# WGAN-GP for Skin Lesion Synthesis and AI-Based Diagnosis

Sarya Demir
*Student Number 240488776*
*MSc. Artificial Intelligence, QMUL*
*s.demir@se24.qmul.ac.uk*
Project Supervisor: Dr. Salman Haleem
m.haleem@qmul.ac.uk

***Abstract*—Skin cancer classification remains as a challenging task because of scarce annotated datasets and heavy class imbalance. In this study, we address this problem by using Generative Adversarial Networks (GANs), namely the Wasserstein GAN with Gradient Penalty (WGAN-GP), to generate realistic skin lesion images and augment the ISIC Archive dataset. Feature representation is further improved by incorporating texture descriptors via Local Binary Patterns (LBP) and an ensemble of deep convolutional neural networks is utilized for robust classification. The resulting hybrid dataset, combining real and synthetic 4-channel (RGB + LBP) images, improves diversity, balance, and model generalizability. In addition, a Swift-based mobile application providing real-time lesion analysis was developed. Experimental results shows that proposed WGAN-GP + LBP + ensemble framework yields significantly improved classification accuracy, bringing clinically viable AI-supported skin cancer detection systems closer to reality.**

***Index Terms*—WGAN-GP, Skin Lesion Classification,Ensemble Model, Weighted Voting, Local Binary Pattern (LBP), SHAP, Grad-CAM, Mobile Application.**

## I. Introduction

SKIN cancer is among the most common cancers globally, with increasing incidence rates for both melanoma and non-melanoma types. The World Health Organization (WHO) estimates that 2–3 million non-melanoma and more than 132,000 melanoma skin cancer cases occur worldwide every year. A depletion of 10% in the ozone layer has been estimated to result in 300,000 new non-melanoma and 4,500 melanoma cases every year (World Health Organization 2017).

Early skin cancer detection is critical for successful treatment, and deep learning models have demonstrated great potential for automatic skin lesion classification. However, limited real world variability and dataset imbalance often impair model performance. Especially, malignant lesions are underrepresented with respect to benign lesions in the majority of datasets, which leads to bias and decreases the generalization capacity of classification models in clinical practice.

A promising solution to overcome this difficulty is data augmentation using synthetic image generation. GANs have been shown to be effective at generating realistic images in different medical imaging projects. However, standard GAN architectures often face mode collapse and unstable training dynamics, constraining their achievement of high-quality medical image generation (Saad et al. 2024).

To overcome these problems, this research uses the Wasserstein GAN with Gradient Penalty (WGAN-GP) (Gulrajani et al. 2017). The method aims at stabilizing the training of GANs and improving the diversity of the generated samples, which finally improves the dataset for robust skin lesion classification.

### A. Research Objectives

This study seeks to:

1) Generate synthetic skin lesion images using the WGAN-GP method to enhance dataset diversity and address class imbalance.
2) Create a hybrid dataset by combining synthetic and real images to improve classification outcomes.
3) Incorporate LBP texture analysis in preprocessing to enrich feature extraction.
4) Develop an ensemble-based classification model leveraging the augmented dataset for accurate benign versus malignant skin lesion differentiation.
5) Apply SHAP and Grad-CAM visualization techniques to analyze and explain model predictions to support transparency in potential clinical use.
6) Deploy the trained model into a Swift-based mobile application to enable real time skin lesion classification on user devices.

## II. Related Work

### A. Generative Models for Medical Image Generation

Because GANs are good at approximating complex data distributions, they have been employed extensively for the generation of medical images. For example, DCGAN's work (Radford et al. 2016) proved the power of convolutional GANs for unsupervised feature learning, and StyleGAN2 (Karras et al. 2019) was able to produce high quality and controllable images. However some models, tend to suffer from problems such as mode collapse and unstable training dynamics , specifically in medical application scenarios (Saad et al. 2024). To overcome these limitations, WGAN-GP (Gulrajani et al. 2017) was proposed, employing Wasserstein distance in addition to a gradient penalty term, instead of weight clipping, to allow training stability and enhance output diversity. In this work, WGAN-GP is used as the main framework for generating skin

lesion images due to its improved convergence rate and visual quality.

### B. Synthetic Data for Skin Lesion Classification

It is difficult to classify skin lesions as there is a scarcity of data and an imbalance of different classes in clinical datasets. Datasets such as the HAM10000 dataset (Tschandl et al. 2018) and ISIC dataset (International Skin Imaging Collaboration (ISIC) n.d.) has become a benchmark resource, yet challenges remain in achieving generalizable performance due to imbalance or low amounts of data. Prior studies such as Rashid et al. (2019) and Behara et al. (2023) have explored GAN-based data augmentation to address these issues, showing that synthetic samples can significantly improve classifier performance. This study extends these works by generating synthetic images using WGAN-GP and evaluating their effectiveness in classification and quantitative metrics .

### C. Texture Features in Medical Imaging

Texture features have been shown to be useful in dermoscopic image analysis, since they capture structural configurations that may not be apparent from color information only. The Local Binary Pattern (LBP) (Ojala et al. 1996) is a popular texture descriptor that defines changes in intensity at a pixel-by-pixel level. Akmalia et al. (2019) showed the success of LBP-augmented CNNs for dermatology applications. Building on these observations, the proposed method in this study incorporates LBP as an additional channel to enhance texture sensitivity in both the generator and classifier models.

### D. Deep CNNs and Ensemble Learning

In image classification tasks relevant to dermatology, deep convolutional networks like ResNet (He et al. 2016), EfficientNet (Tan and Le 2019), and GoogLeNet (Szegedy et al. 2015) have been used frequently (Suryakanth B. Ummapure 2023), (Venugopal et al. 2023). Combining these models using ensemble learning provides robustness. Harangi (2018) showed that ensembling CNNs increases diagnostic accuracy in skin lesion analysis. Inspired by this, in this work, a weighted ensemble approach was adopted to combine the predictions of three distinct CNNs in a way that stronger classifiers had more effective votes towards the final prediction.

Prior research by Gokalp and Tasci (2019) showed that using methods of optimization for enhancing the ensemble structures significantly improves the classification accuracy across a number of datasets. Encouraged by this, in this study, the weights are learned based on the performance of each model on validation, aiming to improve reliability for medical image classification tasks.

