Measuring AI Fairness in a Continuum Maintaining Nuances: A Robustness Case Study

Kuniko Paxton, *University of Hull, Hull, East Riding of Yorkshire, HU6 7RX, UK*

Koorosh Aslansefat, *University of Hull, Hull, East Riding of Yorkshire, HU6 7RX, UK*

Dhavalkumar Thakker, *University of Hull, Hull, East Riding of Yorkshire, HU6 7RX, UK*

Yiannis Papadopoulos, *University of Hull, Hull, East Riding of Yorkshire, HU6 7RX, UK*

Abstract—As machine learning is increasingly making decisions about hiring or healthcare, we want AI to treat ethnic and socioeconomic groups fairly. Fairness is currently measured by comparing the average accuracy of reasoning across groups. We argue that improved measurement is possible on a continuum and without averaging, with the advantage that nuances could be observed within groups. Through the example of skin cancer diagnosis, we illustrate a new statistical method that works on multi-dimensional data and treats fairness in a continuum. We outline this new approach and focus on its robustness against three types of adversarial attacks. Indeed, such attacks can influence data in ways that may cause different levels of misdiagnosis for different skin tones, thereby distorting fairness. Our results reveal nuances that would not be evident in a strictly categorical approach.

Keywords: *Fairness, Adversarial attack, Skin lesion classification, Deep learning, Evaluation system*

The use of machine learning (ML) is spreading across critical domains, including sectors like healthcare, necessitating robust governance to ensure safe and practical application. This has spurred increased research into AI fairness, integrating considerations of bias and fairness into the ML implementation process [1]. Sensitive Attributes (SA) that could cause bias can be classified into three types: a) categorical values, i.e. discrete values, such as gender or nationality; b) continuous numerical values, such as age, where each piece of data takes on a single scalar value; and c) continuous data represented in a multidimensional array, for instance, a skin colour image consists of height, width and channel for each pixel. Despite the increasing number of studies on ML fairness regarding these SAs, significant challenges remain unresolved. Firstly, there is relatively little research on non-categorical data. Secondly, non-categorical data such as skin colour has typically been categorised into groups and this process is subject to flaws h iding nuances of bias within groups. Thirdly, to our knowledge, there are no methods to measure the bias in non-categorial data in multi-dimensional arrays, such as skin colour. In this paper, we focus on these issues and propose a new approach for evaluating fairness on multi-dimensional non-categorial data employing statistical distance methods to account for nuances within groups.

A second area of concern addressed in the paper is the robustness of adversarial attacks. Previous ML fairness studies centred on ensuring fairness in classification p erformance. H owever, t here a re now demands to guarantee fairness in various challenging contexts. Of these, a particular issue is robustness to changes in the data. Robustness in ML has two large key dimensions: uncertainty and resistance to adversarial attacks. Adversarial attacks in ML (AAML) can maliciously change predictions. Given that MLs are used in serious settings such as medicine, vulnerability against AAML is a major challenge for real-world applications. Furthermore, it can be said to

be discriminatory if the vulnerability to poor classification performance under AAML varies according to SA values and subgroups. Many studies have studied discrimination against classification performance, but few have assessed fairness against robustness. In this study, we focus on the robustness of ML fairness to adversarial attacks.

Our scenario focuses on skin colour as a sensitive attribute in a skin lesion classification model for the medical diagnosis of cancers such as melanoma. This application encompasses all the complexities previously discussed. First, skin color, closely linked to ethnicity, serves as a critical sensitive attribute. In fairness research, skin colour is categorized using tools like the Fitzpatrick scale [2], which classifies skin into six colour categories [3]. However, skin colour is an elusive characteristic, making such categorisation highly problematic. Categories may not accurately reflect the skin tones analyzed by the ML algorithm. Hidden biases may exist within groups and remain undetected. Furthermore, images, being inherently vulnerable to adversarial attacks, necessitate a thorough exploration of robustness. These challenges are addressed in our study through a new approach where skin colour is treated as a continuum, uncategorized, allowing fairness to be assessed at every point in the spectrum from very dark to very light skin. In experiments, we compare this new approach with current approaches to measuring fairness and demonstrate that this new non-categorical approach may tackle the difficulties discussed in this section. The study also focuses on evaluating fairness in the context of AAML. Specific research questions addressed are as follows.

- **RQ1:** How do statistical distance measures improve the evaluation of fairness in multidimensional data such as skin colour in comparison to existing approaches?

In this context:

- **RQ2:** How can an AAML affect fairness?
- **RQ3:** What types of AAML affect fairness the most?

