
Mitigating Class Imbalance and Promoting Fairness in Medical Image Classification with Limited Data Using Deep Learning

Rahmat Izwan Heroza

A thesis submitted for the degree of Doctor of Philosophy

School of Computer Science and Electronic Engineering

University of Essex

January 2026

Abstract

Class imbalance, limited data, and demographic bias are significant challenges in medical image classification, particularly in skin lesion analysis, where heterogeneous populations and privacy constraints hinder the development of robust and fair models. This thesis proposes deep learning strategies that improve classification performance, robustness, and fairness in both centralised and federated settings.

First, SIA-SMOTE is introduced as a Siamese network-based oversampling method for imbalanced image data. By exploiting feature similarity in high-dimensional spaces, it generates more representative synthetic samples than conventional SMOTE variants, leading to consistent performance gains across benchmark and medical imaging datasets.

Second, a self-attention fusion strategy for Vision Transformers is proposed to enhance skin lesion classification under imbalanced conditions. By fusing complementary attention mechanisms at the final transformer stage while preserving pretrained representations, the method improves balanced accuracy and F1 performance on the ISIC 2017 dataset.

Third, this thesis presents FedCIAL, a federated colour-invariant adversarial learning framework for mitigating demographic bias in decentralised skin lesion classification. Using Fitzpatrick skin type transformations and adversarial domain adaptation, FedCIAL improves both accuracy and fairness while preserving data privacy on the Fitzpatrick17k dataset.

Finally, a fairness-aware federated learning framework combining domain adaptation and class imbalance mitigation is developed and evaluated on the Diverse Dermatology Images (DDI) dataset. Results show that jointly addressing imbalance and domain shift yields synergistic improvements in subgroup fairness and overall performance.

Overall, this thesis advances oversampling, transformer-based representation learning, and fairness-aware federated learning for developing robust and equitable medical imaging systems.

Acknowledgment

All praise and gratitude are due to Allah, the Most Merciful and the Most Compassionate, for granting me strength, patience, and guidance throughout this long and challenging journey. Without His blessings, this work would not have been possible. May blessings and peace always be upon the noble Prophet Muhammad, peace be upon him.

I would like to express my deepest gratitude to my parents, Ibrahim Aliarab and Helmiaty Cik Aba, whose constant prayers, unconditional love, and unwavering belief in me have been my greatest source of strength. Their sacrifices and encouragement have carried me through every stage of my academic journey.

My heartfelt thanks go to my beloved wife, Adithia Hardianti Pertiwi, whose prayers, patience, and countless sacrifices made this journey possible. Her constant support, understanding, and encouragement sustained me through difficult moments. I am also deeply grateful to my children, Hasna Shofiyah, Ibrahim Syakir, and Ahmad Dzakir, who accompanied me throughout this journey. In particular, my youngest, who often sat on my lap while I worked on my laptop late at night, unknowingly helped keep me awake and motivated to complete this study. Those moments will always remain close to my heart.

I would like to sincerely thank my supervisors for their invaluable guidance and support. I am especially grateful to Professor John Q. Gan for opening the door for me, believing in my potential, and accepting my research proposal, which allowed me to begin this PhD journey. His trust and encouragement laid the foundation for everything that followed. I would also like to thank Dr Haider Raza for being an exemplary supervisor, for his consistent guidance, constructive feedback, academic rigour, and genuine care for both my research progress and personal development. His mentorship significantly shaped my growth as a researcher.

I would like to extend my sincere thanks to Professor Luca Citi and the BCI Lab for providing

a stimulating research environment and valuable academic support throughout my PhD. I am particularly grateful for access to the BCI Lab's GPU computing resources, which I used during the later stages of this research under Professor Citi's management and support.

My sincere appreciation goes to Stuart for his outstanding technical support and for administering the Ceres HPC system, which provided the GPU resources used during the early stages of this research. Access to the Ceres HPC played a crucial role in enabling the initial large-scale experiments conducted in this thesis.

I would like to extend my sincere thanks to my Indonesian colleague, Rifqy Hakimi, who was the first person to welcome me to Colchester. I am also deeply thankful to him and his family, as well as all Indonesian families here, who became my extended family during this journey, including the Zulkarnain family, the Zul Arief family, Roman family, and Girindra. Their warmth, friendship, and support made this journey far more meaningful and less lonely.

I would also like to express my sincere appreciation to Hanif Rasyidi, who, during the final stage of this PhD journey, spent a great deal of time discussing my work with me, providing motivation, encouragement, and a much-needed push to ensure that I could complete and submit this thesis on time.

I would like to thank my office mates, Stephen and Dat, for their companionship and support. Our regular weekend game workshops provided a much-needed balance between research and relaxation, helping me maintain both motivation and well-being throughout the PhD.

I am also grateful to my home institution, Universitas Sriwijaya, for its support and for granting me the opportunity to pursue my doctoral studies abroad. Their trust and sponsorship were instrumental in making this academic journey possible.

Finally, I would like to acknowledge everyone who supported me directly or indirectly throughout this journey. Any omissions are unintentional, and I remain deeply grateful to all who contributed in their own way to the completion of this thesis.

Contents

Abstract	2
Contents	5
List of Figures	10
List of Tables	12
1 Introduction	13
1.1 Problem Statement	13
1.2 Research Questions	15
1.3 Research Objectives	16
1.4 Contributions	17
2 Literature Review	18
2.1 Introduction to Medical Image Classification	18
2.2 Class Imbalance in Medical Imaging	20
2.2.1 Causes of Class Imbalance in Medical Imaging Datasets	20
2.2.2 Techniques for Mitigating Class Imbalance	21
2.2.3 Challenges in the Medical Imaging Domain	26
2.3 Fairness in Medical AI	27
2.3.1 Sources of Bias in Medical Image Datasets	28
2.3.2 Approaches to Mitigating Fairness Issues	29
2.3.3 Fairness in Federated Medical Imaging	31
2.4 Learning with Limited Data in Medical Imaging	32

2.4.1	Causes of Limited Data Availability	33
2.4.2	Impact on Deep Learning Performance	33
2.4.3	Existing Strategies for Learning with Limited Data	34
2.4.4	Limitations of Current Methods	34
2.5	Deep Learning Architectures and Techniques in Medical Image Classification .	35
2.5.1	Convolutional Neural Networks in Medical Imaging	35
2.5.2	Oversampling and Synthetic Data Generation	35
2.5.3	Attention Mechanisms and the Emergence of Vision Transformers . . .	36
2.5.4	Federated Learning and Privacy Preserving Model Training	38
2.6	Critical Summary and Research Gaps	39
3	Oversampling High Dimensional Medical Images Using SIA-SMOTE for Class Imbalance Mitigation	42
3.1	Introduction	42
3.2	Related Work	45
3.3	Method	46
3.3.1	Base Sample Selection	47
3.3.2	Image Synthesis	47
3.3.3	Image Selection with a Siamese Network	48
3.3.4	Final Oversampling with SMOTE	49
3.4	Experimental Setup	49
3.4.1	Datasets	50
3.4.2	Evaluation Metrics	51
3.4.3	Experimental Configuration	52
3.5	Results and Discussion	53
3.5.1	Quantitative Results	53
3.5.2	Qualitative Analysis of Filtered Images	54
3.5.3	Statistical Significance Testing	54
3.5.4	Interpretation in the Context of Medical Imaging	54
3.5.5	Strengths and Limitations	55

3.6	Conclusion	55
4	Self-Attention Fusion with Vision Transformer for Imbalanced Skin Lesion Classification	59
4.1	Introduction	59
4.2	Related Work	61
4.3	Method	62
4.3.1	Network Architecture with Self-Attention Fusion	63
4.3.2	Loss Functions	66
4.4	Experimental Setup	68
4.4.1	Dataset and Performance Metrics	68
4.4.2	Experimental Configuration	69
4.5	Results and Discussion	70
4.5.1	Comparison of Loss Functions	70
4.5.2	Attention Mechanism Modification and Visualisation	71
4.5.3	Quantitative Results with Fusion Strategies	72
4.5.4	Statistical Significance Analysis	74
4.5.5	Comparison with ISIC 2017 Challenge Leaderboard	74
4.5.6	Discussion	75
4.6	Conclusion	76
5	Federated Colour-Invariant Adversarial Learning for Fair and Robust Skin Lesion Classification	77
5.1	Introduction	77
5.2	Related Work	80
5.2.1	Feature Skew in Federated Learning	80
5.2.2	Bias Mitigation in Machine Learning	81
5.2.3	Federated Representation Learning	82
5.3	Method	83
5.3.1	Preliminaries	83

5.3.2	Federated Colour-Invariant Adversarial Learning	84
5.4	Experimental Setup	87
5.4.1	Dataset	88
5.4.2	Baseline and Comparative Methods	89
5.4.3	Implementation Details	90
5.5	Results and Discussion	90
5.5.1	Classification Performance	90
5.5.2	Fairness Across Clients	91
5.5.3	Feature Alignment Analysis	91
5.6	Conclusion	92
6	Fairness-Aware Federated Learning with Domain Adaptation and Class Imbalance Mitigation for Skin Lesion Classification	94
6.1	Introduction	94
6.2	Related Work	96
6.2.1	Fairness in Dermatological AI	96
6.2.2	Federated Learning for Medical Imaging	97
6.2.3	Domain Adaptation in Medical Imaging and Fairness	97
6.3	Method	98
6.3.1	Fairness-aware Federated Learning	98
6.4	Experimental Setup	100
6.4.1	Dataset and Setup	100
6.4.2	Experimental Design	101
6.4.3	Fairness Metrics	102
6.5	Results	103
6.5.1	Single-Fold Evaluation	103
6.5.2	Statistical Validation (5-Fold Cross-Validation)	103
6.5.3	Ablation Study: Effect of Domain Adaptation and Imbalance Mitigation	104
6.6	Discussion	105
6.7	Conclusion	105

7 Conclusion	109
---------------------	------------

Bibliography	113
---------------------	------------

List of Figures

2.1	Imbalanced dataset before oversampling	22
2.2	Balanced dataset after oversampling using SMOTE	23
2.3	Borderline-SMOTE Oversampling Example Distribution	23
3.1	Interpolation between images from different classes at the decision boundary. Minority samples near the boundary are interpolated with nearby majority class samples to generate synthetic examples that enrich the decision space.	48
3.2	Ranking of oversampling methods in terms of three performance metrics	56
3.3	Comparison between real minority-class images and synthetic samples generated by SIA-SMOTE. Columns 1 and 3 show original minority-class images, Column 2 presents synthetic samples retained after similarity-aware filtering, and Column 4 shows synthetic samples rejected by the filtering process.	58
4.1	The proposed self-attention fusion architecture. The final n attention layers in the Vision Transformer are replaced with fusion layers combining scaled dot product, multiplicative, and additive attention mechanisms.	63
4.2	(a) Original image. (b) Attention maps from two different heads correspond to a skin lesion with a darker and brighter area. The even columns present the mask mapping with a threshold of 0.3. (c) Attention maps from the proposed method	73
5.1	Overview of the proposed FedCIAL framework. The architecture integrates federated training with colour-invariant adversarial learning, mitigating skin tone bias through gradient reversal and explicit Fitzpatrick-based colour transformations.	85
5.2	Comparison between original, transformed, and anchor images.	92

6.1	Flow diagram of the fairness-aware federated learning. Each training batch first undergoes imbalance mitigation, followed by synthetic sample generation. Both original and synthetic samples are then processed through domain adaptation-based feature extraction to achieve colour-invariant representations, and finally, the original image is passed to the classification layer for prediction.	100
6.2	5-fold cross-validation results showing mean $\pm$ standard deviation for all metrics. Error bars highlight that differences are consistent but not statistically significant at $p < 0.05$	107
6.3	Per-fold performance across 5 folds for balanced accuracy and fairness score. Points indicate individual folds; dashed lines show means. Overlapping fold results reflect variance-limiting statistical power.	108

List of Tables

2.1	Summary of datasets used across all work packages	41
3.1	Comparison of performance of oversampling methods	57
3.2	Friedman’s test result	57
3.3	Conover test result	58
4.1	Performance Comparison of Different Loss Functions	71
4.2	Self-attention Fusion Performance on VIT-B16 and DINO-B16	72
4.3	McNemar Test on VIT-B16	74
4.4	McNemar Test on DINO-B16	74
4.5	Comparison with the ISIC 2017 Leaderboard	74
5.1	Predefined Lab colour values for Fitzpatrick skin types (FST). Values follow established ITA mappings, ensuring perceptual consistency between pigmentation levels and Lab components.	86
5.2	Fitzpatrick17k dataset breakdown by Fitzpatrick skin type	88
5.3	Average balanced accuracy and standard deviation across 5-fold cross-validation.	90
5.4	Distance between original, transformed, and anchor images across different metrics.	93
6.1	DDI dataset breakdown by Fitzpatrick skin type group	100
6.2	Experiment Configurations	102
6.3	Single-Fold Comparison: Baseline vs. Our method	103
6.4	Ablation Study Results (Single-Fold, 200 Epochs)	104

Introduction

1.1 Problem Statement

Medical image classification using deep learning has the potential to revolutionise diagnostics and treatment planning by providing accurate and automated interpretations of complex medical images. However, real-world medical datasets are often highly imbalanced, small, and prone to demographic biases.

Medical image datasets are often characterised by substantial class imbalance. This imbalance, where some classes are underrepresented, arises due to the inherent rarity of certain medical conditions, the difficulty of acquiring sufficient representative cases, and the cost and complexity of expert annotation. As a result, deep learning models trained on such datasets tend to exhibit class imbalance bias, favouring majority classes and underperforming on minority or rare conditions because the model considers the minor class as noise, which attempts to disregard it [1]. This leads to skewed predictions, lower recall for critical cases, and an increased risk of diagnostic errors, particularly when detecting less prevalent but clinically significant conditions. In these conditions, the accuracy of the deep learning technique for classifying the minority class tends to degrade [2]. Additionally, the bias may arise from other properties, such as varying skin types in skin cancer analysis, which further complicates the training process.

Compounding this issue is the problem of fairness. Machine learning models, when trained on imbalanced and demographically unrepresentative data, may exhibit algorithmic

bias, where it underperformed for specific subpopulations such as racial or ethnic minorities, gender groups, or age categories. In the medical context, such biases can result in unequal access to accurate diagnostics and treatment recommendations, thereby exacerbating existing health disparities. Ensuring equitable model performance across diverse patient groups is not just a technical requirement, but an ethical imperative in the deployment of medical AI systems. In this thesis, fairness refers to the ability of a medical image classification model to achieve consistent and reliable performance across different demographic subgroups without systematically disadvantaging underrepresented populations. Bias refers to systematic performance disparities. Two forms of bias are considered throughout this thesis: (1) class imbalance bias, where models favour majority disease classes over minority classes, and (2) demographic bias, where models exhibit unequal performance across skin tone groups due to differences in representation and colour distribution.

Another critical limitation in this domain is the availability of large, labelled datasets. Unlike general computer vision applications, medical image data is often limited due to privacy regulations, high annotation costs, and the need for expert clinical oversight. Consequently, traditional deep learning approaches, which typically rely on large-scale and balanced datasets, struggle to generalise effectively in low-data regimes. While various techniques such as data augmentation, synthetic oversampling, and transfer learning have been proposed, their effectiveness in promoting subgroups fairness and handling extreme class imbalance with limited data remains insufficiently understood and underexplored.

Insufficient data, especially in the context of rare skin conditions on skin cancer analysis, limits the training of robust and generalizable models [3]. Federated learning, a collaborative machine learning approach that allows multiple institutions to train models on decentralised data while preserving privacy, offers a promising solution to the problem of limited training samples [4]. To address these issues, innovative methods are required to enhance the quality and quantity of training data and to improve model architecture and learning techniques.

Recent work has shown that these challenges become even more complex when subgroups fairness is examined in federated settings. Federated clients may contain data from different demographic groups, such as distinct Fitzpatrick skin types, which results in non-IID (non-

independent and identically distributed) distributions across institutions. This can lead to systematic performance disparities between clients and between demographic subgroups. Our earlier work on FedCIAL introduced a federated debiasing method based on domain adaptation, but an open question remained: how does domain adaptation interact with class imbalance mitigation when both demographic bias and distributional heterogeneity coexist simultaneously? Motivated by this gap, this thesis includes an additional work package that extends the previous federated debiasing method by combining domain adaptation with class imbalance handling in a unified fairness-aware federated learning framework.

1.2 Research Questions

1. Novel Oversampling Method:

- What are the limitations of current oversampling methods in handling imbalanced skin cancer image datasets?
- To what extent does the proposed oversampling method enhance the performance of deep learning models in classifying imbalanced skin cancer image datasets?

2. Attention Fusion in Vision Transformers:

- How can attention fusion mechanisms be effectively integrated into Vision Transformer architectures to address the challenge of imbalanced data?
- How does the integration of attention fusion impact the interpretability and diagnostic accuracy of Vision Transformers on diverse skin cancer imaging tasks?

3. Debaised Method in Federated Learning:

- How can a debaised training method be designed to operate effectively within a federated learning framework while preserving data privacy?
- How can federated learning help in overcoming the problem of limited training samples by enabling collaborative model training across multiple institutions?

- What improvements in model fairness, accuracy, and generalisability can be achieved by applying the debiased method in federated learning for skin cancer image analysis?

4. Fairness-aware Federated Learning with Domain Adaptation and Class Imbalance

Mitigation:

- How do domain adaptation and class imbalance mitigation interact in federated dermatological classification?
- Can a combined approach improve both overall accuracy and subgroup fairness across Fitzpatrick skin types?

1.3 Research Objectives

The primary aim of this thesis is to address the intertwined challenges of class imbalance, subgroups fairness, and limited data availability in medical image classification using deep learning. These issues significantly affect the performance, reliability, and ethical deployment of AI systems in clinical settings. To tackle these challenges holistically, the thesis is structured into three research work packages. The specific objectives are as follows:

- O1 To develop an improved oversampling technique that enhances data quality and class balance in high-dimensional medical image datasets.
- O2 To enhance feature representation and classification performance in medical imaging tasks through an attention fusion strategy integrated into transformer-based architectures.
- O3 To develop a fairness-aware, privacy-preserving learning framework that mitigates demographic and domain bias in distributed medical image classification.
- O4 To extend fairness-aware federated learning by jointly integrating domain adaptation and class imbalance mitigation, and to establish a systematic subgroups fairness evaluation framework for skin lesion classification across Fitzpatrick skin tones.

1.4 Contributions

This thesis makes four original and significant contributions toward mitigating class imbalance and promoting subgroups fairness in medical image classification under limited data conditions using deep learning. The central novelty of this work lies in addressing these challenges holistically across data-level, model-level, and system-level perspectives. Specifically, the thesis contributes (i) principled imbalance mitigation techniques tailored to high-dimensional medical images, (ii) enhanced representation learning strategies for data-efficient transformer-based models, and (iii) fairness-aware learning frameworks that operate under privacy-preserving and data-scarce federated settings. Together, these contributions advance the state of the art by demonstrating how accuracy, robustness, and fairness can be jointly improved even when training data are limited, imbalanced, and demographically skewed.

The specific research outputs corresponding to each objective are as follows:

1. To meet Objective 1 (O1), we published the work: Heroza, R.I., Gan, J.Q. and Raza, H., 2023, June. Sia-SMOTE: A SMOTE-based oversampling method with better interpolation on high-dimensional data by using a Siamese network. *In International Work-Conference on Artificial Neural Networks* (pp. 448-460). Cham: Springer Nature Switzerland.
2. To meet Objective 2 (O2), we published the work: Heroza, R.I., Gan, J.Q. and Raza, H., 2024, July. Enhancing skin lesion classification: A self-attention fusion approach with vision transformer. *In Annual Conference on Medical Image Understanding and Analysis* (pp. 309-322). Cham: Springer Nature Switzerland.
3. To meet Objective 3 (O3), we published the work: Heroza, R.I., Gan, J.Q. and Raza, H., 2025. FedCIAL: Federated Colour-Invariant Adversarial Learning for Enhancing Fairness and Performance in Skin Lesion Classification. *In Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 1821-1828).
4. To meet Objective 4 (O4), we submitted the work to the IEEE Conference on Artificial Intelligence (CAI) 2026: Fairness-Aware Federated Learning with Domain Adaptation and Class Imbalance Mitigation for Skin Lesion Classification.

Literature Review

2.1 Introduction to Medical Image Classification

Medical image classification is a fundamental task in computer-aided diagnosis (CAD) and plays a vital role in modern healthcare systems [5], [6]. The goal is to automatically categorise medical images such as X-rays, CT scans, MRIs, and dermoscopic images into predefined clinical categories, such as benign vs. malignant, or normal vs. pathological [7]–[9]. Accurate classification can support early disease detection, improve diagnostic consistency, and enhance the productivity of clinical workflows, particularly in resource-constrained or high-volume settings [8].

In recent years, deep learning with convolutional neural networks (CNNs) has demonstrated exceptional success in medical image analysis [7], [10], [11]. CNNs have the capacity to learn hierarchical feature representations directly from pixel-level data, eliminating the need for manual feature engineering [9], [11], [12]. Applications of CNN-based classifiers span across domains such as radiology, dermatology, pathology, ophthalmology, and cardiology [8], [10], [12]. For instance, deep learning systems have shown dermatologist-level performance in skin lesion classification [13], [14] and have been used to assist radiologists in detecting lung nodules in chest CT scans [15], [16].

However, the deployment of deep learning models in clinical practice presents several challenges that are not commonly encountered in general-purpose computer vision tasks [17]–

[19]. Although publicly available medical imaging datasets have enabled significant advances in deep learning research, they are often limited in size, demographic diversity, or annotation richness. In real-world clinical settings, substantially larger volumes of imaging data may exist. However, access to such data is frequently restricted due to privacy regulations, ethical considerations, institutional data-sharing barriers, and the high cost of expert annotation [18]–[20]. Moreover, class imbalance is a pervasive issue, as rare but critical pathologies are underrepresented compared to common or healthy cases [19]–[21]. These imbalances can bias models toward majority classes, leading to poor generalisation and reduced clinical utility [19], [22].

Another emerging concern in medical image classification is subgroups fairness [23]–[25]. Deep learning models trained on non-representative or biased datasets can exhibit disparities in performance across demographic groups such as skin tone, gender, or age [23]–[25]. For example, dermatology datasets often underrepresent darker skin tones, resulting in models that perform poorly on images from those populations [24]–[27]. Such biases threaten the ethical deployment of AI in healthcare and raise questions about trust, transparency, and accountability [23], [28].

In parallel with these challenges, new model architectures such as Vision Transformers (ViTs) [29]–[31] and training paradigms such as federated learning [32] and self-supervised learning [33], [34] are expanding the methodological landscape. These approaches offer promising opportunities for developing more robust, generalizable, and equitable medical image classifiers, especially in settings where labeled data is scarce or data sharing is restricted [32], [33].

In this context, there is a growing need for comprehensive strategies that go beyond optimising classification accuracy to also address the challenges of class imbalance, limited data, and subgroups fairness [19], [21], [23]. The subsequent sections of this literature review will explore these challenges in greater depth and examine the existing body of research that seeks to overcome them.

Recent work in dermatology has also shown that performance disparities across Fitzpatrick skin types can persist even when using advanced deep models and training pipelines, particularly under non-IID data distributions and federated settings [3], [35], [36], further motivating fairness-aware and domain-adaptive approaches in this thesis.

2.2 Class Imbalance in Medical Imaging

Class imbalance is a persistent and critical issue in medical image classification that significantly hampers the performance and reliability of deep learning models. In many clinical imaging datasets, the distribution of classes is inherently skewed due to the natural prevalence of certain diseases, the difficulty of collecting rare case data, and the constraints surrounding expert annotation. For example, in dermatology or radiology datasets, images of common conditions or normal findings often vastly outnumber those representing rare or life-threatening diseases. This imbalance creates a bias during model training, as standard deep learning algorithms are optimised to perform well on the majority class, frequently at the expense of the minority. As a result, models may demonstrate high overall accuracy while performing poorly on clinically significant but underrepresented classes, a trade-off that is unacceptable in medical contexts where the cost of missing a rare diagnosis can be life-altering [37]–[39].

2.2.1 Causes of Class Imbalance in Medical Imaging Datasets

There are several reasons why class imbalance is especially pronounced in medical imaging. First, many serious diseases, such as malignant tumours or genetic disorders, occur infrequently in the population, resulting in naturally imbalanced datasets. Second, the annotation of medical images is a resource-intensive process that requires expert radiologists, pathologists, or dermatologists, making it difficult and costly to label large volumes of rare cases. Third, the data collection process itself often reflects historical and systemic biases. For instance, retrospective datasets gathered from hospital records may overrepresent certain demographics or conditions while underrepresenting others, further distorting the data distribution. Moreover, privacy regulations, institutional policies, and data-sharing restrictions often limit research access to diverse and comprehensive multi-institutional datasets, exacerbating imbalance across populations and demographic groups [38]–[40].

In dermatological imaging, an additional layer of imbalance emerges when clinical labels intersect with demographic attributes such as skin tone. Datasets may contain sufficient numbers of benign lesions on lighter skin, but very few malignant lesions on darker skin [3], [36]. This joint

imbalance across both disease label and skin type creates scenarios where minority subgroups (for example, malignant lesions on Fitzpatrick types IV-VI) are severely underrepresented, making it particularly challenging for models to learn fair and reliable decision boundaries.

2.2.2 Techniques for Mitigating Class Imbalance

To address class imbalance, researchers have developed various techniques that can be broadly categorised into data-level and algorithm-level methods. Data-level approaches attempt to balance class distributions by manipulating the dataset directly. On the other hand, algorithm level solutions adjust the learning process to account for imbalance.

Data Approach

The data approach for handling the imbalance problem is by resampling the dataset. Random over/undersampling represent the most straightforward approaches to addressing class imbalance. In random oversampling, samples from the minority class are replicated to increase their representation, whereas random undersampling involves the removal of instances from the majority class. Tomek Links, defined as pairs of close but differently labelled instances, enable a more targeted undersampling strategy by eliminating the majority class instance in each pair, thereby improving class balance. The Cluster Centroid method applies the K-Means clustering algorithm to partition the majority class into N clusters, where N equals the number of minority class instances and replaces each cluster with its centroid. This method can effectively reduce the size of the majority class while preserving representative structure. The Synthetic Minority Oversampling Technique (SMOTE) has gained popularity in many applications due to its ability to create new samples by interpolating between existing ones. It generates artificial minority class samples by selecting an instance, identifying its k -nearest neighbours, and creating new points along the line segments connecting the instance to its neighbours, where k is determined by the desired oversampling ratio [41]. Figures 2.1 and 2.2 illustrate the dataset before and after oversampling by SMOTE, with the latter clearly showing the synthetic class-1 points in orange. However, in the context of high-dimensional data such as medical images, SMOTE may inadvertently introduce noise or generate unrealistic samples, particularly when the original

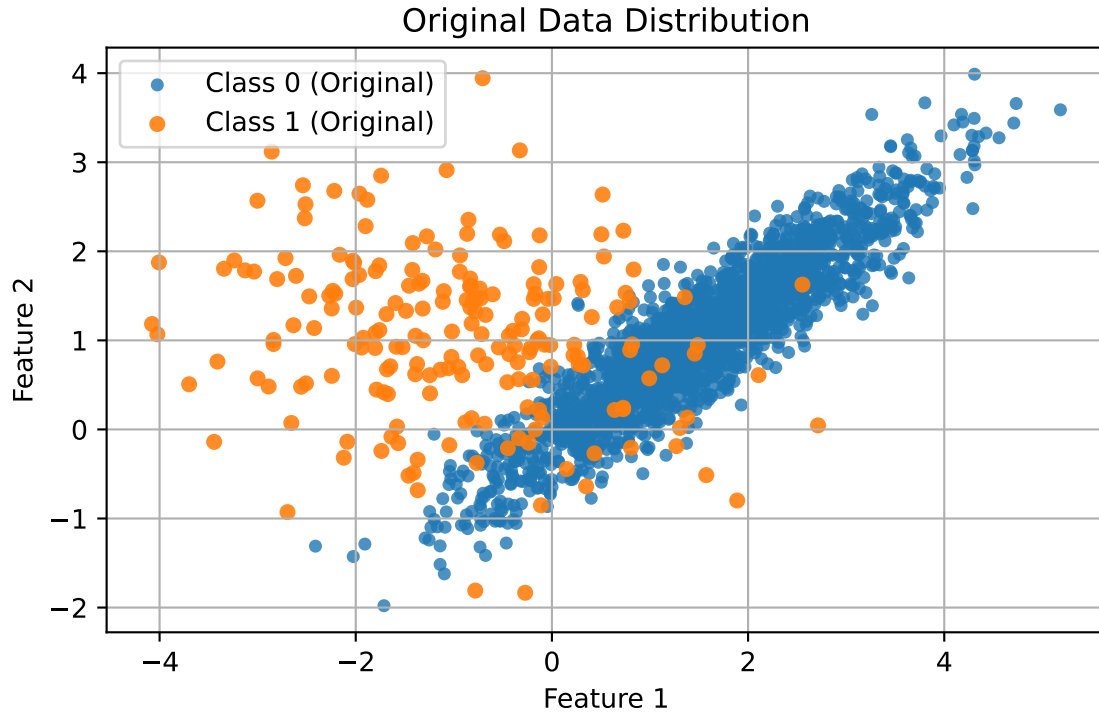


Figure 2.1: Imbalanced dataset before oversampling

minority class samples are sparse or poorly clustered. In such cases, oversampling may not only fail to improve model performance but also increase the risk of overfitting [42]–[44].

Several SMOTE variants have been proposed to improve the quality of generated samples. Borderline-SMOTE focuses on oversampling minority instances located near decision boundaries [45]. In Borderline SMOTE1, synthetic examples are generated from each 'DANGER' minority instance, whose majority of nearest neighbours belong to the majority class. Borderline SMOTE2 extends this by also creating synthetic samples using the nearest negative (majority) neighbours, amplifying the difference by a random factor between 0 and 0.5 to produce points more representative of the minority class. ADASYN (Adaptive Synthetic Sampling) further refines this concept by adaptively generating more synthetic samples for harder-to-learn minority instances than for simpler ones [46]. This is achieved by estimating a density distribution around each minority instance, allowing the method to focus data generation where classification difficulty is greatest.

Although these methods improve sample placement near decision boundaries, they still rely on heuristic neighbourhood definitions [47] and offer limited guarantees that generated samples lie on the true lesion manifold, which is particularly problematic in visually complex medical

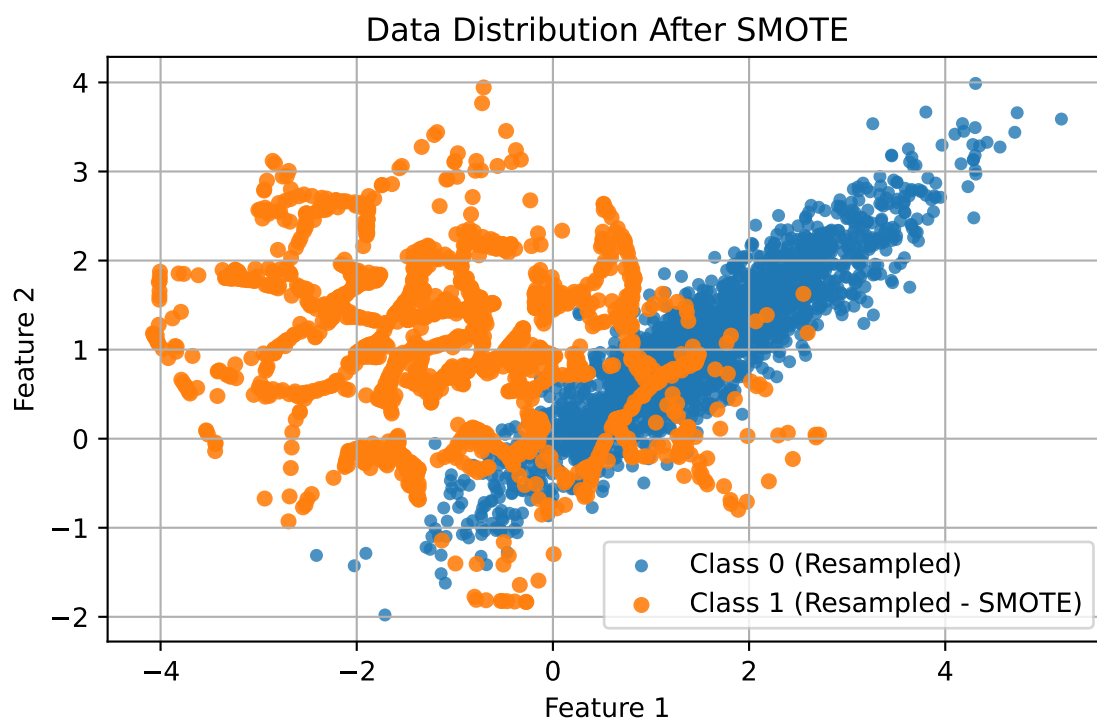


Figure 2.2: Balanced dataset after oversampling using SMOTE

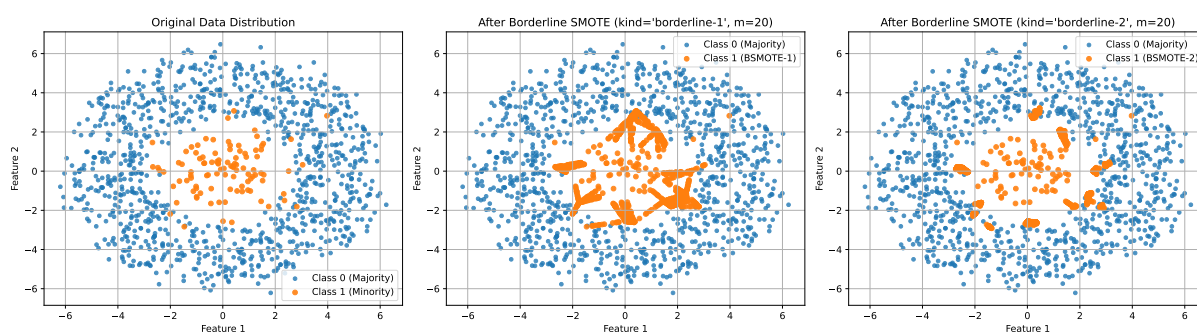


Figure 2.3: Borderline-SMOTE Oversampling Example Distribution

images.

Hybrid approaches have also emerged. One such method integrates SMOTE with k-means clustering, enabling the identification of regions in the feature space where oversampling is most effective. The process consists of three stages: clustering, filtering, and oversampling. First, k-means partitions the input space into k clusters. Next, clusters with a high proportion of minority instances are selected, and the number of synthetic samples assigned to each is determined according to the density of minority samples within the cluster. Finally, SMOTE is applied within each selected cluster to achieve the desired class balance.

