EEE 443 - Project

Final Report

Anomaly Detection for Malignant Skin Cancer Diagnosis Using Convolutional Autoencoders

Yusuf Karacalı - 22203415
Mehreen Irfan - 222011004

January 3, 2026

Abstract

Skin cancer remains a critical global health challenge, with early detection playing a pivotal role in patient outcomes. Traditional diagnostic methods often rely on supervised learning models that require vast, balanced datasets of both benign and malignant lesions – a condition rarely met in clinical practice due to the scarcity of malignant samples. This project proposes an unsupervised anomaly detection framework to address this limitation using a Convolutional Autoencoder (CAE). By training exclusively on benign skin lesions from the ISIC dataset [1], the model learns to reconstruct the structural and textural features of "normal" skin.

During testing, the system identifies malignant lesions as anomalies based on their high reconstruction error, quantified by a hybrid loss function combining Mean Squared Error (MSE) and Structural Similarity Index (SSIM). The methodology incorporates rigorous preprocessing, including digital hair removal and normalization, followed by a deep convolutional architecture with a dense bottleneck to enforce feature compression. Experimental results demonstrate the system's ability to distinguish between benign and malignant classes by establishing a statistical anomaly threshold derived from the benign validation set. This report details the architectural design, implementation strategy, and performance evaluation of the proposed anomaly detection system, highlighting its potential as a robust, data-efficient diagnostic support tool.

Contents

1. Introduction

Skin cancer is one of the most prevalent forms of malignancy worldwide, necessitating the development of accurate, efficient, and accessible diagnostic tools. While dermatoscopy has significantly improved diagnostic precision, it remains subjective and dependent on clinician expertise. In recent years, Computer-Aided Diagnosis (CAD) systems powered by Deep Learning (DL) have emerged as promising assistants in clinical workflows. However, the majority of existing solutions rely on supervised classification models. These models face a significant bottleneck: they require massive, balanced datasets containing thousands of annotated examples for every specific type of cancer. In real-world medical scenarios, data is inherently imbalanced; benign cases (non-cancerous lesions) vastly outnumber malignant ones, leading supervised models to bias toward the majority class or requiring artificial data augmentation that may introduce artifacts.

This project addresses the class imbalance problem by shifting the paradigm from classification to anomaly detection. Instead of teaching the model what "cancer" looks like, we teach it what "normal" looks like [2]. We propose an unsupervised Deep Learning framework utilizing a Convolutional Autoencoder (CAE) trained exclusively on benign skin lesions sourced from the International Skin Imaging Collaboration (ISIC) dataset obtained online. The core hypothesis is that a network trained to efficiently compress and reconstruct healthy skin textures will fail to accurately reconstruct malignant lesions, which often exhibit irregular borders, asymmetry, and chaotic color variation (the "ABCDE" criteria).[3]

The remainder of this report is structured to present the complete lifecycle of the proposed solution. The *Methods* section details the technical implementation, beginning with the preprocessing pipeline which employs OpenCV-based morphological operations for digital hair removal to reduce noise. It further describes the specific Convolutional Autoencoder architecture, featuring a symmetrical encoder-decoder structure with a dense bottleneck designed to capture high-level semantic features of benign skin, and explains the hybrid loss function – a weighted combination of Mean Squared Error (MSE) and Structural Similarity Index (SSIM) – selected to prioritize structural integrity. Following this, the Results section presents the quantitative performance of the model, analyzing the distribution of reconstruction errors across benign and malignant test sets, the effectiveness of the statistical thresholding method, and the generated "error heatmaps" that localize suspicious regions. Finally, the Discussion interprets these findings in the context of clinical applicability, addressing the trade-offs between sensitivity and specificity, the limitations of the current unsupervised approach compared to fully supervised models, and potential avenues for future optimization.

2. Methodology

The proposed framework operates on the principle of unsupervised anomaly detection. Unlike supervised classification, which discriminates between labeled classes, our approach utilizes a Convolutional Autoencoder (CAE) to model the statistical distribution of “normal” (benign) skin texture. The methodology is divided into data partitioning, preprocessing, network architecture, and the hybrid error optimization strategy.

