GraphDerm: Fusing Imaging, Physical Scale, and Metadata in a Population-Graph Classifier for Dermoscopic Lesions

A PREPRINT

Mehdi Yousefzadeh* 1,2, Parsa Esfahanian* 1, Sara Rashidifar 1, Hossein Salahshoor Gavalan 3, Negar Sadat Rafiee Tabatabaee 4, Saeid Gorgin 5, Dara Rahmati 6, and Maryam Daneshpazhooh 7

¹School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

²Department of Physics, Shahid Beheshti University, Tehran, Iran

³Faculty of Science, University of Tehran, Tehran, Iran

⁴Fatemeh Zahra Hospital, Alborz University of Medical Sciences, Eshtehard, Iran

⁵Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, South Korea

⁶Department of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran

⁷Autoimmune Bullous Diseases Research Center, Razi Hospital, Tehran University of Medical Sciences, Tehran, Iran

yousefzadeh.meh@gmail.com, parsa.esfahanian@ipm.ir, rashidifar.sara@gmail.com,
hoseinsalahshoor1990@gmail.com, nsrafiee425@gmail.com, gorgin81@skku.edu,

d_rahmati@sbu.ac.ir, maryamdanesh.pj@gmail.com

September 17, 2025

ABSTRACT

Introduction. Dermoscopy aids melanoma triage, yet image-only AI often ignores patient metadata (age, sex, site) and the physical scale needed for geometric analysis. We present *GraphDerm*, a population-graph framework that fuses imaging, millimeter-scale calibration, and metadata for multiclass dermoscopic classification, to the best of our knowledge the first ISIC-scale application of GNNs to dermoscopy.

Methods. We curate ISIC 2018/2019, synthesize ruler-embedded images with exact masks, and train U-Nets (SE-ResNet-18) for lesion and ruler segmentation. Pixels-per-millimeter are regressed from the ruler-mask two-point correlation via a lightweight 1D-CNN. From lesion masks we compute real-scale descriptors (area, perimeter, radius of gyration). Node features use EfficientNet-B3; edges encode metadata/geometry similarity (fully weighted or thresholded). A spectral GNN performs semi-supervised node classification; an image-only ANN is the baseline.

Results. Ruler and lesion segmentation reach Dice 0.904 and 0.908; scale regression attains MAE 1.5 px (RMSE 6.6). The graph attains AUC 0.9812, with a thresholded variant using $\sim 25\%$ of edges preserving AUC 0.9788 (vs. 0.9440 for the image-only baseline); per-class AUCs typically fall in the 0.97–0.99 range.

Conclusion. Unifying calibrated scale, lesion geometry, and metadata in a population graph yields substantial gains over image-only pipelines on ISIC-2019. Sparser graphs retain near-optimal accuracy, suggesting efficient deployment. Scale-aware, graph-based AI is a promising direction for dermoscopic decision support; future work will refine learned edge semantics and evaluate on broader curated benchmarks.

Keywords Graph Neural Network · Ruler Segmentation · Lesion Segmentation · Dermoscopy

^{*}Equal contribution.

[†]Corresponding author.

1 Introduction

Skin cancer is the most common malignancy worldwide and represents a major public-health challenge. The World Health Organization (WHO) estimates between two to three million new non-melanoma skin cancers (NMSC) and more than 300,000 new cases of melanoma each year, with melanoma alone accounting for over 57,000 deaths annually Sung et al. [2021], Arnold et al. [2022]. Although basal cell carcinoma (BCC) and squamous cell carcinoma (SCC) comprise the majority of skin malignancies, melanoma is far more aggressive and disproportionately responsible for mortality Whiteman et al. [2016]. Its incidence continues to rise, particularly in fair-skinned populations and regions with high ultraviolet exposure WHO [2017], underscoring the urgent need for effective screening and timely intervention. Late-stage diagnosis is strongly associated with poor five-year survival, highlighting the clinical imperative of early detection and precise triage Jones et al. [2019].

Despite its importance, reliable melanoma detection remains challenging. Clinical examination alone is limited by the resemblance of malignant to benign pigmented lesions, as well as similarities to non-cancerous dermatological disorders. To standardize assessment, dermatologists commonly employ the *ABCDE* rule (Asymmetry, Border irregularity, Color variegation, Diameter, Evolving lesion) as a mnemonic for suspicious features Nachbar et al. [1994], Friedman and Rigel [1985]. Lesions larger than 6 mm in diameter are particularly concerning for melanoma. Yet in practice, diagnostic accuracy varies substantially: biopsy accuracy ranges between 49% and 81%, and up to one-third of melanomas are initially misclassified as benignMacKenzie-Wood et al. [1998], Grin et al. [1990].

Dermoscopy (dermatoscopy) has therefore become the *de facto* non-invasive imaging modality for evaluating skin lesions in routine practice. By revealing sub-macroscopic structures, dermoscopy improves both sensitivity and specificity compared to naked-eye inspection, with reported diagnostic accuracies of 75–97% in experienced hands Kittler et al. [2002], Dinnes et al. [2018], Saurat [2004], Celebi et al. [2005]. Nevertheless, performance is highly operator-dependent: manual interpretation is time-consuming, error-prone, and subject to inter- and intra-observer variability, particularly among less experienced clinicians Jones et al. [2019]. These limitations have accelerated the development of Computer-Assisted Diagnosis (CAD) systems to support dermatologists.

Over the last decade, deep learning has transformed medical image analysis, including skin cancer detection. Convolutional Neural Networks (CNNs) have achieved dermatologist-level performance in binary melanoma classification tasks Esteva et al. [2017], Hekler et al. [2019], and dominate lesion segmentation and classification challenges such as ISIC 2017–2019 Codella et al. [2018, 2019], Tschandl et al. [2018], Combalia et al. [2019]. Models leveraging architectures such as U-Net Ronneberger et al. [2015], DenseNet Huang et al. [2017], and their variants have advanced lesion segmentation Yuan and Lo [2017], Yu et al. [2016], while pipelines combining segmentation and classification have demonstrated state-of-the-art performance Lee et al. [2018]. More recent efforts have explored optimization-based dual networks Gomathi et al. [2023], hybrid CNN–RF classifiers Mustafa et al. [2025], and efficient backbones such as MobileNetV3 Kumar Lilhore et al. [2024], reporting accuracies approaching 98% on HAM10000.

Despite these advances, several challenges remain unresolved. First, multiclass lesion classification is substantially more difficult than binary melanoma detection, due to extreme class imbalance and subtle inter-class variability. Second, progress is constrained by limited dataset size, heterogeneous acquisition protocols, and the high cost of expert annotations. Transfer learning and domain adaptation strategies Wang et al. [2019] partially mitigate these issues but have limited generalization in real-world clinical deployment. Finally, many CNN pipelines rely solely on pixel intensities, ignoring clinically salient cues such as lesion geometry and patient context. These gaps motivate exploration of richer modeling paradigms.

Most existing CNN-based methods treat dermoscopic images in isolation, neglecting two critical elements of clinical reasoning: (i) auxiliary metadata such as age, sex, and anatomic site, which contextualize disease likelihood, and (ii) physical scale, essential for quantifying geometric features such as the diameter criterion in the ABCDE rule Nachbar et al. [1994]. Importantly, physical scale cannot be inferred from raw pixels alone and is often absent due to heterogeneous imaging protocols and the lack of calibration markers Codella et al. [2018, 2019], Jones et al. [2019], Gandhi and Kampp [2015].

Population graphs provide a principled way to incorporate these factors. In this paradigm, each patient or lesion is represented as a node with image-derived features, while edges encode inter-case similarity based on metadata or geometric cues. Graph Neural Networks (GNNs) then enable semi-supervised learning and label propagation across the cohort, leveraging relationships that extend beyond individual pixels Kipf and Welling [2016].

In the broader medical imaging literature, population-graph GNNs have demonstrated substantial gains. Parisot et al. [2018] applied GCNs to neuroimaging cohorts for autism spectrum disorder (ABIDE) and Alzheimer's disease (ADNI), showing that integrating phenotypic metadata with MRI features improves classification over image-only models Parisot et al. [2018]. Subsequent work extended these approaches to EEG analysis Song et al. [2018], cerebral

cortex parcellation Gopinath et al. [2020], anatomical segmentation Noh et al. [2020], and multi-modal disease predictionRakhimberdina et al. [2020], Huang et al. [2017]. Collectively, these studies highlight the promise of GNNs for capturing population structure and context—yet dermatology remains underexplored in this respect.