In domains involving image data, SMOTE has been adapted to operate within deep learning frameworks. DeepSMOTE [48] employs an encoder-decoder architecture to map images into a lower-dimensional latent space before oversampling, ensuring that only the most salient features are used in synthetic generation. This reduces the influence of irrelevant features, mitigates overfitting, and improves efficiency. Compared to GAN-based approaches, DeepSMOTE can produce high-quality, information-rich artificial images suitable for visual inspection without requiring a discriminator network.

However, latent space oversampling remains sensitive to representation quality and may propagate existing biases if the learned embedding itself is skewed toward majority classes or demographic groups [49].

Generative adversarial networks (GANs) have also been widely applied to class imbalance problems in computer vision [50]. By learning the underlying data distribution, GANs can generate realistic synthetic images for underrepresented classes. Applications span multiple domains, including DCGAN-based balancing of malaria datasets to reduce false positives [51], CycleGAN-based generation of realistic mammogram images to improve breast mass classification [52], and small-object detection in datasets such as MS COCO [53]. In the latter case, GANs have been used to simulate rare traffic scenarios critical for autonomous driving, as well as to enhance low-resolution small objects into high-resolution ones [54].

Algorithm Approach

Beyond resampling methods, which balance datasets through data-level transformations, algorithm-level strategies address class imbalance directly during training.

Cost-Sensitive Learning (CSL) minimises total misclassification cost by assigning different penalties to different error types [55]. This shifts the focus from overall accuracy to cost minimisation, often penalising false negatives more heavily. Ali et al. [56] proposed a cost-sensitive ensemble feature ranking method for chronic kidney disease classification, outperforming multiple classifiers and feature selection methods. Zieba et al. [57] combined cost-sensitive SVM with boosting to predict post-operative life expectancy in lung cancer patients, achieving superior performance to other imbalance-handling classifiers. Mienye and Sun [58] demonstrated that cost-sensitive variants of logistic regression, random forest, and XGBoost outperformed baseline algorithms on four imbalanced medical datasets.

However, selecting appropriate cost weights is non-trivial, and poorly calibrated costs can destabilise training or degrade performance on the majority classes [59]. Recent work has therefore explored alternative loss rebalancing strategies such as focal loss [60], class-balanced loss [61], and logit adjustment [62], which aim to improve minority performance while maintaining calibration and stability. Despite their effectiveness, these methods often trade improved recall for increased false positives, a trade-off that must be carefully evaluated in clinical settings. Le et al. [63] also introduced Cluster-based Boosting (CBoost), which partitions the majority class into clusters, assigns weights based on distance to the centroid, and iteratively combines weak classifiers. On the Korean Bankruptcy Dataset, CBoost achieved an AUC of 86.8%, outperforming established bankruptcy prediction methods.

Another method is ensemble learning, which combines multiple models to improve predictive performance. Three common forms of ensemble learning are bagging, stacking, and boosting. In bagging, multiple models of the same type are trained on bootstrap samples, with predictions aggregated by voting or averaging. Hong et al. [64] developed two hybrid bagging models for landslide susceptibility prediction, with LADT-Bagging outperforming all baselines. Subasi et al. [65] used bagging with random forests for fetal heart rate classification, achieving 99.02% accuracy.

In stacking, diverse base models are trained on the same data, and a meta-learner combines their outputs. Charoenkwan et al. [66] introduced StackIL6, integrating 12 feature descriptors and five machine learning algorithms to detect IL-6-inducing peptides, outperforming existing approaches. Rajagopal et al. [67] applied stacking for cyberthreat detection, achieving 97% accuracy on real-time network traffic and 94% on emulated datasets.

Boosting train models sequentially, increasing the weight of misclassified samples at each step. Kumari and Toshniwal [68] proposed an Extreme Gradient Boosting Forest and Deep Neural Network (XGBF-DNN) ensemble for solar irradiance forecasting, combined via ridge regression to reduce overfitting. The model maintained high stability across seasonal variations, reducing prediction error by 33-40%.

While ensembles can improve robustness under imbalance, they substantially increase computational and deployment complexity [69], limiting their applicability in resource-constrained clinical environments.

Emerging and Advanced Strategies

Recent advances in deep learning have introduced more sophisticated techniques to deal with class imbalance in medical imaging. These include metric learning approaches using Siamese or triplet networks to better capture similarity between samples, and boundary-aware oversampling strategies that target the generation of synthetic samples near decision boundaries to improve class separability. Additionally, self-supervised pretraining and generative models such as GANs and diffusion models have shown promise in improving representation learning in low-data regimes. However, these methods also face limitations related to complexity, computational cost, and generalisation across datasets [38], [70], [71].

2.2.3 Challenges in the Medical Imaging Domain

Despite the wide range of proposed solutions, several challenges remain when applying them to medical imaging tasks. Medical images are often high-dimensional and complex, which makes naive oversampling or interpolation strategies less effective. The visual variability within a single class can be large, while inter class differences may be subtle, especially in early disease stages. In

addition, the cost of errors is asymmetric, where false negatives for rare but dangerous conditions carry far more risk than false positives for benign findings. Furthermore, class imbalance often intersects with other forms of bias, such as demographic or institutional bias, making it difficult to address one issue in isolation without impacting others [37], [39], [40], [44], [70].

2.3 Fairness in Medical AI

Fairness in medical artificial intelligence (AI) refers to the equitable performance and treatment of all individuals regardless of their demographic attributes, such as race, ethnicity, gender, age, or skin tone. As AI systems become increasingly integrated into clinical decision-making, concerns have emerged about their potential to perpetuate or even exacerbate existing disparities in healthcare. In medical image classification, fairness involves ensuring that deep learning models provide consistently reliable results across all population groups. However, fairness is not simply an ethical concern. It is a practical necessity. A model that performs well on the majority or well-represented groups but poorly on underrepresented groups poses significant risks, including misdiagnosis, delayed treatment, and loss of trust in AI-driven healthcare tools [72]–[74].

In dermatology, several recent studies have provided quantitative evidence of such disparities. Groh et al. [3] showed that state-of-the-art classifiers trained on skin lesion datasets dominated by lighter skin types exhibit markedly lower sensitivity and specificity on images from darker Fitzpatrick skin types. Daneshjou et al. [36] introduced the Diverse Dermatology Images (DDI) dataset to explicitly benchmark performance across skin tones, revealing that models with strong average performance can still underperform substantially on underrepresented FST groups. These findings underline that fairness in medical AI cannot be assessed from global metrics alone and must explicitly consider subgroup performance.

Unfairness in medical AI often becomes evident through disparities in model performance across subgroups. For example, a classification model may achieve high overall accuracy but exhibit significantly lower sensitivity or precision when applied to images from specific ethnic groups or age brackets. These disparities can be especially harmful in medical contexts, where failing to detect a disease in a minority group can contribute to systemic health inequities.

Importantly, traditional evaluation metrics like accuracy or AUC may mask these issues, as they do not capture group-wise performance differences. Consequently, fairness-aware evaluation requires more nuanced metrics such as equal opportunity, demographic parity, or subgroup-specific recall and precision [75], [76].

2.3.1 Sources of Bias in Medical Image Datasets

Many of the fairness challenges in medical AI stem from bias in the underlying data. Clinical datasets are often collected retrospectively from institutions that serve specific geographic or demographic populations. As a result, certain groups, such as individuals with darker skin tones, women, or those from lower socioeconomic backgrounds, may be underrepresented. In imaging tasks like skin lesion classification, the lack of sufficient representation of darker skin can lead to models that struggle to correctly identify dermatological conditions in those individuals. Patients with darker skin (Fitzpatrick types IV-VI) have been consistently underrepresented in dermatological training datasets, resulting in AI systems that demonstrate significantly lower diagnostic accuracy for these populations compared to lighter skin patients [77]. Similarly, women are substantially underrepresented in medical imaging datasets. For example, the widely used NIH Chest-XRay14 dataset and CheXpert dataset contain 43.5% and 40% female patients, respectively, and research has shown that gender imbalance in medical imaging datasets produces significantly degraded performance in minority gender groups. [75]

Beyond demographic representation, individuals from lower socioeconomic backgrounds are often systematically excluded from clinical datasets due to barriers to healthcare access, resulting in incomplete electronic health records that further compound algorithmic bias. This data incompleteness means that AI models trained on these datasets inherit and amplify the disparities embedded within them. [78]

In addition to sampling bias, label bias may arise from subjective annotations or inter-rater variability among clinicians. Such noise can result from systematic errors of the annotator as well as inter-annotator variation. More than three domain experts are typically needed to generate trustworthy annotations, yet the use of single annotators or inadequately reviewed labels remains common in practice [79]. Clinicians may also introduce implicit cognitive biases when

labelling data based on patient demographic characteristics, and these annotator biases can become encoded into the ground truth labels used to train models. [80].

These biases can be unintentionally learned and amplified by data-hungry deep learning models, which tend to prioritise patterns most frequently present in the training data [73], [75]. Deep neural networks possess a high capacity to fit training data, and when trained on biased or incomplete datasets, they systematically learn to amplify the underlying representation disparities. Studies have demonstrated that AI systems show consistent underdiagnosis of underserved populations, including Black, Hispanic, and female patients, falsely categorising already marginalised individuals as healthy and thereby exacerbating existing health disparities [81]. Intersectional disparities are particularly pronounced. For instance, Black female patients experience substantially higher rates of algorithmic underdiagnosis compared to any single identity group [82]. This amplification occurs because models trained on biased labels perpetuate not only differential misclassifications but also encode the original cognitive biases present in their training annotations. Consequently, without deliberate intervention, AI systems risk becoming powerful mechanisms for entrenching and scaling the demographic inequities present in historical healthcare data. [82]

2.3.2 Approaches to Mitigating Fairness Issues

To mitigate unfairness in medical image classification, researchers have developed a variety of fairness-aware learning strategies. One class of methods focuses on preprocessing techniques, such as reweighting or resampling the data to balance group distributions. Reweighting assigns different weights to data samples to ensure that underrepresented groups have a proportionate impact on the model's learning process, while resampling either oversamples minority class instances or undersamples majority class instances to achieve a balanced dataset [74], [83]. Beyond sample-level adjustments, data harmonisation approaches remove sensitive information directly from input images through segmentation to isolate lesions, application of normalisation techniques such as Z-score or ComBat algorithms, differential privacy methods, or transformation of images into sketches that have shown reduced disparities across demographic groups [84].

In-processing approaches incorporate subgroups fairness objectives directly into the model

training process, offering more direct control over the model's behaviour. Fairness-constrained optimisation methods add penalty terms or regularisation constraints to the loss function to encourage demographic parity or equalised odds. More recently, multi-task learning frameworks have emerged as promising in-processing alternatives that view fairness as a task-balancing problem. By dividing predictions on different subpopulations into separate tasks, dynamic reweighting schemes can dynamically adjust the attention weight on each task during backpropagation, achieving significant subgroups fairness improvements (up to 98% disparity reduction) with minimal accuracy loss (less than 4%) [85]. Another prominent in-processing approach, adversarial debiasing, trains the model to learn features informative for the clinical task while penalising the learning of sensitive attributes such as race or gender through an adversarial branch that uses reversed gradients [86]. Adversarial debiasing has demonstrated marked improvements in fairness metrics such as equalised odds across radiology applications, including chest radiography and mammography. However, most adversarial approaches incur some trade-off between subgroups fairness and predictive performance, with evidence of varied outcomes across different clinical tasks. Recent innovations employ Concept Activation Vectors (CAV) to identify convolution layers primarily responsible for encoding sensitive attributes. Partial adversarial debiasing then fine-tunes only from those layers forward, substantially mitigating the typical performance drop while maintaining strong subgroups fairness improvements. For example, in chest X-ray classification, partial debiasing with CAV achieved an AUC of 0.87 compared to baseline 0.57 and standard fine-tuning 0.81, while also maintaining comparable or superior fairness metrics across racial subgroups. [87]

Post-processing techniques modify model outputs or thresholds after training to equalise performance across groups. Calibration methods use different prediction thresholds for each demographic subgroup to achieve fairness criteria such as equalised odds. These approaches are model-agnostic and can be applied to existing black-box systems where the training process cannot be modified [85]. Pruning strategies distil model parameters by identifying and removing neurons or features associated with sensitive attributes, thereby preventing sensitive information from being encoded into the network [84]. However, post-processing methods are generally less effective at addressing root causes of bias compared to in-processing approaches, and the

trade-off between maintaining accuracy and achieving subgroups fairness can be substantial. [85]

Despite their conceptual promise, most fairness mitigation methods have been developed in non-medical domains, while their application to the high-dimensional, multi-modal nature of medical imaging data remains an active area of research. Key challenges specific to medical imaging include the diversity of modalities (2D X-rays, 3D MRI, PET/CT versus the primarily RGB images in facial recognition domains), the complexity of clinical tasks (ranging from classification to segmentation to localisation), and the fundamental difference in how AI scientists versus clinicians conceptualise sensitive attributes [84]. Furthermore, research on AI fairness in medical domains remains sparse relative to other application areas, with a narrow focus on particular demographic attributes and dominance of group fairness approaches over individual fairness notions. Addressing these gaps will require continued development of fairness techniques specifically tailored to the unique characteristics and constraints of medical image analysis. [83], [86]

2.3.3 Fairness in Federated Medical Imaging

Fairness issues are further complicated in federated learning, where data remains decentralised across institutions. Each client may serve a distinct patient population, resulting in client-level demographic skew (for example, one hospital mostly treating lighter FST patients and another mostly treating darker FST patients). When a global model is trained via federated averaging, it may implicitly favour dominant clients or demographics, leading to uneven performance across institutions and subgroups [88].

Fairness-aware FL methods have begun to emerge, such as client reweighting, group-robust optimisation, and fair aggregation strategies [89], [90]. However, most of these approaches focus on client-level fairness (balancing performance across clients) rather than subgroup fairness (balancing performance across demographic groups, such as FST, within and across clients). Moreover, privacy constraints and secure aggregation limit access to individual or group-level statistics, making fairness auditing and enforcement more challenging.

Federated domain adaptation offers a promising direction, where domain alignment is performed across clients without sharing raw data [91]. Building on adversarial learning,

federated colour-invariant approaches use gradient reversal and synthetic colour-transformed images to learn skin tone invariant representations while respecting data localisation and privacy constraints. By leveraging known colour distribution shifts across Fitzpatrick types in a perceptually uniform colour space (for example, Lab), such methods aim to jointly improve accuracy and subgroups fairness across heterogeneous clients, and form the conceptual basis for the FedCIAL and fairness-aware federated frameworks developed later in this thesis.

Despite increasing attention to fairness in medical AI, several key challenges remain unresolved. First, there is a lack of standardised benchmarks and fairness metrics tailored to medical imaging tasks, making it difficult to assess and compare methods. Second, many fairness-aware techniques assume access to explicit group labels, which may be missing or unreliable in clinical datasets. Third, existing methods often address subgroups fairness in isolation, without considering its interaction with other constraints such as class imbalance and limited data availability. In practice, these factors frequently co-occur and exacerbate one another. Addressing subgroups fairness effectively requires models that are not only demographically robust but also data efficient and able to perform well under imbalanced and distributed training scenarios [73], [76], [92], [93].

2.4 Learning with Limited Data in Medical Imaging

One of the most persistent constraints in applying deep learning to medical image classification is the scarcity of labelled data. Unlike general computer vision tasks, where massive datasets such as ImageNet or COCO are available, publicly available medical image datasets are often limited in size and demographic coverage because patient data cannot be freely shared due to privacy regulations, ethical considerations, and institutional governance constraints. Acquiring medical images requires specialised equipment, trained technicians, and patient consent, while producing reliable labels demands expert clinicians. These factors combine to create datasets that are orders of magnitude smaller than those typically required to train deep learning models effectively, making overfitting and poor generalisation recurring concerns [94], [95].

2.4.1 Causes of Limited Data Availability

Several factors contribute to the limited data challenge in medical imaging. Privacy laws such as HIPAA in the United States and GDPR in the European Union restrict the sharing of patient data and limit opportunities for large-scale centralisation of datasets [96]. Annotation complexity is another significant barrier[94]. Tasks like lesion segmentation or disease grading require extensive domain knowledge and are time-consuming. Moreover, manual annotation of medical images is labour-intensive and prone to biases, with annotation rates significantly lagging the rate of image data accumulation. Additionally, the rarity of conditions means that certain diseases yield few examples for databases, while institutional siloing leads to fragmented datasets that may not be representative of the wider patient population. Domain shift further exacerbates data scarcity challenges, as medical images often exhibit significant distribution differences due to variations in imaging equipment, scanning protocols, and patient demographics. [97], [98]

2.4.2 Impact on Deep Learning Performance

Data scarcity is especially detrimental in deep learning, as modern architectures often contain millions of parameters and require large amounts of training data to generalise robustly. When trained on small datasets, these models tend to memorise training examples instead of learning meaningful patterns, resulting in poor performance on unseen data. Small datasets also tend to have incomplete coverage of disease subtypes, limited demographic diversity, and inconsistencies in image acquisition protocols, further degrading model reliability in real-world settings [94], [95]. Class imbalance compounds these problems, where most medical image datasets suffer from unequal class distributions, causing models to favour majority classes and fail to accurately detect rare conditions or pathological subtypes. Standard evaluation metrics such as accuracy and AUROC may mask poor minority class performance in imbalanced scenarios, necessitating complementary metrics such as area under the precision-recall curve (AUC-PR) and balanced accuracy score for accurate clinical assessment. [22], [37]

2.4.3 Existing Strategies for Learning with Limited Data

Researchers employ several strategies to address the challenge of limited data. Transfer learning from models pretrained on large natural image datasets like ImageNet is widely adopted, leveraging generic low-level features for adaptation to medical domains. However, the domain shift from natural to medical images can limit effectiveness, so domain-specific self-supervised or supervised pretraining is increasingly favoured. Data augmentation, including traditional geometric transformations and advanced techniques using generative models such as GANs and diffusion models, synthetically expands the training set. Few-shot learning and meta learning aim to train models capable of adapting from very few labelled examples, while self-supervised learning extracts useful feature representations from unlabeled medical images, reducing reliance on annotated data [99]–[102].

While these approaches improve data efficiency within individual institutions, they are inherently limited by the size and diversity of local datasets. Federated and collaborative learning address this constraint by enabling models to learn from distributed data across institutions without sharing sensitive patient information. Federated learning (FL) enables multiple institutions to collaboratively train a shared model without exchanging raw patient data, preserving privacy while increasing the effective training data size. Model updates, not data, are shared and aggregated centrally. Federated setups introduce challenges such as data heterogeneity, communication overhead, and subgroups fairness concerns, but combining FL with domain adaptation or advanced representation learning can mitigate these issues and make FL a promising strategy for large-scale, privacy-preserving medical AI [103]–[105].

2.4.4 Limitations of Current Methods

Existing techniques have shown promise but are not without drawbacks. Transfer learning from natural images may fail to capture domain-specific features relevant to medical imaging, while synthetic augmentation may produce unrealistic examples if not properly validated. Few-shot and meta learning, though conceptually appealing, are still immature for high-dimensional medical data. Federated learning, though it addresses data access limitations, can exacerbate biases

if participating institutions have imbalanced or non-representative datasets. These limitations highlight the need for methods that operate effectively in low data regimes and tackle class imbalance and subgroups fairness together [103], [106], [107].

2.5 Deep Learning Architectures and Techniques in Medical Image Classification

2.5.1 Convolutional Neural Networks in Medical Imaging

Convolutional Neural Networks (CNNs) have been the dominant architecture in medical image analysis for over a decade, owing to their ability to automatically learn hierarchical spatial features from pixel data. CNNs excel at capturing local spatial dependencies through convolutional kernels, making them well-suited for detecting anatomical structures and pathological patterns. Architectures such as VGGNet, ResNet, and DenseNet have been widely adopted in medical classification tasks, often serving as backbones for more complex pipelines. However, CNNs are limited in their ability to model long-range dependencies across the image, and their performance is highly dependent on the availability of large, balanced datasets, which is a rare condition met in medical imaging [11], [108]–[110].

2.5.2 Oversampling and Synthetic Data Generation

Addressing class imbalance in medical image datasets often requires synthetic data generation. Traditional approaches like the Synthetic Minority Oversampling Technique (SMOTE) interpolate between minority class samples to generate additional data points. While effective in tabular domains, vanilla SMOTE struggles with high-dimensional medical image data, as interpolation between pixel vectors can produce unrealistic or noisy samples. Deep learning based extensions, such as adversarial oversampling with GANs or autoencoder-based data synthesis, attempt to produce more realistic images but can still generate artefacts if trained on noisy inputs. Incorporating feature learning into oversampling, such as using a Siamese network to filter synthetic samples, offers a promising avenue for improving data quality and downstream

classification performance [111]–[113].

2.5.3 Attention Mechanisms and the Emergence of Vision Transformers

Attention mechanisms have emerged as powerful tools for enhancing deep learning models by allowing them to focus on the most informative regions of an image. In medical image classification, attention modules help highlight relevant pathological features while suppressing irrelevant background information. Recent advances such as the Vision Transformer (ViT) [114] architecture extend this concept by replacing convolutional operations with self-attention, enabling the model to capture both local and global contextual relationships.

Self-attention is a mechanism that has played a pivotal role in the recent advancements of natural language processing and deep learning. It is a key component in the Transformer model, a revolutionary architecture that has reshaped the landscape of tasks such as machine translation, text summarisation, and language understanding. At its core, self-attention enables a model to weigh the importance of different words in a sequence when processing a particular word, allowing the model to capture contextual dependencies efficiently [115]. This mechanism is particularly powerful in handling long-range dependencies, a challenge that traditional recurrent neural networks struggled with [115].

In the realm of computer vision, self-attention has emerged as a transformative mechanism for enhancing the contextual understanding of visual data. Unlike traditional convolutional neural networks (CNNs) that rely on fixed receptive fields, self-attention allows models to dynamically focus on different regions of an image while processing specific pixels or regions [115]. This dynamic attention mechanism proves particularly valuable for tasks like image captioning, object detection, and semantic segmentation, where capturing long-range dependencies and contextual information is essential. By enabling the model to assign varying levels of importance to different spatial locations within an image, self-attention enhances the model's ability to discern complex relationships and patterns, contributing to more accurate and nuanced visual interpretations. The application of self-attention in computer vision reflects its versatility as a mechanism that goes beyond natural language processing, providing a promising avenue for advancing the field's capabilities in image understanding and analysis.

Several variants of self-attention, including additive attention [116], multiplicative attention [117], and scaled dot product attention [118], have been instrumental in refining the capabilities of deep learning models. Additive attention introduces learnable parameters to calculate attention scores, offering a flexible approach to capturing relationships within sequences. Multiplicative attention, on the other hand, employs element-wise products to determine attention weights, providing an alternative perspective on modelling dependencies. Scaled dot product attention, a fundamental variant, combines the benefits of efficiency and robustness. It scales the dot product of the query and key by the square root of the dimensionality, preventing the attenuation of gradients during training. These attention mechanisms, each with its unique characteristics, underscore the versatility of self-attention. Their adoption in diverse contexts demonstrates how attention mechanisms can be tailored to specific requirements, enhancing the capacity of models to discern intricate patterns and relationships in a variety of data domains, from natural language processing to computer vision.

In addition to traditional attention mechanisms like additive, multiplicative, and scaled dot product attention, recent advancements have introduced novel variants of self-attention. One such innovation is relative positional encoding [119], which enhances the modelling of sequential relationships by considering the relative positions of elements within a sequence. Linformer [120], known for its linearised self-attention mechanism, presents a streamlined approximation of the complete self-attention operation. This approach effectively mitigates the quadratic complexity inherent in standard attention mechanisms, thereby improving scalability without sacrificing the model's capacity to grasp long-range dependencies. Performer attention [121], inspired by the idea of kernelised self-attention, utilises random feature approximation techniques to achieve linear complexity, making it computationally more efficient for handling long sequences. Sinkhorn attention [122] incorporates ideas from optimal transport theory to improve the stability of attention distributions. These newer self-attention variants showcase the ongoing evolution of attention mechanisms, addressing computational challenges and exploring diverse mathematical foundations. As the field of deep learning continues to evolve, these innovations contribute to the refinement and specialisation of attention mechanisms for various applications, fostering more efficient and effective models in natural language processing, computer vision, and beyond.

ViT has been pretrained using the extensive ImageNet-21k collection of 224x224 images. The model was fine-tuned on ImageNet1K (ILSVRC2012). With a [CLS] token at the start of each sequence for classification tasks and a position embedding before feeding the sequence to the layers of the Transformer encoder, images are presented to the model as a series of fixed-size patches (resolution 16x16) that are linearly embedded.

There are several implementations of vision transformers such as ViT B16, DeiT and DINO. ViT B16 is a vision transformer base variant with 16x16 input patch size, which is described in the ViT paper [114]. DeiT (data efficient image transformers) model was proposed by Facebook AI in [123] use a distillation token to effectively learn from a ConvNet as the teacher. Facebook AI has also created a technique called DINO for the self-supervised training of Vision Transformers [124]. In contrast to convolutional models, vision transformers trained using the DINO approach exhibit some extremely intriguing characteristics. Despite never having had any training, they can segment objects.

While ViTs have demonstrated strong performance in natural image benchmarks, their application to medical imaging is still developing, with challenges including data scarcity and the need for domain-specific attention strategies. Hybrid models that combine CNN feature extractors with transformer-based attention have shown promise in bridging the gap between the two paradigms[29], [125].

2.5.4 Federated Learning and Privacy Preserving Model Training

To overcome these issues, researchers have explored domain adaptation techniques, personalised FL strategies, and adversarial learning frameworks to align feature spaces across clients while preserving privacy [126], [127]. Split-learning-based frameworks such as FeSViBS [128] distribute a Vision Transformer between client and server, with early layers on-device and later transformer blocks and classifiers on the server, reducing communication cost while enabling more expressive models. Building on these ideas, recent fairness-aware FL approaches integrate colour-invariant or domain-invariant representation learning into such pipelines, for example, by using gradient reversal layers and FST-guided colour transformations to encourage skin tone invariant features. These directions directly motivate the federated colour-invariant and

fairness-aware frameworks proposed and evaluated in the later work packages of this thesis.

2.6 Critical Summary and Research Gaps

The literature reviewed in this chapter demonstrates significant progress in medical image classification using deep learning. However, several important limitations remain unresolved, particularly in settings where class imbalance, demographic bias, and limited data coexist simultaneously. Existing studies often address these challenges independently, resulting in methods that may perform well under isolated conditions but remain suboptimal in realistic clinical environments.

In the context of class imbalance, traditional oversampling approaches such as SMOTE and its variants have shown effectiveness in low-dimensional tabular data. However, these methods rely on direct interpolation between samples and often fail to preserve semantic consistency in high-dimensional image spaces. In medical imaging, naive interpolation may generate unrealistic or noisy synthetic samples that negatively affect classifier performance. Although GAN-based approaches can improve visual realism, they often require large training datasets and substantial computational resources, limiting their applicability in limited-data medical settings. For representation learning, convolutional neural networks and more recently Vision Transformers have achieved strong performance in medical image classification tasks. Nevertheless, transformer architectures typically require large-scale datasets to fully exploit their modeling capacity. Existing attention mechanisms also tend to rely on a single attention formulation, potentially limiting their ability to capture complementary lesion characteristics in highly variable dermatological images.

Fairness in medical AI remains another critical challenge. Current fairness-aware approaches primarily focus on centralised training settings and often assume access to sensitive demographic information during training. Furthermore, many methods evaluate subgroups fairness independently from data imbalance, although minority demographic groups are frequently also minority disease groups. This creates compounded bias effects that are not adequately addressed by existing fairness interventions.

Federated learning has emerged as a promising solution for privacy-preserving medical

AI, but current federated fairness methods remain limited in several ways. Many approaches focus solely on reducing demographic bias without explicitly considering class imbalance across distributed clients. Additionally, domain shifts caused by differences in skin tone, acquisition devices, or institutional protocols can significantly degrade subgroups fairness and generalisation performance in federated environments.

These limitations motivate the contributions of this thesis. First, this research proposes SIA-SMOTE, a similarity-aware oversampling framework designed to improve synthetic sample quality in high-dimensional image data. Second, a self-attention fusion strategy is introduced to enhance representation learning in limited-data dermatological classification tasks. Third, the FedCIAL framework integrates fairness-aware adversarial learning into federated skin lesion classification to mitigate colour-related demographic bias. Finally, this thesis investigates the combined effect of domain adaptation and imbalance mitigation within federated learning, addressing the interconnected nature of subgroups fairness and imbalance in medical image classification. Table 2.1 summarises the datasets, tasks, class distributions, and split strategies used throughout the work packages presented in this thesis.

Overall, this thesis argues that reliable medical image classification systems cannot be achieved by addressing imbalance, subgroups fairness, or limited data in isolation. Instead, these challenges must be considered jointly to develop robust, equitable, and clinically deployable AI systems.

Table 2.1: Summary of datasets used across all work packages

Dataset	Task	Total Samples Used	Classification Class Ratio	Split Strategy	Used In
MNIST	Binary classification (3 vs 8)	7,280	1:25	80/20 split + cross-validation	WP1
Fashion-MNIST	Binary classification (T-shirt/top vs Trouser)	7,175	1:40	80/20 split + cross-validation	WP1
ISIC 2018	Binary skin lesion classification (Melanoma vs Nevus)	7,818	1:6	7,818 training / 144 test + cross-validation	WP1
ChestXray Pneumonia	Binary classification (Normal vs Pneumonia)	5,863	1:1.28	80/20 split + cross-validation	WP1
BreakHis	Binary breast cancer classification (Malignant vs Benign)	7,909	1:2.19	80/20 split + cross-validation	WP1
ISIC 2017	Task 1: Melanoma vs Others; Task 2: Seborrheic Keratosis vs Others	2,750	Task 1: 1:4; Task 2: 1:7	2000 train / 150 validation / 600 test	WP2
Fitzpatrick17k	Multi-class classification (Malignant / Benign / Non-neoplastic)	16,577	0.99:1:5.34	cross-validation	WP3
DDI	Binary classification (Malignant vs Benign)	640	1:2.31	cross-validation	WP4

Oversampling High Dimensional Medical Images Using SIA-SMOTE for Class Imbalance Mitigation

3.1 Introduction

Deep learning has emerged as a transformative approach in image analysis and pattern recognition, achieving state-of-the-art performance in various computer vision tasks. In particular, convolutional neural networks (CNNs) and related architectures have demonstrated remarkable success in image classification, benefiting from their ability to automatically learn hierarchical feature representations from raw pixel data. These advances have been especially prominent in domains where large-scale, well-balanced datasets are available. For example, benchmark datasets such as MNIST, containing 70,000 grayscale images of handwritten digits in 10 balanced classes, and CIFAR-10, consisting of 60,000 colour images evenly distributed across 10 object categories, have enabled the training of deep neural networks that achieve near-perfect classification accuracy [129]. The availability of such balanced datasets allows models to learn equally representative features for all classes, leading to robust and generalizable performance.

However, many real-world applications present a starkly different scenario. In several domains, including agriculture [130], medical imaging [22], [131], hyperspectral image analysis [132], remote sensing [133], and construction defect detection [134], datasets often exhibit class imbalance, where certain categories contain far fewer examples than others. In medical

imaging in particular, this imbalance is common due to the rarity of some conditions, ethical constraints in data collection, and limitations in patient recruitment. This uneven distribution poses a significant challenge for deep learning models, as they tend to prioritise learning patterns from the majority class. Consequently, the model's decision boundaries become biased, often resulting in poor performance when predicting minority class samples [1], [2]. In the context of medical diagnosis, such class imbalance bias is unacceptable, as it may lead to poor diagnosis of rare but critical conditions.

Two main strategies exist to address the class imbalance problem: data approaches and algorithm approaches. Data approaches involve resampling the dataset to achieve a more balanced class distribution. Examples include random oversampling, random undersampling, and more sophisticated techniques such as the Synthetic Minority Oversampling Technique (SMOTE) [135]. Recently, generative adversarial networks (GANs) have also been used to synthesise realistic images for minority classes [50]. In contrast, algorithm approaches modify the learning process to give greater weight to minority class samples, either through cost-sensitive learning, specialised loss functions such as Class Rectification Loss (CRL) [136], Large Margin Local Embedding (LMLE) [137]), or ensemble methods that combine multiple base classifiers [138].

This work focuses on the data approach, specifically oversampling techniques for high-dimensional datasets such as images. While SMOTE has proven to be an effective and widely adopted method for generating synthetic minority class samples through interpolation, it faces challenges when applied to image data. In high-dimensional feature spaces, interpolating between noisy or borderline samples can produce synthetic images that do not accurately reflect the underlying minority class distribution, potentially introducing artefacts and misleading the classifier. This problem is exacerbated in medical imaging, where even subtle differences in pixel patterns may correspond to clinically significant variations.

To address these limitations, this chapter presents SIA-SMOTE, a novel oversampling method that combines SMOTE with a Siamese network-based filtering mechanism. The proposed method operates in two stages:

- Decision Boundary Exploration, where synthetic samples are generated near the decision boundary between minority and majority classes to improve representation of complex

class borders.

- Siamese Network Filtering, where a Siamese neural network is trained to measure feature similarity between images and used to assess and select high-quality synthetic samples. Only those synthetic images that exhibit strong similarity to genuine minority class instances are retained, reducing the risk of introducing noisy data.

By integrating these two stages, SIA-SMOTE not only increases the quantity of minority class samples but also enhances their quality, ensuring that the augmented dataset better reflects the true data distribution. In addition to the boundary-generated samples, standard SMOTE is also applied to generate supplementary synthetic images from minority class pairs, thereby creating a richer and more balanced training set.

To evaluate the effectiveness of SIA-SMOTE, we conducted comprehensive experiments on both general-purpose datasets (MNIST, Fashion-MNIST) and three medical imaging datasets: a pneumonia chest X-ray dataset (ChestXray), a histopathological breast cancer dataset (BreakHis), and a dermoscopic skin lesion dataset (ISIC 2018). Performance was assessed using three key metrics, namely G-Mean, F1-score, and AUC. Three established oversampling techniques are compared, namely random oversampling, standard SMOTE, and ASN-SMOTE [139]. Results consistently demonstrate that SIA-SMOTE outperforms these methods, particularly in improving minority class recall without sacrificing overall accuracy.

The work presented in this chapter is based on the paper "SIA-SMOTE: A SMOTE-based Oversampling Method with Better Interpolation on High-Dimensional Data by Using a Siamese Network," which was presented at the 17th International Work-Conference on Artificial Neural Networks (IWANN), Ponta Delgada, Portugal, June 2023 [140]. The published version is available at https://link.springer.com/chapter/10.1007/978-3-031-43085-5_35.

The remainder of this chapter is structured as follows: Section 3.2 presents related work on oversampling techniques for imbalanced image classification, with emphasis on high-dimensional data. Section 3.3 details the proposed SIA-SMOTE methodology. Section 3.4 describes the experimental design, datasets, and evaluation metrics. Section 3.5 presents and discusses the results. Finally, Section 3.6 summarises the key findings of this work.