Data Partitioning and Anomaly Logic

The core hypothesis of this study is that a neural network trained exclusively on benign skin lesions will learn to compress and reconstruct healthy skin features (e.g., regular pigment networks, smooth borders) with high fidelity. Because the model is never exposed to malignant cases (e.g., Melanoma or Basal Cell Carcinoma) during the training phase, it lacks the latent capacity to encode the complex, chaotic visual structures associated with malignancy [4, 5, 6, 7].

Consequently, when a malignant image is passed through the trained CAE, the system attempts to reconstruct it using only the learned “benign” feature set. This results in a significant reconstruction error (residual). We partition the ISIC dataset such that the training set X_{train} consists exclusively of benign classes (Nevus, Dermatofibroma, etc.), while the test set X_{test} contains both unseen benign samples and all malignant samples. The separation ensures that the Autoencoder acts as a one-class classifier, where high reconstruction error signifies an anomaly.

Data Preprocessing

Prior to entering the network, raw dermoscopic images undergo a rigorous preprocessing pipeline to remove artifacts that could hinder feature learning.

1. **Digital Hair Removal:** Hair strands in dermoscopy images introduce high-frequency noise. We apply a morphological Black-Hat transformation to isolate dark hair structures against the lighter skin background. These structures are thresholded to create a binary mask, and the occluded pixels are restored using Telea’s inpainting algorithm (Navier-Stokes based equation).

2. **Normalization:** Images are resized to 128×128 pixels and pixel intensities are normalized to the range $[0, 1]$ to ensure numerical stability during gradient descent.

Convolutional Autoencoder Architecture

The architecture follows a symmetrical encoder-decoder structure designed to force dimensionality reduction, compelling the network to learn semantic features rather than memorizing pixel values.

Encoder

The encoder compresses the input space $I \in \mathbb{R}^{128\times128\times3}$ into a low-dimensional latent vector. It consists of three convolutional blocks. Each block applies a 2D convolution operation followed by a Rectified Linear Unit (ReLU) activation and Max-Pooling. Mathematically, for the l-th layer with input x, the convolution operation is:

$$y^l_{i,j,k} = \text{ReLU}\left(\sum_{u,v,c} W^l_{u,v,c,k} \cdot x^{l-1}_{i+u,j+v,c} + b^l_k\right) \tag{1}$$

where W represents the learnable kernels (filters of size 3×3) and b is the bias. The encoder progressively increases feature depth (32 $\rightarrow$ 64 $\rightarrow$ 128 filters) while reducing spatial resolution via 2×2 Max-Pooling.

Bottleneck

The flattened output of the final encoder block is passed to a fully connected (Dense) layer with 128 neurons. This bottleneck represents the latent space z, compressing the input by a factor of approximately 250 (from 32,768 features to 128). An $L1$ activity regularizer ($\lambda = 10^{-7}$) is applied to this layer to enforce sparsity, ensuring that only the most salient features of benign skin are encoded.

Decoder

The decoder mirrors the encoder to reconstruct the image $\hat{I}$ from the latent vector z. It employs Transposed Convolutions (Deconvolution) to learn the upsampling parameters. The final layer uses a Sigmoid activation function to constrain the output pixel values to the valid range $[0, 1]$.

Hybrid Loss Function

To train the model, we employ a hybrid loss function that balances pixel-wise accuracy with structural integrity. Standard Mean Squared Error (MSE) often produces blurry reconstructions. To mitigate this, we combine MSE with the Structural Similarity Index (SSIM). The total loss function L_{total} is defined as:

$$L_{total} = w_{mse} \cdot \mathcal{L}_{MSE} + w_{ssim} \cdot \mathcal{L}_{SSIM} \tag{2}$$

where the weights are empirically set to $w_{mse} = 0.4$ and $w_{ssim} = 0.6$.