### E. Confidence Calibration and Explainability

Strong deployment of AI systems into medicine not only must be accurate but also must be well calibrated in terms of confidence (Sambyal et al. 2023). Temperature scaling, a post-hoc calibration technique, has been shown to reduce the overconfidence of neural networks (Guo et al. 2017). In this study, temperature scaling is applied to normalize probability outputs.

To address the medical image classification uncertainty and ensure clinical decision making to be more reliable, earlier research works used ROC-based thresholding techniques (Ghesu et al. 2021). Adopting the same principle, this study uses a confidence thresholding technique by ROC-AUC to reject low confidence predictions, which minimizes the chance of incorrect decisions in clinical environments (Ruopp et al. 2008). In addition, explainable AI methods like Grad-CAM (Selvaraju et al. 2017) and SHAP (Sree et al. 2024) are used to give clear visual and feature based explanations of what the model is predicting. These methods help verify that the models focus on lesion relevant regions rather than being misled by background artifacts.

### F. Ethical and Deployment Considerations

Driven by the desire to translate research into practical healthcare solutions, the efforts of Krohling et al. (2021) served as an inspiration. They created a smartphone application for skin cancer classification through deep learning and clinical data. Their focus on making an AI system that could be used as mobile and accessible tools made the promise of connecting academic studies with real world medical solutions even stronger. In line with this approach, a prototype of a real time mobile application was implemented to demonstrate the feasibility of the proposed system.

This project not only supports clinical decision making but also aligns with ethical principles outlined in major AI policy frameworks such as the UK's Data Ethics Framework (UK Government 2020), the Data Protection Act (2018) (UK Government 2018), the Council of Europe's CAHAI Elements (Council of Europe 2021), and the European AI Act (European Commission 2021). By integrating a mobile application for accessibility, SHAP and Grad-CAM for transparency, ROC-based thresholding for harm prevention, and ensuring human oversight and data privacy, the system reflects a responsible-by-design approach. These design choices aim to ensure fairness, safety, and accountability which are core requirements for trustworthy medical AI.

## III. Methodology

The methodology is structured into six core components: dataset and preprocessing, synthetic image generation using WGAN-GP, hybrid dataset construction, ensemble based classification, explainability analysis and Mobile application development. To visually summarize the entire methodology pipeline, a graphical abstract is presented in Figure 1

### A. Dataset and Preprocessing

Initially, the HAM10000 dataset (Tschandl et al. 2018) was considered for this study. However, due to some images being of low quality, the ISIC Archive dataset was ultimately selected. Comprising 2357 images of benign and malignant oncological diseases, the ISIC dataset, curated by the International Skin Imaging Collaboration (ISIC), offers a more

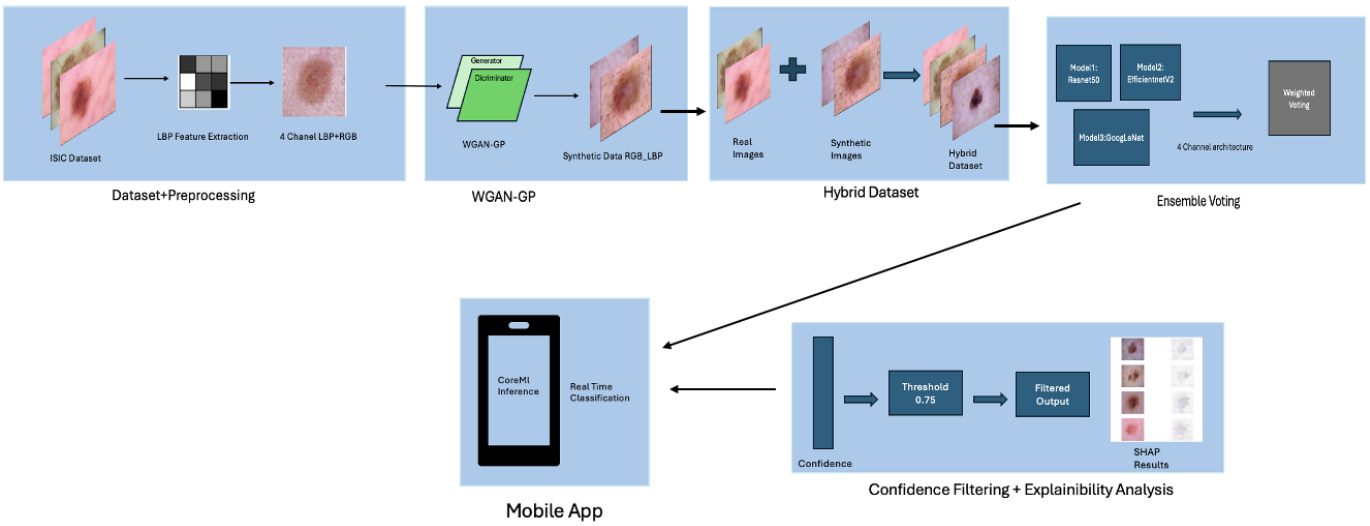


Figure 1: Graphical abstract summarizing methodology pipeline.

comprehensive and clinically relevant collection of annotated skin lesion images, which supports better model training and evaluation (Sahay 2022).

The dataset was divided into subsets for training, validation, and testing. The original training set was further split into a new training subset and a validation subset using an 80-20 ratio to guarantee representative class distributions. To provide an objective final evaluation benchmark, the test set was left unaltered during training.

Preprocessing steps were applied uniformly across all subsets. Images were resized to 224×224 pixels for classification tasks, providing sufficient resolution for Convolutional Neural Networks (CNNs) to extract meaningful lesion features. For training generative adversarial networks (GANs), images were resized to 64×64 pixels to strike a balance between computational efficiency and texture preservation.

A critical enhancement involved the extraction of texture features using Local Binary Pattern (LBP) analysis. Ojala et al. (1996) introduced the classical parameter configuration $P = 8$ (sampling points) and $R = 1$ (radius), which remains the most widely adopted setup in the literature due to its balance between computational efficiency and rotational invariance in a 3×3 neighborhood. This configuration has also been successfully applied in medical imaging and texture classification tasks by studies such as Akmalia et al. (2019) and García-Olalla et al. (2013). In this research, the grayscale LBP map was concatenated to the RGB channels to form a four channel representation, enriching the feature space with complementary texture and color information.