In this work, we propose *GraphDerm*, a population-graph framework for multiclass dermoscopic lesion classification. To the best of our knowledge, this is the first application of GNNs to ISIC-scale dermoscopy. The framework fuses image-derived features with explicit physical scale information and patient metadata, addressing a key limitation of conventional CNN-only pipelines that operate on images in isolation.

To enable scale-aware modeling, we curate the ISIC 2019 dataset by systematically identifying images with embedded rulers and synthesizing ruler-bearing dermoscopic images from ruler-free sources. This provides exact ruler masks by construction, which serve as supervision for a dedicated ruler-segmentation module. From these masks, we estimate pixels-per-millimeter using a two-point correlation function (TPCF) signature regressed by a lightweight CNN, yielding accurate lesion geometry in physical units.

Building on this, we compute real-scale lesion descriptors including area, perimeter, and a radius of gyration in millimeters directly from segmentation masks. These geometric features are then combined with auxiliary metadata such as age, sex, anatomic site, and dataset source to define the similarity structure of the population graph. We systematically explore multiple edge construction and sparsification strategies, including fully weighted, thresholded, random, and identical edge schemes. The fully weighted graph achieves the best performance with AUC = 0.9812, while a thresholded variant at T=0.7 retains nearly identical AUC = 0.9788 using only about 25% of edges.

Across all experimental settings, neighborhoods informed by metadata and real-scale geometry consistently outperform image-only CNN baselines on ISIC 2019. These results demonstrate that integrating physical scale and cohort structure into dermoscopic classification yields measurable gains over conventional pipelines and highlights the promise of graph-based approaches for clinical dermatology.

Contributions. In short, the contributions of this research are:

- Scale-aware population graph. We introduce *GraphDerm*, a framework that unifies image features, calibrated lesion geometry, and patient metadata within a single graph-based model.
- Calibration and geometry pipeline. We develop a practical procedure to obtain ruler signals and segmentation masks (including synthetic augmentation), estimate physical scale, and derive real-scale geometric descriptors.
- **Graph design and sparsification.** We investigate alternative neighborhood constructions and sparsity levels, showing that carefully structured sparse graphs retain performance close to dense counterparts.
- **Empirical validation.** We demonstrate consistent gains over image-only baselines on a large multiclass dermoscopy benchmark, with stable behavior across classes.

This paper is organized as follows. Section 2 details dataset curation, ruler synthesis, segmentation models, the TPCF-based scale estimator, real-scale geometric feature extraction, and graph construction and GNN design. Section 3 reports segmentation and scale-estimation accuracy, ablations over edge construction and sparsification, and end-to-end classification performance. Section 4 examines the implications and limitations of our approach, including learned edge semantics and broader benchmarks. And finally, section 5 synthesizes the key findings, final takeaways, and avenues for clinical translation.

2 Methods

This section details the *GraphDerm* pipeline for multiclass dermoscopic lesion classification, integrating image analysis, physical calibration, and cohort-level context. The framework is designed to address the limitations of conventional CNN-only pipelines by explicitly modeling real-scale lesion geometry and patient metadata within a graph-based framework.

We first curate dermoscopic corpora from the ISIC 2018 and ISIC 2019 challenges to form the working dataset. From ruler-free images, we synthesize ruler-bearing counterparts with paired ruler masks, which enable supervised training of dedicated U-Net models for ruler and lesion segmentation. Predicted ruler masks are then used to estimate millimeters-per-pixel via a two-point correlation function regressed by a lightweight CNN, yielding precise calibration of physical scale. With this calibration, we compute lesion geometric descriptors including area, perimeter, and radius of gyration in millimeters.

Finally, we construct a population graph where each node encodes imaging-derived features together with scale-aware geometry, and edges capture inter-patient similarity defined by metadata such as age, sex, anatomic site, and dataset source. Multiple edge-weighting schemes are investigated—fully weighted, thresholded, random, and identical—before applying a spectral Graph Neural Network (GNN) for semi-supervised multiclass classification.

2.1 Datasets

The *GraphDerm* dataset was constructed by curating, merging, and processing two public dermoscopy cohorts: the 2018 ISIC challenge set Tschandl et al. [2018], Codella et al. [2019] and the 2019 ISIC challenge set Tschandl et al. [2018], Combalia et al. [2019]. *GraphDerm* contains synthesized dermoscopic images with embedded rulers together with pixel-wise segmentation masks for the skin lesion and the ruler, plus associated metadata.

Within *GraphDerm*, the ISIC 2018 and ISIC 2019 data serve two complementary goals: lesion segmentation and ruler segmentation, respectively. The dataset includes eight diagnostic classes; their distribution is reported in Supplementary Table 1, and one representative example from each class is shown in Supplementary Figure 1.

Lesion segmentation data (ISIC 2018). The ISIC 2018 Task 1 dataset comprises 2,594 dermoscopic images, each accompanied by a suggested lesion segmentation mask that we use as ground truth. Examples are provided in Figure 1. A U-Net model Ronneberger et al. [2015] with an encoder pretrained on ImageNet was trained on this set for lesion mask prediction. Standard dermoscopy preprocessing and augmentations were applied.

Source pool for ruler work (ISIC 2019). To preserve acquisition consistency, we retained the 25,331 raw dermoscopic images from ISIC 2019 across the eight diagnostic classes listed in Supplementary Table 1.

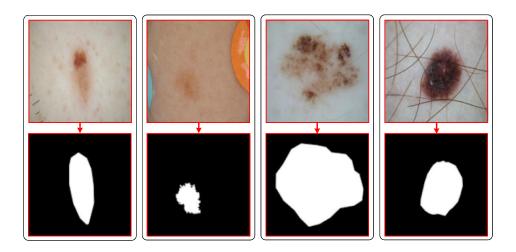


Figure 1: Example dermoscopic images from ISIC 2018 Task 1 alongside their provided lesion segmentation masks used as ground truth.

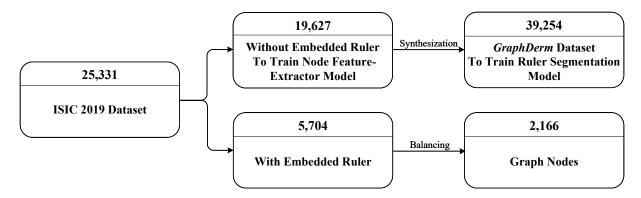


Figure 2: Taxonomy of how ISIC 2019 is utilized within the GraphDerm framework.

Images that already contained visible embedded rulers but lacked ruler masks were treated as noisy and removed via manual screening. After this cleaning, 19,627 ruler-free images remained and were reserved for ruler synthesis (see Supplementary Figure 2). Among the remaining 5,704 images, class counts were imbalanced; we therefore kept all minority-class samples and subsampled the majority classes to approximately balance labels. A subset of 2,166 images was further used to construct the population graph; the overall taxonomy and usage of ISIC 2019 within *GraphDerm* are summarized in Figure 2.

2.2 Synthetic Data Generation for Ruler Segmentation

For training a robust ruler-segmentation model, we synthesized ruler-bearing dermoscopy images from the 19,627 cleaned ruler-free ISIC 2019 images. We designed seven custom rulers and eight millimeter markers (Supplementary Figure 2); each ruler image was stored as a transparent PNG and resized to 512×512 prior to embedding.

Synthesis pipeline. Each synthesized image was produced by the following stochastic procedure:

- 1. Given that real ruler appearances most closely match the top ruler in Supplementary Figure 2, we select that ruler with probability 50%; the other six rulers share the remaining 50% uniformly.
- 2. With probability 20%, one of the millimeter markers in Supplementary Figure 2 is selected uniformly at random and overlaid on the ruler.
- 3. A contiguous segment of the chosen ruler is cropped along its length to a random fraction between 30% and 100% of its original length.
- 4. The resulting ruler segment is embedded onto the dermoscopic image at a random position.
- 5. With probability 50%, a black circular occlusion mask is added to the image.
- 6. Additive white noise with random intensity drawn from a uniform distribution is applied to the image.
- 7. With probability 50%, the entire image is rotated by a random angle in $[0^{\circ}, 360^{\circ}]$ chosen uniformly.
- 8. The image is scaled up or down uniformly at random by 20%.
- 9. A Gaussian filter with σ sampled from a normal distribution and clipped to [0.5, 5.5] is applied.
- 10. The synthesized image is accepted only if at least 20% of ruler line pixels (black ticks) remain after transformations; otherwise the process is repeated.

