ISSN 2007-9737

Confidence Calibration of CNNs in Medical Image Databases

Nancy López-Miguel*, Raquel Diaz-Hernández, Leopoldo Altamirano-Robles

Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla,
Mexico

{robles, nancy.lopez, raqueld}@inaoep.mx

Abstract. Disease classification with convolutional neural networks (CNNs) has evolved significantly, providing practical tools to address this challenge in medical imaging. Notwithstanding these advances, there is a significant gap in evaluating the confidence of the results provided by the networks. Consequently, this paper proposes to perform confidence calibration of the predictions in these models. Some authors have proposed approaches to solve this problem. However, there is still a lack of evaluation of these confidence calibration methods in medical contexts. In this paper, two confidence calibration methods (Mixup and Temperature Scaling) are applied on three different medical image bases (MIDBs), in addition to evaluating a base case with the Geometric Shapes Dataset. The MIDBs analyzed are BCS-DBT, BreakHis, and lung disease. Our results demonstrate the importance of confidence calibration in medical image classification for the following reasons: 1) Model predictions in medical imaging are crucial to be reliable and backed by an accurate measure of their confidence. 2) Calibration methods help identify erroneous or unreliable predictions the model makes. 3) Implementing confidence calibration methods in the models decreases the overconfidence predictions and, in general, improves the predictions made by the model. In the three bases analyzed, the Mixup procedure and the one combined with Temperature Scaling have the best results, obtaining ECE values between 0.0037 and 0.0671 and Accuracy values of 77.91 as the lowest value and 97.52 as the highest.

Keywords. Confidence calibration, mixup, TS.

1 Introduction

Deep Learning disease detection and classification techniques have evolved, providing powerful tools to address this challenge.

However, despite the progress of these approaches in classification tasks, for example, there is a significant lack of confidence in the calibration of the predictions generated by convolutional neural networks. Confidence calibration is essential in all domains of deep learning and applying it in medical imaging is critical for reliable prediction.

There are different calibration methods; among the best known is the Mixup method proposed by Karimi and Gholipour [1] to reduce the overconfidence of Artificial Intelligence (AI) methods in erroneous predictions using different medical image bases.

Rao et al. [2] used this same method, plus CutMix and CutOut to understand the effects of modern data augmentations on the calibration of CNNs for medical image analysis and improve reliability.

Ge et al. [3] propose a modified Bootstrapping loss function with Mixup for model calibration in disease diagnosis.

Gao et al. [4] perform reliability analysis on CNN for confidence calibration in lung cancer images using methods such as Mixup, label smoothing, cross-entropy loss, and temperature scaling.

Another work that uses temperature scaling is Kurz et al. [5] for calibrating CNNs for histological classification of colorectal cancer. Others are not frequently implemented, such as weight scaling, proposed by Frenkel and Goldberger [6], similar to temperature scaling, but in this case, weight controls the confidence of the calibrated prediction, instead of temperature, among others.

It is worth mentioning that there are different calibration methods, and they can be classified into different categories, such as parametric and non-parametric methods [7], within the parametric methods we can find Temperature Scaling, Platt Scaling, and Beta Calibration, among others. In the

doi: 10.13053/CyS-29-1-5510

non-parametric methods, we can find Isotonic Regression and Histogram Binning.

In this work, we propose to use the temperature scaling method, Mixup, and the Temperature Scaling + Mixup method, which have been shown to obtain good results in calibrating convolutional neural networks using databases of geometric images, histological images of breast cancer and digital tomosynthesis of breast and lung X-ray images. It should be noted that the literature indicates that no work has performed confidence calibration on classifying breast cancer images.

This paper is structured as follows: Section 2 provides a comprehensive overview of the experiment, including the data sets used, the calibration methods, and the evaluation method. Section 3 details the methodology employed. Section 4 presents the results, followed by an analysis and discussion of the work performed. Finally, Section 5 offers a conclusion and outlines potential avenues for future research.

2 Experimental Details

2.1 Data Sets

Geometric Shapes Dataset. It is a three-class dataset, each class representing a type of geometric shape (triangle, square, and circle), and it consists of 10,000 images generated for each class. In the generation of this dataset, the perimeter, the position of each shape, the rotation angle, the background color, and the fill color of each image are randomly and independently selected [8].

