Ingénierie des Systèmes d'Information
Vol. 31, No. 3, March, 2026, pp. 791-803
Journal homepage: http://iieta.org/journals/isi

Critical Assessment of DCGAN-Based Data Balancing in Deep Learning for Keratitis Identification: Trade-Off Between Fidelity, Fairness, and Performance

Moch. Syahrir[1*], Rifqi Hammad[1], Bahtiar Imran[2], Muhammad Syairozi Hidayat[3]

[1] Faculty of Engineering, Bumigora University, Mataram 83127, Indonesia
[2] Faculty of Engineering, Mataram University of Technology, Mataram 83115, Indonesia
[3] North Lombok Regency General Hospital, North Lombok Regency 83352, Indonesia

Corresponding Author Email: moch.syahrir@universitasbumigora.ac.id

https://doi.org/10.18280/isi.310313

Received: 1 October 2025
Revised: 16 November 2025
Accepted: 21 November 2025
Available online: 31 March 2026

Keywords:
keratitis identification, deep learning, convolutional neural network, generative adversarial networks, data balancing

ABSTRACT

Keratitis, an inflammation of the cornea, presents significant diagnostic challenges due to overlapping clinical symptoms and the limited availability of expert ophthalmological evaluation, especially in early-stage presentations. Timely diagnosis is crucial to prevent permanent corneal damage and vision loss. Clinical datasets used for automated keratitis detection are often highly imbalanced, which can introduce classification bias and impair the generalization capability of deep learning–based models. To address this issue, synthetic data generation using generative adversarial networks (GANs) has been widely adopted to balance minority classes. However, the effectiveness of such data augmentation in improving both model performance and fairness remains underexplored. A comparative evaluation was performed using convolutional neural networks (CNNs) under two scenarios: (i) imbalanced training with real data and (ii) balanced training incorporating GAN-generated images. MobileNet, trained on the imbalanced dataset, achieved the highest accuracy of 94.53%, demonstrating high sensitivity for the dominant classes. In contrast, Xception trained on the GAN-balanced dataset exhibited a lower accuracy of 80.76%, accompanied by inconsistent recall across the target classes, despite the improved class distribution. Structural similarity index measure (SSIM) analysis indicated suboptimal visual fidelity of the synthetic images, particularly for underrepresented categories. These findings reveal that class balancing via GANs does not consistently enhance classification performance and may introduce representational noise, leading to a trade-off between class balance and overall accuracy. By emphasizing the importance of jointly evaluating model fairness and the fidelity of GAN-synthesized data, the results highlight the need for more adaptive and quality-aware augmentation strategies in medical image classification tasks.

1. INTRODUCTION

Keratitis is the fifth leading cause of corneal blindness globally [1-4]. One of the most severe forms of this condition is fungal keratitis (FK), which accounts for nearly half of microbial keratitis cases in developing countries, particularly in tropical regions [5-8]. FK frequently leads to severe visual impairment or permanent blindness, and approximately 25% of affected patients require invasive and costly surgical intervention [9]. Therefore, early diagnosis is essential to prevent serious complications. In vivo confocal microscopy (IVCM) is recognized as an effective imaging modality for keratitis diagnosis because it provides real-time, non-invasive visualization of corneal microstructures [7, 10]. However, the clinical adoption of IVCM remains limited due to high cost, restricted equipment availability, and reliance on operator expertise and patient cooperation [4, 11, 12]. With recent advances in artificial intelligence, deep learning–based methods have emerged as promising solutions for automated diagnosis, particularly for the analysis of complex medical images.

However, deep learning–based keratitis classification faces major challenges, particularly class imbalance in clinical datasets. This imbalance biases models toward majority classes and often overlooks minority categories, which is critical for accurate diagnosis, especially under conditions of limited access to large medical datasets [13]. To mitigate dataset imbalance, oversampling strategies are commonly employed. These approaches are considered more representative than undersampling because they preserve information from majority classes while enhancing minority-class representation [14]. Unlike random oversampling or the Synthetic Minority Oversampling Technique (SMOTE), generative adversarial networks (GANs) synthesize new minority-class samples rather than duplicating existing data. As a result, GANs provide a more advanced and flexible approach to handling data imbalance. They have been widely applied for synthetic data generation and for addressing image-level imbalance in classification tasks [15-18].

Research on keratitis classification using deep learning has progressed rapidly in recent years, with a primary focus on improving the accuracy and efficiency of image-based

diagnosis. Several studies have explored various convolutional neural network (CNN) architectures to differentiate keratitis types, particularly FK and bacterial keratitis (BK). For example, Kuo et al. [19] employed the DenseNet architecture and reported an average accuracy of 70%. Ghosh et al. [20] compared multiple CNN models and found that VGG19 achieved the highest F1 score of 78% in detecting FK and BK. Hung et al. [21] further demonstrated that DenseNet161 performs effectively in analyzing slit-lamp images, achieving an accuracy of 78.6%. Notably, Redd et al. [5] showed that MobileNet can outperform corneal specialists in keratitis classification, highlighting the clinical potential of automated diagnostic systems. Another study proposed a multiscale CNN, in which Mayya et al. [22] achieved an accuracy of 88.96% in distinguishing FK from non-FK cases. The integration of slit-lamp and smartphone images has also been shown to improve classification performance, reaching accuracies of up to 90% [23]. Keratitis detection based on anterior segment images using ResNet-50 achieved an accuracy of 81% [24]. Furthermore, an ensemble approach combining DenseNet121, MobileNetV2, and SqueezeNet1_0 improved accuracy to 84.52% [25].

Furthermore, IVCM image–based approaches are increasingly adopted because they can capture detailed morphological features of fungal hyphae and corneal microstructures with high accuracy. For example, Inception-ResNet V2 successfully detected Fusarium and Aspergillus fungi, achieving accuracies of 81.7% and 75.7%, respectively. Lv et al. [26] reported an accuracy of 93.64% for FK detection using fungal hyphae as the primary diagnostic indicator. Essalat et al. [10] demonstrated that DenseNet161 applied to IVCM images achieved superior diagnostic performance, with an accuracy of 93.55%, precision of 92.52%, recall of 94.77%, and an F1-score of 96.93%. These results confirm the higher diagnostic potential of IVCM images compared to slit-lamp imaging. The IVCM dataset used in their study, which also serves as the baseline dataset in this research, exhibits a highly imbalanced class distribution. It consists of 1,391 acanthamoeba keratitis (AC), 863 FK, 536 normal (N), and only 231 non-specific keratitis (NSK) images. To mitigate this imbalance, a Weighted Random Sampler was applied. However, this approach only replicates minority samples without introducing new variations, which may increase the risk of overfitting. As an alternative, data synthesis–based methods such as GANs have gained increasing attention. In this study, a deep convolutional GAN (DCGAN) is employed to generate realistic synthetic images through adversarial learning between a generator and a discriminator. DCGANs have demonstrated strong performance in various medical imaging applications [27-29], including image augmentation tasks [2, 30]. The key advantage of GAN-based augmentation lies in its ability to increase the diversity of minority-class samples without altering their labels, thereby potentially improving model fairness.

Although DCGANs have been widely applied to address data imbalance in medical imaging and other domains, their effectiveness in keratitis image classification has not yet been systematically investigated. This study conducts a comparative evaluation of DCGAN-based data balancing in CNN models, involving architectures such as MobileNet, XceptionNet, DenseNet169, ResNet50, VGG19, and EfficientNet-B0. All models are evaluated on both the original imbalanced dataset and the DCGAN-balanced dataset. The evaluation considers classification performance metrics, including accuracy, precision, recall, and F1-score, as well as image quality assessed using the structural similarity index measure (SSIM). The investigation focuses on three key aspects: fairness, fidelity, and performance trade-off. Fairness refers to the model's ability to maintain balanced sensitivity across all classes, including minority categories. Sensitivity is defined as the ability to correctly identify positive samples within each class. Fidelity evaluates the visual and semantic similarity between synthetic and real images. The performance trade-off captures potential accuracy degradation caused by distributional noise introduced by synthetic data. This study aims to provide deeper insights into the effectiveness and limitations of DCGAN-balanced data for developing fair and representative deep learning–based keratitis classification systems.

2. RESEARCH METHODS

This methodology is systematically designed to achieve the primary objective of developing and evaluating deep learning–based image classification models. The research process is divided into several structured stages, including data collection and preprocessing, model architecture selection, training under different data conditions (imbalanced and balanced), and performance evaluation using accuracy-related metrics. The complete workflow is illustrated in Figure 1, which presents the main steps of the proposed experimental framework. Each stage plays a crucial role in ensuring the reliability and validity of the classification results, particularly in addressing data imbalance, a major challenge in this domain.

