DFCA: Disentangled Feature Contrastive Learning and Augmentation for Fairer Dermatological Diagnostics

Pengcheng Zhao , **Xiaowei Ding**[*]
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China
{pc_zhao, dingxiaowei}@sjtu.edu.cn

Abstract

With the increasing integration of AI in medical research and applications, the issue of fairness has become as critical as diagnostic accuracy. In dermatology diagnosis, the challenge of class-imbalanced data, which is sometimes limited and contains demographic attributes, results in an imbalanced and insufficient representation within the feature space of deep learning models. Besides, feature entanglement within deep learning models confuses skin tone and disease condition information, impairing model performance among vulnerable groups. Moreover, feature entanglement often constrains the efforts to mitigate unfairness, entailing a trade-off between fairness and diagnostic accuracy. This paper introduces the Disentangled Feature Contrastive learning and Augmentation framework (DFCA), aiming to enhance fairness in dermatological diagnoses without compromising accuracy. Initially, DFCA disentangles skin images into disease related and skin-tone features. Subsequently, the two sets of features are projected into normalized spaces for contrastive learning, each modeled by a mixture of von Mises-Fisher (vMF) distributions. DFCA then samples from these vMF distributions to inversely augment the feature space. To further evaluate the fairness-accuracy balance, we propose a new metric, the Accuracy-Fairness Balance Degree (AFBD). Extensive experiments demonstrate that DFCA significantly improves both fairness and accuracy compared to state-of-the-art methods.

1 Introduction

With the rapid development of artificial intelligence (AI), deep learning-based disease diagnosis systems have played a vital role in universal healthcare [Zhou *et al.*, 2021; Archana and Jeevaraj, 2024; Li *et al.*, 2024], especially in remote and impoverished areas where medical resources are severely limited. In the domain of dermatological diagnostics, for example, deep learning models trained on expert-annotated data are now accessible through mobile phones [Lee *et al.*, 2022]. As a result, individuals from underdeveloped regions can even capture skin images using their smartphone cameras to conduct a skin condition diagnosis.

However, these diagnosis models may exhibit distinct performance across different demographic subgroups, particularly among populations with different skin color, which namely unfairness [Mehrabi *et al.*, 2021]. Unfair dermatological diagnosis not only hinder the promotion of AI models but also exacerbate the health gap between different races, potentially tearing society apart and raising serious ethical and legal issues.

Despite numerous researchers have devoted to the enhancing the performance of deep learning models [Mirikharaji *et al.*, 2023; Ye *et al.*, 2024; Choy *et al.*, 2023; Noronha *et al.*, 2023], research on mitigating unfairness remain rare, comparatively. The existing literature for fairness can be divided into three main groups: pre-processing [Kamiran and Calders, 2012; Krasanakis *et al.*, 2018; Puyol-Antón *et al.*, 2021], in-processing [Du *et al.*, 2022; Kamishima *et al.*, 2012; Deng *et al.*, 2023; Xu *et al.*, 2023b] and post-processing methods [Wu *et al.*, 2022]. Pre-processing methods reduce possible bias in data before training the model, in-processing methods intervene during training phase, post-processing adjust the output produced by model to mitigate biased predictions. Pre-processing methods often face the challenge due to limited data [Maluleke *et al.*, 2022]. The latter two methods, by contrast, seemingly direct operate models' feature space or output to remove the useless demographic information, showing a higher ceiling. However, the reduction of demographic information in the feature space will also harm the disease information because of the feature coupling, which ultimately leading to a phenomenon described as accuracy-fairness trade-off [Chen *et al.*, 2023].

From the view of representation learning, there are two main reasons accounting for unfairness in dermatological diagnostics: class imbalanced data and entangled feature in deep learning model. Specifically, different quantities data among various race (maybe also among various skin disease) cause unbalanced disease-relevant feature representations. Besides, deep learning models entangle disease-relevant features and skin type statistic information, which further heaven

[*] Corresponding author.

the degree of unbalance between different race groups in feature space. Moreover, public available skin disease datasets that considering demographic attributes especially race are rare and some with limited data size [Xu *et al.*, 2023a; Xu *et al.*, 2024], natively leading to inadequate representation of disease features in deep learning model. Ultimately, dermatological diagnostics algorithms show unfairness.

In this paper, we introduce the Disentangled Feature Contrastive learning and Augmentation framework (DFCA), which aims to reduce unfairness in AI-based dermatological diagnostics while maintaining accuracy. DFCA employs a symmetric architecture to disentangle features into disease-related and disease-unrelated components, capturing lesion information and demographic statistics (e.g., skin color) separately. Considering the fact that lesion color is in connection with both disease condition and skin type, we map these components into von Mises-Fisher (vMF) distribution spaces for supervised contrastive learning. This approach, which has been proven effective in long-tailed data recognition, ensures that representations of the same disease condition to be close and those of different classes distant regardless of skin tone. The same as skin tone features. To address limited data, DFCA samples from the learned vMF distributions to augment the feature space using a reversible flow model. A perfect diagnosis model should both consider the accuracy and fairness. However, most of the existing methods mainly focus on the fairness metrics, at the cost of diagnosis performance. To emphasize the importance of the balance of accuracy and fairness when researchers designing models, we propose a new metric for assessing the Accuracy-Fairness Balance Degree (AFBD). The results of extensive experiments on two datasets shows that DFCA, by combing disentangled feature contrastive learning and augmentation, improves both fairness and accuracy compared to SOTA methods.

