Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
19/2 (2026), 141-156. DOI: http://dx.doi.org/10.21609/jiki.v19i2.1598

Comparative Analysis on Data Balancing and Augmentation in Skin Cancer Image Classification Using Multiple Datasets with Explainable AI

Dimas Fanny Hebrasianto Permadi, Annisaa Utami, Muhammad Raafi'u Firmansyah

Faculty of Informatics, Telkom University, Purwokerto, Indonesia
Jl. DI Panjaitan No.128, Kabupaten Banyumas, Jawa Tengah, 53147, Indonesia

E-mail: dimasfhp@telkomuniversity.ac.id

Abstract

Skin cancer is one of the most common types of cancer worldwide, with a continuously increasing incidence rate. Early detection and accurate classification of skin lesions are crucial for improving patient survival, particularly for melanoma. This study presents a comparative analysis of data-balancing strategies for skin cancer image classification across multiple publicly available datasets, including ISIC 2024, ISIC 2019, ISIC 2020, HAM10000, and PROVe-AI. The ConvNeXt-Tiny architecture is employed as the classification model and evaluated under three dataset configurations, namely RAW without balancing, Undersampling, and Oversampling via data augmentation, combined with two Learning Rate (LR) settings of 0.001 and 0.0001. In addition to reporting classification performance, this study emphasizes a comparative evaluation of accuracy, training stability, and computational efficiency across different balancing strategies. The experimental results indicate that the RAW dataset with an LR=0.0001 provides the best trade-off between performance and efficiency, achieving a validation accuracy of 99.88% and an F1-score of 0.9988. Oversampling via augmentation achieves the highest performance, with a validation accuracy of 99.93% and an F1-score of 0.9993, but requires substantially higher computational resources. Undersampling enables faster training with lower resource consumption, although it results in a slight performance degradation. Furthermore, Explainable Artificial Intelligence techniques, including Grad-CAM and LIME, demonstrate that models trained with an appropriate learning rate consistently focus on clinically relevant lesion regions, thereby improving interpretability and trustworthiness.

Keywords: *Performance Evaluation, Skin Cancer, Data Balancing, ConvNeXt, GradCAM-XAI, LIME-XAI*

1. Introduction

Skin cancer is one of the most common types of cancer worldwide, and its incidence continues to increase every year. Early detection and accurate classification of skin lesions play a crucial role in improving patient survival rates, as malignant skin cancer, such as melanoma, can be life-threatening if not diagnosed at an early stage [1–3]. In recent years, deep learning, particularly Convolutional Neural Networks (CNNs) [4–7], has demonstrated remarkable performance in medical image analysis [2, 8–10], including dermatology. CNN-based models are capable of learning discriminative features directly from images, reducing the need for handcrafted features traditionally used in computer-aided diagnosis systems [8, 10, 11].

Despite these advances, several challenges remain. One of the most critical issues is the class imbalance problem in medical image datasets. In skin cancer datasets, benign cases often vastly outnumber malignant cases, causing CNN models to be biased toward majority classes. In Figure 1, it can be seen that the dataset from ISIC 2024 [12] has data class imbalance. This imbalance reduces the ability of models to generalize and accurately identify minority classes, which in this context are often the most clinically significant [13–15]. Several strategies have been proposed to address imbalances, such as undersampling, oversampling, and data augmen-

tation [13–15]. However, comparative evaluations across different balancing methods remain limited, particularly when applied to state-of-the-art CNN architectures.

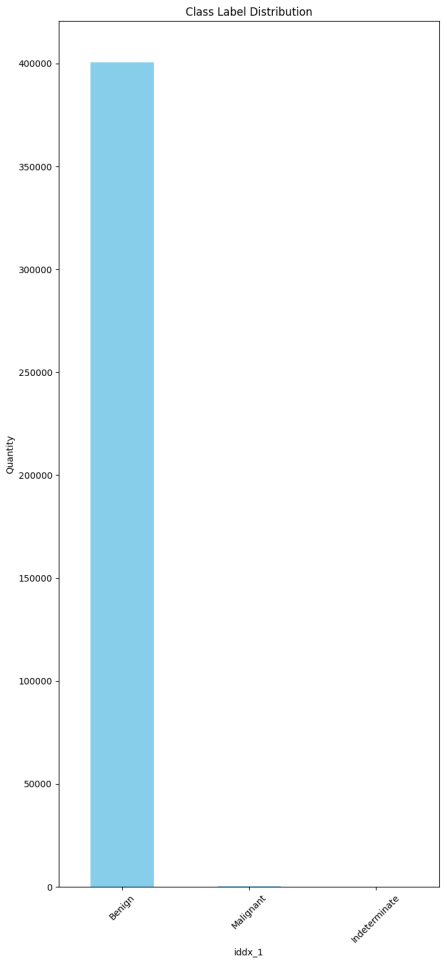

Figure 1. Class distribution of Benign, Malignant, Indeterminate.

Another important challenge is the lack of interpretability of deep learning models in medical applications [6]. Although CNNs can achieve high classification accuracy, they are often criticized as ”black boxes” [16], making it difficult for clinicians to trust and validate the decision-making process [17]. To address this issue, Explainable Artificial Intelligence (XAI) [18] methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) [11, 18], Local Interpretable Model-agnostic Explanations (LIME) [19], and SHapley Additive exPlanations (SHAP) [20, 21] have been introduced. These approaches provide visual or feature-level explanations that highlight important regions of skin lesions used in classification, thereby improving the interpretability and trustworthiness of AI systems in clinical practice.

In this study, ConvNeXt-Tiny is employed as the primary classification model due to its modern convolutional architecture, which integrates design principles inspired by vision transformers while maintaining the efficiency of conventional convolutional neural networks. This characteristic makes ConvNeXt-Tiny particularly suitable for large-scale medical image classification tasks that require both high accuracy and reasonable computational cost. ConvNeXt adopts a purely convolutional design, incorporating several architectural refinements inspired by transformer models [22], such as large kernel sizes, inverted bottlenecks, and layer normalization. Compared to conventional CNNs such as ResNet and DenseNet, ConvNeXt demonstrates superior performance on large-scale vision benchmarks (e.g., ImageNet [23]) while maintaining computational efficiency, making it particularly suitable for medical imaging tasks where both accuracy and resource efficiency are essential. The Tiny variant is chosen as it provides an optimal balance between performance and computational cost, allowing its application in real-world clinical environments with limited hardware resources [24, 25].

Therefore, this research proposes a comprehensive evaluation of ConvNeXt Tiny for skin cancer image classification, addressing both performance optimization through data RAW, Undersampling, and Oversampling via Augmentation dataset configurations and model interpretability through XAI integration. This study makes several important contributions to the field of skin cancer image classification. First, it provides a comprehensive comparative analysis of different dataset configurations, including RAW data, Undersampling, and Oversampling via data augmentation, using multiple publicly available skin lesion datasets. Second, it evaluates the performance of the ConvNeXt-Tiny model across varying data-balancing strategies and learning rate settings, with a particular focus on classification accuracy, training stability, and computational efficiency. Third, this study analyzes the trade-offs between performance and resource consumption to support practical deployment considerations in clinical environments. Finally, Explainable Artificial Intelligence techniques, namely Grad-CAM and LIME, are integrated to validate model predictions and enhance interpretability, thereby improving the clinical reliability of the proposed approach.

2. Research Method

This section describes the research workflow, data sources, preprocessing, balancing strategies as the main contribution, ConvNeXt-Tiny training scenarios, evaluation protocol, XAI analyses, and procedures for comparison and discussion. The descriptions in Figure 2 include details of implementations.

Figure 2 presents the overall research workflow used in this study. The RAW dataset in subsubsection 2.3.1 is obtained directly after preprocessing and serves as the baseline configuration without any balancing operations. The main contribution is a method for balancing class imbalance using Undersampling and Oversampling via the Augmentation Technique after preprocessing. Detailed explanation can be seen in subsection 2.3 and for each balancing strategy for Undersampling in subsubsection 2.3.2 and Oversampling via Augmentation in subsubsection 2.3.3. Each balancing strategy is trained and evaluated independently using the same model architecture and experimental settings to ensure a fair comparison.

2.1. Dataset Collection

The main dataset on this research is ISIC 2024 [12], which is explained in section 1. This dataset was combined with dataset from ISIC 2019 [26], ISIC 2020 [27], HAM10000 [28] and PROVe-AI [29]. For each dataset, include the same name and label for every class in Benign and Malignant. The additional dataset only adds the Malignant Class to balance the class distribution in the ISIC 2024 dataset. Additionally, we excluded the Indeterminate Class for optimal classification results. The combined results are shown in Figure 3, and they imply an imbalance in class conditions, with Benign labeled 0 and Malignant labeled 1.