3.2 Related Work

The challenge of class imbalance in machine learning is well documented across a wide range of application domains, from agriculture and environmental monitoring to medical diagnostics and hyperspectral imaging. In medical image classification in particular, the problem is amplified due to the natural rarity of certain pathological conditions, ethical constraints on data collection, and variability in image acquisition across institutions. These limitations often lead to datasets where minority classes contain only a fraction of the samples available in majority classes, resulting in biased learning behaviour and degraded generalisation performance of deep learning models.

One of the earliest and simplest approaches for mitigating class imbalance is random oversampling [135], in which samples from the minority class are duplicated until the class distribution is balanced. While straightforward to implement, this method is prone to overfitting, as the classifier may memorise the duplicated samples instead of learning robust decision boundaries. Overfitting is particularly problematic in medical image analysis, where generalisation to unseen patient data is critical for clinical adoption.

To address the shortcomings of naive duplication, the Synthetic Minority Oversampling Technique (SMOTE) [41] was introduced. Rather than replicating existing samples, SMOTE synthesises new minority class examples by interpolating between a given sample and one of its nearest minority neighbours in feature space. This interpolation produces synthetic samples located along the line segment connecting the two original samples, thereby increasing variability in the minority class and reducing overfitting risk compared to random oversampling. SMOTE has been widely applied in medical imaging contexts, such as rare cancer subtype detection and lesion segmentation, and has shown improved classifier performance in many cases.

However, SMOTE's reliance on interpolation also introduces potential vulnerabilities. If the original minority samples contain noisy instances, mislabeled data, or outliers, these imperfections can propagate into the generated synthetic samples. This is especially concerning in medical datasets, where image quality may vary due to acquisition noise, artefacts, or manual annotation errors. As a result, SMOTE can unintentionally produce unrealistic or unrepresentative synthetic samples, degrading performance on unseen data. The effect is more pronounced in high-

dimensional spaces, such as those encountered when dealing directly with pixel intensities in medical images, where the curse of dimensionality can obscure the meaningful proximity relationships between samples.

To address this limitation, Adaptive Semi Neighbourhood SMOTE (ASN-SMOTE) [139] introduced a noise filtering mechanism prior to interpolation. ASN-SMOTE identifies minority class samples whose nearest neighbours predominantly belong to the majority class and excludes them from the oversampling process, thereby reducing the influence of borderline or noisy examples. New synthetic samples are generated only between pairs of minority classes that are closer to each other than to any instance of the majority class. This ensures that interpolation occurs in regions more representative of the minority distribution, reducing the likelihood of producing ambiguous or misleading synthetic examples. [139] demonstrated ASN-SMOTE's superior performance over nine state-of-the-art oversampling methods, including ADASYN, Borderline-SMOTE variants, k-means SMOTE, SVM-SMOTE, SWIM-RBF, and LoRAS on 24 benchmark datasets from the KEEL and UCI repositories using multiple classifiers such as K-Nearest Neighbours (KNN), Support Vector Machines (SVM), and Random Forests.

In summary, while SMOTE and its variants have advanced the state of oversampling techniques, challenges remain in effectively generating high-quality, representative synthetic samples, particularly for high-dimensional medical image data. This gap motivates the development of enhanced oversampling strategies, such as the proposed SIA-SMOTE, that combine interpolation-based synthesis with robust filtering mechanisms to improve sample quality and ultimately boost classification performance in imbalanced medical imaging tasks.

3.3 Method

The proposed method, termed SIA-SMOTE, integrates the classical SMOTE oversampling approach with a Siamese network-based filtering strategy to improve the quality of synthetic samples in high-dimensional datasets such as medical images. Unlike traditional oversampling methods that risk propagating noise or creating unrealistic samples, SIA-SMOTE aims to generate synthetic images that are both representative of the minority class and consistent with its underlying distribution. The method is composed of four sequential stages: (1) selection of

borderline minority samples as base instances, (2) synthesis of new images through interpolation near the decision boundary, (3) filtering of synthesised images using a Siamese network, and (4) application of SMOTE to balance the training dataset.

3.3.1 Base Sample Selection

Traditional oversampling methods often treat borderline samples, which are situated close to the decision boundary, as unreliable or noisy and exclude them from the synthesis process [139]. In contrast, the proposed method recognises that borderline samples contain crucial discriminative information about the overlap region between classes. These samples represent cases where the model’s classification decision is most uncertain, and thus, oversampling them can provide valuable insights for refining the decision boundary. Ignoring borderline samples may lead to biased models that fail to capture class boundaries accurately, thereby compromising generalisation.

In SIA-SMOTE, the base samples are chosen from the minority class where the nearest neighbour belongs to a different class. This ensures that the synthesis process explicitly targets samples located at or near the class boundary, thereby enriching the minority class distribution in the most informative regions of the feature space. By contrast, methods such as ASN-SMOTE [139] treat such borderline samples as noise, discarding them from the oversampling process. The decision in SIA-SMOTE to preserve and leverage these borderline cases forms the foundation of its boundary-aware synthesis strategy.

3.3.2 Image Synthesis

The next stage involves generating synthetic images by interpolating between a borderline minority sample (base sample) and one or more majority class neighbours located near the decision boundary, as illustrated in Figure 3.1. Unlike the classical SMOTE algorithm, which interpolates exclusively between two minority class samples, SIA-SMOTE deliberately incorporates majority class neighbours to encourage exploration of the decision boundary space. This design choice is motivated by the observation that many misclassifications occur in regions where class boundaries are ambiguous. By creating synthetic images in these regions, the classifier is exposed to more

challenging examples, improving its robustness.

Formally, given a base minority sample x_b and its nearest majority neighbour x_m , interpolation is performed to generate a new synthetic image x_{new} according to:

$$x_{new} = x_b + \lambda(x_m - x_b),$$

where λ is a random interpolation coefficient constrained by a hyperparameter $d \in (0, 1]$. The parameter d defines the maximum allowable distance of the interpolation point from the base sample. For instance, if $d = 0.5$, interpolation occurs between the base sample and the midpoint of the line connecting x_b and x_m , ensuring that the generated image remains more similar to the base minority sample. This constraint reduces the risk of synthetic samples drifting too far toward the majority class, which could otherwise compromise minority class integrity.

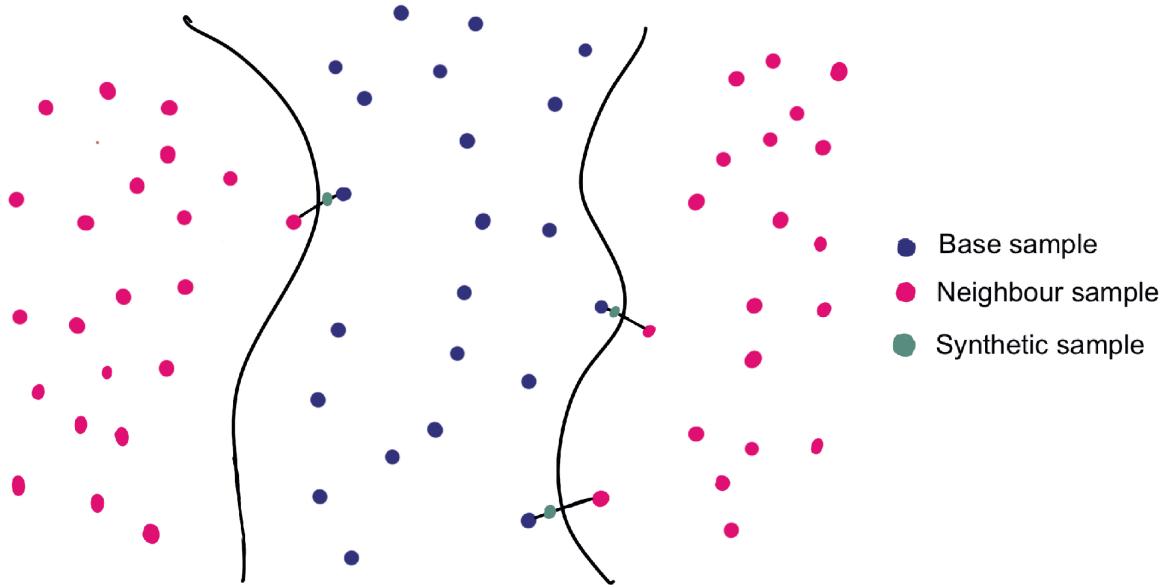


Figure 3.1: Interpolation between images from different classes at the decision boundary. Minority samples near the boundary are interpolated with nearby majority class samples to generate synthetic examples that enrich the decision space.

3.3.3 Image Selection with a Siamese Network

Since interpolation between classes risks producing unrealistic or noisy images, a selection mechanism is introduced to filter out low-quality synthetic samples. For this purpose, a Siamese

network [141] is employed. A Siamese network consists of two identical subnetworks with shared weights that embed input images into a feature space. By comparing the feature vectors of two images, the network outputs a similarity score indicating whether they belong to the same class.

The Siamese network is trained using a contrastive loss function that minimises the distance between embeddings of samples from the same class and maximises the distance for samples from different classes. This training regime allows the network to serve as a robust similarity assessor. During inference, each newly synthesised sample is paired with its corresponding base minority sample, and their similarity score is computed. A hyper-parameter t is defined as the similarity threshold. Only synthetic images with a similarity score above t are accepted into the augmented training set. In this way, the Siamese network acts as a quality gate, ensuring that only synthetic images consistent with minority class features are preserved for training.

3.3.4 Final Oversampling with SMOTE

After the Siamese network-based selection, the number of newly generated samples may still be insufficient to fully balance the class distribution. To address this, a final SMOTE oversampling step is applied to the filtered dataset. Importantly, the synthetic samples produced by this final SMOTE operation are again subjected to the same Siamese filtering process to remove low-quality or noisy images. This dual-layer approach, decision boundary synthesis followed by standard SMOTE with quality filtering, ensures that the augmented dataset not only achieves class balance but also maintains high fidelity to the underlying minority distribution.

3.4 Experimental Setup

To validate the effectiveness of the proposed SIA-SMOTE method, experiments were conducted on both benchmark datasets and real-world medical imaging datasets. The experimental design aims to assess the performance of the proposed method in addressing class imbalance in high-dimensional image data, with a particular focus on the medical imaging domain, where data scarcity and imbalance are common.

3.4.1 Datasets

Five datasets were used in this study, comprising two widely used benchmark datasets, namely MNIST and Fashion-MNIST (FMNIST) and three medical image datasets: ISIC 2018 (skin cancer), ChestXray (pneumonia detection), and BreakHis (histopathological breast cancer images). The benchmark datasets were selected to provide a controlled environment for initial evaluation, while the medical datasets were chosen to reflect real-world clinical challenges involving imbalanced and heterogeneous image data. In all experiments, binary classification tasks were considered for consistency and comparability across methods. For datasets with more than two classes, only subsets containing two classes were used.

MNIST (Modified National Institute of Standards and Technology) is a grayscale image dataset of handwritten digits with a resolution of 28×28 pixels, comprising 60,000 training images and 10,000 test images across 10 classes. In this study, digit '3' was designated as the majority class and digit '8' as the minority class. To simulate a highly imbalanced scenario, images from the minority class were randomly sampled to achieve an imbalance ratio of 1:25.

Fashion-MNIST (FMNIST) is a grayscale image dataset containing fashion product images from Zalando, with the same resolution and number of samples as MNIST. Of its original 10 classes, 'T-shirt/top' (label 0) was used as the majority class and 'Trouser' (label 1) as the minority class. The minority class was randomly sampled to achieve an imbalance ratio of 1:40, making it an even more challenging imbalance setting than MNIST.

ISIC 2018 Skin Lesion Dataset originates from the International Skin Imaging Collaboration (ISIC) Challenge 2018. It contains 10,015 colour dermoscopic training images and 193 test images, classified into seven lesion categories: Melanocytic nevus, Melanoma, Benign keratosis, Basal cell carcinoma, Actinic keratosis/Bowen's disease, Vascular lesion, and Dermatofibroma. In this study, the Melanocytic nevus class served as the majority class and the Melanoma class as the minority class, yielding an imbalance ratio of 1:6. This dataset reflects the clinical challenge of detecting melanoma from benign nevi.

ChestXray Pneumonia Dataset [142] consists of 5,863 chest radiographs categorised as Pneumonia or Normal. Following an 80%/20% train/test split, the Pneumonia class was set as the majority and the Normal class as the minority, with an imbalance ratio of 1:1.28. Although less

imbalanced than other datasets, this case demonstrates the importance of improving sensitivity for minority class detection in respiratory disease screening.

BreakHis Breast Cancer Dataset [143] contains 9,109 histopathological images of breast tumour tissues from 82 patients, captured at magnifications of 40X, 100X, 200X, and 400X. Of these, 2,480 images are benign, and 5,429 are malignant. All images were resised to 128×128 pixels for uniform processing. Detecting malignant tumours from histopathological slides is a critical task in oncology, and the imbalance ratio in this dataset is representative of clinical screening settings.

3.4.2 Evaluation Metrics

Given that class imbalance can lead to misleading accuracy scores, three metrics were selected to provide a more balanced assessment: geometric mean (G-mean), F1-score, and area under the receiver operating characteristic curve (AUC-ROC). These metrics are particularly relevant for imbalanced datasets, where sensitivity to minority class performance is critical.

The G-mean measures the balance between sensitivity and specificity, ensuring that a high score is achieved only when both metrics are strong:

$$G\text{-mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \quad (3.4.1)$$

The F1-score, also known as the Dice similarity coefficient (DSC), captures the trade-off between precision and recall. It is defined as:

$$F1\text{-score} = \frac{2TP}{2TP + FP + FN} \quad (3.4.2)$$

AUC-ROC evaluates the classifier's ability to distinguish between the two classes across different decision thresholds. It is computed as the area under the ROC curve, which plots the true positive rate (TPR) against the false positive rate (FPR).

These three metrics collectively provide a comprehensive evaluation: G-mean rewards balanced performance, F1-score emphasises correct detection of the minority class, and AUC-ROC measures overall discriminative ability.

3.4.3 Experimental Configuration

SIA-SMOTE was compared against three other oversampling techniques: random oversampling, SMOTE [41], and ASN-SMOTE [139]. Implementations of SMOTE and random oversampling were sourced from the imbalanced-learn Python module, while ASN-SMOTE was implemented using the author’s publicly available code¹. For SMOTE and ASN-SMOTE, the number of nearest neighbours was varied over {3, 5, 7, 9}.

Although GAN-based oversampling approaches can generate realistic-looking medical images, they were not included as primary baselines in this study for several practical reasons. First, GANs typically perform better with larger and more diverse training sets and can become unstable or produce low-fidelity or clinically unrealistic samples under severe data scarcity [144], which is the focus of this thesis. Second, the objective of the proposed SIA-SMOTE framework differs from full image-generation approaches. SIA-SMOTE improves the quality and semantic consistency of interpolation-based synthetic samples through similarity-aware filtering rather than synthesising entirely new images, so direct comparison to full-image GAN pipelines would conflate two distinct methodological goals. Finally, GAN frameworks often add substantially higher computational cost and hyperparameter sensitivity [145], which complicates fair, reproducible comparisons within the limited scope and resources of this work. For these reasons, this study focuses on comparisons against established interpolation-based oversampling methods that are more closely aligned with the proposed approach.

For SIA-SMOTE, the Siamese network employed ResNet-50 as the base feature extractor for three-channel (RGB) images and a custom VGG-based architecture for grayscale images. The VGG-based network consisted of three convolutional blocks with 32, 64, and 128 channels, respectively, each followed by max pooling. The Euclidean distance metric was used to measure similarity between image embeddings. Training was conducted for 100 epochs with a batch size of 64, using the RMSProp optimiser and contrastive loss. During inference, similarity thresholds $t \in \{0.02, 0.05, 0.1, 0.3, 0.5\}$ were tested for filtering synthetic samples.

For final classification, the same feature extractors used in the Siamese network were applied. All images were flattened into one-dimensional arrays before oversampling. Five-fold stratified

¹<https://github.com/yixinkai123/ASN-SMOTE/>

cross-validation was employed to ensure balanced representation of classes in each fold. Each fold was repeated three times to obtain stable performance estimates. All hyperparameters were optimised based on cross-validation performance. The implementation of SIA-SMOTE is available at <https://github.com/heroza/sia-smote>.

3.5 Results and Discussion

3.5.1 Quantitative Results

Table 3.1 presents the average performance and standard deviation of the four oversampling methods, namely Random Oversampling, SMOTE, ASN-SMOTE, and the proposed SIA-SMOTE, across three evaluation metrics (G-mean, F1-score, and AUC) on both validation and testing datasets. The rankings of each method, from best (rank 1) to worst (rank 4), are illustrated in Figure 3.2.

Across the majority of datasets and metrics, SIA-SMOTE achieves the highest overall performance, consistently placing first in average rank on both validation and test sets. On benchmark datasets such as MNIST and FMNIST, SIA-SMOTE shows competitive or superior performance to other methods, with notable improvements in G-mean and AUC. On medical datasets ISIC 2018, ChestXray, and BreakHis, SIA-SMOTE often yields the best G-mean and AUC scores, indicating better balance between sensitivity and specificity, as well as stronger discriminative ability.

One particularly strong result is observed on the ISIC 2018 melanoma classification task, where SIA-SMOTE achieves the highest performance across all three metrics for both validation and test sets. This is clinically relevant because correctly identifying melanoma in dermoscopic images is a high-stakes decision, and improvements in sensitivity without sacrificing specificity can directly impact patient outcomes.

However, the method does not uniformly dominate all datasets. On FMNIST validation data, for example, Random Oversampling achieves slightly higher G-mean and F1-score. This indicates that in certain low-noise, lower-dimensional feature spaces, simpler resampling techniques may still compete effectively.

3.5.2 Qualitative Analysis of Filtered Images

To illustrate the role of the Siamese network in maintaining sample quality, Figure 3.3 shows examples of synthetic images that were either accepted (filtered-in) or rejected (filtered-out) during augmentation. The first two columns depict samples generated through decision boundary interpolation, while the last two columns correspond to SMOTE-generated samples.

Rejected images often exhibit artefacts, unrealistic structures, or distortions inconsistent with their class label. By excluding these low-quality samples, SIA-SMOTE prevents the introduction of noise into the training data, which in turn helps maintain classifier robustness.

3.5.3 Statistical Significance Testing

To formally assess differences between methods, Friedman’s test was applied to the validation set performance rankings (Table 3.2). All p-values were below the 0.05 threshold, indicating statistically significant differences in performance among the four methods.

Subsequent Conover post-hoc tests (Table 3.3), with a Bonferroni-corrected significance level of 0.0083, reveal that SIA-SMOTE performs significantly better than ASN-SMOTE for all three metrics. Differences with Random Oversampling and SMOTE were not statistically significant at this threshold, although SIA-SMOTE still achieved better mean rankings.

The lack of significant differences with Random Oversampling on some datasets is noteworthy. It suggests that in highly separable feature spaces or with relatively clean data, the benefit of sophisticated oversampling strategies may be less pronounced. Conversely, in complex and noisy high-dimensional spaces such as medical images, the filtering advantage of SIA-SMOTE becomes more impactful.

3.5.4 Interpretation in the Context of Medical Imaging

The improvements seen with SIA-SMOTE are particularly relevant to the medical domain, where imbalanced datasets are the norm rather than the exception. For example:

- In melanoma detection (ISIC 2018), improved G-mean and AUC scores translate to better sensitivity for rare malignant lesions while maintaining specificity for benign cases.

- In breast cancer histopathology (BreakHis), higher F1-scores mean fewer false negatives, which is critical for early detection and treatment.
- For pneumonia detection in ChestXray, the method achieves high validation performance, though its test performance is comparable to other methods, highlighting the challenge of domain shift in real-world clinical data.

By filtering out poor-quality synthetic images and augmenting borderline regions of the decision space, SIA-SMOTE effectively exposes the classifier to more informative patterns, reducing bias toward the majority class.

3.5.5 Strengths and Limitations

A key strength of SIA-SMOTE lies in its dual augmentation strategy: leveraging decision boundary interpolation to generate challenging samples, and applying SMOTE to further balance the dataset, with both stages filtered by a similarity-based Siamese network. This not only improves minority class representation but also reduces the risk of overfitting to noisy synthetic data.

However, limitations exist. The method’s reliance on the Siamese network introduces additional computational cost during both training and inference for sample filtering. Furthermore, performance gains are dataset dependent, with smaller or less noisy datasets sometimes benefiting less from the method’s complexity, as observed in FMNIST.

3.6 Conclusion

This study introduces SIA-SMOTE, an enhanced oversampling technique that integrates SMOTE with Siamese network-based quality filtering and decision boundary interpolation. Evaluations on both benchmark and medical imaging datasets show that SIA-SMOTE achieves top-ranked performance across G-mean, F1-score, and AUC metrics, often significantly outperforming existing methods.

By selectively generating and filtering synthetic samples, SIA-SMOTE not only addresses the imbalance problem but also improves the representativeness and quality of augmented datasets.

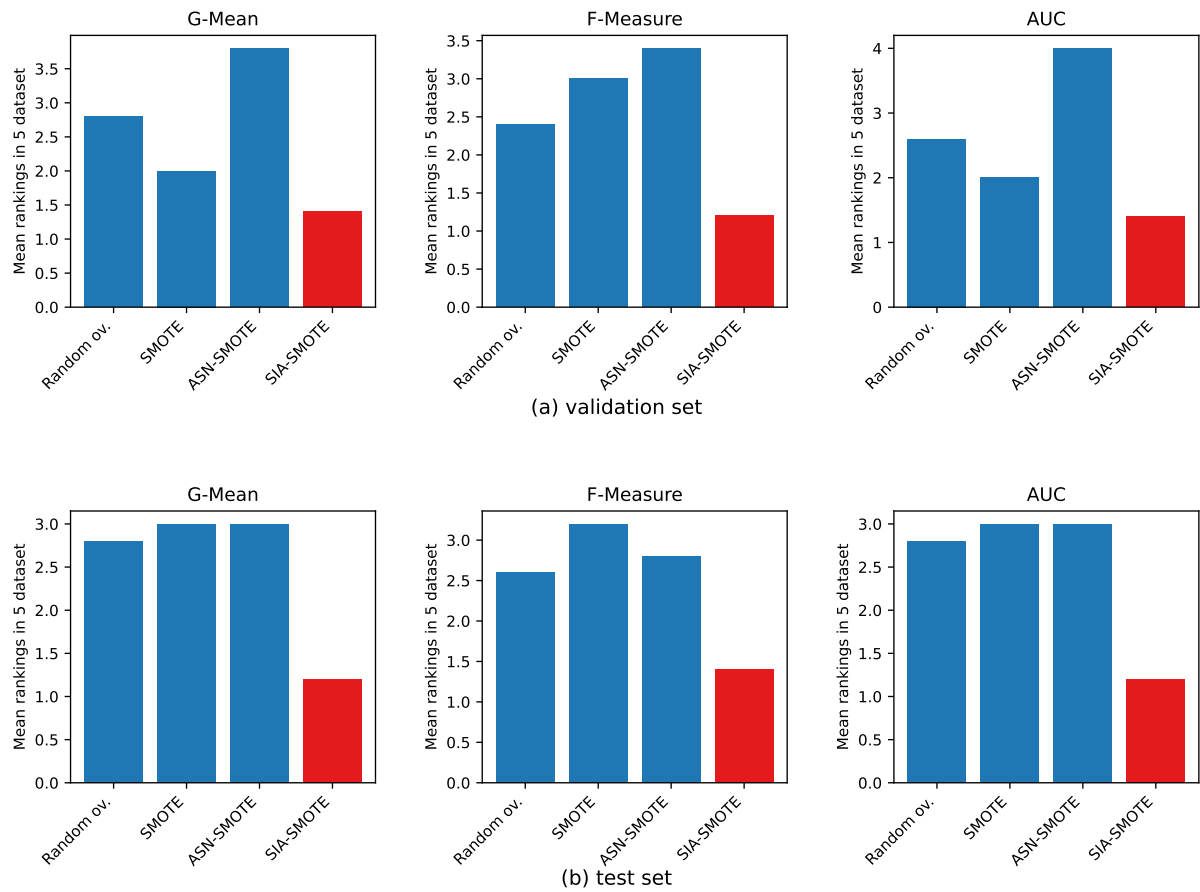


Figure 3.2: Ranking of oversampling methods in terms of three performance metrics

This makes it particularly suitable for high-dimensional, noise-prone domains such as medical imaging, where both sensitivity to minority cases and robustness to noise are critical for reliable automated diagnosis.

Table 3.1: Comparison of performance of oversampling methods

Dataset	Method	G-mean	F-measure	AUC
MNIST				
Validation set	Random oversampling	99.134 $\pm$ 0.652	94.704 $\pm$ 2.577	99.137 $\pm$ 0.649
	SMOTE	99.184 $\pm$ 0.619	92.479 $\pm$ 4.584	99.187 $\pm$ 0.617
	ASN-SMOTE	99.001 $\pm$ 0.904	95.385 $\pm$ 2.918	99.007 $\pm$ 0.896
	SIA-SMOTE	99.786 $\pm$ 0.320	96.818 $\pm$ 2.185	99.787 $\pm$ 0.318
	Random oversampling	99.752 $\pm$ 0.411	97.899 $\pm$ 1.580	99.753 $\pm$ 0.409
	SMOTE	99.614 $\pm$ 0.673	96.655 $\pm$ 3.434	99.617 $\pm$ 0.665
	ASN-SMOTE	99.249 $\pm$ 0.988	97.478 $\pm$ 1.740	99.257 $\pm$ 0.971
	SIA-SMOTE	99.826 $\pm$ 0.329	97.763 $\pm$ 2.169	99.827 $\pm$ 0.328
FMNIST				
Validation set	Random oversampling	99.592 $\pm$ 0.658	96.875 $\pm$ 1.344	99.594 $\pm$ 0.652
	SMOTE	99.579 $\pm$ 0.545	92.736 $\pm$ 4.821	99.581 $\pm$ 0.541
	ASN-SMOTE	98.073 $\pm$ 0.961	93.335 $\pm$ 3.164	98.092 $\pm$ 0.945
	SIA-SMOTE	99.247 $\pm$ 0.806	96.282 $\pm$ 2.617	99.253 $\pm$ 0.800
	Random oversampling	96.325 $\pm$ 1.400	91.157 $\pm$ 2.387	96.393 $\pm$ 1.347
	SMOTE	97.564 $\pm$ 0.689	92.517 $\pm$ 4.515	97.590 $\pm$ 0.668
	ASN-SMOTE	99.059 $\pm$ 0.956	94.477 $\pm$ 5.266	99.067 $\pm$ 0.947
	SIA-SMOTE	99.933 $\pm$ 0.065	97.464 $\pm$ 2.441	99.933 $\pm$ 0.065
ISIC2018				
Validation set	Random oversampling	81.243 $\pm$ 2.101	73.734 $\pm$ 3.107	82.589 $\pm$ 1.783
	SMOTE	82.346 $\pm$ 1.454	74.406 $\pm$ 1.252	83.471 $\pm$ 1.179
	ASN-SMOTE	81.446 $\pm$ 3.131	67.523 $\pm$ 1.935	82.338 $\pm$ 2.473
	SIA-SMOTE	83.034 $\pm$ 1.468	74.934 $\pm$ 2.374	84.036 $\pm$ 1.264
	Random oversampling	86.267 $\pm$ 5.103	80.969 $\pm$ 6.930	87.093 $\pm$ 4.501
	SMOTE	88.857 $\pm$ 3.436	83.170 $\pm$ 4.763	89.311 $\pm$ 3.118
	ASN-SMOTE	85.846 $\pm$ 4.761	74.000 $\pm$ 5.077	86.438 $\pm$ 4.226
	SIA-SMOTE	89.045 $\pm$ 4.479	83.750 $\pm$ 4.015	89.578 $\pm$ 3.969
ChestXray				
Validation set	Random oversampling	98.200 $\pm$ 0.516	97.499 $\pm$ 0.684	98.207 $\pm$ 0.513
	SMOTE	98.463 $\pm$ 0.519	97.633 $\pm$ 0.540	98.468 $\pm$ 0.515
	ASN-SMOTE	97.579 $\pm$ 0.623	95.324 $\pm$ 1.758	97.590 $\pm$ 0.616
	SIA-SMOTE	98.477 $\pm$ 0.509	97.713 $\pm$ 0.644	98.480 $\pm$ 0.505
	Random oversampling	66.748 $\pm$ 2.845	61.498 $\pm$ 3.655	72.265 $\pm$ 1.916
	SMOTE	65.516 $\pm$ 2.194	59.994 $\pm$ 2.878	71.458 $\pm$ 1.487
	ASN-SMOTE	70.626 $\pm$ 7.193	64.885 $\pm$ 8.209	74.722 $\pm$ 4.916
	SIA-SMOTE	66.927 $\pm$ 3.550	61.779 $\pm$ 4.531	72.433 $\pm$ 2.380
Breakhis				
Validation set	Random oversampling	95.895 $\pm$ 0.525	94.602 $\pm$ 0.633	95.916 $\pm$ 0.515
	SMOTE	96.010 $\pm$ 0.308	94.581 $\pm$ 0.641	96.027 $\pm$ 0.306
	ASN-SMOTE	94.313 $\pm$ 1.394	91.599 $\pm$ 1.585	94.371 $\pm$ 1.336
	SIA-SMOTE	96.073 $\pm$ 0.532	94.887 $\pm$ 0.815	96.095 $\pm$ 0.529
	Random oversampling	96.627 $\pm$ 0.394	95.224 $\pm$ 0.522	96.633 $\pm$ 0.391
	SMOTE	96.401 $\pm$ 0.527	94.758 $\pm$ 0.774	96.406 $\pm$ 0.526
	ASN-SMOTE	94.750 $\pm$ 1.348	91.880 $\pm$ 2.176	94.813 $\pm$ 1.295
	SIA-SMOTE	96.648 $\pm$ 0.464	95.300 $\pm$ 0.663	96.655 $\pm$ 0.461

Table 3.2: Friedman’s test result

Metric	p-value
G-Mean	0.02110
F-Measure	0.04057
AUC	0.010891

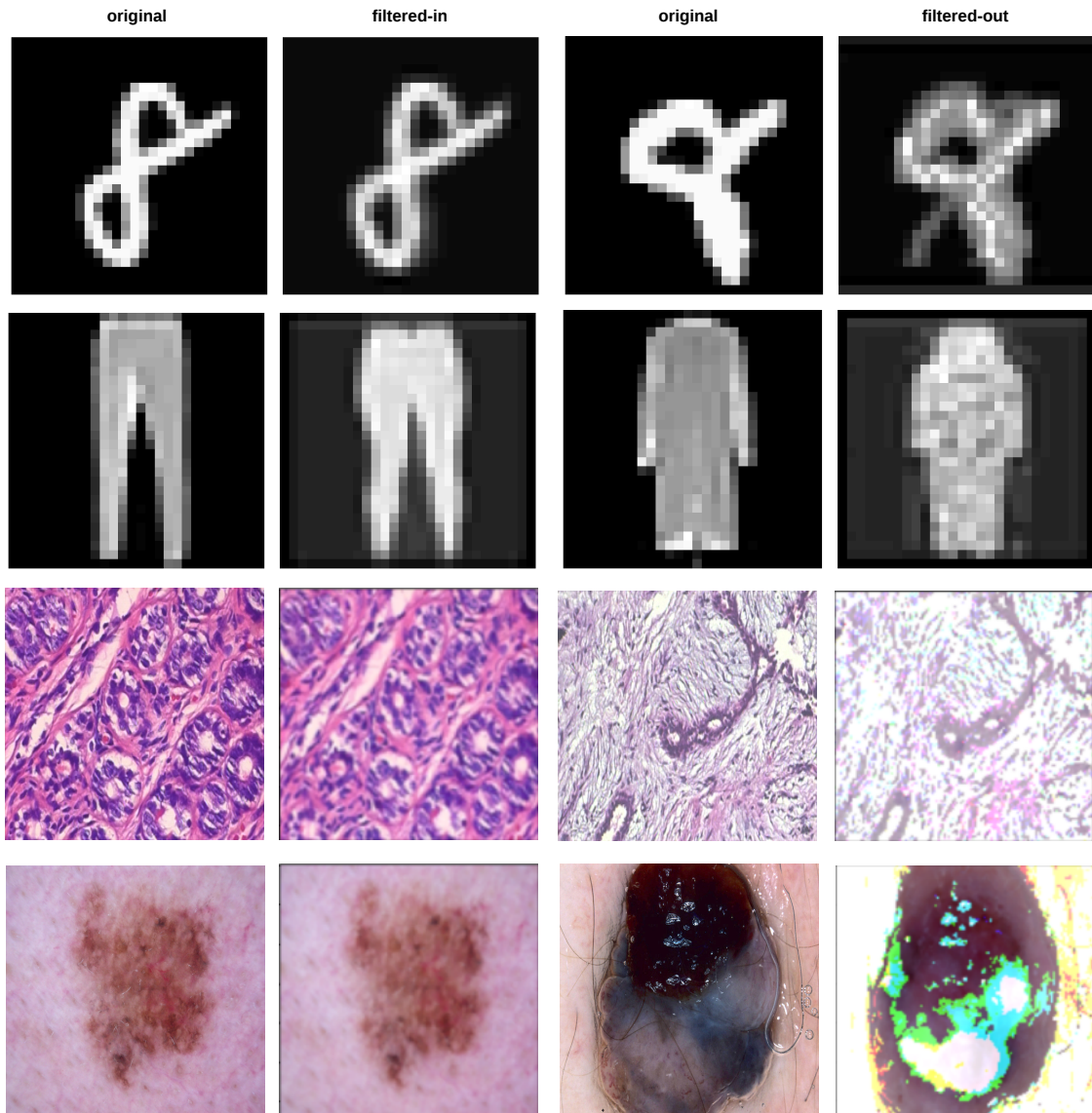


Figure 3.3: Comparison between real minority-class images and synthetic samples generated by SIA-SMOTE. Columns 1 and 3 show original minority-class images, Column 2 presents synthetic samples retained after similarity-aware filtering, and Column 4 shows synthetic samples rejected by the filtering process.