Mean Squared Error (MSE)

Measures the average squared difference between the input pixels x and reconstruction $\hat{x}$:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \tag{3}$$

Structural Similarity Index (SSIM)

The Structural Similarity Index captures changes in luminance (l), contrast (c), and structure (s). The loss is derived as $1 - \text{SSIM}$, ensuring the model minimizes structural divergence:

$$\mathcal{L}_{SSIM} = 1 - \frac{(2\mu_x\mu_{\hat{x}} + C_1)(2\sigma_{x\hat{x}} + C_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + C_2)} \tag{4}$$

[8] This hybrid approach ensures the model is sensitive to the structural irregularities (e.g., jagged borders) typical of malignant lesions.

Anomaly Scoring and Threshold Determination

The final stage of the pipeline involves defining a decision boundary to classify an image as benign or malignant based on its reconstruction error. After training, the model processes the benign validation set to establish a statistical baseline. The anomaly threshold T is defined as:

$$T = \mu_{benign} + \alpha \cdot \sigma_{benign} \tag{5}$$

where μ_{benign} and σ_{benign} are the mean and standard deviation of the reconstruction scores for healthy skin. The coefficient α acts as a sensitivity parameter. While standard outlier detection often employs $\alpha = 3$ (the three-sigma rule), our empirical analysis revealed that the variance of reconstruction errors for malignant lesions significantly exceeds that of benign lesions ($\sigma_{malignant} \approx 2\sigma_{benign}$). Consequently, a standard wide margin would fail to detect subtle anomalies. To address this, we optimized α empirically to approximately 1.3. This value was derived based on the ratio of class deviations, specifically observing that a tighter threshold is necessary when the anomaly class exhibits higher volatility. This ensures the system remains sensitive to malignant features that deviate only slightly from the learned normal distribution.

3. Results

The performance of the proposed anomaly detection framework was evaluated on the ISIC dataset using the hybrid MSE-SSIM loss function. The Convolutional Autoencoder (CAE) was trained on benign samples, and the best-performing model was saved as `SkinCancerAutoencoder.h5` for validation. The results are categorized into quantitative metrics, statistical distribution analysis, and qualitative visual inspection.

Quantitative Performance

The system's ability to discriminate between benign and malignant lesions was tested on a separate test set containing 64 malignant and 54 benign images. Using the empirically optimized threshold ($T = \mu_{benign} + 1.3\sigma_{benign}$), the model achieved the following classification accuracies:

- **Malignant Detection Sensitivity:** 65.62% (42 out of 64 malignant cases correctly identified as anomalies).
- **Benign Specificity:** 57.41% (31 out of 54 benign cases correctly classified as normal).

While the sensitivity demonstrates the model's capability to flag nearly two-thirds of cancerous lesions without ever seeing a malignant training sample, the specificity indicates a moderate rate of false positives. This trade-off is consistent with the anomaly detection paradigm, where unseen benign variations can sometimes trigger high reconstruction errors.

Statistical Analysis of Reconstruction Errors

A key validation of our hypothesis lies in the statistical separation between the reconstruction error distributions of the two classes. The model, having learned the latent representation of normal skin, yielded significantly lower error scores for benign samples compared to malignant ones.

Metric	**Benign (Normal)**	**Malignant (Anomaly)**
Mean Reconstruction Score (μ)	0.1041	0.1727
Standard Deviation (σ)	0.0308	0.0629

Table 3.1: Reconstruction Error Statistics

Notably, the standard deviation for malignant scores ($\sigma_{mal} \approx 0.0629$) is more than double that of the benign scores ($\sigma_{ben} \approx 0.0308$). This confirms that malignant lesions exhibit far greater structural variance and unpredictability, whereas benign lesions cluster more tightly around the learned "normal" manifold. On the training set validation, this thresholding logic successfully filtered 89.65% of benign images (1013/1130) as non-anomalous, while correctly flagging 64.29% of the malignant training samples (713/1109) as anomalies.

Correct Classifications

To evaluate the model's performance in a real-world inference scenario, an MLOps testing pipeline was developed to automate image preprocessing, inference, error map generation, and anomaly scoring for individual samples. Qualitative inspection of these outputs provides critical insight into the model's decision-making process.

