

DOI: <https://doi.org/10.15276/ict.03.2026.53>

UDC 004.93'1:004.032.26

Application of a diffusion model with improved adaptive class-balancing criteria for imbalanced datasets

Kyrylo B. Mikhalev¹⁾

ORCID: <https://orcid.org/0009-0007-8058-4569>; kirill.mikhalev.9035@gmail.com. Scopus Author ID: 59390689500

Anatolii O. Nikolenko¹⁾

ORCID: <https://orcid.org/0000-0002-9849-1797>; anatolyn@ukr.net. Scopus Author ID: 57197491335

¹⁾ Odesa Polytechnic National University, Shevchenko Ave. 1, Odesa, 65044, Ukraine

ABSTRACT

This article addresses the problem of class imbalance, which remains one of the key limitations in building image recognition systems in computer vision applications. The imbalance between classes, in which some categories are represented by thousands of samples while others contain only a few dozen, leads to a systematic bias of models towards dominant classes and to poor recognition of rare categories, which is unacceptable in a clinical context. Generative models have become a common instrument for addressing this problem, and diffusion models in particular now offer a more stable and higher-quality alternative to generative-adversarial networks. However, one component of such pipelines has received considerably less attention than the generative part itself: the criteria that decide how a limited number of synthetic images should be distributed among classes. Existing approaches derive this decision from the accuracy of the classifier measured on a single epoch, which is unreliable for classes represented by a few dozen samples and carries no information about which classes the errors are directed to. The article proposes improved adaptive class-balancing criteria that address both shortcomings. The first component is an exponentially smoothed estimate of per-class accuracy, which preserves sensitivity to persistent weakness of a class while suppressing fluctuations caused by measurement on very few samples. The second component is an error-concentration term derived from the confusion matrix, which distinguishes a class whose errors are directed almost entirely at one neighbouring category from a class whose errors are spread across many, and adjusts its share of generated images accordingly. The proposed criteria were evaluated on the HAM10000 dataset, which contains over 10 000 dermatoscopic images of seven classes with an imbalance ratio of 58:1, using a class-conditional diffusion model as the generator and a single shared pool of synthetic images, so that the criteria could be isolated from every other factor. The results demonstrate that the proposed criteria improve balanced accuracy, macro-averaged F1-score and the Matthews correlation coefficient compared with the accuracy-based criterion, with the improvement concentrated on the smallest classes and the recognition of the dominant class left unchanged. The variability of the per-class allocation across update points is reduced by 63% on average and by 95% for the vascular lesion class, which confirms directly the instability of accuracy-based criteria. The results obtained show that the way synthetic data is distributed among classes affects the outcome more than the fact of its addition, and that careful design of the allocation criteria can compensate for a substantial part of the computational budget normally spent on generation.

Keywords: class imbalance; diffusion models; adaptive balancing criteria; synthetic data generation; image recognition

For citation: Mikhalev K. B., Nikolenko A. O. “Application of a diffusion model with improved adaptive class-balancing criteria for imbalanced datasets”. *Informatics. Culture. Technology*. 2026; Vol. 3 No. 2 (4): 697–707. DOI: <https://doi.org/10.15276/ict.03.2026.53>

Introduction. Image classification is one of the most widely used tasks in modern computer vision. It is applied in medical diagnostics, industrial quality control, agriculture, and remote sensing. The performance of deep learning models on this task depends heavily on having enough labeled data, spread evenly across all target classes.

In practice, this condition is rarely met. Many real-world datasets are strongly imbalanced: a small number of classes holds most of the available samples, while other classes – often the most important ones in practice, such as rare disease types or uncommon defect categories – have very few. Classifiers trained on such data become biased toward the larger classes. They reach high overall accuracy, but this number is misleading, because the models perform poorly on exactly the classes that matter most.

To solve this problem, researchers increasingly use generative models to create synthetic samples for the smaller classes. Generative adversarial networks were the main tool for this task for a long time, but diffusion models have recently become a more stable alternative that produces higher-quality images, and they are now actively used for class-imbalance correction.

One part of such pipelines has received little attention: the criteria that decide how many synthetic samples each class should receive. Existing methods usually base this decision on the classifier's accuracy for each class, recalculated after every training epoch. This idea is simple

and reasonable, but it has two weaknesses. First, for classes with only a few dozen samples, a single-epoch accuracy value changes a lot from one epoch to the next. Second, accuracy alone says nothing about how the errors of a class are spread across the other classes.

The goal of this study is to find out whether improving these allocation criteria – by making the accuracy signal more stable over time and by adding information about how a class's errors are distributed – increases the effectiveness of diffusion-based class balancing, while the generative model itself stays unchanged.

The research question is therefore: how much does the choice of adaptive sampling criteria affect class-balancing effectiveness and the final classification performance on strongly imbalanced image datasets.

The purpose of the study. The purpose of this study is to improve the criteria that govern the allocation of synthetic images among classes in diffusion-based balancing pipelines, so that the allocation remains stable for classes represented by a small number of samples and takes into account the distribution of the classifier errors. To achieve this purpose, the following tasks are addressed: to formalise an allocation criterion that combines a smoothed accuracy estimate with an error-concentration term; to implement the criterion within a class-conditional diffusion pipeline; and to evaluate it experimentally against the existing accuracy-based criterion under identical conditions.