Table 3.3: Conover test result

Method to be compared with SIA-SMOTE (sig. level after correction: 0.0083)	G-Mean		F-Measure		AUC	
	t-stat	p-value	t-stat	p-value	t-stat	p-value
Random oversampling	2.5849	0.0238	1.9639	0.0731	2.5980	0.0233
SMOTE	1.1078	0.2896	2.9459	0.0122	1.2990	0.2183
ASN-SMOTE	4.4312	0.0008	3.6005	0.0036	5.6291	0.0001

Self-Attention Fusion with Vision Transformer for Imbalanced Skin Lesion Classification

4.1 Introduction

Skin cancer is one of the most prevalent types of cancer worldwide, with incidence rates continuing to rise annually [146]. Early detection and accurate diagnosis are crucial for improving patient outcomes, as timely intervention significantly increases survival rates [147]. For instance, delays in detecting suspicious skin lesions can reduce five-year survival rates by up to 20%. To address this, many healthcare systems have adopted fast-track diagnostic pathways, such as the two-week rule in the United Kingdom, which mandates that patients with suspected cancer be seen by a specialist within two weeks of referral. However, the demand for these rapid assessments has surged dramatically, with referrals increasing from 159,430 in 2009/2010 to 506,456 in 2019/2020 [148]. This growth has placed considerable strain on healthcare resources, often leading to longer waiting times for non-urgent cases, which sometimes exceed 18 weeks. The COVID-19 pandemic further exacerbated these delays by prioritising urgent cases and postponing routine appointments [149]. With an ageing population, the number of skin cancer referrals is expected to continue increasing in the coming years.

The process of skin cancer analysis typically involves the visual inspection of suspicious skin lesions by dermatologists, aided by dermoscopy and, increasingly, AI-based image analysis

systems. Deep learning models, in particular, have shown promising results in assisting clinicians with the detection and classification of skin cancer [150]. However, several significant challenges remain in developing reliable AI systems for this domain.

One of the primary challenges is the issue of class imbalance in skin lesion datasets. In many publicly available datasets, benign lesions are far more common than malignant ones, resulting in highly skewed class distributions [150]. This imbalance can bias deep learning models towards the majority class, leading to reduced sensitivity for detecting rare but clinically important classes such as melanoma. From a fairness perspective, this bias can result in unequal diagnostic performance across patient groups, which is particularly concerning in medical applications where false negatives can have severe consequences.

Another challenge lies in the adaptation of state-of-the-art architectures, such as the Vision Transformer (ViT) [151], to the medical imaging domain. While ViTs have achieved remarkable success in natural image classification by leveraging self attention to capture long range dependencies, their performance in medical image tasks can be limited by two factors: (1) the scarcity of labeled training data in medical domains compared to large scale natural image datasets, and (2) the difficulty of transferring features learned from natural images to the unique texture, colour, and structural patterns of medical images [152]. Training ViTs from scratch on small medical datasets typically leads to overfitting and poor generalisation, while fine-tuning pretrained models can be hindered by differences in domain-specific features.

This chapter addresses these challenges by proposing a method that enhances ViT performance on imbalanced skin cancer datasets. The proposed approach combines three strategies: (1) data level augmentation techniques to increase the representation of minority classes [153]; (2) algorithm level adjustments through weighted and class-balanced loss functions to prioritise rare classes during training [152]; and (3) architectural modification of the ViT by replacing its last n attention layers with a self attention fusion module, which integrates multiple attention mechanisms to improve the extraction of discriminative features from dermatological images. The combination of these strategies aims to mitigate the class imbalance problem while preserving the ViT's capacity for capturing global contextual information, leading to fairer and more accurate skin cancer classification.

The remainder of this chapter is structured as follows. Section 4.2 reviews existing work on deep learning for skin cancer classification, class imbalance handling, and attention mechanism design. Section 4.3 describes the proposed self-attention fusion architecture and the imbalance-aware loss functions used in this study. Section 4.4 presents the datasets, experimental setup, and evaluation metrics. Section 4.5 reports the experimental results and provides a detailed discussion. Finally, Section 4.6 summarises the contributions and outlines potential future directions.

4.2 Related Work

Skin cancer classification using deep learning has been an active research area due to the increasing demand for accurate, automated diagnostic tools [150], [154]. Early works primarily relied on convolutional neural networks (CNNs) trained on dermoscopic images, with architectures such as ResNet [155], DenseNet [156], and EfficientNet [157] achieving strong performance on large-scale image datasets. These CNN-based approaches exploit local receptive fields and hierarchical feature extraction to model dermatological patterns effectively. However, CNNs are inherently limited in modelling long-range dependencies between image regions, which can be important for capturing global contextual cues in skin lesion images [151].

The introduction of Vision Transformers (ViTs) [151] brought a paradigm shift in computer vision by leveraging the self-attention mechanism originally popularised in natural language processing [118]. ViTs divide an image into fixed-size patches, embed them into a latent space, and process the sequence of patches using transformer encoders. By modelling pairwise relationships between all patches, ViTs can capture long-range dependencies and global structural information. However, ViTs typically require large amounts of labelled training data to achieve competitive performance. When trained from scratch on small datasets, ViTs tend to underperform CNNs due to insufficient supervision and poor inductive bias for locality and translation equivariance [158].

Transfer learning has been a widely adopted strategy to overcome the data requirements of ViTs in specialised domains such as medical imaging. Pretraining on large datasets such as ImageNet-21k followed by fine-tuning on target medical datasets has shown substantial performance improvements. Self-supervised learning methods, such as DINO [124] and MoCo-v3 [159], further enhance pretraining efficiency by learning generalizable visual representations

without label supervision. Nevertheless, domain adaptation remains a challenge since features learned from natural images may not fully transfer to medical image distributions.

Another major issue in skin cancer image analysis is class imbalance, where certain lesion types (e.g., melanoma) are significantly underrepresented compared to benign classes [150]. This imbalance can bias the model toward majority classes, reducing sensitivity to minority classes that are often clinically critical. Several strategies have been proposed to address class imbalance in deep learning. Data level methods, such as oversampling and augmentation [153], aim to increase minority class representation. Algorithm-level methods include cost-sensitive learning, where weighted loss functions such as weighted cross-entropy, focal loss [60], and class-balanced loss [61] emphasise minority classes during training. More advanced approaches combine both strategies to maximise the benefits.

Attention mechanisms have also been adapted to better handle imbalanced data and improve feature representation. For example, hybrid attention architectures fuse multiple attention types such as additive [116], multiplicative [117], and scaled dot-product attention [118] to capture complementary information. In skin lesion classification, such fusion has the potential to enhance the discriminative ability of models by integrating salient local patterns with global contextual cues, especially when rare lesion features are subtle.

Despite these advancements, relatively few studies have explored the combination of attention mechanism fusion within ViT architectures specifically tailored for imbalanced medical datasets. The proposed method in this chapter addresses this gap by modifying the last n attention layers of a ViT with a self-attention fusion module and integrating imbalance-aware loss functions, aiming to improve classification fairness and accuracy for underrepresented skin cancer types.

4.3 Method

This section details the proposed framework for enhancing skin cancer classification using a Vision Transformer (ViT) backbone augmented with a novel self-attention fusion mechanism. The method is designed to address the dual challenges of imbalanced datasets and domain-specific feature extraction in dermatological imaging. The approach consists of two main components: (1) a modified network architecture that integrates multiple attention mechanisms via a fusion strategy,

and (2) tailored loss functions that emphasise minority classes while mitigating overfitting.

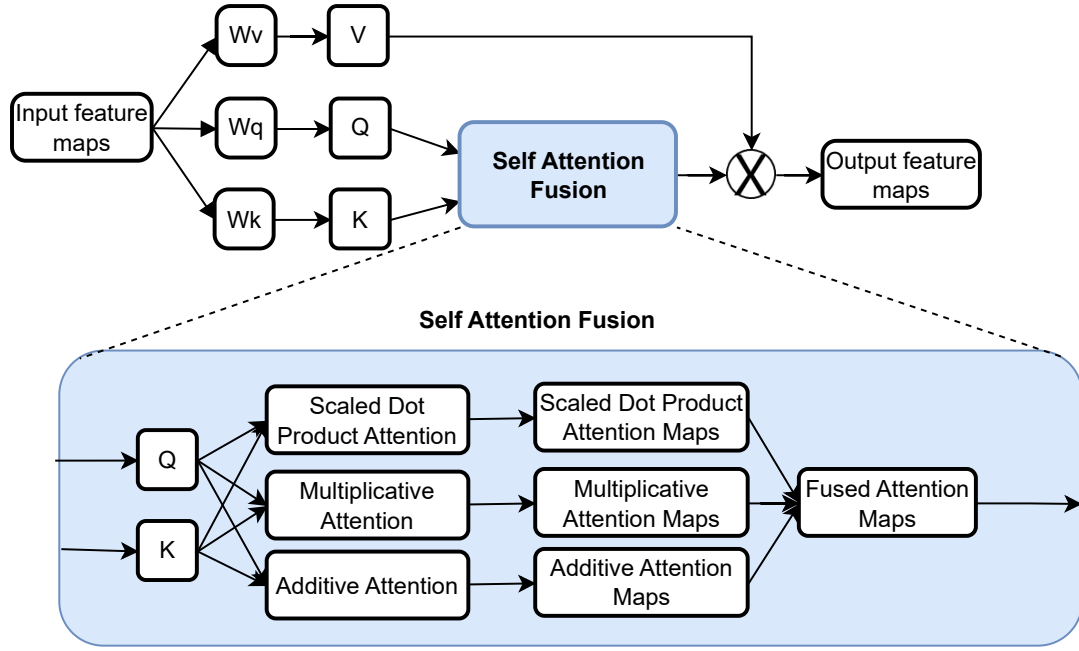


Figure 4.1: The proposed self-attention fusion architecture. The final n attention layers in the Vision Transformer are replaced with fusion layers combining scaled dot product, multiplicative, and additive attention mechanisms.

4.3.1 Network Architecture with Self-Attention Fusion

In our proposed method for skin cancer classification, the Vision Transformer B16 (ViT-B16) architecture [151] is adopted as the backbone. The choice of ViT-B16 is motivated by its proven capability to capture global contextual information through self-attention, enabling it to handle complex spatial relationships in medical images. Unlike Convolutional Neural Networks (CNNs), which rely on local receptive fields, the ViT operates on non-overlapping image patches, allowing for the modelling of long-range dependencies from the earliest stages of processing.

Two distinct pre-training regimens are considered for the backbone network:

- **Supervised Pre-Training:** The first variant is pre-trained on ImageNet-21k (14 million images, 21,843 classes) followed by fine-tuning on ImageNet-1k (1.2 million images, 1,000 classes). This mirrors the training process in the original ViT study [151].
- **Self-Supervised Pre-Training:** The second variant leverages the DINO self-supervised

learning paradigm [124] on ImageNet-1k, enabling the network to learn robust visual representations without reliance on explicit labels, which is advantageous when transferring to medical domains with limited annotated data.

To further enhance model generalisation and mitigate overfitting, particularly important when training on relatively small medical datasets, RandAugment [160] is employed. This augmentation strategy applies a set of randomly selected transformations with dynamically chosen magnitudes, thereby increasing variability in the training set without manual augmentation tuning.

Attention Fusion Layer Design. To adapt the ViT for dermatological image analysis, where both local texture details and global contextual cues are critical, we propose a fusion of three complementary attention mechanisms within each modified attention layer. The selected mechanisms were chosen not only because they capture different types of feature interactions, but also because they produce attention maps with compatible dimensional structures. This compatibility allows straightforward aggregation using simple yet effective fusion operations such as averaging, element-wise multiplication, and maximum selection without requiring additional projection or alignment layers. Such a design keeps the architecture lightweight and computationally efficient, which is important in limited-data medical imaging settings where overly complex modules may increase the risk of overfitting.

The three selected mechanisms are:

1. **Scaled Dot Product Attention** [118] is efficient for modelling global context and long-range dependencies, calculated via the dot product of query and key matrices scaled to prevent gradient vanishing. This mechanism forms the core of the original Transformer architecture and provides strong capability for capturing global semantic relationships across image patches.
2. **Multiplicative Attention** [117] introduces element-wise multiplication between query and key matrices, enabling selective amplification of interdependent features. In dermatological imaging, where subtle visual structures such as lesion borders or pigmentation patterns may

carry diagnostic importance, multiplicative interactions can strengthen highly correlated discriminative features.

3. **Additive Attention** [116] computes attention weights via a learnable combination of query and key features followed by a non-linear transformation. Compared with purely multiplicative formulations, additive attention provides greater flexibility in modelling complex feature relationships and may better capture subtle non-linear interactions commonly found in medical images.

These mechanisms were selected because they offer complementary behaviors while remaining structurally compatible for direct fusion. We did not incorporate other attention modules such as squeeze and excitation blocks [161], or convolutional block attention modules [162] because they operate on different feature representations or dimensional structures and would typically require additional adaptation layers for integration. Likewise, more complex mechanisms such as deformable attention [163] or hierarchical/window-based self-attention [164] introduce substantially higher computational overhead and architectural complexity, which can be prohibitive for relatively small medical imaging datasets where sample efficiency and low overfitting risk are priorities. Therefore, the objective of this work was to balance representational diversity, architectural simplicity, and computational efficiency when choosing attention components.

The resulting attention maps from these mechanisms are integrated through one of three fusion strategies:

$$\text{Avg Fusion}(i, j) = \frac{1}{n} \sum_{k=1}^n \text{Attmap}_k(i, j) \quad (4.3.1)$$

$$\text{Max Fusion}(i, j) = \max_k \text{Attmap}_k(i, j) \quad (4.3.2)$$

$$\text{Prod Fusion}(i, j) = \prod_{k=1}^n \text{Attmap}_k(i, j) \quad (4.3.3)$$

The average fusion strategy equally weights contributions from all mechanisms, promoting balanced feature integration and reducing dependence on any single attention formulation. The maximum fusion strategy highlights the most salient spatial locations detected by any mechanism, ensuring that strong discriminative cues are preserved even if only one mechanism responds strongly to a region. The multiplication fusion strategy emphasises features consistently

attended to across mechanisms, potentially reducing noise by reinforcing consensus regions shared among all attention maps. Collectively, these fusion strategies enable the model to exploit complementary attention behaviours while maintaining computational simplicity and compatibility with pretrained ViT representations.

Selective Layer Replacement. Rather than replacing all self-attention layers in the ViT, only the last n layers ($n \in \{1, 2, 3\}$ in our experiments) are substituted with attention fusion layers. This design decision is grounded in transfer learning principles: the earlier layers of a pre-trained model capture generic low-level patterns (e.g., edges, textures) that are transferable across domains, whereas later layers are more task specific and thus benefit most from adaptation. This selective modification preserves useful generic representations while injecting domain-specific attention capabilities in the higher-level processing stages, ensuring effective adaptation to skin lesion classification.

4.3.2 Loss Functions

Addressing class imbalance is critical in skin cancer classification, as malignant lesions often represent a minority class. This imbalance can bias models towards majority classes, leading to poor sensitivity in detecting rare but clinically important cases. To counter this, we evaluate and integrate several loss functions that adjust the training dynamics to prioritise minority class performance.

The following loss functions are explored:

Cross-Entropy Loss (CE):

$$CE(p_t) = -\log(p_t) \quad (4.3.4)$$

The baseline objective measures the negative log-likelihood of the correct class probability p_t .

Weighted Cross-Entropy Loss (WCE):

$$\text{Weighted } CE(p_t) = -\alpha \log(p_t) \quad (4.3.5)$$

Weights $\alpha \in \{0.2, 0.5, 0.7, 0.8, 0.9\}$ are applied to the minority class to increase its contribution to the total loss.

Focal Loss (FL):

$$FL(p_t) = -\alpha(1 - p_t)^\gamma \log(p_t) \quad (4.3.6)$$

Proposed by Lin et al. [60], focal loss down-weights easy examples and focuses training on hard-to-classify cases. Parameters $\alpha \in \{0.2, 0.5, 0.7, 0.8, 0.9\}$ and $\gamma \in \{1, 2\}$ control class weighting and focusing intensity, respectively.

Class-Balanced Loss (CB):

$$CB(p_t) = -\frac{1 - \beta}{1 - \beta^{n_t}} (1 - p_t)^\gamma \log(p_t) \quad (4.3.7)$$

This loss [61] explicitly accounts for the number of samples n_t per class, with $\beta \in \{0.999, 0.9999\}$ controlling the extent of re-weighting.

Label Smoothing (LS): As proposed by Szegedy et al. [165], label smoothing replaces hard target labels with softened distributions, using $\alpha \in \{0.1, 0.25, 0.5\}$ to reduce overconfidence and improve generalisation.

Implementation Considerations. During training, each loss function is evaluated independently with the same model architecture to ensure fair comparison. All hyperparameters are tuned based on validation performance. The best performing loss function is selected for final model evaluation. This systematic approach ensures that the selected loss not only mitigates imbalance but also synergises effectively with the attention fusion architecture.

4.4 Experimental Setup

4.4.1 Dataset and Performance Metrics

The experimental evaluation is conducted using the publicly available International Skin Imaging Collaboration (ISIC) 2017 dataset [166], which is a benchmark dataset for skin lesion analysis. The dataset consists of dermoscopic images collected from multiple clinical centres, capturing a wide range of lesion appearances, imaging conditions, and patient demographics. Each image is annotated with a ground truth diagnosis provided by expert dermatologists. The dataset is partitioned into three disjoint subsets: 2,000 images for training, 150 images for validation, and 600 images for testing. All samples are categorised into one of three diagnostic classes: melanoma, seborrheic keratosis, or nevus.

A major challenge inherent in this dataset is the severe class imbalance: melanoma cases are underrepresented compared to benign lesions. Such an imbalance can bias model training towards the majority class, making it essential to adopt evaluation metrics that reflect performance across all classes rather than overall accuracy alone.

To provide a comprehensive evaluation, we employ the following metrics: Balanced Accuracy (BA), Area Under the ROC Curve (AUC), Average Precision (AP), Overall Accuracy (Acc), and the F1 Score (F1). Each metric offers a different perspective on performance, enabling a more nuanced assessment of the model's effectiveness, particularly in handling minority classes.

Balanced Accuracy (BA) accounts for both sensitivity and specificity, providing a fairer metric for imbalanced datasets:

$$BA = \frac{\text{Sensitivity} + \text{Specificity}}{2} \quad (4.4.1)$$

where Sensitivity (True Positive Rate, TPR) and Specificity (True Negative Rate, TNR) are given by:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (4.4.2)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4.4.3)$$

Here, TP , TN , FP , and FN denote the counts of true positives, true negatives, false positives, and false negatives, respectively.

The AUC measures the model's ability to discriminate between positive and negative samples across varying decision thresholds, with higher values indicating better separability.

Average Precision (AP) summarises the precision-recall curve by integrating the precision values at different recall levels:

$$AP = \sum_n (\text{Recall}_n - \text{Recall}_{n-1}) \times \text{Precision}_n \quad (4.4.4)$$

where Recall_n and Precision_n denote the recall and precision values at the n -th threshold, respectively.

Overall Accuracy (Acc) measures the proportion of correct predictions:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.4.5)$$

The F1 Score, the harmonic mean of precision and recall, provides a balanced measure of both metrics:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.4.6)$$

By reporting all five metrics, we ensure that the evaluation captures both general classification ability and sensitivity to minority classes.

4.4.2 Experimental Configuration

The experimental design is structured to systematically analyse the influence of two main factors: (1) the choice of loss function, and (2) the replacement of the final attention layers with the proposed attention fusion mechanism.

We follow the ISIC 2017 Challenge protocol, which defines two binary classification subtasks:

- **Classification Task 1:** Distinguish melanoma (positive class) from nevus and seborrheic keratosis (negative class), yielding an imbalance ratio of approximately 1:4.
- **Classification Task 2:** Distinguish seborrheic keratosis (positive class) from nevus and

melanoma (negative class), yielding a more severe imbalance ratio of approximately 1:7.

For primary performance comparisons of the proposed method, we focus on Classification Task 2, as it presents a greater challenge due to its higher imbalance ratio. Subsequently, we also report results from a combined evaluation of both tasks to benchmark against the official ISIC 2017 Challenge leaderboard.

For each experiment, the ViT-B16 architecture [151] is used as the baseline model. Two initialisation strategies are tested: (i) supervised pre-training on ImageNet-21k followed by ImageNet-1k fine-tuning, and (ii) self-supervised pre-training on ImageNet-1k via the DINO framework [124]. All models are trained for 20 epochs using the AdamW optimiser [167] with a fixed learning rate of 2×10^{-5} . RandAugment [160] is applied for on-the-fly data augmentation to improve generalisation.

During training, the best-performing checkpoint is selected based on the lowest validation loss. All reported results are the average performance across the two subtasks on the held-out test set. This evaluation methodology ensures that performance gains are consistent across different imbalance conditions and not confined to a single classification task.

4.5 Results and Discussion

The evaluation of the proposed self-attention fusion approach was carried out in a systematic manner, focusing first on the comparison of different loss functions, followed by an in-depth investigation of the modified attention mechanisms, and finally benchmarking against the ISIC 2017 Challenge leaderboard. This multi stage analysis allows us not only to understand the effect of design choices individually, but also to position the overall contribution of the proposed method in the broader context of skin lesion classification research.

4.5.1 Comparison of Loss Functions

In the first stage of our investigation, we systematically scrutinised the performance of the proposed skin cancer classification model with different loss functions. The results of this comparative study are summarised in Table 4.1, comparing the model performance on classifying

Table 4.1: Performance Comparison of Different Loss Functions

Loss Function	BA	AUC	AP	Acc	F1
CE	0.6975	0.8535	0.5964	0.8467	0.5354
Weighted CE	0.6952	0.8611	0.6103	0.8533	0.5368
Focal Loss	0.5708	0.7868	0.4991	0.8200	0.2603
Class Balanced	0.7166	0.7794	0.4960	0.6950	0.4903

(a) seborrheic keratosis versus (b) a combination of nevus and melanoma.

As shown in Table 4.1, the weighted cross-entropy (CE) loss with $\alpha = 0.8$ achieved superior performance across most metrics, outperforming the standard CE, focal loss, and class-balanced variants. This result underscores the importance of class re-weighting in the context of highly imbalanced medical datasets such as ISIC 2017, where some lesion types occur far less frequently than others. While focal loss was expected to improve the handling of difficult examples, in practice, it led to performance degradation, most notably with a sharp drop in the F1 score. This suggests that in a dataset where the minority class is already extremely underrepresented, aggressively down-weighting easy examples might destabilise learning. Similarly, the class-balanced loss, while achieving the highest balanced accuracy (BA), showed weaker performance on AUC and AP, highlighting a trade-off between fairness across classes and discriminative capability.

These findings align with prior studies in medical imaging, where weighted loss functions have consistently been shown to improve performance on rare disease detection [150]. Therefore, weighted CE was selected as the loss function for subsequent experiments. It is worth noting that our result provides further evidence that properly tuned class weights are often more reliable in practice than more complex imbalance-aware losses.

4.5.2 Attention Mechanism Modification and Visualisation

Having identified the optimal loss function, we next turned to evaluating the effect of modifying the attention mechanism by introducing fusion strategies in the last n layers of the Vision Transformer. To build an intuitive understanding of the improvements, we first visualised the attention maps generated by different attention mechanisms, as shown in Fig. 4.2. Column 1 and 3 illustrate the raw attention maps from different heads, while column 2 and column 4 depict

Table 4.2: Self-attention Fusion Performance on ViT-B16 and DINO-B16

Fusion Strategy	BA	AUC	AP	Acc	F1
ViT-B16					
Base	0.6952	0.8611	0.6103	0.8533	0.5368
Prod. Fusion (Ours)	0.6899	0.8658	0.6277	0.8500	0.5263
Max Fusion (Ours)	0.7005	0.8465	0.6305	0.8567	0.5474
Avg Fusion (Ours)	0.6919	0.8449	0.6167	0.8533	0.5319
DINO-B16					
Base	0.6962	0.8543	0.6180	0.8550	0.5397
Prod. Fusion (Ours)	0.7698	0.8607	0.6448	0.8483	0.6224
Max Fusion (Ours)	0.7768	0.8787	0.6516	0.8700	0.6518
Avg Fusion (Ours)	0.7665	0.8714	0.6439	0.8483	0.6192

their corresponding mask mappings after thresholding. Areas highlighted in red correspond to regions that the model pays more attention to, while darker areas are largely ignored.

These visualisations demonstrate that each attention mechanism tends to highlight different aspects of the lesion, with some mechanisms better capturing brighter regions and others better capturing darker ones. This complementary behaviour motivated our fusion approach, which integrates the outputs of different mechanisms using average, maximum, or product-based operations. By combining them, the fused model benefits from richer and more comprehensive feature representations, thereby reducing the risk of missing subtle diagnostic cues.

4.5.3 Quantitative Results with Fusion Strategies

The performance comparison between baseline and fusion-based models is presented in Table 4.2. Both supervised ViT-B16 and self-supervised DINO-B16 models were evaluated, with and without fusion strategies.

From Table 4.2, it is clear that all fusion strategies led to consistent improvements over the baselines in terms of average precision (AP), which is particularly important in imbalanced datasets where precision/recall trade offs dominate clinical usability. In the ViT-B16 models, maximum fusion delivered the best overall performance with notable gains in F1 score, while in the DINO-B16 models, maximum fusion again outperformed others across all metrics, showing a remarkable 12% improvement in F1 over the baseline. These results indicate that the proposed fusion mechanism is architecture agnostic and synergises especially well with the self-supervised

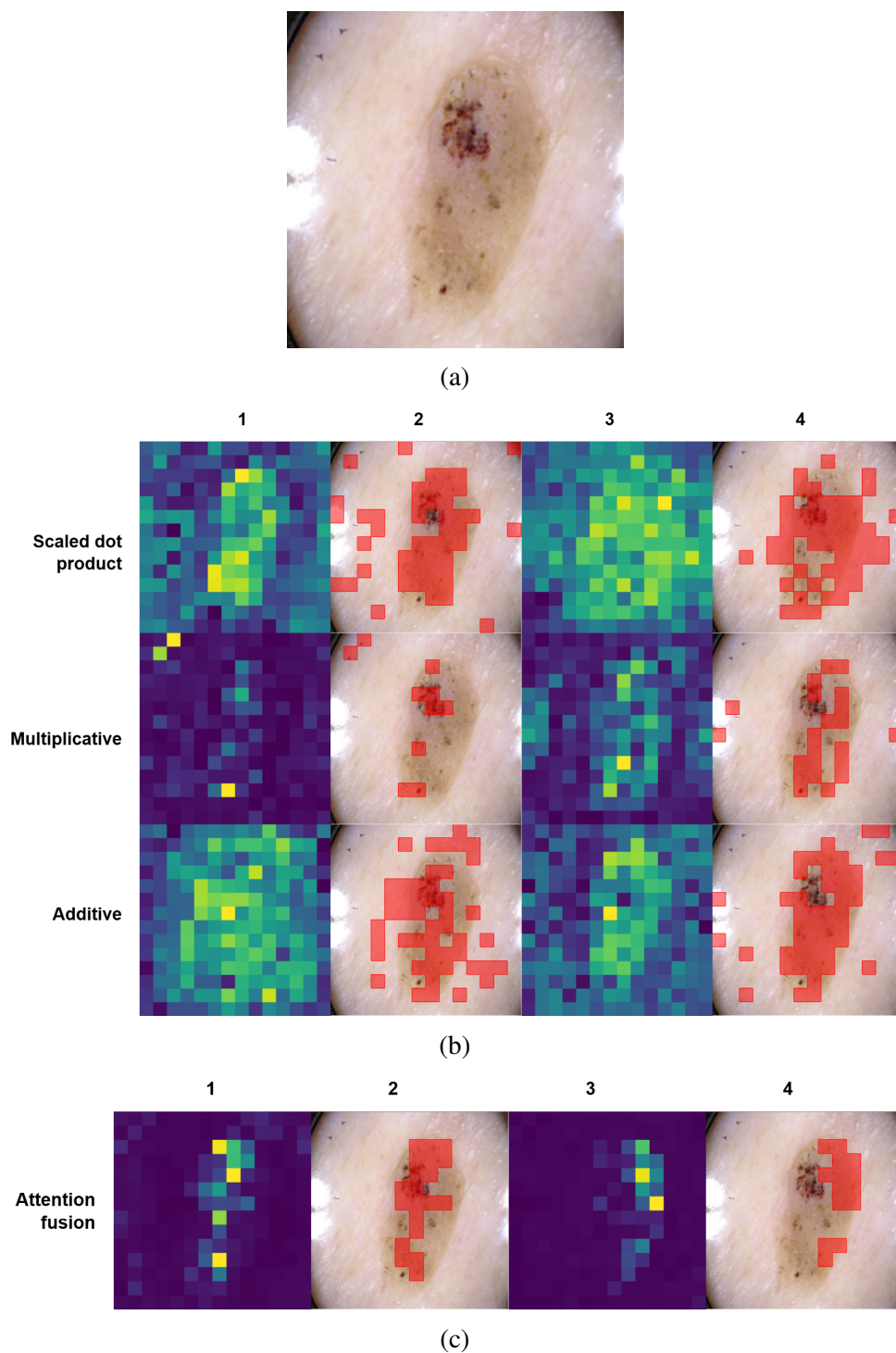


Figure 4.2: (a) Original image. (b) Attention maps from two different heads correspond to a skin lesion with a darker and brighter area. The even columns present the mask mapping with a threshold of 0.3. (c) Attention maps from the proposed method

Table 4.3: McNemar Test on ViT-B16

Contingency Table	Ours - Correct	Ours - Incorrect
Base - Correct	497	15
Base - Incorrect	17	71
p -value: 0.8596		
Fail to Reject Null Hypothesis		

Table 4.4: McNemar Test on DINO-B16

Contingency Table	Ours - Correct	Ours - Incorrect
Base - Correct	476	22
Base - Incorrect	43	59
p -value: 0.0131		
Reject Null Hypothesis		

features of DINO.

4.5.4 Statistical Significance Analysis

To validate the robustness of these improvements, we employed the McNemar test [168], which evaluates whether two models differ significantly in their error rates. The results for ViT-B16 and DINO-B16 are presented in Tables 4.3 and 4.4. While improvements in the ViT case were not statistically significant, the DINO case showed clear superiority with $p < 0.05$, thereby rejecting the null hypothesis.

4.5.5 Comparison with ISIC 2017 Challenge Leaderboard

Finally, to benchmark our proposed model against state-of-the-art methods, we compared our results with those of the ISIC 2017 Challenge leaderboard. Table 4.5 summarises this comparison.

Table 4.5: Comparison with the ISIC 2017 Leaderboard

Model	AUC	BA	AP	Acc	F1
Rank 1	0.911	0.831	0.750	0.816	0.612
Rank 2	0.910	0.883	0.748	0.849	0.242
Rank 3	0.908	0.844	0.754	0.883	0.564
DINO Fusion (Ours)	0.899	0.805	0.695	0.890	0.679
Rank 4	0.896	0.843	0.733	0.888	0.612
ViT Fusion (Ours)	0.896	0.770	0.699	0.871	0.606
Rank 5	0.886	0.847	0.667	0.873	0.608

Although our method ranks 4th in terms of the AUC metric, the primary evaluation criterion in ISIC 2017, it is important to emphasise that our approach used ViT-B16 under memory constraints, whereas the top leaderboard entries leveraged larger backbones and additional segmentation masks. Notably, our model achieved the best accuracy (Acc) and F1 score among all entries, demonstrating a more balanced performance between sensitivity and precision. This balance is of particular clinical importance: a high F1 score implies fewer missed melanomas (high recall) while avoiding unnecessary false alarms (high precision).

4.5.6 Discussion

Taken together, these results highlight three main contributions of our approach: (1) weighted cross-entropy effectively mitigates class imbalance without sacrificing discriminative power; (2) attention fusion enriches the feature space and improves lesion focus, especially when combined with self-supervised features from DINO; and (3) our model generalises competitively against challenge benchmarks, even under constrained resources. The visualisation of attention maps further demonstrates the interpretability of the model, a critical factor in clinical adoption of AI tools.

Nevertheless, some limitations must be acknowledged. The reliance on ViT-B16 was primarily due to memory restrictions. Larger variants may yield further gains. Additionally, while ISIC 2017 is a widely used benchmark, real-world datasets often contain more heterogeneity, including images with lower quality or from underrepresented patient groups. Future work should extend evaluations to larger, more diverse datasets and explore the integration of lesion segmentation to further enhance attention targeting.

Overall, our findings support the effectiveness of self-attention fusion as a promising strategy for improving transformer-based models in medical imaging and demonstrate its potential for deployment in skin cancer screening workflows.

4.6 Conclusion

This thesis introduced a novel self-attention fusion approach for enhancing skin lesion classification using Vision Transformers. The proposed method strategically replaces the last n attention layers with a fusion of diverse attention mechanisms, allowing the model to leverage the strengths of different attention variants while retaining the powerful representations of pre-trained ViTs. By integrating average, maximum, and product-based fusion strategies, the model learns richer and more discriminative features tailored to dermatological image analysis.

Experimental results demonstrated that the weighted cross-entropy loss function provided the best trade-off for addressing severe class imbalance inherent in skin lesion datasets. The proposed fusion-based models consistently outperformed baseline ViTs and achieved particularly strong results when combined with self-supervised DINO pretraining. Notably, the maximum fusion strategy yielded significant gains, including a 12% improvement in F1 score compared to the baseline, highlighting the method's ability to balance precision and recall. Statistical analysis using the McNemar test confirmed that these improvements were significant in the case of DINO-based models.

In conclusion, the proposed self-attention fusion framework represents a simple yet powerful enhancement to Vision Transformer architectures for medical image analysis. It advances both performance and interpretability in skin lesion classification, providing a strong foundation for future research. With further exploration of scalable attention mechanisms, multimodal fusion, and clinical validation, this approach has the potential to contribute meaningfully to the development of trustworthy AI systems in dermatology and beyond.

Federated Colour-Invariant Adversarial Learning for Fair and Robust Skin Lesion Classification

5.1 Introduction

Federated learning (FL) has emerged as a decentralised machine learning paradigm that enables multiple clients, such as hospitals, medical centres, or edge devices, to collaboratively train a shared global model without exchanging raw data [169]–[171]. Instead of transmitting sensitive datasets to a central repository, each client trains a local model on its private data and only communicates model updates or gradients to a central server. The server then aggregates these updates to produce an improved global model. This approach offers two key benefits: (i) it preserves data privacy by preventing the transfer of raw patient information, and (ii) it leverages heterogeneous data sources distributed across multiple institutions, which would otherwise remain siloed due to ethical, legal, or regulatory restrictions.

The healthcare domain has been one of the most prominent beneficiaries of federated learning, given the strict privacy and confidentiality requirements surrounding patient data [53], [172], [173]. In medical imaging tasks, for instance, hospitals and research centres often accumulate valuable datasets but are prohibited from sharing them directly because of regulatory frameworks such as the General Data Protection Regulation (GDPR) in Europe and the Health Insurance Portability and Accountability Act (HIPAA) in the United States [174], [175]. FL provides a viable

solution by allowing these institutions to collaboratively train machine learning models, ensuring both data security and model improvement through exposure to diverse patient populations.