This pipeline was executed twice per source image, yielding 39,254 ruler-embedded synthetic images in total. Because the embedding transformations are controlled, we also generate the corresponding ruler segmentation masks at step 4; any global transforms in steps 7 and 8 are applied identically to the masks. The synthesized images closely mimic real cases while providing precise, high-quality ruler masks.

Ruler-segmentation model and training. Two U-Net variants were evaluated with MobileNetV2 Sandler et al. [2018] and SE-ResNet-18 Hu et al. [2018] backbones; the latter performed best and was adopted. Due to the importance of spatial detail for thin ruler ticks, inputs were 512×512 with batch size 16. Models were trained for 60 epochs using Adam Kingma and Ba [2014] with an initial learning rate of 2e-3 decayed over epochs. Loss function was a custom

sum of binary cross-entropy and Dice:

BCE + Dice =
$$\left(-\frac{1}{N}\sum_{i=1}^{N} [T_i \log(P_i) + (1 - T_i) \log(1 - P_i)]\right) + \left(1 - \frac{2 \times P \cap T}{P + T}\right),$$
 (1)

where P and T denote the predicted and true pixel labels, respectively.

2.3 Statistical and Geometric Properties

To support screening and quantification, lesion descriptors such as area, perimeter, and a radius of gyration are derived in physical units (mm). This is done by first estimating the pixel scale via a two-point correlation analysis of the predicted ruler mask, and then applying geometric computations to the predicted lesion mask. An overview of the *GraphDerm* pipeline is shown in Figure 3.

Pixel-scale estimation via the two-point correlation function. Let $\delta(\mathbf{x}) \in \{0, 1\}$ be the binary ruler-mask indicator and A the image area in pixels. The (isotropic) Two-Point Correlation Function (TPCF) Jiao et al. [2007], Torquato [2002] is

$$\xi_2(\Delta) = \frac{1}{A} \int d^2x \, \delta(\mathbf{x}) \, \delta(\mathbf{x} + \Delta), \tag{2}$$

with $\Delta = \|\mathbf{x}_1 - \mathbf{x}_2\|_2$ the Euclidean separation in pixels. When evaluated on ruler masks, $\xi_2(\Delta)$ exhibits peaks at separations corresponding to tick spacing (see Supplementary Figure 4). We compute ξ_2 on 300 discrete separations and feed the resulting 1D signature into a lightweight CNN regressor (three 1D-convolutional layers with max pooling and batch normalization, followed by a linear head) to predict ρ = pixels-per-millimeter. Supervision for ρ is obtained directly from the synthesis process by tracking the applied ruler transformations and known physical tick spacing.

Geometric features from the lesion contour. Let $M \in \{0,1\}^{H \times W}$ be the predicted lesion mask, let ρ be the estimated pixels-per-millimeter, and let $\alpha = 1/\rho$ denote pixels-per-millimeter. From M we extract one or more closed iso-contours at level ν (marching squares), each returned as an ordered sequence of sub-pixel points $\{(x_k, y_k)\}_{k=1}^K$ with $(x_{K+1}, y_{K+1}) \equiv (x_1, y_1)$ Mantz et al. [2008]. Using these vertices, we compute:

Perimeter (mm): the Euclidean length of the polyline,

$$P_{\text{lesion}} = \alpha \sum_{k=1}^{K} \sqrt{(x_{k+1} - x_k)^2 + (y_{k+1} - y_k)^2},$$
 (3)

which matches the implementation that sums consecutive point-to-point distances. (For strictly grid-adjacent steps this reduces to the familiar 8-connected chain-code weights $\{1, \sqrt{2}\}$ Dorst and Smeulders [1987].)

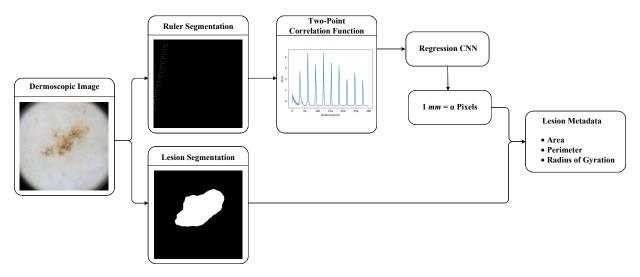


Figure 3: Overview of the *GraphDerm* pipeline. From each dermoscopic image, ruler and lesion segmentation masks are predicted. The two-point correlation function on the ruler mask is input to a CNN regressor that estimates pixels-per-millimeter. Combining this scale with the lesion mask yields geometric lesion features.

Area (mm²): the signed polygon area via Green's theorem (shoelace formula),

$$A_{\text{lesion}} = \alpha^2 \frac{1}{2} \left| \sum_{k=1}^{K} \left(x_k y_{k+1} - y_k x_{k+1} \right) \right|, \tag{4}$$

i.e., the absolute value of the oriented contour area scaled by α^2 . For multiple disjoint components, areas are summed; if interior holes are present, their (oppositely oriented) contour areas subtract naturally O'Rourke [1998].

Radius of Gyration (mm): measures the spatial dispersion of contour points about their centroid,

$$R_{\rm g}^2 = \frac{1}{N} \sum_{i=1}^{N} \left[(x_i - x_c)^2 + (y_i - y_c)^2 \right], \qquad x_c = \frac{1}{N} \sum_{i=1}^{N} x_i, \ y_c = \frac{1}{N} \sum_{i=1}^{N} y_i, \tag{5}$$

which is translation-invariant and, when normalized by size, provides a compact geometric descriptor complementary to area and perimeter for characterizing lesion extent, elongation, and boundary irregularity, often improving discrimination in wound classification and healing-monitoring tasks Hosseinabadi et al. [2012].

2.4 Graph Neural Network

Inspired by [Parisot et al., 2018], an overview of constructing the graph structure is shown in Figure 4.

For the graph structure, a population of N samples S_1, S_2, \ldots, S_N , each corresponding to a single dermoscopic image. Our goal is to predict the classification of each sample based on its imaging and auxiliary data.

The population sample can be denoted as a sparse weighted graph G = (V, E, W), where W is the graph adjacency matrix. Each dermoscopic image is denoted as a node (vertex) $v \in V$, corresponds to a sample S_v of the population, and is accompanied by a d-dimensional feature vector \mathbf{x}_v obtained from imaging data. The graph edges E quantify the similarity between the samples by incorporating relevant auxiliary and imaging data.

By assigning a label $l \in \{0, 1, 2, \dots, 7\}$ to each graph node to indicate a normal label or the presence of any of the aforementioned skin abnormalities, we model skin lesion prediction as a node classification problem. For training the model, we assume a semi-supervised learning scheme, in which the graph neural network receives the complete population graph along with all the node and edge features and only an initial subset of the graphs nodes are used during the training and optimization process.

Intuitively, the population graph acts as a regulator, encouraging vertices connected with high weights to help filter their neighbors features for maximizing label propagation. Therefore, constructing the population graph is a key aspect of the method, as an ill-structured population graph (that does not accurately describe the similarity between samples and their feature vectors) will fail to exploit the power of the graph neural networks. Moreover, an improperly structured population graph can even perform worse compared to a linear classifier. Similarly, this equates to performing an image convolution operation on unrelated pixels (*e.g.* randomly selected) instead of a local image patch.

Constructing the population graph structure requires two primary decisions;

1. The definition of the feature vector \mathbf{x}_v that describes each node,

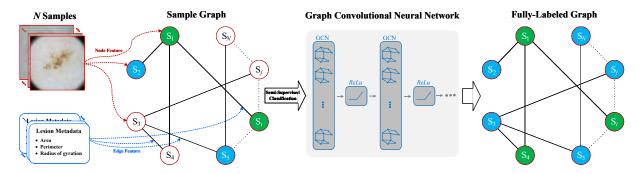


Figure 4: Overview of the proposed graph-based framework for classification using graph neural networks. Imaging data are used to construct node features, while metadata provide the basis for edge features. In the semi-supervised setting, only a subset of nodes have classification labels during training; after training, labels are inferred for all nodes.