Analysis of recent research. Class imbalance is a long-standing obstacle to reliable image classification, because standard classifiers fit the larger classes well and the smaller ones poorly. Recent surveys give a broad overview of how the field addresses this: [1] reviews deep long-tailed learning and groups existing methods into class re-balancing, information augmentation and module improvement, [2] focuses specifically on class imbalance in computer vision tasks, and [3] examines how far deep learning itself has actually solved the problem. Common solutions work either at the data level, by resampling the training set, or at the loss level, by re-weighting the contribution of each class, but both have limits when a class has only a few dozen samples.

A different data-level strategy is generative oversampling: instead of copying or interpolating existing samples of a small class, a generative model creates new ones. Generative adversarial networks were the standard tool for this task for several years [4], but their adversarial training is unstable and prone to mode collapse, which hurts exactly those classes that already have few samples. Diffusion models offer a more stable alternative, and their rapid growth in computer vision and in medical imaging is documented in [5] and [6]. Later work made sampling much faster [7] and improved the training objective [8], while classifier guidance [9] and classifier-free guidance [10] made it possible to control which class is generated. Latent diffusion [11] reduced the cost of producing high-resolution images, and diffusion models have also been shown to work well as a general-purpose data augmentation tool [12].

Several recent studies adapt diffusion models directly to imbalanced data. Class-Balancing Diffusion Models [13] adjust the conditional transfer probability during sampling so that the generator itself is less biased toward the larger classes. Bias-aware prior adjusting [14] instead gives each class its own noise-sampling prior during training, to counter the accumulation effect that favours head classes. The Anisotropic Diffusion Probabilistic Model [15] sets the forward-process noise level per class using error-analysis theory and reports gains on several imbalanced image benchmarks. Overlap optimization [16] reduces the feature-space overlap between large and small classes during training.

In medical image classification, the Iterative Online Image Synthesis framework [17] combines classifier-guided generation with an accuracy-adaptive sampling module: the share of synthetic images given to each class is proportional to one minus that class's accuracy, recalculated after every training epoch. Other studies combine diffusion generation with multi-scale feature fusion [18], prompt-driven joint image and mask synthesis [19], and detail-preserving augmentation for skin images [20].

In almost all of this work, the criteria that decide how many synthetic samples each class receives are taken for granted. They are usually based either on a fixed class frequency or on the

accuracy from a single epoch. Neither option accounts for how unstable such an accuracy estimate is when a class has only a few dozen samples, and neither uses information about which classes the errors go to. This gap is the starting point of the present study.

Proposed method. This paper proposes a method for balancing imbalanced image datasets that combines a class-conditional diffusion model with improved adaptive class-balancing criteria. The central idea is that the effectiveness of generative balancing depends not only on the quality of the synthetic images, but also on the rule that decides how the limited number of generated images is divided between the classes. The proposed method consists of three components: class-conditional synthetic image generation, improved criteria for estimating how much each class needs additional data, and the distribution of the generated images according to these criteria. The second component is the main contribution of this work, because it replaces a noisy single-epoch accuracy signal with a stable estimate that also takes into account how the errors of a class are distributed across the other classes.

Synthetic image generation. Synthetic images are produced by a class-conditional denoising diffusion model trained once on the training part of the dataset. The network is a U-Net with three resolution levels (64, 32 and 16 pixels) with 64, 128 and 128 channels, two residual blocks at each level and a self-attention block at the lowest resolution, which gives approximately 5.1 million parameters. The timestep is encoded with a sinusoidal embedding, and the class label is encoded with a learned embedding that is added to the timestep embedding and passed to every residual block. The model is trained to predict the noise added to an image, with the mean squared error between the predicted and the true noise as the loss function and a cosine noise schedule over 1 000 steps. During training the class label is replaced with an empty label with a probability of 10%, so that the same network learns both a conditional and an unconditional prediction. At sampling time, an image of the chosen class is generated with 30 deterministic DDIM steps, and classifier-free guidance with a scale of 2.0 combines the two predictions to strengthen the correspondence to the target class; no external classifier is used during generation. Training images are downsampled to 64×64 pixels with random horizontal flipping, and the model is optimised with AdamW at a learning rate of $2 \cdot 10^{-4}$ and a batch size of 128 for 60 epochs. The generative part is kept identical in all configurations considered below, so that any change in the final result can be attributed to the balancing criteria and not to the generator. Improved adaptive class-balancing criteria.

The classifier is evaluated on a validation subset at fixed evaluation points t .

For a problem with K classes, the accuracy of class k at such a point is:

$$a_k(t) = \frac{TP_k(t)}{n_k}, \quad (1)$$

where $TP_k(t)$ is the number of correctly recognised validation samples of class k and n_k is the total number of validation samples of that class. A widely used adaptive weight is obtained directly from this value:

$$w_k^{AAS}(t) = 1 - a_k(t), \quad (2)$$

so that a class recognised worse receives a larger share of the generated images. For classes represented by only a few dozen samples this weight is unreliable, because the accuracy is computed on very few samples and changes strongly between neighbouring evaluation points.

