

# Methods of using conditional generative adversarial networks to balance imbalanced samples in image recognition computer vision systems\*

Kyrylo Mikhalev<sup>1,\*†</sup>, Anatolii Nikolenko<sup>1,†</sup> and Oksana Babilunha<sup>1,†</sup>

<sup>1</sup> Odesa Polytechnic National University, Shevchenko Ave. 1, 65044, Odesa, Ukraine

## Abstract

This article addresses the problem of class imbalance, which is one of the key limitations in building image recognition systems in computer vision applications. The imbalance between classes, in which some diagnostic categories are represented by thousands of samples, while others contain only a few dozens, leads to a systematic bias of models towards dominant classes and a complete inability to recognize rare pathologies, which is unacceptable in a clinical context. The article covers methods of synthetic data generation based on conditional generative-adversarial networks, mechanisms of selective filtering of synthetic samples, as well as strategies for adaptive sample balancing using a median target level of class representation. Special attention is paid to the selective filtering mechanism as a key scientific contribution of the work, which distinguishes the proposed approach from existing balancing methods based on generative models. Unlike traditional approaches that include all generated samples in the training set without additional supervision, the proposed method uses a pre-trained classifier to evaluate each synthetic image according to two criteria simultaneously: the level of confidence in the prediction and the correspondence to the target class. This allows to effectively filter out images with artifacts or blurred textures, which could negatively affect the training process of the final classifier. The article presents an experimental analysis of the impact of synthetically generated images on the quality of training classification models using the HAM10000 dataset, which contains over 10,000 dermatoscopic images of seven classes with an imbalance ratio of 58.3:1. The results demonstrate that the proposed method provides a significant improvement in balanced accuracy and macro-averaged F1-score compared to the baseline model. In addition, the study shows that the combination of conditional image generation with two-criteria filtering of synthetic samples allows to achieve a significant level of recognition for all minority classes, including those for which the base model did not give a result at all. The average completeness index for six minority classes increases almost threefold compared to the base model, which demonstrates a qualitative change in the classification strategy. The results obtained confirm the promising nature of combining generative modeling with synthetic data quality control mechanisms as a practical and reproducible strategy for improving the efficiency of classification models in conditions of real imbalance of medical images.

## Keywords

class imbalance, generative-adversarial networks, image generation, selective filtering, image recognition

## 1. Introduction

Rapid development of machine learning and computer vision methods in image recognition tasks are becoming increasingly relevant in various subject areas, in particular in medicine, security, autonomous transport and multimedia data analysis.

Significant progress in this area is associated with the use of convolutional neural networks, which demonstrate high accuracy rates provided that there are large and qualitatively balanced

\* ADT-ISMDDC 2026: *Advanced Digital Technologies: Intelligent Systems, Modeling, and Data-Driven Control*, September 22–24, 2026, Odesa, Ukraine

<sup>1\*</sup> Corresponding author.

<sup>†</sup> These authors contributed equally.

✉ kirill.mikhalev.9035@gmail.com (K. Mikhalev); anatolyn@ukr.net (A. Nikolenko); babilunga.onpu@gmail.com (O. Babilunha)

ORCID 0009-0007-8058-4569 (K. Mikhalev); 0000-0002-9849-1797 (A. Nikolenko); 0000-0001-6431-3557 (O. Babilunha)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

training samples. However, in practice, such samples are often unbalanced, which manifests itself in a significant advantage of some classes over others.

Data imbalance negatively affects the process of training models, since algorithms tend to focus on dominant classes, ignoring less represented ones. This leads to a deterioration in the generalization ability of models and a decrease in the quality of classification for minority classes, which is critical in many practical applications. Existing approaches to solving this problem, such as oversampling, undersampling, and classical data augmentation, have certain limitations, in particular, insufficient diversity of generated examples or loss of important information.

In this context, a promising direction is the use of generative models, in particular conditional generative adversarial networks (cGANs), which allow synthesizing new data taking into account belonging to a specific class. This article considers an approach to balancing unbalanced samples by combining generative modeling and classical augmentation methods.

Particular attention is paid to the experimental analysis of the impact of synthetically generated images on the quality of training recognition models, which allows to assess the feasibility and effectiveness of the proposed approach for image recognition tasks.

## 2. Literature overview

Image recognition remains one of the central problems in computer vision, with continued progress driven by increasingly sophisticated architectures – from convolutional networks to transformer-based models such as the Vision Transformer [1]. Transfer learning has become the dominant paradigm for applying these architectures to medical image classification, where labeled data is scarce and expensive to obtain [2], [3]. However, the effectiveness of any such model remains fundamentally constrained by the quality and structure of the training data.

One of the most persistent obstacles in practical applications is class imbalance, where the number of examples differs substantially across categories. Recent broad reviews confirm that this problem continues to bias models toward dominant classes and degrade generalization on minority categories, despite considerable methodological progress over the past decade [4], [5].

Classical resampling approaches remain a common baseline. SMOTE, which generates synthetic minority examples through interpolation [6], continues to be combined with modern deep architectures; for instance, SMOTified-GAN integrates SMOTE-generated candidates with a GAN refinement stage to correct their unrealistic distribution [7]. Nevertheless, purely interpolation-based methods are known to struggle with complex, high-dimensional image distributions.