In addressing these RQs, the paper makes the following contributions:

- Through experiments, we show that our method more accurately captures the nuances of skin tone, as demonstrated by comparing the rates of accuracy (and fairness) degradation across these approaches.
- We analyze how predictive accuracy and fairness for different skin colours are impacted by adversarial attack in a comparison of three approaches, including our own.
- We highlight the diagnostic risks associated with skin colour in skin lesion image classification, emphasising the disparity in diagnostic accuracies between lighter and darker skin tones due to adversarial attack in our experimented dataset.

The rest of the paper is organised as follows: first, we discuss related work, and then we present our approach, describe experiments and compare performance against earlier approaches, such as the average Individual Typology Angle (ITA). Our proposal quantifies ITA as a distribution on a per-pixel basis, offering a more detailed and continuous measurement. The experiment and results sections discuss the use of two different models and two different datasets to show the capabilities of the proposed method. Finally, we conclude and discuss future work.

RELATED WORK

We focus on skin colour as a sensitive attribute in fairness studies and shed light on adversarial attacks that could potentially impact fairness.

Multi-dimensional array SA: Skin color

While skin colour can introduce biases, it is widely recognised that skin tone is an unstable measure and introduces challenges in comprehension, representation, and ultimately, the reasoning of ML algorithms [4]. Categorisation methods have been employed to identify biases associated with skin colour. This involved converting skin images into the CIELAB colour space, calculating the average pigment, and determining the ITA value from this average colour. The ITA is defined by the formula displayed in Figure 2 - (A), where 'L' represents lightness and 'b' represents chromaticity, indicating the colour's hue. However, these methods lack the gradations and nuances of colour that exist in the skin colour categories, so they cannot solve the problem of skin colour representation.

Fairness in Skin Lesion Classification

Several studies assessed the fairness of the skin lesion classification task by skin colour. For example, [3] proposed a technique to assign accurate skin-colour tone labels, thereby attempting to improve fairness. This research did not exceed the area of categorisation. Our method does not require labelling and does not require annotation data. FairDisCo was proposed to improve the fairness of lesion classification by separately learning SA information, such as skin colour [5]. In this

paper, the Fitzpatrick-based skin colour category was used for the evaluation. Therefore, discrimination of differences based on individual skin colouration is not guaranteed.

Adversarial attacks to fairness

Several studies have explored adversarial attacks on fairness. Van et al. demonstrated a poison attack that compromised both accuracy and fairness by using adversarial samples, label flipping, and feature modification [6]. Mehrabi et al. investigated two types of attacks targeting fairness [7]: the influence attack, which maximises the covariance between sensitive attributes and class labels, damaging both accuracy and fairness. The anchoring attack altered the decision boundaries and affected fairness without impacting accuracy. Solans et al. developed a poisoning attack designed to impact the fairness of a model without discernible effects on its overall performance, thereby concealing fairness violations from users [8]. These studies indicate that attacks focusing solely on degrading classification performance do not uniformly affect fairness, underscoring the complexity of detecting fairness violations. Ghosh et al. introduced their adversarial attack by targeting a text-to-image ranking system [9]. The attack was specifically designed to aim to elevate individuals with light skin tones to a higher ranking. This attack intentionally succeeded in creating unfairness.

These studies suggest that fairness is not only about classification performance but also about the importance of understanding robustness. Existing research studies on ensuring fairness have used a categorial approach in which sensitive attributes are assigned discrete values that separate groups of people. Similarly, studies focusing on robustness to adversarial attacks on ML with respect to fairness have also adhered to a categorial approach. We depart from this tradition to examine robustness when sensitive attributes are used as uncategorized continuous numerical values. This is the first study that explores how adversarial attacks aimed at accuracy degradation affect fairness using a statistical approach without using traditional categorization.

PROPOSED METHODOLOGY

The methodology for a more nuanced evaluation of fairness is described in Figure 2.

Our key innovation is to treat ITA values as continuous numerical variables and use this to represent the nuance of skin colour in each image as a distribution. More precisely, at first, we calculate the ITA value for individual pixels and instead of averaging values across the image, the computed ITA values are treated as a distribution. Then, we select a baseline image. To simplify the visualisation of results, in this case, we selected images with the minimum mean ITA as a baseline. The distance of an image from the baseline is established by measuring the distance of its ITA distribution from that of the baseline using statistical distance metrics. In the result section, we demonstrate this using the Wasserstein Distance (WD) measure.

Using this new approach, we examined the effect on the fairness of three types of adversarial attacks.