Based on the workflow, two experimental schemes are defined. In the first scheme, CNN models are trained and evaluated using the original imbalanced dataset. In the second scheme, CNN models are trained using a balanced dataset, following Stage 1 and Stage 2. The research begins with the collection of IVCM keratitis image datasets. Subsequently, experiments are conducted according to Scheme 1 and Scheme 2. In Scheme 1, the dataset, which remains imbalanced, proceeds directly to Stage 2, namely data preprocessing, before being used for CNN training. The model achieving the highest classification accuracy on the test dataset is selected as the final model for keratitis diagnosis. In Scheme 2, the dataset undergoes Stage 1 and Stage 2. Stage 1 focuses on balancing the dataset using DCGAN. Prior to DCGAN training, preprocessing is applied by converting images to grayscale and normalizing pixel values. The best DCGAN model is then selected to generate synthetic images using the generator. Model selection is based on the training loss value, where Binary Cross-Entropy (BCE) is employed as the loss function. The BCE formulation is defined in Eq. (1) [31].

$$LD = E_x \sim P_r[-\log(D(x))] + E_{x\sim g}[-\log(1-D(x))],$$
$$LG = E_{x\sim P_g + E_{x\sim g}}[\log(1-D(x))] \quad (1)$$

A generator with a lower BCE loss is considered more effective, as it indicates the ability to generate synthetic images that closely resemble real data. Conversely, when the discriminator struggles to distinguish between real and synthetic images, its loss value increases, reflecting improved quality and realism of the generated data. The synthetic images are then combined with real images to form a balanced dataset,

which is subsequently used for CNN training in Stage 2.

Figure 1 further illustrates each sub-process of the proposed research workflow.

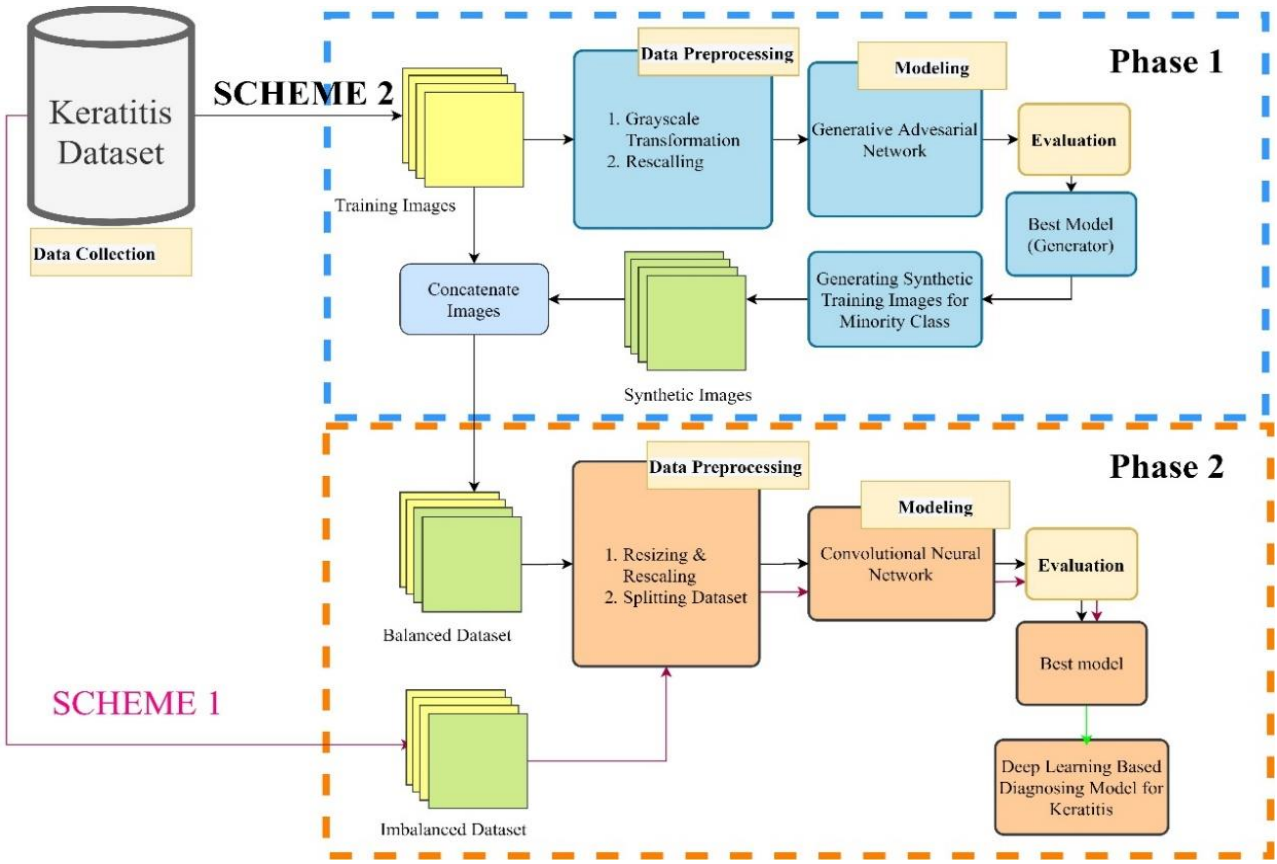

Figure 1. Research flow diagram

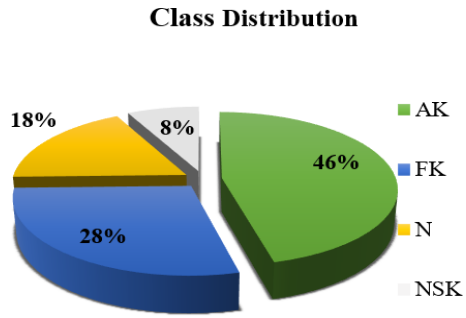

Figure 2. Distribution of keratitis classes

2.1 Dataset collection

In this stage, IVCM examination images are collected for keratitis detection. The dataset consists of four classes, defined according to the observed keratitis types. However, the dataset is highly imbalanced, with an uneven number of images across classes. The class distribution is presented in Figure 2. This imbalance can introduce bias during model training, causing the model to favor majority classes while exhibiting reduced sensitivity toward minority classes. Therefore, in the subsequent stage, data imbalance is explicitly addressed to ensure fairer and more representative classification performance across all classes.

2.2 Data pre-processing

Poor-quality data can significantly reduce model accuracy and lead to incorrect predictions, making it crucial to enhance dataset quality through preprocessing. Model performance can also be improved by introducing variations to image data. Data augmentation encompasses a set of strategies designed to expand and refine the size and characteristics of images [13]. For images, augmentation techniques may include scaling, rotation, horizontal or vertical flipping, shifting, and more [32]. In this study, only scaling was applied, and all images in the dataset were standardized to the same size. Scaling, or normalization, of image dimensions and color channels was performed to meet the requirements of the CNN model architecture. This approach aims to enable the model to recognize visual patterns more accurately while reducing sensitivity to variations in data distribution.

2.3 Model implementation

2.3.1 Generative adversarial networks

GAN is a type of artificial neural network architecture comprising two main models: a generator and a discriminator, which compete with each other during the training process. GANs have proven highly effective in generating realistic synthetic data across a variety of domains, including images, audio, text, and more. The GAN architecture is illustrated in Figure 3. In Figure 3, the generator produces synthetic data from noise, while the discriminator distinguishes between real and synthetic data. Both are trained alternately until the generator can produce data that is challenging for the discriminator to differentiate. This study employs DCGAN, an enhanced GAN variant with convolutional layers as its

backbone [3, 33].

The generator architecture in this study is designed to generate synthetic images resembling the original data from a 100-dimensional random vector input. The architecture begins with a Dense layer that expands the input dimensionality to 65,536, followed by normalization and LeakyReLU activation. The output is then reshaped into a spatial form (16 × 16 × 256) and passed through three consecutive Conv2DTranspose layers to gradually enlarge the spatial dimensionality to (128 × 128 × 1), corresponding to the target image dimensions, interspersed with BatchNormalization and LeakyReLU activation to maintain stability and non-linearity during training. This architecture is shown in Table 1. Meanwhile, Table 2 presents the discriminator architecture.

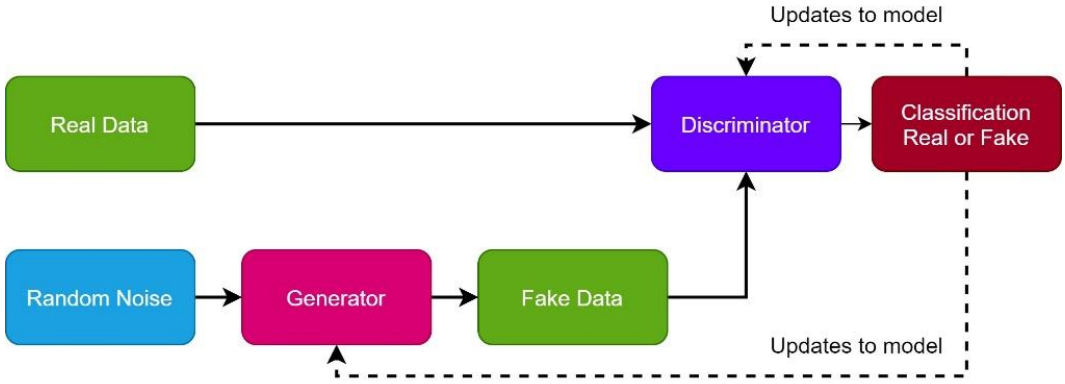

Figure 3. Generative adversarial networks (GAN) architecture [18]