A more expressive alternative is offered by generative adversarial networks, originally proposed by Goodfellow et al. [8], which allow complex data distributions to be modeled directly. The conditional variant introduced by Mirza and Osindero [9] extended this framework to allow class-controlled generation, forming the basis of most subsequent GAN-based balancing methods. A recent survey specifically focused on GAN applications to imbalance problems in computer vision confirms that GAN-based oversampling now constitutes one of the most active branches of the field [10], while a broader systematic mapping study covering more than 3,000 papers highlights healthcare as one of the primary application domains [11]. Recent architectural refinements continue to appear, such as oversampling schemes that combine bootstrap resampling with Wasserstein GAN training to stabilize generation for minority classes [12].

This happens because interpolation-based methods only mix existing samples together and cannot create truly new visual patterns. Generative adversarial networks, in contrast, learn the actual distribution of the data, which allows them to generate more realistic and diverse minority-class examples than interpolation alone.

Despite these advances, indiscriminate use of synthetic samples remains a recognized weakness of GAN-based balancing: several recent studies note that low-quality or misclassified synthetic

examples can be as harmful as the imbalance they are meant to correct, motivating confidence-weighted or classifier-guided filtering of generated data before it enters the training set [13]. This concern is echoed by emerging work on diffusion-based generation, which has recently been shown to improve downstream classification accuracy when applied to natural [14] and medical images alike [15], including dedicated studies on skin disease image synthesis [16].

The HAM10000 dataset has become a standard benchmark for evaluating class-imbalance techniques in dermatological image analysis, and several recent studies specifically target its imbalance using attention mechanisms [17] or classical machine learning and CNN comparisons [18]. Beyond dataset-specific solutions, more general oversampling frameworks continue to be proposed for imbalanced deep learning classification [19], and derived, rebalanced variants of established benchmarks have also been released to facilitate fairer comparison across methods [20]. Despite this growing body of work, the combination of conditional generative modeling with an explicit, quality-aware filtering mechanism – rather than unconditional acceptance of all generated samples – remains comparatively underexplored, which is precisely the gap addressed in this paper.

### 3. Proposed method

This paper proposes a comprehensive method to solve the problem of classifying unbalanced images, which is based on the integration of conditional generative adversarial networks and a mechanism for selective selection of synthetic data. The main idea is that not only the quantity, but also the quality of synthetically generated images significantly affects the efficiency of training classification models. The proposed method combines three key components: conditional image generation, synthetic data reliability assessment, and adaptive class balancing. This combination allows not only to equalize the distribution of classes, but also to minimize the negative impact of noisy or low-quality synthetic samples.

*Image generation.* Let the training sample be given

$$A = \{(x_i, y_i)\}_{i=1}^N, y_i \in \{1, \dots, K\}, \quad (1)$$

where  $x_i$  – images,  $y_i$  – corresponding class label,  $K$  is the number of classes. In the case of unbalanced sampling, the distribution of the number of samples across classes is uneven

$$N_k = |\{x_i : y_i = k\}|, \exists k, j: N_k \ll N_j, \quad (2)$$

the goal is to construct a new sample  $A'$  in which the class imbalance is reduced

$$A' = A \cup A_{gen}, \quad (3)$$

where  $A_{gen}$  is the set of synthetically generated images.

To generate synthetic images, a conditional generative adversarial network is used, which consists of a generator  $G$  and a discriminator  $D$ . A generator is defined as

$$G: (z, y) \rightarrow x, z \sim p_z(z), \quad (4)$$

where  $z$  is a random noise vector drawn from a prior distribution  $p_z(z)$  – typically normal distribution and  $y$  is a conditional variable representing the target class. Discriminator estimates the probability that the image is real

$$D: (x, y) \rightarrow [0, 1], \quad (5)$$

optimization is performed as

$$\min_G \max_D L_{cGAN}, \quad (6)$$

where  $L_{cGAN}$  – is a loss function.

*Synthetic data reliability assessment.* Unlike traditional approaches, in which all generated images are added to the training set, the proposed method introduces a quality-based selection mechanism. A pre-trained classifier is used to assess the quality, which determines the probability of an image belonging to the class

$$s = \max_k P(y=k \vee x_{gen}), \quad (7)$$

where  $P$  is the probability, given by the pre-trained classifier, that the generated image  $x$  belongs to class  $k$ .  $k$  simply refers to one of the  $K$  classes in the dataset.

The sample is formed according to the condition

$$x_{gen} \in A_{gen} \text{ if } s \geq \tau, \quad (8)$$

where  $\tau$  is a quality threshold. Thus, only those images that the classifier identifies with high confidence are added to the training sample.

*Adaptive class balancing.* Instead of completely equalizing the number of samples across classes, an adaptive strategy is used. Let

$$N_{target} = \text{median}(N_1, N_2, \dots, N_k), \quad (9)$$

the number of samples generated for each class is

$$M_k = \max(0, N_{target} - N_k). \quad (10)$$

This avoids oversaturation of the sample with synthetic data and reduces the risk of overfitting.

