# Improving Balanced Accuracy for Minority Plant Species under Data Imbalance

## Mejorando la exactitud balanceada para especies de plantas minoritarias con datos desbalanceados

Ruben Gonzalez-Villanueva[1], Jose Carranza-Rojas[2]

Gonzalez-Villanueva, R; Carranza-Rojas, J. Improving balanced accuracy for minority plant species under data imbalance. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 22-27.

https://doi.org/10.18845/tm.v37i7.7293

[1] Costa Rica Institute of Technology. Costa Rica.
rgonzalezv@estudiantec.cr
https://orcid.org/0000-0001-8044-3474

[2] Costa Rica Institute of Technology. Costa Rica.
jcarranza@itcr.ac.cr
https://orcid.org/0000-0002-9177-9173

## Keywords

Imbalanced datasets;long-tail distribution;automatic plant identification;balanced metrics;deep learning; minority classes;classification.

## Abstract

Regardless of the widely known success of deep learning in classification, such models are commonly measured by metrics that do not account for data imbalance, especially in terms of predictions per class, ignoring minority classes. This can be a problem, as minority classes are often the most difficult to predict and collect data for. In the plant domain, for example, species with fewer samples are often the ones that are hardest to collect and predict in the field. As we continue to identify more and more plant species, more of them become minority species, making it increasingly difficult to accurately classify them using traditional machine learning methods. To address this issue, we explore the combination of traditional data and machine learning approaches with deep learning techniques such as self-supervision in a preprocessing stage. By using self-supervised training together with different sampling algorithms and class weights, we were able to improve the balanced accuracy metric for minority plant species by between 7.9% and 13% without affecting general accuracy. This shows that using deep learning techniques in combination with traditional machine learning methods can help to improve the accuracy of predictions for minority classes, even in domains where data is limited.

## Palabras clave

Conjuntos de datos desbalanceados; distribución de cola larga; identificación automática de plantas; métricas balanceadas; aprendizaje profundo; clases minoritarias; clasificación.

## Resumen

A pesar del ampliamente conocido éxito del aprendizaje profundo en tareas de clasificación, estos modelos se miden comúnmente con métricas que no tienen en cuenta el desbalance de datos, especialmente en términos de predicciones por clase, ignorando las clases minoritarias. Esto puede ser un problema, ya que las clases minoritarias suelen ser las más difíciles de predecir y en términos de recolección de datos. En el dominio de las plantas, por ejemplo, las especies con un menor número de muestras son a menudo las más difíciles de recolectar y predecir en el campo. A medida que se siguen identificando más y más especies de plantas, más de ellas se vuelven minoritarias, lo que dificulta cada vez más la clasificación precisa utilizando métodos tradicionales de aprendizaje automático. Para abordar este problema, se explora la combinación de enfoques de los datos y tradicionales de aprendizaje automático con técnicas de aprendizaje profundo, como la auto-supervisión en una etapa de preprocesamiento. Al utilizar el entrenamiento auto supervisado junto con diferentes algoritmos de muestreo y pesos de clase, logramos mejorar la métrica de exactitud balanceada para las especies de plantas minoritarias entre el 7.9% y el 13% sin afectar la datos general. Esto demuestra que el uso de técnicas de aprendizaje profundo en combinación con métodos tradicionales de aprendizaje automático puede ayudar a mejorar la precisión de las predicciones para clases minoritarias, incluso en dominios donde los datos son limitados.

## Introduction

Imbalanced datasets affect models to classify species more fairly [1]. Training a classifier with a long-tailed distribution may achieve good results on the general accuracy metric, but not on a metric that considers the predictive ability of each class, even minority ones, as the balanced accuracy metric does [2].

Different approaches have been tried to solve this problem and find a configuration to more fairly classify an imbalanced dataset, but these approaches only use traditional ML methods [3],[4]. In this short paper we use these ML approaches, combine them with DL approaches such as self-supervision pre-training, to get better results on the balanced accuracy metric for minority species without decreasing overall accuracy in a plants dataset.

## Matherials and methods

### Materials

*Dataset*

The CRLeaves dataset contains images of leaves from Costa Rica from 255 species. It was taken from [5]. It contains a total of 6,938 images but has a long-tail distribution in which the species with the most images contain 89, while the species with the fewer images contains only 4. All images have an uniform background as shown in the figure 1.