2. The edges of the graph E and their weights W, which describe the similarity between the vertices and their feature vectors.

As the objective of this research was to study the improvement on image classification performance when additional information is used in companion with the imaging data, we will only utilize the imaging data of a node's corresponding sample for deriving its feature vector.

We employ two different methods. In the first method, the model's initial weights are pre-trained only on the ImageNet [Deng et al., 2009] dataset. However in the second method, in addition to ImageNet transfer-learning, we also train the model on the 19,627 samples mentioned in subsection 2.1. In both methods, the imaging data with dimensions 300×300 and basic ImageNet normalization is fed to the feature-extractor block of an EfficientNet-B3 model [Tan and Le, 2019] and the output vector of length 1536 is used as the feature vector of the node corresponding to the image's sample. A breakdown on the number of samples used for the training and validation sets used for training the feature-extractor model can be found in Supplementary Table 1.

Similar to the pixel neighborhood logic used by the convolutional filter in CNNs, the graph structure provides a wider field of view for label propagation and filters the value of a feature with respect to its neighbors instead of examining each feature separately. Thus, for an accurate modeling of interactions between the nodes feature vectors, a carefully designed graph edge structure is required. Our hypothesis is that non-image auxiliary data can provide key information to explain the relationship between the feature vectors of the samples. The purpose of using this information is to define an accurate neighborhood system (similar to the convolution filter in the image) that improves the performance of the convolution operation in the graph structure.

Four methods of deriving edge features are explored in this research; fully-weighted, threshold, random, and identical. In all four methods, the graph structure is assumed fully-connected and each edge is then weighted accordingly.

In the fully-weighted method, inspired by Parisot et al. [2018], the weight of each edge is determined by applying a similarity function on the auxiliary data of the node samples at the two ends of the edge. Given a set of H non-imaging auxiliary metrics $M = \{m_h | h \in [1, H]\}$ (e.g. gender or age of the sample), the adjacency matrix of the population graph w is defined as:

$$W(v, w) = Sim(\mathbf{x}_v, \mathbf{x}_w) \sum_{h=1}^{H} \mathbf{w}_h \cdot \gamma(m_h(v), m_h(w)),$$
(6)

where \mathbf{w}_h is the impact weight of the auxiliary metric m_h , γ is an operator measuring the distance between auxiliary metrics, and $Sim(\mathbf{x}_v, \mathbf{x}_w)$ is a measure of similarity between the feature vector of vertices, which increases the weight of edges between the most similar vertices of the graph.

The function to calculate the similarity of node feature vectors $Sim(\mathbf{x}_v, \mathbf{x}_w)$ is an integral component in the construction of Equation 6 as seen in Parisot et al. [2018]. However in this research, we omit this function from the edge weight definition. As previously discussed, the samples' auxiliary data is obtained from training a naturally error-prone model on synthesized data resulted from other error-prone methods. Even if the final result is accurate for independent classification and segmentation problems, the resulting sample feature vector can introduce unwanted and unexpected errors in the diffusion mechanism of the graph structure.

Function γ is defined differently depending on the type of the auxiliary data metric. Each sample in the dataset is associated with three named metrics; gender, anatomical site of the dermoscopy image on the body, name of the source dataset, and four numerical metrics; age, lesion area, lesion perimeter, and lesion radius of gyration.

For named metrics, γ is defined as a Kronecker delta function with the following formula:

$$\gamma(m_h(v), m_h(w)) = \begin{cases} 1 & \text{if } m_h(v) = m_h(w), \\ 0 & \text{otherwise.} \end{cases}$$
 (7)

The definition of γ is however slightly more contrived for numerical metrics. For each of the four numerical metric, the following procedure is applied:

First, a cumulative distribution function is calculated for all the metric values of the entire sample population. To deal with outliers, values below the 1st percentile and above the 99th percentile are clipped to the respective percentile values. Next, for each node pair $v, w \in V$, the set $F = \{f_{vw} \mid f_v w = |m_h(v) - m_h(w)|\}$ is calculated and a z-score normalization is performed to bring the mean and variance values to 0 and 1.

By denoting the newly normalized set as F', the function γ for numerical metric m_h is defined as the following:

$$\gamma(m_h(v), m_h(w)) = Norm(f'_{vw}) \tag{8}$$

for which, $Norm(x) = -\tanh(x)$ was selected which results in the value of γ consequently falling in the [-1,1] range.

For the threshold method of deriving edge features, the edge weight is calculated is as

$$W(v,w) = \begin{cases} 1 & \text{if } W(v,w) \ge T, \\ 0 & \text{otherwise.} \end{cases}$$
 (9)

in which W(v,w) is the weight calculated by Equation 6 using operator γ from Equation 8 and T is a manually selected threshold value.

In the random method, each edge receives a random weight value in the range [-1,1] drawn from a uniform distribution. Then, from the set of edged sorted ascending by their weight value, the top n edges are selected, where n is the number of edges that have a non-zero value in the threshold method that had the best performance in Table 3.

Finally in the identical method, all the edge weights have the same value of 1.

For *GraphDerm*, the graph neural network architecture involves 32 graph convolution hidden layers derived from the spectral-based definition of graph convolution in Kipf and Welling [2016] with a Rectified Linear Unit (ReLU) activation and a 0.5 dropout rate.

The graph neural network model is trained on the whole of the population graph as its input. The training set, following the semi-supervised learning scheme, involves a labeled subset of vertices at the beginning and the validation subset (the unlabeled vertices) receive their features during the training process.

The model was trained with a batch size of 256 and using Adam optimizer Kingma and Ba [2014] with an initial learning rate of 1% for 300 epochs, which stops training if the performance does not improve for 50 consecutive epochs.

A 5-fold cross-validation scheme and also a weighted cross-entropy loss function are used for the optimization process. The weight of the loss function for the presence (+) or the lack there of (-) a class i is calculated using

$$w_i^{+/-} = \frac{\text{\#classes all in samples all}}{\text{\# +/- class in sample } i}$$
 (10)

and the final value of the loss function is obtained using the following:

$$loss = \frac{1}{8} \sum_{i=0}^{7} w_i^{+/-} loss_i.$$
 (11)

After training the graph neural network model, The raw predicted values passed through a softmax function are calculated on the test set, and labels are assigned to the unlabeled vertices accordingly. For the final model, general details of the utilized parameters and the number of samples in each class alongside their corresponding loss function weights can be found in Supplementary Table 2 and Supplementary Table 3, respectively.

All the methodology in this research have been implemented using the Python programming language version 3.7 Van Rossum and Drake Jr [1995]. The neural network models have been developed using Tensorflow version 2.7 Abadi et al. [2015], Keras version 2.7 Chollet et al. [2015], and have been trained using an Nvidia GeForce RTX 5000. The NetworkX package Hagberg et al. [2008] was utilized for processing and analyzing the graphs, the scikit-learn package Pedregosa et al. [2011] for statistical analysis metrics mentioned in Van der Walt et al. [2014], and finally the Matplotlib package was used for generating the plots.

Table 1: The performance results of the two models for the problems of skin lesion segmentation and ruler segmentation with two different backbones.

Problem	Backbone	Dice Score
Lesion Segmentation	MobileNetV2 SE-ResNet-18	0.906 ± 0.21 0.908 ± 0.19
Ruler Segmentation	MobileNetV2 SE-ResNet-18	0.900 ± 0.29 0.904 ± 0.22

3 Results

This section reports the empirical evaluation of the constituent modules and the proposed population-graph classifier. We follow the referencing style established earlier for figures, tables, and citations.

3.1 Evaluation Protocol

To adopt a conservative and reproducible assessment, we report precision, recall, and the area under the ROC curve (AUC) for all experiments, together with confusion matrices when appropriate. Each underlying class is treated as a binary label (present/absent). Using a Bayesian formulation, we compute 95% marginal credible intervals for all summary metrics, and we assess statistical significance via p-values, adopting a conservative 3σ decision level. Because ground-truth labels are binary whereas model outputs are probabilistic, a single operating point is selected to favor high specificity; ROC curves are provided to contextualize this choice.

3.2 Synthesis Pipeline Performance

The data-synthesis process begins with automated screening of ISIC 2019 images for the presence of embedded rulers using an EfficientNet-B2 classifier. The classifier achieves excellent discrimination with an AUC of 0.99 (see confusion matrix and ROC in Supplementary Figure 3).