Despite these advantages, federated learning introduces unique challenges compared to conventional centralised learning. A major difficulty lies in the fact that client data is often non-independent and identically distributed (non-IID). In practical terms, this means that the statistical properties of data vary substantially across institutions due to differences in demographic distributions, imaging devices, acquisition protocols, or clinical practices [171], [176], [177]. For example, dermatological image datasets collected in different hospitals may reflect varying skin tone distributions, as categorised by the Fitzpatrick skin type scale [3]. A federated model trained naively on such data might achieve high accuracy on lighter skin tones while performing poorly on darker ones, leading to subgroups fairness concerns and potential health disparities in clinical applications. Addressing these distribution shifts is therefore essential to ensuring that federated learning produces fair, unbiased, and clinically reliable models.

One promising strategy for mitigating distribution-induced bias is to design feature extractors that are invariant to nuisance variables such as skin tone or colour variations. An invariant feature extractor is trained to capture task-relevant information (e.g., lesion morphology) while discarding spurious correlations tied to demographic or environmental factors [178]–[180]. A common approach to achieving this is through adversarial learning, in which a domain classifier attempts to predict the nuisance variable (such as skin type), while the feature extractor is trained adversarially to prevent this prediction from succeeding. Gradient reversal layers (GRL) [181] make this possible by reversing the gradient signal during backpropagation, effectively forcing the feature extractor to produce domain-invariant features. Domain-Adversarial Neural Networks (DANN) represent one of the most well-known frameworks for this approach [181]. However, DANN was originally designed for centralised training environments where all data is pooled in one place, and thus does not naturally extend to federated settings where data is distributed and privacy must be maintained.

Several recent works have sought to incorporate colour-invariant learning into medical imaging tasks. For example, CIRCLE [182] proposed a colour-invariant representation learning method that reduces demographic bias arising from heterogeneous imaging conditions in skin lesion

classification. While effective in centralised settings, CIRCLE assumes access to skin-type labels across all training samples, which is unrealistic in federated contexts where labels are fragmented and direct sharing of sensitive metadata is restricted. This creates a gap between the need for fairness-aware learning in federated settings and the limitations of existing colour-invariant approaches.

In this work, we address this gap by introducing **FedCIAL** (Federated Colour-Invariant Adversarial Learning), a federated framework specifically designed for fair and robust skin lesion classification. The core idea is to exploit known colour distribution shifts, such as those reflected in Fitzpatrick skin types, to guide the adversarial learning process. Rather than relying on external transformers or centralised access to skin-tone annotations, FedCIAL integrates a colour-invariant adversarial module directly into the federated learning pipeline. This module enforces the feature extractor to discard skin-type specific information, thereby producing representations that are both privacy-preserving and fair across diverse client populations. By aligning feature spaces across decentralised clients, FedCIAL ensures that classification performance is not disproportionately skewed toward particular subgroups.

Skin lesion classification is a particularly critical application for this work, as skin cancer remains one of the most prevalent cancers worldwide, and its early detection is crucial for improving survival outcomes [183]–[185]. However, existing diagnostic systems risk exacerbating health inequalities if they are biased toward specific skin tones or imaging conditions. By embedding subgroups fairness into the federated training process, FedCIAL directly addresses this concern. The key contributions of this work can be summarised as follows:

- We propose a novel federated adversarial learning framework that achieves colour-invariant feature extraction without requiring centralised access to sensitive demographic labels.
- We demonstrate that FedCIAL mitigates demographic fairness concerns by reducing performance disparities across different Fitzpatrick skin types while preserving overall classification accuracy.
- We validate the effectiveness of the proposed method through comprehensive experiments on real-world dermatological datasets, showing improvements in both performance and

subgroups fairness over existing baselines.

In doing so, FedCIAL advances the state of fairness-aware federated learning for medical image analysis, paving the way toward more equitable and clinically reliable AI systems.

5.2 Related Work

The proposed work builds upon three major areas of research: (i) the study of feature skew and data heterogeneity in federated learning, (ii) bias mitigation strategies in machine learning, particularly in medical imaging, and (iii) federated frameworks designed to enhance representation learning in privacy-preserving settings. This section reviews the most relevant contributions from these areas, highlighting their relevance to fairness aware federated skin lesion classification.

5.2.1 Feature Skew in Federated Learning

A central challenge in federated learning arises from non-IID data distributions, which manifest in several forms, including feature skew, label skew, and quantity skew. Among these, feature skew is particularly problematic for healthcare applications because the distribution of input features often varies significantly across clients due to demographic, environmental, or device-related factors. Feature skew is a case of covariate shift [186], which occurs when the marginal distribution of inputs differs across clients while the conditional distribution of outputs given the inputs remains consistent [187], [188]. In practice, this means that although two hospitals may label data using the same clinical definitions, the underlying appearance of data samples (e.g., imaging modalities, skin tones, acquisition devices) may differ drastically. Such heterogeneity leads to generalisation difficulties and a higher error rate on unseen or underrepresented distributions [189].

Several studies have demonstrated the adverse effects of feature skew in federated medical imaging. Liu *et al.* [190] developed a federated prostate segmentation model using heterogeneous MRI datasets collected from multiple clinical sites. Variability in MRI acquisition protocols across sites produced strong domain discrepancies, reducing model generalisation. Similarly, Li *et al.* [186] investigated functional brain imaging classification using data from the ABIDE I study, which pooled datasets from four medical institutions. They observed substantial heterogeneity

due to diverse scanning devices and protocols, resulting in models that failed to generalise across sites.

In the context of dermatology, feature skew becomes particularly acute when considering variations in Fitzpatrick skin types. Xu *et al.* [35] analysed federated classification of skin lesion images across clients with different distributions of skin tones. They found that models performed well on skin types represented in local training data but generalised poorly to underrepresented skin tones, confirming that feature skew can embed demographic biases into federated models. Such findings are consistent with broader fairness concerns raised in dermatology AI research [3], where underrepresentation of darker skin tones often leads to disparate model performance. These observations underscore the importance of explicitly addressing feature skew in federated healthcare, not only to improve robustness but also to promote equitable outcomes across diverse patient populations.

5.2.2 Bias Mitigation in Machine Learning

Feature skew directly contributes to demographic and class imbalance bias in machine learning models. In medical imaging, bias mitigation has become a crucial research direction as biased models can exacerbate existing health disparities. For skin lesion classification in particular, bias often arises from disproportionate reliance on skin colour features. Models trained predominantly on lighter skin tones may misclassify or underperform on images from individuals with darker skin tones, where lesion appearance is more subtle or confounded by pigmentation. Studies have shown that darker skin tones are significantly underrepresented in dermatological datasets, leading to systematic disparities in performance [3], [191]. Addressing this imbalance requires methods that explicitly reduce reliance on skin tone as a predictive factor.

Domain adversarial approaches represent one influential line of work in this area. Domain-Adversarial Neural Networks (DANN) [181] introduced the concept of using a gradient reversal layer to enforce domain invariance in learned representations. The idea is to jointly train a feature extractor with two competing objectives: one to minimise task loss (e.g., lesion classification) and another to maximise the error of a domain classifier attempting to predict nuisance variables such as skin tone. By reversing the gradient during backpropagation, the feature extractor learns

to suppress domain-specific cues, yielding more invariant and generalizable representations. Although highly effective in centralised scenarios where pooled data is available, DANN’s direct applicability to federated settings is limited, as federated environments require strict data privacy and often lack access to global demographic annotations.

To address colour-specific biases, CIRCLe [182] introduced a colour-invariant representation learning framework for centralised skin lesion classification. CIRCLe trains a skin colour transformer that explicitly normalises the influence of colour variations, thereby mitigating performance disparities linked to heterogeneous imaging conditions. However, this method assumes centralised access to skin type labels across all training data, an assumption incompatible with federated learning, where skin tone metadata cannot be freely shared between institutions. Consequently, while effective in reducing bias under centralised training, CIRCLe falls short of addressing subgroups fairness in decentralised and privacy-sensitive federated contexts.

5.2.3 Federated Representation Learning

Recent research has explored federated frameworks that integrate advanced representation learning strategies to better handle data heterogeneity. One notable example is FeSViBS [128], which combines federated learning with split learning to balance privacy preservation and computational efficiency. FeSViBS introduces a block sampling mechanism that selects intermediate features from Vision Transformer (ViT) [114] layers, compressing them into pseudo class tokens. These tokens, which capture essential local information while discarding raw data, are transmitted to the server for aggregation. By leveraging intermediate features rather than final layer embeddings, FeSViBS enhances generalisation across clients and mitigates information loss that often arises in heterogeneous federated environments. Furthermore, this strategy ensures that valuable cross-client knowledge is shared without compromising data privacy.

In the present work, we build on the FeSViBS framework as a baseline and extend it with a novel colour invariant adversarial learning module. Our method introduces federated colour invariant adversarial learning (FedCIAL), which integrates adversarial feature alignment directly into the federated pipeline to mitigate bias associated with Fitzpatrick skin types. Unlike prior centralised approaches such as DANN or CIRCLe, FedCIAL does not require centralised access

to demographic metadata. Instead, it leverages the known colour distribution shifts as implicit domain signals within a federated adversarial learning strategy, enabling the model to produce fairer and more robust representations across decentralised, heterogeneous datasets.

5.3 Method

This chapter presents the methodology behind the proposed Federated Colour-Invariant Adversarial Learning (FedCIAL) framework. The section begins with preliminaries on federated learning and the challenges of feature skew, before describing the design of the FedCIAL model. The architecture integrates domain adversarial techniques with colour-invariant representation learning, specifically addressing the problem of bias induced by skin tone variations in federated skin lesion classification.

5.3.1 Preliminaries

Federated Learning

Federated learning (FL) is a collaborative paradigm for training machine learning models in a decentralised fashion, where sensitive data remains stored locally on client devices. In the FL setting, there are N clients denoted by $C = \{C_1, C_2, \dots, C_N\}$, and a central server coordinates the training process. Each client C_k , where $k \in [1, N]$, trains a local model w_k using its private dataset D_k . The local optimisation is performed using a loss function $\mathcal{L}_k$, formally expressed as:

$$\mathcal{L}_k(w) = \frac{1}{|D_k|} \sum_{i \in D_k} \ell(w; x_i, y_i), \quad (5.3.1)$$

where (x_i, y_i) is a training sample and $\ell(\cdot)$ is typically a cross-entropy loss. At the end of each communication round, clients send their updated local models to the server, which aggregates them to form a new global model. The global objective function is defined as:

$$\min_w \mathcal{L}(w) = \sum_{k=1}^N \alpha_k \mathcal{L}_k(w), \quad (5.3.2)$$

where the aggregation weight α_k is proportional to the client's dataset size:

$$\alpha_k = \frac{|D_k|}{\sum_{j=1}^N |D_j|}. \quad (5.3.3)$$

This weighted averaging process, commonly known as FedAvg [169], enables global knowledge integration while maintaining data privacy, since raw data never leaves local devices.

Feature Skew in Federated Learning

One of the primary challenges in federated learning is heterogeneity across clients, particularly in the form of feature skew. Feature skew occurs when the marginal distributions of input features differ between clients, formally described as $P_i(x) \neq P_j(x)$ for clients C_i and C_j , even if the conditional label distribution remains similar, i.e., $P_i(y|x) \approx P_j(y|x)$.

In healthcare, feature skew frequently arises due to demographic differences, imaging protocols, or device-specific acquisition settings. For dermatological images, a major source of feature skew stems from Fitzpatrick skin type variations across patient populations. Such discrepancies degrade the generalisation performance of the aggregated global model, as the model struggles to align feature spaces across heterogeneous domains. Therefore, mitigating feature skew is crucial to ensure subgroups fairness and robustness in federated medical AI systems.

5.3.2 Federated Colour-Invariant Adversarial Learning

The proposed FedCIAL framework addresses the demographic fairness challenges of federated learning by introducing colour-invariant adversarial learning. Unlike prior work such as CIRCLE [182], which relies on centralised access to skin-type labels to train a colour transformer, FedCIAL directly integrates prior knowledge of Fitzpatrick colour distributions into a federated adversarial learning pipeline. This is achieved by extending the DANN approach [181] to a decentralised setting, where a gradient reversal layer (GRL) encourages the feature extractor to learn representations invariant to skin tone while still discriminative for the classification task.

An overview of the FedCIAL framework is shown in Figure 6.1, while the federated training process and local adversarial updates are detailed in Algorithm 1 and Algorithm 2.

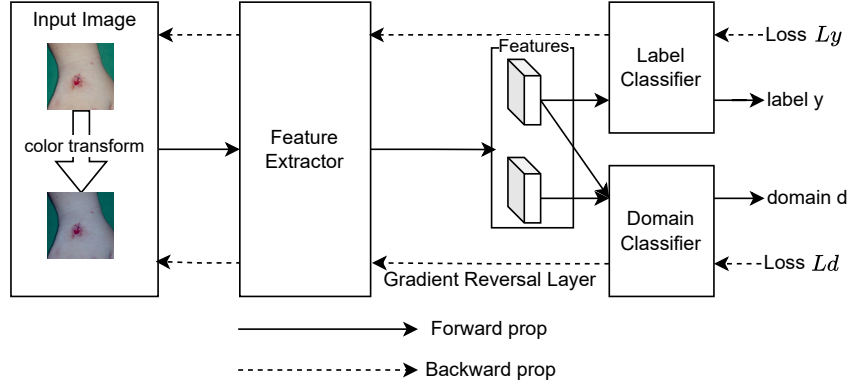


Figure 5.1: Overview of the proposed FedCIAL framework. The architecture integrates federated training with colour-invariant adversarial learning, mitigating skin tone bias through gradient reversal and explicit Fitzpatrick-based colour transformations.

Algorithm 1 Main FL Training Loop

Require: Number of clients N , communication rounds T

- 1: Initialise global model weights w_0
- 2: **for** each round $t = 1, 2, \dots, T$ **do**
- 3: Server selects a subset of clients $\mathcal{S}_t \subseteq \{1, \dots, N\}$
- 4: **for** each client $k \in \mathcal{S}_t$ **in parallel do**
- 5: Client receives global model w_t
- 6: Client performs **Colour-Invariant Adversarial Learning** local updates with w_t
- 7: Send updated model w_{t+1}^k to the server
- 8: **end for**
- 9: Server aggregates models:

$$w_{t+1} = \sum_{k \in \mathcal{S}_t} \frac{n_k}{\sum_{j \in \mathcal{S}_t} n_j} w_{t+1}^k$$

10: **end for**

Colour-Invariant Representation Learning

Colour-invariant representation learning forms the core of FedCIAL. The aim is to ensure that the learned features are unaffected by systematic colour variations in skin lesion images, which often arise due to differences in Fitzpatrick skin types or acquisition conditions. Unlike abstract sources of bias, colour distributions can be explicitly quantified and manipulated. FedCIAL leverages this property by incorporating known Lab space colour statistics associated with Fitzpatrick types during training.

The transformation process converts RGB inputs into the Lab colour space, separating luminance (L) from chromatic components (a, b). Lab is chosen because it closely aligns

Algorithm 2 Colour-Invariant Adversarial Learning**Require:** Model weights w_t , local epochs E , learning rate η , gradient reversal coefficient λ

- 1: **for** each local epoch $e = 1, \dots, E$ **do**
- 2: **for** each mini-batch $\{x_i, y_i, d_i\}$ from local dataset D_k **do**
- 3: Sample a random target domain label $d'_i \neq d_i$
- 4: Generate transformed input $x'_i = \text{Transform}(x_i, d'_i)$
- 5: Extract features: $f_i = F(x_i)$, $f'_i = F(x'_i)$
- 6: Predict class: $\hat{y}_i = C(f_i)$
- 7: Predict domains: $\hat{d}_i = D(f_i)$, $\hat{d}'_i = D(f'_i)$
- 8: Compute losses:

$$\mathcal{L}_{\text{cls}} = \ell(\hat{y}_i, y_i)$$

$$\mathcal{L}_{\text{dom}} = \ell(\hat{d}_i, d_i) + \ell(\hat{d}'_i, d'_i)$$

- 9: Total loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{dom}}$$

- 10: Backpropagate $\mathcal{L}_{\text{cls}}$ normally from classifier C to feature extractor F
- 11: Backpropagate $\mathcal{L}_{\text{dom}}$ via GRL, multiplying gradient by $-\lambda$
- 12: **end for**
- 13: **end for**

with human visual perception [192]. Images are then shifted toward predefined Lab values corresponding to Fitzpatrick skin types, as shown in Table 5.1. These values are empirically derived based on ITA (Individual Typology Angle) ranges [3], providing a principled mapping between pigmentation levels and perceptual colour components.

Table 5.1: Predefined Lab colour values for Fitzpatrick skin types (FST). Values follow established ITA mappings, ensuring perceptual consistency between pigmentation levels and Lab components.

FST	L	a	b
1	85	4	12
2	78	6	18
3	69	9	25
4	62	14	33
5	55	19	42
6	48	23	51

By systematically transforming inputs across Fitzpatrick distributions, the model is exposed to synthetic variations of each image, encouraging the feature extractor to ignore skin tone information and instead focus on lesion-specific cues.

Integration of Domain-Adversarial Techniques

The adversarial component of FedCIAL is inspired by Domain-Adversarial Neural Networks (DANN) [181], which employ a Gradient Reversal Layer (GRL) to enforce invariance. In FedCIAL, the GRL sits between the feature extractor and the domain classifier. During forward passes, it behaves as the identity function, while during backpropagation, it multiplies domain gradients by $-\lambda$. This process forces the feature extractor to discard colour-specific features, producing domain-agnostic representations.

To further strengthen adaptation, FedCIAL introduces transformed images (aligned with target Fitzpatrick Lab values) as explicit auxiliary samples. Unlike classical DANN, which assumes shared target domains, these transformed samples serve as consistent adversarial targets in a federated context. To avoid destabilising training, we schedule the domain adaptation coefficient λ using a sigmoid-based schedule:

$$\lambda = \frac{2}{1 + \exp\left(-10 \times \frac{r}{\text{epochs}}\right)} - 1,$$

where r denotes the current training round. This ensures that adversarial influence starts weak and grows stronger over training, gradually enforcing invariance.

Finally, to simplify domain classification and reflect melanin-related grouping, Fitzpatrick types 1-3 (lighter tones) are grouped as domain 0, while types 4-6 (darker tones) are grouped as domain 1. This binary split allows effective adversarial alignment across broad skin tone categories without requiring fine-grained centralised annotations.

5.4 Experimental Setup

This section presents the experimental evaluation of the proposed Federated Colour-Invariant Adversarial Learning (FedCIAL) framework. We begin by introducing the dataset and partitioning strategy, followed by a description of the baseline and comparative models used in our study. Finally, we outline the implementation details, including training parameters, optimisation strategy, and evaluation protocol, which ensure a fair and reproducible comparison across all

methods.

5.4.1 Dataset

The experiments are conducted on the Fitzpatrick17k dataset [3], which is one of the most comprehensive dermatological datasets annotated with Fitzpatrick skin type (FST). The dataset contains approximately 16,018 clinical images, with each image labelled according to one of six FST categories. This annotation provides a valuable foundation for studying demographic fairness in medical image classification, since skin tone is a major source of bias in dermatological diagnosis. In addition to skin tone information, the dataset covers a diverse range of lesion categories, broadly grouped into benign, malignant, and non-neoplastic conditions, ensuring coverage of clinically relevant cases. Table 5.2 presents the distribution of samples across Fitzpatrick skin types after excluding the unknown skin type category.

Table 5.2: Fitzpatrick17k dataset breakdown by Fitzpatrick skin type

FST	Non-Neoplastic	Benign	Malignant	Total
Type I	2054	445	457	2956
Type II	3394	670	742	4806
Type III	2380	474	457	3311
Type IV	2114	366	301	2781
Type V	1220	159	147	1526
Type VI	532	45	61	638
Total	11694	2159	2165	16018

To simulate the non-IID nature of federated learning, we follow the partitioning strategy of Xu *et al.* [35]. Specifically, the dataset is divided across six clients, where each client exclusively contains images from a single Fitzpatrick skin type. This partitioning reflects a realistic healthcare scenario in which medical institutions may serve patient populations with uneven distributions of skin tones. Such a setup naturally induces feature skew and provides an appropriate benchmark for testing fairness-oriented federated learning methods.

Before model training, we apply a series of preprocessing and augmentation steps to improve robustness and mitigate overfitting. Images are resized to a consistent resolution and normalised using ImageNet statistics [193], ensuring compatibility with pretrained Vision Transformer backbones. In addition, standard geometric augmentations are applied, including random

rotations, flips, and shear transformations. These augmentations increase dataset diversity and help the model generalise better across heterogeneous data distributions.

5.4.2 Baseline and Comparative Methods

For the backbone architecture, we adopt the Federated Split Learning of Vision Transformer with Block Sampling (FeSViBS) framework proposed by Almalik *et al.* [128]. FeSViBS combines federated split learning with Vision Transformers (ViTs), and introduces block sampling to reduce communication and computational costs. The use of ViTs allows the model to capture long-range dependencies in images, which is particularly beneficial in medical imaging tasks.

We consider FeSViBS as the baseline model and progressively extend it to evaluate the impact of different bias mitigation strategies. Specifically, we compare the following methods:

1. **Baseline (FeSViBS only):** The original FeSViBS model trained in a federated setting without explicit bias mitigation.
2. **Colour-invariant representation learning:** An extension of the baseline where the feature extractor is trained using regularisation with colour augmented samples, but without adversarial adaptation.
3. **Adversarial learning without target samples:** The baseline integrated with adversarial training through gradient reversal layers, but without using transformed target samples.
4. **FedCIAL (proposed):** The full framework, which integrates colour-invariant adversarial learning by explicitly incorporating Fitzpatrick-based transformations as target samples during adversarial training.

This comparative design allows us to disentangle the contributions of adversarial learning and explicit colour-invariant training, while also demonstrating the cumulative effect of their integration in FedCIAL.

5.4.3 Implementation Details

All experiments are implemented using the PyTorch framework. For both lesion classification and domain classification, we use cross-entropy loss. The batch size is fixed at 32 for all clients. Optimisation is performed using the Adam optimiser [194] with a learning rate of 1×10^{-4} .

Training is conducted over 100 communication rounds, where each round corresponds to one local epoch per client, followed by model aggregation on the central server. To ensure robust evaluation, we employ 5-fold cross-validation, partitioning the dataset into five folds and rotating the test set across runs. Results are reported as averages over the five folds, with standard deviation included to quantify variability across splits.

5.5 Results and Discussion

This section presents the experimental results obtained using the proposed FedCIAL framework, followed by an analysis of their implications. We report performance improvements over the baseline, evaluate statistical significance, and discuss the effectiveness of FedCIAL in enhancing demographic fairness and robustness across diverse skin types in federated learning settings.

5.5.1 Classification Performance

The average balanced accuracy across all experiments is reported in Table 5.3. Balanced accuracy is chosen as the primary metric since it accounts for class imbalance across lesion types and provides a fairer measure of performance compared to standard accuracy.

Table 5.3: Average balanced accuracy and standard deviation across 5-fold cross-validation.

Method	Bal. Acc.
Baseline (FeSViBS)	0.7666 ± 0.0054
Colour-invariant rep. learning	0.7682 ± 0.0122
Adversarial learning w/o target samples	0.7686 ± 0.0056
FedCIAL (proposed)	0.7754 ± 0.0098

The results clearly show that FedCIAL outperforms the baseline and other comparative methods. FedCIAL achieves a balanced accuracy of 0.7754, representing a statistically significant

improvement over the baseline (0.7666). To confirm this, we conducted a paired t-test between the baseline and FedCIAL results. The Shapiro-Wilk test confirmed normality of the distribution ($p = 0.522$), validating the use of parametric tests. The paired t-test yielded a p-value of 0.044, confirming that the observed improvement is statistically significant at the 5% level.

We also benchmarked FedCIAL against the centralised setting reported by Groh *et al.* [3], who achieved 62.4% accuracy on Fitzpatrick17k. Importantly, FedCIAL surpasses this centralised benchmark despite operating under federated and privacy-preserving constraints. This highlights the potential of our approach to deliver robust and fair classification performance in realistic decentralised healthcare environments.

5.5.2 Fairness Across Clients

Subgroups fairness in federated learning is measured by consistency of performance across clients. A model that achieves high accuracy but performs poorly for certain skin tones would be unsuitable for deployment in clinical practice. FedCIAL improves demographic fairness by reducing the variability of client performance. Specifically, the average standard deviation of client-level balanced accuracy under FedCIAL is 0.044, compared to 0.053 for the baseline, indicating more equitable performance across skin tone groups.

5.5.3 Feature Alignment Analysis

To further investigate FedCIAL’s effectiveness, we measure the feature-level distance between original images, their colour-transformed variants, and anchor images. Figure 5.2 illustrates examples of these transformations, while Table 5.4 reports quantitative distances using Manhattan (L1), Euclidean (L2), and cosine similarity metrics.

The results show that FedCIAL consistently reduces distances between original and transformed images, particularly for metric c , which measures the distance between original and colour-transformed inputs. This suggests that FedCIAL successfully aligns feature representations across different Fitzpatrick skin tones, ensuring that skin colour variations do not disproportionately affect the model’s predictions. Such feature alignment is a key step toward fair and bias-resistant medical AI systems.

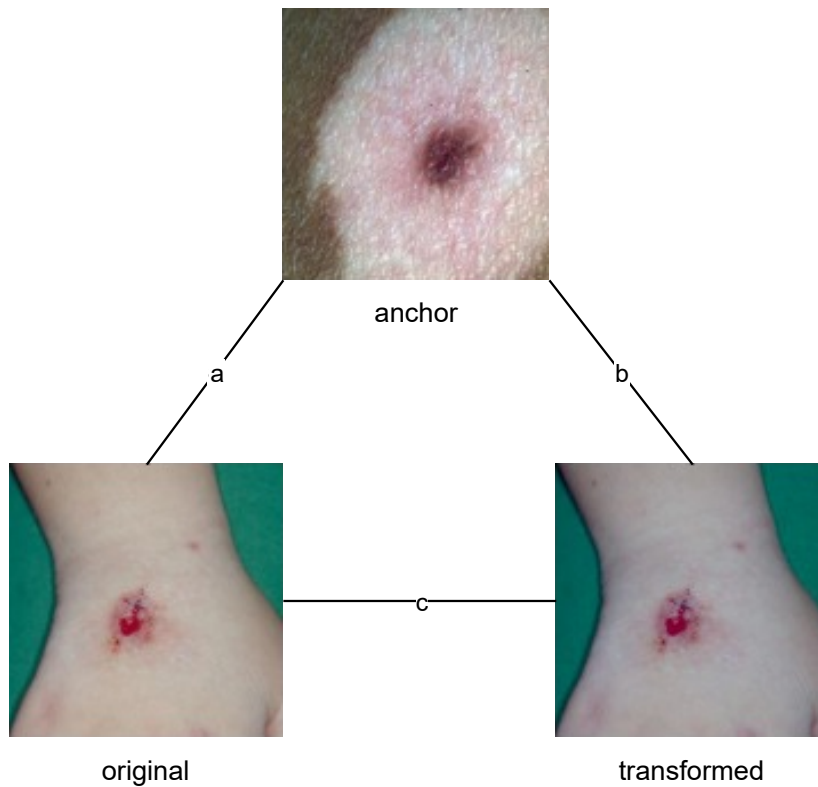


Figure 5.2: Comparison between original, transformed, and anchor images.

5.6 Conclusion

In this work, we introduced FedCIAL, a novel Federated Colour-Invariant Adversarial Learning framework designed to address demographic bias and performance disparities in skin lesion classification under federated learning settings. Unlike traditional approaches that rely on centralised skin type labels or colour transformers, FedCIAL leverages explicit knowledge of Fitzpatrick skin type distribution shifts and integrates them into an adversarial learning framework. By combining colour-invariant representation learning with a gradient reversal layer, our method effectively mitigates colour-induced biases while preserving the privacy-preserving advantages of federated learning.

Extensive experiments on the Fitzpatrick17k dataset demonstrate that FedCIAL achieves superior performance compared to the state-of-the-art FeSViBS baseline, both in terms of higher balanced accuracy and reduced performance variance across clients. This not only

Table 5.4: Distance between original, transformed, and anchor images across different metrics.

Metric	a	b	c
Baseline			
Manhattan (L1)	24.7	27.7	11.8
Euclidean (L2)	1.1	1.3	0.5
Cosine Distance	0.11	0.14	0.03
FedCIAL			
Manhattan (L1)	22.3	24.6	9.8
Euclidean (L2)	1.0	1.1	0.4
Cosine Distance	0.09	0.11	0.01

highlights the robustness of the proposed method but also its ability to promote demographic fairness across diverse populations with varying skin tones. Furthermore, the reduction in feature distance between original and transformed samples indicates that FedCIAL successfully aligns representations, contributing to more equitable diagnostic outcomes in dermatological applications.

Looking ahead, FedCIAL offers a promising foundation for future research on fairness-aware federated learning. Potential directions include extending the framework to multimodal clinical data, exploring alternative adversarial strategies beyond gradient reversal, and testing scalability across larger federated networks in real-world healthcare systems. By continuing to integrate subgroups fairness considerations into the design of federated learning models, this line of research can advance equitable and trustworthy AI solutions for medical imaging and beyond.

Fairness-Aware Federated Learning with Domain Adaptation and Class Imbalance Mitigation for Skin Lesion Classification

6.1 Introduction

Skin lesion classification is a critical task in dermatology, where early and accurate detection can significantly improve patient outcomes [185]. However, deep learning models often exhibit performance disparities across different skin tones, primarily due to demographic underrepresentation and colour distribution differences associated with darker skin types [3]. The Fitzpatrick Skin Type (FST) classification system [101] provides a standardised means of categorising skin tone, making it an effective framework for studying and mitigating demographic bias and subgroup performance disparities in dermatological AI systems.

Federated Learning (FL) offers a privacy-preserving solution for collaborative model training across distributed datasets [195]. The FeSViBS (Federated Split Vision Blocks) framework [128] provides an efficient foundation for federated medical image classification using split learning architectures. Yet, when applied to dermatological datasets, it faces two persistent challenges: (1) class imbalance bias, where certain diagnostic categories are underrepresented and models become biased toward majority classes, and (2) demographic bias, where model performance

varies systematically across FST groups. These issues are particularly pronounced in federated settings, where data heterogeneity and limited client samples can further amplify subgroup performance disparities and minority-class underperformance.

FedCIAL (Federated Colour-Invariant Adversarial Learning) [196] introduces a fairness-driven domain adaptation mechanism that aims to achieve colour-invariant feature learning across different skin tones. It leverages inter-group colour distribution differences in the LAB colour space to generate synthetic domain-shifted images and employs adversarial domain adaptation to encourage the model to learn representations that are invariant to skin tone variations, all without requiring direct data sharing between clients. The LAB colour space is a device-independent, perceptually uniform model defined by three components L^* (lightness), a^* (green-red axis), and b^* (blue-yellow axis), allowing adjustments in lightness without altering hue or saturation and supporting accurate colour correction and conversion across devices [197]. While prior work validated its feasibility, open questions remain regarding how its domain adaptation component interacts with imbalance mitigation strategies for handling class imbalance, and how these elements collectively influence subgroups fairness outcomes.

To address these gaps, this work proposes a fairness-aware federated learning framework that extends our prior approaches by jointly integrating domain adaptation and class imbalance handling to enhance equitable performance across diverse skin tones while rigorously assessing subgroups fairness in federated dermatological classification. The proposed experimental framework is organised into three stages. First, a single-fold comparison is performed as a preliminary experiment to compare baseline and fairness-enhanced configurations, in order to assess the initial impact of these interventions. Second, a cross-validation analysis is conducted using a 5-fold setup to evaluate the robustness and statistical significance of the observed improvements. Paired t-tests are applied across folds to ensure consistent performance trends. Finally, an ablation study employs a controlled 2x2 experimental design to isolate the effects of domain adaptation and imbalance mitigation strategies by systematically adding or removing each component within the proposed framework. Additional variants incorporating alternative loss functions and sampling strategies are also explored to identify the optimal configuration.

This work makes the following key contributions:

- We propose a fairness-aware federated learning framework that jointly integrates domain adaptation and class imbalance mitigation to enhance equitable performance across Fitzpatrick skin tone groups.
- We propose a three-stage evaluation pipeline to disentangle and quantify the effects of domain adaptation and imbalance mitigation: single-fold comparison, 5-fold statistical validation, and controlled ablation analysis.
- We introduce a systematic evaluation protocol for federated dermatological classification that incorporates multi-dimensional fairness metrics, including worst-group balanced accuracy, fairness score, and fairness gap.

6.2 Related Work

6.2.1 Fairness in Dermatological AI

Fairness has emerged as a critical concern in dermatological AI, as several studies have demonstrated systematic disparities in model performance across skin tones and demographic groups. Groh *et al.* [3] highlighted that dermatological classifiers trained predominantly on lighter skin types often underperform on darker tones, primarily due to the underrepresentation of these groups in benchmark datasets. Subsequent works [36], [198] have emphasised dataset diversification and the evaluation of demographic fairness through subgroup accuracy and calibration analyses. However, these approaches largely assume access to centralised datasets, which is often impractical in medical contexts due to privacy regulations and institutional data silos. Recent efforts have begun exploring algorithmic fairness interventions, such as domain adaptation [199] and invariant feature learning [200], to mitigate colour and demographic biases in skin lesion classification. Yet, these solutions remain underexplored in distributed learning paradigms, where client-specific data heterogeneity exacerbates subgroups fairness challenges.

6.2.2 Federated Learning for Medical Imaging

FL has gained traction in medical imaging as a paradigm that enables multi-institutional collaboration without direct data sharing [176], [201]. Applications span radiology, pathology, and dermatology, demonstrating improvements in model generalizability and data privacy preservation. Despite these advantages, FL inherits the underlying biases present in local datasets and can even amplify them due to non-independent and identically distributed (non-IID) data across clients [88]. Several fairness-aware FL frameworks have been proposed to address this, such as FedFair [89] and AFL [90], which uses mixture of anonymised client data to which allows most centralised debiasing methods to be used. However, these methods typically assume uniform data distributions across clients and focus on client-level fairness, rather than addressing subgroup disparities such as those arising from skin tone variation within each client. Moreover, privacy-preserving techniques like differential privacy [202] and secure aggregation [203] can further limit the visibility of subgroup statistics, complicating fairness auditing in federated setups. This underscores the need for frameworks that integrate subgroups fairness mechanisms into the learning process itself.

6.2.3 Domain Adaptation in Medical Imaging and Fairness

Domain adaptation (DA) addresses distributional shifts between training and testing data, a common issue in medical imaging due to variations in devices, acquisition settings, or patient demographics [98], [204]. Deep learning-based DA methods, such as adversarial domain adaptation [205], promote domain-invariant feature learning by aligning latent representations across domains. These techniques have improved generalisation, but they typically rely on centralised data access, which limits their use in privacy-sensitive contexts.