Table 1. Generator architecture

Layer	Input	Output
Dense	(None,100)	(None,65536)
BatchNormalization	(None,65536)	(None,65536)
LeakyRelu	(None,65536)	(None,65536)
Reshape	(None,65536)	(None,16,16,256)
Conv2DTranspose	(None,16,16,256)	(None,32,32,128)
BatchNormalization	(None,32,32,128)	(None,32,32,128)
LeakyRelu	(None,32,32,128)	(None,32,32,128)
Conv2DTranspose	(None,32,32,128)	(None,64,64,64)
BatchNormalization	(None,64,64,64)	(None,64,64,64)
LeakyRelu	(None,64,64,64)	(None,64,64,64)
Conv2DTranspose	(None,128,128,1)	(None,128,128,1)

Table 2. Generator architecture

Layer	Input	Output
Conv2D	(None,128,128,1)	(None,64,64,64)
LeakyRelu	(None,64,64,64)	(None,64,64,64)
Dropout	(None,64,64,64)	(None,64,64,64)
Conv2D	(None,64,64,64)	(None,32,32,128)
LeakyRelu	(None,32,32,128)	(None,32,32,128)
Dropout	(None,32,32,128)	(None,32,32,128)
Flatten	(None,32,32,128)	(None, 131072)
Dense	(None,131072)	(None,1)

The discriminator is responsible for distinguishing between the original images and the fake images generated by the generator. This network receives input images of size 128 × 128 × 1 and processes them through two consecutive Conv2D layers that reduce resolution while extracting features, followed by LeakyReLU activation and dropout to prevent overfitting [34]. The resulting feature maps are then flattened into a one-dimensional vector and classified by a dense layer into a single scalar output, representing the probability that the image is real or synthetic. These two networks are trained adversarially to achieve a balance between image quality and detection accuracy. The GAN was trained for 3,000 epochs with a batch size of 32, using binary cross-entropy loss and the Adam optimizer with a learning rate of $1e^{-4}$. The choice of 3,000 epochs was empirically determined to allow sufficient adversarial convergence between the generator and discriminator, ensuring that the model achieves stable training without premature saturation [35, 36]. The learning rate of $1e^{-4}$ was selected to maintain gradient stability and prevent oscillations during adversarial updates.

2.3.2 CNN

CNNs are specifically designed for processing image data and consist of multiple layers, including convolutional, pooling, and activation layers. A key advantage of CNNs is their ability to automatically extract features from images without the need for manual feature engineering [37]. This capability makes CNNs highly effective for image classification. The CNN architecture used for classification is illustrated in Figure 4. The core component of a CNN is the convolutional layer, which employs kernels or filters to generate relevant features (feature maps) from the input image.

The convolution operation can be interpreted as the repeated application of a function to the output of another function. In the context of CNNs, it facilitates the extraction of important patterns from images. After passing through multiple convolutional and pooling layers, the CNN output is converted into a one-dimensional vector using a flattening layer. This step is essential for preparing the data before it is fed into the fully connected layers for final classification. The CNN models employed in this study include DenseNet169, MobileNet, ResNet50, VGG19, XceptionNet, and EfficientNetB0. These architectures were selected to represent a range of model complexities and design principles. MobileNet and EfficientNetB0 serve as lightweight and efficient models, whereas DenseNet169, ResNet50, VGG19, and XceptionNet represent deeper, high-capacity networks. This selection allows for a comprehensive evaluation of GAN-based balancing performance across different CNN characteristics. These models serve as baseline architectures for feature extraction from 128 × 128 × 3 input images, as depicted in Figure 5. Each backbone model extracts spatial representations through convolutional layers, which are then passed to the head layer via global average pooling and multiple fully connected layers for final classification into four classes using the Softmax activation function.

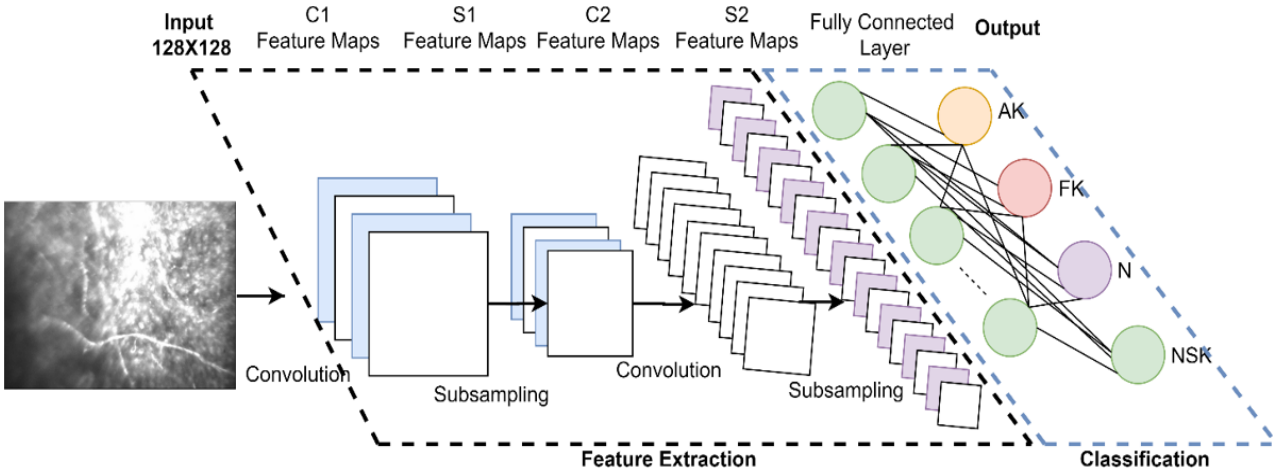

Figure 4. Convolutional neural network (CNN) architecture for image classification [38]

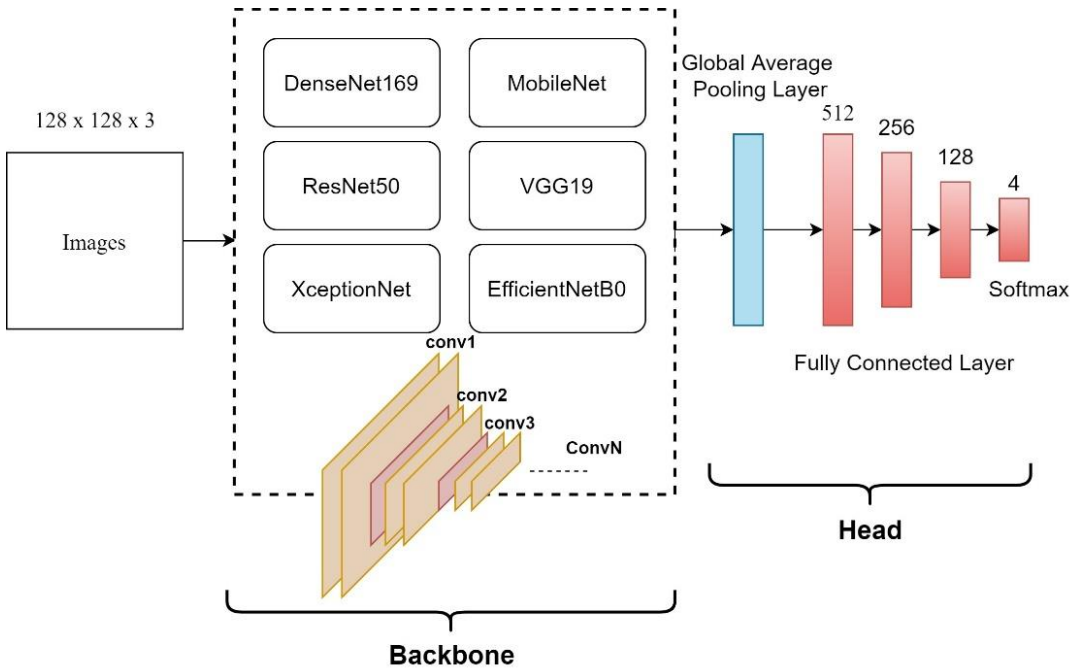

Figure 5. Proposed convolutional neural network (CNN) architecture with backbone as feature extractor and head for four-class classification

2.3.3 Performance evaluation

Model performance evaluation is conducted using the Precision, Recall, F1-Score, and Accuracy metrics. Calculating these metrics requires an understanding of several key terms. True positives (TP) refer to instances of the positive class that are correctly classified as positive. True negatives (TN) are instances of the negative class correctly classified as negative. False positives (FP) describe instances of negative data incorrectly classified as positive, while false negatives (FN) are instances of positive data incorrectly classified as negative. Each evaluation metric is calculated using Eqs. (2)-(5) [10, 32, 39-42]:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