Next, two U-Net models were trained for ruler and lesion segmentation. Backbones based on MobileNetV2 and SE-ResNet-18 were compared. As summarized in Table 1, SE-ResNet-18 offers superior lesion performance and comparable ruler performance at much lower wall-clock time for the 512×512 input setting; we therefore adopt SE-ResNet-18 for both tasks. A qualitative example is shown in Figure 5.

Finally, pixel scale (pixels-per-millimeter) is inferred by regressing on the two-point correlation signature computed from the predicted ruler masks. The regressor attains an RMSE of 6.6 pixels and an MAE of 1.5 pixels. Distributional summaries across ruler types are shown in Supplementary Figure 5.



Figure 5: Example dermoscopy image with predicted ruler mask and predicted lesion mask.

_ 1	0 0				
Method	Initialization	# Edges	Precision (95% CI)	Recall (95% CI)	AUC (95% CI)
ANN	ImageNet ISIC 2019	-	$\begin{array}{c} 0.8372 \pm 0.0791 \\ 0.7416 \pm 0.0229 \end{array}$	0.5211 ± 0.2928 0.7173 ± 0.0168	0.8701 ± 0.0421 0.9440 ± 0.0064
Full-Weight	ImageNet ISIC 2019	2344695 2344695	$\begin{array}{c} 0.8524 \pm 0.0769 \\ 0.8429 \pm 0.0190 \end{array}$	$\begin{array}{c} 0.5307 \pm 0.3075 \\ 0.8652 \pm 0.0185 \end{array}$	$\begin{array}{c} 0.8773 \pm 0.0903 \\ 0.9812 \pm 0.0020 \end{array}$
Threshold	ImageNet ISIC 2019	36651 36651	$\begin{array}{c} 0.7880 \pm 0.1304 \\ 0.8321 \pm 0.0159 \end{array}$	0.6180 ± 0.1395 0.8490 ± 0.0110	$\begin{array}{c} 0.9098 \pm 0.0534 \\ 0.9788 \pm 0.0019 \end{array}$
Random	-	36651	0.7001 ± 0.0220	0.7067 ± 0.0209	0.9330 ± 0.0106
Identical	-	2344695	0.6426 ± 0.0420	0.6047 ± 0.0640	0.9184 ± 0.0119

Table 2: Combined results for node-feature and edge-weighting strategies (number of active edges shown per setting). Best values per metric are highlighted in blue.

3.3 Population-Graph Classification

We evaluate the proposed graph neural network on ISIC 2019 under the semi-supervised setting described previously, comparing two node-feature initializations (ImageNet vs. additional fine-tuning on ISIC 2019 imaging data and four edge-weighting schemes (full-weighted, thresholded, random, identical. The main results are compiled in Table 2.

- Node features. Using features initialized and further trained on ISIC 2019 yields large gains over ImageNetonly for both the ANN baseline and all GNN variants.
- Edge construction. The full-weighted strategy delivers the best overall performance (highest AUC and recall, with strong precision). The identical and random baselines underperform, underscoring the importance of a meaningful neighborhood system.

Training dynamics (loss, precision, recall, and AUC for both training and validation) for the full-weighted and thresholded graphs are provided in Supplementary Figure 6 and Supplementary Figure 7, respectively.

To probe sparsification, we sweep the threshold $T \in [0,1]$ in increments of 0.05 and record performance and the resulting number of edges. As shown in Table 3, T=0.7 yields the best overall trade-off across all metrics; detailed metric trends with 95% error bars are plotted in Figure 6. *Class-wise* precision, recall, and AUC as functions of T are provided in Supplementary Figure 8.

Finally, per-class analyses for the best full-weighted configuration and for the thresholded graph at T=0.7 are reported in Table 4 and Table 5. Both strategies perform particularly well for *benign keratosis* and *vascular lesion* across metrics.

Summary. Across all settings, leveraging auxiliary information to define graph neighborhoods significantly improves lesion classification over purely image-based baselines. The full-weighted graph attains the strongest overall performance; a carefully chosen thresholded graph (at T=0.7) matches it closely while using $\sim 25\%$ of the edges.

Table 3: Performance of the threshold edge feature deriving method with node feature extraction method based on ISIC 2019 initialization for different thresholds. The best threshold performance for each criterion is highlighted in blue.

Threshold	# Edges	Precision (95% CI)	Recall (95% CI)	AUC (95% CI)
0.0	1045454	0.6808 ± 0.0345	0.6102 ± 0.0613	0.9199 ± 0.0132
0.5	158138	0.6727 ± 0.0401	0.6388 ± 0.0478	0.9197 ± 0.0154
0.55	110307	0.7530 ± 0.0731	0.6457 ± 0.1347	0.9286 ± 0.0326
0.6	66647	0.7162 ± 0.0840	0.6771 ± 0.1171	0.9288 ± 0.0309
0.65	47054	0.8057 ± 0.0622	0.8088 ± 0.0908	0.9683 ± 0.0228
0.7	36651	0.8321 ± 0.0159	0.8490 ± 0.0110	0.9788 ± 0.0019
0.75	26957	0.8054 ± 0.0079	0.8305 ± 0.0079	0.9731 ± 0.0019
0.8	17493	0.7984 ± 0.0151	0.8143 ± 0.0152	0.9711 ± 0.0030
0.85	7959	0.7567 ± 0.0097	0.7658 ± 0.0174	0.9579 ± 0.0058

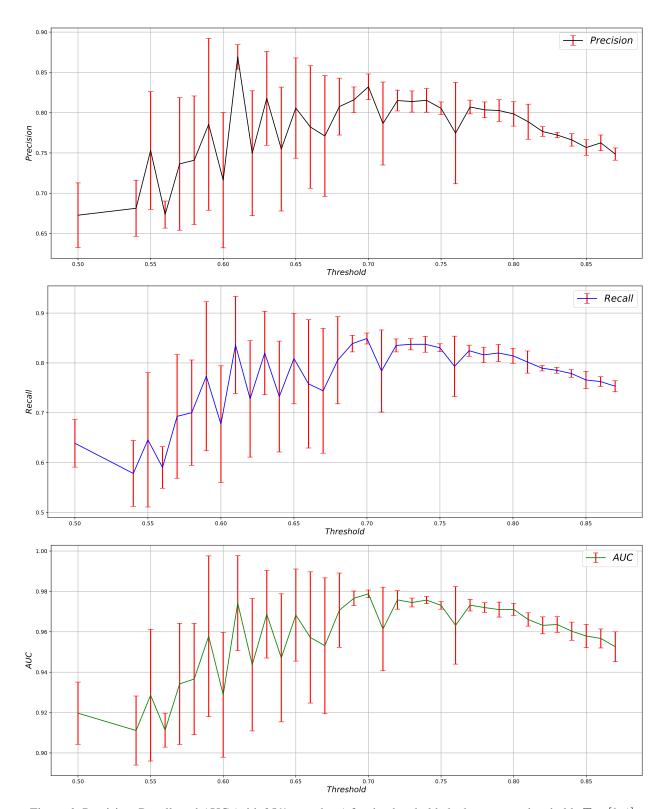


Figure 6: Precision, Recall, and AUC (with 95% error bars) for the thresholded edges across thresholds $T \in [0,1]$.

Table 4: Per-class precision, recall, and AUC for the full-weighted edges with ISIC 2019-initialized node features.

Classes	Precision (95% CI)	Recall (95% CI)	AUC (95% CI)
Melanoma	0.8723 ± 0.0305	0.8464 ± 0.0263	0.9754 ± 0.0033
Melanocytic Nevus	0.8794 ± 0.0317	0.8806 ± 0.0322	0.9837 ± 0.0037
Basal Cell Carcinoma	0.8668 ± 0.0197	0.8942 ± 0.0280	0.9837 ± 0.0016
Actinic Keratosis	0.8040 ± 0.0497	0.8252 ± 0.0334	0.9746 ± 0.0044
Benign Keratosis	0.9033 ± 0.0172	0.8910 ± 0.0166	0.9863 ± 0.0025
Dermatofibroma	0.6778 ± 0.0393	0.7473 ± 0.0841	0.9813 ± 0.0031
Vascular Lesion	0.8400 ± 0.0655	0.9000 ± 0.0640	0.9891 ± 0.0007
Squamous Cell Carcinoma	0.6960 ± 0.0363	0.8383 ± 0.0695	0.9754 ± 0.0051
Average	0.8174 ± 0.0362	0.8529 ± 0.0443	0.9812 ± 0.0031

Table 5: Per-class precision, recall, and AUC for the *thresholded* edges at T=0.7 with ISIC 2019–initialized node features.