2.2. Preprocessing

All processes are implemented by loading images, resizing, normalizing, cleaning, and saving metadata. The sample of Benign images is shown in Figure 4 and the malignant images are in Figure 5. All images are resized into 224×224 pixels for the standard ConvNeXt models. Also, normalize the images within Convert to PyTorch tensors and apply ImageNet normalization (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) only at the loader transform stage. The cleaning process drops NaN and Indeterminate class labels. All of them save into metadata for each name in CSV format.

2.3. Balancing Strategies

This study employs three dataset configurations, namely RAW, Undersampling, and Oversampling via Augmentation. The RAW dataset is obtained after preprocessing and does not involve any class-balancing operations. The Undersampling configuration randomly reduces the majority class to match the minority class sample count, thereby achieving a fully oversampled dataset via augmentation at the cost of reduced data diversity. The Oversampling via Augmentation configuration augments samples from the minority class to reduce class imbalance; however, due to practical constraints, a slight class imbalance persists. These three configurations are evaluated independently to enable a fair comparative analysis.

Three dataset configurations are used in this study, each with a dedicated metadata CSV file: RAW, Undersampling, and Oversampling via Augmentation. As illustrated in Figure 3, the RAW dataset comprises 400,513 benign and 14,829 malignant samples. In this study, undersampling and oversampling via augmentation strategies are applied only to their respective dataset configurations, while the RAW dataset is used solely as a baseline, without any class-balancing operations. The Undersampling results can be seen in Figure 6, while the Oversampling via Augmentation results are shown in Figure 7. After applying the undersampling strategy, both the benign (0) and malignant (1) classes contain 14,829 samples each. The results of the Oversampling via Augmentation in Benign (0) are 400,513 data, and Malignant (1) is 399,821 data. The dataset comparison graph can be seen in Figure 8.

2.3.1. RAW. The RAW is combined from only 5 datasets. This section has no process within augmentation or other processes. The process in this section occurs only in subsection 2.2. The results of class distribution in Figure 3.

2.3.2. Undersampling. This section is like what RAW has done. But this randomly samples the majority class down to the minority class count. The random state is set to 42 for reproducibility. The results of class distribution in Figure 6.

2.3.3. Oversampling via Augmentation. Examples of the augmentation results are shown in Figure 9 to illustrate the visual effects of the applied geometric and photometric transformations on minority-class samples. The process includes geometric transforms: RandomHorizontalFlip, RandomRotation (±20°), RandomResizedCrop (optional),

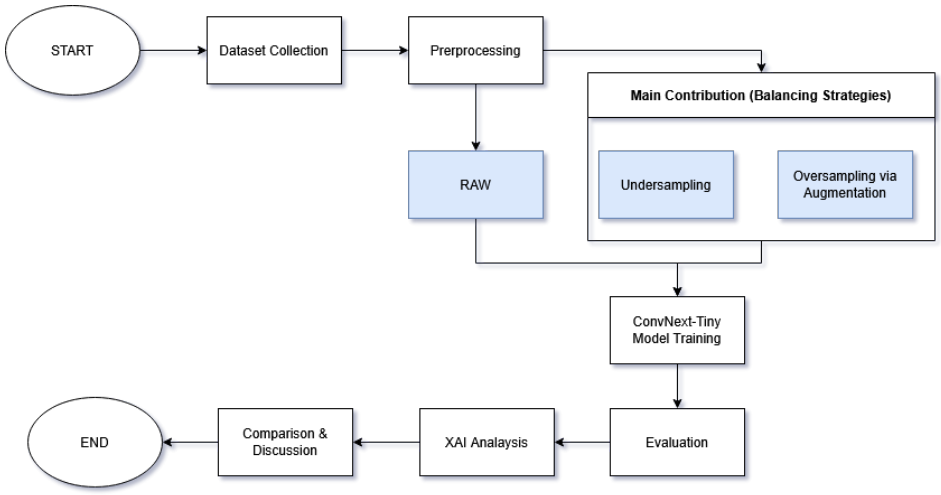

Figure 2. Research method.

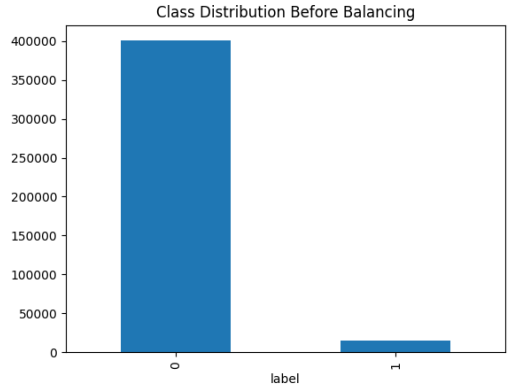

Figure 3. Class distribution after combined and before balancing.

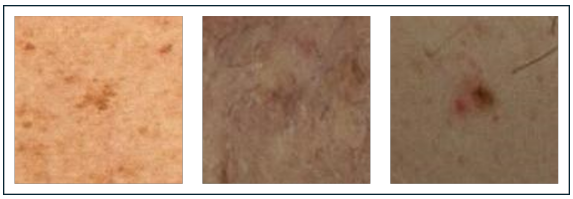

Figure 4. Sample images of Benign.

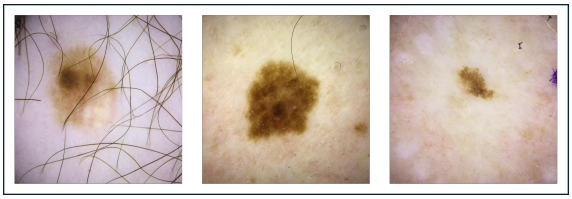

Figure 5. Sample images of Malignant.

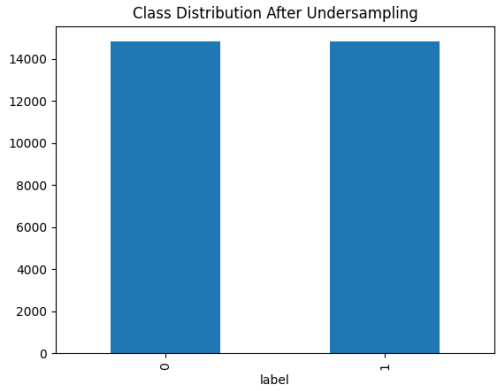

Figure 6. Undersampling class distribution.

RandomVerticalFlip (optional), photometric transforms: ColorJitter (brightness=0.2, contrast=0.2, saturation=0.15), blur variants (randomly apply one for each): GaussianBlur with kernel sizes chosen randomly from [3,5,7], BoxBlur with radius uniform in [0.5, 2.0], and Combined: GaussianBlur then BoxBlur with a smaller radius. For reproducibility, use a fixed seed and, where possible, deterministic augmentation sampling, and log the augmentation parameters used per generated image. The augmentation results can be seen in Figure 9. The results of class distribution in Figure 7.

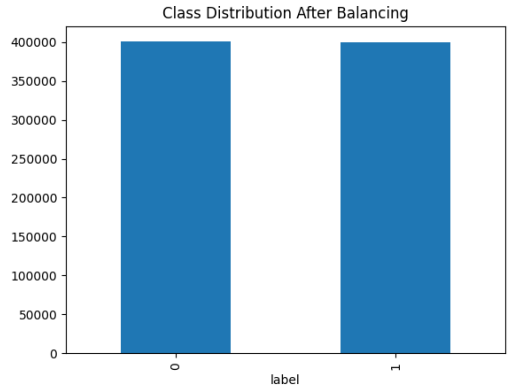

Figure 7. Class distribution after applying oversampling via augmentation.

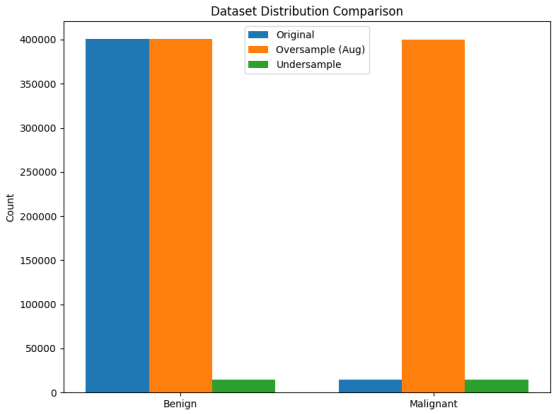

Figure 8. Dataset comparison.