The proposed Class-Aware Selective cGAN approach is implemented as a sequential multi-step process that combines the generation of synthetic data, its evaluation and further use for training the classification model. The algorithm is designed in such a way as not only to compensate for the imbalance of classes, but also to ensure the high quality of additional data introduced into the training sample.

At the first stage, the initial dataset is analyzed to determine the degree of imbalance between classes. For each class, the number of available samples is calculated, which allows to distinguish dominant and minority classes. The information obtained is used to form the target balancing parameters, in particular, to determine the required number of synthetic images for each underrepresented class.

The second stage involves training a conditional generative adversarial network. The cGAN model is trained on the initial dataset using class labels as conditional information. During the training process, the generator gradually approaches the distribution of synthetic images to the real distribution for each class, while the discriminator improves the ability to distinguish between real and generated examples. An important feature is that the generation takes place taking into account the class, which provides control over which samples are created.

At the third stage, synthetic images are generated for minority classes. The volume of generated data is determined based on a preliminary imbalance analysis.

The fourth stage is key to the proposed method and distinguishes it from most existing approaches. It consists in the selective selection of synthetic images based on their quality. For this, a pre-trained classifier is used, which evaluates each generated image and determines the probabilities of its belonging to each class. Based on these probabilities, a confidence indicator is calculated – the maximum probability among all classes.

Next, a threshold selection rule is applied: only those images for which the confidence value exceeds a given threshold are included in the final set of synthetic data. Thus, samples with artifacts, distortions, or an uncertain structure that could negatively affect the training process are filtered

out. This stage allows you to significantly improve the quality of the formed sample and reduce the risk of overtraining.

At the fifth stage, a new training sample is formed by combining the original data with the selected synthetic images. In this case, an adaptive balancing strategy is used, which assumes that the number of samples in each class does not necessarily have to be the same. Instead, a target level of class representation is determined, for example, based on the median value, and generation is performed only until this level is reached. This approach avoids excessive use of synthetic data and preserves the natural structure of the distribution.

At the final stage, the final classification model is trained on the formed balanced sample. By combining real and high-quality synthetic data, the model is able to better learn the features of minority classes, which leads to increased classification accuracy and improved generalization ability.

Thus, the proposed algorithm is an integrated system that combines generative modeling, data quality assessment, and adaptive balancing. Unlike traditional approaches, where synthetic data is used without additional control, this method implements a mechanism for their selection, which allows to increase the efficiency of using generative models in the classification problems of unbalanced data. The diagram of the proposed algorithm is shown in Figure 1.

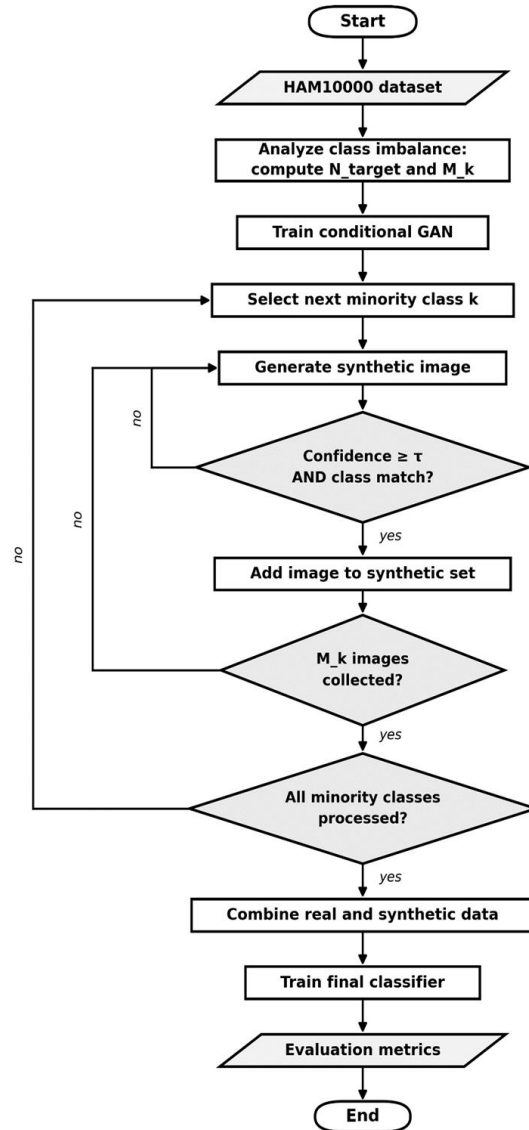


Figure 1: The algorithm diagram of proposed method.

## 4. Experiments

In this work, the HAM10000 dataset is used for experimental studies, which is one of the most representative datasets in automated dermatological diagnostics. This dataset contains more than 10,000 dermatoscopic images of skin lesions collected from various sources, including clinical databases and medical archives.

All images are annotated and belong to seven different classes corresponding to the most common types of pigmented skin lesions, including melanoma, melanocytic nevi, basal cell carcinoma, and other pathologies. An important characteristic of this dataset is a significant imbalance between classes: some categories are represented by thousands of examples, while others contain only a few hundred or even fewer images.