- **Adversarial Random Noise (RN):** The attack applies random noise to the entire test image which does not require access to the model itself. It is generally accepted that random noise attacks are relatively weaker compared to perturbation-based models.
- **Fast Gradient Sign Method (FGSM):** FGSM is a well-known adversarial attack method [10]. The attack takes the loss function of Deep Learning(DL) to create images that maximise the loss value. It adds slight perturbations on a pixel basis.
- **Adversarial Saliency Map Patch (SMP):** Firstly, we identified the model attention area by the saliency map proposed by [11]. Noise is added only on the focus area to cause model misclassification. We call this an SMP attack. We generated the SMP to monitor the behaviour of local area attacks rather than applying it to all image areas.

Three measures of AAML success rates were considered:

1) **Attack Success Rate (SR):** The ratio of results incorrectly predicted under AAML that were correctly predicted in the original image.
2) **Attack False Positive Success Rate (FP SR):** The percentage of false positives specifically caused by AAML over correct predictions of no cancer in the original undisturbed images.
3) **Attack False Negative Success Rate (FN SR):** The percentage of false negatives specifically caused by AAML over correct predictions of cancer in the original images.

Statuses other than those listed above are not regarded as AAML successes.

EXPERIMENT

Dataset

Two publicly available datasets are used to demonstrate the capabilities and limitations of the proposed approach. A) Human Against Machine with 10000 training images (HAM10000) [12]: This dataset was synthetically created with input from several hospitals and validation by medical professionals. The HAM10000 dataset is a comprehensive collection of dermatoscopic images aimed at supporting research in the automated diagnosis of skin lesions, particularly melanoma. It consists of 10,000 dermatoscopic images labelled with various skin lesion types, making it a valuable resource for training and evaluating skin ML in dermatology. B) Fitzpatrick17K [13], [9]: This is a skin lesion image dataset with Fitzpatrick skin type labels assigned by dermatology experts, obtained from two dermatology atlas.

The HAM10000 exhibits a notable bias towards lighter skin tones, which reflects the higher incidence rates of skin cancer in these populations [14]. We focused on two classes among seven classes because other classes had the possibility of not containing dark-coloured skin (skin colour type "Dark" was only 0.7%, and type "Brown" was 0.9% of all datasets) in the process of data split due to the number of data. The two classes are "Melanocytic Nevi", which is non-cancerous and "Melanoma", which is one type of cancer. The number of data is for training (n=2,000), validation (n=1,563) and test (n=1,564). In terms of Fitzpatrick17K, we chose "benign" and "malignant" out of three_partition_label to make it a binary classification task. The number of data is for training (n=2,000), validation (n=864) and test (n=864). There are differences between the HAM10000 and the Fitzpatrick 17K dataset: the HAM10000 is an image with the lesion centred and with some uniformity. Furthermore, the segmentation data provided makes it very easy and accurate to identify lesions and skin-tone areas. In comparison, the Fitzpatrick 17K had random lesion locations and contained images in which non-skin-coloured areas occupy the majority of the image.

Model Architecture

Two model architectures are Convolutional Neural Network (CNN) and Residual Neural Network (ResNet). The setting details of those two models are available on our GitHub (https://github.com/Kuniko925/FairRobustness).
General performances are in the middle of Figure 1.

TABLE 1. AAML Success Rates (%) each Model and Dataset

Dataset	HAM10000		Fitzpatrick17K	
Model	CNN	ResNet50	CNN	ResNet50
RN	63.80	11.1	26.3	34.0
RN FP	63.4	0.01	18.2	0.8
RN FN	0.3	10.9	8.1	33.2
FGSM	43.0	24.8	59.8	63.3
FGSM FP	32.2	16.3	27.0	28.9
FGSM FN	10.8	8.5	32.7	34.3
SMP	33.9	3.1	17.7	8.1
SMP FP	32.3	1.4	12.6	7.4
SMP FN	1.5	1.7	5.0	0.7
Test ACC	0.77	0.88	0.6	0.8
Test F1-Score	0.66	0.79	0.6	0.79
Skin Colour Types				
Very Light(n)	1345		186	
Light(n)	124		259	
Intermediate(n)	53		205	
Tan(n)	29		135	
Brown(n)	5		63	
Dark(n)	8		16	

Adversarial examples

In accordance with the outlined methodology, three adversarial examples were generated. We randomly decided hyperparameters in adversarial examples to prevent intentionality. The hyperparameters are constant in all models and datasets. The detailed settings are in Figure 2 - (B). The adversarial examples' images are Figure 2 - (C). We attacked models with adversarial examples and observed the impact of performance on skin colour.