2.4. ConvNeXt-Tiny Model Training

This section describes the architecture and training configuration of the ConvNeXt-Tiny model. The explanation is structured into two parts: model architecture and training scenario.

2.4.1. ConvNeXt-Tiny Model Architecture. ConvNeXt is a family of modern CNN models designed to be more efficient, simple, and modular. ConvNext can incorporate several design elements from the Transformer architecture (e.g., Layer Normalization, Multi-Layer Perceptron (MLP) block expansion), such as the Vision Transformer (ViT) and Swin Transformer. However, it still maintains the convolutional structure by using a 7×7 kernel, an inverted bottleneck from MobileNetV2, Layer Normalization replacing Batch Normalization, and the use of GELU instead of ReLU [30, 31].

Some important points of the ConvNeXt-Tiny design are the use of stem patchify style (replacing the initial large convolution layer with patch/stride = 4), the basic block uses depthwise convolution + pointwise convolution (inverted bottleneck style), the addition of LayerScale and stochastic depth as regularization, normalization of the LN layer (LayerNorm) instead of BatchNorm in some parts, and the block division scheme per stage follows a more Oversampling via Augmentation pattern than ResNet [31].

In Figure 10, the general block structure of ConvNeXt-Tiny is Depthwise convolution (kernel 7×7), Pointwise 1×1 convolution becomes expansion (e.g., factor 4), Activation (GELU), Pointwise 1×1 convolution becomes back projection, and Residual connection plus LayerScale and stochastic depth. A normalization layer (LayerNorm) is generally applied before or after convolution, depending on the implementation. The ConvNeXt-Tiny model also uses downsampling between stages, increasing the number of channels and decreasing the spatial resolution (as in most CNNs).

ConvNeXt-Tiny has advantages, including maintaining the computational efficiency of convolution (more efficient than attention in some scenarios) and Being Modular and flexible — easily used as a backbone for classification, detection, and segmentation. ConvNeXt-Tiny is also compatible with powerful training techniques (augmentation, regularization). ConvNeXt-Tiny also faces challenges due to its dependence on the number of parameters and dataset capacity. Selecting the right layer for XAI is difficult due to the residual structure and the presence of multiple blocks. Integration of self-supervised or transfer learning techniques sometimes requires architectural adjustments [22, 30, 31].

2.4.2. ConvNeXt-Tiny Model Scenario Training and Evaluation Metrics. In this research, model testing was conducted using the ConvNeXt Tiny architecture across several training scenarios to evaluate performance under varying datasets and hyperparameter settings. The datasets were divided into three forms: the original dataset (RAW), an Oversampling via Augmentation dataset generated via oversampling techniques, and a controlled undersampled dataset. Each dataset was then trained with the ConvNeXt Tiny model at learning rates of 1e-3 and 1e-4. The training process was run for 30 epochs, with evaluation metrics recorded at each iteration. These metrics included train loss, train accuracy, validation loss, validation accuracy, F1-Score, precision, recall, memory usage, and execution time per epoch. The best model in each scenario was saved in a ".pth" file

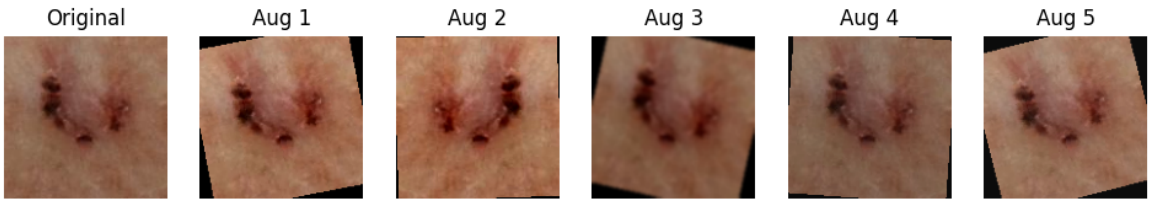

Figure 9. Augmentation results.

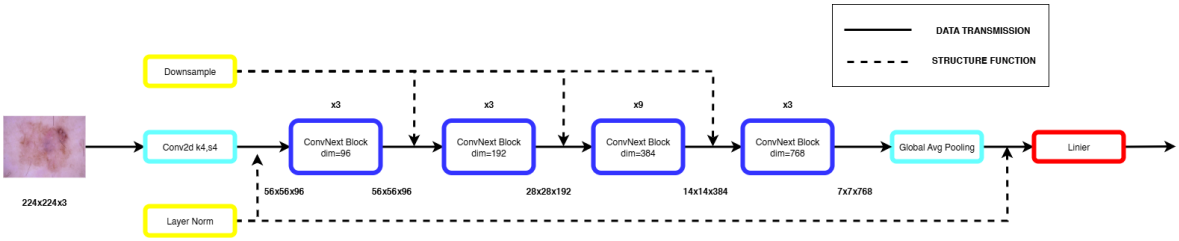

Figure 10. ConvNeXt-Tiny architecture.

for further analysis using the Explainable AI (XAI) method. With this scenario design, a comparative analysis can be conducted of the effects of dataset balancing strategies and learning rate variations on the generalization ability of ConvNeXt Tiny for skin cancer image classification.

2.5. Explainable AI (XAI) Methods

To ensure the CNN method using ConvNeXt-Tiny yields accurate results, we used XAI to analyze and demonstrate the possible outcomes of skin area images for lesion classes. This XAI method is useful for interpreting image outputs by highlighting images processed by the CNN model and providing a confidence value indicating the accuracy of the results. This XAI method uses Grad-CAM (Gradient-weighted Class Activation Mapping) and LIME (Local Interpretable Model-agnostic Explanations).

2.5.1. Grad-CAM (Gradient-weighted Class Activation Mapping). Grad-CAM is an XAI method that generates activation maps and uses the gradient of the target score to highlight salient regions that influence the CNN's decision. The basic principle of Grad-CAM is to utilize the gradient flowing from the output layer to the final convolutional layer to determine how much each spatial feature contributes to the prediction of a particular class. In other words, Grad-CAM produces a heatmap showing the areas in the image that most influence the classification decision. Mathematically, in Equation 1, Grad-CAM calculates the weighted importance of α_k^c for each map feature A^k in the last convolutional layer by averaging the gradients of the class scores y^c [11, 21, 32].

$$\alpha_k^c = \frac{1}{z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

The variable y^c is the score before softmax for the target class c, then A_{ij}^k is the activation of the position in (i, j) of the feature map k, and Z is the number of units in the feature map (normalized average gradient). Then, to select only those features with a positive influence on the class of interest, a weighted combination of the forward activation maps is performed, and the rectified linear unit (ReLU) function is applied in Equation 2:

$$L_{\text{GradCAM}}^c = \text{ReLU}(\sum_k \alpha_k^c A^k) \tag{2}$$

The resulting activation map is a weighted visualization of the regions of greatest importance for classification [21].

2.5.2. LIME (Local Interpretable Model-agnostic Explanations). Local Interpretable Model-agnostic Explanations (LIME) is a model-agnostic method that explains individual predictions by building simpler, interpretable models (e.g., linear regression) around specific input samples. The main idea is that while complex models such as CNNs or Transformers are difficult to interpret globally, their predictions

for a single data instance can be locally approximated by simpler models. LIME works by perturbing the input data (e.g., by masking or shuffling image pixels) and observing how these perturbations affect the model's output. In doing so, LIME generates a visual map that highlights which parts of the input contribute most to the prediction. This makes it particularly effective because it does not depend on a specific model architecture. Importantly, LIME does not attempt to provide a global explanation of the model; instead, it focuses on local fidelity by approximating the decision boundary near a single prediction. The mathematical formulation of LIME is given in Equation 3 [19, 33, 34].

$$\xi(x) = \text{argmin} L(f.g.\pi_x) + \Omega(g) \tag{3}$$

The variable g is the interpretable surrogate model that approximates the original complex model f. The term π_x is a proximity measure that assigns higher weights to perturbed samples closer to the original instance x. $\Omega(g)$ is a complexity measure to ensure the model g remains interpretable, and $L(f.g.\pi_x) + \Omega(g)$ is the loss function that measures how well the model approximates the original model f, which is achieved by Equation 4 [19, 33, 34].

$$L(f.g.\pi_x) + \Omega(g) = \sum_{z.z' \epsilon Z} \pi_z(z)(f(z) - g(z'))^2 \tag{4}$$

The variable z', represents the perturbed version of the original instance x in the interpretable feature space.