To address data scarcity and class imbalance, a total of 1000 malignant and 883 benign synthetic images were generated using a modified WGAN-GP framework (Gören and Çınarer 2023). Each generated image was stored as a NumPy array, including both RGB and LBP components as a 4 channel input. Prior to integration, all synthetic images were visually inspected and normalized. These synthetic samples were then merged with the original training data to construct a hybrid dataset, ensuring diversity and improving generalization capabilities.

Normalization strategies were tailored to the specific requirements of each model type. For GAN inputs, pixel values were scaled to the range $[-1, 1]$ to stabilize training dynamics. For CNN classifiers, pixel intensities were normalized to the $[0, 1]$ range to accelerate and stabilize convergence. This unified preprocessing strategy allowed both GAN-based generation and CNN-based classification pipelines to leverage complementary spatial and textural features effectively.

This careful dataset curation and preprocessing pipeline aimed to maximize model performance while preserving clinical relevance and data integrity.

### *B. Synthetic Image Generation using WGAN-GP*

In order to address data scarcity, class imbalance and to improve classification performances prevalent in medical imaging datasets, Wasserstein GAN with Gradient Penalty (WGAN-GP) was used as the generative model. WGAN-GP, which was first proposed by Gulrajani et al. (2017), stabilizes the training of GAN by applying the Lipschitz condition on the critic (discriminator) using a gradient penalty rather than weight clipping. This results in smoother convergence and less mode collapse.

$$\mathcal{L}_D = \underbrace{\mathbb{E}_{\tilde{\mathbf{x}}\sim\mathbb{P}_g}\left[D(\tilde{\mathbf{x}})\right] - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_r}\left[D(\mathbf{x})\right]}_{\text{critic loss}} + \underbrace{\lambda\,\mathbb{E}_{\hat{\mathbf{x}}\sim\mathbb{P}_{\hat{\mathbf{x}}}}\left[\left(\left\|\nabla_{\hat{\mathbf{x}}}D(\hat{\mathbf{x}})\right\|_2 - 1\right)^2\right]}_{\text{gradient penalty}} \tag{1}$$

Here, WGAN-GP was used because it provides better training dynamics than the classical variants of GAN, particularly in sensitive fields such as skin lesion synthesis where diversity and stable gradients are critical.

To adapt the architecture to the medical context, a few changes were implemented relative to the original WGAN-GP:

- The output of the generator was increased from 3 channels (RGB) to 4 channels (RGB + Local Binary Pattern) to increase texture based information that is vital in this study for dermatological examination.
- Input noise vector size was decreased from 128 to 100 from empirical experiments, which enhanced convergence stability in the environment.
- Instance normalization (InstanceNorm2d) was used in place of batch normalization in the discriminator, following Gulrajani et al. (2017)'s suggestion that batch normalization can conflict with gradient penalty mechanisms.
- The $\beta_1$ parameter of the Adam optimizer was changed to 0.5 (from 0.0), a change motivated by DCGAN and empirical practices, which provided smoother generator gradient flow in the initial stages of training in the experiments.

Table I: Key Architectural and Training Differences

| Component | Original WGAN-GP | Proposed Version |
|---|---|---|
| Input Noise Dim. | 128 | 100 |
| Generator Output | 3 (RGB) | 4 (RGB + LBP) |
| Normalization (D) | None | InstanceNorm2d |
| Optimizer $\beta_1$ | 0.0 | 0.5 |
| Normalization (G) | BatchNorm2d | BatchNorm2d |
| Training Epochs | Until convergence | 1000 |
| Input to Critic | $3 \times 64 \times 64$ | $4 \times 64 \times 64$ |

These changes maintained the theoretical structure of WGAN-GP but customized it to the particular requirements of this studies skin lesion synthesis.

*1) Training Configuration:* A batch size of 64, a latent dimension of 100, and an Adam optimizer with a learning rate of 0.0001 were used to train the WGAN-GP (Negi et al. 2020). To maintain stability, the generator was modified every five discriminator iterations. 10 was used as the gradient penalty coefficient. The training lasted 1000 epochs in total.

### C. Classification with Ensemble

*1) Ensemble Framework and Prediction Strategy:* Ensemble learning combines several classifiers to enhance robustness and generalization through model diversity, by variance reduction and correction of individual mistakes. (Dietterich 2000; Opitz and Maclin 1999). In this research, an ensemble of three convolutional neural networks (CNNs): EfficientNetV2, ResNet50, and GoogLeNet was used. All models were separately fine tuned on the same 4-channel dataset and outputs class probabilities.

To enhance the reliability of such outputs, *temperature scaling* , a post-hoc calibration technique introduced by Guo et al. (2017) is used. It adjusts softmax outputs by dividing the logit vector $\mathbf{z}_i$ by a scalar temperature $T > 0$, producing smoother probabilities:

$$\hat{p}_i^{(k)} = \frac{\exp(z_i^{(k)}/T)}{\sum_{j=1}^{K} \exp(z_i^{(j)}/T)} \tag{2}$$

Where $\hat{p}_i^{(k)}$ denotes the calibrated probability of class $k$ from model $i$.In this study, a fixed temperature $T = 1.5$ is used. Model weights are derived from validation accuracy as:

$$w_i = \frac{a_i}{\sum_{j=1}^{M} a_j} \tag{3}$$

Where $a_i$ is models validation accuracy and $\sum_{j=1}^{M} a_j$ is sum of validation accuracies.

Ensemble predictions are computed by taking a weighted average of the calibrated probabilities from all $M = 3$ models:

$$\bar{p}(x) = \sum_{i=1}^{M} w_i \cdot \hat{p}_i(x) \tag{4}$$

where $w_i$ is the normalized validation accuracy of model $i$, and $\sum w_i = 1$.

The final decision rule is defined as:

$$F(x) = \begin{cases} \arg\max_c \bar{p}_c(x), & \text{if } \max_c \bar{p}_c(x) \geq \tau \\ \text{rejected}, & \text{otherwise} \end{cases} \tag{5}$$

Where $\max_c \bar{p}_c(x)$ denotes the maximum predicted probability across all classes $c$, representing the model's confidence, and $\arg\max_c \bar{p}_c(x)$ returns the class label $c$ for which the predicted probability $\bar{p}_c(x)$ is highest.