RESULTS

Results show variability in attack success rates and model performance in Table 1. In the two models we selected, ResNet50, a relatively newly proposed model, had better classification accuracy than CNN. ResNet50 had accuracies of over 80 % in both the HAM10000 and Fitzpatrick 17K datasets. The F1 score was also high, at 79 % for both classes. In the CNN model, the accuracy kept high performance, but the F1 score dropped to around 66 % (HAM10000) and 60 % (Fitzpatrick17K). Here, we discuss further analysis of each skin colour measure.

Conventional Method: Skin Tone

The AAML success rate for the traditional skin colour measurement with HAM10000 is shown in Figure 1. HAM10000 has general RN attacks at 87% and 81% for skin colour types Light and Intermediate. Very Light and Brown, which differ greatly in the number of data, had comparable success rates. FGSM was the only one with successful attacks in all skin-tone categories.

The highest success rate was 50%, which is not a high success rate overall, but even FN showed a success rate of 12% in Very Light, which is also large. SMP attacks showed a high success rate of 80% in Brown, and the success rate decreased as the colour tone became lighter. Furthermore, in our noise range, the type of attack success is concentrated in FP. Next, the Fitzpatrick17K results showed comparable AAML success rates for all skin colour types in all attacks, albeit slightly higher for darker skin colours. In particular, FGSM success rates were higher for all skin tone types equally. There were no differences between the models.

Scalar ITA Value: Average Skin Colour

The Scalar ITA Value is a method used between traditional categorisation and our proposed multi-dimensional array of skin colour. There is no categorisation but a continuous scalar value by averaging the colour tones. The results are in Figure 3. As we see, the AAML success rate is basically similar to that of categorised conventional methods. The combination of ResNet and HAM10000 made it easier to identify trends in skin colour effects compared to categorisation when measured by ITA values. Both adversarial attacks had higher success rates as skin colour became fairer. Furthermore, success rates vary for AAML, even within the same Very Light group. The ITA range for Very Light is greater than 55. In this group, success rates for the combination of CNN and HAM10000 fluctuate. The ITA values most vulnerable to RN attacks are 60 - 80. ITA angles exceeding 100 resulted in robustness against AAML. In the Fitzpatrick17K, there was no significant d ifference b etween a ny o f t he m odels and their AAMLs; the SMP success rate was extremely low and did not discriminate by skin colour. The correlation between ITA angles and AAML success rates is shown in Table 2. HAM10000 showed strong correlations with skin colour for RN and FGSM attacks; the CNN had a negative correlation for RN and a positive correlation for FGSM; the ResNet50 had a positive correlation for both RN and FGSM; the correlation with SMP was low. The SMP was unsuccessful. No differences by skin colour were seen in Fitzpatrick17K. This is because the dataset contained many images that restricted us from extracting the correct skin tones, as shown in Figure 2 - (D). For example, some were zoomed out and had too large a proportion of the background, some had multiple subjects, and some were zoomed-in photographs in which the focus was on something other than the skin.

Multi-dimensional array (Our core approach): Skin Colour Gradation

The results of our method are as Figure 3 and Table 2. The combination of HAM10000 and ResNet tends to have a markedly lower success rate with lighter skin types. Darker skin nuances increase the likelihood of non-cancer images being diagnosed as cancer by the FGSM. RN is a simple black box attack, but it is effective in that darker skin tones can misdiagnose cancer. The case of HAM10000 and ResNet also shows a different trend from ITA values and categorisation. We further explore the correlation between skin colour and attack success rate using our method: for HAM10000, a very strong correlation was shown between the FP success rate for the RN attack and the FN success rate for the FGSM attack. The absolute values of these correlations are higher than for the scalar ITA values: for CNN, the FP success of RN attacks differs by 6 and for ResNet by 14. The maximum correlation for the ITA values is 0.71, whereas our method shows a correlation strength of -0.81. Our method detects a correlation of -0.53 for SMP attacks with HAM10000 and ResNet but only 0.36 for the averaged ITA values. By taking into account our skin colour gradient, the colour feature in the image is more likely to show up as a correlation. As with the average ITA method, Fitzpatrick17K did not allow our method to capture the effect of skin colour due to the difficulty of extracting skin colour.

