INTERFACE

royalsocietypublishing.org/journal/rsif

Correction

Cite this article: Jiménez-García B, Abellán N, Baquedano E, Cifuentes-Alcobendas G, Domínguez-Rodrigo M. 2020 Corrigendum to 'Deep learning improves taphonomic resolution: high accuracy in differentiating tooth marks made by lions and jaguars'. *J. R. Soc. Interface* **17**: 20200782. http://dx.doi.org/10.1098/rsif.2020.0782

Subject Areas:
evolution, computational biology

Corrigendum to 'Deep learning improves taphonomic resolution: high accuracy in differentiating tooth marks made by lions and jaguars'

Blanca Jiménez-García, Natalia Abellán, Enrique Baquedano, Gabriel Cifuentes-Alcobendas and Manuel Domínguez-Rodrigo

J. R. Soc. Interface **17**, 20200446 (2020; Published 29 July 2020) (doi:10.1098/rsif.2020.0446)

1. Introduction

In our previous paper [1], we presented some convolutional neural network (CNN) models to classify images of tooth scores made by lions and jaguars through deep learning computer vision. In that work, we reached an accuracy of 82% of the testing set correctly classified. However, such an accuracy is biased, since the original sample was highly unbalanced. The accuracy reported was impacted by a better classification of the larger lion sample than the smaller jaguar sample. In our original study, to compensate for the unbalanced tooth mark samples, we created a more balanced subsample composed of 42 images of tooth marks from jaguars and 42 tooth marks from lions. This was done by shuffling the original image dataset and randomly sampling 42 images of each agent. In this smaller sample, accuracy decreased slightly. When confusion matrices for this subsample were inspected, it appeared that the classification of tooth marks from both agents was balanced. For the most successful model (VGG19; accuracy = 75.6), the F1-score factor was 0.71. We neglected checking the accuracy balance in the larger sample. We did that posteriorly and realized a divergence between precision and recall in our models. In the larger sample, tooth marks of lions were well classified, but a significant portion of those of jaguars were misclassified. The reason for this is that a few of the tooth marks made by lions display microscopic features that are very similar to those documented in jaguar tooth scores. When using a small randomly sampled set of images, the probability of including that minor part of the lion sample is small and, hence, the similar values obtained for accuracy and F1-score factors. However, when using the much larger lions sample, that jaguar-looking portion of tooth scores is enough to produce a low precision/recall for the jaguar testing sample, because the algorithm sees those marks similar to those documented in lions.

A subsequent analysis of multiple carnivore tooth marks underscored this problem by showing systematic misclassification of the jaguar tooth scores [2]. They were mostly classified as lion tooth marks. We thought that this indicated that both types of tooth marks were situated in different parts of a general felid tooth mark spectrum, but that they overlapped enough to make their differentiation difficult (against our own previous work). Additionally, we also realized that in most of the transfer knowledge models used in Jiménez-García *et al.*'s study [1], we used the same pre-processing standard function, instead of using the model-specific pre-processing functions. We thought these might have produced different results. In the multiple carnivore study, we also realized that augmentation, usually considered a panacea for avoiding overfitting and producing higher accuracy models [3], did not universally do so, and several models yielded higher accuracy when not using image augmentation. For these reasons, we considered adequate to address whether the unbalanced classification problem of the published lion-jaguar models, which rendered their utility rather limited, could be overcome by using non-augmented architectures and using

THE ROYAL SOCIETY PUBLISHING

the pre-processing functions specifically designed for each of the transfer knowledge models. We also thought that ensemble learning, which is known to produce more balanced results, could also improve the precision–recall imbalance documented in our previous modelling. Here, we present the results, which correct the problems of the previously published models by producing more balanced classifications and also by achieving higher accuracy than previously reported.

2. Methods

We used the same image dataset used by Jiménez-García *et al*. [1] and selected the following models: VGG19, Densenet 201, ResNet50, Inception V3 and InceptionResNetV2. For a description of these architectures, see our previous publication. The method implemented was the same as before, with the following modifications. First, we realized that the non-augmented models yielded similar or superior accuracy. For this reason, in this correction to the original models, we discarded image augmentation. Also, we ran the models using a common standard normalization procedure and then with the specific pre-processing functions for each of the transfer models. We observed that the specific pre-processing functions yielded similar accuracy for the ResNet50, Inception V3 and InceptionResNetV2 models; higher accuracy for the Densenet 201 model (by 4 points) and substantially lower accuracy for the VGG19 model (by more than 8 points). We selected the version with the highest accuracy and lowest loss for each of the five models. Then we compared the five models individually. In this process, we paid as much attention to accuracy as to balanced classification (F1 score) and area under the curve (AUC)).