Overall, our contributions lie in effectively disentangling dermatologic images into disease-related and skin-tone-related features, thereby minimizing the interference of demographic attributes in dermatological diagnostics. Additionally, we introduce spaces based on mixture of von Mises-Fisher distributions for contrastive learning, which facilitates the feature disentanglement and clustering. Moreover, we samples from the learned vMF distributions to inversely augment the feature space. To emphasize the importance of balancing fairness and accuracy, we propose a new metric named Accuracy-Fairness Balance Degree (AFBD). Extensive experiments demonstrate that our framework achieves state-of-the-art performance in both fairness and accuracy.

2 Related Works

Despite deep learning has been widely applied in dermatological diagnosis and has achieved notably performance, its generalisability remain limited by poor capture of demographic information[Choy *et al.*, 2023]. The phenomena of unfairness is that those diagnosis models may exhibit distinct performance across different demographic subgroups [Mehrabi *et al.*, 2021]. In this section, we give a concise survey of unfairness mitigation methods and analyze the position of this paper within the field.

Existing unfairness mitigation can be categorized into pre-processing, in-processing, and post-processing methods according to the implementation phase [Xu *et al.*, 2024].

Pre-processing methods reduce possible bias in data before training the model, their common practice is to restructure the training data. The representative work is [Kamiran and Calders, 2012], where the authors proposed four methods to pre-processing the dataset, including removing sensitive attribute, massaging the dataset, reweighing and resampling. Similarly, [Zhang *et al.*, 2022; Abernethy *et al.*, 2020; Puyol-Antón *et al.*, 2021] also upsampled minority groups to ensure equal presence of each subgroup. [Krasanakis *et al.*, 2018] proposed an adaptive sensitive reweighting mechanism model for fair classify. Another technical approach is data generation. [Burlina *et al.*, 2021] augmented the training data by generative models for debiasing in retinal diagnostics. To mitigate the impact of demographic imbalance, [Pombo *et al.*, 2023] employed generative model to synthesise counterfactual volumetric brain imaging, conditioning on original image and demographic attributes (sex and age). Although pre-processing methods can achieve high accuracy, they can not solve feature under representation among vulnerable groups and may face instability problems in generation due to limited data. Our method makes the advantage of the learned vMF distribution in contrastive space and samples from that to inversely conduct augmentation in feature space.

In-processing methods intervene during training phase, they usually combine the architecture of model and extra losses to mitigate unfairness. [Stanley *et al.*, 2022] added an adversarial branch to minimize the influence of sex and race in brain MRI image classification. FairAdaBN [Xu *et al.*, 2023b] reduces unfairness by adding extra adapters to the original model which can adaptively adjust the mean and variance of the feature vector according to the sensitive attribute. [Pakzad *et al.*, 2022] adopted a domain invariant representation learning method to remove the skin tone information in the classifer. [Du *et al.*, 2022] is most closely to our work, which proposed FairDisCo framework to mitigate the negative effects of skin tone in dermatology diagnosis. Similarly, FairDisCo also used contrastive learning to boost the diagnoise accuracy and disentangled learning with different branches to reduce impede of skin tone information. However, its contrastive learning constrained by limited data and different branches constrained by highly coupled feature in the feature extractor. Our method adopts a symmetric architecture for disentanglement. Besides, we uses the vMF distribution based contrastive learning to further promote the diagnosis performance.

Post-processing methods adjust the output produced by model to mitigate biased predictions. [Pleiss *et al.*, 2017] employed calibration for specific subgroup thresholds in fair classification. [Wu *et al.*, 2022] utilized a pruning strategy to remove the sensitive information associated with a specific subgroup. [Huang *et al.*, 2023] considered age and gender and also proposed a pruning method. However, these methods may face the accuracy-fairness trade-off because of the feature coupling. A portion of the model parameters may be associated with both disease condition and skin type, which can be settled by feature disentanglement in our method.