To remove this instability, the raw accuracy is replaced by an exponentially smoothed estimate:

$$s_k(t) = \gamma s_k(t-1) + (1 - \gamma)a_k(t), \quad (3)$$

where $\gamma \in (0,1)$ is a memory coefficient and the estimate is initialised as $s_k(0) = a_k(0)$. Larger values of γ give a slower and more stable signal, smaller values follow the recent behaviour of the classifier more closely but keep more of its noise.

The second improvement uses information that accuracy alone does not contain, namely the direction of the errors. Let $C_{kj}(t)$ be the number of validation samples of class k that are assigned to class j .

The total number of errors of the class and the largest number of errors falling on one single other class are:

$$c_k^{\text{tot}}(t) = \sum_{j \neq k} C_{kj}(t), \quad (4)$$

$$c_k^{\text{max}}(t) = \max_{j \neq k} C_{kj}(t), \quad (5)$$

and their ratio defines the concentration of the errors:

$$r_k(t) = \frac{c_k^{\text{max}}(t)}{c_k^{\text{tot}}(t)}. \quad (6)$$

A value of $r_k(t)$ close to 1 means that the class is confused almost only with one similar class, which is usually a local problem of two neighbouring categories. A low value means that the errors are spread over many classes, which indicates a weaker representation of the class as a whole and a stronger need for additional data.

The two parts are combined into the final criteria:

$$w_k(t) = (1 - s_k(t)) (1 + \lambda(1 - r_k(t))), \quad (7)$$

where $\lambda \geq 0$ controls the influence of the second term. With $\lambda = 0$ the criteria reduce to a smoothed accuracy rule, and if in addition the smoothing is disabled the expression becomes identical to (2), which allows the schemes to be compared under otherwise equal conditions.

Distribution of synthetic images between classes.

The weights are normalised over all classes, and a fixed total number of synthetic images N is distributed proportionally to them:

$$N_k(t) = \text{round} \left(N \cdot \frac{w_k(t)}{\sum_{j=1}^K w_j(t)} \right), \quad (8)$$

where rounding is applied to obtain integer counts, so that:

$$\sum_{k=1}^K N_k(t) = N. \quad (9)$$

The generated images are added to the training set, the classifier is trained further on the updated data, and at the next evaluation point the criteria are recomputed from the new accuracy and confusion values.

Algorithm of the proposed method. The proposed approach is implemented as a cyclic multi-stage process, in which generation and training are not separated but alternate with each other. The algorithm is built so that the decision about the amount of additional data for every class is made from the current behaviour of the classifier, and not fixed in advance from the initial class sizes.

At the first stage, the initial dataset is analysed and divided into a training, a validation and a test part, and the class-conditional diffusion model is trained on the training part. In parallel, a classifier is trained on the same original data; it provides the accuracy and confusion values from which the criteria are computed and does not take part in the generation.

At the second stage, the classifier is evaluated on the validation part. For every class, the accuracy from (1) and the confusion counts are obtained, the smoothed estimate (3) is updated, and the concentration of the errors (6) is calculated. At the first evaluation point the smoothed value is equal to the measured accuracy, so the criteria start from the real state of the model.

At the third stage, the weights (7) are computed for all classes and normalised, and the number of images to be generated for each class is obtained from (8). Classes that are recognised poorly and, at the same time, are confused with many different classes receive the largest share, while a class that is confused mostly with one neighbouring class receives less, because for it the problem is more likely to be the similarity of two categories than the lack of data.

At the fourth stage, the diffusion model generates the required number of images for each class under the guidance of the classifier, and the generated images are added to the training set. The

classifier is then trained further on this extended set until the next evaluation point, after which the process returns to the second stage.

The cycle is repeated until all N synthetic images are generated or the training limit is reached, after which the final model is evaluated on the test part, which is not used at any earlier stage. Since the diffusion model, the total number of synthetic images and the interval between evaluation points are the same in all compared configurations, the difference between the results reflects only the balancing criteria.

Experimental setup. All experiments were carried out on the HAM10000 dermatoscopic dataset, which contains 10 015 images distributed over seven diagnostic classes. The distribution is strongly skewed: the largest class (melanocytic nevi) holds 6 705 images, while the smallest one (dermatofibroma) holds only 115, giving an imbalance ratio of roughly 58:1.

The dataset was divided into training, validation and test parts in a 70/10/20 proportion using sampling stratified by class, which yielded 7 010, 1 001 and 2 004 images respectively. The split was performed at the level of individual images, following the protocol of the reference framework [17], so that the class composition of the test part closely matches the one used there and the results remain comparable with the published values. The same split, fixed by a single random seed, was used in every configuration. HAM10000 contains several images of the same skin lesion, identified by the lesion_id field, while patient identifiers are not provided; with an image-level split, images of the same lesion may fall into different parts, which can make the absolute values of the metrics somewhat optimistic. This effect, however, concerns all configurations in the same way, since they were trained and evaluated on exactly the same partition, and it is not expected to favour one allocation criterion over another. An evaluation with a lesion-level split is left for further research. All images were resized to 256×256 pixels and cached once before training; the classifier receives 224×224 crops.

ResNet-50 initialised with ImageNet weights was used as the classifier in every configuration. Training was performed with SGD (momentum 0.9, weight decay 10^{-4}) at an initial learning rate of 0.0125, with a batch size of 32 and mixed-precision arithmetic. The learning rate was reduced by a factor of ten after 30%, 60% and 90% of the training epochs. Geometric augmentation only was applied: random resized cropping, horizontal and vertical flipping, and rotation within $\pm 30^\circ$. Each configuration was trained for 60 epochs, and the checkpoint with the highest macro F1 on the validation part was selected for the final evaluation on the test part.

