β-VAE for Skin Cancer Detection

Alison Oyome, Afreen Naveen and Max Pierson
Group #: 13
School of Engineering and Applied Sciences
University at Buffalo
Buffalo, NY 142603
{aoyome;afreenna;maxpiers}@buffalo.edu

Abstract

This paper discusses the use of Variational Autoencoders for early detection of skin cancer using the ISIC 2024 challenge dataset. This competition focuses on classifying skin lesions using cellphone-like images in order to allow for detection without requiring focused attention from a licensed dermatologist, allowing those without easy access to medical services to know when proper diagnosis is required. This is done with the implementation of a β-VAE to map the latent space of benign skin lesions, allowing for the generation of reconstructed images within the benign latent space from raw input images. Using features calculated from the latent and reconstruction losses, input images are then classified using a raw anomaly score with weak results, then repurposed into a vector of features for SVM classification, producing competitive results with existing implementation on the ISIC 2024 dataset.

1 Introduction

Skin cancer is an unusual growth of cells in the body and can be quite problematic if not treated early on. This cancer is not widely discussed, however there is projected to be a 30% increase in cases by 2030. Some people get bruises or other types of lesions and don't stop to think if it could be cancerous. Other people are too lazy to get it checked out by the doctor. Even when seen by a doctor, there is a long process to determine if the lesion is benign or malignant. Artificial Intelligence and computer vision help to ease the detection of skin cancer and reduce mortality. Because of the rarity of skin cancer, a large quantity of benign skin lesion images is available while only a small portion of skin lesion images are ever diagnosed as malignant. Our approach combines VAE and SVM to map a latent space using the large quantity of benign images, allowing for classification of skin lesions. The latent space in the VAE allows us to capture characteristics of the lesions using the training dataset, theoretically resulting in limited reconstruction of anomalies. SVM allows us to use our simplified feature set to distinguish between similar extracted features.

2 Related works

In the history of artificial intelligence and skin cancer detection, most researchers have used supervised learning with different neural networks. However, there are a few that have used unsupervised learning differently. Datasets used include a HAM or ISIC dataset. In 2021, B. Ahmad et al. used VAE and GAN togetherAhmad et al. [2021]. They trained the VAE on the dataset and then used multiple GANs to test the generated images ending with a CNN to classify and create new images. In 2022, M. Zia Ur Rehman et al. used a CNN called MobileNetV2 to obtain a confusion matrix for their classificationZia Ur Rehman et al. [2022]. M. S. Islam and S. Panta tested five different CNNs on the ISIC dataset to compare the results and find which network had the best scoresIslam

and Panta [2024]. Similarly to others, we will be using the ISIC dataset for 2024. For the challenge on Kaggle with this dataset, the participants commonly used architectures like ResNet, EfficientNet, VGG, etc., which motivates us to try a VAE instead. Unlike the previous approaches where CNNs were mainly utilized, we will be using a VAE for lesion classification and generation. We plan to generate new versions of the train and test images mapped to the latent space and use reconstruction loss to classify the input images.

3 Data

(a) Benign Sample Image

(b) Malignant Sample Image

The dataset used for this project is the ISIC 2024 – Skin Cancer Detection with 3D-TBP available on Kaggle [ISIC]. It was created for a competition to develop the best algorithm for early detection of skin cancer using cellphone photos. This dataset is comprised of cropped JPEG images of individual skin lesions from 3D Total Body Photos with resolutions adjusted to resemble close-up smartphone photos. Accompanying each image is a feature set describing the skin lesion's shape, size , color, approximate location on the body, and a target value representing the official diagnosis: 0 for benign and 1 for malignant.

The images are all 3 channel RGB, with varying sizes ranging from 100x100 to 175x175. Due to the relative rarity of malignant skin lesions compared to benign tissue, the dataset is heavily imbalanced with 400,666 benign samples and 393 malignant samples.