In the second stage, we carried out a stacking ensemble analysis. This machine learning method uses a procedure consisting of stacking a collection of classification models. This ensemble learning method uses the predictions of a set of supervised algorithms to generate an aggregated final prediction. Stacking has the advantage over other ensemble methods, such as bagging or boosting, of enabling the combination of diverse algorithm types. The final prediction is made through a multiple (usually double) layer model. The baseline layer is the algorithms' predictions taken as inputs. The upper layer is the transformation of those inputs via a classifier (e.g. a logistic regression or a decision tree). Sometimes, an additional layer can be implemented using a different classifier over the previous one. Stacking is known to be more successful at classifying than other ensemble methods and especially, than single-trained models [4,5]. This statement must be nuanced when dealing with small sample sizes. In these situations, individual models may be as good or even superior to ensemble approaches.

Here, we applied a stacked ensemble model using the five transfer learning algorithms described above as the baseline and a random forest as the classifier. The random forest was tuned to produce 100 trees. No maximum depth was specified. The number of features selected was specified via the square root of the feature range. The resulting model was contrasted against the testing set, with special emphasis on the degree of balanced classification.

3. Results

The five models yielded similar accuracy as in [1]. Four of the models yielded an accuracy of 82% of the testing set and F1 scores < 0.5, indicating a very unbalanced classification (table 1), as was also the case in Jiménez-García *et al*.'s [1] analysis. This resulted from all the tooth scores of lions (larger sample) having been correctly classified and those of jaguars (smaller sample) having been misclassified. One model (VGG19) yielded a higher accuracy (88%) and a more balanced classification (F1 score = 0.66). This resulted from all the lion marks and half of the jaguar marks having been correctly classified.

By contrast, the stacked model resulted in an overall higher accuracy (89%) and a much better classification (F1 score = 0.84), with 90.5% of all the testing lion marks and 83.3% of all the testing jaguar marks correctly classified (table 2). This underscores the greater efficiency in classification of ensemble learning methods, especially with unbalanced datasets.

Table 1. Classification indicators for each of the individual models.

	accuracy	loss	F1 (macro avg)	AUC
VGG19	0.88	0.17	0.71	0.66
Densenet 201	0.82	0.019	0.45	0.5
Inception V3	0.82	0.019	0.45	0.5
InceptionresnetV2	0.82	0.0071	0.45	0.5
Resnet 50	0.82	0.023	0.45	0.5

Table 2. Classification indicators for the stacked model using a random forest as an upper-level classifier.

	precision	recall	F1 score	support
Jaguar	0.67	0.83	0.74	12
Lion	0.96	0.91	0.93	53
micro avg	0.89	0.89	0.89	65
macro avg	0.81	0.87	0.84	65
weighted avg	0.91	0.89	0.90	65

4. Conclusion

The unbalanced classifications reported by Jiménez-García *et al*. [1] were due to an artefact of method and not to both carnivores being similar in their tooth morphologies or in their behaviour during carcass consumption. This corrected ensemble analysis reinforces the original conclusions reported in [1], claiming that lions and jaguars are different in the tooth-marking patterns documented on bones from carcasses consumed by both types of taphonomic agents. The reason may lie in the fact that lions are flesh eating carnivores, whereas jaguars are more durophagous in their carcass consumption behaviours, resulting in more highly modified bone remains [6]. Therefore, tooth marks imparted with force on bones are more likely to show a wider range of shape and size than those created accidentally during defleshing only.

Data accessibility. The image dataset and the resulting models can be found at https://doi.org/10.7910/DVN/A0ZACG.

Competing interests. We declare we have no competing interests.

Funding. We thank the Spanish Ministry of Education and Science for funding this research (HAR2017-82463-C4-1-P).

Acknowledgements. We wish to thank Ilkka Sipila for methodological suggestions.

References

1. Jiménez-García B, Aznarte J, Abellán N, Baquedano E, Domínguez-Rodrigo M. 2020 Deep learning improves taphonomic resolution: high accuracy in differentiating tooth marks made by lions and jaguars. *J. R. Soc. Interface* **17**, 20200446. (doi:10.1098/rsif.2020.0446)
2. Abellán N, Jiménez-García B, Aznarte J, Baquedano E, Dominguez-Rodrigo M. 2020 Deep learning classification of tooth marks made by different carnivores: achieving high accuracy when comparing African carnivore taxa and testing the hominin shift in the balance of power. *Arch. Anthop. Sci.* (submitted).
3. Chollet F. 2017 *Deep learning with Python*. New York, USA: Manning Publications Company.
4. Wolpert DH. 1992 Stacked generalization. *Neural Netw.* **5**, 241–259. (doi:10.1016/S0893-6080(05)80023-1)
5. Ghorbani AA, Owrangh K. 2001 Stacked generalization in neural networks: generalization on statistically neutral problems. In *IJCNN'01. Int. Joint Conf. on Neural Networks. Proc. (Cat. No.01CH37222), Washington DC, July 15–19 2001*, vol. 3, pp. 1715–1720.
6. Dominguez-Rodrigo M, Yravedra J, Organista E. 2015 A new methodological approach to the taphonomic study of paleontological and archaeological faunal assemblages: a preliminary case study from Olduvai Gorge. *J. Archa. Sci.* **59**, 35–53. (doi:10.1016/j.jas.2015.04.007)