applied sciences

MDPI

Article

Use of Generative Adversarial Networks (GAN) for Taphonomic Image Augmentation and Model Protocol for the Deep Learning Analysis of Bone Surface Modifications

Manuel Domínguez-Rodrigo [1,2,*], Ander Fernández-Jaúregui [1], Gabriel Cifuentes-Alcobendas [1,2] and Enrique Baquedano [1,3]

[1] Institute of Evolution in Africa (IDEA), Alcalá University, Covarrubias 36, 28010 Madrid, Spain; anderfernandezj@gmail.com (A.F.-J.); gabrcifu@ucm.es (G.C.-A.); enrique.baquedano@madrid.org (E.B.)
[2] Area of Prehistory (Department History and Philosophy), University of Alcalá, 28801 Alcalá de Henares, Spain
[3] Regional Archaeological Museum of Madrid, Plaza de las Bernardas s/n, 28001 Alcalá de Henares, Spain
* Correspondence: manuel.dominguezr@uah.es

Abstract: Deep learning models are based on a combination of neural network architectures, optimization parameters and activation functions. All of them provide exponential combinations whose computational fitness is difficult to pinpoint. The intricate resemblance of the microscopic features that are found in bone surface modifications make their differentiation challenging, and determining a baseline combination of optimizers and activation functions for modeling seems necessary for computational economy. Here, we experiment with combinations of the most resolutive activation functions (relu, swish, and mish) and the most efficient optimizers (stochastic gradient descent (SGD) and Adam) for bone surface modification analysis. We show that despite a wide variability of outcomes, a baseline of relu–SGD is advised for raw bone surface modification data. For imbalanced samples, augmented datasets generated through generative adversarial networks are implemented, resulting in balanced accuracy and an inherent bias regarding mark replication. In summary, although baseline procedures are advised, these do not prevent to overcome Wolpert's "no free lunch" theorem and extend it beyond model architectures.

Keywords: generative adversarial networks; optimizer; activation function; neural networks; computer vision; taphonomy

Citation: Domínguez-Rodrigo, M.; Fernández-Jaúregui, A.; Cifuentes-Alcobendas, G.; Baquedano, E. Use of Generative Adversarial Networks (GAN) for Taphonomic Image Augmentation and Model Protocol for the Deep Learning Analysis of Bone Surface Modifications. *Appl. Sci.* **2021**, *11*, 5237. https://doi.org/10.3390/app11115237

Academic Editor: Miguel Ángel Maté-González

Received: 20 April 2021
Accepted: 1 June 2021
Published: 4 June 2021

1. Introduction

The use of computer vision (CV) through deep learning (DL) methods has substantially modified the resolution of taphonomic studies. It initially showed that, on small samples ($n \leq 100$), CV yielded an accuracy of classification > 90% when human experts were systematically producing < 60% correct identifications of tested bone surface modifications (BSM). In larger samples, the difference between CV and experts becomes exponential. Preliminary models have been produced, using deep convolutional neural networks (DCNN), that correctly identify images of tooth, cut, and trampling marks [1]. DCNN models go as far as to differentiate cut marks imparted on bones when carcasses were fleshed or defleshed [2]. These methods are even capable of detecting BSM morphing through dynamic impact of biostratinomic abrasion processes [3]. Traditional taphonomic studies do not have the power to identify carnivore agency when carnivorous mammals affect bone assemblages. DL models have successfully and variably differentiated among several diverse carnivore types [4]. Some DL models have even successfully classified tooth marks from different felid types [5,6]. All this shows the promising path ahead in the use of these techniques for taphonomic research.

However, optimal DL model construction is not easy, since it involves combinations of multiple variables, namely the architecture of the model, the choice of transfer and

ensemble learning procedures, the selection of activation functions and the choice of optimizers, among others. There currently is no protocol for which of these combinations is most adequate in the use of BSM images for taphonomic analyses. Most of the models mentioned above are based on imbalanced samples and their balanced accuracy is variable. In the present work, our objectives are two-fold: we intend to provide a baseline protocol based on combinations of different architectures, activation functions and optimizers, and we will show the convenience of including, in these protocols, data augmentation procedures for coping with unbalanced samples.

