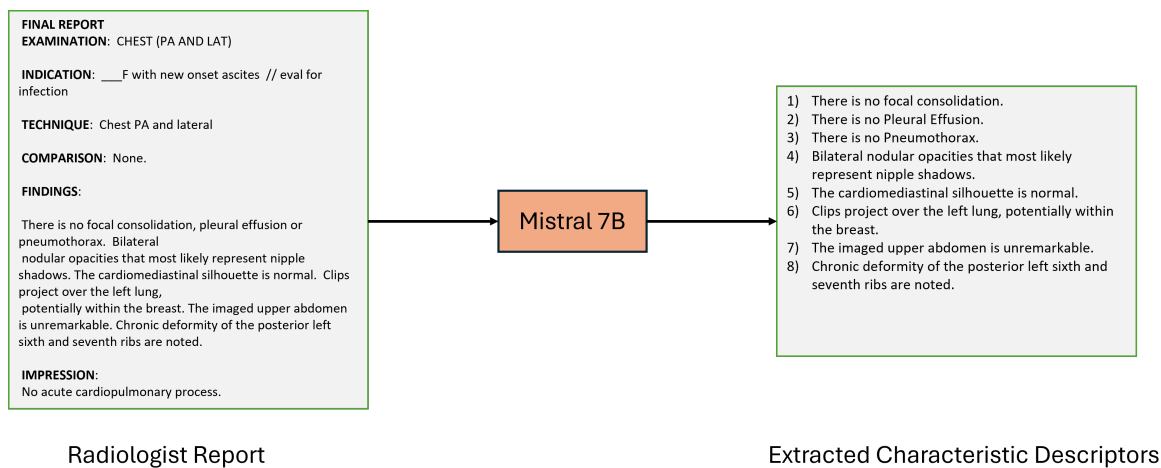


Figure 4.3: An example of extracting characteristic descriptors from radiologist reports.

Given a Chest X-ray, its corresponding report, and label we calculate the cosine similarity between the X-ray and the characteristic descriptor embeddings to obtain concept scores, for obtaining a concept label we set the concepts present in the report to max value in the calculated concept scores. L1 loss calculated between the concept scores and concept scores label is used fine-tuning the model to give appropriate explanations. The concept scores are further passed through a linear layer to get the logits used for making the diagnosis prediction. As the standard, we use the cross-entropy loss for classification.

## 4.4 Results

We used the pre-trained CLIP image encoder along with a linear layer which predicts the presence of Plueral Effusion as a baseline, and compared our proposed explainable architecture against it. We observe from table 4.2 that the classification performance of our proposed architecture is similar to that of the black box while our approach has an added advantage of providing explanations. We also evaluate the explanations provided by our model with the radiologist reports to prove the fidelity. We use two metrics for our evaluation, i) cosine similarity of the embeddings extracted from text encoder of the explanations and ground

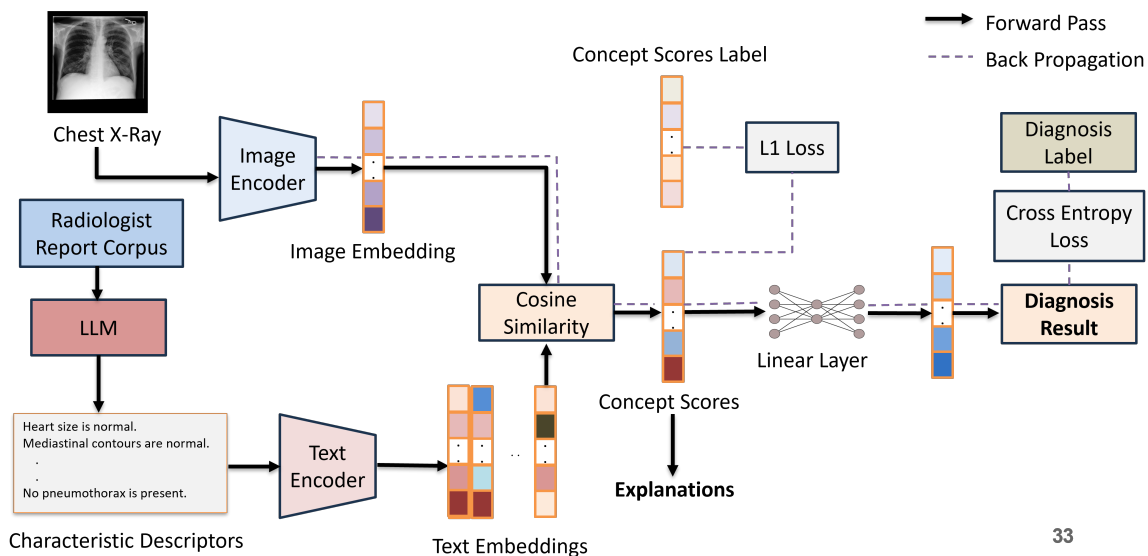


Figure 4.4: An Overview of the Proposed Methodology for Explainable Chest X-ray Diagnosis.

truth, ii) Rouge-L score for measuring the overlap. The evaluation results in table 4.1 show that we achieve a cosine similarity between the explanations and the ground truth exhibiting the fidelity of our explanations.