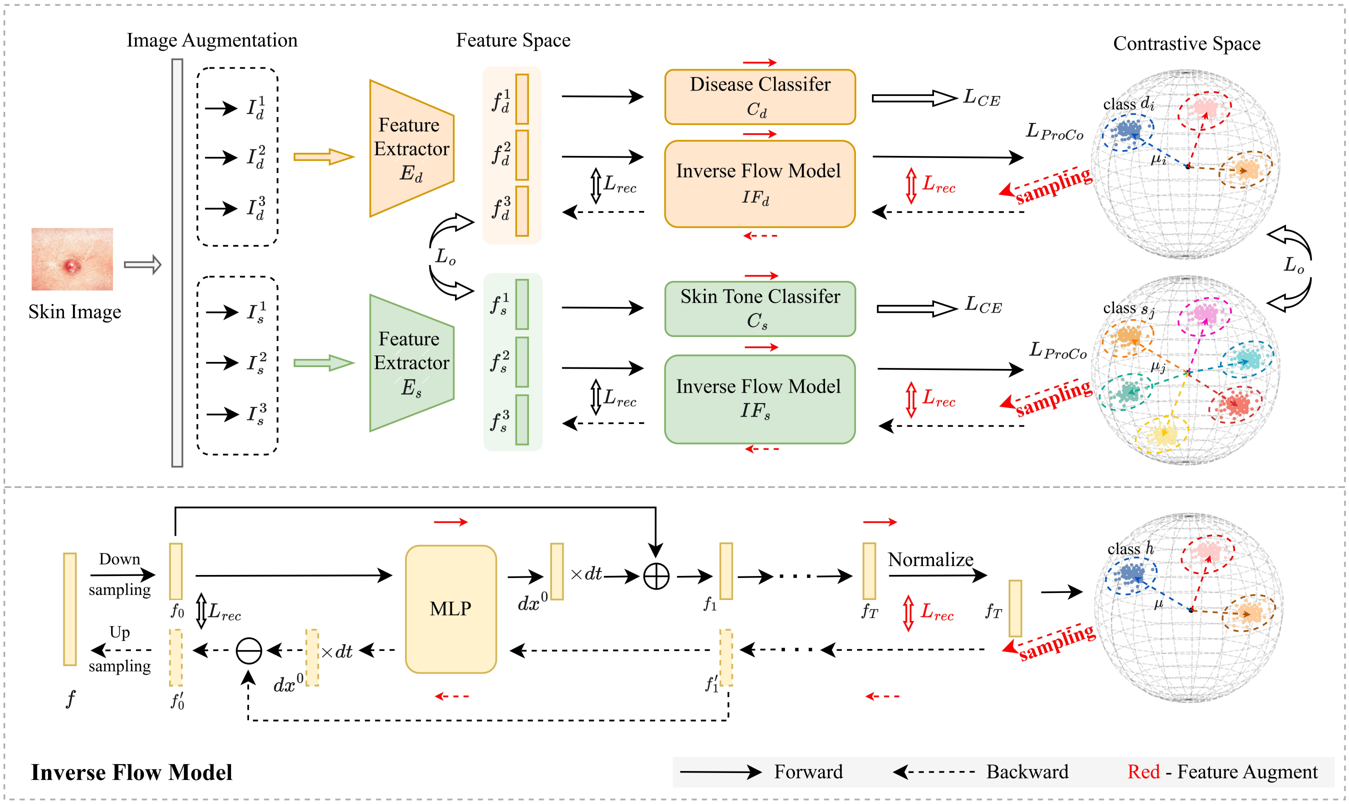

Figure 1: Overview of the Disentangled Feature Contrastive learning and Augmentation (DFCA) Framework. The upper box illustrates the symmetric architecture and training process of DFCA. Given a mini-batch of real skin images, DFCA generates six augmented versions for each image. I^2 and I^3 use the same augmentation method, distinct from that of I^1. Each I_d differs from I_s due to diverse tasks. DFCA then disentangles the input images into disease-related and skin-type-related feature spaces using E_d and E_s. The disentangled features f_d^1 and f_s^1 are fed to the Disease Classifier C_d and Skin Tone Classifier C_s for supervised learning. Meanwhile, f_d^2, f_d^3, f_s^2, and f_s^3 are projected into contrastive spaces by the Inverse Flow Models IF_d and IF_s for contrastive learning. The bottom box shows the Inverse Flow Model and its training process. This model consists of two inverse processes to learn the conversion between feature space and contrastive space. In the forward stage, an input feature vector f is down-sampled and passed through a series of MLP blocks to capture minute changes dx at each step. The final feature f_T is normalized to a unit space modeled by a mixture of vMF distributions. In the backward stage, f_T is used to reconstruct the original feature f for training the Inverse Flow Model. Red arrows indicate the feature augmentation process. After training with real skin data, we sample from the vMF distributions for feature augmentation, incorporating real images into the training process.

3 Methods

As mentioned previously, unbalanced data and feature entanglement are the primary causes of unfairness. Class imbalanced data brings the imbalanced feature representation and feature entanglement in deep learning model further mix the disease condition information and skin tone information, thereby cause unfairness dermatological diagnosis. Our goal is to reduce the performance differences of the diagnostic model across different populations while maintaining accuracy, thereby alleviating unfairness. To achieve this, DFCA is proposed. To eliminate the impact of skin tone on disease diagnosis, we employ a symmetric disentangled structure to obtain disease-related features and skin tone related features. To overcome the imbalanced feature representation, DFCA adopts surpervised contrastive learning based on a mixture of von Mises-Fisher (vMF) distributions, as described in Section 3.1. Moreover, we samples from the vMF distribution-based contrastive spaces and inversely generate features to further augment the feature space and boost diagnosis performance, as described in Section 3.2.

3.1 Feature Disentanglement and Contrastive Learning

We consider the problem of dermatological diagnosis under the influence of skin tone information. Let $X = \{x_1, \ldots, x_N\}$ be the training set, where x_i denotes the ith sample. X has two sets of labels: $Y_D = \{y_d^1, \ldots, y_d^N\}$ and $Y_S = \{y_s^1, \ldots, y_s^N\}$, where y_d^i, y_s^i correspond to the disease label and skin tone label of x_i, respectively. $Y_d^i \in \{0, 1, \ldots, C\}$ and $Y_s^i \in \{0, 1, \ldots, S\}$ The perfect situation is that the diagnostic results are independent of the skin tone labels of the data:

$$P(\hat{Y}_D = i | Y_D = i, Y_S = a) = P(\hat{Y}_D = i | Y_D = i, Y_S = b) \tag{1}$$

where $a \neq b$ and $a, b \in Y_S$. We strive to achieve this goal by feature disentanglement and contrastive learning.

A Feature Disentanglement

As shown in Figure 1, given a sample from a mini batch of training data, we initially employ a distinctive data augmentation to get six versions. Each three of them form a group as the inputs to the following designed network, denoted as $I_D = \{I_d^1, I_d^2, I_d^3\}$ and $I_S = \{I_s^1, I_s^2, I_s^3\}$. In each group, I^2 and I^3 employ the same weak augmentation method, distinct from a strong augmentation method of I^1, as the manner of [Zhu *et al.*, 2022]. To maximize the reduction of the impact of skin tone information in the data on disease diagnosis, we adopt a symmetric architecture to disentangle the feature into disease-related part and disease-unrelated part, which mainly contain the disease condition information and skin tone information, respectively. We named them the branch d and branch s. Correspondingly, we adopt different augmentation methods between I_D and I_S. While I_D focus on diverse gray scale variation and flip for robust lesion feature analysis ability and I_S focus on different fuzzy degree for the judgment of skin tone.