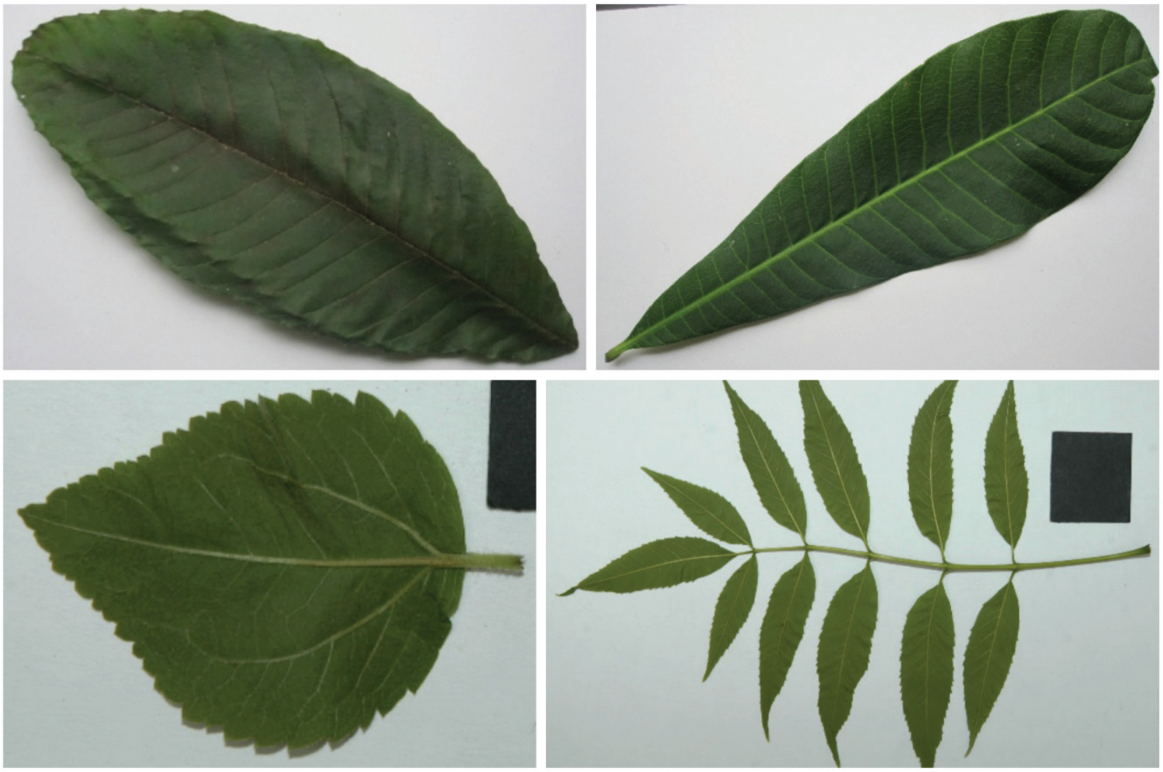

**Figure 1.** Samples from CR Leaves dataset of Clethra costaricensis, Anacardium excelsum, Calea urticifolia and Tecoma Stans. Source: [5]

*Architecture*

The chosen deep learning architecture is ResNet50 [6]. It contains 50 deep residuals layers and keeps good results in computer vision for our dataset. This is the baseline for the experiments.

### *Metrics*

We used the general accuracy that calculates the hit rate of the predictions and labels for all the samples [2]. Additionally, we focus on balanced accuracy that uses weights in the general accuracy to measure the predictions by class [7]. Balanced accuracy for minority species is calculated only for the species with less than 10 samples.

## Methods

### *Pre-Training Methods*

The chosen pre-training method is SimCLR [8]. It is a self-supervision approach of encoder-decoder used to pre-train the model with no labels. SimCLR contains two parts: training method and the architecture change that consists in a new fully connected (FC) layer at the end of the model. We test how both training methods work with the baseline architecture (ResNet50) and with the SimCLR FC.

### *Imbalanced Algorithms*

The chosen algorithms to address the imbalance problem are from two different kind of approaches. The first one is the data approach in which we used random undersampling and random oversampling to obtain a similar number of images per species [3]. We explored two implementations:

### Sampling A

We use the implementation of sklearn [9]. It adjusts the weights inversely proportional to class frequencies as shown in Equation 1, where  is the weight for class , *S* the number of samples and *C* the number of classes.

$$w_i = \frac{S}{C \ * \ S_i} \tag{1}$$

### Sampling B

It is described in Equation 2. It divides a 1 with the number of samples in the class to obtain the inverse frequency. We use NumPy’s implementation.

$$w_i = \frac{1}{S_i} \tag{2}$$

### Class Weights

We additionally use class weights, which are based on calculating the weights for each species and give this information to the cross-entropy loss [4].

## Results

Table 1 describes the factors for the experiments, which combined sum up to 12 experiment runs. We ran 3 repetitions for each combination for statistical validity. We ran the SimCLR self-supervision pre-training models for 100 epochs to use the weights in the experiments. For the supervised training, we used a distribution of 50:50 for training and testing.