This strategy integrates the strengths of each CNN and improves predictive stability. Although validation accuracies were close (EfficientNetV2: 89%, ResNet50: 89%, GoogLeNet: 91%), their prediction errors varied, justifying their complementary combination. Overall, the ensemble improved F1 score by over 9% compared to the best individual model, confirming the effectiveness of confidence-weighted voting (Gokalp and Tasci 2019).

*2) Confidence Thresholding via Youden Index:* To ensure that predictions are only made when the model is sufficiently confident, a *confidence-aware filtering* mechanism based on the Youden Index (Kallner 2018; Ruopp et al. 2008) is implemented. The Youden Index $J$ is a statistic for evaluating the performance of binary classifiers and is defined as:

$$J(c) = \text{Sensitivity}(c) + \text{Specificity}(c) - 1 \tag{6}$$

Where:

- Sensitivity (True Positive Rate) is defined as:

$$\text{Sensitivity}(c) = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

- Specificity (True Negative Rate) is defined as:

$$\text{Specificity}(c) = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{8}$$

Since the False Positive Rate (FPR) is $1 - \text{Specificity}$, the index can equivalently be written as:

$$J(c) = \text{True Positive Rate} - \text{False Positive Rate} \tag{9}$$

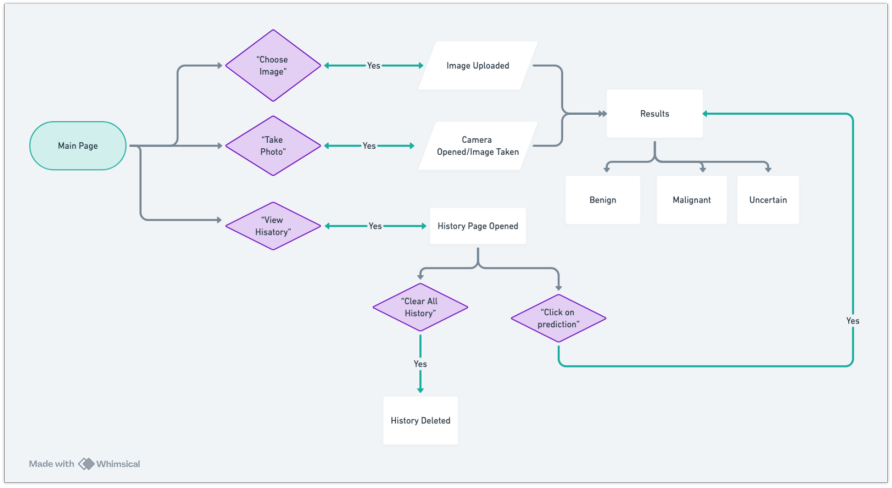


Figure 2: Flow diagram of the mobile application pipeline.

The threshold $c$ that maximizes $J(c)$ on the ROC curve represents the best trade-off between sensitivity and specificity. In this study optimal cutoff was found to be $\tau = 0.75$.

This threshold was then used in the Equation 5 predictions with confidence scores below this threshold were labeled as *uncertain* and excluded from evaluation metrics.

By preventing confident but inaccurate predictions, this filtering technique, when combined with temperature scaling and weighted ensemble voting, increased the system's precision and clinical reliability. In high stakes medical applications, it made sure that decisions were only made on cases that were properly confident.

### D. Explainability Analysis

SHAP (SHapley Additive exPlanations) was applied using GradientExplainer on the ResNet50 model to determine pixel level feature importance. Grad-CAM was also used for highlighting the discriminative regions in lesion classification (Selvaraju et al. 2017).

### E. Mobile Application Deployment

A mobile application was developed to bridge the gap between the deep learning algorithm development and practical clinical deployment (Krohling et al. 2021) using Swift and Apple's Core ML framework. The three trained models (ResNet50, GoogLeNet, and EfficientNetV2) were converted into .mlmodel format compatible with Core ML, allowing efficient on device inference without the need for internet connectivity. This facilitates real time skin lesion classification directly on the user's iOS device.

The application provides a user interface where clinicians or users can upload skin lesion images easily. Upon image submission, the app performs ensemble inference by taking predictions from the three models, leveraging weighted voting combined with temperature scaled confidence calibration. Predictions with confidence scores below a 0.75 threshold trigger a low confidence alert, advising users to seek professional medical advice, thus enhancing the safety and reliability of the tool in real world clinical scenarios.

Figure 2 shows the flow diagram of the developed mobile application. This pipeline includes image upload, real time ensemble classification, confidence aware filtering, and historical result management. The application integrates the trained ensemble model and ensures user friendly operation in clinical settings.

## IV. Results

### A. Dataset Composition

The original training set (2109 real images) was augmented with synthetic images produced using the proposed WGAN-GP model in order to enhance the diversity and size. This produced a better balanced and enriched dataset for classifier training, with a final training set of 3992 images (2035 benign, 1957 malignant).

### B. Impact of GAN-Based Data Augmentation

Two training settings, actual data only (No GAN) and real data added using GAN synthetic images, were used in a comparative experiment to evaluate the impact of synthetic data augmentation (Table II). Both processes employed same architecture (ensemble of EfficientNetV2, ResNet50, GoogLeNet), preprocessing , and assessment parameters.

Table II: Effect of GAN augmentation on classification performance

| Setting | Accuracy | F1 Score |
|---|---|---|
| No GAN | 84.39% | 84.37% |
| WGAN-GP (RGB + LBP) | >89% | >88% |
| StyleGAN | >89% | >89% |
| DCGAN | >89% | >89% |

The inclusion of augmented samples improved classification performances, accuracy and F1 score.

### C. Comparison of GAN-based Data Augmentation Strategies

Four GAN architectures were quantitavely compared: Original WGAN-GP proposed by Gulrajani et al. (2017) Proposed WGAN-GP, DCGAN, and StyleGAN, using four metrics: Jensen-Shannon Divergence (JSD), Wasserstein Distance, Distance Correlation (dCor), and Fréchet Inception Distance (FID). These metrics evaluate both distributional similarity and visual fidelity between real and synthetic samples (Table III).