The Error Heatmaps presented below visualize the pixel-wise difference between the input and the reconstruction. The color spectrum ranges from Blue (low error, high similarity) to Red (high error, low similarity). Red regions indicate structural anomalies where the autoencoder failed to compress and restore the visual features, signaling potential malignancy.

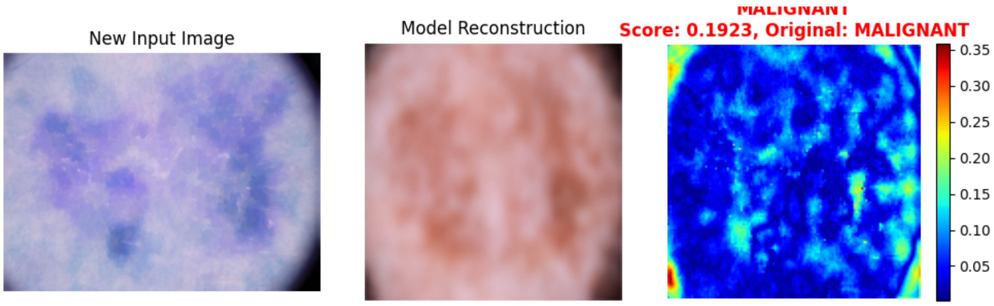

Fig. 3.1: True Positive (Malignant detected as Anomaly)

As seen in Fig. 3.1, the model fails to reconstruct the irregular, chaotic texture of the malignant lesion. The heatmap shows intense red regions around the borders and center, resulting in a high anomaly score of **0.1923**. This confirms our hypothesis that malignant lesions possess high-frequency spatial details and irregular boundaries that the autoencoder—trained only on smooth, regular benign skin—struggles to reproduce, leading to high residuals.

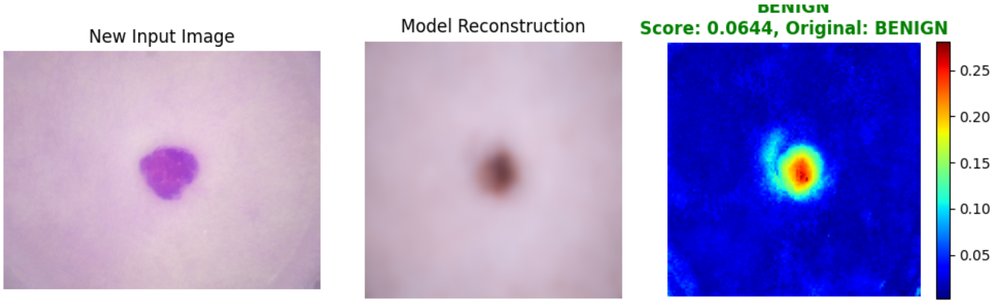

Fig. 3.2: True Negative (Benign correctly identified as Normal)

Conversely, Fig. 3.2 demonstrates the model's ability to successfully reconstruct the smooth, regular structure of a benign lesion. The error map is predominantly blue, indicating that the input and output are nearly identical. The resulting score of **0.0644** falls well below the anomaly threshold, correctly validating the lesion as non-cancerous.

Misclassifications

Analyzing where the model fails is equally important for understanding its limitations.

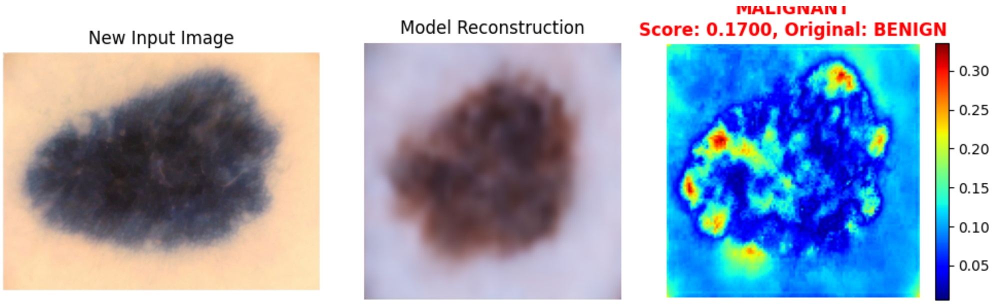

Fig. 3.3: False Positive (Benign misclassified as Malignant)

Fig. 3.3 illustrates a **False Positive**, where a benign lesion generated a high anomaly score of **0.1700**. The heatmap reveals that the model struggled with the lesion's large size and the complex, hair-like artifacts at the periphery which the preprocessing step may have failed to fully remove. The autoencoder interpreted these unlearned textural complexities as anomalies, triggering a false alarm.