**Table 1.** Factors for experiments.

| Pre-Training Methods | Imbalanced Algorithms |
|---|---|
| None | None |
| SimCLR Training | Sampling A |
| Sim CLR Training + FC | Sampling B |
| | Class Weights |

We report the P-Value from the pairwise T test, as well as the results for the metric General Accuracy in Table 2. The baseline model without any techniques for imbalance data obtained 0.87 of general accuracy. The only combination that was statistically different from the baseline was SimCLR Training + FC + class weights with an improvement. In the other combinations there was no statistical difference. This means the combinations of algorithms to address the data imbalance are not hurting drastically the general accuracy.

In contrast, the results for Balanced Accuracy Minority Species are also described in Table 2. The baseline obtained a value of 0.657, been this a difference of 0.213 compared to general accuracy. There were 3 combinations that were statistical better than the baseline: Sampling A, Sampling B and Class Weights, all combined with SimCLR Training.

**Table 2.** Average balanced accuracy for minority species, average general accuracy, and p-values for interactions of factors vs the baseline with the testing subset. ↑ denotes statistical improvement against the baseline.

| Pre-Training Method | Imbalanced Algorithm | Average Balanced Accuracy Minority Species / P-Value | Average General Accuracy / P-Value |
|---|---|---|---|
| None | None | 0.657 | 0.870 |
| | Sampling A | 0.713 / 0.390 (-) | 0.865 / 0.580 (-) |
| | Sampling B | 0.713 / 0.270 (-) | 0.867 / 0.640 (-) |
| | Class Weights | 0.722 / 0.180 (-) | 0.869 / 0.890 (-) |
| SimCLR Training | None | 0.694 / 0.270 (-) | 0.883 / 0.068 (-) |
| | Sampling A | 0.787 / 0.020 (↑) | 0.881 / 0.190 (-) |
| | Sampling B | 0.736 / 0.023 (↑) | 0.877 / 0.260 (-) |
| | Class Weights | 0.750 / 0.042 (↑) | 0.880 / 0.160 (-) |
| Sim CLR Training + FC | None | 0.690 / 0.500 (-) | 0.895 / 0.097 (-) |
| | Sampling A | 0.727 / 0.110 (-) | 0.885 / 0.11 (-) |
| | Sampling B | 0.755 / 0.120 (-) | 0.886 / 0.042 (↑) |
| | Class Weights | 0.713 / 0.210 (-) | 0.886 / 0.051 (-) |

## Conclusions and recommendations

The results show that there was only one experiment that had an improvement in general accuracy, but this combination did not improve the balanced accuracy for minority species. For the balanced accuracy for minority species metric, there was no improvement using pre-training methods or the imbalanced algorithms individually. We found three combinations that had an improvement on imbalanced accuracy, in particular when there was a combination of SimCLR Training as Pre-Training Method and with sampling methods and class weights. This shows that it is possible to improve imbalanced metrics without heavily affecting balanced metrics.

Finally, for this dataset the best way we found to improve the metric is to combine SimCLR Training and the different imbalanced algorithms used. Future work includes performing these experiments on a larger dataset with more species, images, and even more imbalance. In addition, finding completely new approaches to further improve the results obtained, which could be based on loss functions using the plant hierarchy.

## Acknowledgments

We would like to thank the Costa Rica Institute of Technology for its support during this research.

## References

[1] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, 2012.

[2] N. Bressler, “How to check the accuracy of your machine learning model,” Feb 2022. [Online]. Available: https://deepchecks.com/how-to- check-the-accuracy-of-your-machine-learning-model/

[3] Y. Pristyanto, I. Pratama, and A. F. Nugraha, “Data level approach for imbalanced class handling on educational data mining multiclass classification,” in 2018 International Conference on Information and Communications Technology (ICOIACT), 2018, pp. 310–314.

[4] S. Lu, F. Gao, C. Piao, and Y. Ma, “Dynamic weighted cross entropy for semantic segmentation with extremely imbalanced data,” in 2019 Interna- tional Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), 2019, pp. 230–233.

[5] J. Carranza-Rojas and E. Mata-Montero, “Combining leaf shape and texture for costa rican plant species identification,” CLEI Electronic journal, vol. 19, no. 1, pp. 7–7, 2016.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[7] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, “The balanced accuracy and its posterior distribution,” in 2010 20th International Conference on Pattern Recognition, 2010, pp. 3121–3124.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple frame- work for contrastive learning of visual representations,” in International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[9] G. King and L. Zeng, “Logistic regression in rare events data,” Political analysis, vol. 9, no. 2, pp. 137–163, 2001.