After the image augmentation, DFCA employ E_d, E_s to encode $\{I_d^1, I_d^2, I_d^3\}$, $\{I_s^1, I_s^2, I_s^3\}$ into the feature space $\{f_d^1, f_d^2, f_d^3\}$ and $\{f_s^1, f_s^2, f_s^3\}$. E_d and E_s have the same structure but without shared weights. Suppose that the features of sample x_i are $\{f_d^{i1}, f_d^{i2}, f_d^{i3}\}$ and $\{f_s^{i1}, f_s^{i2}, f_s^{i3}\}$ We employ an orthogonality loss to ensure that the two feature spaces are mutually independent and semantically continuous:

$$L_{of} = \sum_{i=1}^{N}\{\sum_{j=1}^{3}\sum_{a=1}^{3} < f_d^{ij}, f_s^{ia} > + < \sum_{j=1}^{3} f_d^{ij}, \sum_{a=1}^{3} f_s^{ia} >\} \tag{2}$$

Next, we take the branch d as an example while the branch s undergoing the same operations. Given the $\{f_d^1, f_d^2, f_d^3\}$, DFCA send f_d^1 into the disease classifer C_d and training in a surpervised manner:

$$L_{cd} = -\frac{1}{N}\sum_{i=1}^{N} y_d^i log(\hat{y}_d^i) \tag{3}$$

and the same with L_{cs} for branch s.

Considering the fact that lesion color is in connection with both disease condition and skin type, we assigned a small weight to L_{of} in practice. Moreover, we design a inverse flow model takeing f_d^2 and f_d^3 as input for supervised contrastive learning, making the distance of representation from same disease condition close and from different class far away no matter which skin tone it belongs to, and the same with f_s^2 and f_s^3.

B Supervised Contrastive Learning

The *von Mises-Fisher* (vMF) distribution is often described as the Normal Gaussian distribution on a hypersphere and is sometimes used as a substitute for Y in some scenarios. Recently, vMF distribution has also demonstrated its potential for application in the contrastive learning especially for the long-tailed recognition problem [Du *et al.*, 2024]. Similar to Gaussian distribution, it can be parameterized by μ indicating the mean direction and k representing the concentration around μ. The larger the value of k, the greater is the clustering aruond the mean direction μ. When $k = 0$, the distribution is uniform. The probability density function of the vMF distribution for a random d-dimensional unit vector z is defined as:

$$f(z;\mu,k) = (\frac{k}{2})^{p/2-1}\frac{1}{\Gamma(p/2)I_{p/2-1}(k)}e^{k\mu^T x} \tag{4}$$

$$C_p(k) = (\frac{k}{2})^{p/2-1}\frac{1}{\Gamma(p/2)I_{p/2-1}(k)} = \frac{k^{p/2-1}}{(2\pi)^{p/2}I_{P/2-1}(k)} \tag{5}$$

where $k \geq 0$, $\|\mu\|^2 = 1$, $C_p(k)$ is the normalizing constant and I_v denotes the modified Bessel function of the first kind at order v:

$$I_{P/2-1}(z) = \sum_{K=0}^{\infty}\frac{1}{k!\Gamma(p/2-1+k+1)}(\frac{z}{2})^{2k+p/2-1} \tag{6}$$

In this paper, we introduce a mixture of vMF distributions as the distribution in the contrastive space for each branch. To model the conversion from feature space to contrastive space, we adopt a simple but effective inverse flow model. This paves the way for feature enhancement, which will be introduced in the next section.

The bottom box of Figure 1 exhibits the structure of our inverse flow model for supervised contrastive learning. IF_d has two reversible flow process. Given the d-dimensional feacture vetor f, we firstly downsample it to a $d/2$-dimensional vector f_0 as the input at the starting moment ($t = 0$). Through the repeated MLP blocks, the model learn the minute changes from feature space to contrastive space:

$$f_1 = f_0 + M(f_0) * dt, \ldots, f_T = f_{T-1} + M(f_{T-1}) * dt \tag{7}$$

where $dt = \frac{1}{T}$, $M()$ stands for MLP block and f_t is the vector at the t th moment. Note that MLP blocks would not change the dimensionality of the vectors. The final vector f_T is subsequently projected into the contrastive space by normalization. We also use an orthogonality loss to ensure the independence of two contrastive spaces:

$$L_{oc} = \sum_{i=1}^{N}\{\sum_{j=2}^{3}\sum_{a=2}^{3} < f_{dT}^{ij}, f_{sT}^{ia} > + < \sum_{j=2}^{3} f_{dT}^{ij}, \sum_{a=2}^{3} f_{sT}^{ia} >\} \tag{8}$$

[Du *et al.*, 2024] has given a closed form of expected supervised contrastive loss L_{ProCo} based on the estimated vMF distribution when the sampling number tends to infinite. L_{ProCo} can address the issue of insufficient performance in contrastive learning due to a small number of contrastive sample pairs. We use the L_{ProCo} as:

$$L_{ProCod} = -log(\pi_{y_j} \frac{C_p(\hat{k}_{y_j})}{C_p(k_{y_j})}) + log(\sum_{j=1}^{C} \pi_j \frac{C_p(\hat{k}_j)}{C_p(k_j)}) \quad (9)$$

where π_{y_j} is the class frequency in the training set. $\{f_{dT}^2, f_{dT}^3\}$ are used to calculate the L_{ProCod} loss and the same with L_{ProCos} for $\{f_{dS}^2, f_{dS}^3\}$.