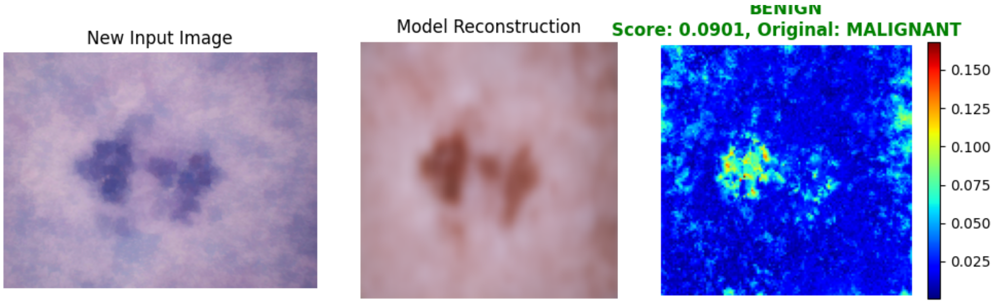

Fig. 3.4: False Negative (Malignant misclassified as Benign)

Fig. 3.4 shows a **False Negative**, a critical failure case where a malignant lesion produced a low score of **0.0901**. Visually, this specific malignant sample lacks the chaotic "ABCDE" features typical of cancer; it appears relatively smooth and symmetrical. Because its visual features overlap significantly with the learned distribution of benign skin, the autoencoder reconstructed it with high fidelity, failing to flag it as an anomaly. This highlights a limitation of unsupervised methods: if a cancer looks "too normal," it may escape detection.

4. Discussion

The methodolgy is implemented using Python with TensorFlow and Keras libraries for building and training the autoencoder model. The dataset used consists of benign and malignant skin lesion images, which are preprocessed and normalized before being fed into the model.

The dataset is split into training and testing sets to evaluate the model's performance. The images are coverted from BGR to RGB format, hair removal is performed, and the images are resized to a uniform size of 128x128 pixels. The pixel values are normalized to the range [0, 1] to facilitate better training of the neural network.

The autoencoder weights for SSIM and MSE losses are iterated over multiple values to find the optimal combination that yields the best reconstruction quality. The model is trained for 50 epochs with a batch size of 32, and the training process is monitored using validation data to prevent overfitting.

- 0.2 SSIM and 0.8 MSE weights tested.
- 0.4 SSIM and 0.6 MSE weights tested.
- 0.6 SSIM and 0.4 MSE weights tested.
- 0.8 SSIM and 0.2 MSE weights tested.

The performance of the trained autoencoder is evaluated using metrics such as Mean Squared Error (MSE) and malignant marking accuracy. The best performing model is 0.6 SSIM and 0.4 MSE weights. The performance tested against the hair removal option. The hair removal option showed better results in terms of reconstruction quality and malignant marking accuracy compared to the model trained without hair removal. This indicates that preprocessing steps like hair removal can significantly impact the effectiveness of the autoencoder in reconstructing skin lesion images. The model is tested for overfitting by setting:

- Total Epochs: 100 and Batch Size: 8
- Total Epochs: 25 and Batch Size: 64

The results indicated that the mean error on the square error is most distinct when the model is trained for 50 epochs with a batch size of 32 and hair removal preprocessing.

The ISIC dataset includes so close and similar images and there are many imbalanced classes in the dataset. This makes it difficult for the autoencoder to generalize well on unseen data, leading to overfitting easily.

The model's perforamce was not satisfactory on the first sight. This lead to the test of different approaches such as CNN clasifier.

By removing decoder and adding dense layers for classification, the model is converted to a CNN classifier. The CNN classifier showed worse performance such that average 13.72% accuracy on the test set. This test raised the question that whether the autoencoder is a suitable approach for this problem.