Among the three, proposed WGAN-GP outperformed the others across all metrics. It achieved the lowest JSD (0.2210) and lowest Wasserstein Distance (0.0242), indicating the best alignment with the real data distribution. It also had the lowest FID score (160.16), suggesting higher perceptual quality compared to StyleGAN and DCGAN. Although StyleGAN and Original WGAN-GP showed moderate performance, DCGAN lagged behind significantly, especially in terms of JSD and FID.

Table III: Quantitative comparison of GAN architectures: Jensen-Shannon Divergence (JSD), Wasserstein Distance, Distance Correlation (dCor), and Fréchet Inception Distance (FID).

| GAN Type | JSD | Wasserstein | dCor | FID |
|---|---|---|---|---|
| Original WGAN-GP | 0.2263 | 0.0466 | 0.0605 | 415.18 |
| Proposed WGAN-GP | 0.2210 | 0.0242 | 0.0351 | 160.16 |
| DCGAN | 0.8579 | 0.2031 | 0.0367 | 441.96 |
| StyleGAN | 0.2875 | 0.0402 | 0.0440 | 172.75 |

### D. Visual Comparison with Other GAN Architectures

To highlight the impact of GAN architecture choice, a limited number of synthetic images using DCGAN and StyleGAN2 generated for qualitative comparison Figure 3. DCGAN produced low-resolution outputs with mode collapse and repetitive structures. Background textures appeared noisy and unrealistic. Although the images produced by StyleGAN2 looks promising, quantitative comparison showed that images produces by WGAN-GP were superior. WGAN-GP produced diverce and realistic samples, effectively balancing global structure.

### E. Impact of LBP Channel on Classification Accuracy

To evaluate the effect of texture enhanced inputs, each CNN architecture was trained under two conditions: with RGB only inputs and with RGB + LBP inputs. Results are provided in Table IV.

Table IV: Effect of LBP Channel on Accuracy Across Models

| Model | RGB Only (3 channels) | RGB + LBP (4 channels) |
|---|---|---|
| ResNet50 | ∼86.9% | **89.0%** |
| EfficientNetV2 | ∼87.2% | **89.0%** |
| GoogLeNet | ∼88.5% | **91.0%** |
| ViT | ∼83.5% | **84.0%** |

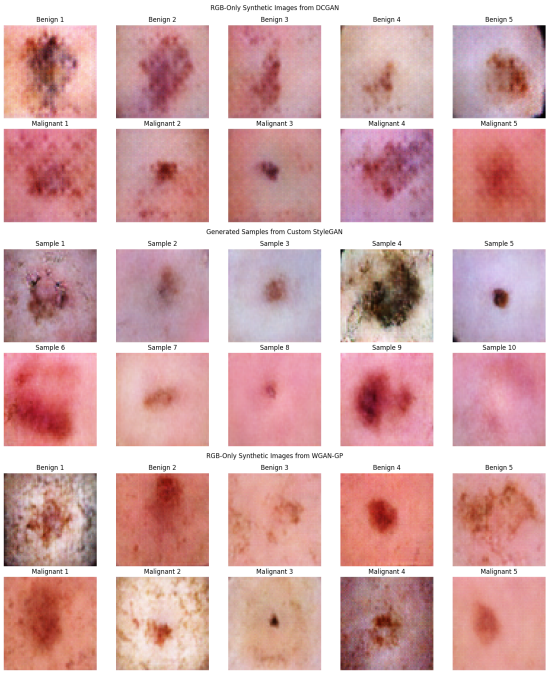


Figure 3: Comparison of synthetic samples from DCGAN (top), StyleGAN2 (middle), and WGAN-GP (bottom).

### F. Ensemble Performance and Optimization

In order to maximize classification performance, a number of ensemble techniques and confidence thresholds were evaluated. As explained in Section III, the final decision process used confidence aware weighted voting, where predictions with a confidence level of less than 0.75 were rejected after calibrated probability scores from each model were averaged using validation based weights.

This setup produced the highest F1 score, as Table V demonstrates. The system's clinical reliability was enhanced and false positives were decreased by rejecting low-confidence predictions.

Table V: Effect of Ensemble Strategies and Confidence Thresholding on Classification Performance

| Method | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Baseline (No Voting, No Filtering) | ∼89.09% | 84.55% | 93.00% | 88.57% |
| Majority Voting | 89.09% | 84.55% | 93.00% | 88.57% |
| Weighted Voting (0.33/0.33/0.34) | 89.39% | 84.85% | 93.33% | 88.89% |
| Weighted Voting + Filtering (Threshold 0.65) | 93.40% | 90.24% | 95.93% | 93.00% |
| Weighted Voting + Filtering (Threshold 0.70) | 94.93% | 92.00% | 97.31% | 94.58% |
| Fine-Tuned Weights + Threshold 0.70 | 94.93% | 92.00% | 97.31% | 94.58% |
| **Improved Weighted Voting (Threshold 0.75)** | **96.98%** | **94.96%** | **98.79%** | **96.84%** |

### G. Hair Removal Experiments

Initially several hair removal methods to remove occluding hair artifacts were investigated to enhance input image quality. Every technique, showed some levels of issues that impacted lesion visibility or texture preservation negatively. Table VI lists the noticed restrictions.

Table VI: Comparison of Hair Removal Methods and Their Limitations

| Method | Problem Observed |
|---|---|
| VAE | Blurred both hair and lesion features |
| DAE | Minor hair removal but significant texture loss |
| Gaussian Blur | Smoothed out hair but removed lesion edges |
| Classical Contour | Lost lesion shape, distorted fine details |

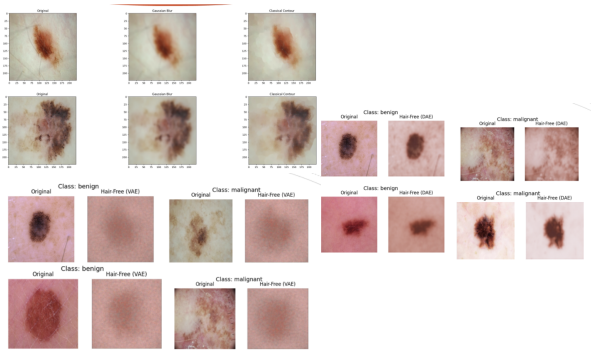

Figure 4: Visual outputs from various hair removal methods (VAE, DAE, Gaussian Blur, Classical Contour). Original and processed samples are shown side-by-side.