For this project, all 393 malignant samples were extracted for a classification set. To maintain the class imbalance, the classification set used a 10:1 ratio of benign to malignant samples; this set was then split into train, validation, and test sets with a 60:20:20 ratio. The remaining 394,386 benign samples were saved for latent space mapping, and split into 60:40 vae_training and vae_validation sets. During preprocessing, all images were resized to 112x112, the categorical feature representing body location was one hot encoded, and all other features were normalized using the means and standard deviations calculated from the vae_training set.

4 Methods

4.1 VAE

Due to the significant class imbalance of the dataset, this project uses a β-VAE architecture to map the latent space of benign image samplesLi et al. [2021]. Given enough samples of a class, a VAE is able to map a latent space representing the possible images of the class. New samples within the class can then be mapped to this latent space and reconstructed with little error, while samples outside this class should fail to reconstruct effectively. First, the 112x112 images are encoded to a

Field Name	Description
$clin_size_longdiam_mm$	Maximum diameter of the lesion (mm).
$tbp_lv_areaMM2$	Area of lesion (mm^2).
$tbp_lv_area_perim_ratio$	Border jaggedness, ratio between perimeter and area.
$tbp_lv_color_std_mean$	Color irregularity.
tbp_lv_deltaA	Average A contrast (inside vs. outside lesion).
tbp_lv_deltaB	Average B contrast (inside vs. outside lesion).
tbp_lv_deltaL	Average L contrast (inside vs. outside lesion).
$tbp_lv_deltaLBnorm$	Contrast between lesion and surrounding skin.
$tbp_lv_eccentricity$	Eccentricity.
$tbp_lv_location_simple$	Simplified anatomical location classification.
$tbp_lv_minorAxisMM$	Smallest lesion diameter (mm).
$tbp_lv_nevi_confidence$	Nevus confidence score (0–100).
$tbp_lv_norm_border$	Border irregularity (0–10).
$tbp_lv_norm_color$	Color variation (0–10).
$tbp_lv_perimeterMM$	Perimeter of lesion (mm).
$tbp_lv_radial_color_std_max$	Color asymmetry.
tbp_lv_stdL	Standard deviation of L inside lesion.
$tbp_lv_stdLExt$	Standard deviation of L outside lesion.
$tbp_lv_symm_2axis$	Border asymmetry (0–10).
$tbp_lv_symm_2axis_angle$	Lesion border asymmetry angle.

Table 1: ISIC dataset tabular fields and their descriptions.

vector of length 256. For encoding, the ConvNeXt architecture with pretrained weights is used due to its strong performance in feature embedding compared to other CNN architectures Woo et al. [2023]. With the encoded vector, gaussian noise is introduced while sampling using the reparamaterization trick

$$z = \mu + \sigma \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

This allows sampling of the latent space for the decoder while maintaining gradient flow between the encoder and decoder. For decoding, a DCGAN style upsampling architecture is used Radford et al. [2015], resulting in an image reconstruction of size 112x112, symmetric with input. Though the selected decoder results in blurrier generated images, this method focuses on the general structure of latent space samples, improving the ability to ignore unseen features when reconstructing which should improve anomaly detection.

Once the encoder generates a reconstruction x_hat from the input, the reconstruction loss is calculated as the mean of the L1 Loss

$$\mathbf{E} = |\hat{x} - x|$$

KL Divergence is calculated from the latent mean and variance

$$\mathrm{KL}_j = -\frac{1}{2}\left(1 + \log \sigma_j^2 - \mu_j^2 - \sigma_j^2\right)$$

, and the total loss is calculated with scaling

$$\mathcal{L} = \mathbb{E}_{x \sim \mathcal{D}}\left[\alpha\, \mathcal{L}_{\text{recon}} + \beta\, \mathcal{L}_{\text{KL}}\right]$$