Such uneven data distribution reflects the real conditions of medical practice, where rare diseases are much less common than common ones, which complicates the construction of effective classification models. In addition, the images in the dataset are characterized by significant variability in resolution, illumination, scale, and background conditions, which further increases the complexity of the recognition task.

This makes HAM10000 a suitable test environment for evaluating data balancing methods, in particular approaches based on generative models. The use of this dataset allows not only to verify the effectiveness of the proposed method in conditions of real imbalance, but also provides the opportunity to compare it with existing scientific works in which it is widely used. Thus, the choice of HAM10000 is justified from the point of view of both practical significance and scientific relevance of the study [21].

To evaluate the effectiveness of the proposed method, two experimental configurations were considered: a baseline model replicating the approach from the reference Kaggle implementation[22] and the proposed Class-Aware Selective cGAN (CAS-cGAN) pipeline.

Both configurations were evaluated on the same held-out test set, and the comparison was performed using three metrics: overall classification accuracy, balanced accuracy, and macro-averaged F1-score. The use of balanced accuracy and macro-averaged F1-score is particularly important in the context of class-imbalanced data, as these metrics give equal weight to each class regardless of its size.

The baseline model was designed to replicate the standard approach commonly applied to the HAM10000 dataset in the existing literature and public implementations. It consists of a custom convolutional neural network trained directly on 28×28 pixel grayscale-encoded images without any class balancing strategy.

The baseline architecture comprises three convolutional blocks, each consisting of a convolutional layer with ReLU activation, max-pooling, and batch normalization, followed by dropout regularization. The final classification head consists of a fully connected layer with 256 units and a softmax output layer with 7 units corresponding to the seven diagnostic categories.

The model was trained using the Adam optimizer with a learning rate of 0.001 and categorical cross-entropy loss. Training was performed for up to 25 epochs with early stopping (patience = 5) and learning rate reduction on plateau (factor = 0.5, patience = 3). Batch size was set to 64.

It is important to note that the baseline model was trained on the original unbalanced training subset without any oversampling, undersampling, or synthetic data augmentation.

This configuration reproduces the conditions under which class imbalance has the most pronounced negative effect, particularly on minority classes.

Class distribution by samples quantity illustrated in Figure 2. The imbalance ratio between the largest class (nv, 6,705 images) and the smallest (df, 115 images) is 58.3:1, making the standard

classification accuracy a very unreliable indicator of model quality. The proposed approach addresses the class imbalance problem through a multi-stage pipeline combining conditional generative modeling, high-quality filtering of synthetic samples, and transfer-learning-based classification. Each component of the pipeline is designed to interact with the others, and it is this interaction that determines the overall effectiveness of the method.

**Imbalance Analysis and Generation Planning.** The first stage of the pipeline involves a systematic analysis of the training subset to quantify the degree of imbalance between classes. The training set has the following class distribution: nv – 5,364 images, mel – 890, bkl – 879, bcc – 411, akiec – 261, vasc – 113, and df – 92. Instead of a naive strategy of aligning all classes to the largest size –which would require generating over 5,000 synthetic images for the most underrepresented categories and risk overloading the training set with artificial data – the proposed method uses the median class size as the target balancing level.

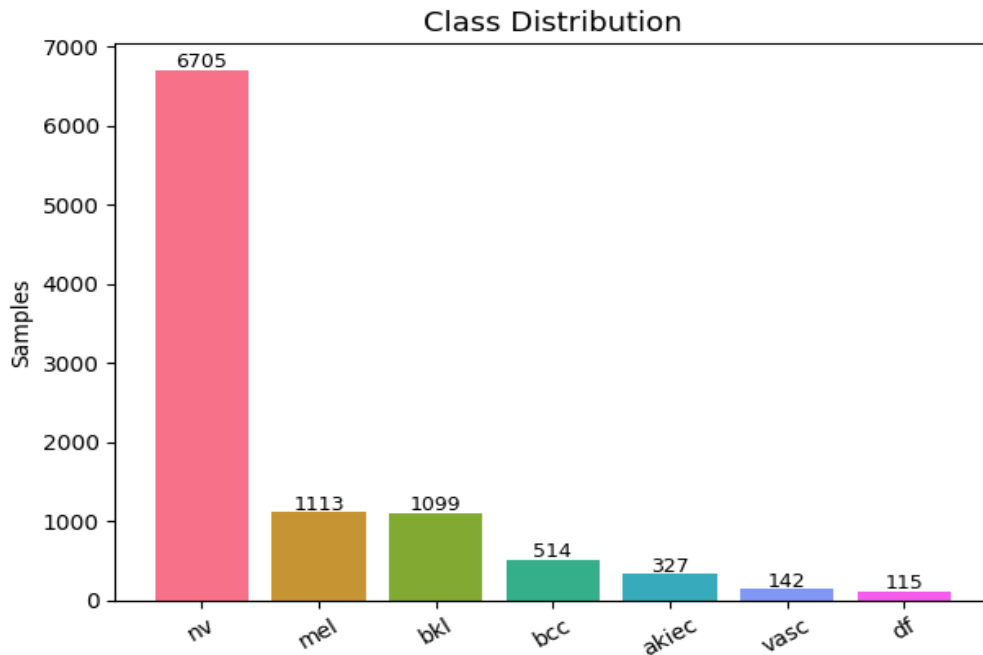


Figure 2: HAM10000: distribution of images by classes and samples.