In the final part of this section, we respond to our RQs. **RQ1:** The statistical distance measure, using the pigment gradient across skin pixels as the measurement criterion, was able to identify differences in performance within the same subgroup that could not be ascertained by categorisation. Moreover, it strengthened features and sharpened correlations more than a single averaged ITA value. **RQ2:** AAML showed a different degradation of classification performance with skin colour in our method. This suggested a bias in robustness depending on the lightness or darkness of the skin colour and its characteristics. **RQ3:** RN attacks can easily and extensively affect diagnosis. However, for medical applications such as skin lesion classification, it is necessary to consider defences against FGSM attacks, which have high FN success rates, lead to overlooking malignant cancer under the measure by our method.

CONCLUSION

This study introduced a new method for assessing fairness in machine learning models used for skin cancer diagnosis, focusing on how these models per-

TABLE 2. Correlation between AAML Success Rates and ITA or Skin gradation

Average skin color (Scalar)										
Dataset	Model	RN	RN FP	RN FN	FGSM	FGSM FP	FGSM FN	SMP	SMP FP	SMP FN
HAM10000	CNN	-0.53	-0.53	0.10	0.59	0.16	0.55	-0.21	-0.26	0.31
	ResNet50	0.55	0.00	0.56	0.71	0.64	0.51	0.36	0.31	0.24
Fitzpatrick17K	CNN	-0.01	-0.20	0.30	-0.14	-0.19	0.02	0.35	0.18	0.23
	ResNet50	-0.11	0.11	-0.13	0.08	0.14	-0.04	0.22	0.16	0.10
The Proposed Method: Skin gradation (Multi-dimensional Array)										
Dataset	Model	RN	RN FP	RN FN	FGSM	FGSM FP	FGSM FN	SMP	SMP FP	SMP FN
HAM10000	CNN	0.59	0.59	-0.04	-0.57	-0.13	-0.66	0.15	0.22	-0.33
	ResNet50	-0.69	-0.01	-0.69	-0.81	-0.75	-0.58	-0.53	-0.39	-0.37
Fitzpatrick17K	CNN	0.01	0.28	-0.34	0.23	0.29	-0.06	-0.27	-0.13	-0.27
	ResNet50	0.14	-0.15	0.16	0.04	-0.01	0.05	-0.17	-0.15	-0.09

form under three types of adversarial attack and taking into account variations in skin colour more precisely than previous methods. The three types of AAML are FGSM, a gradient-based white-box attack; SMPs using explainability; and random noise, a simple black-box attack. We compared the effects of those attacks on accuracy and fairness. Results indicate that the model's accuracy and fairness can be significantly impacted by these attacks, with varying risks of misdiagnosis based on skin colour. Our method exposed vulnerabilities under AAML within the same skin colour group that cannot be detected by conventional skin colour categorisation methods. Compared to scalar ITA values, which average skin image, our method, which takes into account the nuances of skin gradation, succeeded in finding sharper correlations in the skin's characteristics.

FUTURE WORK

We consider four future works from our results: (1) We build a bias mitigation system which is fair regarding robustness and performance for the nuanced, sensitive attribute values. (2) When unfairness is detected, models should be in place to trace the causality. We apply explainability techniques to track down the causes by comparing the difference between the original skin image and a counterfactual image with a different skin colour. (3) Our mechanism should be applied not only to image classifiers but also to image-to-image generative models. (4) We will incorporate expert evaluations to further assess the effectiveness of our proposed method.

CODE AVAILABILITY

Regarding the research reproducibility, codes and functions supporting this paper are published online at GitHub: https://github.com/FairRobustness.

ACKNOWLEDGMENTS

The authors would like to thank the Dependable Intelligence Systems Lab, the Responsible AI Hull Research Group, and the Data Science, Artificial Intelligence, and Modelling (DAIM) Institute at the University of Hull for their support. Furthermore, the author extends heartfelt gratitude to Professor Balaraman Ravindran of the Indian Institute of Technology Madras, whose invaluable provision of the initial research idea has been the cornerstone of this study.