BreakHis. This data set is a set of breast cancer histopathological images (BreakHis) obtained from 82 patients. It consists of 9,109 microscopic images of breast tumor tissue with different magnification factors. In addition, it contains 480 benign and 5,429 malignant specimens [9]; in this case, three types of malignant images were used.

BCS-DBT. This dataset consists of patients who underwent digital breast tomosynthesis (DBT) examination. DBT volumes were obtained from the Duke Health System from 5,060 patients [10], with 19,148 images labeled for training, of which 76 images are of malignant tumors. Due to the database's imbalance, data augmentation was performed for tumor images with different techniques such as rotation, displacement, inversion, brightness, and contrast, among others, obtaining 6,528 images.

Lung Disease. This data set is a set of lung X-ray images obtained from Kaggle [11] and collected from different hospitals, clinics, and healthcare institutions. There are 3,475 images in total, divided into three classes: normal (1,250 images), which are images of healthy lungs; Lung Opacity (1,125 images), which include images with varying degrees of lung abnormalities; and Viral Pneumonia (1,100 images) associated with cases of viral pneumonia**.**

2.2 Calibration Methods

The calibration method implemented in this work is temperature scaling, which is essential for confidence calibration in deep-learning prediction models. The technique helps to determine the predicted confidence scores for different classes, which helps assess the reliability of model predictions. The temperature scaling equation uses the Softmax function, which converts raw scores to probabilities and adds the parameter T [5]. It is denoted as:

$$f_i(y) = \frac{e^{y_i/_T}}{\sum_j e^{y_i/_T}}, \tag{1}$$

where T is the temperature parameter that controls the sharpness of the predicted probabilities, adjusting T allows the equation to maintain this sharpness, which can benefit different applications and domains.

The Mixup method was also used. This method combines values to calculate a new example based on the original x or y value and a weighted combination of other related x_j or $y_j$values [2]:

$$x = \lambda * x_i + (1-\lambda)\, x_j, \tag{2}$$

$$y = \lambda * y_i + (1-\lambda)\, y_j, \tag{3}$$

where λ represents a parameter ranging from 0 to 1. x_i represents a specific value or variable related to x, and $(1-\lambda)x_j$ represents a weighted value of x_j, where $(1-\lambda)$ determines the weight.

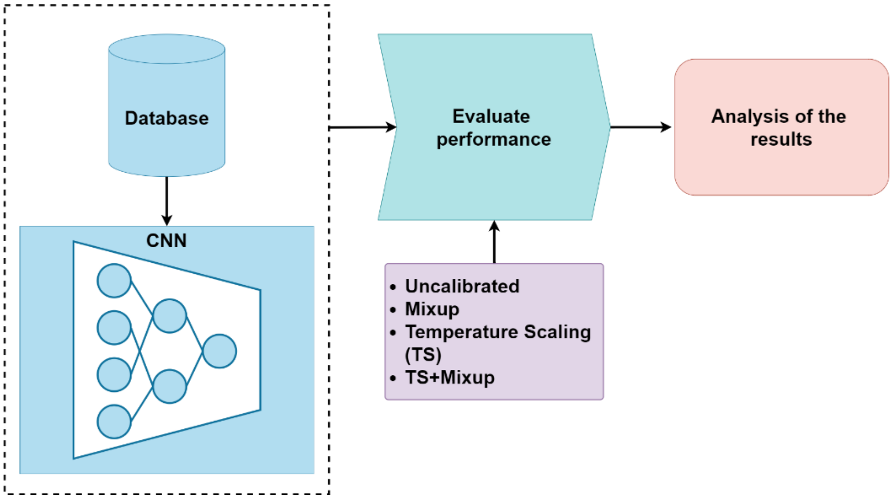

Fig. 1. General diagram of this research

The evaluation metrics are Accuracy and ECE (Expected Calibration Error).

The standard evaluation metric is Accuracy, which describes how the model performs in all classes, i.e., the measure reflects how reliable the model is in classifying samples as positive.