Generative model. A compact class-conditional denoising diffusion model was trained to produce the synthetic images. The network is a U-Net with approximately 5.1 million parameters operating at a resolution of 64×64, with a cosine noise schedule over 1 000 steps. Class conditioning was implemented through an embedding added to the timestep embedding, and the model was trained with 10% label dropout so that classifier-free guidance could be applied at sampling time. Optimisation used AdamW at a learning rate of $2 \cdot 10^{-4}$ with a batch size of 128.

The diffusion model was trained once, on the training part only, and the same trained model was used in every configuration that involves synthesis. Sampling was performed with 30 DDIM steps and a guidance scale of 2.0. Rather than generating images repeatedly during classifier training, a single pool of 900 images per class was produced in advance and shared by all configurations. This design choice is deliberate: because the generator, the sampling procedure and the available synthetic images are identical across configurations, any observed difference in the final result can be attributed to the allocation criteria alone, and not to variation in the generative process.

The synthetic images serve only as additional training data: all evaluation is carried out exclusively on real test images, so the usefulness of the synthetic data is assessed directly through the classification quality on real data. The study does not claim that the generated images are clinically realistic; at a resolution of 64×64 pixels the generator does not reproduce fine dermatoscopic structures, and some generated samples deviate in colour from real dermatoscopic images. Since the same pool of images is used in both synthesis configurations, these limitations affect them in the same way and do not change the comparison between the allocation criteria. A

detailed evaluation of the generator, including quality and diversity metrics, per-class analysis and a check for memorised training images, is left for further research.

Configurations. Three configurations were compared. The first is a baseline in which the classifier is trained on the original data with cross-entropy loss and no balancing of any kind. The second applies the accuracy-adaptive criterion, in which the weight of a class is taken directly as one minus its current accuracy. The third applies the criteria proposed in this work, combining the exponentially smoothed accuracy estimate of (3) with the error-concentration term of (6), using $\gamma = 0.75$ and $\lambda = 0.5$. The set of configurations follows from the research question: the contribution of this work is a criterion for allocating synthetic images, so it is compared with the existing allocation criterion under identical conditions – the same generator, the same pool and the same number of synthetic images. Balancing methods of other types, such as loss re-weighting or resampling, do not allocate synthetic data and can be combined with any allocation criterion; results for cross-entropy, focal and class-balanced losses, resampling and GAN-based synthesis under the same data protocol are reported in the reference work [17]. Classical geometric augmentation is applied in all three configurations. Since switching off both components of the proposed criteria yields exactly the accuracy-based criterion, the comparison between the second and third configurations shows the joint contribution of the two components; a separate evaluation of each component is left for further research.

The classifier learns from the original data alone during the first ten epochs, after which the first allocation is computed and synthetic images enter the training set. The short drop at that point is followed by an immediate recovery, and from the twentieth epoch onwards the validation curve settles into a narrow band, with the best value of 0.8111 reached at epoch 35. The validation curve is shown in the Fig. 1.

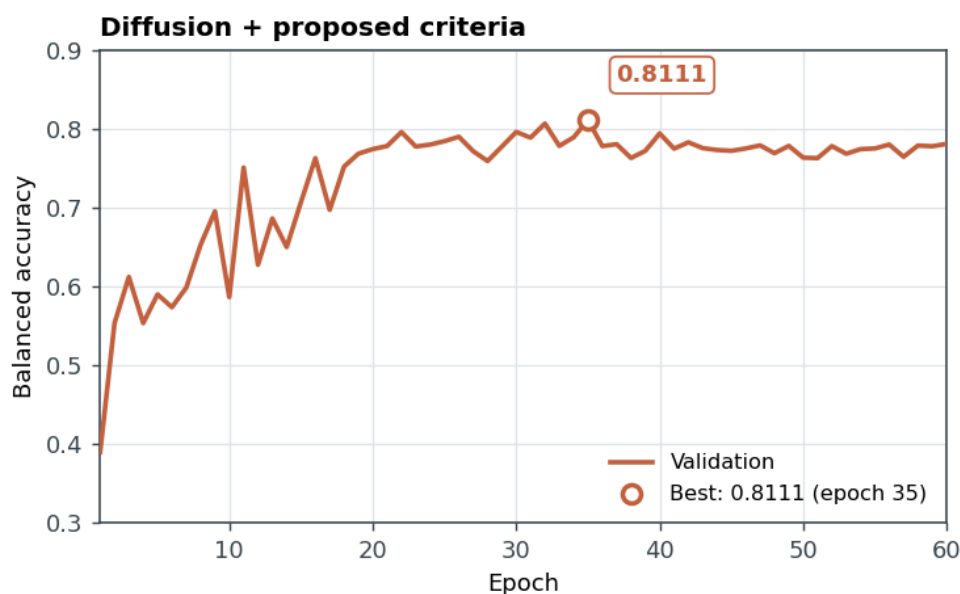


Fig. 1. Graph of validation curve

Source: compiled by authors

Quantitative results. The proposed criteria outperform the accuracy-based criterion on every metric considered: balanced accuracy rises from 0.7810 to 0.7886, macro F1 from 0.8085 to 0.8182, MCC from 0.7764 to 0.7846, and overall accuracy from 0.8857 to 0.8902. The comparison was carried out as a paired experiment: all configurations used the same data split, the same random seed, the same generator and the same pool of synthetic images, so that differences caused by the data split, the initialisation of the classifier and the generative model are excluded, and the configurations differ only in the allocation criteria. The improvement is consistent across all four metrics and is most visible on the minority classes df and bcc, which is the behaviour expected from

a better allocation of synthetic data. A formal assessment of statistical significance over repeated runs with different random seeds is planned as further research.