The SSIM is used in this study to evaluate the quality of synthetic images generated by a GAN-based generative model. SSIM assesses the similarity between the original and synthetic images based on luminance, contrast, and structure. SSIM values range from -1 to 1, with 1 indicating perfect similarity. The calculation is performed by dividing the image into small blocks, calculating local statistics, and aggregating them into an overall score. This approach is considered more perceptually representative than conventional metrics such as mean square error (MSE) or peak signal-to-noise ratio (PSNR), making it more suitable for assessing the visual quality of synthetic data [43]. The SSIM calculation follows Eq. (6). x and y are the two images being compared, μ is the

average pixel intensity of the image, σ^2 is the variance of the image, σ_{xy} is the covariance between the two images, and c is a small constant to avoid division by 0 and stabilize the calculation.

$$\mathrm{SSIM}(x,y)=\frac{\left(2\mu_x\mu_y+c_1\right)\left(2\sigma_{\{xy\}}+c_2\right)}{\left(\mu_x^2+\mu_y^2+c_1\right)\left(\sigma_x^2+\sigma_y^2+c_2\right)} \quad (6)$$

3. RESULTS AND DISCUSSION

The performance evaluation of CNN models in this study was conducted using four main metrics: Precision, Recall, F1-Score, and Accuracy. This evaluation was applied to two scenarios: (1) original imbalanced dataset, and (2) balanced dataset based on GAN. In this study, CNN was trained using 128 × 128 pixel images for 50 epochs with a batch size of 32. The model was fitted with a dropout of 0.3 and optimized using Adam (learning rate of $1e^{-4}$) with a categorical cross entropy loss function. The dataset was split with an 80:20 ratio for training and testing. Table 3 summarizes the performance of six CNN architectures in imbalanced and balanced dataset scenarios. In imbalanced data, almost all models showed high performance. MobileNet and XceptionNet were the top two models with accuracies of 94.53% and 94.20%, respectively, and F1-Scores above 93%, indicating effectiveness in recognizing patterns from the original data despite the imbalanced class distribution.

Table 3. Model testing results under both scenarios

No.	Model	Evaluation Metrices Imbalanced Scenario				Evaluation Metrices Balanced Scenario			
		Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy
1	DenseNet169	90.56%	91.38%	90.64%	91.38%	81.31%	78.06%	73.72%	78.06%
2	MobileNet	94.05%	94.53%	93.97%	94.53%	84.38%	78.24%	72.94%	78.24%
3	ResNet50	91.43%	92.21%	91.67%	92.21%	79.56%	79.14%	76.23%	79.14%
4	VGG19	90.79%	91.38%	89.33%	91.38%	83.81%	78.06%	72.43%	78.06%
5	XceptionNet	94.21%	94.20%	93.19%	94.20%	82.40%	80.76%	78.60%	80.76%
6	EfficientNetB0	51.09%	57.38%	50.04%	57.38%	22.29%	44.33%	29.65%	44.33%

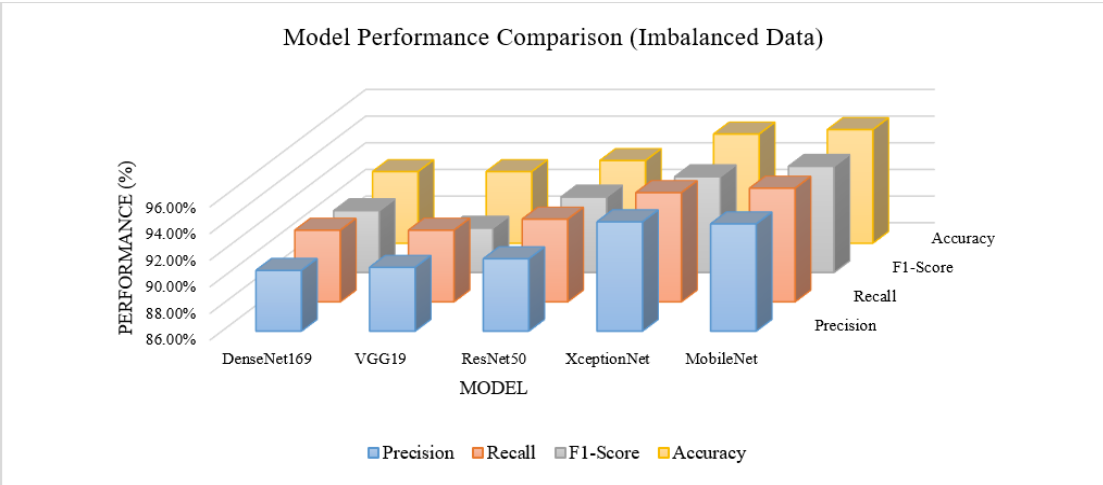

Figure 6. Comparison of model performance on imbalanced datasets

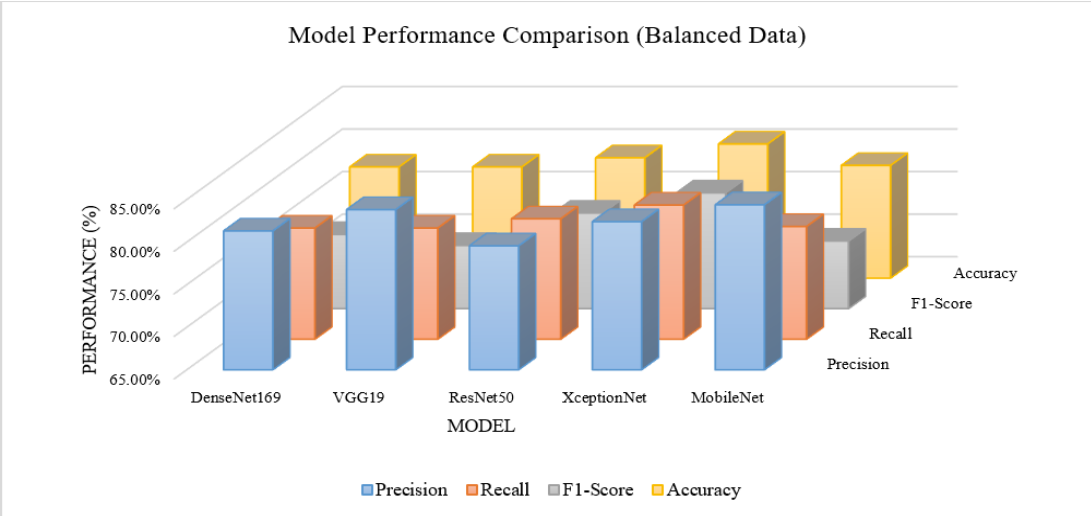

Figure 7. Comparison of model performance on balanced datasets

Table 4. Performance degradation after balancing

Model	Accuracy Imbalanced Data (%)	Accuracy Balanced Data (%)	Decrease in Accuracy (%)
EfficientNetB0	57.38	44.33	22.74
DenseNet169	91.38	78.06	14.58
VGG19	91.38	78.06	14.58
ResNet50	92.21	79.14	14.17
XceptionNet	94.2	80.76	14.27
MobileNet	94.53	78.24	17.23
	Mean		16.26

DenseNet169, ResNet50, and VGG19 also demonstrated strong performance with accuracy above 91%. However, when the data was balanced using synthetic data generated by GANs, performance decreased across all models. DenseNet169's F1 score dropped from 90.64% to 73.72%. Similarly, MobileNet and XceptionNet experienced a decrease in F1 scores. The most significant decrease occurred in EfficientNetB0, where its accuracy dropped drastically from 57.38% to just 44.33%, and its F1 score dropped from 50.04% to 29.65%. Statistical analysis using a paired t-test ($T = 26.843$, $p < 0.001$) confirmed that the difference in model performance between the imbalanced and DCGAN-balanced datasets was statistically significant. This finding suggests that the accuracy drop observed after balancing is not random but rather a systematic effect likely caused by the generative data characteristics. The synthetic images produced by DCGAN may not fully capture the intrinsic variability of real samples, leading to feature mismatches and lower discriminative power across CNN architectures.

Figure 6 presents a visualization comparing the performance of the six models on imbalanced data. It is clear that the majority of models fall within the 90–94% accuracy range, demonstrating consistent and stable performance in the face of imbalanced class distributions. MobileNet, XceptionNet, and ResNet50 models are among the best performers, confirming their ability to extract important features from IVCM images even without any balancing process. This visualization also shows that despite the imbalanced classes, the models are still able to provide accurate predictions by exploiting consistent patterns in the data. This indicates that the training and preprocessing techniques used are quite effective in handling class imbalance without requiring additional intervention.

Figure 7 shows a significant performance drop after balancing by adding synthetic data. All models experienced a decrease in accuracy and F1 score, indicating that balancing did not have a significant positive impact in the context of this dataset. For example, MobileNet and XceptionNet, which previously recorded high accuracy above 94%, experienced a significant decline after balancing. This pattern of decline was consistent across almost all models. This suggests that the GAN-balanced data may not be realistic enough or may even introduce noise into the training process. This finding is supported by studies showing that excessive synthetic data can lead to performance degradation because small modes in the original data appear as noise [44]. The performance degradation reflected in Figure 7 underscores that the generative model-based balancing process requires special attention in validating the quality of synthetic data. Models that initially performed well on the original data actually experienced performance degradation when confronted with the combined data. This visualization supports the argument that balancing does not always guarantee improved accuracy, especially if the distribution and quality of the additional data are suboptimal.