ECE is commonly used to measure the calibration performance of prediction models. The ECE is calculated in two steps: divide the prediction value space into equal-sized bins and calculate the weighted average of the difference between accuracy and confidence for each garbage can [1, 5, 6]. The equation is defined as follows:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc\ (B_m) - conf\ (B_m)|, \quad (4)$$

where n is the total number of samples in the set B_m, B_m is the set of predictions made by the model, $acc\ (B_m)$ is the accuracy of the predictions in the set, $conf\ (B_m)$ is the confidence, the fraction of correct predictions in the set.

3 Methodology

Three Convolutional Neural Networks (CNNs) — Resnet18, Resnet34, and Resnet50—were used for the experiments. Each CNN used the same parameters: a batch size of 32, a learning rate of 0.001, and 20 epochs. The Adam algorithm was used as the optimization method. The databases used were those explained in section 2.1. The methods discussed above are Mixup, Temperature Scaling (TS), and combined TS with Mixup.

As a base case, experiments were carried out with the data set of geometric figures with the three networks, uncalibrated, the calibration methods explained before and their combinations.

Likewise, experiments were performed with the three uncalibrated networks, and the calibration methods were applied using the different medical image bases previously mentioned, producing a total of 36 experiments.

4 Results, Analysis, and Discussion

4.1 Results

Base case: Resnet18, Resnet34, and Resnet50 with the Geometric Shapes Dataset were used in all uncalibrated (UC) experiments, using the calibration methods Mixup, TS, and combined TS with Mixup.

In this case, an Accuracy of 100 and an ECE of 0 were obtained. Due to the good results, we do not consider showing the corresponding tables and graphs necessary.

Table 1. Results of the Resnet18 experiment

	BreakHis		**BCS-DBT**		**Lung Disease**	
	Acc (↑)	ECE (↓)	Acc (↑)	ECE (↓)	Acc (↑)	ECE (↓)
UC	85.26	0.0202	81.58	0.0208	91.67	0.0247
Mixup	**97.34**	**0.0053**	**90.29**	**0.0181**	89.72	0.0073
TS	95.05	0.0254	77.91	0.0671	**92.78**	**0.0067**
TS+Mixup	94.67	0.0101	86.09	0.0235	91.66	0.0114

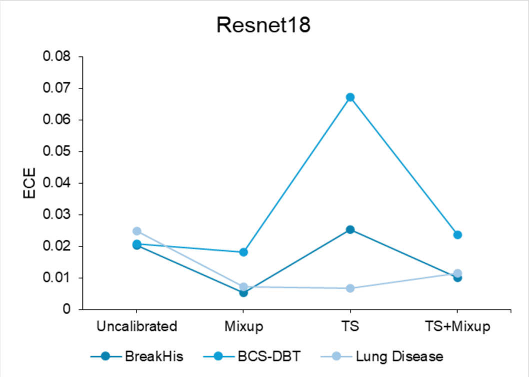

Fig. 2. The experiment's results with the three databases and Resnet18. This graph corresponds to the results of Table 1 for ECE

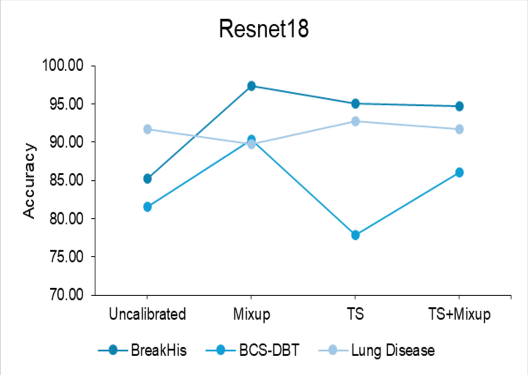

Fig. 3. The experiment's results with the three databases and Resnet18. This graph corresponds to the accuracy results in Table 1

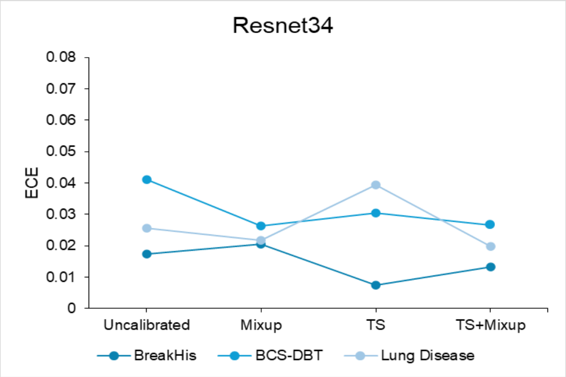

Fig. 4. The experiment's results with the three databases and Resnet34 are shown here. This graph corresponds to the results of Table 2 for ECE