Table 4.1: Comparison of predicted concepts with labels.

Cosine Similarity (Embeddings)	ROUGE-L
0.91	0.41

Table 4.2: Performance of black box model vs the proposed explainable model in detecting Pleural Effusion.

Model	Accuracy	Precision	Recall	F1-Score
Baseline (Black Box)	78%	79%	78%	78%
Proposed (Explainable)	79%	80%	79%	79%



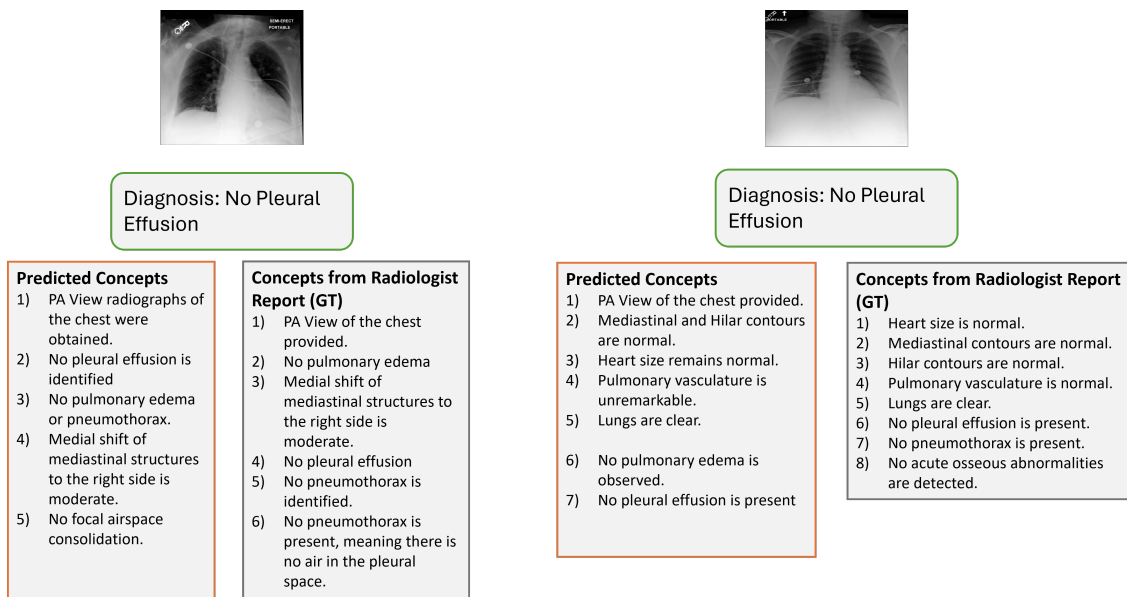


Figure 4.5: Examples of the explanations produced by the model for chest X-Ray Diagnosis.

# Chapter 5

## Conclusion

In this work, we address the problem of explaining face recognition and X-ray diagnosis decisions made by deep learning models using characteristic descriptors. To summarize, our contributions are manifold:

- **Interpretability:** Our approach offers tangible explanations in terms of high-quality human interpretable text.
- **Expert-Level Insight:** The rationales provided are akin to those made by human forensic experts for face matching or radiologists for X-ray diagnosis.
- **Enhances performance through explanatory debugging:** Through counterfactual examples, we shed light on instances of system failure useful for debugging and bias evaluations.

We show that we can design models that give faithful and concrete explanations like a human expert. Moreover, we use counterfactual explanations to debug and understand the biases, and spurious correlations of the model. Through our experiments, we show the performance on benchmark datasets proving the efficacy of our model in providing explanations without significantly affecting the performance in comparison with black-box models. We hope our proposed method with the ability to produce consumable and verifiable descriptions can address transparent and trustworthy face recognition and X-Ray diagnosis systems.