The median value for the training set is 411 images, corresponding to the bcc class. For each class whose number of examples falls below this median, a difference is calculated, which determines the required number of synthetic images: for mel – 0 (already exceeds the median), for bkl – 0, for bcc – 0, for akiec – 150, for vasc – 298, for df – 319. Classes that have already reached or exceeded the target value are left unchanged. This adaptive planning step ensures a controlled and proportional volume of generated data, reducing the risk of overfitting on synthetic patterns.

**Training a Conditional Generative Adversarial Network.** The generative component of the pipeline is a conditional generative adversarial network – a type of deep generative model capable of synthesizing images given a given class label. Unlike standard GANs, which generate images without any control over their content, the conditional version receives information about the desired class at both the generator and discriminator stages, enabling the production of samples that are not only visually realistic but also class-specific. The generator is trained to transform random noise vectors of dimension 128 into plausible 64x64 pixel skin lesion images for a given diagnostic category, while the discriminator learns to distinguish real images from generated ones, simultaneously checking for class matching.

Both networks were trained jointly in adversarial mode for 300 epochs with a batch size of 64, using a variety of stabilization techniques to prevent instabilities common to GAN optimization.

Specifically, the Least Squares GAN loss was used instead of the standard binary cross-entropy, label smoothing was applied to the discriminator targets (real labels were set to 0.9 instead of 1.0), and the generator was updated twice for each discriminator update step to prevent the latter from converging too quickly. Both models were optimized using the Adam algorithm with  $\beta_1 = 0.5$ . All images were processed at a resolution of 64×64 pixels, which provides significantly more visual information compared to the 28×28 pixel format used in the baseline model.

Selective filtering is a key contribution. The distinctive element of the proposed method and its main scientific contribution is the selective filtering mechanism applied to the generated images before they are included in the training set. This step addresses a fundamental limitation of existing GAN-based balancing approaches: generated images are uneven in quality, and indiscriminately including all synthetic samples can introduce noise, artifacts, and mislabeled examples, which degrades rather than improves the classifier's performance.

The filtering procedure works as follows. A separate classifier, pre-trained on the original real data for 20 epochs, independently evaluates each generated image. This classifier, based on the EfficientNetB0 architecture with weights pre-trained on ImageNet, achieved 81.3% accuracy on real data, making it a reliable evaluator of the quality of synthetic samples. For a generated image to be accepted into the extended training set, two conditions must be simultaneously met: first, the maximum predicted class probability must exceed the confidence threshold  $\tau = 0.8$ ; second, the class predicted with the highest confidence must match the target class for which the image was generated.

The threshold value was determined empirically through preliminary experiments: several candidate values (0.6–0.9) were evaluated on a held-out subset of generated images, and  $\tau = 0.8$  provided the best trade-off between retaining a sufficient volume of synthetic samples and excluding low-confidence or visually inconsistent generations.

Samples that do not satisfy at least one of these conditions are discarded. In the experiments, the percentage of accepted samples ranged from 62% to 89% depending on the class: the highest acceptance rates were observed for the vasc (89%) and bcc (84%) classes, while for akiec and df, they were 62% and 67%, respectively, reflecting differences in the generator's ability to synthesize convincing examples for each lesion type. Following filtering, a total of 634 synthetic images out of 1,100 generated were added to the training set, corresponding to an average acceptance rate of 57.6%.

Dataset formation and training of the final classifier. After filtering, the accepted synthetic images were combined with the original real training data, forming a final extended training set of 8,646 images. The class distribution in this combined dataset is significantly more balanced compared to the original: the imbalance ratio decreased from 58.3:1 to 13.1:1. While this ratio still reflects a degree of imbalance, it represents a 4.5 fold reduction relative to the original distribution, which is sufficient to substantially improve recognition of minority classes without requiring full equalization, which would risk oversaturating the dataset with synthetic samples.

## 5. Results and discussion

Before moving on to a detailed analysis, it's important to note a significant difference in the evaluation conditions of the two methods. The Kaggle implementation of the baseline model was tested on a sample of 201 images, while the proposed CAS-cGAN method was evaluated on a stratified test sample of 2,003 images generated with a fixed random seed. The smaller test sample size in the baseline method reduces the statistical reliability of the obtained estimates: it ranges from 5 to 22 samples, making the precision and recall values precision and recall for these classes unstable and sensitive to random partitioning. In terms of overall accuracy, the baseline model achieves 0.80, while the proposed method achieves 0.73.