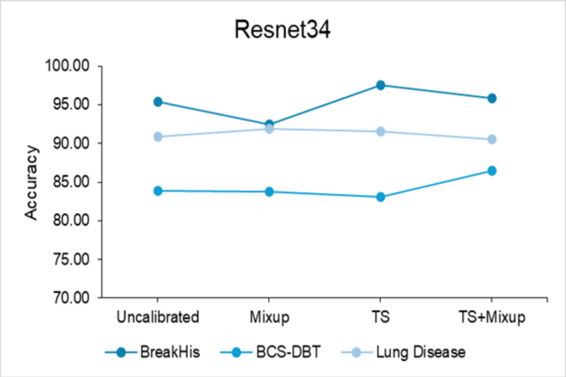

Fig. 5. The experiment's results with the three databases and Resnet34. This graph corresponds to the accuracy results in Table 2 1

Table 2. Results of the Resnet34 experiment

	BreakHis		**BCS-DBT**		**Lung Disease**	
	Acc (↑)	ECE (↓)	Acc (↑)	ECE (↓)	Acc (↑)	ECE (↓)
UC	95.37	0.0175	83.88	0.0411	90.87	0.0257
Mixup	92.51	0.0206	83.72	0.0263	91.89	0.0217
TS	**97.52**	**0.0076**	83.10	0.0305	91.59	0.0393
TS+Mixup	95.85	0.0132	**86.47**	**0.0266**	**90.49**	**0.0198**

Table 3. Results of the Resnet50 experiment

	BreakHis		BCS-DBT		Lung Disease	
	Acc (↑)	ECE (↓)	Acc (↑)	ECE (↓)	Acc (↑)	ECE (↓)
UC	97.61	0.0037	81.88	0.0523	91.01	0.0199
Mixup	96.81	0.0095	80.09	0.0444	91.29	0.0167
TS	94.69	0.0075	76.23	0.0400	88.96	0.0142
TS+Mixup	95.13	0.0072	80.96	0.0306	91.25	0.0186

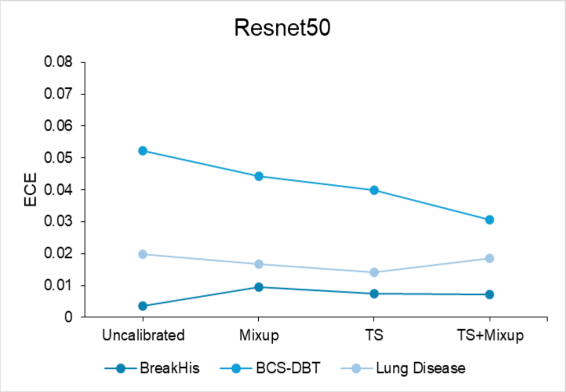

Fig. 6. Results of the experiment with the three databases and Resnet50. This graph corresponds to the results of Table 3 for ECE

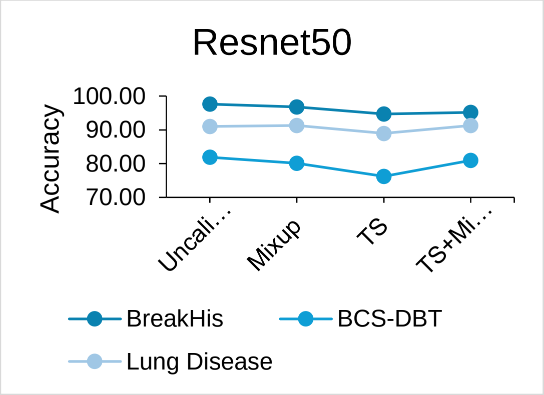

Fig. 7. Results of the experiment with the three databases and Resnet50. This plot correlates with the results of Table 3 for Accuracy

Resnet18: In these experiments, Resnet18 was used with the three medical image bases and the calibration methods Mixup, TS, and TS+Mixup.

Resnet34: Resnet34 was used with the three media image bases and the calibration methods Mixup, TS, and TS+Mixup in these experiments.

Resnet50: Resnet50 was used with the three medical image databases and the calibration methods Mixup, TS, and TS+Mixup in these experiments.