Federated domain adaptation (FDA) extends DA to decentralised settings by aligning client-specific feature spaces without sharing raw data [91]. However, the FDA focus on improving generalisation rather than subgroups fairness, and may inadvertently amplify biases when domain shifts correspond to sensitive attributes such as skin tone. To address this, fairness-aware

domain adaptation seeks to remove correlations between demographic factors and task-relevant features. FedCIAL [196] builds on this idea by introducing colour-invariant adversarial learning in a federated framework, aligning LAB-space representations to mitigate skin tone bias. Yet, the interaction between such domain adaptation mechanisms and fairness-oriented imbalance mitigation remains underexplored, which is the gap this study aims to address.

6.3 Method

We propose a fairness-aware federated learning framework that extends existing approaches by jointly integrating domain adaptation and class imbalance mitigation to enhance equitable performance across skin tones. Specifically, we (i) implement colour-invariant domain adaptation within a federated setting with multiple strategies for class imbalance, (ii) evaluate its subgroups fairness behaviour across Fitzpatrick groups on the DDI dataset [36], and (iii) run a controlled 2x2 ablation to isolate the effects of domain adaptation versus imbalance mitigation strategies.

6.3.1 Fairness-aware Federated Learning

We extend the FeSViBS framework [128] to enable fairness-aware federated learning in dermatological image classification. FeSViBS used a split-learning-based architecture that combines the efficiency of Vision Transformers (ViTs) with the privacy benefits of federated learning. In this framework, each client hosts a local feature extractor (comprising the early ViT layers and a ResNet50 stem), while the server maintains a shared classification head (later transformer blocks and the classifier). During training, clients forward intermediate representations rather than raw images, allowing the server to complete the forward and backward passes. This design significantly reduces communication overhead and preserves data privacy.

While FeSViBS effectively addresses scalability and privacy, it does not explicitly account for subgroups fairness across demographic subgroups, particularly in dermatological datasets where skin tone variations introduce domain shifts. To address this limitation, we follow the colour-invariant adversarial learning strategy introduced in [196] by leveraging known colour distribution shifts between FST groups. For each client representing the FST group i , we estimate

its colour distribution parameters:

$$\theta_i = (\mu_{L_i}, \sigma_{L_i}, \mu_{A_i}, \sigma_{A_i}, \mu_{B_i}, \sigma_{B_i}).$$

During training, we generate synthetic samples from target FST groups by applying colour transformations:

$$x_t = T(\theta_s \rightarrow \theta_t, x_s),$$

which enables domain adaptation without requiring actual data from other clients.

We then train a domain classifier to distinguish between source and target-domain samples, while simultaneously training the feature extractor to fool the domain classifier. This adversarial interplay encourages the model to learn colour-invariant feature representations with the following loss functions:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{adv}, \quad (6.3.1)$$

$$\mathcal{L}_{adv} = \mathcal{L}_{dom}(D(f(x_s)), 0) + \mathcal{L}_{dom}(D(f(x_t)), 1), \quad (6.3.2)$$

where f is the feature extractor, D is the domain classifier, and λ controls the trade-off between classification and adversarial alignment. This approach encourages feature extractors to be invariant to skin-tone differences, thereby improving fairness across FST groups without requiring data sharing.

In contrast to the original formulation [196], we extend it with multiple imbalance-handling strategies (i.e., weighted loss, weighted sampling, focal loss, and SMOTE) and evaluate its effect on both overall accuracy and subgroup fairness. We apply our method in a realistic federated partition of DDI, where clients correspond to different Fitzpatrick Skin Types, enabling direct measurement of fairness across skin tone domains under federated constraints. This constitutes a new evaluation setting and allows us to attribute subgroups fairness gains to specific components. The flow diagram of the proposed framework is presented in [Figure 6.1](#).

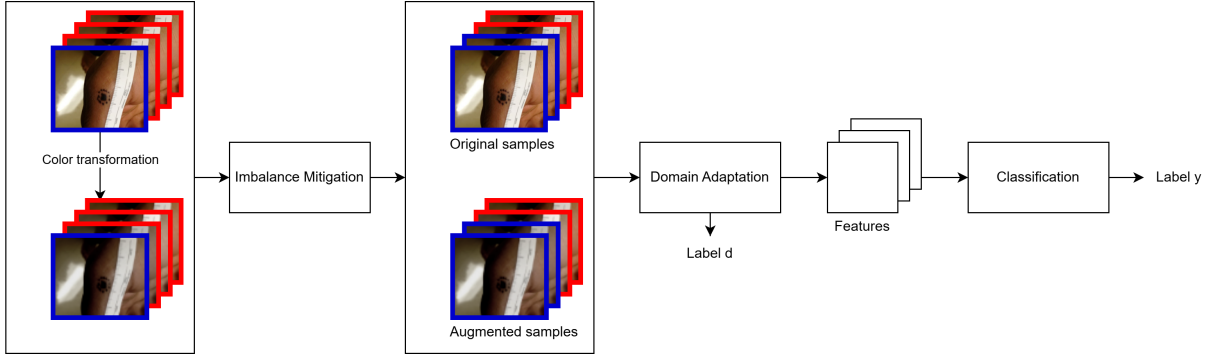


Figure 6.1: Flow diagram of the fairness-aware federated learning. Each training batch first undergoes imbalance mitigation, followed by synthetic sample generation. Both original and synthetic samples are then processed through domain adaptation-based feature extraction to achieve colour-invariant representations, and finally, the original image is passed to the classification layer for prediction.

6.4 Experimental Setup

6.4.1 Dataset and Setup

We evaluate our framework on the Diverse Dermatology Images (DDI) dataset [36], which contains skin lesion images annotated with Fitzpatrick Skin Types for binary classification (benign vs. malignant). The DDI dataset provides a realistic federated scenario where data is naturally partitioned by demographic characteristics. Table 6.1 summarises the benign and malignant sample distribution across Fitzpatrick skin type groups in the DDI dataset.

Table 6.1: DDI dataset breakdown by Fitzpatrick skin type group

FST Group	Benign	Malignant	Total
FST I–II (Light)	153	47	200
FST III–IV (Medium)	166	74	240
FST V–VI (Dark)	154	46	200
Total	473	167	640

We partition the DDI dataset into 3 clients representing different FST groups:

- Client 0: FST [0, 1] (Light skin types) - 160 train + 40 test images
- Client 1: FST [2, 3] (Medium skin types) - 192 train + 48 test images
- Client 2: FST [4, 5] (Dark skin types) - 160 train + 40 test images

This naturally heterogeneous partition reflects real-world scenarios where different medical institutions may serve demographically distinct populations, making it an ideal testbed for fairness-aware federated learning.

6.4.2 Experimental Design

To rigorously evaluate the model and quantify the source of its subgroups fairness gains, we propose a three-stage evaluation protocol.

Stage 1: Single-Fold Comparison

The first experiment establishes a preliminary comparison between the baseline and the proposed method on a single data fold.

- **Baseline:** Vanilla FeSViBS trained with standard cross-entropy loss and no sampling modification.
- **Our method:** Fairness-aware FL with domain adaptation and class imbalance mitigation.

This single-fold evaluation serves to confirm whether the addition of domain adaptation and imbalance mitigation yields visible improvements before performing more computationally intensive cross-validation. All configurations are presented in [Table 6.2](#)

Stage 2: Statistical Validation via 5-Fold Cross-Validation

To assess the statistical significance and robustness of the observed improvements, we perform 5-fold cross-validation using the two configurations identified in Stage 1.

- **Evaluation Protocol:** Each fold is trained for 200 epochs with a fixed random seed (42), ensuring consistent experimental conditions across folds.
- **Statistical Testing:** Paired t -tests are conducted on per-fold results (balanced accuracy and fairness score) to determine statistical significance of performance differences between FeSViBS and our method.

This stage provides quantitative evidence of whether our method achieves consistent and statistically reliable performance gains.

Stage 3: Ablation Study (Effect of Imbalance Mitigation and Domain Adaptation)

Finally, an ablation study is conducted to isolate the individual and combined effects of imbalance mitigation and domain adaptation. We train multiple configurations organised in a 2x2 matrix by adding or omitting imbalance mitigation strategies, including resampling and loss functions, to determine how each component contributes to overall subgroups fairness and accuracy. All configurations on all stages are trained for 200 epochs.

Table 6.2: Experiment Configurations

Configuration	Loss Type	Sampling
vanilla_fesvibs	CE	none
<i>Our method</i>		
weighted_ce_only	Weighted CE (bal. weights)	none
weighted_ce&sampling	Weighted CE (bal. weights)	weighted
focal_gamma_1.0	Focal Loss ($\gamma = 1.0$)	none
focal_gamma_1.5	Focal Loss ($\gamma = 1.5$)	none
focal_gamma_2.0	Focal Loss ($\gamma = 2.0$)	none
label_smoothing	Label Smoothing ($\alpha = 0.1$)	none
focal_weighted	Focal Loss ($\gamma = 2.0$)	weighted
ce_weighted_sampling	CE	weighted
ce_smote_sampling	CE	SMOTE

6.4.3 Fairness Metrics

We adopt comprehensive fairness metrics, including:

- **Overall Performance:** Accuracy, Balanced Accuracy
- **Fairness Metrics:** Accuracy Gap (Max-Min), Standard Deviation, Coefficient of Variation
- **Worst-Group Performance:** Minimum balanced accuracy across FST groups
- **Calibration Metrics:** Brier Score, Expected Calibration Error (ECE)
- **FST-Specific Performance:** Per-group accuracy, balanced accuracy, and F1-score

6.5 Results

6.5.1 Single-Fold Evaluation

We first conducted a single-fold experiment to establish the baseline and to verify the effect of domain adaptation and class imbalance mitigation for skin lesion classification. This same single fold is later used for the ablation study.

Table 6.3: Single-Fold Comparison: Baseline vs. Our method

Method	BAcc	Fairness Score
Baseline	0.732	0.693
Our method	0.776	0.759
Difference	+0.044 (+6.0%)	+0.066 (+9.5%)

This single-fold comparison in Table 6.3 shows a clear initial improvement when domain adaptation is combined with imbalance mitigation strategies. Our method consistently improves both balanced accuracy (+6.0%) and fairness score (+9.5%) relative to the baseline, motivating further validation using cross-validation and ablation analysis.

6.5.2 Statistical Validation (5-Fold Cross-Validation)

To evaluate whether the observed improvements generalise across data partitions, we conducted 5-fold cross-validation using the same two configurations. Each fold was trained for 200 epochs with a fixed random seed (42), and statistical significance was assessed using paired t -tests across folds. The summary of the cross-validation and the performance per fold is shown in Figure 6.2 and Figure 6.3, respectively. Although the differences did not reach statistical significance ($p > 0.05$, $n = 5$ folds), the results show medium effect sizes (Cohen’s $d=0.45$ - 0.63), suggesting practically meaningful improvements. The lack of significance primarily reflects the higher variance in our method’s performance across folds rather than an absence of consistent gains.

6.5.3 Ablation Study: Effect of Domain Adaptation and Imbalance Mitigation

The ablation study, conducted on the same single fold as the initial comparison, isolates the effects of domain adaptation and class imbalance handling. The 2x2 core matrix compares all combinations of these interventions, with additional variants exploring alternative loss and sampling strategies, as shown in [Table 6.4](#).

Table 6.4: Ablation Study Results (Single-Fold, 200 Epochs)

Rank	Configuration	BAcc	Fairness Score
<i>Core 2x2 Matrix</i>			
1	Baseline + DA + IM	0.776	0.759
6	Baseline + IM	0.746	0.723
8	Baseline + DA	0.720	0.712
9	Baseline	0.732	0.693
<i>Additional Variants: Baseline + DA + IM</i>			
2	+ IM (focal_gamma_1.5)	0.763	0.748
3	+ IM (focal_weighted)	0.768	0.732
4	+ IM (ce_weighted_sampling)	0.734	0.732
5	+ IM (focal_gamma_1.0)	0.734	0.732
7	+ IM (label_smoothing)	0.751	0.719
10	+ IM (weighted_ce)	0.706	0.692
11	+ IM (focal_gamma_2.0)	0.742	0.684
12	+ IM (ce_smote_sampling)	0.652	0.616

The ablation study reveals several key findings. First, domain adaptation alone improves the fairness score from 0.693 to 0.712, a gain of 2.7%. Second, adding weighted cross-entropy and weighted sampling to the baseline increases the fairness score from 0.693 to 0.723, representing a 4.3% improvement. When both interventions are combined, they achieve a fairness score of 0.759, exceeding the additive effect of either component alone, confirming a clear synergistic benefit. Among all tested configurations, the combination of weighted cross-entropy and weighted sampling performs best, outperforming alternative strategies such as focal loss, label smoothing, and SMOTE.

6.6 Discussion

The results align closely with the three-stage experimental design. The single-fold comparison confirmed that combining domain adaptation and imbalance mitigation substantially improves both balanced accuracy (from 0.732 to 0.776) and fairness score (from 0.693 to 0.759). In the cross-validation stage, our method consistently outperformed the baseline across folds, showing average gains of +3.9% in balanced accuracy and +4.7% in fairness score. Although these improvements were not statistically significant due to the limited number of folds, the medium effect sizes (Cohen’s $d = 0.45$ - 0.63) indicate practical, meaningful improvements. Finally, the ablation study using the same single fold revealed that imbalance mitigation alone contributed larger gains (+4.3%) than domain adaptation alone (+2.7%), while their combination produced synergistic effects, yielding a total improvement of +9.5%.

The absence of statistical significance is primarily due to the limited dataset size and number of folds rather than the absence of a real effect, as reflected by the medium effect sizes. Imbalance mitigation methods, particularly weighted cross-entropy and weighted sampling, consistently enhance subgroups fairness, underscoring the importance of addressing class imbalance before applying fairness-oriented adaptations. For practical deployment, our method provides the best overall trade-off between subgroups fairness and accuracy, while implementing imbalance mitigation strategy on the baseline offers a simpler and more stable alternative for federated environments with limited computational or data resources.

6.7 Conclusion

We develop a fairness-aware FL with domain adaptation and class imbalance mitigation for skin lesion classification, evaluate it on the demographically diverse DDI dataset in a federated setting, and introduce a controlled 2x2 ablation to separate the effects of domain adaptation and class imbalance handling. We reveal several important and nuanced findings:

- Our method with optimised imbalance mitigation technique (weighted CE + weighted sampling) achieves consistent improvements across all primary metrics on DDI on 5-folds

cross-validation balanced accuracy (+3.9%), fairness score (+4.7%), and worst-group performance (+4.2%). While not statistically significant at $p < 0.05$ level (p-values: 0.23-0.68), medium effect sizes (Cohen's d : 0.45-0.63) indicate practically meaningful differences.

- The lack of statistical significance despite consistent improvements highlights challenges inherent to limited data scenarios (640 images, 5 folds). The variance-to-difference ratio prevents achieving $p < 0.05$, underscoring the importance of larger datasets for definitive statistical conclusions in federated medical imaging research.
- Our ablation study on single-fold data provides clear evidence that addressing class imbalance (+4.3%) is more important than domain adaptation alone (+2.7%). The baseline with imbalance mitigation outperforms the baseline with domain adaptation (ranks #6 vs #8), confirming imbalance mitigation's fundamental importance.

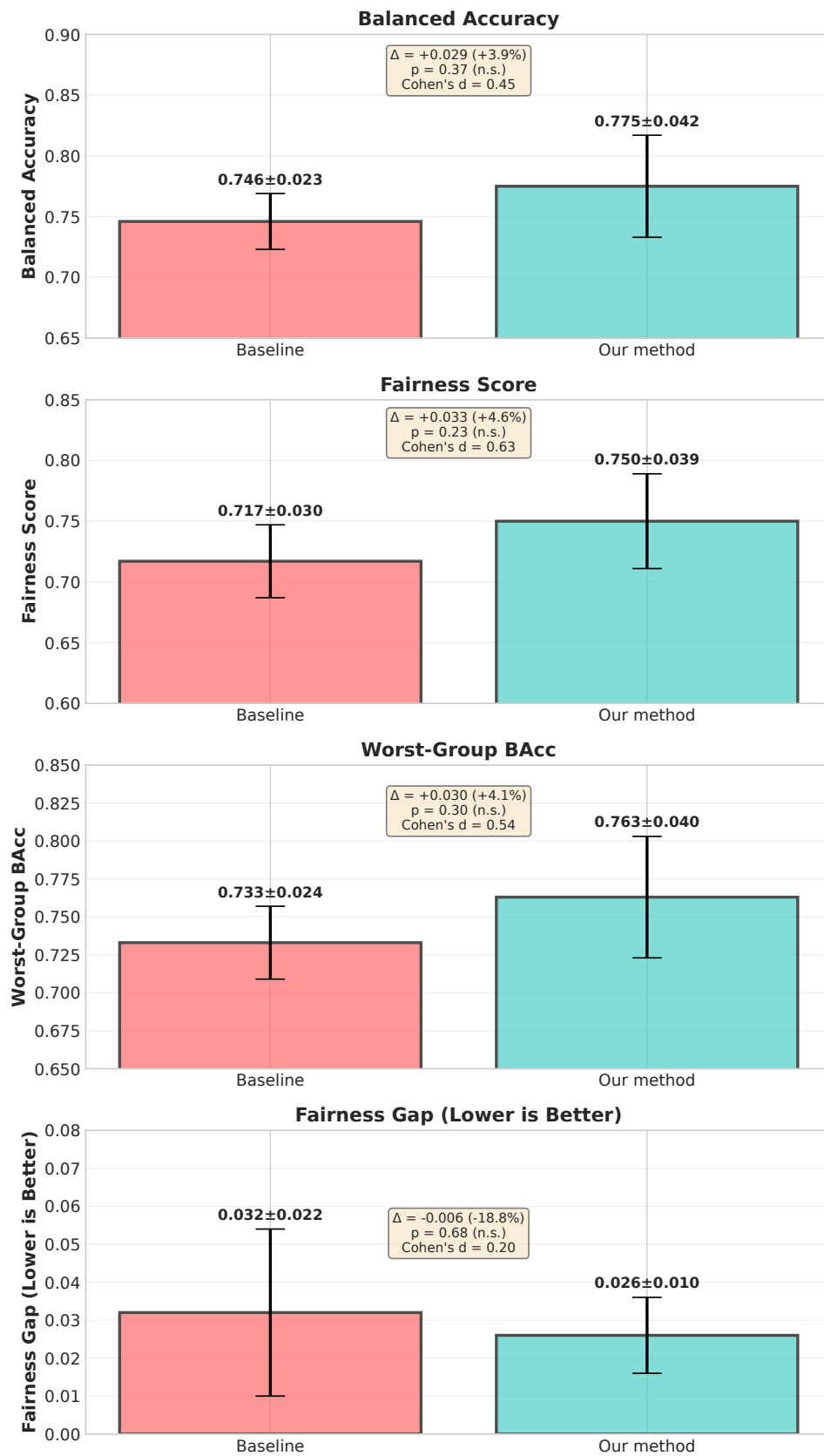


Figure 6.2: 5-fold cross-validation results showing mean $\pm$ standard deviation for all metrics. Error bars highlight that differences are consistent but not statistically significant at $p < 0.05$.

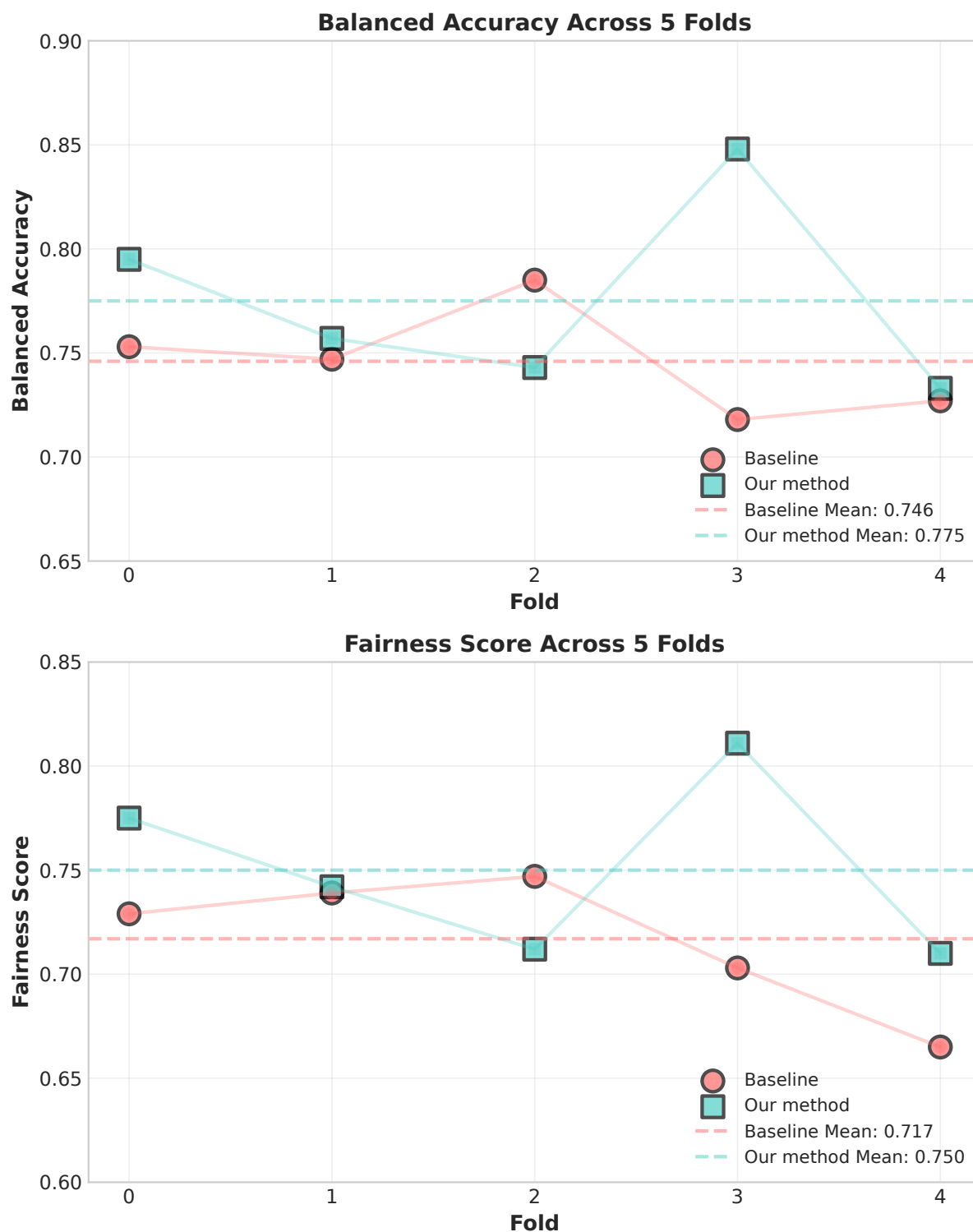


Figure 6.3: Per-fold performance across 5 folds for balanced accuracy and fairness score. Points indicate individual folds; dashed lines show means. Overlapping fold results reflect variance-limiting statistical power.

Conclusion

This thesis explored advanced methods to address the intertwined challenges of data imbalance, bias, subgroups fairness, and limited data in medical image classification, with a particular focus on dermatological imaging. Through four complementary work packages, we developed and evaluated novel approaches that contribute both technically and practically to the field of medical artificial intelligence (AI). The overarching aim was to enhance model performance while ensuring fairness and inclusivity across diverse populations, a critical requirement for equitable healthcare deployment in both centralised and federated settings.

Summary of Contributions

The first work package introduced **SIA-SMOTE**, a SMOTE-based oversampling method enhanced by a siamese network for generating more representative synthetic samples in high-dimensional image data. By leveraging similarity-aware interpolation, SIA-SMOTE improved classification performance across several benchmark datasets, particularly under severe class imbalance conditions. This work demonstrated that principled oversampling can mitigate imbalances in medical image data while preserving feature integrity.

The second work package presented a **self-attention fusion strategy** for vision transformers in skin lesion classification. By replacing the final attention layer with a fusion of diverse attention mechanisms, the method improved the model's capacity to capture complementary

feature representations. Evaluated on the ISIC 2017 dataset, this approach outperformed standard transformer baselines and achieved competitive results against the challenge leaderboard, even under computational constraints. This contribution underscored the potential of transformer-based architectures, enhanced with attention fusion, in advancing dermatological image analysis.

The third work package introduced **FedCIAL**, a federated colour-invariant adversarial learning framework, to address subgroups fairness concerns in federated skin lesion classification. FedCIAL incorporated explicit knowledge of Fitzpatrick skin type distribution shifts into an adversarial representation learning framework, mitigating demographic bias while preserving privacy in decentralised environments. Experiments on the Fitzpatrick17k dataset demonstrated that FedCIAL not only improved accuracy compared to the FeSViBS baseline but also reduced performance disparities across skin types, thereby enhancing fairness in model predictions.

Building on FedCIAL, the fourth work package developed a **fairness-aware federated learning** framework with domain adaptation and class-imbalance mitigation for skin lesion classification. This extension integrated colour-invariant domain adaptation with multiple imbalance-handling strategies and evaluated their joint effect on both overall performance and subgroup fairness. Using the Diverse Dermatology Images (DDI) dataset partitioned by Fitzpatrick skin type, the framework introduced a three-stage evaluation pipeline comprising single-fold comparison, 5-fold statistical validation, and a controlled 2x2 ablation study. The results showed that (i) imbalance mitigation is at least as important as domain adaptation for subgroups fairness, (ii) combining both yields synergistic improvements in balanced accuracy and fairness scores, and (iii) gains are practically meaningful even when statistical significance is limited by dataset size. This work contributed a systematic methodology for disentangling and quantifying subgroups fairness mechanisms in federated dermatological classification.

Limitations of The Research

Despite the encouraging results, we acknowledge several limitations of this research. First, dataset size and diversity remain a fundamental constraint. While multiple public dermatological datasets were used, including fitzpatrick17k and DDI, these datasets are still relatively small compared to natural image benchmarks and may not fully capture the diversity of real-world

clinical populations. This limitation is particularly evident in the federated experiments, where the number of clients and subgroup samples restricts the statistical power of cross-validation and hypothesis testing.

Second, subgroups fairness evaluation is inherently dependent on available demographic annotations. The focus of this thesis was primarily on skin tone fairness using the fitzpatrick scale, as this is one of the few demographic attributes consistently annotated in dermatology datasets. Other important sources of bias, such as age, sex, socioeconomic status, or comorbidities, were not explicitly modelled due to missing or unreliable labels. As a result, the subgroups fairness conclusions are necessarily scoped to observable skin tone disparities.

Third, the proposed oversampling and imbalance mitigation methods introduce additional modeling assumptions. Techniques such as sia-smote rely on the assumption that meaningful interpolation can be achieved in learned feature spaces. While empirical results support this assumption, there remains a risk that synthetic samples may not fully reflect the underlying clinical variability, particularly in extremely rare disease subtypes.

Finally, computational and communication constraints limit scalability in federated settings. Although split learning and efficient aggregation strategies were employed, adversarial training and multiple imbalance-handling configurations increase training complexity. These factors may pose challenges when scaling to very large federated networks or resource-constrained clinical environments.

Future Directions

Building on the findings and limitations of this thesis, several promising directions for future research emerge. First, integration of imbalance mitigation with representation learning warrants deeper investigation. Combining oversampling strategies such as SIA-SMOTE with transformer-based or self-supervised representation learning may further improve robustness in low-data regimes, particularly when labelled data are scarce.

Second, extending the self-attention fusion approach to multimodal medical data. In real clinical workflows, diagnostic decisions are rarely based on a single imaging modality. Future research could investigate transformer-based fusion architectures capable of jointly

modeling heterogeneous inputs combining dermoscopy, histopathology, and clinical metadata, where modality-specific encoders capture complementary feature representations and attention mechanisms dynamically weigh their relative importance. Such multimodal fusion may improve diagnostic accuracy, robustness, and interpretability, while also enabling subgroups fairness analysis across modalities.

Third, the fairness-aware federated frameworks developed in this thesis can be expanded to address other sources of bias beyond skin tone, such as age, gender, socioeconomic status, or imaging device differences, which are known to influence model performance in medical imaging. Future work could explore multi-attribute fairness settings, where multiple sensitive attributes are considered simultaneously. In addition, validating these frameworks across larger and more diverse federated networks spanning multiple hospitals and geographic regions would be essential to assess scalability, robustness, and real-world applicability.

Finally, larger and more diverse datasets are essential. Future work should evaluate fairness-aware and imbalance-aware methods on multi-institutional datasets with richer demographic annotations, ideally collected prospectively. This would enable stronger statistical conclusions and more comprehensive fairness analyses across multiple axes of bias.

Closing Remarks

In conclusion, this thesis demonstrates that advancing medical AI requires not only improving accuracy but also addressing subgroups fairness and inclusivity. Through SIA-SMOTE, attention fusion in transformers, FedCIAL, and the extended fairness-aware federated framework, we have proposed novel solutions to key challenges in imbalanced learning, representation learning, and federated subgroups fairness. Together, these contributions represent a step toward building AI systems that are not only technically powerful but also aligned with the ethical and practical imperatives of modern healthcare.

Bibliography

- [1] A. Ozdemir, K. Polat, and A. Alhudhaif, "Classification of imbalanced hyperspectral images using smote-based deep learning methods," *Expert Systems with Applications*, vol. 178, p. 114986, 2021, ISSN: 0957-4174. DOI: <https://doi.org/10.1016/j.eswa.2021.114986>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417421004279>.
- [2] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018, ISSN: 0893-6080. DOI: <https://doi.org/10.1016/j.neunet.2018.07.011>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0893608018302107>.
- [3] M. Groh, C. Harris, L. Soenksen, *et al.*, "Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1820–1828, Apr. 2021, ISSN: 21607516. DOI: [10.1109/CVPRW53098.2021.00201](https://doi.org/10.1109/CVPRW53098.2021.00201). [Online]. Available: <https://arxiv.org/abs/2104.09957v1>.
- [4] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, Feb. 2016. [Online]. Available: <https://arxiv.org/abs/1602.05629v4>.
- [5] S. Yang, X. Liu, Z. Zheng, W. Wang, and X. Ma, "Fusing medical image features and clinical features with deep learning for computer-aided diagnosis," Mar. 2021. DOI: [10.48550/arXiv.2103.05855](https://doi.org/10.48550/arXiv.2103.05855). [Online]. Available: <https://arxiv.org/pdf/2103.05855>.

- [6] Y. A. Kadhim, M. U. Khan, and A. Mishra, “Deep learning-based computer-aided diagnosis (cad): Applications for medical image datasets,” *Sensors* 2022, Vol. 22, Page 8999, vol. 22, p. 8999, 22 Nov. 2022, ISSN: 1424-8220. DOI: [10.3390/S22228999](https://doi.org/10.3390/S22228999). [Online]. Available: <https://www.mdpi.com/1424-8220/22/22/8999/htm%20https://www.mdpi.com/1424-8220/22/22/8999>.
- [7] C. Chen, N. A. M. Isa, and X. Liu, “A review of convolutional neural network based methods for medical image classification,” *Computers in Biology and Medicine*, vol. 185, Feb. 2025, ISSN: 18790534. DOI: [10.1016/j.combiomed.2024.109507](https://doi.org/10.1016/j.combiomed.2024.109507). [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/39631108/>.
- [8] G. K. Thakur, A. Thakur, S. Kulkarni, N. Khan, and S. Khan, “Deep learning approaches for medical image analysis and diagnosis,” *Cureus*, vol. 16, e59507, 5 May 2024. DOI: [10.7759/CUREUS.59507](https://doi.org/10.7759/CUREUS.59507). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11144045/>.
- [9] M. K. Santos, J. R. F. Junior, D. T. Wada, A. P. M. Tenorio, M. H. N. Barbosa, and P. M. D. A. Marques, “Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: Advances in imaging towards to precision medicine,” *Radiologia Brasileira*, vol. 52, p. 387, 6 Nov. 2019, ISSN: 01003984. DOI: [10.1590/0100-3984.2019.0049](https://doi.org/10.1590/0100-3984.2019.0049). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7007049/>.
- [10] I. D. Mienye, T. G. Swart, G. Obaido, M. Jordan, and P. Ilono, “Deep convolutional neural networks in medical image analysis: A review,” *Information* 2025, Vol. 16, Page 195, vol. 16, p. 195, 3 Mar. 2025, ISSN: 2078-2489. DOI: [10.3390/INFO16030195](https://doi.org/10.3390/INFO16030195). [Online]. Available: <https://www.mdpi.com/2078-2489/16/3/195/htm%20https://www.mdpi.com/2078-2489/16/3/195>.
- [11] S. Mathi, D. Deb, A. K. Chaudhuri, L. Joseph, and S. Biswas, “A review of convolutional neural networks for medical image analysis: Trends and future directions,” *Journal of Neonatal Surgery*, vol. 14, pp. 90–97, 7 May 2025. DOI: [10.63682/jns.v14i7.5117](https://doi.org/10.63682/jns.v14i7.5117). [Online]. Available: <https://www.jneonatsurg.com/index.php/jns/article/view/5117>.

- [12] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Annals of Translational Medicine*, vol. 8, p. 713, 11 Jun. 2020, ISSN: 23055839. DOI: [10.21037/ATM.2020.02.44](https://doi.org/10.21037/ATM.2020.02.44). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7327346/>.
- [13] L. Mevorach, A. Farcomeni, G. Pellacani, and C. Cantisani, "A comparison of skin lesions' diagnoses between ai-based image classification, an expert dermatologist, and a non-expert," *Diagnostics*, vol. 15, p. 1115, 9 May 2025, ISSN: 20754418. DOI: [10.3390/DIAGNOSTICS15091115](https://doi.org/10.3390/DIAGNOSTICS15091115). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12071753/>.
- [14] J. K. Winkler, A. Blum, K. Kommoss, *et al.*, "Assessment of diagnostic performance of dermatologists cooperating with a convolutional neural network in a prospective clinical study: Human with machine," *JAMA Dermatology*, vol. 159, pp. 621–627, 6 Jun. 2023, ISSN: 21686084. DOI: [10.1001/JAMADERMATOL.2023.0905](https://doi.org/10.1001/JAMADERMATOL.2023.0905),. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/37133847/>.
- [15] J. Y. Kim, W. S. Ryu, D. Kim, and E. Y. Kim, "Better performance of deep learning pulmonary nodule detection using chest radiography with pixel level labels in reference to computed tomography: Data quality matters," *Scientific Reports*, vol. 14, pp. 1–9, 1 Dec. 2024, ISSN: 20452322. DOI: [10.1038/S41598-024-66530-Y](https://doi.org/10.1038/S41598-024-66530-Y);SUBJMETA=117,1421,1770,639,692,700,705;KWRD=COMPUTER+SCIENCE,RADIOGRAPHY. [Online]. Available: <https://www.nature.com/articles/s41598-024-66530-y>.
- [16] W. Hendrix, N. Hendrix, E. T. Scholten, *et al.*, "Deep learning for the detection of benign and malignant pulmonary nodules in non-screening chest ct scans," *Communications Medicine*, vol. 3, 1 Dec. 2023, ISSN: 2730664X. DOI: [10.1038/S43856-023-00388-5](https://doi.org/10.1038/S43856-023-00388-5),. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/37891360/>.
- [17] R. Aggarwal, V. Sounderajah, G. Martin, *et al.*, "Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis," *npj Digital Medicine* 2021 4:1, vol. 4, pp. 1–23, 1 Apr. 2021, ISSN: 2398-6352. DOI: [10.1038/s41746-021-00438-z](https://doi.org/10.1038/s41746-021-00438-z). [Online]. Available: <https://www.nature.com/articles/s41746-021-00438-z>.