4.2 Anomaly Score

Initial classification was completed using an anomaly score created with four features calculated from the latent space and reconstruction. From the per-pixel reconstruction error

$$\mathbf{E} = |\hat{x} - x|$$

The mean reconstruction error, variance, and maximum were extracted to capture the global, statistical, and local reconstruction failures. Additionally, the latent Mahalanobis distance

$$f_{\text{mahal}} = (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)$$

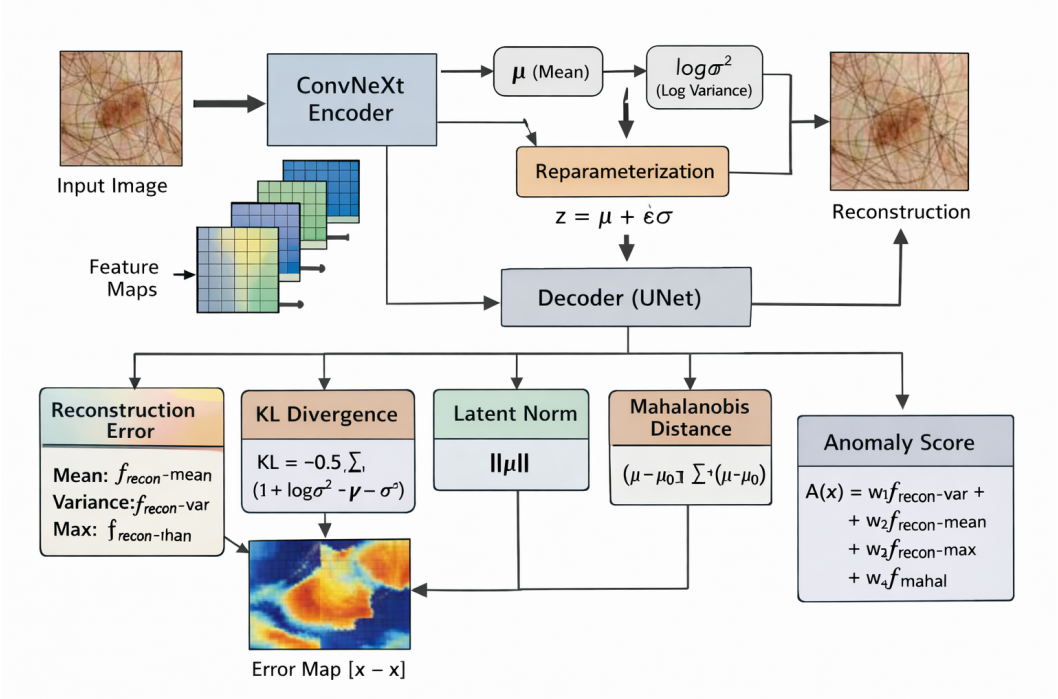


Figure 2: VAE & Anomaly Feature Extraction Structure. This diagram was created by the authors using an AI-assisted image generation tool.

was used to capture the deviation from the trained latent distribution. These values were normalized against train set statistics and used to calculate an anomaly scoreŠmídl et al. [2019]:

$$\mathcal{A}(x) = 0.55\, f_{\text{recon-var}} + 0.25\, f_{\text{recon-mean}} + 0.10\, f_{\text{recon-max}} + 0.10\, f_{\text{mahal}}$$

Which was compared against a threshold to classify samples.