3. Results and Discussion

3.1. Training Environment and Experimental Setup

All experiments were conducted using Python 3.12 and PyTorch 2.4, with CUDA 12.9 for GPU acceleration. Models were trained with a batch size of 32, the Adam optimizer, the GELU activation function, and Cross-Entropy Loss at loss rates of 0.001 and 0.0001. The training and validation processes are carried out by dividing the data into 80% for training and 20% for validation. The training process was also conducted on an NVIDIA GeForce RTX 4090 GPU with 24 GB of VRAM, ensuring all experiments ran efficiently without memory constraints. The training environment was run on Ubuntu 24.04 LTS with the NVIDIA driver version 575.57.08.

3.2. Training and Validation Performance Analysis

This section observes the effect of the learning rate on model convergence. However, there are differences in validation losses between models and data scenarios in the Undersample, RAW, and Oversampling via Augmentation dataset types. The training process was run for 30 epochs for each dataset scenario (RAW, Oversampling via Augmentation, Undersampling) with two learning rates (0.001 and 0.0001). A summary of the overall performance results is shown in Table 1 and Figure 11. The computational and memory efficiency results are shown in Table 2 and Figure 12.

3.2.1. RAW Dataset. The training results of each model on Raw with a LR=0.001 in Figure 13 and Raw with a LR=0.0001 in Figure 14. The RAW dataset achieved near-perfect performance when trained with a LR=0.0001. The model achieved a validation accuracy of 99.89%, an F1-Score of 0.999, and nearly identical precision and recall. Training loss and validation loss were both extremely low, indicating strong convergence without signs of overfitting. In terms of computational efficiency, training required 1.8 GB of memory and 1220 seconds (20 minutes), making it highly efficient relative to other strategies. This shows that ConvNeXt-Tiny can handle the class imbalance in the RAW dataset without significant degradation in minority-class detection.

3.2.2. Oversampling Dataset (Augmentation). The Oversampling via Augmentation dataset demonstrates strong classification performance when trained with a LR=0.0001, achieving near-perfect validation accuracy and F1-score. However, this improvement is accompanied by a significant increase in computational cost due to the augmented dataset size.

The balancing strategy via augmentation produced similarly strong results, with a validation accuracy of 99.93% and an F1-Score of 0.9993 at LR=0.0001. Precision and recall remained at ≈0.999 with Oversampling via Augmentation. This demonstrates that augmenting the minority class can yield models as performant as those trained on the RAW dataset, with the added advantage of fairness across classes. However, this improvement comes at a high computational cost. The Oversampling via Augmentation dataset required 4,381–6,103 seconds (≈1h–1h 40m) for training, and memory usage ranged from 1.25 GB to 2.05 GB, higher than RAW. Moreover, when trained with LR=0.001, the

Table 1. Evaluation results of the ConvNeXt-Tiny model on various dataset scenarios.

Dataset	Epoch	LR	Train Loss	Train Accuracy (%)	Val Loss	Val Accuracy (%)	F1-Score	Val Precision	Val Recall
RAW	30	0.001	0.016	99.654	0.014	99.709	0.997	0.997	0.997
Oversampling via Augmentation	30	0.001	0.693	50.048	0.694	49.766	0.331	0.248	0.498
Undersample	30	0.001	0.054	98.431	0.075	97.807	0.978	0.979	0.978
RAW	30	0.0001	0.001	99.986	0.012	99.882	0.999	0.999	0.999
Oversampling via Augmentation	30	0.0001	0.000	99.992	0.008	99.899	0.999	0.999	0.999
Undersample	30	0.0001	0.004	99.873	0.097	98.988	0.990	0.990	0.990

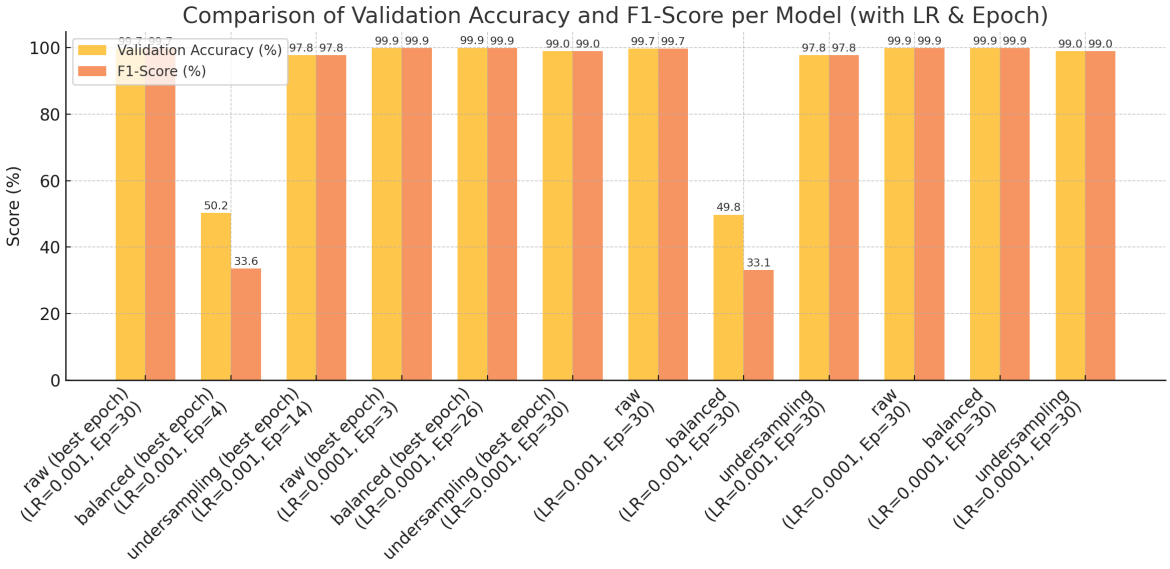

Figure 11. Comparison of validation accuracy and F1-Score per model (with LR & Epoch).

Table 2. Evaluation results of the ConvNeXt-Tiny model on computation time and memory.

Dataset	Epoch	LR	Memory (MB)	Time (s)
RAW	30	0.001	1810.25	1216.52
Oversampling via Augmentation	30	0.001	2066.90	4172.01
Undersample	30	0.001	772.82	194.94
RAW	30	0.0001	1812.43	1222.03
Oversampling via Augmentation	30	0.0001	1287.92	6102.93
Undersample	30	0.0001	1325.97	150.59

Oversampling via Augmentation model collapsed, achieving only ≈50% accuracy and an F1-Score of ≈0.33, highlighting the importance of careful hyperparameter tuning when using heavy augmentation. The Oversampling via Augmentation results in a LR=0.001 in Figure 15, Oversampling via Augmentation with a LR=0.0001 in Figure 16 in consecutive.

3.2.3. Undersampling Dataset. The undersampling strategy provided efficient training, with memory usage between 0.77–1.3 GB and training time of 150–287 seconds (≈2–5 minutes). However, the reduction in data diversity led to slightly lower performance than in the RAW and Oversampling via Augmentation modes. At the best LR=0.0001, the model achieved 98.99% accuracy, an F1-Score of 0.990, and precision and recall of 0.999 and 0.999, respectively, under Oversampling via Augmentation. While this accuracy is strong, the ≈1% gap compared to RAW/Oversampling via Augmentation models indicates the cost of discarding a significant portion of the dataset. Thus, undersampling is suitable for