- [18] M. M. Ahsan, S. A. Luna, and Z. Siddique, “Machine-learning-based disease diagnosis: A comprehensive review,” *Healthcare*, vol. 10, p. 541, 3 Mar. 2022, ISSN: 22279032. DOI: [10.3390/HEALTHCARE10030541](https://doi.org/10.3390/HEALTHCARE10030541). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8950225/>.
- [19] Y. Zhao, Z. S. Y. Wong, and K. L. Tsui, “A framework of rebalancing imbalanced healthcare data for rare events’ classification: A case of look-alike sound-alike mix-up incident detection,” *Journal of Healthcare Engineering*, vol. 2018, p. 6 275 435, 2018, ISSN: 20402309. DOI: [10.1155/2018/6275435](https://doi.org/10.1155/2018/6275435). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5987310/>.
- [20] N. Hasani, F. Farhadi, M. A. Morris, *et al.*, “Artificial intelligence in medical imaging and its impact on the rare disease community: Threats, challenges and opportunities,” *PET clinics*, vol. 17, p. 13, 1 Jan. 2022, ISSN: 18799809. DOI: [10.1016/J.CPET.2021.09.009](https://doi.org/10.1016/J.CPET.2021.09.009). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8764708/>.
- [21] A. Galdran, G. Carneiro, and M. A. G. Ballester, “Balanced-mixup for highly imbalanced medical image classification,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12905 LNCS, pp. 323–333, Sep. 2021. DOI: [10.1007/978-3-030-87240-3_31](https://doi.org/10.1007/978-3-030-87240-3_31). [Online]. Available: <http://arxiv.org/abs/2109.09850> http://dx.doi.org/10.1007/978-3-030-87240-3_31.
- [22] L. Gao, L. Zhang, C. Liu, and S. Wu, “Handling imbalanced medical image data: A deep-learning-based one-class classification approach,” *Artificial Intelligence in Medicine*, vol. 108, p. 101 935, Aug. 2020, ISSN: 0933-3657. DOI: [10.1016/J.ARTMED.2020.101935](https://doi.org/10.1016/J.ARTMED.2020.101935).
- [23] M. Lin, T. Li, Y. Yang, *et al.*, “Improving model fairness in image-based computer-aided diagnosis,” *Nature Communications* 2023 14:1, vol. 14, pp. 1–9, 1 Oct. 2023, ISSN: 2041-1723. DOI: [10.1038/s41467-023-41974-4](https://doi.org/10.1038/s41467-023-41974-4). [Online]. Available: <https://www.nature.com/articles/s41467-023-41974-4>.
- [24] A. O’malley, M. Veenhuizen, and A. Ahmed, “Ensuring appropriate representation in artificial intelligence-generated medical imagery: Protocol for a methodological approach

- to address skin tone bias,” *JMIR AI*, vol. 3, e58275, 2024, ISSN: 28171705. DOI: [10.2196/58275](https://doi.org/10.2196/58275). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11635324/>.
- [25] N. Alipour, T. Burke, and J. Courtney, “Skin type diversity in skin lesion datasets: A review,” *Current Dermatology Reports*, vol. 13, p. 198, 3 Sep. 2024, ISSN: 21624933. DOI: [10.1007/S13671-024-00440-0](https://doi.org/10.1007/S13671-024-00440-0). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11343783/>.
- [26] S. Du, B. Hers, N. Bayasi, G. Hamarneh, and R. Garbi, “Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning,” *Lecture Notes in Computer Science*, vol. 13804 LNCS, pp. 185–202, Aug. 2022, ISSN: 16113349. DOI: [10.1007/978-3-031-25069-9_13](https://doi.org/10.1007/978-3-031-25069-9_13). [Online]. Available: <https://arxiv.org/pdf/2208.10013>.
- [27] G. Kleinberg, M. J. Diaz, S. Batchu, and B. Lucke-Wold, “Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare,” *Journal of Biomed Research*,; *Volume 3(Issue 1)*:42-47, vol. Volume 3, pp. 42–47, Issue 1 Dec. 2022, ISSN: 2693-5910. DOI: [10.46439/BIOMEDRES.3.025](https://doi.org/10.46439/BIOMEDRES.3.025). [Online]. Available: <https://www.probiologists.com/article/racial-underrepresentation-in-dermatological-datasets-leads-to-biased-machine-learning-models-and-inequitable-healthcare>.
- [28] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, “Addressing bias in big data and ai for health care: A call for open science,” *Patterns*, vol. 2, p. 100347, 10 Oct. 2021, ISSN: 26663899. DOI: [10.1016/J.PATTER.2021.100347](https://doi.org/10.1016/J.PATTER.2021.100347). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8515002/>.
- [29] A. Halder, S. Gharami, P. Sadhu, P. K. Singh, M. Wozniak, and M. F. Ijaz, “Implementing vision transformer for classifying 2d biomedical images,” *Scientific Reports*, vol. 14, pp. 1–19, 1 Dec. 2024, ISSN: 20452322. DOI: [10.1038/S41598-024-63094-9](https://doi.org/10.1038/S41598-024-63094-9);SUBJMETA=1042, 117,258,639,705;KWRD=COMPUTATIONAL+SCIENCE,COMPUTER+SCIENCE, INFORMATION+TECHNOLOGY. [Online]. Available: <https://www.nature.com/articles/s41598-024-63094-9>.

- [30] K. Al-hammuri, F. Gebali, A. Kanan, and I. T. Chelvan, "Vision transformer architecture and applications in digital health: A tutorial and survey," *Visual Computing for Industry, Biomedicine, and Art*, vol. 6, p. 14, 1 Dec. 2023, ISSN: 25244442. DOI: [10.1186/S42492-023-00140-9](https://doi.org/10.1186/S42492-023-00140-9). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10333157/>.
- [31] S. Aburass, O. Dorgham, J. A. Shaqsi, M. A. Rumman, and O. Al-Kadi, "Vision transformers in medical imaging: A comprehensive review of advancements and applications across multiple diseases," *Journal of Imaging Informatics in Medicine*, Mar. 2025, ISSN: 2948-2933. DOI: [10.1007/S10278-025-01481-Y](https://doi.org/10.1007/S10278-025-01481-Y). [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/40164818/%20https://pubmed.ncbi.nlm.nih.gov/40164818/?fc=20240423223220&ff=20250403000316&v=2.18.0.post9+e462414>.
- [32] R. Haripriya, N. Khare, and M. Pandey, "Privacy-preserving federated learning for collaborative medical data mining in multi-institutional settings," *Scientific Reports* 2025 15:1, vol. 15, pp. 1–30, 1 Apr. 2025, ISSN: 2045-2322. DOI: [10.1038/s41598-025-97565-4](https://doi.org/10.1038/s41598-025-97565-4). [Online]. Available: <https://www.nature.com/articles/s41598-025-97565-4>.
- [33] F. C. Ghesu, B. Georgescu, A. Mansoor, *et al.*, "Contrastive self-supervised learning from 100 million medical images with optional supervision," *Journal of Medical Imaging*, vol. 9, p. 064 503, 6 Nov. 2022, ISSN: 23294310. DOI: [10.1117/1.JMI.9.6.064503](https://doi.org/10.1117/1.JMI.9.6.064503). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9710476/>.
- [34] S. C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari, "Self-supervised learning for medical image classification: A systematic review and implementation guidelines," *npj Digital Medicine*, vol. 6, pp. 1–16, 1 Dec. 2023, ISSN: 23986352. DOI: [10.1038/S41746-023-00811-0;SUBJMETA=1046,117,1421,639,692,700,705;KWRD=COMPUTER+SCIENCE,MEDICAL+IMAGING,SCIENTIFIC+DATA](https://doi.org/10.1038/S41746-023-00811-0;SUBJMETA=1046,117,1421,639,692,700,705;KWRD=COMPUTER+SCIENCE,MEDICAL+IMAGING,SCIENTIFIC+DATA). [Online]. Available: <https://www.nature.com/articles/s41746-023-00811-0>.
- [35] G. Xu, Y. Wu, J. Hu, and Y. Shi, "Achieving fairness in dermatological disease diagnosis through automatic weight adjusting federated learning and personalization," Aug. 2022. [Online]. Available: <https://arxiv.org/abs/2208.11187v1>.

- [36] R. Daneshjou, K. Vodrahalli, R. A. Novoa, *et al.*, “Disparities in dermatology ai performance on a diverse, curated clinical image set,” *Science Advances*, vol. 8, p. 6147, 32 Aug. 2022, ISSN: 23752548. DOI: [10.1126/SCIADV.ABQ6147](https://doi.org/10.1126/SCIADV.ABQ6147). [Online]. Available: [/doi/pdf/10.1126/sciadv.abq6147](https://doi.org/10.1126/sciadv.abq6147).
- [37] C. Mosquera, L. Ferrer, D. H. Milone, D. Luna, and E. Ferrante, “Class imbalance on medical image classification: Towards better evaluation practices for discrimination and calibration performance,” *European Radiology*, vol. 34, pp. 7895–7903, 12 Dec. 2024, ISSN: 14321084. DOI: [10.1007/S00330-024-10834-0](https://doi.org/10.1007/S00330-024-10834-0). [Online]. Available: <https://link.springer.com/article/10.1007/s00330-024-10834-0>.
- [38] W. Albattah and R. U. Khan, “Impact of imbalanced features on large datasets,” *Frontiers in Big Data*, vol. 8, p. 1455442, 2025, ISSN: 2624909X. DOI: [10.3389/FDATA.2025.1455442](https://doi.org/10.3389/FDATA.2025.1455442). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11948280/>.
- [39] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, “Handling imbalanced medical datasets: Review of a decade of research,” *Artificial Intelligence Review*, vol. 57, pp. 1–57, 10 Oct. 2024, ISSN: 15737462. DOI: [10.1007/S10462-024-10884-2](https://doi.org/10.1007/S10462-024-10884-2). [Online]. Available: <https://link.springer.com/article/10.1007/s10462-024-10884-2>.
- [40] S. Iqbal, A. N. Qureshi, J. Li, I. A. Choudhry, and T. Mahmood, “Dynamic learning for imbalanced data in learning chest x-ray and ct images,” *Heliyon*, vol. 9, e16807, 6 Jun. 2023, ISSN: 2405-8440. DOI: [10.1016/J.HELİYON.2023.E16807](https://doi.org/10.1016/J.HELİYON.2023.E16807). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2405844023040148>.
- [41] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, ISSN: 1076-9757. DOI: [10.1613/JAIR.953](https://doi.org/10.1613/JAIR.953). [Online]. Available: <https://www.jair.org/index.php/jair/article/view/10302>.
- [42] J. Singh, C. Beeche, Z. Shi, *et al.*, “Batch-balanced focal loss: A hybrid solution to class imbalance in deep learning,” *Journal of Medical Imaging*, vol. 10, p. 051809, 5 Jun. 2023, ISSN: 23294310. DOI: [10.1117/1.JMI.10.5.051809](https://doi.org/10.1117/1.JMI.10.5.051809). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10289178/>.

- [43] W. Qu, I. Balki, M. Mendez, J. Valen, J. Levman, and P. N. Tyrrell, "Assessing and mitigating the effects of class imbalance in machine learning with application to x-ray imaging," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, pp. 2041–2048, 12 Dec. 2020, ISSN: 18616429. DOI: [10.1007/S11548-020-02260-6](https://doi.org/10.1007/S11548-020-02260-6). [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/32965624/>.
- [44] D. Schaudt, R. von Schwerin, A. Hafner, *et al.*, "Augmentation strategies for an imbalanced learning problem on a novel covid-19 severity dataset," *Scientific Reports*, vol. 13, pp. 1–16, 1 Dec. 2023, ISSN: 20452322. DOI: [10.1038/S41598-023-45532-2](https://doi.org/10.1038/S41598-023-45532-2);SUBJMETA=114,1305,1421,2413,631,692,700;KWRD=MACHINE+LEARNING,MEDICAL+IMAGING,PREDICTIVE+MEDICINE. [Online]. Available: <https://www.nature.com/articles/s41598-023-45532-2>.
- [45] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," *Lecture Notes in Computer Science*, vol. 3644, pp. 878–887, PART I 2005, ISSN: 03029743. DOI: [10.1007/11538059_91](https://doi.org/10.1007/11538059_91)/COVER. [Online]. Available: https://link.springer.com/chapter/10.1007/11538059_91.
- [46] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," *Proceedings of the International Joint Conference on Neural Networks*, pp. 1322–1328, 2008. DOI: [10.1109/IJCNN.2008.4633969](https://doi.org/10.1109/IJCNN.2008.4633969).
- [47] Y. Li, Y. Yang, P. Song, L. Duan, and R. Ren, "An improved smote algorithm for enhanced imbalanced data classification by expanding sample generation space," *Scientific Reports* 2025 15:1, vol. 15, pp. 23 521–, 1 Jul. 2025, ISSN: 2045-2322. DOI: [10.1038/s41598-025-09506-w](https://doi.org/10.1038/s41598-025-09506-w). [Online]. Available: <https://www.nature.com/articles/s41598-025-09506-w>.
- [48] D. Dablain, B. Krawczyk, and N. V. Chawla, "Deepsmote: Fusing deep learning and smote for imbalanced data," *IEEE transactions on neural networks and learning systems*, vol. PP, 2022, ISSN: 2162-2388. DOI: [10.1109/TNNLS.2021.3136503](https://doi.org/10.1109/TNNLS.2021.3136503). [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/35085094/>.
- [49] S. Chatterjee, S. Bhattacharjee, K. Ghosh, A. K. Das, and S. Banerjee, "Class-biased sarcasm detection using bilstm variational autoencoder-based synthetic oversampling,"

- Soft Computing* 2023 27:9, vol. 27, pp. 5603–5620, 9 Mar. 2023, ISSN: 1433-7479. DOI: [10.1007/S00500-023-07956-W](https://doi.org/10.1007/S00500-023-07956-W). [Online]. Available: <https://link.springer.com/article/10.1007/s00500-023-07956-w>.
- [50] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, “Data augmentation using generative adversarial networks (cycleGAN) to improve generalizability in CT segmentation tasks,” *Scientific Reports* 2019 9:1, vol. 9, pp. 1–9, 1 Nov. 2019, ISSN: 2045-2322. DOI: [10.1038/s41598-019-52737-x](https://doi.org/10.1038/s41598-019-52737-x). [Online]. Available: <https://www.nature.com/articles/s41598-019-52737-x>.
- [51] L. M. Shoohi and J. H. Saud, “DcGAN for handling imbalanced malaria dataset based on over-sampling technique and using CNN,” *Medico Legal Update*, vol. 20, pp. 1079–1085, 1 Apr. 2020, ISSN: 0974-1283. DOI: [10.37506/MLU.V20I1.514](https://doi.org/10.37506/MLU.V20I1.514). [Online]. Available: <https://ijop.net/index.php/mlu/article/view/514>.
- [52] C. Muramatsu, M. Nishio, T. Goto, *et al.*, “Improving breast mass classification by shared data with domain transformation using a generative adversarial network,” *Computers in Biology and Medicine*, vol. 119, p. 103698, Apr. 2020, ISSN: 0010-4825. DOI: [10.1016/j.compbiomed.2020.103698](https://doi.org/10.1016/j.compbiomed.2020.103698).
- [53] W. Xu, N. Souly, and P. P. Brahma, “Reliability of GAN generated data to train and validate perception systems for autonomous vehicles,” *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2021.
- [54] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, “Perceptual generative adversarial networks for small object detection,” *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 1951–1959, Nov. 2017. DOI: [10.1109/CVPR.2017.211](https://doi.org/10.1109/CVPR.2017.211).
- [55] “Cost-sensitive learning,” *Encyclopedia of Machine Learning*, pp. 231–235, 2011. DOI: [10.1007/978-0-387-30164-8_181](https://doi.org/10.1007/978-0-387-30164-8_181). [Online]. Available: https://link.springer.com/referenceworkentry/10.1007/978-0-387-30164-8_181.
- [56] S. I. Ali, H. S. M. Bilal, M. Hussain, *et al.*, “Ensemble feature ranking for cost-based non-overlapping groups: A case study of chronic kidney disease diagnosis in developing

- countries,” *IEEE Access*, vol. 8, pp. 215 623–215 648, 2020, issn: 21693536. doi: [10.1109/ACCESS.2020.3040650](https://doi.org/10.1109/ACCESS.2020.3040650).
- [57] M. Zieba, J. M. Tomczak, M. Lubicz, and J. Swiatek, “Boosted svm for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients,” *Applied Soft Computing*, vol. 14, pp. 99–108, PART A Jan. 2014, issn: 1568-4946. doi: [10.1016/J.ASOC.2013.07.016](https://doi.org/10.1016/J.ASOC.2013.07.016).
- [58] I. D. Mienye and Y. Sun, “Performance analysis of cost-sensitive learning methods with application to imbalanced medical data,” *Informatics in Medicine Unlocked*, vol. 25, p. 100 690, Jan. 2021, issn: 2352-9148. doi: [10.1016/J.IMU.2021.100690](https://doi.org/10.1016/J.IMU.2021.100690).
- [59] X. Zhang and B. G. Hu, “A new strategy of cost-free learning in the class imbalance problem,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, pp. 2872–2885, 12 Dec. 2014, issn: 10414347. doi: [10.1109/TKDE.2014.2312336](https://doi.org/10.1109/TKDE.2014.2312336).
- [60] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 318–327, 2 Aug. 2017, issn: 19393539. doi: [10.1109/TPAMI.2018.2858826](https://doi.org/10.1109/TPAMI.2018.2858826). [Online]. Available: <https://arxiv.org/abs/1708.02002v2>.
- [61] Y. Cui, M. Jia, T. Y. Lin, Y. Song, and S. Belongie, “Class-balanced loss based on effective number of samples,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 9260–9269, Jan. 2019, issn: 10636919. doi: [10.1109/CVPR.2019.00949](https://doi.org/10.1109/CVPR.2019.00949). [Online]. Available: <https://arxiv.org/abs/1901.05555v1>.
- [62] A. Krishna, M. Sadeep, J. Ankit, *et al.*, “Long-tail learning via logit adjustment,” 2021. doi: [10.48550/arXiv.2007.07314](https://doi.org/10.48550/arXiv.2007.07314). [Online]. Available: https://github.com/google-research/google-research/tree/master/logit_adjustment..
- [63] T. Le, L. H. Son, M. T. Vo, M. Y. Lee, and S. W. Baik, “A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset,” *Symmetry 2018, Vol. 10, Page 250*, vol. 10, p. 250, 7 Jul. 2018, issn: 2073-8994. doi: [10.3390/SYM10070250](https://doi.org/10.3390/SYM10070250).

- [Online]. Available: <https://www.mdpi.com/2073-8994/10/7/250/htm%20https://www.mdpi.com/2073-8994/10/7/250>.
- [64] H. Hong, J. Liu, and A. X. Zhu, “Modeling landslide susceptibility using logitboost alternating decision trees and forest by penalizing attributes with the bagging ensemble,” *Science of The Total Environment*, vol. 718, p. 137 231, May 2020, ISSN: 0048-9697. DOI: [10.1016/J.SCITOTENV.2020.137231](https://doi.org/10.1016/J.SCITOTENV.2020.137231).
- [65] A. Subasi, B. Kadasa, and E. Kremic, “Classification of the cardiotocogram data for anticipation of fetal risks using bagging ensemble classifier,” *Procedia Computer Science*, vol. 168, pp. 34–39, Jan. 2020, ISSN: 1877-0509. DOI: [10.1016/J.PROCS.2020.02.248](https://doi.org/10.1016/J.PROCS.2020.02.248).
- [66] P. Charoenkwan, W. Chiangjong, C. Nantasenamat, M. M. Hasan, B. Manavalan, and W. Shoombuatong, “Stackil6: A stacking ensemble model for improving the prediction of il-6 inducing peptides,” *Briefings in Bioinformatics*, vol. 22, 6 Nov. 2021, ISSN: 14774054. DOI: [10.1093/BIB/BBAB172](https://doi.org/10.1093/BIB/BBAB172). [Online]. Available: <https://academic.oup.com/bib/article/22/6/bbab172/6271998>.
- [67] S. Rajagopal, P. P. Kundapur, and K. S. Hareesha, “A stacking ensemble for network intrusion detection using heterogeneous datasets,” *Security and Communication Networks*, vol. 2020, 2020, ISSN: 19390122. DOI: [10.1155/2020/4586875](https://doi.org/10.1155/2020/4586875).
- [68] P. Kumari and D. Toshniwal, “Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance,” *Journal of Cleaner Production*, vol. 279, p. 123 285, Jan. 2021, ISSN: 0959-6526. DOI: [10.1016/J.JCLEPRO.2020.123285](https://doi.org/10.1016/J.JCLEPRO.2020.123285).
- [69] N. L. Rane, S. P. Choudhary, and J. Rane, “Ensemble deep learning and machine learning: Applications, opportunities, challenges, and future directions,” *Studies in Medical and Health Sciences*, vol. 1, pp. 18–41, 2 Jul. 2024, ISSN: 3007-3707. DOI: [10.48185/SMHS.V1I2.1225](https://doi.org/10.48185/SMHS.V1I2.1225). [Online]. Available: <https://sabapub.com/index.php/SMHS/article/view/1225>.
- [70] L. Cui, M. Xu, C. Liu, *et al.*, “Towards reliable healthcare imaging: A multifaceted approach in class imbalance handling for medical image segmentation,” *Interdisciplinary*

- Sciences, Computational Life Sciences*, vol. 17, p. 614, 3 Sep. 2025, ISSN: 18671462. DOI: [10.1007/S12539-025-00726-2](https://doi.org/10.1007/S12539-025-00726-2). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12289783/>.
- [71] I. Park, W. H. Kim, and J. Ryu, “Style-kd: Class-imbalanced medical image classification via style knowledge distillation,” *Biomedical Signal Processing and Control*, vol. 91, p. 105 928, May 2024, ISSN: 1746-8094. DOI: [10.1016/J.BSPC.2023.105928](https://doi.org/10.1016/J.BSPC.2023.105928). [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S1746809423013617>.
- [72] S. Carey, A. Pang, and M. de Kamps, “Fairness in ai for healthcare,” *Future Healthcare Journal*, vol. 11, p. 100 177, 3 Sep. 2024, ISSN: 2514-6645. DOI: [10.1016/J.FHJ.2024.100177](https://doi.org/10.1016/J.FHJ.2024.100177). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2514664524015674>.
- [73] D. Ueda, T. Kakinuma, S. Fujita, *et al.*, “Fairness of artificial intelligence in healthcare: Review and recommendations,” *Japanese Journal of Radiology*, vol. 42, p. 3, 1 Jan. 2023, ISSN: 1867108X. DOI: [10.1007/S11604-023-01474-3](https://doi.org/10.1007/S11604-023-01474-3). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10764412/>.
- [74] R. J. Chen, J. J. Wang, D. F. Williamson, *et al.*, “Algorithmic fairness in artificial intelligence for medicine and healthcare,” *Nature Biomedical Engineering* 2023 7:6, vol. 7, pp. 719–742, 6 Jun. 2023, ISSN: 2157-846X. DOI: [10.1038/s41551-023-01056-8](https://doi.org/10.1038/s41551-023-01056-8). [Online]. Available: <https://www.nature.com/articles/s41551-023-01056-8>.
- [75] G. Stanovsky, N. A. Smith, and L. Zettlemoyer, “Evaluating gender bias in machine translation,” *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, vol. 117, pp. 1679–1684, 23 Jun. 2020. DOI: [10.18653/v1/p19-1164](https://doi.org/10.18653/v1/p19-1164). [Online]. Available: [/doi/pdf/10.1073/pnas.1919012117?download=true](https://doi.org/10.1073/pnas.1919012117?download=true).
- [76] Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, and M. Ghassemi, “The limits of fair medical imaging ai in real-world generalization,” *Nature Medicine*, vol. 30, pp. 2838–2848, 10 Oct. 2024, ISSN: 1546170X. DOI: [10.1038/S41591-024-03113-4;SUBJMETA=139](https://doi.org/10.1038/S41591-024-03113-4;SUBJMETA=139),

- 1421, 1770, 692, 700; KWRD=DIAGNOSIS, RADIOGRAPHY. [Online]. Available: <https://www.nature.com/articles/s41591-024-03113-4>.
- [77] T. Dowie, “Exploring the diagnostic capability of artificial intelligence in dermatology for darker skin tones: A narrative review,” *Cureus*, vol. 17, e94909, 10 Oct. 2025, ISSN: 2168-8184. DOI: [10.7759/CUREUS.94909](https://doi.org/10.7759/CUREUS.94909). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12624499/>.
- [78] Y. J. Juhn, E. Ryu, C. I. Wi, *et al.*, “Assessing socioeconomic bias in machine learning algorithms in health care: A case study of the houses index,” *Journal of the American Medical Informatics Association : JAMIA*, vol. 29, p. 1142, 7 Jul. 2022, ISSN: 1527974X. DOI: [10.1093/JAMIA/OCAC052](https://doi.org/10.1093/JAMIA/OCAC052). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9196683/>.
- [79] S. Wang, C. Li, R. Wang, *et al.*, “Annotation-efficient deep learning for automatic medical image segmentation,” *Nature Communications*, vol. 12, p. 5915, 1 Dec. 2021, ISSN: 20411723. DOI: [10.1038/S41467-021-26216-9](https://doi.org/10.1038/S41467-021-26216-9). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8501087/>.
- [80] J. L. Cross, M. A. Choma, and J. A. Onofrey, “Bias in medical ai: Implications for clinical decision-making,” *PLOS Digital Health*, vol. 3, e0000651, 11 Nov. 2024, ISSN: 27673170. DOI: [10.1371/JOURNAL.PDIG.0000651](https://doi.org/10.1371/JOURNAL.PDIG.0000651). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11542778/>.
- [81] M. K. Cho, “Rising to the challenge of bias in health care ai,” *Nature medicine*, vol. 27, p. 2079, 12 Dec. 2021, ISSN: 1546170X. DOI: [10.1038/S41591-021-01577-2](https://doi.org/10.1038/S41591-021-01577-2). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11017306/>.
- [82] Y. Yang, Y. Liu, X. Liu, *et al.*, “Demographic bias of expert-level vision-language foundation models in medical imaging,” *Science Advances*, vol. 11, 13 Mar. 2025, ISSN: 23752548. DOI: [10.1126/SCIADV.ADQ0305](https://doi.org/10.1126/SCIADV.ADQ0305); PAGEGROUP:STRING:PUBLICATION. [Online]. Available: [/doi/pdf/10.1126/sciadv.adq0305?download=true](https://doi.org/10.1126/sciadv.adq0305?download=true).
- [83] M. Liu, Y. Ning, S. Teixayavong, *et al.*, “A scoping review and evidence gap analysis of clinical ai fairness,” *npj Digital Medicine* 2025 8:1, vol. 8, pp. 360–, 1 Jun. 2025,

- ISSN: 2398-6352. DOI: [10.1038/s41746-025-01667-2](https://doi.org/10.1038/s41746-025-01667-2). [Online]. Available: <https://www.nature.com/articles/s41746-025-01667-2>.
- [84] Z. Xu, J. Li, Q. Yao, H. Li, M. Zhao, and S. K. Zhou, “Addressing fairness issues in deep learning-based medical image analysis: A systematic review,” *npj Digital Medicine*, vol. 7, pp. 1–16, 1 Dec. 2024, ISSN: 23986352. DOI: [10.1038/S41746-024-01276-5](https://doi.org/10.1038/S41746-024-01276-5);SUBJMETA=179,3935,648,692,700,706;KWRD=ETHICS,MEDICAL+ETHICS. [Online]. Available: <https://www.nature.com/articles/s41746-024-01276-5>.
- [85] C. Li, S. Ding, N. Zou, X. Hu, X. Jiang, and K. Zhang, “Multi-task learning with dynamic re-weighting to achieve fairness in healthcare predictive modeling,” *Journal of biomedical informatics*, vol. 143, p. 104 399, Jul. 2023, ISSN: 15320464. DOI: [10.1016/J.JBI.2023.104399](https://doi.org/10.1016/J.JBI.2023.104399). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10665114/>.
- [86] J. Waithira, R. Chweya, and R. M. Cyprian, “Adversarial debiasing for bias mitigation in healthcare ai systems: A literature review,” *Journal 2025*, vol. 12, p. 13 340, 2025, ISSN: 2333-9705. DOI: [10.4236/oalib.1113340](https://doi.org/10.4236/oalib.1113340). [Online]. Available: <https://doi.org/10.4236/oalib.1113340>.
- [87] R. Correa, K. Pahwa, B. Patel, C. M. Vachon, J. W. Gichoya, and I. Banerjee, “Efficient adversarial debiasing with concept activation vector - medical image case-studies,” *Journal of biomedical informatics*, vol. 149, Jan. 2024, ISSN: 1532-0480. DOI: [10.1016/J.JBI.2023.104548](https://doi.org/10.1016/J.JBI.2023.104548). [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/38043883/>.
- [88] N. T. Madathil, F. K. Dankar, M. Gergely, A. N. Belkacem, and S. Alrabae, “Revolutionizing healthcare data analytics with federated learning: A comprehensive survey of applications, systems, and future directions,” *Computational and Structural Biotechnology Journal*, vol. 28, pp. 217–238, Jan. 2025, ISSN: 2001-0370. DOI: [10.1016/J.CSBJ.2025.06.009](https://doi.org/10.1016/J.CSBJ.2025.06.009). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2001037025002223#br2140>.
- [89] H. Zheng, T. Zhang, and J. Chen, “Fedfair: A debiasing algorithm for federated learning systems,” *Proceedings of the IEEE International Conference on Trust, Security and*

- Privacy in Computing and Communications, TrustCom*, pp. 2322–2326, 2024 2024, ISSN: 23249013. DOI: [10.1109/TRUSTCOM63139.2024.00320](https://doi.org/10.1109/TRUSTCOM63139.2024.00320).
- [90] M. Mohri, G. Sivek, and A. T. Suresh, “Agnostic federated learning,” *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 8114–8124, Feb. 2019. [Online]. Available: <https://arxiv.org/pdf/1902.00146>.
- [91] X. Peng, Z. Huang, Y. Zhu, and K. Saenko, “Federated adversarial domain adaptation,” *8th International Conference on Learning Representations, ICLR 2020*, Nov. 2019. [Online]. Available: <https://arxiv.org/pdf/1911.02054>.
- [92] S. V. Chinta, Z. Wang, A. Palikhe, *et al.*, “Ai-driven healthcare: A review on ensuring fairness and mitigating bias,” *PLOS Digital Health*, vol. 4, e0000864, 5 May 2025, ISSN: 2767-3170. DOI: [10.1371/JOURNAL.PDIG.0000864](https://doi.org/10.1371/JOURNAL.PDIG.0000864). [Online]. Available: <https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000864>.
- [93] I. G. B. Mahendra, “Fairness of ai: Review and recommendations,” *Journal of Public Health*, vol. 47, e220–e221, 2 May 2025, ISSN: 1741-3842. DOI: [10.1093/PUBMED/FDAE201](https://doi.org/10.1093/PUBMED/FDAE201). [Online]. Available: <https://dx.doi.org/10.1093/pubmed/fdae201>.
- [94] L. C. Bell and E. Shimron, “Sharing data is essential for the future of ai in medical imaging,” *Radiology: Artificial Intelligence*, vol. 6, e230337, 1 Jan. 2023, ISSN: 26386100. DOI: [10.1148/RYAI.230337](https://doi.org/10.1148/RYAI.230337). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10831510/>.
- [95] D. C. Castro, I. Walker, and B. Glocker, “Causality matters in medical imaging,” *Nature Communications 2020 11:1*, vol. 11, pp. 1–10, 1 Jul. 2020, ISSN: 2041-1723. DOI: [10.1038/s41467-020-17478-w](https://doi.org/10.1038/s41467-020-17478-w). [Online]. Available: <https://www.nature.com/articles/s41467-020-17478-w>.
- [96] J. Hu, Z. Yang, P. Wang, *et al.*, “Federated learning for medical image analysis: Privacy-preserving paradigms and clinical challenges,” 2025. DOI: [10.53941/tai.2025.100010](https://doi.org/10.53941/tai.2025.100010). [Online]. Available: <https://doi.org/10.53941/tai.2025.100010>.

- [97] X. Zeng, N. Abdullah, and P. Sumari, "Self-supervised learning framework application for medical image analysis: A review and summary," *BioMedical Engineering OnLine*, vol. 23, p. 107, 1 Dec. 2024, ISSN: 1475925X. DOI: [10.1186/S12938-024-01299-9](https://doi.org/10.1186/S12938-024-01299-9). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11514943/>.
- [98] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE transactions on bio-medical engineering*, vol. 69, p. 1173, 3 Mar. 2022, ISSN: 15582531. DOI: [10.1109/TBME.2021.3117407](https://doi.org/10.1109/TBME.2021.3117407). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9011180/>.
- [99] R. Schafer, T. Nicke, H. Hofener, *et al.*, "Overcoming data scarcity in biomedical imaging with a foundational multi-task model," *Nature Computational Science*, vol. 4, pp. 495–509, 7 Jul. 2024, ISSN: 26628457. DOI: [10.1038/S43588-024-00662-Z;SUBJMETA=117,308,639,692,705,794;KWRD=COMPUTER+SCIENCE,MEDICAL+RESEARCH,SOFTWARE](https://doi.org/10.1038/S43588-024-00662-Z;SUBJMETA=117,308,639,692,705,794;KWRD=COMPUTER+SCIENCE,MEDICAL+RESEARCH,SOFTWARE). [Online]. Available: <https://www.nature.com/articles/s43588-024-00662-z>.
- [100] T. Islam, M. S. Hafiz, J. R. Jim, M. M. Kabir, and M. F. Mridha, "A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions," *Healthcare Analytics*, vol. 5, p. 100340, Jun. 2024, ISSN: 2772-4425. DOI: [10.1016/J.HEALTH.2024.100340](https://doi.org/10.1016/J.HEALTH.2024.100340). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S277244252400042X>.
- [101] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review," *BMC Medical Imaging* 2022 22:1, vol. 22, pp. 1–13, 1 Apr. 2022, ISSN: 1471-2342. DOI: [10.1186/S12880-022-00793-7](https://doi.org/10.1186/S12880-022-00793-7). [Online]. Available: <https://bmcmmedimaging.biomedcentral.com/articles/10.1186/s12880-022-00793-7>.
- [102] P. Kora, C. P. Ooi, O. Faust, *et al.*, "Transfer learning techniques for medical image analysis: A review," *Biocybernetics and Biomedical Engineering*, vol. 42, pp. 79–107, 1 Jan. 2022, ISSN: 0208-5216. DOI: [10.1016/J.BBE.2021.11.004](https://doi.org/10.1016/J.BBE.2021.11.004). [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0208521621001297>.