# Bibliography

- [1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "*Why Should I Trust You?*": *Explaining the Predictions of Any Classifier*. 2016. arXiv: 1602.04938 [cs.LG].
- [2] Scott M. Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, 4768–4777. ISBN: 9781510860964.
- [3] David Alvarez Melis and Tommi Jaakkola. "Towards robust interpretability with self-explaining neural networks". In: *Advances in neural information processing systems* 31 (2018).
- [4] Mengnan Du, Ninghao Liu, and Xia Hu. *Techniques for Interpretable Machine Learning*. 2019. arXiv: 1808.00033 [cs.LG].
- [5] Ruoyu Chen et al. "Sim2word: Explaining similarity with representative attribute words via counterfactual explanations". In: *ACM Transactions on Multimedia Computing, Communications and Applications* 19.6 (2023), pp. 1–22.
- [6] Pietro Barbiero et al. "Entropy-Based Logic Explanations of Neural Networks". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.6 (June 2022), 6046–6054. ISSN: 2159-5399. DOI: 10.1609/aaai.v36i6.20551. URL: <http://dx.doi.org/10.1609/aaai.v36i6.20551>.
- [7] Abubakar Abid, Mert Yuksekgonul, and James Zou. "Meaningfully debugging model mistakes using conceptual counterfactual explanations". In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022. URL: <https://proceedings.mlr.press/v162/abid22a.html>.
- [8] Jonathan R Williford, Brandon B May, and Jeffrey Byrne. "Explainable face recognition". In: *European conference on computer vision*. Springer. 2020, pp. 248–263.
- [9] P. Jonathon Phillips and Mark A. Przybocki. "Four Principles of Explainable AI as Applied to Biometrics and Facial Forensic Algorithms". In: *CoRR* abs/2002.01014 (2020).
- [10] <https://www.nytimes.com/2023/08/06/business/facial-recognition-false-arrest.html>. 2023.

- 
- [11] <https://www.nytimes.com/2019/12/19/technology/facial-recognition-bias.html>. 2023.
- [12] Yu-Sheng Lin et al. “xCos: An explainable cosine metric for face verification task”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17.3s (2021), pp. 1–16.
- [13] Haoran Jiang and Dan Zeng. “Explainable face recognition based on accurate facial compositions”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1503–1512.
- [14] Martin Knoche et al. “Explainable model-agnostic similarity and confidence in face verification”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 711–718.
- [15] Cynthia Rudin et al. *Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges*. 2021. arXiv: 2103.11251 [cs.LG].
- [16] Zachary C. Lipton. *The Mythos of Model Interpretability*. 2017. arXiv: 1606.03490 [cs.LG].
- [17] Anirban Sarkar et al. *A Framework for Learning Ante-hoc Explainable Models via Concepts*. 2021. arXiv: 2108.11761 [cs.LG].
- [18] Domingo Mery. “True black-box explanation in facial analysis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 1596–1605.
- [19] Domingo Mery and Bernardita Morris. “On black-box explanation for face verification”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 3418–3427.
- [20] Joao Brito and Hugo Proença. “A deep adversarial framework for visually explainable periocular recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1453–1461.
- [21] Marco Huber et al. “Efficient explainable face verification based on similarity score argument backpropagation”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 4736–4745.
- [22] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *ICML*. 2021.
- [23] Chao Jia et al. “Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision”. In: *CoRR* abs/2102.05918 (2021).

- [24] Mitchell Wortsman et al. “Robust fine-tuning of zero-shot models”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 7949–7961. DOI: 10.1109/CVPR52688.2022.00780.
- [25] Peng Gao et al. “CLIP-Adapter: Better Vision-Language Models with Feature Adapters”. In: *International Journal of Computer Vision* 132 (Sept. 2023), pp. 1–15. DOI: 10.1007/s11263-023-01891-x.
- [26] Sarah Pratt, Rosanne Liu, and Ali Farhadi. “What does a platypus look like? Generating customized prompts for zero-shot image classification”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2022), pp. 15645–15655. URL: <https://api.semanticscholar.org/CorpusID:252111028>.
- [27] Liunian Harold Li et al. “DesCo: Learning Object Recognition with Rich Language Descriptions”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023. URL: <https://openreview.net/forum?id=J2Cso0wWZX>.
- [28] A. Yan et al. “Learning Concise and Descriptive Attributes for Visual Recognition”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, 2023, pp. 3067–3077. DOI: 10.1109/ICCV51070.2023.00287. URL: <https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00287>.
- [29] Mayug Maniparambil et al. “Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts”. In: *arXiv preprint arXiv:2307.11661* (2023).
- [30] Yue Yang et al. *Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification*. 2023. arXiv: 2211.11158 [cs.CV].
- [31] Tuomas Oikarinen et al. *Label-Free Concept Bottleneck Models*. 2023. arXiv: 2304.06129 [cs.LG].
- [32] Sachit Menon and Carl Vondrick. *Visual Classification via Description from Large Language Models*. 2022. arXiv: 2210.07183 [cs.CV].
- [33] [https://fiswg.org/FISWG\\_Morph\\_Analysis\\_Feature\\_List\\_v2.0\\_20180911.pdf](https://fiswg.org/FISWG_Morph_Analysis_Feature_List_v2.0_20180911.pdf). 2023.
- [34] Pang Wei Koh et al. “Concept Bottleneck Models”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 5338–5348. URL: <https://proceedings.mlr.press/v119/koh20a.html>.