Based on both qualitative outputs (Figure 4) and due to decreased performance, all hair removal methods were excluded from the final pipeline.

### H. Classification Performance

The final ensemble classifier, which combines EfficientNetV2, ResNet50, and GoogLeNet using temperature scale, weighted voting, achieved strong results on the test set. Performance metrics were calculated only for predictions that passed the confidence threshold ($\geq 0.75$):

- **Accuracy:** 96.98%
- **Precision:** 94.96%
- **Recall:** 98.79%
- **F1 Score:** 96.84%
- **Confident Predictions:** 530 out of 660

Predictions falling below the threshold were labeled as "uncertain" and excluded from final metric calculations to ensure clinical reliability.

### I. Confidence Thresholding

In line with the application's clinical safety objective a confidence threshold of 0.75 was determined based on the ROC validation analysis to reduce overconfident classifications. This strategy enabled the system to identify low confidence predictions for human review (Fig. 5).

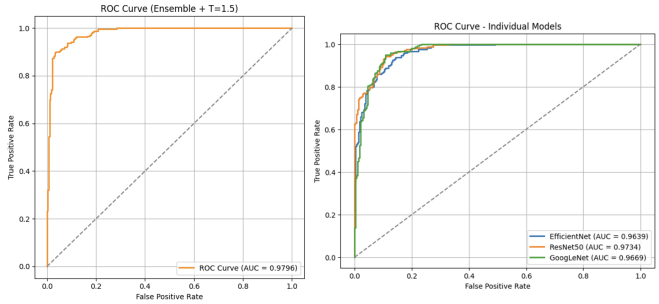


Figure 5: ROC curve of the ensemble model with temperature scaling (T=1.5), and ROC curve of the hree classifiers used in the ensemble training.

### J. Explainability with SHAP and Grad-CAM

The SHAP (SHapley Additive explanations) was tested on the ResNet50 model in the ensemble to learn more about the process of decision making. For every prediction, pixel-level feature attribution was given by SHAP values. Examples of comparing input dermoscopic images (left) with SHAP overlays (right) are shown in Fig. 6.

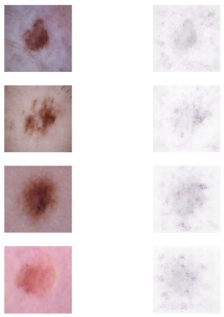

Figure 6: SHAP explanations of model predictions.

As shown, the model usually focuses on the edges of lesions. Non lesion artifacts occasionally affected the prediction.

Additionally, Grad-CAM (Gradient-weighted Class Activation Mapping) was used to better understand the CNN-based decision process. Grad-CAM emphasized areas of the image that had a significant influence on the final classification. Grad-CAM overlays are shown in Fig. 7 for both benign and malignant examples.

The images show that the model looks at clinically important details like uneven edges, color changes, and asymmetry in malignant cases.

### K. Mobile Application Outputs

To demonstrate the practical implementation and user experience of the developed mobile application, screenshots from the app are provided in Figure 8. The application interface

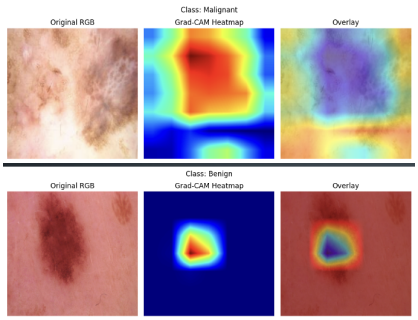


Figure 7: Grad-CAM visualizations. Left: Malignant lesion with focused attention on irregular structure. Right: Benign lesion with attention over smooth central region.

includes functionality for image upload, ensemble-based prediction, and uncertainty-aware warnings. These outputs ensure that the AI-based system can be used confidently in real clinical settings.

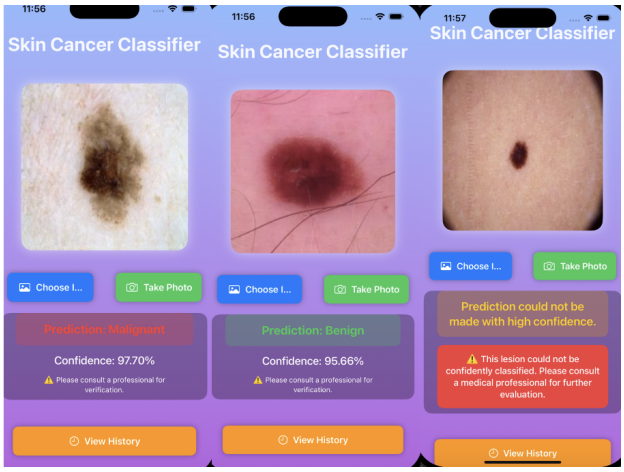


Figure 8: Prediction result screen showing classification outputs malignant, benign, uncertain with confidence level.

## V. Discussion

### A. Impact of GAN-Driven Data Augmentation and Classification

This research proved that WGAN-GP is a powerful generative model for dermoscopic image generation, overcoming issues with GAN instability reported in previous work (Saad et al. 2024). In comparison with reference architectures like DCGAN (Radford et al. 2016) and StyleGAN2 (Karras et al. 2019), WGAN-GP showed better divergence and image quality, evidenced by quantitative measures (Table III), consistent with theoretical benefits. The use of WGAN-GP in data augmentation achieved improved classifier accuracy and F1 scores, confirming that stable GANs with better distributional alignment can augment training diversity and classifier generalizability (Rashid et al. 2019).

### B. Learning with LBP Features Using Texture-Aware

Texture descriptors such as Local Binary Patterns (LBP) (Ojala et al. 1996) were added as a fourth input channel to encode lesion surface patterns and structural detail. This is based on previous work (Akmalia et al. 2019) of the effectiveness of LBP in dermatological CNN pipelines. Findings (Table IV) indicate that RGB+LBP inputs improved classification accuracy across all architectures, most notably GoogLeNet and ResNet50, confirming LBP's contribution to improved texture sensitivity of deep learning models (Rashid et al. 2019).