This performance decline is further clarified in Table 4, which quantitatively shows the magnitude of the accuracy drop. EfficientNetB0 experienced the highest decline at 22.74%, while ResNet50 recorded the lowest decline at 14.17%. Overall, the average accuracy decline among the six models was 16.26%. This indicates that balancing using synthetic data has not fully improved the model's generalization ability. This finding suggests that the quality of synthetic data needs to be improved to more closely reflect the characteristics and distribution of the original data, both in terms of texture, morphology, and clinical variations in IVCM keratitis images.

Table 5 displays the average SSIM score of synthetic images generated by the GAN model for each class. SSIM is used to assess how structurally similar synthetic images are to the original images. The results show that the N class has the highest score, at 73.76%, indicating that the quality of synthetic images in this class is closer to the original data than other classes. Meanwhile, the NSK and FK classes obtained relatively low scores, at 57.35% and 58.54%, respectively, indicating a structural mismatch between synthetic and original images in these classes. When it comes to balancing requirements, it can be seen that the NSK class requires the most synthetic images, followed by the N class, and the FK class. The correlation between data requirements and SSIM scores indicates that the more synthetic data required, the more difficult it is for the model to generate images with structures that closely resemble the original data. This is evident in the NSK class, which, in addition to having the highest data requirements, also recorded the lowest SSIM score. In general, low SSIM scores in minority classes may indicate that GANs are not yet able to replicate the complexity and texture variation of these classes with a limited number of training samples. Therefore, these results emphasize the need to improve GAN training capacity and stability, particularly for generating representative synthetic data for classes with highly skewed distributions.

Table 5. SSIM evaluation of synthetic images

No.	Class	GAN Average SSIM Score
1	N	73.76%
2	NSK	57.35%
3	FK	58.54%

Figure 8 depicts the relationship between model complexity (in terms of the number of parameters) and model accuracy on imbalanced and balanced data. Figure 8 illustrates that the relationship between model complexity (number of parameters) and accuracy is non-linear, especially when the data is balanced using synthetic images. MobileNet, with a total of 3,917,360 parameters, shows high accuracy on imbalanced data due to the efficiency of its lightweight

architecture, but its performance drops significantly after balancing, indicating its limitations in handling variations in synthetic data. EfficientNetB0, despite having only 4,869,139 parameters, produces the lowest accuracy in both scenarios, indicating that parameter efficiency alone is not sufficient to guarantee robustness to new distributions. In contrast, DenseNet169, with 13,659,056 parameters, shows high accuracy and relative stability, reflecting the advantage of dense connectivity in maintaining robust feature representations amid changing data distributions.

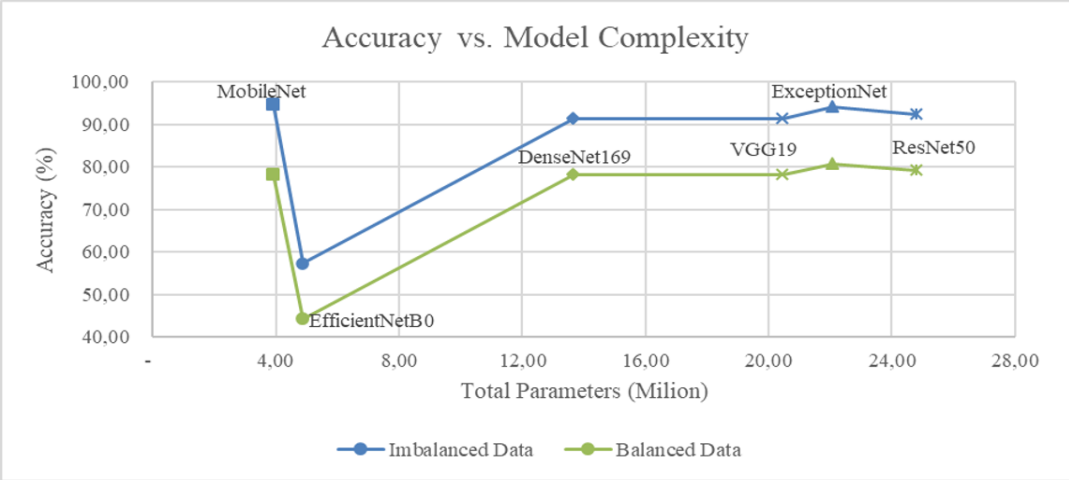

Figure 8. Comparison of model accuracy versus model complexity

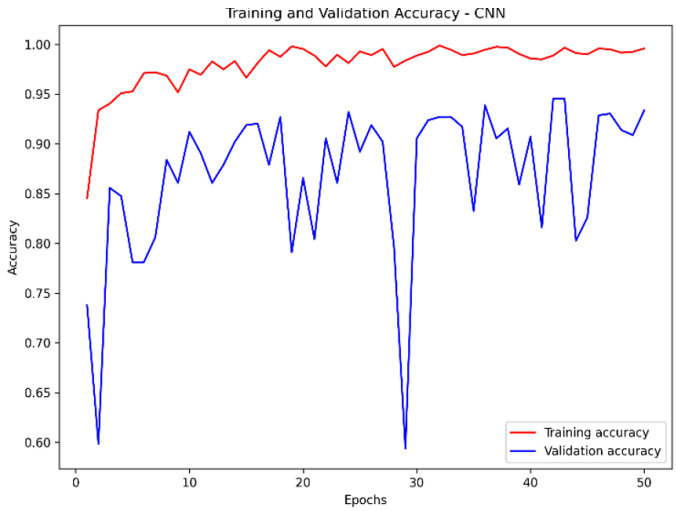

Figure 9. Training and validation accuracy curves of MobileNet on imbalanced data

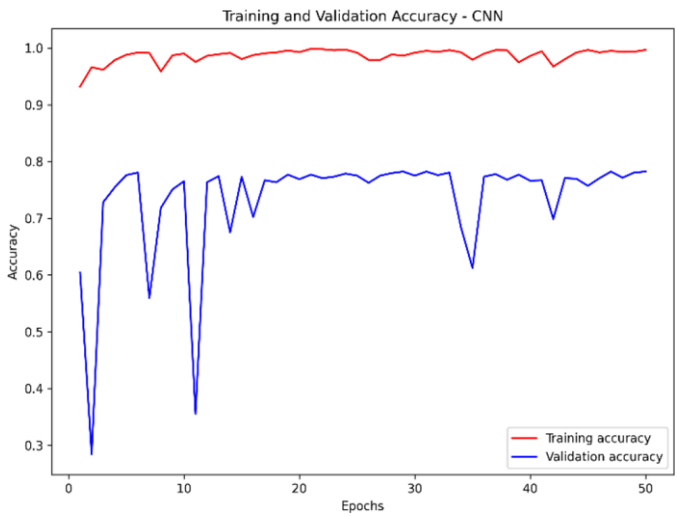

Figure 10. Training and validation accuracy curves of MobileNet on balanced data

High-complexity models like ResNet50 (24,800,496 parameters), VGG19 (20,450,736), and XceptionNet (22,074,264) achieve high accuracy but show signs of performance saturation after the balancing process, which could be caused by noise or a mismatch in the characteristics of the synthetic data. XceptionNet remains consistent thanks to its efficient and depthwise separable convolutional architecture, while ResNet50 excels thanks to its deep shortcut connections. However, VGG19 tends to stagnate due to its conventional architecture and lack of adaptability to new distributions. This analysis emphasizes the importance of choosing an architecture with the right balance between capacity and generalization, especially when handling GAN-based synthetic data.

Based on the training and validation accuracy curves shown in Figure 9 and Figure 10, the models demonstrated stable convergence by the end of the 50 training epochs. In the imbalanced dataset condition (Figure 9), the training accuracy increased rapidly and approached a value close to 1, while the validation accuracy exhibited relatively high fluctuations up to around epoch 30. However, after that point, the oscillations gradually diminished and stabilized around an accuracy of 0.95, indicating that the learning process had reached a balance between fitting and generalization. In the balanced dataset condition with DCGAN (Figure 10), a similar trend was observed, where the validation accuracy showed more noticeable oscillations in the early epochs but became progressively stable after epoch 30, suggesting that the model had achieved adequate convergence. Moreover, there were no significant signs of overfitting, as the gap between training and validation accuracy remained relatively small toward the end of training.

Figure 11 shows the generator and discriminator loss values in the GAN models for each class. The results indicate that the normal class has the lowest generator loss (0.4357) and the highest discriminator loss (2.3918), suggesting that the generator successfully produces synthetic images that are realistic enough to be difficult for the discriminator to distinguish [36]. In contrast, for the NSK class, the generator

loss value was the highest (0.6856) with the lowest discriminator loss (1.4812). This suggests that the generator has difficulty representing the characteristics of this class, and the discriminator can easily distinguish synthetic images from real images. This reflects a possible failure of the generator in optimally capturing the distribution of the NSK class, which could be caused by feature complexity or limited original data for that class. For the FK class, the loss value is between the other two classes, with a generator loss of 0.5912 and a discriminator loss of 1.7276, indicating relatively stable training but still not achieving optimal synthesis quality. These findings suggest that although the chosen configuration enabled partial learning convergence, instability remained during adversarial training. This limitation likely contributed to the relatively poor classification performance observed after DCGAN-based balancing.