Classes	Precision (95% CI)	Recall (95% CI)	AUC (95% CI)
Melanoma	0.8037 ± 0.0178	0.8357 ± 0.0454	0.9584 ± 0.0047
Melanocytic Nevus	0.8142 ± 0.0569	0.8020 ± 0.0339	0.9574 ± 0.0088
Basal Cell Carcinoma	0.9095 ± 0.0195	0.8694 ± 0.0323	0.9816 ± 0.0034
Actinic Keratosis	0.8337 ± 0.0557	0.9156 ± 0.0229	0.9854 ± 0.0065
Benign Keratosis	0.9087 ± 0.0439	0.8071 ± 0.0191	0.9720 ± 0.0121
Dermatofibroma	0.7809 ± 0.0599	0.8273 ± 0.0661	0.9923 ± 0.0040
Vascular Lesion	0.7451 ± 0.0937	0.9400 ± 0.1200	0.9972 ± 0.0029
Squamous Cell Carcinoma	0.7117 ± 0.0182	0.9097 ± 0.0566	0.9861 ± 0.0052
Average	0.8321 ± 0.0159	0.849 ± 0.0110	0.9788 ± 0.0019

4 Discussion

This work investigated the integration of population-graph learning with dermoscopic image analysis, where imaging-derived node representations were augmented with auxiliary metadata and structured through weighted edges to model cohort relationships. Representing the dataset as a sparse weighted graph G = (V, E, W) enabled semi-supervised label propagation across the cohort, allowing the learning process to exploit dependencies that are not recoverable from pixel intensities alone. In the context of eight-class lesion classification on ISIC 2019, our results demonstrated that neighborhoods constructed from metadata and geometric scale information consistently improved performance over image-only baselines (cf. Table 2, Table 4, Table 5).

A central aspect of this study concerned the design of neighborhoods in the adjacency matrix W (Equation 6). We operationalized metadata similarity through a structured definition of $\gamma(\cdot,\cdot)$, incorporating principled normalization for continuous variables (Equation 8) alongside Kronecker-delta treatment of categorical fields. The empirical evaluation indicated that fully weighted graphs achieved the strongest aggregate performance, while thresholded variants with T=0.7 preserved most of these gains using only a fraction of edges (Table 3). This finding suggests that sparse neighborhoods, when carefully designed, can retain near-optimal accuracy while reducing graph density, consistent with prior work on medical population graphs Parisot et al. [2018].

Two practical implications emerge for real-world deployment. First, the choice of node features is critical: representations derived from models fine-tuned on dermoscopy data (ISIC initialization) consistently outperformed features transferred solely from generic natural image pretraining (ImageNet). Second, the semantics of edges matter substantially. Graphs without informative neighborhood structure—whether defined identically or at random—performed markedly worse, illustrating that observed improvements stem not merely from model capacity but from the meaningful incorporation of metadata into graph construction.

This work also has limitations. Chief among them is the reliance on ISIC 2019 as the primary dataset. An ideal benchmark would combine sufficiently large sample size, richly annotated metadata, and standardized protocols for fair comparison across methods. While other population-graph approaches have been validated on large-scale neuroimaging cohorts such as ABIDE Di Martino et al. [2014] and ADNI Jack Jr et al. [2008], the combination of domain expertise, data accessibility, and the need to address dermatology-specific challenges motivated our focus on dermoscopy.

A second limitation is the absence of a curated, held-out test set with expert-verified lesion masks and ruler annotations, which would provide external validation of the full end-to-end pipeline. Assembling such a dataset remains an open task requiring coordinated expert annotation.

As we did not conduct end-to-end external validation on a held-out clinical cohort with real, in-distribution ruler images, generalization across devices, acquisition protocols, and clinics remains unverified. While experiments on ISIC 2019 demonstrated that our method is effective and yields improved results, the absence of a cohort with expert-verified annotations spanning the entire pipeline—specifically, expert labels for the ruler-derived pixel-to-millimeter scale and geometric features in millimeters used to construct the graph and its edges—limited our ability to explore alternative graph-construction strategies and to further increase accuracy; however, this did not alter the evidence that the graph component positively impacts performance. Assembling such a dataset remains an open task requiring coordinated expert annotation.

Finally, robustness assessments would benefit from extending evaluation to broader graph-learning benchmarks, including PubMed Sen et al. [2008], PPI Zitnik and Leskovec [2017], and Reddit Hamilton et al. [2017], which have served as canonical references for evaluating GNN generalization in non-medical domains.

Several avenues for future research follow naturally from the present findings. With respect to node features, an important direction is to replace static CNN-derived embeddings with representations learned jointly through autoencoder bottlenecks optimized end-to-end with the GNN classifier, as explored in related graph-based medical imaging studies Parisot et al. [2018]. Alternative backbones such as ResNet, DenseNet, or MobileNet families could also be investigated for their differing trade-offs between accuracy and efficiency.

Regarding edge weights, our formulation in Equation 6 treated metadata channels uniformly; optimization strategies such as metaheuristic search or differentiable weighting could allow the relative contribution of each metadata field to be learned during training. Robust similarity measures also warrant exploration: although explicit feature-based similarity was avoided to reduce error propagation from noisy embeddings, bounded or regularized kernels may offer a means of capturing finer-grained relationships while mitigating noise sensitivity.

For numerical metadata, sensitivity to outlier clipping and the normalization function $Norm(\cdot)$ in Equation 8 suggests the value of adaptive distribution-aware approaches to further stabilize edge weights. Finally, the GNN architecture itself was fixed to a strong baseline in this study. Systematic evaluation of architectural depth, hidden width, activation

functions, normalization schemes, and regularization strategies (*e.g.*, dropout, edge dropout, MixUp for graphs) could provide further gains, while extending comparisons to spatial as well as spectral GNN variants would contextualize the benefits of our design.

5 Conclusion

We have presented a population-graph framework for multiclass dermoscopic lesion classification that fuses imaging features with auxiliary metadata and explicit physical scale within a unified GNN. Experiments on ISIC 2019 demonstrated that metadata-informed neighborhoods significantly improve precision, recall, and AUC over image-only baselines, with fully weighted graphs performing best and thresholded graphs at T=0.7 achieving nearly identical accuracy while requiring far fewer edges (Table 2, Table 3).

These improvements, together with stable training behavior (Supplementary Figure 6, Supplementary Figure 7) and high-quality segmentation and pixel-scale estimation, underscore the promise of graph-based learning for dermoscopic analysis. Looking forward, the development of curated test sets, particularly with real, expert-verified lesion masks and ruler annotations, along with external validation on clinical cohorts, remains essential steps toward translating graph neural networks into clinical dermatology. The integration of learned edge semantics, and systematic benchmarking against broader datasets, will further establish the robustness and transportability of this framework.

Acknowledgement

The authors would like to thank the doctors and medical staff at Razi Hospital, Tehran University of Medical Sciences, Tehran, Iran, as well as Dr. Farideh Beyki. The authors also acknowledge the use of *ChatGPT* (*OpenAI*, chatgpt.com) for assistance with language editing and improving the readability of the manuscript. The research ideas, analysis, and conclusions are entirely the responsibility of the authors.