REFERENCES

1. S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Computing Surveys*, 2020.
2. T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types i through vi," *Archives of dermatology*, vol. 124, no. 6, pp. 869–871, 1988.
3. P. J. Bevan and A. Atapour-Abarghouei, "Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification," in *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pp. 1–11, Springer, 2022.
4. C. M. Heldreth, E. P. Monk, A. T. Clark, C. Schumann, X. Eyee, and S. Ricco, "Which skin tone measures are the most inclusive? an investigation of skin tone measures for artificial intelligence," *ACM Journal on Responsible Computing*, vol. 1, no. 1, pp. 1–21, 2024.
5. S. Du, B. Hers, N. Bayasi, G. Hamarneh, and R. Garbi, "Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning," in *European Conference on Computer Vision*, pp. 185–202, Springer, 2022.
6. M.-H. Van, W. Du, X. Wu, and A. Lu, "Poisoning attacks on fair machine learning," in *International Conference on Database Systems for Advanced Applications*, pp. 370–386, Springer, 2022.
7. N. Mehrabi, M. Naveed, F. Morstatter, and A. Galstyan, "Exacerbating algorithmic bias through fair-

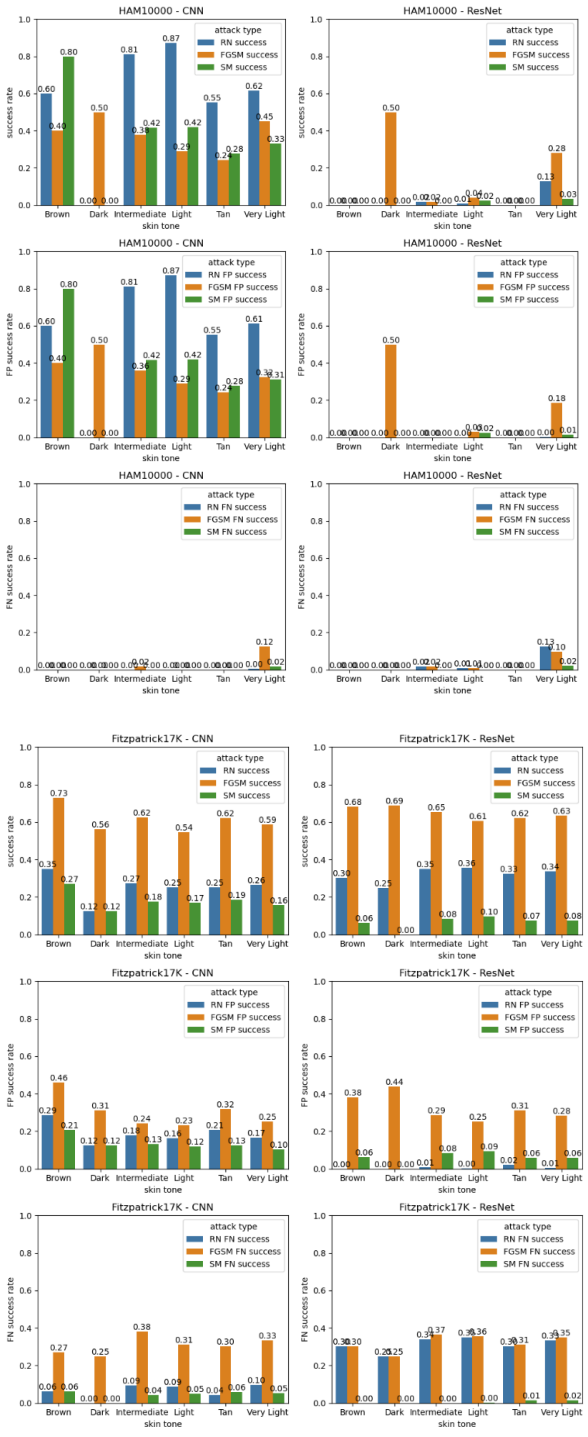


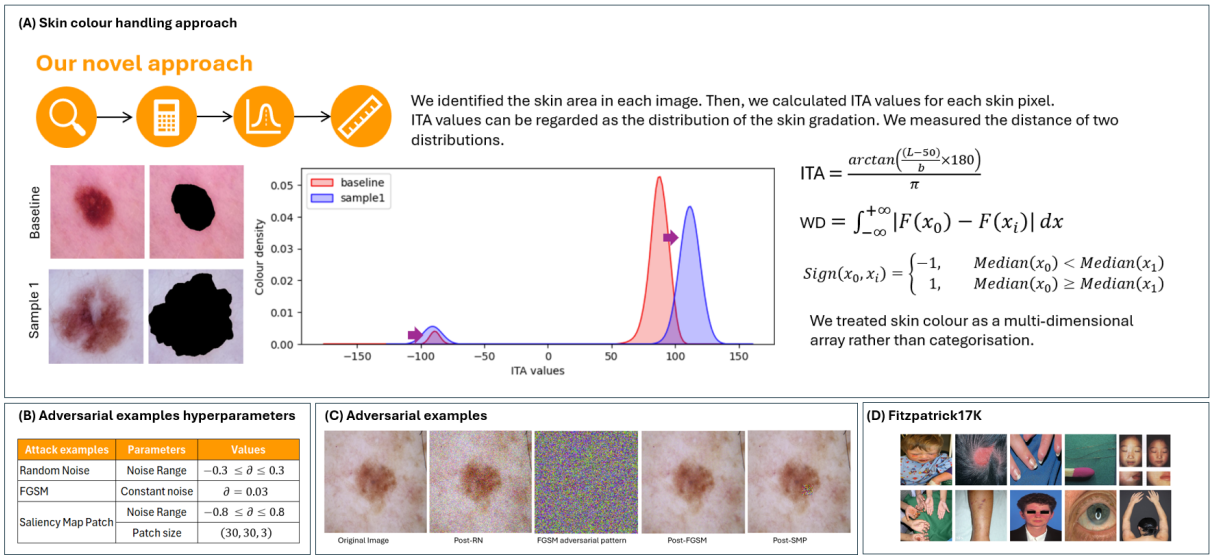
FIGURE 1. Traditional method: Adversarial success rate vs Skin tones. The first 6 plots are with HAM10000, and the later 6 plots are with Fitzpatrick17K.