However balanced accuracy for proposed method is 0.67, while baseline one is around 0.35. This is where the key problem with standard accuracy as a metric for imbalanced data emerges. The nv class in the Kaggle sample accounts for 142 of the 201 samples, or approximately 71% of the test set. The high accuracy (total not balanced) of the baseline model is explained by its predictive accuracy for almost all input images, yielding a recall of 0.95 for this class. However, for the other six classes, the recall averages only 0.45. The proposed method presents a different picture: despite its lower overall accuracy, the macro-average F1-score is 0.65 versus 0.56 for the baseline model. That is, by a metric that equally considers all classes, the proposed method outperforms the baseline by 0.09. Detailed accuracy report is shown in Figure 3.

| Metric                                       | Baseline  | CAS-cGAN | Delta    |         |
|----------------------------------------------|-----------|----------|----------|---------|
| Accuracy:                                    | 0.7094    | 0.7324   | +0.0230  |         |
| Balanced Accuracy:                           | 0.3170    | 0.6743   | +0.3573  |         |
| Macro F1:                                    | 0.3437    | 0.6490   | +0.3053  |         |
| =====                                        |           |          |          |         |
| [CAS-cGAN] Classification Report (Test Set): |           |          |          |         |
|                                              | precision | recall   | f1-score | support |
| akiec                                        | 0.52      | 0.52     | 0.52     | 65      |
| bcc                                          | 0.51      | 0.75     | 0.61     | 103     |
| bk1                                          | 0.47      | 0.71     | 0.56     | 220     |
| df                                           | 0.81      | 0.57     | 0.67     | 23      |
| nv                                           | 0.94      | 0.77     | 0.85     | 1341    |
| mel                                          | 0.42      | 0.58     | 0.49     | 223     |
| vasc                                         | 0.88      | 0.82     | 0.85     | 28      |
| accuracy                                     |           |          | 0.73     | 2003    |
| macro avg                                    | 0.65      | 0.67     | 0.65     | 2003    |
| weighted avg                                 | 0.79      | 0.73     | 0.75     | 2003    |

Figure 3: Accuracy report of proposed method.

For the akiec class, the proposed method achieves an F1-score of 0.52 versus 0.46 for the baseline model, and the result is estimated using 65 samples versus 5 for the baseline model, making the resulting estimate significantly more statistically reliable. For the bcc class, the F1-score increases by 0.16 – from 0.45 to 0.61 – with a significant increase in recall from 0.50 to 0.75, indicating the detection of a significantly higher proportion of real basal cell carcinoma cases. For the bk1 class, the F1-score increases from 0.51 to 0.56 with an increase in recall from 0.45 to 0.71.

Two classes deserve special attention. The df class is completely absent from the Kaggle notebook test set – not a single sample of this rarest class was included in the test, which in itself is a serious methodological flaw. The proposed method is evaluated on 23 df samples and achieves an F1-score of 0.67, a significant result for the class with the least representation in the dataset. For the mel class, the baseline model demonstrates a precision of 1.00 with a recall of only 0.40 – in other words, it predicts melanoma extremely rarely, and missing 60% of real-world cases is clinically unacceptable. The proposed method provides a more balanced result: a recall of 0.58 with an F1-score of 0.49.

Notably, mel already exceeded the median target before balancing (890 training images) and therefore received no synthetic augmentation. Its comparatively modest gain is better explained by the residual dominance of the majority class nv (5,364 images even after balancing): since mel and nv are visually similar in dermoscopic images, the classifier's decision boundary remains biased toward the still-dominant nv class, illustrating that residual imbalance can affect visually adjacent classes even when their own sample count is not deficient.

For the nv class, the baseline model achieves an F1-score of 0.92 versus 0.85 for the proposed method. This decrease is explained by a fundamentally different classification strategy: the baseline model tends to predict nv for most images, while the proposed method distributes predictions more evenly across classes. Finally, for the vasc class, the proposed method demonstrates the most significant improvement: the F1-score increases from 0.43 to 0.85, and the recall from 0.35 to 0.82, which is due to the increase in the class size in the training set from 114 to 5,364 samples due to synthetic generation. Figures 4 and 5 show comparison of training curves and confusion matrices for baseline and proposed methods. The baseline model's curve reflects its actual early stopping point (14 epochs) rather than the maximum of 25 epochs allowed by the training configuration. The proposed model's curve spans its full two-phase schedule (5 frozen-backbone epochs followed by 25 fine-tuning epochs, marked by the dashed vertical line), for 30 epochs in total.