- [103] M. R. Bujotzek, U. Akunal, S. Denner, *et al.*, “Real-world federated learning in radiology: Hurdles to overcome and benefits to gain,” *Journal of the American Medical Informatics Association*, vol. 32, pp. 193–205, 1 Jan. 2025, ISSN: 1527974X. DOI: [10.1093/JAMIA/OCAE259](https://doi.org/10.1093/JAMIA/OCAE259). [Online]. Available: <https://dx.doi.org/10.1093/jamia/ocae259>.
- [104] E. Darzidehkalani, M. Ghasemi-rad, and P. M. van Ooijen, “Federated learning in medical imaging: Part i: Toward multicentral health care ecosystems,” *Journal of the American College of Radiology*, vol. 19, pp. 969–974, 8 Aug. 2022, ISSN: 1546-1440. DOI: [10.1016/J.JACR.2022.03.015](https://doi.org/10.1016/J.JACR.2022.03.015). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1546144022002800>.
- [105] I. N. Tzortzis, A. Gutierrez-Torre, S. Sykiotis, *et al.*, “Towards generalizable federated learning in medical imaging: A real-world case study on mammography data,” *Computational and Structural Biotechnology Journal*, vol. 28, pp. 106–117, Jan. 2025, ISSN: 2001-0370. DOI: [10.1016/J.CSBJ.2025.03.031](https://doi.org/10.1016/J.CSBJ.2025.03.031). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S200103702500100X>.
- [106] E. Pachetti and S. Colantonio, “A systematic review of few-shot learning in medical imaging,” *Artificial Intelligence in Medicine*, vol. 156, p. 102949, Oct. 2024, ISSN: 0933-3657. DOI: [10.1016/J.ARTMED.2024.102949](https://doi.org/10.1016/J.ARTMED.2024.102949). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S093336572400191X>.
- [107] L. Alzubaidi, M. Al-Amidie, A. Al-Asadi, *et al.*, “Novel transfer learning approach for medical imaging with limited labeled data,” *Cancers*, vol. 13, 7 Apr. 2021, ISSN: 20726694. DOI: [10.3390/CANCERS13071590](https://doi.org/10.3390/CANCERS13071590). [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/33808207/>.
- [108] H. Jia, J. Zhang, K. Ma, X. Qiao, L. Ren, and X. Shi, “Application of convolutional neural networks in medical images: A bibliometric analysis,” *Quantitative Imaging in Medicine and Surgery*, vol. 14, pp. 3501–3518, 5 May 2024, ISSN: 22234306. DOI: [10.21037/QIMS-23-1600/COIF](https://doi.org/10.21037/QIMS-23-1600/COIF). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11074758/>.

- [109] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights into Imaging*, vol. 9, pp. 611–629, 4 Aug. 2018, ISSN: 18694101. DOI: [10.1007/S13244-018-0639-9](https://doi.org/10.1007/S13244-018-0639-9). [Online]. Available: <https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9>.
- [110] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, and M. J. Deen, "Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives," *Neurocomputing*, vol. 444, pp. 92–110, Jul. 2021, ISSN: 0925-2312. DOI: [10.1016/J.NEUCOM.2020.04.157](https://doi.org/10.1016/J.NEUCOM.2020.04.157). [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0925231221001314>.
- [111] A. Koivu, M. Sairanen, A. Airola, and T. Pahikkala, "Synthetic minority oversampling of vital statistics data with generative adversarial networks," *Journal of the American Medical Informatics Association*, vol. 27, pp. 1667–1674, 11 Nov. 2020, ISSN: 1527974X. DOI: [10.1093/JAMIA/OCAA127](https://doi.org/10.1093/JAMIA/OCAA127). [Online]. Available: <https://dx.doi.org/10.1093/jamia/ocaa127>.
- [112] A. Tucker, Z. Wang, Y. Rotalinti, and P. Myles, "Generating high-fidelity synthetic patient data for assessing machine learning healthcare software," *npj Digital Medicine* 2020 3:1, vol. 3, pp. 1–13, 1 Nov. 2020, ISSN: 2398-6352. DOI: [10.1038/s41746-020-00353-9](https://doi.org/10.1038/s41746-020-00353-9). [Online]. Available: <https://www.nature.com/articles/s41746-020-00353-9>.
- [113] B. Khosravi, F. Li, T. Dapamede, *et al.*, "Synthetically enhanced: Unveiling synthetic data's potential in medical imaging research," *eBioMedicine*, vol. 104, p. 105 174, Jun. 2024, ISSN: 2352-3964. DOI: [10.1016/J.EBIOM.2024.105174](https://doi.org/10.1016/J.EBIOM.2024.105174). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352396424002093>.
- [114] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://github.com/>.

- [115] S. Khan, M. Naseer, S. Khan, *et al.*, “Transformers in vision: A survey,” *ACM Computing Surveys (CSUR)*, vol. 54, 10 Sep. 2022, ISSN: 15577341. DOI: [10.1145/3505244](https://doi.org/10.1145/3505244). [Online]. Available: <https://dl.acm.org/doi/10.1145/3505244>.
- [116] D. Bahdanau, K. H. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Sep. 2014. [Online]. Available: <https://arxiv.org/abs/1409.0473v7>.
- [117] M. T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Aug. 2015. DOI: [10.18653/v1/d15-1166](https://arxiv.org/abs/1508.04025v5). [Online]. Available: <https://arxiv.org/abs/1508.04025v5>.
- [118] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 5999–6009, Jun. 2017, ISSN: 10495258. [Online]. Available: <https://arxiv.org/abs/1706.03762v7>.
- [119] P. Shaw, J. Uszkoreit, and A. Vaswani, “Self-attention with relative position representations,” *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 2, pp. 464–468, Mar. 2018. DOI: [10.18653/v1/n18-2074](https://arxiv.org/abs/1803.02155v2). [Online]. Available: <https://arxiv.org/abs/1803.02155v2>.
- [120] S. Wang, J. Sun, I. Mehmood, C. Pan, Y. Chen, and Y. D. Zhang, “Cerebral micro-bleeding identification based on a nine-layer convolutional neural network with stochastic pooling,” *Concurrency and Computation: Practice and Experience*, vol. 32, e5130, 1 Jan. 2020, ISSN: 1532-0634. DOI: [10.1002/CPE.5130](https://onlinelibrary.wiley.com/doi/10.1002/cpe.5130). [Online]. Available: <https://onlinelibrary.wiley.com/doi/full/10.1002/cpe.5130%20https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.5130%20https://onlinelibrary.wiley.com/doi/10.1002/cpe.5130>.
- [121] K. Choromanski, V. Likhoshesterov, D. Dohan, *et al.*, “Rethinking attention with performers,” *ICLR 2021 - 9th International Conference on Learning Representations*, Sep. 2020. [Online]. Available: <https://arxiv.org/abs/2009.14794v4>.

- [122] Y. Tay, D. Bahri, L. Yang, D. Metzler, and D. C. Juan, “Sparse sinkhorn attention,” *37th International Conference on Machine Learning, ICML 2020*, vol. PartF168147-13, pp. 9380–9389, Feb. 2020. [Online]. Available: <https://arxiv.org/abs/2002.11296v1>.
- [123] H. Touvron, M. Cord, M. Douze, *et al.*, “Training data-efficient image transformers & distillation through attention,” Dec. 2020, ISSN: 26403498. [Online]. Available: <https://arxiv.org/abs/2012.12877v2>.
- [124] M. Caron, H. Touvron, I. Misra, *et al.*, “Emerging properties in self-supervised vision transformers,” 2021. [Online]. Available: <https://github.com/facebookresearch/dino>.
- [125] S. Takahashi, Y. Sakaguchi, N. Kouno, *et al.*, “Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review,” *Journal of Medical Systems*, vol. 48, pp. 1–22, 1 Dec. 2024, ISSN: 1573689X. DOI: [10.1007/S10916-024-02105-8](https://doi.org/10.1007/S10916-024-02105-8). [Online]. Available: <https://link.springer.com/article/10.1007/s10916-024-02105-8>.
- [126] D. Ng, X. Lan, M. M. S. Yao, W. P. Chan, and M. Feng, “Federated learning: A collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets,” *Quantitative Imaging in Medicine and Surgery*, vol. 11, p. 852, 2 Feb. 2021, ISSN: 22234306. DOI: [10.21037/QIMS-20-595](https://doi.org/10.21037/QIMS-20-595). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7779924/>.
- [127] M. H. U. Rehman, W. H. L. Pinaya, P. Nachev, J. T. Teo, and M. J. Cardoso, “Federated learning for medical imaging radiology,” *The British Journal of Radiology*, vol. 96, p. 20220890, 1150 Oct. 2023, ISSN: 1748880X. DOI: [10.1259/BJR.20220890](https://doi.org/10.1259/BJR.20220890). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10546441/>.
- [128] F. Almalik, N. Alkhunaizi, I. Almakky, and K. Nandakumar, “Fesvibs: Federated split learning of vision transformer with block sampling,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14221 LNCS, pp. 350–360, Jun. 2023, ISSN: 16113349. DOI: [10.1007/978-3-031-43895-0_33](https://doi.org/10.1007/978-3-031-43895-0_33). [Online]. Available: <https://arxiv.org/abs/2306.14638v1>.

- [129] W. Dai, D. Li, D. Tang, H. Wang, and Y. Peng, “Deep learning approach for defective spot welds classification using small and class-imbalanced datasets,” *Neurocomputing*, vol. 477, pp. 46–60, 2022, ISSN: 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2022.01.004>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231222000042>.
- [130] G. Sambasivam and G. D. Opiyo, “A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks,” *Egyptian Informatics Journal*, vol. 22, pp. 27–34, 1 Mar. 2021, ISSN: 1110-8665. DOI: [10.1016/J.EIJ.2020.02.007](https://doi.org/10.1016/J.EIJ.2020.02.007).
- [131] C. Zhao, R. Shuai, L. Ma, W. Liu, and M. Wu, “Improving cervical cancer classification with imbalanced datasets combining taming transformers with t2t-vit,” *Multimedia Tools and Applications*, pp. 1–36, Mar. 2022, ISSN: 15737721. DOI: [10.1007/S11042-022-12670-0/TABLES/18](https://doi.org/10.1007/S11042-022-12670-0/TABLES/18). [Online]. Available: <https://link.springer.com/article/10.1007/s11042-022-12670-0>.
- [132] W. Feng, G. Dauphin, W. Huang, *et al.*, “Dynamic synthetic minority over-sampling technique-based rotation forest for the classification of imbalanced hyperspectral data,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, pp. 2159–2169, 7 Jul. 2019, ISSN: 21511535. DOI: [10.1109/JSTARS.2019.2922297](https://doi.org/10.1109/JSTARS.2019.2922297).
- [133] Y. Wang, L. Cui, C. Zhang, W. Chen, Y. Xu, and Q. Zhang, “A two-stage seismic damage assessment method for small, dense, and imbalanced buildings in remote sensing images,” *Remote Sensing 2022, Vol. 14, Page 1012*, vol. 14, p. 1012, 4 Feb. 2022, ISSN: 2072-4292. DOI: [10.3390/RS14041012](https://doi.org/10.3390/RS14041012). [Online]. Available: <https://www.mdpi.com/2072-4292/14/4/1012/htm%20https://www.mdpi.com/2072-4292/14/4/1012>.
- [134] J. Chen, H. Huang, A. G. Cohn, M. Zhou, D. Zhang, and J. Man, “A hierarchical dcnn-based approach for classifying imbalanced water inflow in rock tunnel faces,” *Tunnelling and Underground Space Technology*, vol. 122, p. 104 399, Apr. 2022, ISSN: 0886-7798. DOI: [10.1016/J.TUST.2022.104399](https://doi.org/10.1016/J.TUST.2022.104399).

- [135] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modelling under imbalanced distributions," *Association for Computing Machinery*, vol. 49, 2 May 2016. [Online]. Available: <https://arxiv.org/abs/1505.01658v2>.
- [136] Q. Dong, S. Gong, and X. Zhu, "Class rectification hard mining for imbalanced deep learning," *Proceedings of the IEEE International Conference on Computer Vision*, 2017.
- [137] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [138] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting methods for multi-class imbalanced data classification: An experimental review," *Journal of Big Data*, vol. 7, pp. 1–47, 1 Dec. 2020, ISSN: 21961115. DOI: [10.1186/S40537-020-00349-Y/FIGURES/5](https://doi.org/10.1186/S40537-020-00349-Y/FIGURES/5). [Online]. Available: <https://link.springer.com/articles/10.1186/s40537-020-00349-y%20https://link.springer.com/article/10.1186/s40537-020-00349-y>.
- [139] X. Yi, Y. Xu, Q. Hu, S. Krishnamoorthy, W. Li, and Z. Tang, "Asn-smote: A synthetic minority oversampling method with adaptive qualified synthesizer selection," *Complex & Intelligent Systems*, vol. 8, pp. 2247–2272, 3 Jun. 2022, ISSN: 2199-4536. DOI: [10.1007/s40747-021-00638-w](https://doi.org/10.1007/s40747-021-00638-w). [Online]. Available: <https://link.springer.com/10.1007/s40747-021-00638-w>.
- [140] R. I. Heroza, J. Q. Gan, and H. Raza, "Sia-smote: A smote-based oversampling method with better interpolation on high-dimensional data by using a siamese network," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14134 LNCS, pp. 448–460, 2023, ISSN: 1611-3349. DOI: [10.1007/978-3-031-43085-5_35](https://doi.org/10.1007/978-3-031-43085-5_35). [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-43085-5_35.
- [141] J. Bromley, I. Guyon, Y. Lecun, *et al.*, "Signature verification using a "siamese" time delay neural network," *Advances in Neural Information Processing Systems*, vol. 6, 1993.
- [142] D. S. Kermany, M. Goldbaum, W. Cai, *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, 1122–1131.e9, 5

- Feb. 2018, ISSN: 10974172. DOI: [10.1016/j.cell.2018.02.010](https://doi.org/10.1016/j.cell.2018.02.010). [Online]. Available: [http://www.cell.com/article/S0092867418301545/fulltext%20http://www.cell.com/article/S0092867418301545/abstract%20https://www.cell.com/cell/abstract/S0092-8674\(18\)30154-5](http://www.cell.com/article/S0092867418301545/fulltext%20http://www.cell.com/article/S0092867418301545/abstract%20https://www.cell.com/cell/abstract/S0092-8674(18)30154-5).
- [143] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, “A dataset for breast cancer histopathological image classification,” *IEEE Transactions on Biomedical Engineering*, vol. 63, pp. 1455–1462, 7 Jul. 2016, ISSN: 15582531. DOI: [10.1109/TBME.2015.2496264](https://doi.org/10.1109/TBME.2015.2496264).
- [144] W. Cagas, C. Ko, B. Hsiao, *et al.*, “Medical imaging complexity and its effects on gan performance,” Oct. 2024. [Online]. Available: <https://arxiv.org/pdf/2410.17959v1>.
- [145] Y. Skandarani, P. M. Jodoin, and A. Lalande, “Gans for medical image synthesis: An empirical study,” *Journal of Imaging*, vol. 9, 3 May 2021, ISSN: 2313433X. DOI: [10.3390/jimaging9030069](https://doi.org/10.3390/jimaging9030069). [Online]. Available: <https://arxiv.org/pdf/2105.05318>.
- [146] A. M. Anderson, M. Matsumoto, M. I. Saul, A. M. Secrest, and L. K. Ferris, “Accuracy of skin cancer diagnosis by physician assistants compared with dermatologists in a large health care system,” *JAMA Dermatology*, vol. 154, pp. 569–573, 5 May 2018, ISSN: 2168-6068. DOI: [10.1001/JAMADERMATOL.2018.0212](https://doi.org/10.1001/JAMADERMATOL.2018.0212).
- [147] M. Pacifico, R. Pearl, and R. Grover, “The uk government two-week rule and its impact on melanoma prognosis: An evidence-based study,” *The Annals of The Royal College of Surgeons of England*, vol. 89, no. 6, pp. 609–615, 2007. DOI: [10.1308/003588407x205459](https://doi.org/10.1308/003588407x205459).
- [148] L. Smith, N. Sansom, S. Hemphill, *et al.*, “Trends and variation in urgent referrals for suspected cancer 2009/2010–2019/2020,” *British Journal of General Practice*, vol. 72, no. 714, pp. 34–37, 2022. DOI: [10.3399/bjgp22X718217](https://doi.org/10.3399/bjgp22X718217).
- [149] C. Garbe, T. Amaral, K. Peris, *et al.*, “European consensus-based interdisciplinary guideline for melanoma. part 1: Diagnostics: Update 2022,” *European Journal of Cancer*, vol. 170, pp. 236–255, 2022. DOI: [10.1016/j.ejca.2022.03.008](https://doi.org/10.1016/j.ejca.2022.03.008).
- [150] N. Gessert, T. Sentker, F. Madesta, *et al.*, “Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting,” *IEEE Transactions on*

- Biomedical Engineering*, vol. 67, pp. 495–503, 2 Feb. 2020, ISSN: 15582531. DOI: [10.1109/TBME.2019.2915839](https://doi.org/10.1109/TBME.2019.2915839).
- [151] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *ICLR 2021 - 9th International Conference on Learning Representations*, Oct. 2020. [Online]. Available: <https://arxiv.org/abs/2010.11929v2>.
- [152] S. Ayas, “Multiclass skin lesion classification in dermoscopic images using swin transformer model,” *Neural Computing and Applications*, vol. 35, pp. 6713–6722, 9 Mar. 2023, ISSN: 14333058. DOI: [10.1007/S00521-022-08053-Z/TABLES/7](https://doi.org/10.1007/S00521-022-08053-Z/TABLES/7). [Online]. Available: <https://link.springer.com/article/10.1007/s00521-022-08053-z>.
- [153] C. Zhao, R. Shuai, L. Ma, W. Liu, D. Hu, and M. Wu, “Dermoscopy image classification based on stylegan and densenet201,” *IEEE Access*, vol. 9, pp. 8659–8679, 2021, ISSN: 21693536. DOI: [10.1109/ACCESS.2021.3049600](https://doi.org/10.1109/ACCESS.2021.3049600).
- [154] P. Tschandl and T. Wiesner, “Advances in the diagnosis of pigmented skin lesions,” *British Journal of Dermatology*, vol. 178, pp. 9–11, 1 Jan. 2018, ISSN: 13652133. DOI: [10.1111/BJD.16109;CTYPE:STRING:JOURNAL](https://doi.org/10.1111/BJD.16109;CTYPE:STRING:JOURNAL). [Online]. Available: [/doi/pdf/10.1111/bjd.16109%20https://onlinelibrary.wiley.com/doi/abs/10.1111/bjd.16109%20https://onlinelibrary.wiley.com/doi/10.1111/bjd.16109](https://doi.org/10.1111/bjd.16109%20https://onlinelibrary.wiley.com/doi/abs/10.1111/bjd.16109%20https://onlinelibrary.wiley.com/doi/10.1111/bjd.16109).
- [155] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dec. 2015, ISSN: 10636919. DOI: [10.48550/arxiv.1512.03385](https://doi.org/10.48550/arxiv.1512.03385). [Online]. Available: <https://arxiv.org/abs/1512.03385v1>.
- [156] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, Aug. 2016. DOI: [10.1109/CVPR.2017.243](https://doi.org/10.1109/CVPR.2017.243). [Online]. Available: <https://arxiv.org/pdf/1608.06993>.

- [157] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 10 691–10 700, May 2019. [Online]. Available: <https://arxiv.org/pdf/1905.11946>.
- [158] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in vision: A survey,” *ACM Computing Surveys*, vol. 54, 10 Jan. 2021, ISSN: 15577341. DOI: [10.1145/3505244](https://arxiv.org/pdf/2101.01169). [Online]. Available: <https://arxiv.org/pdf/2101.01169>.
- [159] X. Chen, S. Xie, and K. He, “An empirical study of training self-supervised vision transformers,” *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9620–9629, Apr. 2021, ISSN: 15505499. DOI: [10.1109/ICCV48922.2021.00950](https://arxiv.org/pdf/2104.02057). [Online]. Available: <https://arxiv.org/pdf/2104.02057>.
- [160] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, “Randaugment: Practical automated data augmentation with a reduced search space,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2020-June, pp. 3008–3017, Sep. 2019, ISSN: 21607516. DOI: [10.1109/CVPRW50498.2020.00359](https://arxiv.org/abs/1909.13719v2). [Online]. Available: <https://arxiv.org/abs/1909.13719v2>.
- [161] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Dec. 2018, ISSN: 10636919. DOI: [10.1109/CVPR.2018.00745](https://ieeexplore.ieee.org/document/8578843). [Online]. Available: <https://ieeexplore.ieee.org/document/8578843>.
- [162] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11211 LNCS, pp. 3–19, Jul. 2018, ISSN: 16113349. DOI: [10.1007/978-3-030-01234-2_1](https://arxiv.org/pdf/1807.06521). [Online]. Available: <https://arxiv.org/pdf/1807.06521>.
- [163] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” *ICLR 2021 - 9th International Conference on Learning Representations*, Oct. 2020. [Online]. Available: <https://arxiv.org/pdf/2010.04159>.

- [164] Z. Liu, Y. Lin, Y. Cao, *et al.*, “Swin transformer: Hierarchical vision transformer using shifted windows,” *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9992–10 002, Mar. 2021, ISSN: 15505499. DOI: [10.1109/ICCV48922.2021.00986](https://doi.org/10.1109/ICCV48922.2021.00986). [Online]. Available: <https://arxiv.org/pdf/2103.14030>.
- [165] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens, “Rethinking the inception architecture for computer vision,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.
- [166] “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic),” *Proceedings - International Symposium on Biomedical Imaging*, vol. 2018-April, pp. 168–172, May 2018, ISSN: 19458452. DOI: [10.1109/ISBI.2018.8363547](https://doi.org/10.1109/ISBI.2018.8363547).
- [167] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *7th International Conference on Learning Representations, ICLR 2019*, Nov. 2017. [Online]. Available: <https://arxiv.org/abs/1711.05101v3>.
- [168] T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” *Neural Computation*, vol. 10, pp. 1895–1923, 7 Oct. 1998, ISSN: 0899-7667. DOI: [10.1162/089976698300017197](https://doi.org/10.1162/089976698300017197).
- [169] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, *Communication-efficient learning of deep networks from decentralized data*, Apr. 2017. [Online]. Available: <https://proceedings.mlr.press/v54/mcmahan17a.html>.
- [170] Q. Li, Z. Wen, Z. Wu, *et al.*, “A survey on federated learning systems: Vision, hype and reality for data privacy and protection,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, pp. 3347–3366, 4 Jul. 2019. DOI: [10.1109/TKDE.2021.3124599](https://doi.org/10.1109/TKDE.2021.3124599). [Online]. Available: <http://arxiv.org/abs/1907.09693%20http://dx.doi.org/10.1109/TKDE.2021.3124599>.
- [171] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” *IEEE Signal Processing Magazine*, vol. 37, pp. 50–60, 3 Aug.

2019. DOI: [10.1109/MSP.2020.2975749](https://doi.org/10.1109/MSP.2020.2975749). [Online]. Available: <http://arxiv.org/abs/1908.07873><http://dx.doi.org/10.1109/MSP.2020.2975749>.
- [172] Z. L. Teo, L. Jin, S. Li, *et al.*, “Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture,” *Cell Reports Medicine*, vol. 5, p. 101419, 2 Feb. 2024, ISSN: 26663791. DOI: [10.1016/J.XCRM.2024.101419](https://doi.org/10.1016/J.XCRM.2024.101419). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10897620/>.
- [173] N. Rieke, J. Hancox, W. Li, *et al.*, “The future of digital health with federated learning,” *npj Digital Medicine* 2020 3:1, vol. 3, pp. 1–7, 1 Sep. 2020, ISSN: 2398-6352. DOI: [10.1038/s41746-020-00323-1](https://doi.org/10.1038/s41746-020-00323-1). [Online]. Available: <https://www.nature.com/articles/s41746-020-00323-1>.
- [174] K. Meduri, G. S. Nadella, A. R. Yadulla, *et al.*, “Leveraging federated learning for privacy-preserving analysis of multi-institutional electronic health records in rare disease research,” *Journal of Economy and Technology*, vol. 3, pp. 177–189, Nov. 2025, ISSN: 2949-9488. DOI: [10.1016/J.JECT.2024.11.001](https://doi.org/10.1016/J.JECT.2024.11.001). [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2949948824000544>.
- [175] S. Pati, S. Kumar, A. Varma, *et al.*, “Privacy preservation for federated learning in health care,” *Patterns*, vol. 5, p. 100974, 7 Jul. 2024, ISSN: 2666-3899. DOI: [10.1016/J.PATTER.2024.100974](https://doi.org/10.1016/J.PATTER.2024.100974).
- [176] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” Dec. 2018. [Online]. Available: <https://arxiv.org/abs/1812.06127v5>.
- [177] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, “Scaffold: Stochastic controlled averaging for federated learning,” *37th International Conference on Machine Learning, ICML 2020*, vol. PartF168147-7, pp. 5088–5099, Oct. 2019. [Online]. Available: <https://arxiv.org/abs/1910.06378v4>.
- [178] S. Mei, R. Jiang, J. Ji, J. Sun, Y. Peng, and Y. Zhang, “Invariant feature extraction for image classification via multi-channel convolutional neural network,” *2017 International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2017 -*

- Proceedings*, vol. 2018-January, pp. 491–495, Jul. 2017. DOI: [10.1109/ISPACS.2017.8266528](https://doi.org/10.1109/ISPACS.2017.8266528).
- [179] W. Lu, J. Wang, H. Li, Y. Chen, and X. Xie, “Domain-invariant feature exploration for domain generalization,” Jul. 2022. [Online]. Available: <https://arxiv.org/abs/2207.12020v2>.
- [180] C. Gao and W. Li, “An invariant feature extraction for multi-modal images matching,” Nov. 2023. [Online]. Available: <https://arxiv.org/abs/2311.02842v1>.
- [181] Y. Ganin, E. Ustinova, H. Ajakan, *et al.*, “Domain-adversarial training of neural networks,” *Journal of Machine Learning Research*, vol. 17, pp. 1–35, 2016.
- [182] A. Pakzad, K. Abhishek, and G. Hamarneh, “Circle: Color invariant representation learning for unbiased classification of skin lesions,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13804 LNCS, pp. 203–219, 2023, ISSN: 16113349. DOI: [10.1007/978-3-031-25069-9_14/FIGURES/5](https://doi.org/10.1007/978-3-031-25069-9_14/FIGURES/5). [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-25069-9_14.
- [183] A. Adebisi, N. Abdalnabi, E. J. Simoes, *et al.*, “Transformers in skin lesion classification and diagnosis: A systematic review,” *medRxiv*, p. 2024.09.19.24314004, Sep. 2024. DOI: [10.1101/2024.09.19.24314004](https://doi.org/10.1101/2024.09.19.24314004). [Online]. Available: <https://www.medrxiv.org/content/10.1101/2024.09.19.24314004v2%20https://www.medrxiv.org/content/10.1101/2024.09.19.24314004v2.abstract>.
- [184] R. I. Heroza, J. Q. Gan, and H. Raza, “Enhancing skin lesion classification: A self-attention fusion approach with vision transformer,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14860 LNCS, pp. 309–322, 2024, ISSN: 16113349. DOI: [10.1007/978-3-031-66958-3_23/TABLES/5](https://doi.org/10.1007/978-3-031-66958-3_23/TABLES/5). [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-66958-3_23.
- [185] M. Azeem, K. Kiani, T. Mansouri, and N. Topping, “Skinlesnet: Classification of skin lesions and detection of melanoma cancer using a novel multi-layer deep convolutional

- neural network,” *Cancers*, vol. 16, p. 108, 1 Jan. 2023, ISSN: 20726694. DOI: [10.3390/CANCERS16010108](https://doi.org/10.3390/CANCERS16010108). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10778045/>.
- [186] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, “Fedbn: Federated learning on non-iid features via local batch normalization,” *ICLR 2021 - 9th International Conference on Learning Representations*, Feb. 2021. [Online]. Available: <https://arxiv.org/abs/2102.07623v2>.
- [187] Y. Yan, H. Fu, Y. Li, *et al.*, “A simple data augmentation for feature distribution skewed federated learning,” Jun. 2023. [Online]. Available: <https://arxiv.org/abs/2306.09363v2>.
- [188] T. Zhou and E. Konukoglu, “Fedfa: Federated feature augmentation,” *11th International Conference on Learning Representations, ICLR 2023*, Jan. 2023. [Online]. Available: <https://arxiv.org/abs/2301.12995v1>.
- [189] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *32nd International Conference on Machine Learning, ICML 2015*, vol. 1, pp. 448–456, Feb. 2015. [Online]. Available: <https://arxiv.org/abs/1502.03167v3>.
- [190] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, “Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data,” *IEEE Transactions on Medical Imaging*, vol. 39, pp. 2713–2724, 9 Feb. 2020. DOI: [10.1109/TMI.2020.2974574](https://doi.org/10.1109/TMI.2020.2974574). [Online]. Available: <http://arxiv.org/abs/2002.03366> <http://dx.doi.org/10.1109/TMI.2020.2974574>.
- [191] M. Bencevic, M. Habijan, I. Galic, D. Babin, and A. Pizurica, “Understanding skin color bias in deep learning-based skin lesion segmentation,” *Computer Methods and Programs in Biomedicine*, vol. 245, p. 108 044, Mar. 2024, ISSN: 0169-2607. DOI: [10.1016/J.CMPB.2024.108044](https://doi.org/10.1016/J.CMPB.2024.108044).
- [192] G. Jung, S. Kim, and S. Yoo, “Skin tone analysis through skin tone map generation with optical approach and deep learning,” *Skin Research and Technology*, vol. 30,

- e70088, 10 Oct. 2024, ISSN: 16000846. DOI: [10.1111/SRT.70088](https://doi.org/10.1111/SRT.70088). [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11452249/>.
- [193] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, pp. 84–90, 6 May 2017, ISSN: 15577317. DOI: [10.1145/3065386](https://doi.org/10.1145/3065386). [Online]. Available: <https://dl.acm.org/doi/10.1145/3065386>.
- [194] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Dec. 2014. [Online]. Available: <https://arxiv.org/abs/1412.6980v9>.
- [195] H. Guan, P. T. Yap, A. Bozoki, and M. Liu, “Federated learning for medical image analysis: A survey,” *Pattern Recognition*, vol. 151, p. 110 424, Jul. 2024, ISSN: 0031-3203. DOI: [10.1016/J.PATCOG.2024.110424](https://doi.org/10.1016/J.PATCOG.2024.110424). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320324001754>.
- [196] R. I. Heroza, J. Q. Gan, and H. Raza, “Fedcial: Federated color-invariant adversarial learning for enhancing fairness and performance in skin lesion classification,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1812–1819, 2025, ISSN: 21607516. DOI: [10.1109/CVPRW67362.2025.00170](https://doi.org/10.1109/CVPRW67362.2025.00170).
- [197] T. Suzuki, C. Ito, K. Kitano, and T. Yamaguchi, “Cielab color space as a field for tracking color-changing chemical reactions of polymeric ph indicators,” *ACS Omega*, 2024, ISSN: 24701343. DOI: [10.1021/ACSOMEGA.4C05320](https://doi.org/10.1021/ACSOMEGA.4C05320). [Online]. Available: [/doi/pdf/10.1021/acsomega.4c05320](https://doi.org/10.1021/acsomega.4c05320).
- [198] M. A. R. Lara, C. Mosquera, E. Ferrante, and R. Echeveste, “Towards unraveling calibration biases in medical image analysis,” May 2023. DOI: [10.48550/arXiv.2305.05101](https://doi.org/10.48550/arXiv.2305.05101). [Online]. Available: <https://arxiv.org/pdf/2305.05101>.
- [199] J. Wang, Y. Zhang, Z. Ding, and J. Hamm, “Can domain adaptation improve accuracy and fairness of skin lesion classification?” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5157–5166, Jul. 2023, ISSN:

21607516. DOI: [10.1109/CVPRW63382.2024.00523](https://doi.org/10.1109/CVPRW63382.2024.00523). [Online]. Available: <https://arxiv.org/pdf/2307.03157>.
- [200] A. Pakzad, K. Abhishek, and G. Hamarneh, “Circle: Color invariant representation learning for unbiased classification of skin lesions,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13804 LNCS, pp. 203–219, Aug. 2022, ISSN: 16113349. DOI: [10.1007/978-3-031-25069-9_14](https://doi.org/10.1007/978-3-031-25069-9_14). [Online]. Available: <https://arxiv.org/abs/2208.13528v1>.
- [201] M. J. Sheller, B. Edwards, G. A. Reina, *et al.*, “Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data,” *Scientific reports*, vol. 10, 1 Dec. 2020, ISSN: 2045-2322. DOI: [10.1038/S41598-020-69250-1](https://doi.org/10.1038/S41598-020-69250-1). [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/32724046/>.
- [202] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” *Foundations and Trends in Theoretical Computer Science*, vol. 9, pp. 211–487, 3-4 2014, ISSN: 15513068. DOI: [10.1561/04000000042](https://doi.org/10.1561/04000000042). [Online]. Available: <https://dl.acm.org/doi/10.1561/04000000042>.
- [203] E. Hosseini, S. Chen, and A. Khisti, “Secure aggregation in federated learning using multiparty homomorphic encryption,” Mar. 2025. [Online]. Available: <https://arxiv.org/pdf/2503.00581>.
- [204] S. H. Wang, P. Phillips, Y. Sui, B. Liu, M. Yang, and H. Cheng, “Classification of alzheimer’s disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling,” *Journal of Medical Systems*, vol. 42, pp. 1–11, 5 May 2018, ISSN: 1573689X. DOI: [10.1007/S10916-018-0932-7/FIGURES/11](https://doi.org/10.1007/S10916-018-0932-7/FIGURES/11). [Online]. Available: <https://link.springer.com/article/10.1007/s10916-018-0932-7>.
- [205] Y. Ganin, E. Ustinova, H. Ajakan, *et al.*, “Domain-adversarial training of neural networks,” *Advances in Computer Vision and Pattern Recognition*, vol. 17, pp. 189–209, 9783319583464 May 2015, ISSN: 21916594. DOI: [10.1007/978-3-319-58347-1_10](https://doi.org/10.1007/978-3-319-58347-1_10). [Online]. Available: <https://arxiv.org/pdf/1505.07818>.