ness attacks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 8930–8938, 2021.

8. D. Solans, B. Biggio, and C. Castillo, "Poisoning attacks on algorithmic fairness," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 162–177, Springer, 2020.
9. A. Ghosh, M. Jagielski, and C. Wilson, "Subverting fair image search with generative adversarial perturbations," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 637–650, 2022.
10. J. Sen and S. Dasgupta, "Adversarial attacks on image classification models: Fgsm and patch attacks and their impact," *arXiv preprint arXiv:2307.02055*, 2023.
11. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833, Springer, 2014.
12. P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, *et al.*, "Human–computer collaboration for skin cancer recognition," *Nature Medicine*, vol. 26, no. 8, pp. 1229–1234, 2020.
13. M. Groh, C. Harris, R. Daneshjou, O. Badri, and A. Koochek, "Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–26, 2022.
14. D. Wen, S. M. Khan, A. J. Xu, H. Ibrahim, L. Smith, J. Caballero, L. Zepeda, C. de Blas Perez, A. K. Denniston, X. Liu, *et al.*, "Characteristics of publicly available skin cancer image datasets: a systematic review," *The Lancet Digital Health*, vol. 4, no. 1, pp. e64–e74, 2022.

Kuniko Paxton is a PhD candidate in Computer Science at the University of Hull, and the funding institution is DAIM. The research interests are Fairness in AI, explainability and Adversarial attacks. Contact her at k.azuma-2021@hull.ac.uk

Koorosh Aslansefat is an assistant professor of computer science at the University of Hull, HU6 7RX Hull, U.K., affiliated with the Dependable Intelligent System Group. His research interests span artificial intelligence safety, Markov modelling, and real-time dependability analysis. Aslansefat received his PhD in computer science from the University of Hull. He is a Member of IEEE. Contact him at K.Aslansefat@hull.ac.uk



Attack examples	Parameters	Values
Random Noise	Noise Range	$-0.3 \leq \partial \leq 0.3$
FGSM	Constant noise	$\partial = 0.03$
Saliency Map Patch	Noise Range	$-0.8 \leq \partial \leq 0.8$
	Patch size	$(30, 30, 3)$

FIGURE 2. (A) This is a process of measuring skin colour nuance and the samples. Original images masked skin lesion pixels. After calculating the ITA values of each pixel and regarding the distributions, the baseline and sample 1 distribution have a gap between images. WD is used to measure the distance between distributions. The distribution medians are compared to identify the direction of distance. (B). The table summarizes adversarial examples of hyperparameters. These parameter values are selected after several trials. (C) of images are adversarial examples. Left is the original image. The 3 types of adversarial patterns were applied to the original image. (D) Example images in the Fitzpatrick17K dataset cannot detect skin pixels properly.