References

- Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.
- Melina Arnold, Desraj Singh, Mathieu Laversanne, Jérôme Vignat, Jacques Ferlay, and Freddie Bray. Global burden of cutaneous melanoma in 2020 and projections to 2040. *JAMA Dermatology*, 158(5):495–503, 2022. doi:10.1001/jamadermatol.2022.0160.
- David C Whiteman, Adele C Green, and Catherine M Olsen. The growing burden of invasive melanoma: Projections of incidence rates and numbers of new cases in six susceptible populations through 2031. *Journal of Investigative Dermatology*, 136(6):1161–1171, 2016. doi:10.1016/j.jid.2016.01.035.
- WHO. Radiation: Ultraviolet (uv) radiation and skin cancer. https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-(uv)-radiation-and-skin-cancer, 2017. Accessed: October 16, 2017.
- OT Jones, LC Jurascheck, MA Van Melle, S Hickman, NP Burrows, PN Hall, J Emery, and FM Walter. Dermoscopy for melanoma detection and triage in primary care: a systematic review. *BMJ open*, 9(8):e027529, 2019.
- Franz Nachbar, Wilhelm Stolz, Tanja Merkle, Armand B Cognetta, Thomas Vogt, Michael Landthaler, Peter Bilek, Otto Braun-Falco, and Gerd Plewig. The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30(4):551–559, 1994.
- Robert J Friedman and Darrell S Rigel. The clinical features of malignant melanoma. *Dermatologic clinics*, 3(2): 271–283, 1985.
- Andrew R MacKenzie-Wood, Gerry W Milton, and John W de Launey. Melanoma: accuracy of clinical diagnosis. *Australasian journal of dermatology*, 39(1):31–33, 1998.
- Caron M Grin, Alfred W Kopf, Bruce Welkovich, Robert S Bart, and Marcia J Levenstein. Accuracy in the clinical diagnosis of malignant melanoma. *Archives of dermatology*, 126(6):763–766, 1990.
- Harald Kittler, H Pehamberger, K Wolff, and Michael Binder. Diagnostic accuracy of dermoscopy. *The Lancet Oncology*, 3(3):159–165, 2002. doi:10.1016/S1470-2045(02)00679-4.
- Jonathan Dinnes, Jonathan J Deeks, Naomi Chuchu, et al. Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults. *Cochrane Database of Systematic Reviews*, (12):CD011902, 2018. doi:10.1002/14651858.CD011902.pub2.
- JH Saurat. Dermoscopy of pigmented lesions: a valuable tool in the diagnosis of melanoma. *Swiss Medical Weekly*, 134(0708):83–90, 2004.
- M Emre Celebi, Y Alp Aslandogan, and Paul R Bergstresser. Unsupervised border detection of skin lesion images. In *International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II*, volume 2, pages 123–128. IEEE, 2005.
- Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.
- Achim Hekler, Jochen S Utikal, Alexander H Enk, Axel Hauschild, Michael Weichenthal, Roman C Maron, Carola Berking, Sebastian Haferkamp, Joachim Klode, Dirk Schadendorf, et al. Superior skin cancer classification by the combination of human and artificial intelligence. *European Journal of Cancer*, 120:114–121, 2019.
- Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. 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). In 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), pages 168–172. IEEE, 2018.

- Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv* preprint arXiv:1902.03368, 2019.
- Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.
- Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint* arXiv:1908.02288, 2019.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- Yading Yuan and Yeh-Chi Lo. Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks. *IEEE journal of biomedical and health informatics*, 23(2):519–526, 2017.
- Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng-Ann Heng. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE transactions on medical imaging*, 36(4):994–1004, 2016.
- Yeong Chan Lee, Sang-Hyuk Jung, and Hong-Hee Won. Wonderm: Skin lesion classification with fine-tuned neural networks. *arXiv preprint arXiv:1808.03426*, 2018.
- E Gomathi, M Jayasheela, M Thamarai, and M Geetha. Skin cancer detection using dual optimization based deep learning network. *Biomedical Signal Processing and Control*, 84:104968, 2023.
- Saleem Mustafa, Arfan Jaffar, Muhammad Rashid, Sheeraz Akram, and Sohail Masood Bhatti. Deep learning-based skin lesion analysis using hybrid resunet++ and modified alexnet-random forest for enhanced segmentation and classification. *PloS one*, 20(1):e0315120, 2025.
- Umesh Kumar Lilhore, Sarita Simaiya, Yogesh Kumar Sharma, Kuldeep Singh Kaswan, KBV Brahma Rao, VVR Maheswara Rao, Anupam Baliyan, Anchit Bijalwan, and Roobaea Alroobaea. A precise model for skin cancer diagnosis using hybrid u-net and improved mobilenet-v3 with hyperparameters optimization. *Scientific reports*, 14(1):4299, 2024.
- Zengmao Wang, Bo Du, and Yuhong Guo. Domain adaptation with neural embedding matching. *IEEE transactions on neural networks and learning systems*, 31(7):2387–2397, 2019.
- Sumul Ashok Gandhi and Jeremy Kampp. Skin cancer epidemiology, detection, and management. *The Medical Clinics of North America*, 99(6):1323–1335, 2015.
- Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv* preprint *arXiv*:1609.02907, 2016.
- Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero, Ben Glocker, and Daniel Rueckert. Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease. *Medical image analysis*, 48:117–130, 2018.
- Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2018.
- Karthik Gopinath, Christian Desrosiers, and Herve Lombaert. Graph domain adaptation for alignment-invariant brain surface segmentation. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 152–163. Springer, 2020.
- Kyoung Jin Noh, Sang Jun Park, and Soochahn Lee. Combining fundus images and fluorescein angiography for artery/vein classification using the hierarchical vessel graph network. In *International conference on medical image computing and computer-assisted intervention*, pages 595–605. Springer, 2020.
- Zarina Rakhimberdina, Xin Liu, and Tsuyoshi Murata. Population graph-based multi-model ensemble method for diagnosing autism spectrum disorder. *Sensors*, 20(21):6001, 2020.
- Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018. doi:10.1109/CVPR.2018.00474.
- Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. doi:10.1109/CVPR.2018.00745.

- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
- Yang Jiao, Frank H. Stillinger, and Salvatore Torquato. Modeling heterogeneous materials via two-point correlation functions: Basic principles. *Physical Review E*, 76(3):031110, 2007. doi:10.1103/PhysRevE.76.031110.
- Salvatore Torquato. Random Heterogeneous Materials: Microstructure and Macroscopic Properties. Springer, New York, 2002. doi:10.1007/978-1-4757-6355-3.
- Hubert Mantz, Karin Jacobs, and Klaus Mecke. Utilizing minkowski functionals for image analysis: a marching square algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, (12):P12015, 2008. doi:10.1088/1742-5468/2008/12/P12015.
- Leo Dorst and Arnold W. M. Smeulders. Length estimators for digitized contours. *Computer Vision, Graphics, and Image Processing*, 40(3):311–333, 1987. doi:10.1016/0734-189X(87)80145-7.
- Joseph O'Rourke. Computational Geometry in C. Cambridge University Press, Cambridge, 2 edition, 1998. ISBN 0521649765.
- S Hosseinabadi, MA Rajabpour, M Sadegh Movahed, and SM Vaez Allaei. Geometrical exponents of contour loops on synthetic multifractal rough surfaces: Multiplicative hierarchical cascade p model. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 85(3):031113, 2012.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.
- Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
- Guido Van Rossum and Fred L Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.
- Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.
- François Chollet et al. Keras. https://keras.io, 2015.
- Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 15, Pasadena, CA USA, 2008.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.
- Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.
- Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.
- Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.
- Marinka Zitnik and Jure Leskovec. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, 33(14):i190–i198, 2017.
- Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

6 Supplementary Material

Supplementary Table 1: Class distribution in the ISIC 2019 dataset, along with the number of samples allocated to the training and validation sets used for training the feature-extraction model.

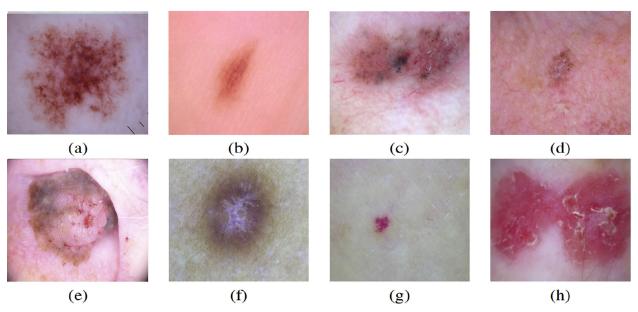
Classes	Total	Train	Validation	
Melanoma	4522	2586	601	
Melanocytic Nevus	12875	8514	2168	
Basal Cell Carcinoma	3323	2027	511	
Actinic Keratosis	867	480	113	
Benign Keratosis	2624	1446	353	
Dermatofibroma	239	126	39	
Vascular Lesion	253	168	44	
Squamous Cell Carcinoma	628	354	97	

Supplementary Table 2: Details and parameters of the graph neural network model architecture and training.