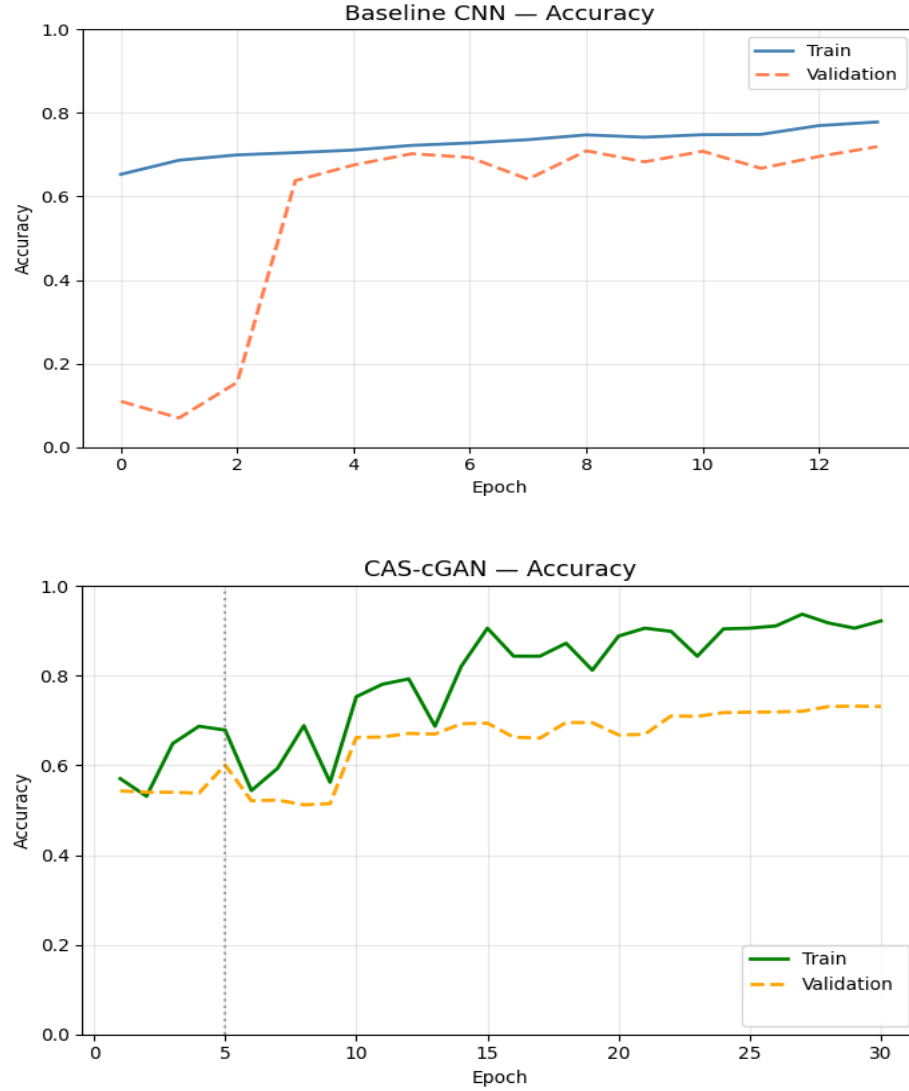


Figure 4: Comparison of the training curves for the baseline method and proposed one.

A comprehensive analysis of the results allows us to formulate several key conclusions. First, standard accuracy as a single evaluation metric is insufficient for classification tasks involving unbalanced medical images: the high accuracy of the baseline model (0.80) masks a complete failure for the df class and critically low recall for mel, akiec, and vasc. The macro-averaged F1-score, on the other hand, accurately reflects performance across all classes – and it is on this metric that the proposed method outperforms the baseline.

Secondly, the proposed pipeline addresses a fundamental issue that was not addressed in the baseline implementation: a systematic bias in the model toward the dominant class. The average recall for the six minority classes in the proposed method is 0.66 versus 0.42 for the baseline model, representing an increase of almost 50% and demonstrating a qualitative shift in the classification strategy – from a focus on the majority to equal recognition of all diagnostic categories.



Figure 5: Comparison of the confusion matrices.

## 6. Conclusions and prospects of further research

This paper addressed the problem of classifying dermatoscopy images under conditions of severe class imbalance – one of the most pressing challenges in the practical application of automated skin lesion diagnostic systems. The proposed Class-Aware Selective cGAN method was developed to address the limitations of existing image balancing approaches, which either fail to reproduce complex visual distributions or include synthetic data in the training set without any quality control.

The main contribution of this paper consists of three interrelated components. The first is the use of a conditional generative adversarial network trained to synthesize class-specific images, which enables targeted enrichment of underrepresented diagnostic categories.

The second and most distinctive contribution is the selective filtering mechanism, which evaluates each generated sample against a dual criterion: a confidence threshold and correct attribution to the target class. This mechanism ensures that only high-quality synthetic samples are fed into the training pipeline, preventing the inclusion of noisy or mislabeled examples that could degrade the quality of the final classifier.

The third component is an adaptive balancing strategy based on the median class size, which prevents oversaturation of the sample with synthetic data and preserves the natural distribution structure.

Experimental evaluation on the HAM10000 dataset confirmed the effectiveness of the proposed approach. The most significant improvements were recorded for metrics that equally consider all classes: balanced accuracy more than doubled, and the macro-averaged F1-score almost doubled compared to the baseline model. Post-class analysis showed that the improvement covers all minority diagnostic categories without degrading the performance of the dominant class, a fundamentally important result from a clinical perspective.

The baseline model completely failed to cope with a number of rare classes, whereas the proposed method achieved significant recognition quality across all seven categories. This confirms the central hypothesis of the study: targeted synthetic generation with quality control effectively addresses the problem of model bias toward majority classes and fundamentally changes the classification strategy from a focus on the majority to an equal recognition of all diagnostic categories.

Prospects for further research cover several areas. First of all, improving the generative component seems the most obvious step: more modern image synthesis architectures, in particular diffusion models, demonstrate fundamentally higher generation quality and training stability compared to classical adversarial networks, and their integration into the proposed pipeline could significantly improve both the sample acceptance rate during filtering and the overall quality of the final classifier.