- [35] Gabriel Ilharco et al. *OpenCLIP*. Version 0.1. If you use this software, please cite it as below. July 2021. DOI: 10.5281/zenodo.5143773. URL: <https://doi.org/10.5281/zenodo.5143773>.
- [36] Minchul Kim, Anil K. Jain, and Xiaoming Liu. *AdaFace: Quality Adaptive Margin for Face Recognition*. 2023. arXiv: 2204.00964 [cs.CV].
- [37] Jiankang Deng et al. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.10 (Oct. 2022), 5962–5979. ISSN: 1939-3539. DOI: 10.1109/tpami.2021.3087709. URL: <http://dx.doi.org/10.1109/TPAMI.2021.3087709>.
- [38] Gary B. Huang et al. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Tech. rep. 07-49. University of Massachusetts, Amherst, 2007.
- [39] “Frontal to profile face verification in the wild”. English (US). In: *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*. 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016. Institute of Electrical and Electronics Engineers Inc., May 2016. DOI: 10.1109/WACV.2016.7477558.
- [40] Stylianos Moschoglou et al. “AgeDB: The First Manually Collected, In-The-Wild Age Database”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2017.
- [41] Tianyue Zheng, Weihong Deng, and Jiani Hu. *Cross-Age LFW: A Database for Studying Cross-Age Face Recognition in Unconstrained Environments*. 2017. arXiv: 1708.08197 [cs.CV].
- [42] Tianyue Zheng, Weihong Deng, and Jiani Hu. “Cross-Age LFW: A Database for Studying Cross-Age Face Recognition in Unconstrained Environments”. In: *CoRR* abs/1708.08197 (2017). arXiv: 1708.08197. URL: <http://arxiv.org/abs/1708.08197>.
- [43] Hao Wang et al. “CosFace: Large Margin Cosine Loss for Deep Face Recognition”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 5265–5274. DOI: 10.1109/CVPR.2018.00552.
- [44] T. Tanida et al. “Interactive and Explainable Region-guided Radiology Report Generation”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2023, pp. 7433–7442. DOI: 10.1109/CVPR52729.2023.00718. URL: <https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00718>.

- [45] Haibo Jin et al. *PromptMRG: Diagnosis-Driven Prompts for Medical Report Generation*. 2024. arXiv: 2308.12604 [cs.CV].
- [46] T. Pham et al. “I-AI: A Controllable and Interpretable AI System for Decoding Radiologists’ Intense Focus for Accurate CXR Diagnoses”. In: *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, 2024, pp. 7835–7844. DOI: 10.1109/WACV57701.2024.00767. URL: <https://doi.ieeecomputersociety.org/10.1109/WACV57701.2024.00767>.
- [47] Ying Wang, Tim G. J. Rudner, and Andrew Gordon Wilson. *Visual Explanations of Image-Text Representations via Multi-Modal Information Bottleneck Attribution*. 2023. arXiv: 2312.17174 [cs.CV].
- [48] Yiming Lei et al. “CLIP-Lung: Textual Knowledge-Guided Lung Nodule Malignancy Prediction”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan et al. Cham: Springer Nature Switzerland, 2023, pp. 403–412. ISBN: 978-3-031-43990-2.
- [49] C. Wu et al. “MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, 2023, pp. 21315–21326. DOI: 10.1109/ICCV51070.2023.01954. URL: <https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01954>.
- [50] Francesco Dalla Serra et al. “Controllable Chest X-Ray Report Generation from Longitudinal Representations”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 4891–4904. DOI: 10.18653/v1/2023.findings-emnlp.325. URL: <https://aclanthology.org/2023.findings-emnlp.325>.
- [51] Matthew Watson, Bashar Awwad Shiekh Hasan, and Noura Al Moubayed. “Learning How to MIMIC: Using Model Explanations to Guide Deep Learning Training”. In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 1461–1470. DOI: 10.1109/WACV56688.2023.00151.
- [52] Benedikt Boecking et al. “Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing”. In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan et al. Cham: Springer Nature Switzerland, 2022, pp. 1–21. ISBN: 978-3-031-20059-5.