The second observation is just as important. The accuracy-based criterion, applied to the same generator and the same pool of synthetic images, does not improve on the baseline: its macro F1 (0.8085) and MCC (0.7764) are in fact slightly below the values obtained without any balancing at all (0.8120 and 0.7776). This means that the gain reported here does not come from adding synthetic data as such – it comes from the way that data is distributed among the classes. The generative model, the sampling procedure, the synthetic pool, the total number of injected images, the update interval, the training schedule and the random seed were the same in both synthesis configurations, so the difference between them comes from the allocation criteria alone.

Stability of the allocation. The clearest confirmation of the hypothesis stated in the problem formulation is found in the allocations themselves. Measured as the standard deviation of the per-class allocation across the five update points, the proposed criteria reduce the variability of the allocation by 63% on average compared with the accuracy-based criterion. The effect is strongest exactly where it was expected to matter most, on the smallest classes: for the vascular lesion class the variability falls by 95% (from 40.4 to 1.9 images), and for dermatofibroma by 74% (from 47.3 to 12.4 images). Under the accuracy-based criterion the allocation for the vascular class moved between 197, 92, 93, 105 and 97 images at successive update points, while under the proposed criteria it stayed within a range of five images over the whole training run. The instability described in the problem formulation is therefore not only a theoretical concern but a measurable property of the accuracy-based rule, and the exponential smoothing of the accuracy estimate removes it.

Comparison with the reference framework. The accuracy-adaptive sampling module improved in this work was introduced as part of the IOIS framework [17], which reports results on the same dataset and with the same three metrics. For its cross-entropy baseline that work reports a macro F1 of 0.7778, a balanced accuracy of 0.7678 and an MCC of 0.7611, and for its best configuration, combining online image synthesis with accuracy-adaptive sampling, 0.8197, 0.8150 and 0.8064 respectively.

Several points of comparison stand out. First, the baseline obtained in the present study is much stronger: a macro F1 of 0.8120 against 0.7778, that is, 3.4 percentage points higher, and an MCC of 0.7776 against 0.7611. A stronger starting point leaves less room for improvement, so any gain above it is achieved under harder conditions.

Second, the macro F1 of 0.8182 reached by the proposed criteria comes within 0.15 percentage points of the best value reported for the reference framework, and it is obtained with far fewer computational resources. The classifier here was trained for 60 epochs against 200 in the reference work, a schedule shorter by a factor of 3.3. The generative model is a compact network of 5.1 million parameters working at a resolution of 64×64, and it was trained once and reused; the synthetic images were produced in advance as a single shared pool instead of being regenerated during training, which removes the repeated sampling cost of online synthesis. The remaining difference in balanced accuracy and MCC should be read against this much shorter schedule and smaller generator.

Third, and most relevant to the contribution claimed here, the reference framework does not examine how stable its allocation is. Its sampling module recomputes the class distribution from single-epoch accuracy after every epoch, and the effect of doing so on classes represented by a few dozen samples is not measured. The present study measures this property directly and shows that it can be improved by a wide margin without changing the generative component at all.

The overall performance of the three configurations on the test part is summarised below, with the best value of each metric highlighted. The proposed criteria lead on all four measures, while the accuracy-based criterion stays at the level of the baseline. Following metrics are shown in the Fig. 2.

Metric	Baseline	AAS	Proposed
Accuracy	0.8872	0.8857	0.8902
Balanced accuracy	0.7785	0.7810	0.7886
Macro F1	0.8120	0.8085	0.8182
MCC	0.7776	0.7764	0.7846

Fig. 2. Overall performance of the three configurations on the test part

Source: compiled by authors

The row-normalised confusion matrices of the two synthesis configurations show where the difference between the criteria actually appears. The proposed criteria improve recognition of the two smallest classes – dermatofibroma rises from 0.59 to 0.68 and basal cell carcinoma from 0.83 to 0.87 – while the recall of the dominant class is left unchanged. The confusion matrix is shown in the Fig. 3.