The backward process of inverse flow model follows:

$$f'_{T-1} = f_T - M(f_T) * dt, \ldots, f'_0 = f'_1 - M(f'_1) * dt \quad (10)$$

which is just the inverse process of Eq.(7). Through the backward process, f_0 is gradually recovered from f_T denoted as f'_0. A recontruction loss L_{recd} is ultilized to train the inverse flow model IF_d and the same with L_{recs}:

$$L_{recd} = \|f - IF_d^{-1}(IF_d(f))\|_2 \quad (11)$$

3.2 vMF distribution based Feature Augmentation

Through feature disentanglement and contrastive Learning mentioned before, DFCA learns a mixture of vMF distributions in contrastive space. Despite L_{ProCo} has given a closed form of expected supervised contrastive loss when the sampling number tends to infinite, we observe that its capability becomes constrained in some situations where the dataset contains a very limited number of samples. To take advantage of the learned vMF distributions, we sample from contrastive space to inversely augment the feature space for further enhancing the capability of disease classifer.

Consider the vMF distribution:

$$f(z; \mu, k) = C_p(k) e^{k<\mu^T, x>} \quad (12)$$

where $C_p(k)$ is the normalizing constant. Given the μ_j of vMF distribution of class j, the density depends on x only through $\mu^T x$. We decomposition x as:

$$x = \omega\mu + (1 - \omega^2)^{\frac{1}{2}} v \quad (13)$$

where $\mu \perp v$ and ω can be seen as the cosine of the angle between x and μ. The probability density function of ω is given in [Mardia and Jupp, 2000]:

$$f(\omega) = (\frac{k}{2})^{p/2-1} \{\Gamma(\frac{p-1}{2}) \Gamma(\frac{1}{2} I_{(p-1)/2}(k))\}^{-1} \cdot e^{k\omega}((1 - \omega^2)^{(p-3)/2} \quad (14)$$

with $\omega \in [-1, 1]$. We calculate the $f^{-1}(\omega)$ to get the distribution of ω. Then we can sample the x from the vMF distribution through Eq.13, defined as f_{samT}.

Red arrows in Figure 1 shows the training process in this stage. We firstly inverse the f_{samdT} to feature vector y_{samd}. Afterwards, y_{samd} is sent into the disease classifer with the loss functions:

$$L_{csamd} = -\frac{1}{M} \sum_{i=1}^{M} y_{samd}^i log(\hat{y}_{samd}^i) \quad (15)$$

where M is the number of sampled features and $\hat{y}_{samd}$ is the label of sampled feature y_{samd}. Similar with L_{recd}, we inversely reconstruct the f_{samdT} with the constrain of:

$$L_{recsamd} = \|f_{samdT} - IF_d(IF_d^{-1}(f_{samdT}))\|_2 \quad (16)$$

while the same with $L_{recsams}$ and L_{csams}. Noted that we also use the real data for mixed training in this stage.

3.3 Learning Loss

Briefly, we define the loss fuctions by: $L_o = L_{of} + L_{oc}$, $L_d = L_{cd} + L_{ProCod} + L_{csamd}$, $L_s = L_{cs} + L_{ProCos} + L_{csams}$ and $L_{rec} = L_{recd} + L_{recs} + L_{recsamd} + L_{recsams}$. Therefore, the joint loss function can be represented as:

$$L = L_{rec} + \alpha L_d + \beta L_s + \gamma L_o \quad (17)$$

where α, β, γ are the weights of the loss terms.

4 Experiment

4.1 Datasets and Evaluation Metrics

We use two well-known dermatology datasets to evaluate our proposed method: Fitzpatrick-17k dataset [Groh *et al.*, 2021] and DDI dataset [Daneshjou *et al.*, 2022]. Both of the datasets contain skin tone attribute.

The Fitzpatrick-17k dataset [Groh *et al.*, 2021] contains 16,577 dermatology images with disease condition labels and skin tone labels. They have two kinds of methods for disease condition classification: 3 (malignant, non-neoplastic, benign) and 9. There are six categories of skin tones labeled by 1-6. The higher the number, the darker the skin tone.

The DDI dataset [Daneshjou *et al.*, 2022] contains 656 dermatology images with disease condition labels and skin tone labels. Its disease condition labels contains malignant (0) or not (1). And its skin tones are categorised into 3 groups: 1, 2, 3, corresponding to the skin tone of {1, 2}, {3, 4}, {5, 6} in Fitzpatrick-17k dataset, respectively.

We use three common metircs for evaluating the fairness toward the effection of skin tone attributes: Predictive Quality Disparity (PQD), Demographic Disparity (DP) and Equality of Opportunity (EO). PQD measures the prediction quality difference between each subgroup, DP computes the percentage diversities of positive outcomes for each subgroup and EO asserts that different subgroups should have similar true positive rates, just as [Du *et al.*, 2022].

Moreover, a perfect diagnosis model should both consider the accuracy and fairness. However, most of the existing methods mainly focus on the fairness metrics. To emphasize the importance of the balance of accuracy and fairness when researchers designing models, we propose a new metric for assessing the Accuracy-Fairness Balance Degree (AFBD):

Method	Avg	T1	T2	T3	T4	T5	T6	PQD	DP	EO	AFBD
RESM	85.45	82.27	82.63	86.45	**89.39**	90.11	89.17	91.31	46.98	67.56	20.25
REWT	85.23	81.77	83.04	85.57	**89.39**	89.40	89.17	91.47	48.92	70.64	21.08
FairAdaBN	84.78	82.99	82.52	85.60	86.32	86.45	85.76	**93.21**	56.32	69.99	33.78
FairPrune	85.19	84.29	82.50	86.32	87.45	87.88	83.48	91.78	52.43	66.32	29.41
LFF	86.26	83.27	83.74	87.50	89.67	91.26	84.21	91.25	51.87	65.12	22.30
FairDisCo	85.76	84.45	82.94	86.60	88.49	90.11	87.50	92.04	55.53	75.62	26.00
DFCA(ours)	**87.94**	**87.28**	**85.91**	**88.63**	88.73	**91.28**	**89.37**	89.89	**59.57**	**77.62**	**35.29**

Table 1: In-domain comparative analysis of accuracy across different methods on Fitzpatrick-17k dataset. Avg means micro average accuracy computed over all skin tones. T1-T6 represent the classfication accuracy on each skin tone. The larger the values, the better they are regarded. All values are expressed as percentages, except for the last four columns.