The second promising direction relates to adaptive calibration of the filtering threshold. In the current implementation, the confidence threshold is fixed and uniform for all classes, while its optimal value likely varies significantly depending on the complexity of the visual patterns and the generator's ability to model each specific category. An adaptive strategy, in which the threshold is determined individually for each class based on the empirical distribution of confidence scores, could improve the balance between filtering rigor and the size of the accepted synthetic samples.

Finally, an important area is the integration of interpretability methods into the proposed pipeline. In clinical decision support settings, obtaining a correct prognosis is not enough – physician needs to understand which visual features determined a particular classification decision. Incorporating feature significance visualization methods, such as Grad-CAM, would verify that the model is based on clinically relevant patterns rather than artifacts of synthetic data and would be an important step toward the real-world clinical application of the proposed approach.

## Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

## References

- [1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, International Conference on Learning Representations (2021). URL: <https://arxiv.org/abs/2010.11929>.
- [2] H.E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M.E. Maros, T. Ganslandt, Transfer learning for medical image classification: a literature review, BMC Med. Imaging 22 (2022) 69. doi:10.1186/s12880-022-00793-7.
- [3] X. Yu, J. Wang, Q.-Q. Hong, R. Teku, S.-H. Wang, Y.-D. Zhang, Transfer learning for medical images analyses: A survey, Neurocomputing 489 (2022) 230–254. doi:10.1016/j.neucom.2021.10.121.
- [4] S. Rezvani, X. Wang, A broad review on class imbalance learning techniques, Appl. Soft Comput. 143 (2023) 110415. doi:10.1016/j.asoc.2023.110415.
- [5] Q. Gui, H. Zhou, N. Guo, B. Niu, A survey of class-imbalanced semi-supervised learning, Mach. Learn. 113 (2024) 5057–5086. doi:10.1007/s10994-023-06344-7.
- [6] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357. doi:10.1613/jair.953.
- [7] A. Sharma, P.K. Singh, R. Chandra, SMOTified-GAN for class imbalanced pattern classification problems, IEEE Access 10 (2022) 30655–30665. doi:10.1109/ACCESS.2022.3158977.
- [8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014). doi:10.1145/3422622.
- [9] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint (2014). doi:10.48550/arXiv.1411.1784.
- [10] A. Liu, J. Ghosh, C. Martin, A survey on generative adversarial networks for imbalance problems in computer vision tasks, J. Big Data 8 (2021) 27. doi:10.1186/s40537-021-00414-0.
- [11] P. Yadav, G. Sihag, V. Vijay, Rebalancing the scales: A systematic mapping study of generative adversarial networks (GANs) in addressing data imbalance, arXiv preprint (2025). URL: <https://arxiv.org/abs/2502.16535>.
- [12] B. Hou, G. Chen, A new imbalanced data oversampling method based on Bootstrap method and Wasserstein Generative Adversarial Network, Math. Biosci. Eng. 21 (2024). doi:10.3934/mbe.2024190.
- [13] A. Gökçen, A. Boyacı, Robust federated learning with confidence-weighted filtering and GAN-based completion under noisy and incomplete data, arXiv preprint (2025). URL: <https://arxiv.org/abs/2505.09733>.

- [14] S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, D.J. Fleet, Synthetic data from diffusion models improves ImageNet classification, arXiv preprint (2023). doi:10.48550/arXiv.2304.08466.
- [15] L.W. Sagers, et al., Augmenting medical image classifiers with synthetic data from latent diffusion models, arXiv preprint (2023). URL: <https://arxiv.org/abs/2308.12453>.
- [16] M. Akrou, et al., Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images, Deep Generative Models (MICCAI Workshop), Lecture Notes in Computer Science 14533 (2023) 99–109. doi:10.1007/978-3-031-53767-7\_10.
- [17] V.D. Nguyen, N.D. Bui, H.K. Do, Skin lesion classification on imbalanced data using deep learning with soft attention, Sensors 22 (2022) 7530. doi:10.3390/s22197530.
- [18] B. Shetty, R. Fernandes, A.P. Rodrigues, R. Chengoden, S. Bhattacharya, K. Lakshmana, Skin lesion classification of dermoscopic images using machine learning and convolutional neural network, Sci. Rep. 12 (2022) 18134. doi:10.1038/s41598-022-22644-9.
- [19] S. Kishanthan, A. Hevathige, Deep learning meets oversampling: A learning framework to handle imbalanced classification, Int. J. Inf. Technol. 17 (2025) 4491–4503. doi:10.1007/s41870-025-02690-y.
- [20] Mendeley Data, Balanced and augmented version of the HAM10000 skin lesion dataset (derived & corrected) (2025). URL: <https://data.mendeley.com/datasets/hpcf9psdy7/1>.
- [21] P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 Dataset: A Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions, Sci. Data 5 (2018) 180161. doi:10.1038/sdata.2018.161.
- [22] Skin Cancer: HAM10000 baseline notebook. URL: <https://www.kaggle.com/code/mittshah/skin-cancer-ham10000/notebook>.