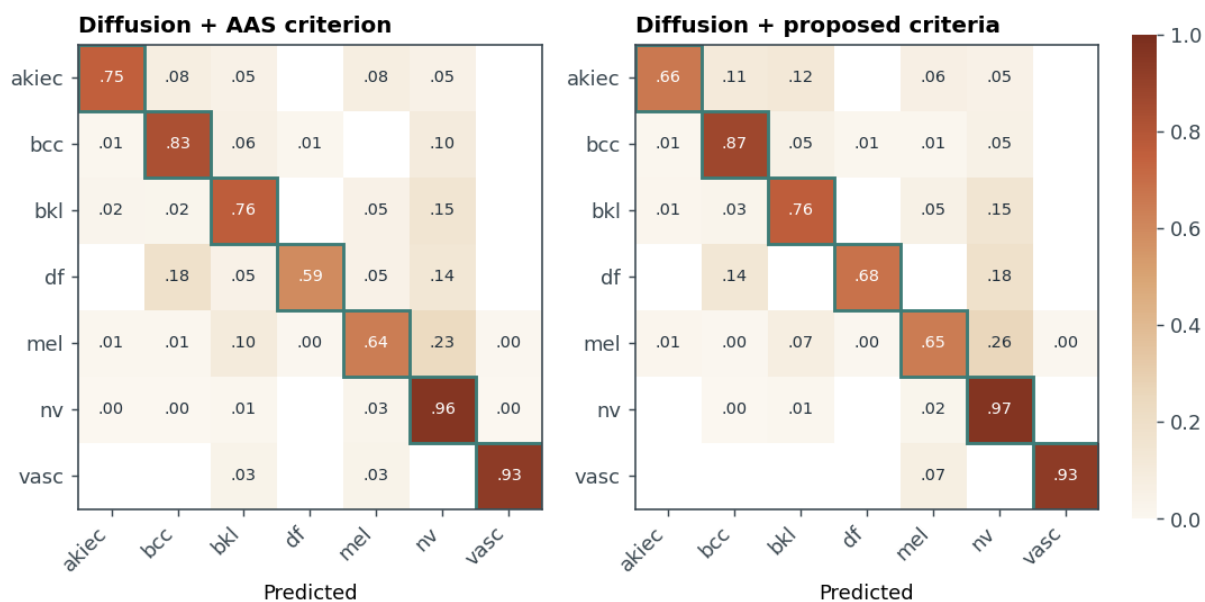


Fig. 3. Confusion matrices for the accuracy adaptive sampling method and the proposed criteria

Source: compiled by authors

Summary of the results. The experiments show that the way a fixed number of synthetic images is distributed among classes matters more than the fact of adding them. With the generator, the synthetic pool and every training setting held constant, replacing the accuracy-based criterion with the proposed one raised balanced accuracy by 0.76 percentage points, macro F1 by 0.97 and MCC by 0.82, while the accuracy-based criterion by itself stayed at the level of the unbalanced baseline.

The improvement is concentrated on the smallest classes, which is where a balancing method is expected to act. Alongside this, the proposed criteria made the allocation itself far steadier, reducing its variability across update points by 63 % on average and by up to 95% for the vascular lesion class, which confirms directly the weakness of single-epoch accuracy identified in the problem formulation. These results were obtained with a training schedule 3.3 times shorter and a considerably smaller generative model than in the reference framework, which suggests that careful design of the allocation criteria can compensate for a substantial part of the computational budget usually spent on generation.

Conclusions. This paper addressed the problem of classifying images under severe class imbalance, focusing on a component of generative balancing pipelines that has received little attention: the criteria that decide how a limited number of synthetic images should be distributed

among classes. The generative side of such pipelines has advanced considerably in recent years, from adversarial networks to diffusion models, and the quality of the images that can now be synthesised is far higher than it was only a few years ago. The rule that decides how many of those images each class should receive, however, has remained largely unchanged. Existing approaches derive this decision from the classifier's accuracy on a single epoch, which is unreliable for classes represented by a few dozen samples and carries no information about which classes the errors are directed to. Both shortcomings act most strongly on exactly the categories that a balancing method is supposed to help.

The main contribution of this work consists of two interrelated modifications of these criteria. The first is an exponentially smoothed estimate of per-class accuracy, which retains the responsiveness of the adaptive approach to genuine, persistent weakness of a class while suppressing the fluctuations that arise when accuracy is measured on very few samples. The memory coefficient of this estimate controls the trade-off between the two: larger values produce a slower and steadier signal, smaller values follow recent behaviour more closely but retain more of its noise. The second modification is an error-concentration term derived from the confusion matrix, which distinguishes a class whose errors are directed almost entirely at one neighbouring category from a class whose errors are spread across many, and adjusts its share of synthetic images accordingly. The distinction matters because the two situations have different causes: errors concentrated on a single neighbour usually indicate visual similarity between two categories, whereas errors spread broadly point to a weaker representation of the class as a whole. Both terms are combined into a single allocation weight, and the resulting criteria reduce exactly to the accuracy-based rule when the smoothing is disabled and the confusion term is set to zero, which allows the two schemes to be compared under fully identical conditions. The experimental design was arranged so that the criteria could be isolated from every other factor. The diffusion model was trained once and reused, the synthetic images were produced in advance as a single pool shared by all configurations, and the total number of injected images, the update interval, the training schedule and the random seed were kept identical. Any difference between the configurations therefore reflects the allocation criteria and nothing else.

Experimental evaluation on the HAM10000 dataset confirmed the effectiveness of the proposed approach. The proposed criteria improved balanced accuracy from 0.7810 to 0.7886, macro F1 from 0.8085 to 0.8182 and MCC from 0.7764 to 0.7846 relative to the accuracy-based criterion, with all four metrics moving in the same direction. The improvement is concentrated on the smallest classes, while the recall of the dominant class remains unchanged, which is the behaviour expected of a balancing method rather than of a method that simply shifts the decision threshold. Notably, the accuracy-based criterion applied to the same generator did not improve on the unbalanced baseline at all, and on two of the four metrics fell marginally below it. This shows that the benefit reported here comes from the distribution of synthetic data and not from its addition, and it suggests that a poorly chosen allocation can cancel out the advantage that synthesis is expected to bring.