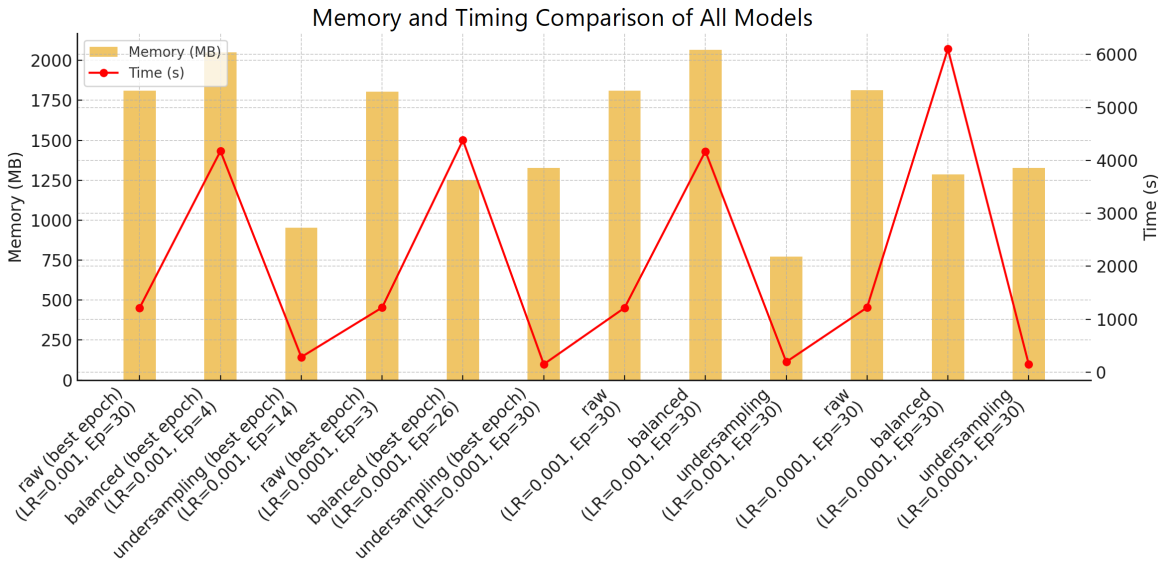

Figure 12. Memory and timing comparison of all models.

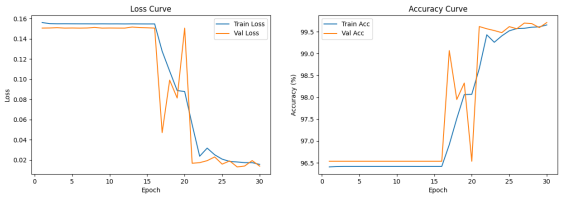

Figure 13. Training results dataset RAW with LR=0.001.

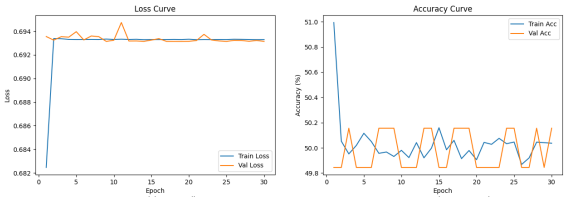

Figure 15. Training results dataset Oversampling via Augmentation with LR=0.001.

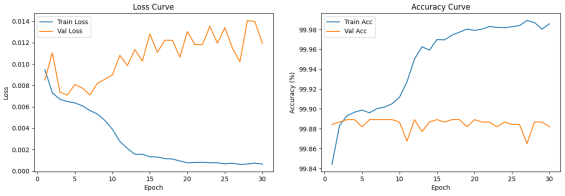

Figure 14. Training results dataset RAW with LR=0.0001.

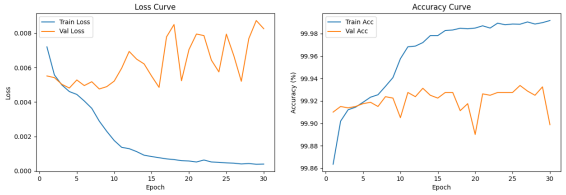

Figure 16. Training results dataset Oversampling via Augmentation with LR=0.0001.

resource-constrained scenarios but not optimal for maximizing classification performance. The Undersample results with a LR=0.001 in Figure Figure 17, Undersample with a LR=0.0001 in Figure 18.

3.2.4. Computational and Memory Efficiency.

Computational efficiency was analyzed by comparing GPU memory usage and average epoch execution time across model scenarios. The results showed significant differences between the models across the RAW, Oversampling via Augmentation, and Undersampling datasets, as well as between the best- and last-epoch conditions. The comparison can be seen in Figure 19 and Table 3.

In both RAW datasets, the average time per epoch was 40–41 seconds, with an average memory usage of around 60 MB/epoch. This indicates that the model achieved stable convergence without excessive computational overhead. This dataset is efficient and stable, with the best trade-off between accuracy and computational cost.

On the Oversampling via Augmentation dataset, the computational overhead increased significantly. For the best epoch (e.g., LR=(0.001) with 4 epochs),

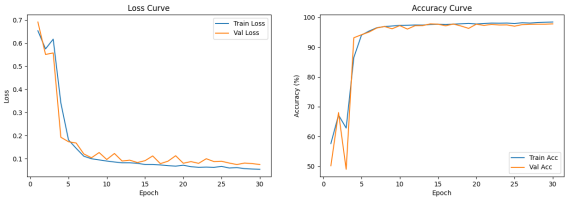

Figure 17. Training results dataset Undersample with LR=0.001.

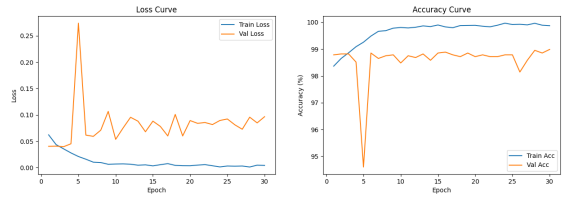

Figure 18. Training results dataset Undersample with LR=0.0001.

the average time per epoch reached >1000 seconds, and despite the small number of epochs, the memory overhead remained high (500 MB/epoch). This dataset is very slow, with high memory usage due to oversampling, and fails to improve accuracy (stuck at ≈50%). When the model was trained to the last epoch (30 epochs), the time per epoch remained high (>200 seconds) and the memory footprint was relatively large (40–50 MB/epoch). This is consistent with the theory that oversampling and augmentation increase the effective size of the dataset, making the training process slower and more memory-intensive. But, the Oversampling via Augmentation dataset with LR=0.0001 is still computationally heavy, performance is highly sensitive to the learning rate, and it achieves a more stable, reasonable time per epoch and high accuracy (≈99.9%).

Conversely, the Undersample dataset exhibits the most efficient computational performance. At the best epoch (14–30 epochs and LR=0.001), the average training time is only 5–20 seconds per epoch, with a smaller average memory footprint (25–45 MB/epoch). This is due to the reduction in most of the data, which shrinks the dataset size and makes the feed-forward and backpropagation processes lighter. However, this efficiency comes at the cost of a slight decrease in validation accuracy compared to the RAW dataset. The Undersample dataset with LR=0.0001 is the most efficient, providing very fast training with an acceptable accuracy of about 98%. Very efficient, maintains high accuracy (≈99%), but loses information about the majority class. Also, the results are consistently fast, but validation accuracy is slightly lower compared to the RAW dataset. Overall, these results show that the Oversampling via Augmentation dataset is the most computationally expensive, RAW is at an intermediate level with an optimal trade-off between accuracy and efficiency, and Undersample is the most efficient but may entail information loss due to data reduction.

3.3. XAI Results

To improve model interpretability, this study also used two XAI methods: GradCAM and LIME. This section presents an analysis of the ConvNeXt-Tiny model using three dataset variations, namely RAW, Undersampling, and Oversampling via Augmentation, combined with two learning rate values of 0.001 and 0.0001.

The prediction results in Table 4 indicate that all samples were correctly classified according to their original labels (true_label = 0 and pred_label = 0).

1) RAW dataset: very high (0.9945–0.9999).
2) Oversampling via Augmentation dataset: low for LR=0.001 (≈ 0.5055, close to random guessing), but maximum (1.0) at LR=0.0001.
3) Undersample dataset: high (0.9159–1.0).

This shows that the Oversampling via Augmentation dataset with LR=0.001 still fails to achieve stable learning, even though the model provides correct predictions with low confidence.

3.3.1. GradCAM Visualization. Based on the experimental results, the GradCAM visualization of the ConvNeXt-Tiny model shows a pattern consistent with the model's quantitative metrics (see Figure 20).

In the RAW dataset (0.001 & 0.0001), GradCAM exhibits a strong focus on the main lesion area. An LR of 0.001 indicates that the heatmap focuses on the lesion core, with the red area extensively covering the melanoma. This supports high confidence (0.9945–0.9981) and suggests the model can distinguish the lesion from the background. Meanwhile, an LR of 0.0001 indicates sharper activation, localized to the most critical parts of the lesion. Confidence is nearly perfect (0.9998–0.9999), indicating optimal convergence.

In the Oversampling via Augmentation Dataset, LR=0.001 indicates that the GradCAM Heatmap spreads to the background rather than focusing on the lesion. The Confidence value is low (≈0.5055). This is consistent with the validation accuracy of ≈50% and indicates underfitting. Meanwhile, LR=0.0001 indicates that activation becomes highly focused, with high heat intensity in the melanoma

Table 3. Summary of computational and memory efficiency across dataset variants.