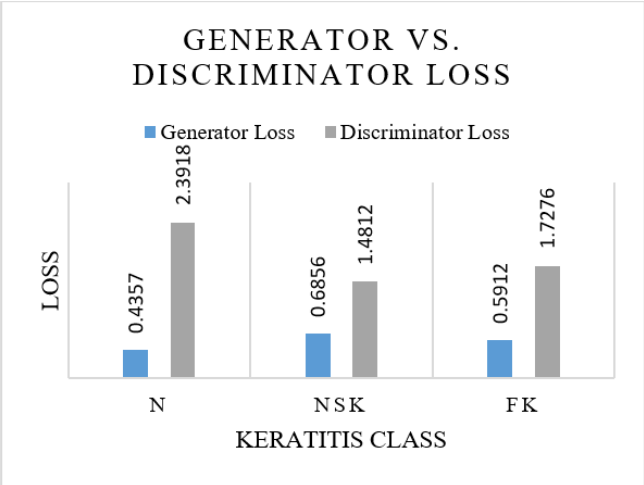

Figure 11. Loss values of the generator and discriminator in the GAN models for each class

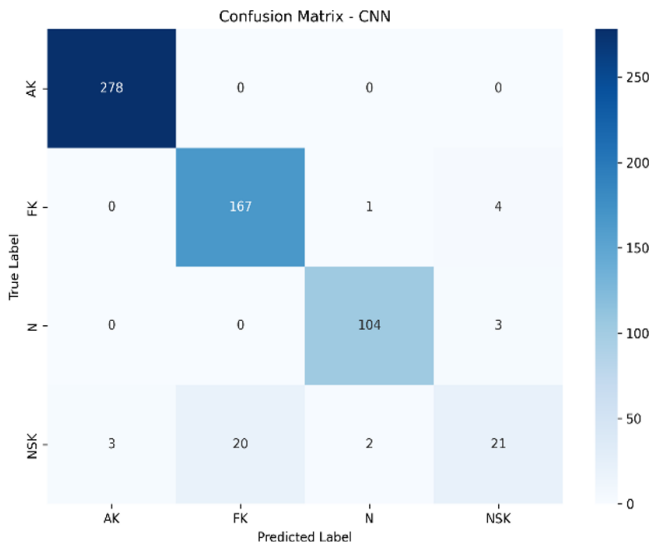

Figure 12. Confusion matrix of MobileNet trained on imbalanced data

Figure 12 shows the confusion matrix of a MobileNet model trained on the imbalanced dataset. Under these conditions, the model tends to be biased toward the majority classes, particularly AK and FK, which achieve nearly perfect predictions (278 and 167, respectively), while the NSK class exhibits a high number of misclassifications, with only 21 correctly classified samples. This indicates that minority samples, such as NSK, are not adequately represented in the learning process. After applying balancing (Figure 13), the model's performance against the majority class remains excellent, with AK, FK, and N correctly classifying 274, 273, and 274 samples, respectively. The number of misclassifications across these three classes is minimal, indicating that the model has high sensitivity and good fairness toward the AK, FK, and N classes. This shows that balancing does not impair the model's performance on these major classes and, in fact, strengthens its classification stability.

However, the NSK class is different. Despite the increase in the number of NSK samples, the classification results still show a high error rate, with only 49 samples correctly classified out of a total of 278. The majority of the NSK samples are misclassified into the AK (27), FK (161), and N (41) classes. This confirms that the NSK class, as a category of NSK, naturally has a broad and unstructured scope and tends to overlap with the characteristics of other classes. This phenomenon is also consistent with the low SSIM value in the augmentation results for the NSK class, which indicates that the quality of synthetic images or visual representations of this class is indeed less consistent or homogeneous. Overall, the balancing process successfully improved the sensitivity and fairness significantly for the AK, FK, and N classes, but still faces challenges in the NSK class due to limited visual representation and unclear category boundaries, which are indeed the nature of the "non-specific" class. A similar pattern occurs with XceptionNet in Figures 14 and 15. However, interestingly, XceptionNet is able to classify more NSK samples than the MobileNet model.

Figure 16 shows an example of the prediction results from the best model, MobileNet, trained on the imbalanced dataset. It can be observed that the model correctly classifies images from the AK, FK, and N classes, demonstrating high accuracy for these classes. However, misclassification occurs only for data from the NSK class, which is incorrectly predicted as AK or FK. This indicates that the characteristics of NSK images are not consistently represented by the model. This phenomenon also aligns with previous evaluation results that showed low sensitivity to NSK classes, reinforcing the importance of approaches that consider semantic fidelity when dealing with broad and undefined classes.