4.2 Analysis and Discussion

In both Resnet18 and Resnet34, it was observed that better results were obtained with the Mixup scoring method and the combination of TS with Mixup. The ECE and Accuracy tend to be the best regardless of the database used, as shown in Figures 1, 2, 3, and 4.

With Resnet18, better results are obtained in both Accuracy and ECE with the BreakHis and BCS-DBT databases, and only in lung disease were better results obtained with the TS method; on the other hand, in the three databases, the TS+Mixup method is second place. Resnet34, using the TS+Mixup method, proved to be efficient with the BCS-BDT and Lung Disease databases, and only in BreakHis were better results obtained with the TS method. In addition, the second-best method is Mixup in all three databases.

Using Resnet50, better results are obtained with the Mixup method with the BreakHis and Lung Disease databases, and with BCS-DBT, better results are obtained using the TS+Mixup method, with these two being the best confidence calibration methods. On the other hand, using this network, it is observed that, uncalibrated, with Mixup, TS, and the combination of TS with Mixup, the results are similar, as shown in Figures 5 and 6.

In the particular case of the BreakHis image database with this network, uncalibrated, superior results are obtained (ECE equal to 0.0037 and Accuracy equal to 97.61) compared to any of the calibration methods used; however, this case represents a singularity compared to the remaining 35 experimental cases.

As an example of one of the experiments performed, we can see in Figure 8 that the BreakHis image base, using Resnet34, with the TS

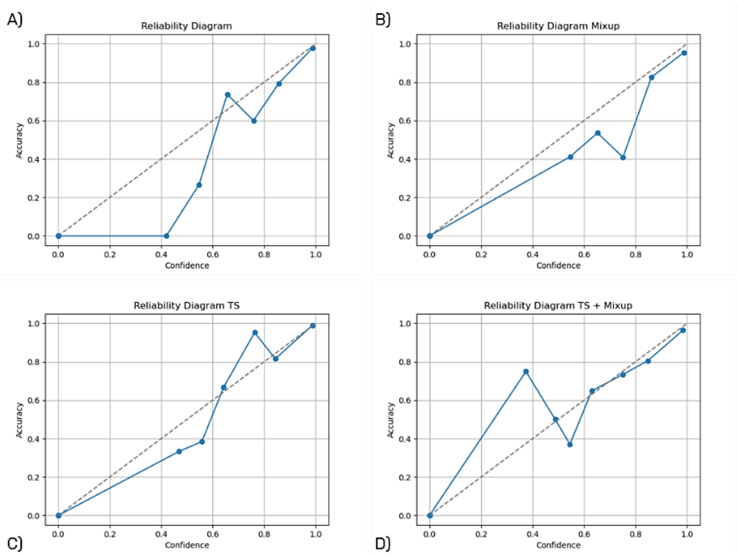

Fig. 8. Reliability diagrams using BreakHis and Resnet34. A) Without using any calibration method. B) Using Mixup. C) Using Temperature Scaling (TS) and D) Combination of TS with Mixup

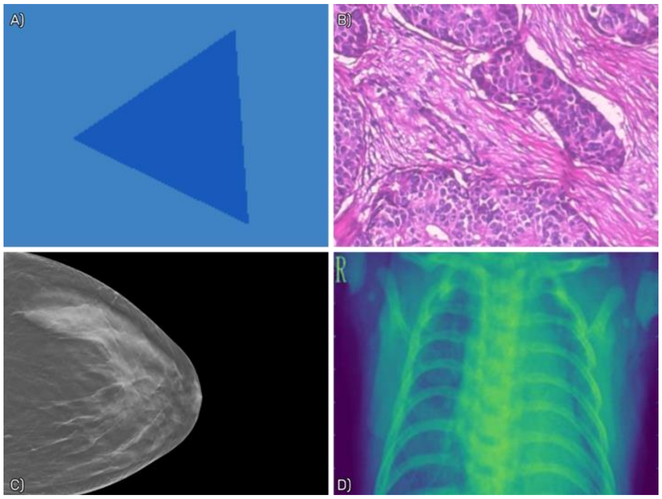

Fig. 9. Examples of images from each medical image database used. A) Geometric figures. B) BreakHis. C) BCS-DBT and D) Lung Disease

method shown in item C) achieves the best result in both Accuracy and ECE.