Conflicts of Interest: The authors declare no conflict of interest, financial, personal, authorship or any other, that could influence the research and its results presented in this article

Financing: The research was conducted without financial support

Data availability: The study uses the publicly available HAM10000 dataset (<https://doi.org/10.7910/DVN/DBW86T>). The code, configuration files with all hyperparameters, the per-class numbers of synthetic images at every update point and the training logs are available from the corresponding author upon reasonable request

Use of Artificial Intelligence (AI): The authors confirm that they did not use AI tools for writing this paper

REFERENCES

1. Zhang, Y., Kang, B., Hooi, B., Yan, S. & Feng, J. “Deep long-tailed learning: A survey”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2023; 45 (9): 10795–10816. DOI: <https://doi.org/10.1109/TPAMI.2023.3268118>.

2. Saini, M. & Susan, S. “Tackling class imbalance in computer vision: A contemporary review”. *Artificial Intelligence Review*. 2023; 56: S1279–S1335. DOI: <https://doi.org/10.1007/s10462-023-10557-6>.
3. Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B. & Japkowicz, N. “The class imbalance problem in deep learning”. *Machine Learning*. 2024; 113(7): 4845–4901. DOI: <https://doi.org/10.1007/s10994-022-06268-8>.
4. Iglesias, G., Talavera, E. & Díaz-Álvarez, A. “A survey on GANs for computer vision: Recent research, analysis and taxonomy”. *Computer Science Review*. 2023; 48: 100553. DOI: <https://doi.org/10.1016/j.cosrev.2023.100553>.
5. Croitoru, F.-A., Hondru, V., Ionescu, R. T. & Shah, M. “Diffusion models in vision: A survey”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2023; 45 (9): 10850–10869. DOI: <https://doi.org/10.1109/TPAMI.2023.3261988>.
6. Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacıhaliloglu, I. & Merhof, D. “Diffusion models in medical imaging: A comprehensive survey”. *Medical Image Analysis*. 2023; 88: 102846. DOI: <https://doi.org/10.1016/j.media.2023.102846>.
7. Song, J., Meng, C. & Ermon, S. “Denoising diffusion implicit models”. *Proceedings of ICLR*. 2021. DOI: <https://doi.org/10.48550/arXiv.2010.02502>.
8. Nichol, A. & Dhariwal, P. “Improved denoising diffusion probabilistic models”. *Proceedings of ICML*. 2021. p. 8162–8171. DOI: <https://doi.org/10.48550/arXiv.2102.09672>.
9. Dhariwal, P. & Nichol, A. “Diffusion models beat GANs on image synthesis”. *NeurIPS*. 2021; 34: 8780–8794. DOI: <https://doi.org/10.48550/arXiv.2105.05233>.
10. Ho, J. & Salimans, T. “Classifier-free diffusion guidance”. *arXiv*. 2022. DOI: <https://doi.org/10.48550/arXiv.2207.12598>.
11. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. “High-resolution image synthesis with latent diffusion models”. *Proceedings of CVPR*. 2022: 10684–10695. DOI: <https://doi.org/10.1109/CVPR52688.2022.01042>.
12. Trabucco, B., Doherty, K., Gurinas, M. & Salakhutdinov, R. “effective data augmentation with diffusion models”. *Proceedings of ICLR*. 2024. DOI: <https://doi.org/10.48550/arXiv.2302.07944>.
13. Qin, Y., Zheng, H., Yao, J., Zhou, M. & Zhang, Y. “Class-balancing diffusion models”. *Proceedings of CVPR*. 2023: 18434–18443. DOI: <https://doi.org/10.1109/CVPR52729.2023.01768>.
14. Xu, C., Yan, J., Yang, M. & Deng, C. “Rethinking noise sampling in class-imbalanced diffusion models”. *IEEE Transactions on Image Processing*. 2024; 33: 6298–6308. DOI: <https://doi.org/10.1109/TIP.2024.3485484>.
15. Kong, J., Guo, Y., Wang, Y. & Duan, Y. “Anisotropic diffusion probabilistic model for imbalanced image classification”. *arXiv*. 2024. DOI: <https://doi.org/10.48550/arXiv.2409.14313>.
16. Yan, D., Qi, L., Hu, V. T., Yang, M.-H. & Tang, M. “Training class-imbalanced diffusion model via overlap optimization”. *arXiv*. 2024. DOI: <https://doi.org/10.48550/arXiv.2402.10821>.
17. Li, S., Lin, Y., Chen, H. & Cheng, K.-T. “Iterative online image synthesis via diffusion model for imbalanced classification”. *arXiv*. 2024. DOI: <https://doi.org/10.48550/arXiv.2403.08407>.
18. Zhu, Z., Liu, Y., Yuan, C.-A., Qin, X. & Yang, F. “A diffusion model multi-scale feature fusion network for imbalanced medical image classification research”. *Computer Methods and Programs in Biomedicine*. 2024; 256: 108384. DOI: <https://doi.org/10.1016/j.cmpb.2024.108384>.
19. Xu, Z. “SkinDualGen: Prompt-driven diffusion for simultaneous image-mask generation in skin lesions”. *arXiv*. 2025. DOI: <https://doi.org/10.48550/arXiv.2507.19970>.
20. Kim, M., Yoo, J., Kwon, S., et al. “Diffusion-based skin disease data augmentation with fine-grained detail preservation and interpolation for data diversity”. *PLOS ONE*. 2025. DOI: <https://doi.org/10.1371/journal.pone.0331404>.