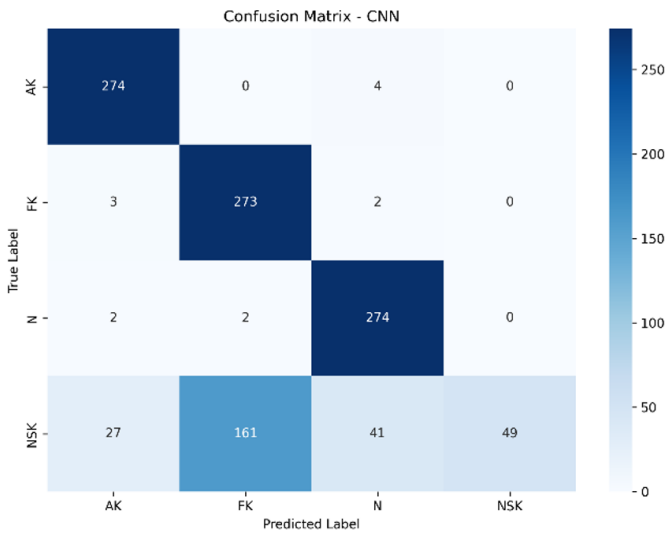

Figure 13. Confusion matrix of MobileNet trained on balanced data

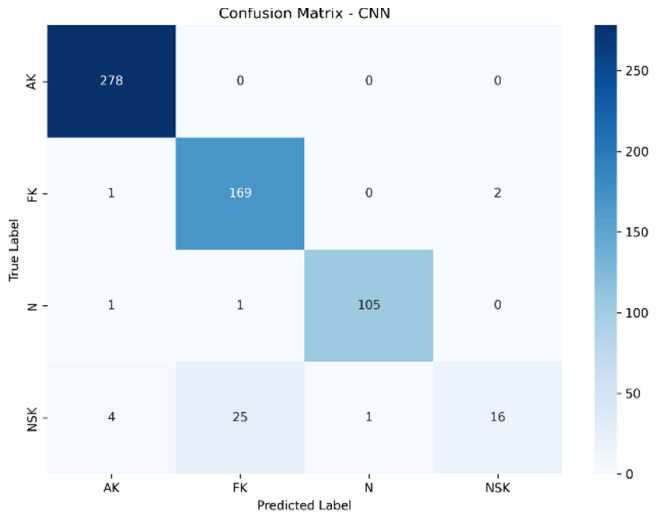

Figure 14. Confusion matrix of XceptionNet trained on imbalanced data

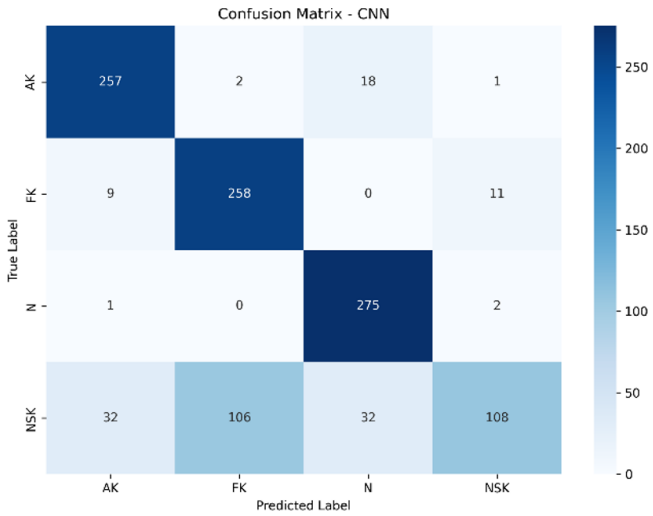

Figure 15. Confusion matrix of XceptionNet trained on balanced data

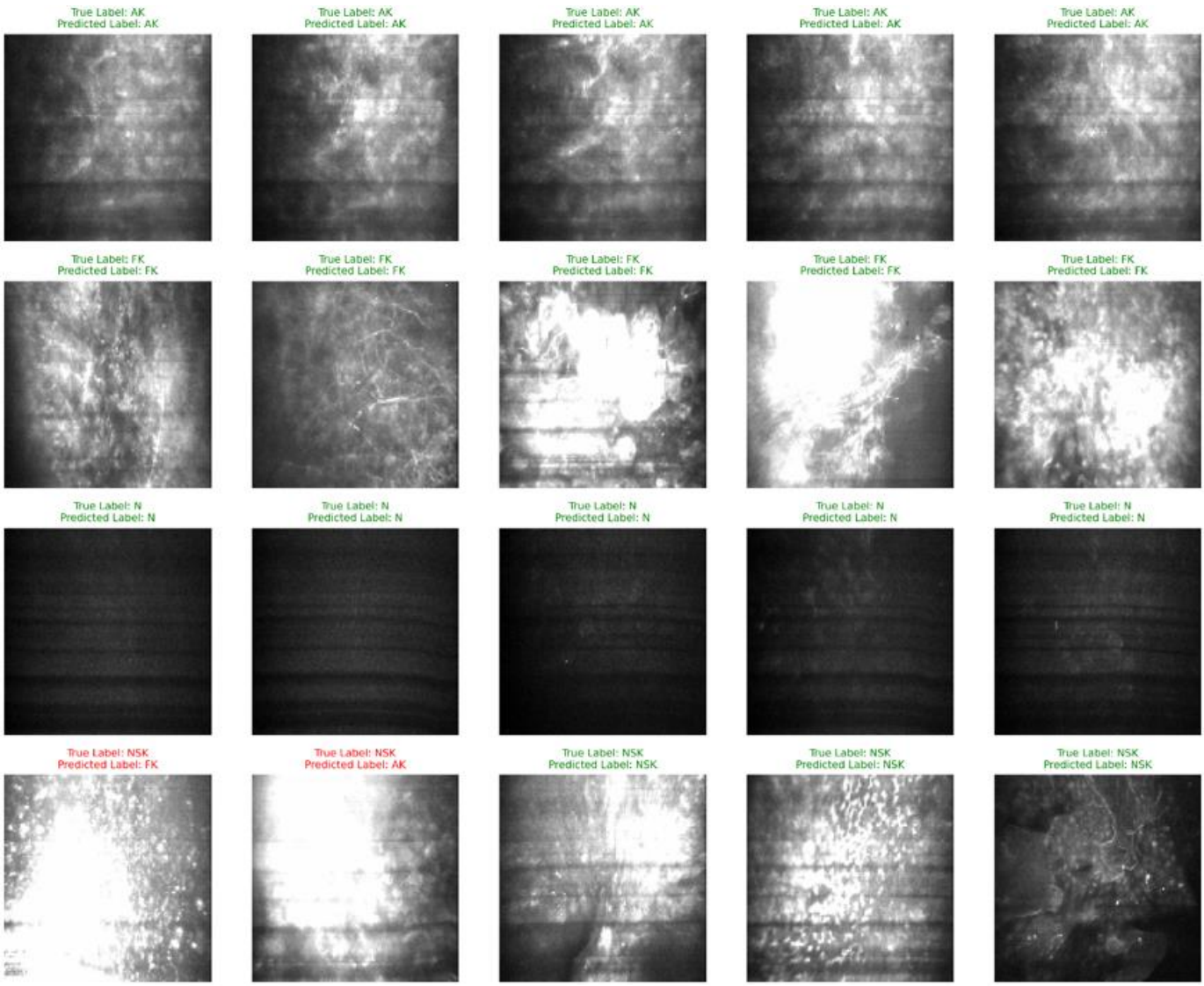

Figure 16. Sample prediction results from the best model (MobileNet trained on imbalanced data)

4. CONCLUSION

This study critically examined the impact of data balancing using DCGAN on automated multiclass keratitis classification with deep convolutional neural networks. Although GAN-based data synthesis has theoretical potential in addressing

data imbalance, the experimental results consistently demonstrated a decline in classification performance after balancing. This was particularly evident in the decrease of overall accuracy, precision, and F1-score compared to models trained on the original imbalanced dataset. For instance, MobileNet trained on the real-world imbalanced data achieved the highest accuracy of 94.53% with strong sensitivity for dominant classes. In contrast, Xception trained on the balanced dataset incorporating synthetic images yielded a significantly lower accuracy of 80.76% and variable recall values across all target classes. Further evaluation using SSIM analysis revealed that the quality and fidelity of GAN-generated images, especially for underrepresented classes, were insufficient. These synthetic images likely introduced noise into the training process, contributing to the decline of the model's generalization capability.

Notably, fairness across AK, FK, and N classes remained high even in the imbalanced setting, with strong sensitivity and low misclassification. However, the broad and ambiguous nature of the NSK class resulted in poor representation and frequent misclassification, emphasizing a trade-off in performance introduced by GAN. This suggests that using GANs for balancing without careful assessment may result in representational dilution rather than meaningful improvements in model fairness. The findings emphasize the importance of not only balancing class distribution but also ensuring the semantic realism and diagnostic relevance of synthetic data in medical image classification tasks.

However, this study is limited to DCGAN and a single medical image dataset, which may restrict the generalizability of the results. Future research should explore other GAN variants (e.g., StyleGAN, CycleGAN, diffusion-based GAN) and more diverse datasets, focusing on adaptive, quality-aware balancing methods or hybrid balancing frameworks that combine standard balancing (SMOTE, Random Weighted, etc.) and GAN-based synthesis. Such approaches can incorporate domain-specific knowledge to enhance robustness, fairness, and diagnostic reliability.

ACKNOWLEDGMENT

We sincerely acknowledge the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, through the Directorate of Research, Technology, and Community Service, for their financial support provided under the Fundamental Grant scheme for the 2025 fiscal year. Our appreciation also goes to the Institute for Research and Community Service (LPPM) of Bumigora University for their continuous support and facilitation throughout this research.

REFERENCES

[1] Ting, D.S.J., Ho, C.S., Deshmukh, R., Said, D.G., Dua, H.S. (2021). Infectious keratitis: An update on epidemiology, causative microorganisms, risk factors, and antimicrobial resistance. Eye, 35(4): 1084-1101. https://doi.org/10.1038/s41433-020-01339-3

[2] Ong, Z.Z., Sadek, Y., Liu, X., Qureshi, R., Liu, S.H., Li, T., Ting, D.S.J. (2023). Diagnostic performance of deep learning in infectious keratitis: A systematic review and meta-analysis protocol. BMJ Open, 13(5): e065537. https://doi.org/10.1136/bmjopen-2022-065537

[3] Sharma, N., Bagga, B., Singhal, D., Nagpal, R., Kate, A., Saluja, G., Maharana, P.K. (2022). Fungal keratitis: A review of clinical presentations, treatment strategies and outcomes. The Ocular Surface, 24: 22-30. https://doi.org/10.1016/j.jtos.2021.12.001

[4] Sharma, S.P., Dwivedi, S., Kumar, S., Dhama, K., Sharma, A.K. (2023). Bacterial and fungal keratitis: Current trends in its diagnosis and management. Current Clinical Microbiology Reports, 10(4): 266-278. https://doi.org/10.1007/s40588-023-00210-9

[5] Redd, T.K., Prajna, N.V., Srinivasan, M., Lalitha, P., Krishnan, T., Rajaraman, R., Song, X. (2022). Image-based differentiation of bacterial and fungal keratitis using deep Convolutional Neural Networks. Ophthalmology Science, 2(2): 100119. https://doi.org/10.1016/j.xops.2022.100119

[6] Montgomery, M.L., Fuller, K.K. (2020). Experimental models for fungal keratitis: An overview of principles and protocols. Cells, 9(7): 1713. https://doi.org/10.3390/cells9071713

[7] Niu, L., Liu, X., Ma, Z., Yin, Y., Sun, L., Yang, L., Zheng, Y. (2020). Fungal keratitis: Pathogenesis, diagnosis and prevention. Microbial Pathogenesis, 138: 103802. https://doi.org/10.1016/j.micpath.2019.103802

[8] Ahmadikia, K., Aghaei Gharehbolagh, S., Fallah, B., Naeimi Eshkaleti, M., Malekifar, P., Rahsepar, S., Mahmoudi, S. (2021). Distribution, prevalence, and causative agents of fungal keratitis: A systematic review and meta-analysis (1990 to 2020). Frontiers in Cellular and Infection Microbiology, 11: 698780. https://doi.org/10.3389/fcimb.2021.698780

[9] Mills, B., Radhakrishnan, N., Rajapandian, S.G.K., Rameshkumar, G., Lalitha, P., Prajna, N.V. (2021). The role of fungi in fungal keratitis. Experimental Eye Research, 202: 108372. https://doi.org/10.1016/j.exer.2020.108372

[10] Essalat, M., Abolhosseini, M., Le, T.H., Moshtaghion, S.M., Kanavi, M.R. (2023). Interpretable deep learning for diagnosis of fungal and acanthamoeba keratitis using in vivo confocal microscopy images. Scientific Reports, 13(1): 8953. https://doi.org/10.1038/s41598-023-35085-9

[11] Ghenciu, L.A., Faur, A.C., Bolintineanu, S.L., Salavat, M.C., Maghiari, A.L. (2024). Recent advances in diagnosis and treatment approaches in fungal keratitis: A narrative review. Microorganisms, 12(1): 161. https://doi.org/10.3390/microorganisms12010161

[12] Donovan, C., Arenas, E., Ayyala, R.S., Margo, C.E., Espana, E.M. (2022). Fungal keratitis: Mechanisms of infection and management strategies. Survey of Ophthalmology, 67(3): 758-769. https://doi.org/10.1016/j.survophthal.2021.08.002

[13] Maharana, K., Mondal, S., Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3(1): 91-99. https://doi.org/10.1016/j.gltp.2022.04.020

[14] Lin, C., Tsai, C.F., Lin, W.C. (2023). Towards hybrid over-and under-sampling combination methods for class imbalanced datasets: An experimental study. Artificial Intelligence Review, 56(2): 845-863. https://doi.org/10.1007/s10462-022-10186-5

[15] Saaim, K., Srinath, S., Fu, S. (2022). Generative models for data synthesis. Doctoral dissertation, University of Alberta, Canada. https://hal.science/hal-03911560v1/file/Final_Report_MM811.pdf.