Furthermore, the second-best method is the combination of TS + Mixup, the graph is shown in item D). Therefore, it can be concretized that both the plots of the tables shown in section 4.1 agree with the individual results of each experiment shown in Figure 8.

On the other hand, an example of each MIDBs used is shown in Figure 9.

As mentioned in section 2. 1 each one has a different number of images, and they are of different types, different acquisition techniques were used, the base case, i.e., geometric figures has the highest number of images and is the one that better balanced compared to medical imaging, so, it can be said that in different learning contexts such as large scale training set versus smaller scale training sets and differences in image, structure can lead to different performance of calibration methods, medical images have some degree of complexity to work with.

5 Conclusions and Future Work

In general, the Mixup method proved efficient with Resnet18 and Resnet50, and the TS method with Mixup proved efficient with Resnet34. Moreover, all three networks efficiently classified diseases independently of the database. However, it should be noted that the classification of medical images presents significant challenges compared to natural images such as geometric figures.

In this sense, confidence calibration is essential for prediction models, especially in the medical context, to identify erroneous or unreliable predictions made by the model. Furthermore, implementing confidence calibration methods in models decreases overconfident predictions and improves the projections made by the model.

Future work will include more confidence calibration methods such as Platt Scaling and Beta Calibration [12].

Acknowledgments

N. López-Miguel is grateful to the Mexican National Council of Humanities, Science and Technology (CONAHCYT) for the Ph.D. scholarship.

References

1. **Karimi, D., Gholipour, A. (2022).** Improving Calibration and Out-of-distribution Detection in Deep Models for Medical Image Segmentation. IEEE Transactions on Artificial Intelligence, Vol. 4, No. 2, pp. 383–397. DOI: 10.1109/ TAI.2022.3159510.
2. **Rao, A., Lee, J.Y., Aalami, O. (2023).** Studying the Impact of Augmentations on Medical Confidence Calibration. In:Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2454–2464.DOI:10.1109/ICCVW60793.2023.00260
3. **Ge, S., Yuan, K., Han, M., Sun, D., Zhang, H., Ye, Q. (2022).** BSM loss: A Superior Way in Modeling Aleatory Uncertainty of Fine_Grained Classification. arXiv preprint arXiv:2206.04479.
4. **Gao, R., Li, T., Tang, Y., Xu, Z., Kammer, M., Antic, S.L., Landman, B. (2022).** A Comparative Study of Confidence Calibration in Deep Learning: from Computer Vision to Medical Imaging. arXiv preprint arXiv:2206.08833.
5. **Kurz, A., Mehrtens, H.A., Bucher, T.C., Brinker, T.J. (2023).** On the Calibration of Neural Networks for Histological Slide-level Classification. DOI: 10.48550/ARXIV.23 12.09719.
6. **Frenkel, L., Goldberger, J. (2022).** Calibration of Medical Imaging Classification Systems with Weight Scaling. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer Nature Switzerland, pp. 642–651. DOI: 10.1007/978-3-031-16452-1_61.
7. **Zhang, J., Kailkhura, B., Han, T.Y.J. (2020).** Mix-n-match: Ensemble and Compositional Methods for Uncertainty Calibration in Deep Learning. International Conference on Machine Learning, pp. 11117–11128.
8. **Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L. (2015).** A Dataset for Breast Cancer Histopathological Image Classification. IEEE Transactions on Biomedical Engineering, Vol. 63, No. 7, pp. 1455–1462. DOI: 10.1109/TBME.2015.2496 264.
9. **Buda, M., Saha, A., Walsh, R., Ghate, S., Li, N., Swiecicki, A., Mazurowski, M. (2020).** Breast Cancer Screening–digital Breast Tomosynthesis (BCS-DBT). Type: dataset, Google Scholar.
10. **Rajaraman, S., Ganesan, P., Antani, S. (2022).** Deep Learning Model Calibration for Improving Performance in Class-imbalanced Medical Image Classification Rasks. PloS one, Vol. 17, No. 1, pp. e0262838. DOI: 10.1371/ journal.pone.0262838.

Article received on 10/06/2024; accepted on 18/08/2024.
**Corresponding author is Nancy López-Miguel.*