### C. Ensemble Learning for Enhanced Classification Robustness

The ensemble of EfficientNetV2 (Tan and Le 2019), ResNet50 (He et al. 2016), and GoogLeNet (Szegedy et al. 2015) performed better than any of the models alone, as supported by Harangi (2018) for CNN ensembles. Use of validation accuracy based weights (Gokalp and Tasci 2019) further contributed to reliability, with a final accuracy of 96.98%. This confirms that the use of architectural diversity reduces the variance of predictions and increases robustness, as necessary for high risk clinical settings.

### D. Confidence Calibration and Filtering

For clinical reliability, temperature scaling (Guo et al. 2017) and ROC-based thresholding was introduced (Ghesu et al. 2021) (Kallner 2018) (Ruopp et al. 2008) for the detection of uncertain predictions. By removing predictions that had a confidence level below 0.75, model maintained high recall and precision levels while prioritizing cautious decision making. This is in line with clinical AI system standards, where maintaining high specificity is central to minimizing diagnostic risk (Sambyal et al. 2023).

### E. Explainability and Trust

To assess transparency , Grad-CAM (Selvaraju et al. 2017) and SHAP (Sree et al. 2024) were employed. These techniques confirmed that model attention was high in areas relevant to the lesion and mostly not in distracting background artifacts. These findings provide evidence of interpretability aligned with best practices for explainable AI in clinical applications.

### F. Mobile Apps and Accessibility

Building on the findings of Krohling et al. (2021) which established the efficacy of mobile devices in skin diagnosis, a prototype mobile app with integrated Core ML functionality was created. This development allows practical application, particularly in areas with limited resources, thereby upholding the principles of accessibility and fairness promoted in the AI Act (European Commission 2021).

## VI. Conclusion

This paper presents an end to end AI pipeline for skin lesion classification integrating WGAN-GP driven synthetic data augmentation, texture-sensitive feature enrichment through LBP, ensemble learning via CNNs, confidence-aware prediction filtering, and mobile deployment for real world usability.

Key conclusions drawn from the findings are:

- **WGAN-GP** produced higher quality and more diverse synthetic images than prior GANs, boosting classification performance and resolving common stability issues
- **LBP-enhanced inputs** significantly improved classification performance in all CNN designs by providing important texture related information required for dermatological analysis, thus proving the relevance of structural features along with RGB information.
- **Ensemble learning**, using EfficientNetV2, ResNet50, and GoogLeNet with validation-weighted voting, reduced prediction variance and improved robustness, important for minimizing error in clinical settings.
- **Confidence calibration and ROC-based thresholding** enhanced clinical safety through the rejection of low confidence predictions and the prevention of overconfident misclassifications, aligned with best practices for medical AI deployment.
- **Mobile deployment** with Core ML enabled real time, offline on device classification, which improved accessibility and low resource clinical environments, bridging research to real world application.

This approach of employing data augmentation via WGAN-GP, texture-aware feature creation, ensemble techniques, calibrated confidence, and mobile accessibility presents a promising solution for secure and precise classification of skin lesions under real world clinical conditions.

## VII. Future Work

While the proposed system had high classification accuracy, several directions can be considered that can enhance its clinical utility and robustness:

- **Dataset Enrichment:** Incorporating more diverse datasets in terms of skin tones, lesions, and environments would increase fairness and generalization.
- **Next Generation Generative Models:** Exploring next generation architectures like StyleGAN3 or diffusion models may yield more realistic synthetic samples for underrepresented lesions.
- **Clinical Validation of Synthetic Data:** Official assessment by dermatologists needs to be performed to establish the diagnostic importance of images generated through GAN.
- **Improved Ensemble Methods:** Ensembling a bigger collection of architectures or hybrid models can reduce error variance more and improve resilience.
- **Optimizing Mobile Deployment:** Enhancing cross-platform compatibility and secure data handling will enhance real world usage.
- **Ethical Compliance:** Future research needs to tackle more of fairness, transparency, and privacy issues in order to satisfy regulatory requirements in healthcare AI.

The purpose of these recommendations is to bridge the gap between safe, fair clinical practices and evidences.


## Acknowledgment

I would like to express my gratitude to my supervisor, Dr. Salman Haleem, for his encouragement and insightful critique during this study. His assistance and encouragement have been invaluable.

I also want to thank my family and friends for their continuous support, understanding, and tolerance during this time. Their faith in me kept me going even when things were tough.

Finally, I want to thank Queen Mary University of London for providing the resources and stimulating academic environment that made this study possible.


## References


Akmalia, Nurul et al. (2019). "Skin Diseases Classification Using Local Binary Pattern and Convolutional Neural Network". In: *2019 3rd International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, pp. 168–173. DOI: 10.1109/ELTICOM47379.2019.8943892.

Behara, Kavita et al. (Aug. 2023). "Skin Lesion Synthesis and Classification Using an Improved DCGAN Classifier". In: *Diagnostics* 13.16. This article belongs to the Special Issue Deep Disease Detection and Diagnosis Models, p. 2635. DOI: 10.3390/diagnostics13162635. URL: https://doi.org/10.3390/diagnostics13162635.

Council of Europe (2021). *CAHAI(2021)09rev – Elements of a legal framework on Artificial Intelligence*. https://rm.coe.int/cahai-2021-09rev-final-eng/1680a4f17d. Accessed: 2025-07-10.

Dietterich, Thomas G. (2000). "Ensemble methods in machine learning". In: *Multiple Classifier Systems*. Ed. by Josef Kittler and Fabio Roli. Accessed: 2025-05-30. Springer, pp. 1–15. URL: https://link.springer.com/chapter/10.1007/3-540-45014-9_1.

European Commission (2021). *Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act)*. Accessed: 2025-07-10. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

García-Olalla, Oscar et al. (Dec. 2013). "Adaptive local binary pattern with oriented standard deviation (ALBPS) for texture Classification". In: *EURASIP Journal on Image and Video Processing* 2013. DOI: 10.1186/1687-5281-2013-31.

Ghesu, Florin C. et al. (Feb. 2021). "Quantifying and leveraging predictive uncertainty for medical image assessment". In: *Medical Image Analysis* 68, p. 101855. DOI: 10.1016/j.media.2020.101855.

Gokalp, Osman and Erdal Tasci (2019). "Weighted Voting Based Ensemble Classification with Hyper-parameter Optimization". In: *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*. IEEE, pp. 1–6. DOI: 10.1109/IDAP46976.2019.9080164.