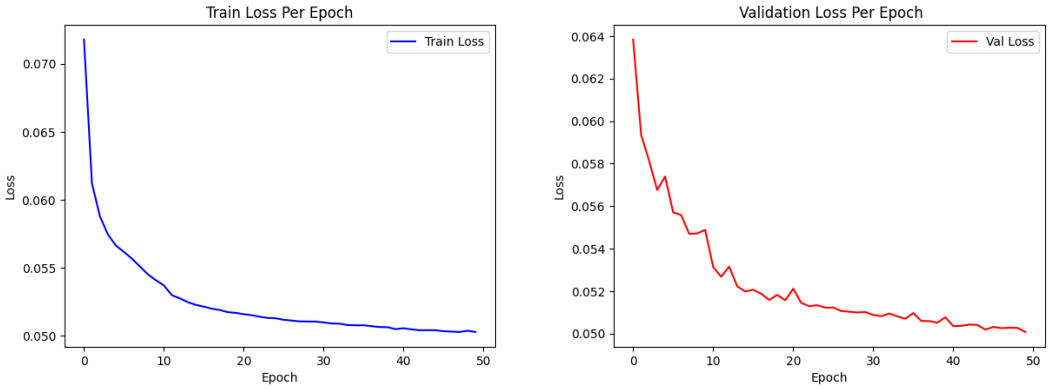
4.3 SVM

The second classification method used these features combined with the latent KL Divergence as a set to represent a simplified image feature vector. This vector was concatenated with tabular features provided by ISIC 2024, resulting in a vector of length 35 which was passed into SVM with an RBF kernel for classification.

5 Experiments and Results

5.1 Latent Space Mapping

Using the VAE architecture with reconstruction $\alpha = 100.0$ and KLD $\beta = 1.0$, the VAE was trained for 50 epochs with a $batchsize = 64$, using the Adam optimizer with $learningrate = 1e-4$. The loss curves and reconstruction samples are shown in figures 3a and 3b, notably the reconstructions are all generalized to the basic shape of the lesion and the skin color. While this generalization is effective for anomaly detection, in this case the effect was reduced, likely by the inability of the model to generate fine detail in the hair and skin textures.



(a) Training loss per epoch

(b) Validation loss per epoch

Figure 3: Training and validation curves (top) and Generated reconstruction samples (bottom).

5.2 Classification

This limited effectiveness is shown in the loss features of the generated reconstructions, whose distribution was inseparable between the two classes (Figure 4), resulting in weak classification ability of the anomaly score method as shown in Table 2.

Class	Precision	Recall	F1-score	Support
Benign	0.92	0.94	0.93	1970
Malignant	0.28	0.21	0.24	197

Table 2: Classification report for Anomaly Scoring Classification method.

However, the SVM classification method greatly improved upon this result by using each feature of the anomaly score calculation in combination with more fine detailed tabular data (Table 3).

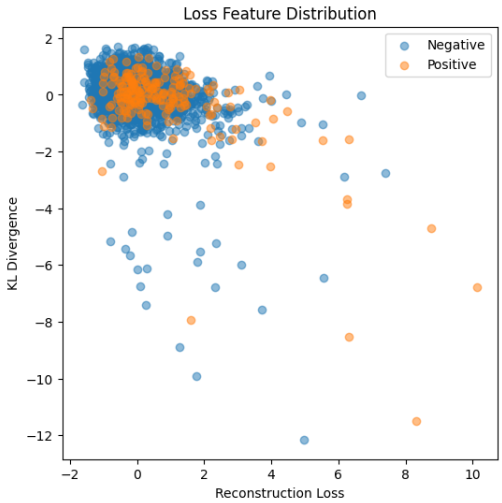
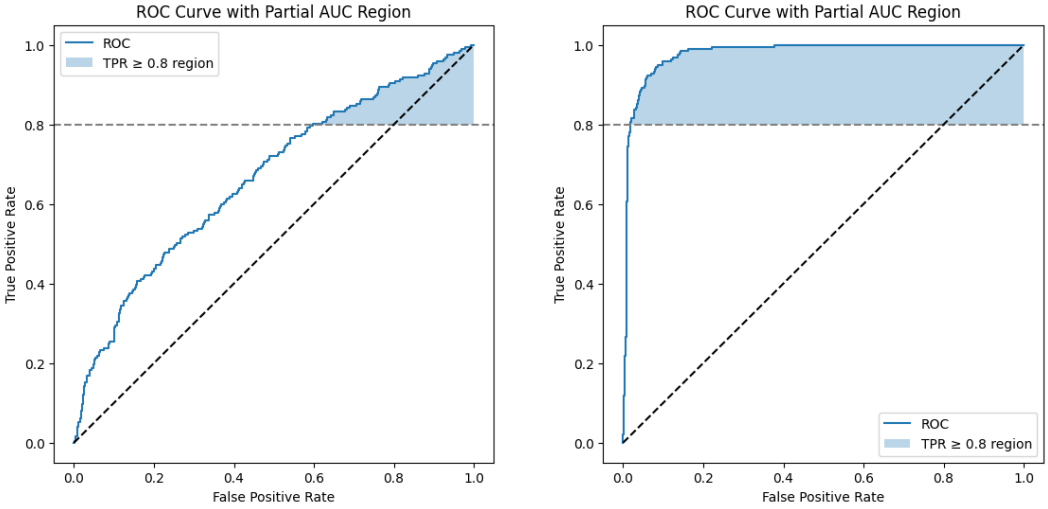