[16] Sampath, V., Maurtua, I., Aguilar Martin, J.J., Gutierrez, A. (2021). A survey on generative adversarial networks for imbalance problems in computer vision tasks. Journal of Big Data, 8(1): 27. https://doi.org/10.1186/s40537-021-00414-0

[17] Dablain, D.A., Bellinger, C., Krawczyk, B., Chawla, N.V. (2023). Efficient augmentation for imbalanced deep learning. In 2023 IEEE 39th International Conference on Data Engineering (ICDE), Anaheim, CA, USA, pp. 1433-1446. https://doi.org/10.1109/ICDE55515.2023.00114

[18] Arora, A., Arora, A. (2022). Generative adversarial networks and synthetic patient data: Current challenges and future perspectives. Future Healthcare Journal, 9(2): 190-193. https://doi.org/10.7861/fhj.2022-0013

[19] Kuo, M.T., Hsu, B.WY., Yin, Y.K., Fang, P.C., Lai, H.Y., Chen, A., Tseng, V.S. (2020). A deep learning approach in diagnosing fungal keratitis based on corneal photographs. Scientific Reports, 10(1): 14424. https://doi.org/10.1038/s41598-020-71425-9

[20] Ghosh, A.K., Thammasudjarit, R., Jongkhajornpong, P., Attia, J., Thakkinstian, A. (2022). Deep learning for discrimination between fungal keratitis and bacterial keratitis: DeepKeratitis. Cornea, 41(5): 616-622. https://doi.org/10.1097/ICO.0000000000002830

[21] Hung, N., Shih, A.K.Y., Lin, C., Kuo, M.T., Hwang, Y.S., Wu, W.C., Hsiao, C.H. (2021). Using slit-lamp images for deep learning-based identification of bacterial and fungal keratitis: Model development and validation with different Convolutional Neural Networks. Diagnostics, 11(7): 1246. https://doi.org/10.3390/diagnostics11071246

[22] Mayya, V., Kamath Shevgoor, S., Kulkarni, U., Hazarika, M., Barua, P.D., Acharya, U.R. (2021). Multi-scale convolutional neural network for accurate corneal segmentation in early detection of fungal keratitis. Journal of Fungi, 7(10): 850. https://doi.org/10.3390/jof7100850

[23] Mita, S., Sujeeth, A., Aiello, G., Patti, D., Gennaro, F., Scelba, G., Cacciato, M. (2022). Power loss modelling of GaN HEMT-based 3L-ANPC three-phase inverter for different PWM techniques. In 2022 24th European Conference on Power Electronics and Applications (EPE'22 ECCE Europe), Hanover, Germany, pp. P.1-P.10.

[24] Won, Y.K., Lee, H., Kim, Y., Han, G., Chung, T.Y., Ro, Y.M., Lim, D.H. (2023). Deep learning-based classification system of bacterial keratitis and fungal keratitis using anterior segment images. Frontiers in Medicine, 10: 1162124. https://doi.org/10.3389/fmed.2023.1162124

[25] Li, D.J., Huang, B.L., Peng, Y. (2023). Comparisons of artificial intelligence algorithms in automatic segmentation for fungal keratitis diagnosis by anterior segment images. Frontiers in Neuroscience, 17: 1195188. https://doi.org/10.3389/fnins.2023.1195188

[26] Lv, J., Zhang, K., Chen, Q., Chen, Q., Huang, W., Cui, L., Lin, H. (2020). Deep learning-based automated diagnosis of fungal keratitis with in vivo confocal microscopy images. Annals of Translational Medicine, 8(11): 706. https://doi.org/10.21037/atm.2020.03.134

[27] Smaida, M., Yaroshchak, S., El Barg, Y. (2021). DCGAN for enhancing eye diseases classification. In CMIS, pp. 22-33. https://www.academia.edu/download/82254652/paper3.pdf.

[28] Gurusubramani, S., Latha, B. (2025). Deep convolutional generative adversarial network for improved cardiac image classification in heart disease diagnosis. Journal of Imaging Informatics in Medicine, 38: 2146-2169. https://doi.org/10.1007/s10278-024-01343-z

[29] La Salvia, M., Torti, E., Leon, R., Fabelo, H., Ortega, S., Martinez-Vega, B., Leporati, F. (2022). Deep convolutional generative adversarial networks to enhance artificial intelligence in healthcare: A skin cancer application. Sensors, 22(16): 6145. https://doi.org/10.3390/s22166145

[30] Draksharam, M., Rao, K.V. (2025). A transformative approach for multi-class glaucoma detection using deep learning attention-vision transformer model. Mathematical Modelling of Engineering Problems, 12(11): 3852-3860. https://doi.org/10.18280/mmep.121110

[31] Yahiro, I., Ishida, T., Yokoya, N. (2023). Flooding regularization for stable training of generative adversarial networks. arXiv preprint arXiv:2311.00318. https://doi.org/10.48550/arXiv.2311.00318

[32] Susan, S., Kumar, A. (2021). The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent state of the art. Engineering Reports, 3(4): e12298. https://doi.org/10.1002/eng2.12298

[33] Saadi, S., Nini, B., Kada, B. (2025). DepthFusion: A depth-guided framework combining GAN and diffusion for high-fidelity 3D reconstruction from single images. Ingénierie des Systèmes d'Information, 30(8): 2157-2163. https://doi.org/10.18280/isi.300821

[34] Switrayana, I.N., Hadi, S., Sulistianingsih, N. (2024). A robust gender recognition system using Convolutional Neural Network on Indonesian speaker. Sistemasi: Jurnal Sistem Informasi, 13(3): 1008-1021. https://doi.org/10.32520/stmsi.v13i3.3698

[35] AbdulRazek, M., Khoriba, G., Belal, M. (2023). GAN-GA: A generative model based on genetic algorithm for medical image generation. arXiv preprint arXiv:2401.00314. https://doi.org/10.3389/978-2-8325-1231-9

[36] Skandarani, Y., Jodoin, P.M., Lalande, A. (2023). Gans for medical image synthesis: An empirical study. Journal of Imaging, 9(3): 69. https://doi.org/10.3390/jimaging9030069

[37] Switrayana, I.N., Maulidevi, N.U. (2022). Collaborative convolutional autoencoder for scientific article recommendation. In 2022 9th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), Semarang, Indonesia, pp. 96-101. https://doi.org/10.1109/ICITACEE55701.2022.9924130

[38] Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., Ghayvat, H. (2021). CNN variants for computer vision: History, architecture, application, challenges and future scope. Electronics, 10(20): 2470. https://doi.org/10.3390/electronics10202470

[39] Liang, S., Zhong, J., Zeng, H., Zhong, P., Li, S., Liu, H., Yuan, J. (2023). A structure-aware convolutional Neural Network for automatic diagnosis of fungal keratitis with in vivo confocal microscopy images. Journal of Digital Imaging, 36(4): 1624-1632.

https://doi.org/10.1007/s10278-021-00549-9
[40] Rudi Kurniawan, S., Mohamad, F.S. (2025). Improved classification of arsenic-affected skin diseases through image. Processing and Transfer Learning, 19(1): 71-88. https://doi.org/10.21512/commit.v19i1.11891
[41] Pradana, R.C., Suhartono, D. (2025). A cost-sensitive hybrid model of albert model and Convolutional Neural Network for personality classification. CommIT (Communication and Information Technology) Journal, 19(1): 89-99. https://journal.binus.ac.id/index.php/commit/article/download/11822/5395.
[42] Diyasa, I.G.S.M., Sunarko, V.I., Puspaningrum, E.Y., Asy'ari, V., Ibrahim, M.Z. (2025). Optimization of multi-section and partially augmented magnetic resonance imaging (MRI) images for brain tumor classification using ResNet-50. CommIT (Communication and Information Technology) Journal, 19(1): 115-128. https://doi.org/10.21512/commit.v19i1.12467
[43] Al Najjar, Y. (2024). Comparative analysis of image quality assessment metrics: MSE, PSNR, SSIM and FSIM. International Journal of Science and Research, 13(3): 110-114. https://dx.doi.org/10.21275/SR24302013533
[44] Lacan, A., Sebag, M., Hanczar, B. (2023). GAN-based data augmentation for transcriptomics: Survey and comparative assessment. Bioinformatics, 39(Supplement_1): i111-i120. https://doi.org/10.1093/bioinformatics/btad239