Gören, E. and G. Çınarer (2023). "Cancer Lesion Classification with GAN-Based Image Augmentation Method from Skin Images". In: *International Conference on Engineering, Natural and Social Sciences*. Vol. 1. Accessed: 27 June 2025, pp. 658–666. URL: https://as-proceeding.com/index.php/icensos/article/view/521.

Gulrajani, Ishaan et al. (2017). "Improved training of Wasserstein GANs". In: *arXiv preprint arXiv:1704.00028*.

Guo, Chuan et al. (2017). "On Calibration of Modern Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. Accessed: 2025-05-30. URL: https://arxiv.org/abs/1706.04599.

Harangi, Balazs (2018). "Skin lesion classification with ensembles of deep convolutional neural networks". In: *Journal of Biomedical Informatics* 86, pp. 25–32. DOI: 10.1016/j.jbi.2018.08.006.

He, Kaiming et al. (2016). "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. URL: https://arxiv.org/abs/1512.03385.

International Skin Imaging Collaboration (ISIC) (n.d.). *ISIC Archive Dataset*. https://www.isic-archive.com. Accessed: 2025-05-30.

Kallner, Anders (2018). "Formulas". In: *Laboratory Statistics (Second Edition)*. Ed. by Anders Kallner. Second Edition. Elsevier, pp. 1–140. ISBN: 978-0-12-814348-3. DOI: https://doi.org/10.1016/B978-0-12-814348-3.00001-0. URL: https://www.sciencedirect.com/science/article/pii/B9780128143483000010.

Karras, Tero et al. (2019). "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Accessed: 2025-05-30. URL: https://arxiv.org/abs/1812.04948.

Krohling, Breno et al. (2021). *A Smartphone based Application for Skin Cancer Classification Using Deep Learning with Clinical Images and Lesion Information*. arXiv preprint arXiv:2104.14353. arXiv: 2104.14353 [eess.IV]. URL: https://doi.org/10.48550/arXiv.2104.14353.

Negi, Anuja et al. (2020). "RDA-UNET-WGAN: An Accurate Breast Ultrasound Lesion Segmentation Using Wasserstein Generative Adversarial Networks". In: *Arabian Journal for Science and Engineering* 45. Research Article - Electrical Engineering, pp. 6399–6410. DOI: 10.1007/s13369-020-04480-z. URL: https://doi.org/10.1007/s13369-020-04480-z.

Ojala, Timo et al. (1996). "A comparative study of texture measures with classification based on featured distributions". In: *Pattern Recognition* 29.1, pp. 51–59. ISSN: 0031-3203. DOI: https://doi.org/10.1016/0031-3203(95)00067-4. URL: https://www.sciencedirect.com/science/article/pii/0031320395000674.

Opitz, David W. and Richard Maclin (1999). "Popular ensemble methods: An empirical study". In: *Journal of Artificial Intelligence Research* 11. Accessed: 2025-05-30, pp. 169–198. URL: https://www.d.umn.edu/~rmaclin/publications/opitz-jair99.pdf.

Radford, Alec et al. (2016). "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434*. Accessed: 2025-05-30. URL: https://arxiv.org/abs/1511.06434.

Rashid, Haroon et al. (2019). "Skin Lesion Classification Using GAN based Data Augmentation". In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 916–919. DOI: 10.1109/EMBC.2019.8857905.

Ruopp, Michael D et al. (2008). "Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection". In: *Biometrical Journal* 50.3, pp. 419–430. DOI: 10.1002/bimj.200710415.

Saad, Muhammad Muneeb et al. (2024). "A survey on training challenges in generative adversarial networks for biomedical image analysis". In: *Artificial Intelligence Review* 57.19. DOI: 10.1007/s10462-023-10624-y. URL: https://doi.org/10.1007/s10462-023-10624-y.

Sahay, Rajiv (2022). *ISIC Skin Cancer Dataset*. https://www.kaggle.com/datasets/rajivaiml/isic-skin-cancer-dataset. Accessed: 2025-07-03.

Sambyal, Abhishek Singh et al. (Dec. 2023). "Understanding calibration of deep neural networks for medical image classification". In: *Computer Methods and Programs in Biomedicine* 242, p. 107816. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2023.107816. URL: http://dx.doi.org/10.1016/j.cmpb.2023.107816.

Selvaraju, Ramprasaath R. et al. (2017). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

Sree, Nichenametla Hima et al. (2024). "Explainable AI Insights into Skin Cancer Detection: A Comparative Study of CNN, DenseNet, and ResNet". In: *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, pp. 1–8. DOI: 10.1109/I2CT61223.2024.10543490.

Suryakanth B. Ummapure Ravindrakumar Tilekar, Satishkumar Mallappa (Jan. 2023). "Skin Cancer Classification using VGG-16 and Googlenet CNN Models". In: *International Journal of Computer Applications* 184.42, pp. 5–9. ISSN: 0975-8887. DOI: 10.5120/ijca2023922497. URL: https://ijcaonline.org/archives/volume184/number42/32587-2023922497/.

Szegedy, Christian et al. (2015). "Going Deeper with Convolutions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9. URL: https://arxiv.org/abs/1409.4842.

Tan, Mingxing and Quoc Le (2019). “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. URL: https://arxiv.org/abs/1905.11946.

Tschandl, Philipp et al. (2018). “The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions”. In: *Scientific Data* 5.180161. Accessed: 2025-05-30. URL: https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000.

UK Government (2018). *Data Protection Act 2018*. https://www.legislation.gov.uk/ukpga/2018/12/contents. Accessed: 2025-07-10.

– (2020). *Data Ethics Framework*. https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-2020. Accessed: 2025-07-10.

Venugopal, Vipin et al. (2023). “A deep neural network using modified EfficientNet for skin cancer detection in dermoscopic images”. In: *Decision Analytics* 10, p. 100278. DOI: 10.1016/j.dajour.2023.100278. URL: https://www.researchgate.net/publication/372139502.

World Health Organization (2017). *Radiation: Ultraviolet (UV) radiation and skin cancer*. https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-(uv)-radiation-and-skin-cancer. Accessed: 2025-07-09. World Health Organization.