Figure 4: Training loss per epoch

Class	Precision	Recall	F1-score	Support
Benign	0.98	0.98	0.98	1960
Malignant	0.79	0.82	0.80	196

Table 3: Classification report for SVM.



(a) ROC for Anomaly Classification method

(b) ROC for SVM

By using the results of each classifier calculate ROC curves against the classified data (Figures 5a and 5b, the partial area under the curve above a TPR of 0.80 could be calculated. This value is used to compare this project's results with the submissions to the ISIC 2024 competition. As previously discussed the anomaly score classifier displayed only weak success and a chieved a $pAUC = 0.0551$. While the SVM classifier was not able to beat all submitted algorithms(Table 4), it proved to be extremely effective with a $pAUC = 0.1799$, suggesting that minor improvements in this implementation could result in a winning submission.

Method	pAUC Score
resnet18_StratifiedGroupKFold_3fold	0.185
EfficientNetV2S ISIC 2024 Dataset v1	0.181
resnet18	0.174
tf-efficientnet-b0	0.152

Table 4: Comparison of methods by partial AUC (pAUC) score.

6 Conclusion and future work

The implemented VAE was able to define a latent space capable of generating extremely generalized images of skin lesions to be used in anomaly detection. Though this project focused on detecting anomalies by limiting the model's reconstructive ability, this showed limited effectiveness due to additional reconstruction losses from varying skin and hair textures. However, using the features extracted from these reconstructions as a simplified representation of the image data, this method proved extremely effective when combined with additional tabular data. Further experimentation would involve replacing the decoder architecture with one more tailored for reconstructing fine details to compare the anomaly detection capabilities. Additionally, the dataset is noted to have limited representation, with all data collected from only 9 hospitals across the world. Future implementations of this dataset should focus on increased diversity to improve detection effectiveness for all possible users of the dataset.

References

B. Ahmad, S. Jun, V. Palade, Q. You, L. Mao, and M. Zhongjie. Improving skin cancer classification using heavy-tailed student t-distribution in generative adversarial networks (ted-gan). *Diagnostics*, 11(11):2147, November 2021. doi: 10.3390/diagnostics11112147.

International Skin Imaging Collaboration (ISIC). Isic 2024 challenge dataset: Skin cancer detection with 3d total body photography. `https://www.kaggle.com/competitions/isic-2024-challenge/data`, 2024. Accessed: 2025-12-16.

M. S. Islam and S. Panta. Skin cancer images classification using transfer learning techniques, 2024. URL `https://arxiv.org/pdf/2406.12954`.

Chengwei Li, Kihyuk Sohn, Tal Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization, 2021. URL `https://arxiv.org/abs/2104.11927`.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2015. URL `https://arxiv.org/abs/1511.06434`.

Václav Šmídl, Jan Bím, and Tomáš Pevný. Anomaly scores for generative models. *arXiv preprint arXiv:1905.11890*, 2019.

S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. URL `https://openaccess.thecvf.com/content/CVPR2023/papers/Woo_ConvNeXt_V2_Co-Designing_and_Scaling_ConvNets_With_Masked_Autoencoders_CVPR_2023_paper.pdf`.

M. Zia Ur Rehman, F. Ahmed, S. A. Alsuhibany, S. S. Jamal, M. Zulfiqar Ali, and J. Ahmad. Classification of skin cancer lesions using explainable deep learning. *Sensors*, 22(18):6915, September 2022. doi: 10.3390/s22186915.