Literary research is done on the ISIC dataset. The repositories on Github that used same dataset are examined. The similar studies showed that the dataset is mostly turned into binary classification problem [9, 10]. Basic approaches are mostly focused on classifying benign and malignant images becasue the imbalanced classes in the dataset. Academic searches use complex structures like Resnet and EfficientNet [11, 12, 11, 13]. Gemini LLM's report on the dataset is indicated that without using complex structures like Resnet or Ensemble it is not possible to achieve classification into each disease.

Not to break the exprimentation in this project these found sources are not used.

The autoencoder structure is kept and a better decision boundary is tried to be found by measuring mean hybrid error on the reconstructed images and their standard deviations. The benign dataset shows a lower mean hybrid error compared to the malignant dataset and standard deviation is also lower for benign dataset. This allowed us to chose a decision boundary based on mean and standard deviation of benign dataset which is closer to benign mean error.

In conclusion, the autoencoder model with around 60% accuracy is actually a good starting point for this complex dataset with a book definition autoencoder structure.

Bibliography

[1] "Skin cancer (9 classes) — isic," Kaggle, accessed: Dec. 13, 2025. [Online]. Available: https://www.kaggle.com/datasets/nodoubttome/skin-cancer9-classesisic/data

[2] Cleveland Clinic, "Skin lesions," Sep 2021, accessed: Dec. 12, 2025. [Online]. Available: https://my.clevelandclinic.org/health/diseases/24296-skin-lesions

[3] National Cancer Institute, "Moles & melanoma tool," accessed: Dec. 12, 2025. [Online]. Available: https://moles-melanoma-tool.cancer.gov/

[4] Mayo Clinic, "Basal cell carcinoma," accessed: Dec. 13, 2025. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/basal-cell-carcinoma/symptoms-causes/syc-20354187

[5] Cleveland Clinic, "Carcinoma," accessed: Dec. 13, 2025. [Online]. Available: https://my.clevelandclinic.org/health/diseases/23180-carcinoma

[6] Mayo Clinic, "Melanoma," accessed: Dec. 13, 2025. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/melanoma/symptoms-causes/syc-20374884

[7] The Skin Cancer Foundation, "Actinic keratosis," accessed: Dec. 13, 2025. [Online]. Available: https://www.skincancer.org/skin-cancer-information/actinic-keratosis/

[8] A. K. Pandey, "Structural similarity index(ssim)," Medium, accessed: Dec. 13, 2025. [Online]. Available: https://medium.com/@akp83540/structural-similarity-index-ssim-c5862bb2b520

[9] M. Toğaçar, Z. Cömert, and B. Ergen, "Intelligent skin cancer detection applying autoencoder, mobilenetv2 and spiking neural networks," *Chaos, Solitons & Fractals*, vol. 144, p. 110714, 3 2021.

[10] M. Organokov, "kabartay/kaggle-siim-isic-melanoma-classification: Kaggle competition to identify melanoma in lesion images." 2021. [Online]. Available: https://github.com/kabartay/kaggle-siim-isic-melanoma-classification

[11] R. Maurya *et al.*, "Skin cancer detection through attention guided dual autoencoder approach with extreme learning machine," *Scientific Reports*, vol. 14, no. 17785, Aug 2024, accessed: Dec. 12, 2025. [Online]. Available: https://www.nature.com/articles/s41598-024-68749-1

[12] T. Tomić, I. Marković, and M. Gruić, "Deep learning methods for skin lesion classification and segmentation: A review," *Biomedicines*, vol. 12, no. 4, p. 326, Apr 2024, accessed: Dec. 13, 2025. [Online]. Available: https://www.mdpi.com/2306-5354/12/4/326

[13] M. Yilmaz and S. Demir, “Deep learning-based skin lesion analysis using convolutional features,” *JISEM - Journal of Information Systems Engineering & Management*, vol. 8, no. 2, pp. 45–60, 2024, accessed: Dec. 13, 2025. [Online]. Available: https://jisem-journal.com/index.php/journal/article/view/1238