Method	Avg	T12	T34	T56	PQD	DP	EO	AFBD
RESM	82.58	83.78	80.39	84.09	95.60	79.28	**93.01**	31.36
REWT	82.58	81.08	78.43	**88.64**	88.49	83.39	82.45	16.84
FairAdaBN	80.67	89.17	74.54	81.24	87.89	81.23	78.56	13.30
FairPrune	74.79	68.48	73.56	85.21	85.02	63.33	62.07	10.70
LFF	84.21	84.09	81.39	86.12	95.82	83.75	82.50	32.18
FairDisCo	83.33	83.78	84.31	81.82	**97.04**	83.62	83.46	42.09
DFCA(ours)	**87.88**	**89.23**	**87.19**	88.53	93.26	**84.11**	82.69	**46.33**

Table 2: In-domain comparative analysis of accuracy across different methods on DDI dataset. Avg means micro average accuracy computed over all skin tones. T12-T56 represent three classes of skin tone in DDI. The larger the values, the better they are regarded. All values are expressed as percentages, except for the last four columns.

$$AFBD = \frac{ACC}{1 + AD}, AD = \frac{1}{N} \sum_{i=1}^{N} |ACC - ACC_i| \quad (18)$$

where AD is the averaged disparity between average accuracy ACC and accuracy of subgroup ACC_i with skin tone i.

4.2 Experiment Settings

We implement our DFCA model by PyTorch. DFCA is trained for 150 epochs firstly with the real datasets and 100 epochs with the mixture of feature augmentation. After all the trainging stages, we only use the Feature Extractor E_d and Disease Classifer C_d for evaluation. Both of the feature extractors are ResNet-101 [He *et al.*, 2016] pretrained on ImageNet. We use the same image augmentation mechanisms as [Zhu *et al.*, 2022]. Our model is trained by Adam optimizer with a learning rate $lr = 0.0001$. The batch size is 32. The weights are $\alpha = 10, \beta = 0.5$ and $\gamma = 1$.

4.3 Baseline Methods

We compare our DFCA with six state-of-the-art models, including the pre-processing methods RESM and REWT [Kamiran and Calders, 2012], in-processing methods FairAdaBN [Xu *et al.*, 2023b], LFF [Wang *et al.*, 2024], FairDisCo [Du *et al.*, 2022], post-processing methods FairPrune [Wu *et al.*, 2022] ,according to the intervene stage for fairness.

4.4 Performance Comparison with SOTA

In this section, we conduct the in-domain experiment for the comparison of fair diagnosis and out-domain experiment for observing the model's generalization ability. Due to the samll datasize of DDI dataset, we only conduct the in-domain experiment on it, just as [Du *et al.*, 2022].

Table 1 and table 2 show the results of in-domain comparison on Fitzpatrick-17k and DDI datasets. We see that our method has the biggest average accuracy and the best diagnosis performance in almost all the skin tones as well as three fairness evaluation metrics, both on Fitzpatrick-17k and DDI datasets. It is noteworthy that in table 2, DFCA achieve a significant enhancement on DDI, suggesting that our method can promote the performance on small dataset. While FairPrune and FairAdaBN gain the worst performance, because of the feature entanglement in their model. As to AFBD, we found that our method also attaines the optimal performance on both of the datasets, indicating that DFCA also taken the accuracy-fairness balance into consideration. Overall, the results demonstrate that our contrastive feature disentangment and augmentation method is able to enhance fairness in dermatological diagnoses as well as accuracy.

Table 3 exhibts the out-domain comparison on Fitzpatrick-17k dataset. '-' represents that we use data of this skin tone for training and test in other skin tones. And we conduct three groups of experiments. Due to the poor performance of FairPrune and FairAdaBN in both the accuracy and fairness metrics especially when data is limted (as seen in Table 2), we discarded them in the out-domain experiment. We can see that our proposed method achieves highest scores in almost every average accuracy and diagnosis accuracy of each skin tone among all of the three groups. Especially in the group two, DFCA has elevated average accuracy metric by nearly 4 percentage points. Moreover, DFCA has also deliv-

Method	Avg	T1	T2	T3	T4	T5	T6	PQD	DP	EO	AFBD
RESM	80.33	-	-	80.20	80.65	79.58	81.42	94.17	79.49	63.29	**51.08**
REWT	79.13	-	-	80.44	79.11	77.76	75.75	**98.60**	69.56	72.30	31.40
LFF	79.97	-	-	80.18	80.89	78.92	80.93	95.23	68.48	62.42	44.80
FairDisCo	80.37	-	-	81.41	81.12	78.02	77.32	94.98	74.34	62.11	28.73
DFCA(ours)	**82.45**	-	-	**83.07**	**82.88**	**80.83**	**82.52**	92.69	**79.52**	**72.41**	**48.93**
RESM	78.74	74.96	78.62	-	-	85.91	81.10	86.94	71.03	74.96	18.07
REWT	77.60	72.89	77.18	-	-	85.58	83.31	85.17	53.97	64.90	13.60
LFF	76.63	71.23	76.98	-	-	80.64	80.07	82.63	50.60	74.33	17.82
FairDisCo	78.61	74.62	78.56	-	-	85.52	80.79	**87.25**	71.44	69.53	18.36
DFCA(ours)	**82.18**	**77.33**	**81.16**	-	-	**86.21**	**83.63**	86.98	**73.23**	**74.97**	**21.41**
RESM	73.70	69.05	71.83	74.56	**80.76**	-	-	85.50	63.23	77.65	15.99
REWT	69.75	61.70	66.95	72.30	79.93	-	-	77.20	56.24	75.50	10.12
LFF	70.81	62.73	70.69	73.11	79.67	-	-	83.42	65.96	74.33	12.13
FairDisCo	73.64	70.63	71.64	73.61	80.25	-	-	**88.01**	70.69	83.69	18.82
DFCA(ours)	**74.22**	**72.32**	**73.02**	**74.62**	80.69	-	-	86.92	**72.11**	**85.32**	**21.25**