Dhaval Thakker is a Professor of Artificial Intelligence (AI) and the Internet of Things (IoT) at the University of Hull, where he leads a group focused on Responsible Artificial Intelligence. His research emphasises AI Explainability, AI Safety, and Fairness. With nearly two decades of experience, Dhaval has been at the forefront of innovative solutions through funded projects. His interdisciplinary research spans Generative AI and the applications of Edge computing alongside IoT technologies. He has a track record in leveraging AI for Social Good, notably in Smart Cities, Digital Health, and the Circular Economy. Contact him at D.Thakker@hull.ac.uk

Yiannis Papadopoulos is Professor of Computer Science and Leader of the Dependable Intelligent Systems (DEIS) Research Group at the University of Hull in the U.K. For over 30 years, Papadopoulos and his research group have pioneered cutting-edge model-based, bio-inspired and statistical technologies for the analysis and design of dependable engineering systems, with a recent intense focus and contributions towards achieving trustworthy, safe AI. Many of the software tools that they have crafted, including HiP-HOPS and EAST-ADL, have become commercial and are used in transport and other industries. Contact him at Y.I.Papadopoulos@hull.ac.uk

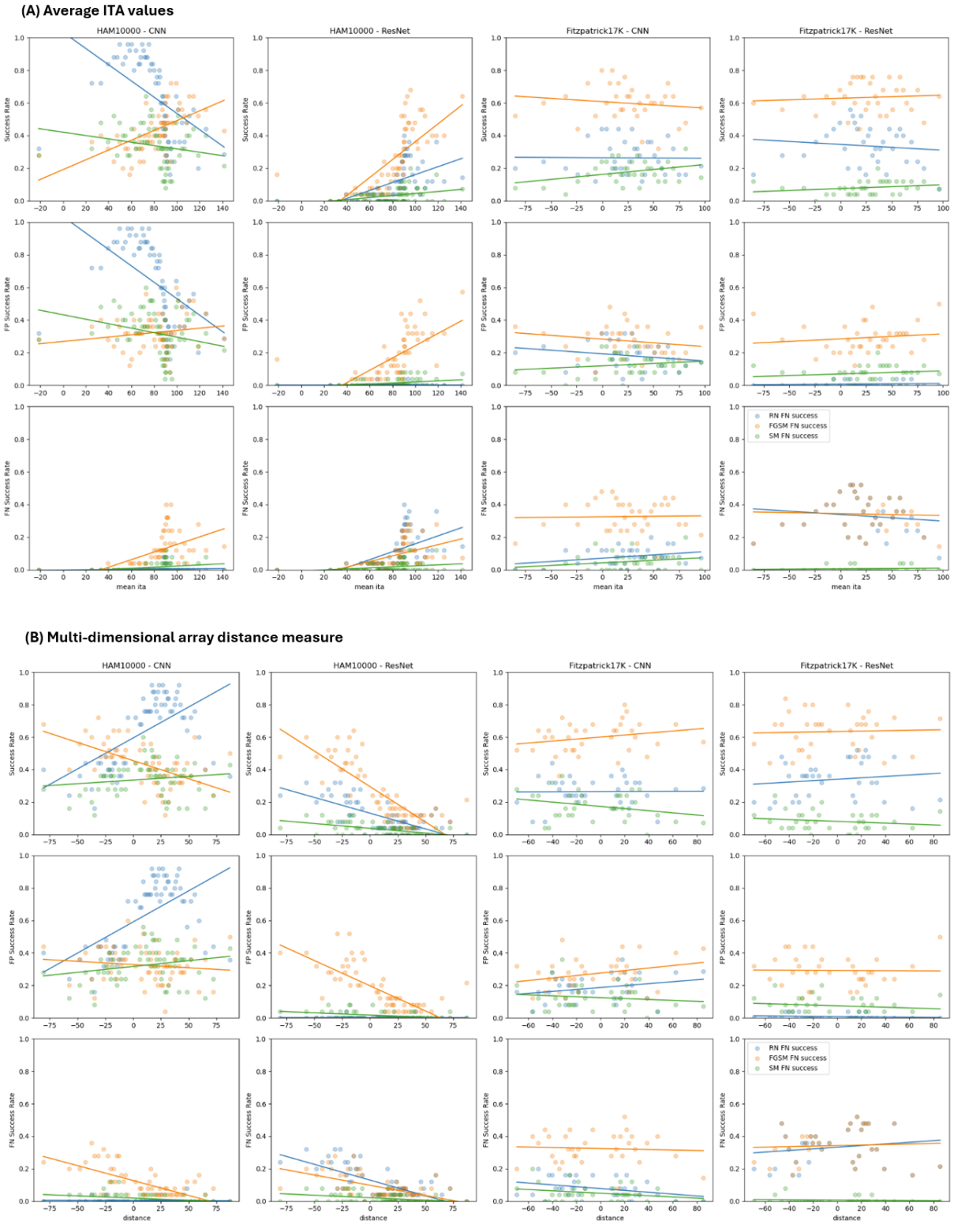


FIGURE 3. AAML results versus (a) ITA scalar skin colour and (b) Skin colour gradation