Dataset Variant	Learning Rate	Epoch	Avg Time/epoch (s)	Avg Mem/epoch (MB)
RAW	0.001 / 0.0001	30 (best & last)	40–41	≈60
Oversampling via Augmentation	0.001 (best)	4 (best)	≈1044	≈512
Oversampling via Augmentation	0.0001 (best)	26 (best)	≈168	≈48
Oversampling via Augmentation	0.001 / 0.0001	30 (last)	≈200–210	≈43–45
Undersample	0.001 (best)	14 (best)	≈20.5	≈68
Undersample	0.0001 (best)	30 (best)	≈5	≈44
Undersample	0.001 / 0.0001	30 (last)	≈6–6.5	≈26–44

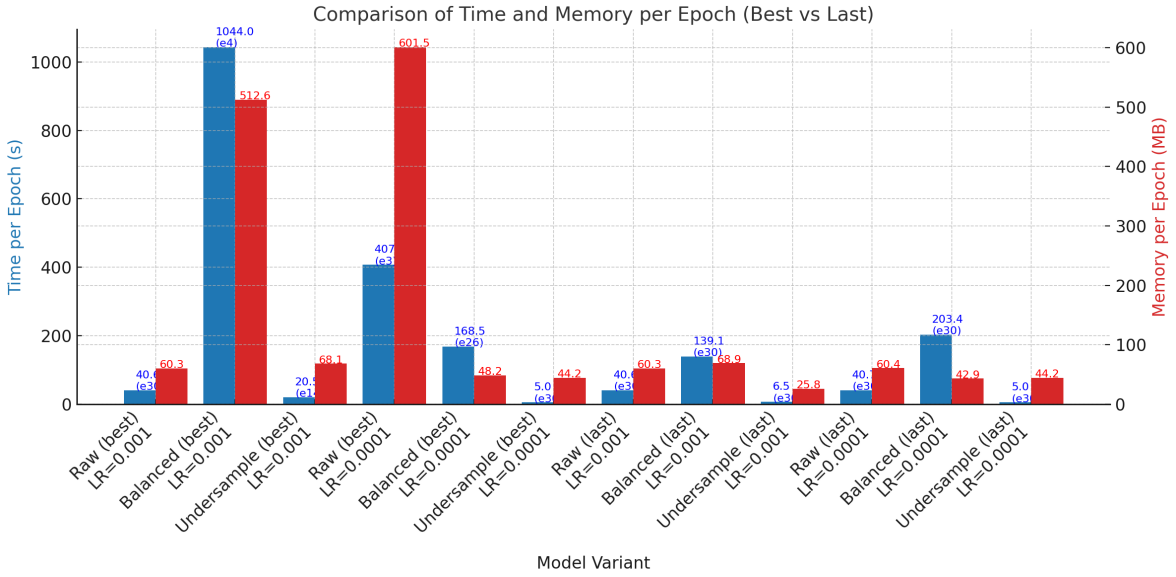

Figure 19. Average time and memory per Epoch (Best vs Last Epoch).

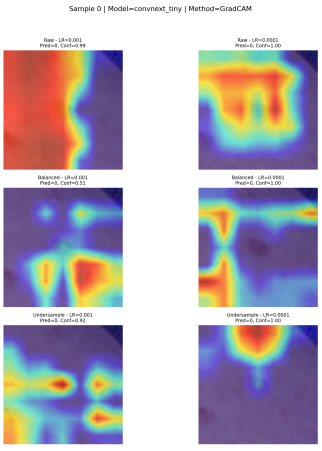

Figure 20. GradCAM visualization results.

area. Confidence reaches 1.0, demonstrating that a lower learning rate corrects errors in the Oversampling via Augmentation model.

In the Undersample Dataset, LR=0.001 indicates that the Heatmap remains focused on the lesion, but extends more widely (over-activation) to the background. Confidence value is quite high (0.9159–0.9924). This pattern indicates limited information due to benign data reduction. LR=0.0001 indicates that activation is more localized to the lesion core, consistent with a confidence value of 1.0. This demonstrates that despite limited data, the model can still learn critical features stably.

3.3.2. LIME (Local Interpretations). The summary results of the LIME analysis are shown in Figure 21. LIME displays the superpixels that contribute to the model's predictions, enabling it to assess whether its decisions are based on the lesion area or the background.

In the RAW Dataset, LR=0.001 indicates that the superpixels supporting the prediction are distributed precisely within the lesion area, with clear contrast boundaries against healthy skin. This aligns with high confidence (0.9945–0.9981). Meanwhile, LR=0.0001 indicates that the superpixels more closely follow the melanoma contour, with denser

Table 4. XAI prediction results using ConvNeXt-Tiny with GradCAM and LIME.

Sample	Model	Variant	LR	True Label	Pred Label	Confidence
0	ConvNeXt-Tiny	RAW	0.001	0	0	0.9945
	ConvNeXt-Tiny	RAW	0.0001	0	0	0.9999
	ConvNeXt-Tiny	Oversampling via Augmentation	0.001	0	0	0.5055
	ConvNeXt-Tiny	Oversampling via Augmentation	0.0001	0	0	1.0000
	ConvNeXt-Tiny	Undersample	0.001	0	0	0.9159
	ConvNeXt-Tiny	Undersample	0.0001	0	0	1.0000
1	ConvNeXt-Tiny	RAW	0.001	0	0	0.9981
	ConvNeXt-Tiny	RAW	0.0001	0	0	0.9998
	ConvNeXt-Tiny	Oversampling via Augmentation	0.001	0	0	0.5055
	ConvNeXt-Tiny	Oversampling via Augmentation	0.0001	0	0	1.0000
	ConvNeXt-Tiny	Undersample	0.001	0	0	0.9408
	ConvNeXt-Tiny	Undersample	0.0001	0	0	1.0000
2	ConvNeXt-Tiny	RAW	0.001	0	0	0.9981
	ConvNeXt-Tiny	RAW	0.0001	0	0	0.9999
	ConvNeXt-Tiny	Oversampling via Augmentation	0.001	0	0	0.5055
	ConvNeXt-Tiny	Oversampling via Augmentation	0.0001	0	0	1.0000
	ConvNeXt-Tiny	Undersample	0.001	0	0	0.9924
	ConvNeXt-Tiny	Undersample	0.0001	0	0	1.0000

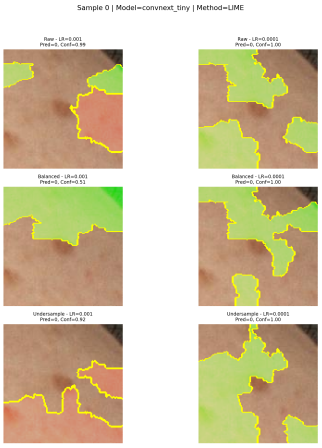

Figure 21. LIME visualization results.

coverage in the lesion core. Confidence is nearly perfect (0.9998–0.9999). This confirms that the model is not only accurate but also interpretable.

In the Oversampling via Augmentation Dataset, LR=0.001 indicates low confidence (≈0.5055). The superpixels in the image are spread out into the background and inconsistently highlight the lesion. This pattern supports underfitting. Meanwhile, LR=0.0001 indicates that confidence increases to 1.0. The superpixel results are more focused on dominant lesions, with a clear distribution in the cancerous areas. These results demonstrate that hyperparameter tuning is crucial for improving results on the Oversampling via Augmentation dataset.

On the Undersample Dataset, LR=0.001 indicates high confidence (0.9159–0.9924). The superpixels are sufficiently focused on the lesion area, although some small melanoma areas are not covered. The limited information can be explained by the cropped dataset. Meanwhile, LR=0.0001 indicates perfect confidence (1.0). The superpixels are denser in the target area, with segmentation consistently highlighting the lesions. This demonstrates that despite the smaller dataset, feature representations can still be learned effectively if the learning rate is optimized.

3.4. Discussion

Experimental results show that the learning rate is the most influential factor affecting convergence stability. Across all dataset variants, LR=0.0001 consistently yielded low training and validation losses and accuracy approaching 100%. While LR=0.001 proved too high for the Oversampling via Augmentation dataset, resulting in underfitting with a loss of ≈0.693 and accuracy approaching 50%. This phenomenon aligns with the theory of two-class cross-entropy loss, where random predictions yield loss values approaching log(2). The RAW dataset remains the best baseline because it is stable at both LR values, while the undersample, while still converging, exhibits higher validation loss due to limited data representation.