Received 02.09.2026

Received after revision 04.10.2026

Accepted 12.10.2026

DOI: <https://doi.org/10.15276/ict.03.2026.53>

УДК 004.93'1:004.032.26

Застосування дифузійної моделі з удосконаленими критеріями адаптивного балансування класів для незбалансованих наборів даних

Міхалев Кирило Богданович¹⁾

ORCID: <https://orcid.org/0009-0007-8058-4569>; kirill.mikhalev.9035@gmail.com. Scopus Author ID: 59390689500

Ніколенко Анатолій Олександрович¹⁾

ORCID: <https://orcid.org/0000-0002-9849-1797>; anatolyn@ukr.net. Scopus Author ID: 57197491335

¹⁾ Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

АНОТАЦІЯ

У статті розглянуто проблему незбалансованості класів, яка залишається одним із ключових обмежень під час побудови систем розпізнавання зображень. Незбалансованість, за якої одні категорії представлені тисячами зразків, а інші містять лише кілька десятків, призводить до систематичного зміщення моделей у бік домінуючих класів і до низької якості розпізнавання рідкісних категорій, що є неприйнятним у клінічному контексті. Генеративні моделі стали поширеним інструментом розв'язання цієї проблеми, і дифузійні моделі зокрема нині забезпечують стабільнішу та якіснішу альтернативу генеративно-змагальним мережам. Проте одному з компонентів таких конвексів приділено значно менше уваги, ніж власне генеративній частині, а саме критеріям, які визначають, як обмежену кількість синтетичних зображень слід розподілити між класами. Наявні підходи виводять це рішення з точності класифікатора, виміряної за одну епоху, що є ненадійним для класів, представлених кількома десятками зразків, і не містить інформації про те, до яких саме класів спрямовані помилки. У роботі запропоновано вдосконалені критерії адаптивного балансування класів, які усувають обидва недоліки. Першим компонентом є експоненційно згладжена оцінка точності за класами, яка зберігає чутливість до стійкої слабкості класу та водночас пригнічує коливання, спричинені вимірюванням на дуже малій кількості зразків. Другим компонентом є концентрація помилок, отримана з матриці невідповідностей, який відрізняє клас, помилки якого спрямовані майже винятково до однієї сусідньої категорії, від класу, помилки якого розподілені між багатьма, і відповідно коригує його частку згенерованих зображень. Запропоновані критерії оцінено на наборі даних, що містить понад десять тисяч дерматоскопічних зображень семи класів із коефіцієнтом незбалансованості п'ятдесят вісім до одного, з використанням класово-обумовленої дифузійної моделі як генератора та єдиного спільного пулу синтетичних зображень, що дозволило відокремити вплив критеріїв від усіх інших чинників. Результати демонструють, що запропоновані критерії підвищують збалансовану точність, макроусереднену міру F1 та коефіцієнт кореляції Метьюза порівняно з критерієм на основі точності, причому покращення зосереджене на найменших класах, а розпізнавання домінуючого класу залишається незмінним. Мінливість розподілу за класами між точками оновлення зменшено в середньому на шістьдесят три відсотки, а для класу судинних уражень на дев'яносто п'ять відсотків, що безпосередньо підтверджує нестабільність критеріїв на основі точності. Отримані результати показують, що спосіб розподілу синтетичних даних між класами впливає на результат більше, ніж сам факт їх додавання.

Ключові слова: незбалансованість класів; дифузійні моделі; адаптивні критерії балансування; генерація синтетичних даних; розпізнавання зображень

ІНФОРМАЦІЯ ПРО АВТОРІВ



Міхалев Кирило Богданович - аспірант кафедри Інформаційних систем. Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

Галузь досліджень: генеративні моделі, класифікація зображень, незбалансовані набори даних, розпізнавання зображень

Mikhalev Kyrylo Bogdanovych - postgraduate student, Department of Information Systems. Odesa Polytechnic National University. 1, Shevchenko Ave. Odesa, 65044, Ukraine

ORCID: <https://orcid.org/0009-0007-8058-4569>; kirill.mikhalev.9035@gmail.com. Scopus Author ID: 59390689500

Research field: generative models, image classification, imbalanced datasets, image recognition



Ніколенко Анатолій Олександрович - кандидат технічних наук, доцент, кафедра Інформаційних систем. Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

Галузь досліджень: інформаційні технології; штучний інтелект; системи підтримки прийняття рішень; машинне навчання; нейронні мережі

Nikolenko Anatolii Oleksandrovych - Candidate of Engineering Sciences, Associate Professor, Department of Information Systems. Odesa Polytechnic National University. 1, Shevchenko Ave. Odesa, 65044, Ukraine

ORCID: <https://orcid.org/0000-0002-9849-1797>; anatolyn@ukr.net. Scopus Author ID: 57197491335

Research field: information technology; artificial intelligence; decision support systems; machine learning; neural networks