Table 3: Out-domain comparative analysis of accuracy across different methods on Fitzpatrick-17k dataset. Avg means micro average accuracy computed over all skin tones. '-' represents we use data of this skin tone for training and test in other skin tones. The larger the values, the better they are regarded. All values are expressed as percentages, except for the last four columns.

	w/o Dis	w/o IF	w/o Ctr	DFCA
ACC	85.45	86.39	86.95	**87.94**
AFBD	29.98	35.16	33.56	**35.29**

Table 4: Abalation study: the contribution of each component

	Diff	MLP	DFCA(IF)
ACC	86.67	87.23	**87.94**
AFBD	31.69	32.76	**35.29**

Table 5: Abalation study: the impact of structure to inverse model

ered a satisfactory performance in terms of fairness metrics and our propose AFBD. Overall, the results demonstrate that our method possess robust generalization capabilities under the fairness dermatological diagnostics as well.

4.5 Abalation Study

To understand the contribution of each component in the DFCA, we perform ablation studies on the Fitzpatrick-17k dataset. Table 4 shows the comparison results of ACC and AFBD. 'w/o Dis' means we conduct the experiment without feature disentanglement, 'w/o IF' means no feature augmentation while with contrastive learning, and 'w/o Ctr' means no contrastive learning while with feature augmentation. The results indicate that feature disentanglement makes the greatest contribution both in diagnosis accuracy and fairness. While the proposed feature augmentation from vMF distribution can enhance diagnosis accuracy while ensuring fairness. And the L_{ProCo} loss of contrastive learning can better take the balance between accuracy and fairness into consideration.

Furthermore, we conduct another abalation study to investigate the impact brought by the structure of inverse model, which is designed to model the inverse conversation between feature space and contrastive space, as exhibited in Table 5. We take three methods for comparision: diffusion model, one MLP network and our proposed inverse flow model IF. The results show that MLP can achieve comparable performance to DFCA in terms of accuracy. However, MLP is relatively weak in terms of fairness, indicating the flow model can better handle the complex conversions between feature space and contrastive space. Diffusion model achieves the worst performance in both metrics, due to its training instability expecially for the certain vMF distribution.

5 Conclusion and Future Work

In this paper, we propose a Disentangled Feature Contrastive learning and Augmentation framework (DFCA) to reduce unfairness in dermatological diagnostics. DFCA employs a symmetric architecture to disentangle features into disease-related and skin-tone-related components, thereby minimizing the impact of skin tone on disease diagnosis. We introduce a contrastive learning framework based on von Mises-Fisher (vMF) distributions and use it to inversely augment the feature space. Extensive comparisons and ablation studies demonstrate the effectiveness of our approach.

Despite our method presents a promising approach to solving the fairer dermatological diagnostics problem, there are several limitations. Dermatological datasets with skin tone attributes remain scarce, which limits the development in this direction. Our proposed inverse flow model for feature augmentation is computationally complex. Further research is needed to provide larger datasets and computationally efficient methods. Additionally, improving interpretability and extending applications to a broader range of diseases are key directions for future work.

Acknowledgments

We are grateful for the insightful comments and suggestions from the reviewers.

References

[Abernethy *et al.*, 2020] Jacob Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, and Jie Zhang. Active sampling for min-max fairness. *arXiv preprint arXiv:2006.06879*, 2020.

[Archana and Jeevaraj, 2024] R Archana and PS Eliahim Jeevaraj. Deep learning models for digital image processing: a review. *Artificial Intelligence Review*, 57(1):11, 2024.

[Burlina *et al.*, 2021] Philippe Burlina, Neil Joshi, William Paul, Katia D Pacheco, and Neil M Bressler. Addressing artificial intelligence bias in retinal diagnostics. *Translational Vision Science & Technology*, 10(2):13–13, 2021.

[Chen *et al.*, 2023] Richard J Chen, Judy J Wang, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6):719–742, 2023.

[Choy *et al.*, 2023] Shern Ping Choy, Byung Jin Kim, Alexandra Paolino, Wei Ren Tan, Sarah Man Lin Lim, Jessica Seo, Sze Ping Tan, Luc Francis, Teresa Tsakok, Michael Simpson, et al. Systematic review of deep learning image analyses for the diagnosis and monitoring of skin disease. *NPJ Digital Medicine*, 6(1):180, 2023.

[Daneshjou *et al.*, 2022] Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022.

[Deng *et al.*, 2023] Wenlong Deng, Yuan Zhong, Qi Dou, and Xiaoxiao Li. On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations. In *International Conference on Information Processing in Medical Imaging*, pages 158–169. Springer, 2023.

[Du *et al.*, 2022] Siyi Du, Ben Hers, Nourhan Bayasi, Ghassan Hamarneh, and Rafeef Garbi. Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning. In *European Conference on Computer Vision*, pages 185–202. Springer, 2022.