In terms of evaluation metrics, the Oversampling via Augmentation using LR=0.0001 combination provides the highest quantitative performance (Val Acc 99.93%, F1-Score 0.9993) while ensuring a more Oversampling via Augmentation class distribution, but at a higher computational cost. The RAW dataset with LR=0.0001 provided the best trade-off between near-perfect accuracy (Val Acc 99.88%) and efficiency, making it more practical for deployment. Undersampling achieved the highest computational efficiency, with the lowest epoch time and memory

requirements, but it showed slightly lower accuracy (Val Acc 98.99%, F1-Score 0.990) and was more susceptible to data bias. These results suggest that selecting a balancing strategy should prioritize fairness, performance, and efficiency.

XAI analysis (GradCAM and LIME) reinforced the quantitative findings with visual evidence. On the RAW dataset, with Oversampling via Augmentation and LR=0.0001, GradCAM produced a sharp heatmap focused on the lesion core. At the same time, LIME produced consistent superpixels around the melanoma, indicating that the model was indeed learning clinically relevant features. In contrast, Oversampling via Augmentation with LR=0.001 produced a diffuse heatmap and out-of-focus superpixels, indicating low confidence ($\approx$0.5055) and validation failure. On Undersample, GradCAM still highlights lesions, but more broadly (over-activation), while LIME does not always cover the entire lesion, reflecting the limitations of a small dataset. Overall, XAI demonstrates model interpretability that is consistent with confidence and validation metrics and confirms the model's trustworthiness for medical applications.

4. Conclusions and Future Works

4.1. Conclusions

Based on experimental results with the ConvNeXt-Tiny architecture across multiple dataset configurations, the learning rate plays a crucial role in determining model performance. A smaller LR=0.0001 consistently provides more stable training and better classification results than a larger LR. In contrast, a large LR=0.001 leads to underfitting on the Oversampling via Augmentation dataset, with an accuracy of only about 50.05%. The RAW dataset with LR=0.0001 is the best overall, offering an optimal trade-off between performance, stability, and computational efficiency (Val Acc 99.88%, F1-Score 0.9988). An Oversampling via Augmentation dataset with LR=0.0001 achieves the highest quantitative performance (Val Acc 99.93%, F1-Score 0.9993) and provides a more representative representation of the Malignant minority class. However, this strategy requires more time and memory. The undersampled dataset provides the best computational efficiency, with the fastest epoch time and the smallest memory usage. Still, accuracy decreases (Val Acc 98.99%, F1-Score 0.990) and shows indications of overactivation in GradCAM. Finally, XAI analysis (GradCAM and LIME) shows that the model with the optimal configuration really focuses on the lesion area. In contrast, the failed model (Oversampling via Augmentation with LR=0.001) exhibits a scattered, inconsistent pattern in its visualization. This emphasizes the importance of combining quantitative and interpretability metrics for medical model validation.

4.2. Future Works

Although the experimental results demonstrate strong classification performance, several limitations of the current study motivate future research directions. The data balancing strategy primarily relies on handcrafted augmentation techniques, which may not fully capture the complex morphological and textural variations of malignant skin lesions encountered in real clinical practice. Future work should therefore investigate alternative strategies to address class imbalance not only at the data level but also at the model and loss function levels. One promising direction is to incorporate class weighting into the loss function, enabling the model to penalize misclassifications of minority classes more heavily without increasing the dataset size. This approach could reduce computational overhead while improving sensitivity to malignant cases.

Another potential extension involves applying label smoothing techniques to mitigate overconfident predictions and improve generalization, particularly in heavily augmented datasets. Label smoothing may help stabilize training by preventing the model from becoming overly certain about noisy or synthetic samples generated through augmentation. In addition, future studies could explore the use of local or region-aware loss functions that emphasize learning from lesion-specific regions rather than treating all image pixels equally. Such loss formulations may better align the optimization process with clinically relevant visual cues, especially when combined with explainability-driven supervision.

From a data sampling perspective, stratified sampling strategies should be evaluated to ensure that Oversampling via Augmentation maintains class representation across the training and validation splits. This approach could improve the reliability of performance estimation and reduce sampling bias, particularly when working with highly imbalanced datasets aggregated from multiple sources. Furthermore, integrating stratified sampling with cross-validation protocols may yield more robust, statistically reliable comparisons of different balancing strategies.

In addition to methodological improvements, future research should incorporate formal statistical significance testing to support comparative conclusions between dataset configurations and training

strategies. Evaluating the proposed approach on multi-institutional and multi-ethnic datasets would further improve generalizability and clinical relevance. Finally, deeper integration of advanced Explainable Artificial Intelligence techniques, including hybrid or hierarchical explanation frameworks, may enhance clinician trust by providing more consistent and localized interpretations of model predictions.

References

[1] K. Hauser, A. Kurz, S. Haggenmüller, R. C. Maron, C. von Kalle, J. S. Utikal, F. Meier, S. Hobelsberger, F. F. Gellrich, M. Sergon, A. Hauschild, L. E. French, L. Heinzerling, J. G. Schlager, K. Ghoreschi, M. Schlaak, F. J. Hilke, G. Poch, H. Kutzner, C. Berking, M. V. Heppt, M. Erdmann, S. Haferkamp, D. Schadendorf, W. Sondermann, M. Goebeler, B. Schilling, J. N. Kather, S. Fröhling, D. B. Lipka, A. Hekler, E. Krieghoff-Henning, and T. J. Brinker, "Explainable artificial intelligence in skin cancer recognition: A systematic review," *European Journal of Cancer*, vol. 167, p. 54–69, May 2022. [Online]. Available: http://dx.doi.org/10.1016/j.ejca.2022.02.025

[2] A. Bello, S.-C. Ng, and M.-F. Leung, "Skin cancer classification using fine-tuned transfer learning of densenet-121," *Applied Sciences*, vol. 14, no. 17, p. 7707, Aug. 2024. [Online]. Available: http://dx.doi.org/10.3390/app14177707

[3] S. Benyahia, B. Meftah, and O. Lézoray, "Multi-features extraction based on deep learning for skin lesion classification," *Tissue and Cell*, vol. 74, p. 101701, Feb. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.tice.2021.101701

[4] D. F. H. Permadi and M. Z. Abdullah, "Implementation of animal face's recognition by convolutional neural network (cnn) algorithm," *JUTI: Jurnal Ilmiah Teknologi Informasi*, p. 29–39, Jan. 2023. [Online]. Available: http://dx.doi.org/10.12962/j24068535.v21i1.a1131

[5] H. Firat, H. Üzen, D. Hanbay, and A. Şengür, "Convnext mixer-based encoder decoder method for nuclei segmentation in histopathology images," *International Journal of Imaging Systems and Technology*, vol. 34, no. 5, Sep. 2024. [Online]. Available: http://dx.doi.org/10.1002/ima.23181

[6] B. A. Chakravarthi and G. Shivakanth, "¡scp¿speech2dementia¡/scp¿: A novel deep learning framework integrating enhanced ¡scp¿cnn¡/scp¿ and large language models for automatic detection of alzheimer's dementia," *Computational Intelligence*, vol. 41, no. 2, Apr. 2025. [Online]. Available: http://dx.doi.org/10.1111/coin.70051

[7] Z. Unnisa, A. Tariq, N. Sarwar, I. Din, M. A. Serhani, and Z. Trabelsi, "Impact of fine-tuning parameters of convolutional neural network for skin cancer detection," *Scientific Reports*, vol. 15, no. 1, Apr. 2025. [Online]. Available: http://dx.doi.org/10.1038/s41598-025-99529-0

[8] M. Fraiwan and E. Faouri, "On the automatic detection and classification of skin cancer using deep transfer learning," *Sensors*, vol. 22, no. 13, p. 4963, Jun. 2022. [Online]. Available: http://dx.doi.org/10.3390/s22134963

[9] R. C. Maron, S. Haggenmüller, C. von Kalle, J. S. Utikal, F. Meier, F. F. Gellrich, A. Hauschild, L. E. French, M. Schlaak, K. Ghoreschi, H. Kutzner, M. V. Heppt, S. Haferkamp, W. Sondermann, D. Schadendorf, B. Schilling, A. Hekler, E. Krieghoff-Henning, J. N. Kather, S. Fröhling, D. B. Lipka, and T. J. Brinker, "Robustness of convolutional neural networks in recognition of pigmented skin lesions," *European Journal of Cancer*, vol. 145, p. 81–91, Mar. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.ejca.2020.11.020