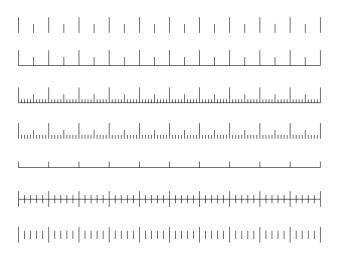
Model	GCN Kipf and Welling [2016]	Loss	Weighted Categorical Cross-Entropy	
Batch Size	256	Validation	5-Fold Cross-Validation	
Optimizer	Adam	# Epochs	300	
Learning-Rate	0.01	# Hidden Layers	32	

Supplementary Table 3: Sample count and loss function weights for all classes.

i	0	1	2	3	4	5	6	7
# Samples	400	400	400	274	400	74	41	177
w_i^+	2.7075	2.7075	2.7075	3.9525	2.7075	14.635	26.414	6.1186
w_i^-	0.6132	0.6132	0.6132	0.5724	0.6132	0.5176	0.5096	0.5444

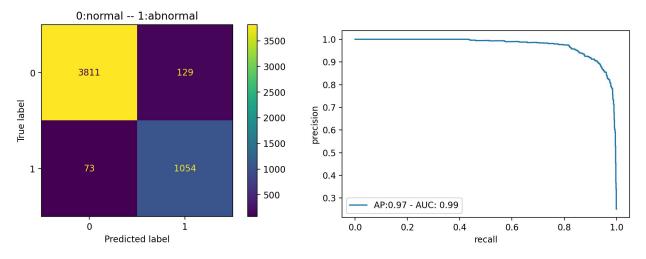


Supplementary Figure 1: One sample image per each of the 8 diagnostic classes in the *GraphDerm* dataset: (a) Melanoma, (b) Melanocytic Nevus, (c) Basal Cell Carcinoma, (d) Actinic Keratosis, (e) Benign Keratosis, (f) Dermatofibroma, (g) Vascular Lesion, and (h) Squamous Cell Carcinoma.

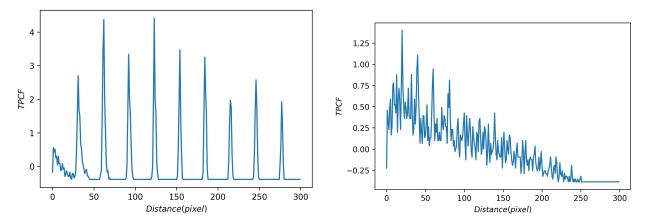


MM mm MM mm MM mm

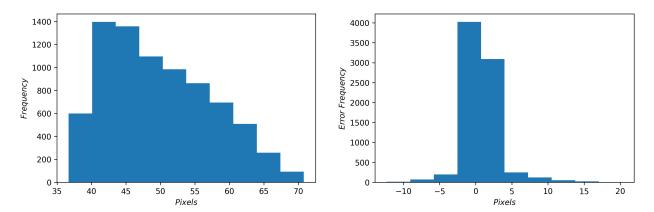
Supplementary Figure 2: Custom-designed rulers and millimeter markers used for embedding into synthesized dermoscopic images.



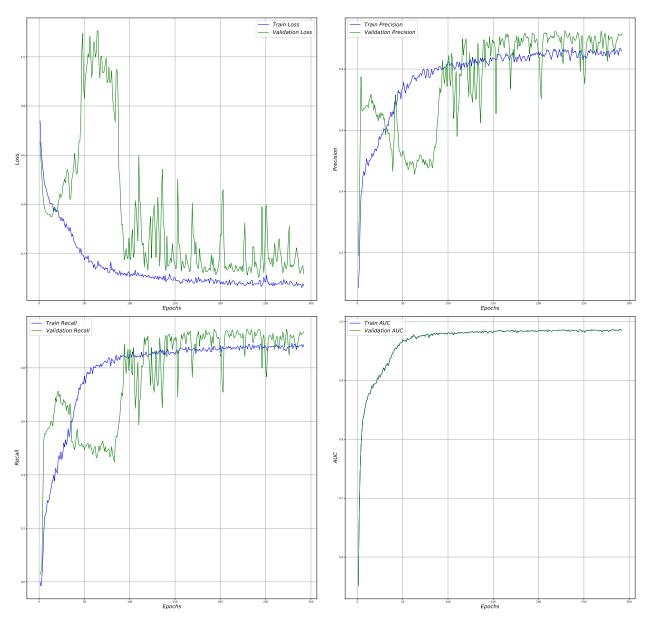
Supplementary Figure 3: Inspection model for detecting embedded rulers: confusion matrix (left) and ROC curve (right).



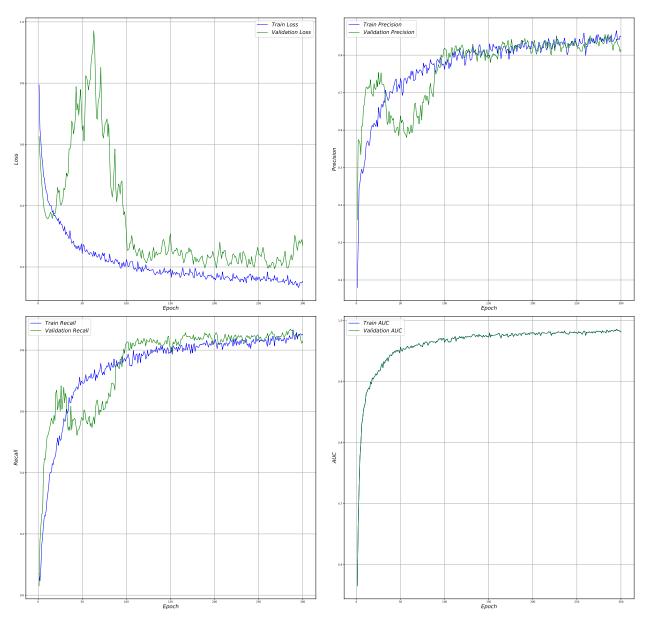
Supplementary Figure 4: Two examples of the two-point correlation function computed on predicted ruler masks. Right: high-quality ruler prediction with clear peaks. Left: noisy prediction with degraded peaks.



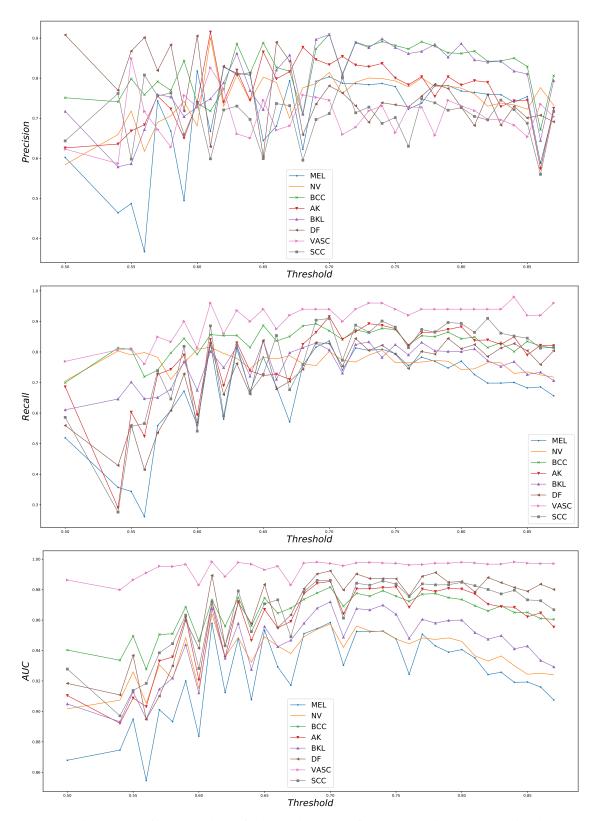
Supplementary Figure 5: Pixel-scale regressor outputs across ruler types: model predictions (left) and prediction error histogram (right).



Supplementary Figure 6: Training/validation curves for the full-weighted edge construction (ISIC 2019–initialized node features): loss, precision, recall, and AUC across epochs.



Supplementary Figure 7: Training/validation curves for the thresholded edge construction (ISIC 2019–initialized node features): loss, precision, recall, and AUC across epochs.



Supplementary Figure 8: Performance chart of the threshold edge feature extraction method (with the ISIC 2019 Initialization-based node feature extraction method in the statistical measures of precision, recall, and area under the receiver operating characteristic curve for different threshold values, separated by dataset classifications.