[Du *et al.*, 2024] Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Groh *et al.*, 2021] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Huang *et al.*, 2023] Yun-Yang Huang, Venesia Chiuwanara, Chao-Hsuan Lin, and Po-Chih Kuo. Mitigating bias in mri-based alzheimer's disease classifiers through pruning of deep neural networks. In *Workshop on Clinical Image-Based Procedures*, pages 163–171. Springer, 2023.

[Kamiran and Calders, 2012] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

[Kamishima *et al.*, 2012] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 35–50. Springer, 2012.

[Krasanakis *et al.*, 2018] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pages 853–862, 2018.

[Lee *et al.*, 2022] Kyungsu Lee, Thiago Coutinho Cavalcanti, Sewoong Kim, Hah Min Lew, Dae Hun Suh, Dong Hun Lee, and Jae Youn Hwang. Multi-task and few-shot learning-based fully automatic deep learning platform for mobile diagnosis of skin diseases. *IEEE Journal of Biomedical and Health Informatics*, 27(1):176–187, 2022.

[Li *et al.*, 2024] Xiang Li, Lin Zhao, Lu Zhang, Zihao Wu, Zhengliang Liu, Hanqi Jiang, Chao Cao, Shaochen Xu, Yiwei Li, Haixing Dai, et al. Artificial general intelligence for medical imaging analysis. *IEEE Reviews in Biomedical Engineering*, 2024.

[Maluleke *et al.*, 2022] Vongani H Maluleke, Neerja Thakkar, Tim Brooks, Ethan Weber, Trevor Darrell, Alexei A Efros, Angjoo Kanazawa, and Devin Guillory. Studying bias in gans through the lens of race. In *European Conference on Computer Vision*, pages 344–360. Springer, 2022.

[Mardia and Jupp, 2000] Kanti V. Mardia and Peter E Jupp. *Directional Statistics*. Directional statistics /, 2000.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[Mirikharaji *et al.*, 2023] Zahra Mirikharaji, Kumar Abhishek, Alceu Bissoto, Catarina Barata, Sandra Avila, Eduardo Valle, M Emre Celebi, and Ghassan Hamarneh. A survey on deep learning for skin lesion segmentation. *Medical Image Analysis*, 88:102863, 2023.

[Noronha *et al.*, 2023] Stephanie S Noronha, Mayuri A Mehta, Dweepna Garg, Ketan Kotecha, and Ajith Abraham. Deep learning-based dermatological condition detection: A systematic review with recent methods, datasets, challenges and future directions. *IEEE Access*, 2023.

[Pakzad *et al.*, 2022] Arezou Pakzad, Kumar Abhishek, and Ghassan Hamarneh. Circle: Color invariant representation learning for unbiased classification of skin lesions. In *European Conference on Computer Vision*, pages 203–219. Springer, 2022.

[Pleiss *et al.*, 2017] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

[Pombo *et al.*, 2023] Guilherme Pombo, Robert Gray, M Jorge Cardoso, Sebastien Ourselin, Geraint Rees, John Ashburner, and Parashkev Nachev. Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3d deep generative models. *Medical Image Analysis*, 84:102723, 2023.

[Puyol-Antón *et al.*, 2021] Esther Puyol-Antón, Bram Ruijsink, Stefan K Piechnik, Stefan Neubauer, Steffen E Petersen, Reza Razavi, and Andrew P King. Fairness in cardiac mr image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 413–423. Springer, 2021.

[Stanley *et al.*, 2022] Emma AM Stanley, Matthias Wilms, and Nils D Forkert. Disproportionate subgroup impacts and other challenges of fairness in artificial intelligence for medical image analysis. In *Workshop on the Ethical and Philosophical Issues in Medical Imaging*, pages 14–25. Springer, 2022.

[Wang *et al.*, 2024] Ke Wang, Ningyuan Shan, Henry Gouk, and Iris Szu-Szu Ho. Skin malignancy classification using patients' skin images and meta-data: Multimodal fusion for improving fairness. In *Medical Imaging with Deep Learning*, 2024.

[Wu *et al.*, 2022] Yawen Wu, Dewen Zeng, Xiaowei Xu, Yiyu Shi, and Jingtong Hu. Fairprune: Achieving fairness through pruning for dermatological disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 743–753. Springer, 2022.

[Xu *et al.*, 2023a] Zikang Xu, Jun Li, Qingsong Yao, Han Li, and S Kevin Zhou. Fairness in medical image analysis and healthcare: A literature survey. *Authorea Preprints*, 2023.

[Xu *et al.*, 2023b] Zikang Xu, Shang Zhao, Quan Quan, Qingsong Yao, and S Kevin Zhou. Fairadabn: Mitigating unfairness with adaptive batch normalization and its application to dermatological disease classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 307–317. Springer, 2023.

[Xu *et al.*, 2024] Zikang Xu, Jun Li, Qingsong Yao, Han Li, Mingyue Zhao, and S Kevin Zhou. Addressing fairness issues in deep learning-based medical image analysis: a systematic review. *npj Digital Medicine*, 7(1):286, 2024.

[Ye *et al.*, 2024] Zichen Ye, Daqian Zhang, Yuankai Zhao, Mingyang Chen, Huike Wang, Samuel Seery, Yimin Qu, Peng Xue, and Yu Jiang. Deep learning algorithms for melanoma detection using dermoscopic imaging: A systematic review and meta-analysis. *Artificial Intelligence in Medicine*, page 102934, 2024.

[Zhang *et al.*, 2022] Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. Improving the fairness of chest x-ray classifiers. In *Conference on health, inference, and learning*, pages 204–233. PMLR, 2022.

[Zhou *et al.*, 2021] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

[Zhu *et al.*, 2022] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2022.