[10] C. Angelina and R. U. Ulfitria, "Classification of skin cancer using resnet and vgg deep learning network," in *Proceedings of the 11th International Applied Business and Engineering Conference, ABEC 2023, September 21st, 2023, Bengkalis, Riau, Indonesia*, ser. ABEC. EAI, 2024. [Online]. Available: http://dx.doi.org/10.4108/eai.21-9-2023.2342881

[11] L. Gamage, U. Isuranga, D. Meedeniya, S. De Silva, and P. Yogarajah, "Melanoma skin cancer identification with explainability utilizing mask guided technique," *Electronics*, vol. 13, no. 4, p. 680, Feb. 2024. [Online]. Available: http://dx.doi.org/10.3390/electronics13040680

[12] N. R. Kurtansky, B. M. D'Alessandro, M. C. Gillis, B. Betz-Stablein, S. E. Cerminara, R. Garcia, M. A. Girundi, E. V. Goessinger, P. Gottfrois, P. Guitera, A. C. Halpern, V. Jakrot, H. Kittler, K. Kose, K. Liopyris, J. Malvehy, V. J. Mar, L. K. Martin, T. Mathew, L. V. Maul, A. Mothershaw, A. M. Mueller, C. Mueller, A. A. Navarini, T. Rajeswaran, V. Rajeswaran, A. Saha,

M. Sashindranath, L. Serra-García, H. P. Soyer, G. Theocharis, A. Vos, J. Weber, and V. Rotemberg, "The slice-3d dataset: 400, 000 skin lesion image crops extracted from 3d tbp for skin cancer detection," *Scientific Data*, vol. 11, no. 1, Aug. 2024. [Online]. Available: http://dx.doi.org/10.1038/s41597-024-03743-w

[13] X. Tang and F. Rashid Sheykhahmad, "Boosted dipper throated optimization algorithm-based xception neural network for skin cancer diagnosis: An optimal approach," *Heliyon*, vol. 10, no. 5, p. e26415, Mar. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.heliyon.2024.e26415

[14] M. A. Albahar, "Skin lesion classification using convolutional neural network with novel regularizer," *IEEE Access*, vol. 7, p. 38306–38313, 2019. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2019.2906241

[15] A. Al Mahmud, S. Azam, I. U. Khan, S. Montaha, A. Karim, A. Haque, M. Zahid Hasan, M. Brady, R. Biswas, and M. Jonkman, "Skinnet-14: a deep learning framework for accurate skin cancer classification using low-resolution dermoscopy images with optimized training time," *Neural Computing and Applications*, vol. 36, no. 30, p. 18935–18959, Aug. 2024. [Online]. Available: http://dx.doi.org/10.1007/s00521-024-10225-y

[16] R. Fernandez, A. Guzmán-Ponce, R. Fernandez-Beltran, and G. García-Mateos, "Symmetry in explainable ai: A morphometric deep learning analysis for skin lesion classification," *Symmetry*, vol. 17, no. 8, p. 1264, Aug. 2025. [Online]. Available: http://dx.doi.org/10.3390/sym17081264

[17] A. Alesmaeil and E. Şehirli, "Real-time nail-biting detection on a smartwatch using three ¡scp¿cnn¡/scp¿ models pipeline," *Computational Intelligence*, vol. 41, no. 1, Jan. 2025. [Online]. Available: http://dx.doi.org/10.1111/coin.70020

[18] S. A. Hamim, M. U. I. Tamim, M. F. Mridha, M. Safran, and D. Che, "Smartskin-xai: An interpretable deep learning approach for enhanced skin cancer diagnosis in smart healthcare," *Diagnostics*, vol. 15, no. 1, p. 64, Dec. 2024. [Online]. Available: http://dx.doi.org/10.3390/diagnostics15010064

[19] M. Waqar, Z. A. Khan, S. T. Khawaja, N. I. Chaudhary, S. Khan, K. M. Cheema, M. F. Khan, S. S. Ahmed, and M. A. Z. Raja, "Explainable clinical diagnosis through unexploited yet optimized fine-tuned convnext models for accurate monkeypox disease classification," *SLAS Technology*, vol. 33, p. 100336, Aug. 2025. [Online]. Available: http://dx.doi.org/10.1016/j.slast.2025.100336

[20] T. Khater, S. Ansari, S. Mahmoud, A. Hussain, and H. Tawfik, "Skin cancer classification using explainable artificial intelligence on pre-extracted image features," *Intelligent Systems with Applications*, vol. 20, p. 200275, Nov. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.iswa.2023.200275

[21] F. Grignaffini, E. De Santis, F. Frezza, and A. Rizzi, "An xai approach to melanoma diagnosis: Explaining the output of convolutional neural networks with feature injection," *Information*, vol. 15, no. 12, p. 783, Dec. 2024. [Online]. Available: http://dx.doi.org/10.3390/info15120783

[22] Z. Li, T. Gu, B. Li, W. Xu, X. He, and X. Hui, "Convnext-based fine-grained image classification and bilinear attention mechanism model," *Applied Sciences*, vol. 12, no. 18, p. 9016, Sep. 2022. [Online]. Available: http://dx.doi.org/10.3390/app12189016

[23] M. B. Hossain, S. H. S. Iqbal, M. M. Islam, M. N. Akhtar, and I. H. Sarker, "Transfer learning with fine-tuned deep cnn resnet50 model for classifying covid-19 from chest x-ray images," *Informatics in Medicine Unlocked*, vol. 30, p. 100916, 2022. [Online]. Available: http://dx.doi.org/10.1016/j.imu.2022.100916

[24] R. A. Perdana, A. M. Arimurthy, and Risnandar, "Remote sensing scene classification using convnext-tiny model with attention mechanism and label smoothing," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 3, p. 389–400, Jun. 2024. [Online]. Available: http://dx.doi.org/10.29207/resti.v8i3.5731

[25] Y. Zhang, A. Xu, D. Lan, X. Zhang, J. Yin, and H. H. Goh, "Convnext-based anchor-free object detection model for infrared image of power equipment," *Energy Reports*, vol. 9, p. 1121–1132, Sep. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.egyr.2023.04.145

[26] International Skin Imaging Collaboration, "Isic 2019 challenge dataset," https://challenge.isic-archive.com/data/, 2019, accessed: 2026-05-18.

[27] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, S. Langer, K. Lioprys, J. Malvehy, S. Musthaq, J. Nanda, O. Reiter, G. Shih, A. Stratigos, P. Tschandl, J. Weber, and H. P. Soyer, "A patient-centric dataset of images and

metadata for identifying melanomas using clinical context," *Sci. Data*, vol. 8, no. 1, p. 34, Jan. 2021. [Online]. Available: https://doi.org/10.34970/2020-ds01

[28] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, Aug. 2018. [Online]. Available: http://dx.doi.org/10.1038/sdata.2018.161

[29] Memorial Sloan Kettering Cancer Center, "Prove-ai," 2022. [Online]. Available: https://api.isic-archive.com/doi/prove-ai/

[30] Q. Pan, K. Liu, S. Zheng, and G. Wang, "A fine-grained image classification method based on convnext heatmap localization and contrastive learning," *IEEE Access*, vol. 13, p. 80123–80132, 2025. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2025.3567488

[31] F. Liu, C. Zang, J. Shi, W. He, Y. Liang, and L. Li, "An improved covid-19 lung x-ray image classification algorithm based on convnext network," *International Journal of Image and Graphics*, vol. 24, no. 03, May 2023. [Online]. Available: http://dx.doi.org/10.1142/S0219467824500360

[32] N. H. Shabrina, D. Gunawan, M. F. Ham, and A. S. Harahap, "Papillary thyroid cancer histopathological image classification using pretrained convnext tiny and grad-cam interpretation," in *2023 IEEE 11th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*. IEEE, Dec. 2023, p. 1836–1842. [Online]. Available: 10.1109/itaic58329.2023.10409019

[33] A. Kohan, A. Zahedi, R. Alizadehsani, R.-S. Tan, and U. R. Acharya, "Application of explainable artificial intelligence (xai) techniques in patients with intracranial hemorrhage: A systematic review," *WIREs Data Mining and Knowledge Discovery*, vol. 15, no. 3, pp. 1–26, 2025, e70031 DMKD-00920.R1. [Online]. Available: https://doi.org/10.1002/widm.70031

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778