Image augmentation has frequently been done using morphing of the currently existing datasets by modifying spatial properties, such as rotation, zoom, flipping, cropping, translation, kernel filters, noise introduction, and others [7–10]. DCNNs need large datasets for training in order to efficiently learn, and small samples hinder the process by either biasing it or by overfitting the training and underperforming on the testing. Data augmentation has been essential to avoid these issues. One major improvement over traditional image augmentation techniques has been the development of generative adversarial networks (GAN) [11,12]. GANs are capable of, not just modifying existing images, but creating new images that could diversify within sample variability. This has boosted the efficiency of medical imaging, where samples for specific pathologies are always limited [13–16]. Here, we will use GAN data augmentation with the goal of improving existing models and making accuracy more balanced. However, we will present some problems associated with GAN methods that require some caution in how these methods are implemented in BSM studies.

2. Methods and Samples

2.1. Phase 1: Parameter Selection and Model Protocol

In order to test the best architecture and parameters for BSM DL analysis, we selected some of the most powerful existing DCNN models, some of them successfully used for BSM classification [1]: VGG16, ResNet50, Densenet, and Jason2. All these models, but the latter, were used through transfer learning. Transfer learning consists of using models that were trained for a different problem and use their feature-learning weights, which are already pre-trained [17], for retraining on new image datasets. Here, some of the most high-performing models trained on more than 1,000,000 images for the 1000-image category ILSVRC competition were used. These pre-trained models were used as standalone feature extractors and classifiers. The layers of each pre-trained model with their weights were integrated within the new models used here containing an output dense layer containing 128 neurons. This was implemented through the Keras API. For a summary of the description of these architectures see [1,5]. In previous modeling, BSM images were high resolution (80 × 400 pixels) [1]. Here, we adopted a lower resolution approach, since experimentation showed that model accuracy was not affected. For this reason, we reshaped the original rectangular BSM images into 64 × 64-pixel images. The original images were captured using a binocular microscope (Optika) at 30×. The resulting BSM image data bank was composed of 488 cut marks, 106 tooth marks, plus 63 marks from trampling experiments. Cut marks were made with simple flakes. Tooth marks were obtained in experiments of bones modified by lions and wolves. Trampling marks were obtained from experiments using different times of bone exposure to diverse sand-grain sizes. For a detailed description of the BSM samples refer to [1]. For a more in-depth description of the experiments, see [2,5,18]. All images were transformed into black and white during image processing in the Keras platform (with a Tensorflow backend), by using bidimensional matrices for standardization and centering. The architectures were designed to address a multiple classification problem. For this reason, the "softmax" activation was used on the last dense layer. This function is specific for multinomial classification. It provides the probabilities of each input element of pertaining to a specific label. During compilation, categorical cross-entropy was used as the loss function. We used Tensorflow 2.4.1 and Keras 2.4.2.

The architectures were trained on 70% of the original BSM image set (n = 459). The resulting models were subsequently tested against the 30% remaining sample, which constituted the testing set (n = 197). Training and testing were performed through mini-batch kernels (size = 32) because it is the minimum recommended. Models were run using a backpropagation process for 50 epochs.

The two parameters selected for intra- and inter- model performance comparison were the activation functions and the optimizers. Following experimentally-based recommendations, we selected the "relu" activation as the baseline function [7,17,19,20]. Relu (rectified linear activation) has become the default activation function for DL models because it deals efficiently with vanishing gradient problems (decreased error through backpropagation), the saturation on the ends and the sensitivity focused in the mid-points of other activation functions like sigmoid or tanh [20]. The "relu" function returns a value equivalent to the input if activated or zero if inactivated. The formula is rather simple: $f(x) = max(0,x)$.

Recently, a purported improvement over "relu" has been suggested by the Google Brain team under a new activation function named "swish", whose formula is $f(x) = x \cdot sigmoid(x)$ [21,22]. Swish is a monotonic smooth function that implements a sigmoid curvature at the base of the ramp and it is more flexible for activation/deactivation. It is supposed to work better in deep architectures (>40 layers) because of the problems in correctly activating deep layers. Here, we will compare the performance of both functions for BSM image analysis.

"Mish" is another recent function innovation. The formula is a little more complex:

$f(x) = x \cdot tanh(\varsigma(x))$ where, $\varsigma(x) = \ln(1 + e\hat{}x)$. This latter is the softplus activation function [23]. "Mish" is not limited in its upper slope values, which avoids saturation and deals efficiently with zero-gradient problems. In contrast, the slope is bounded at the bottom, which enables better performance at avoiding overfitting during training. It has been argued that this function outperforms "relu" and "swish" [23]. We tried alternative functions, such as "tanh" and "LeakyReLU" and the results were not as good as with "relu", "swish" and "mish".

The second parameter that we will analyze and compare is the optimizer type. Optimizers try to minimize the loss function by tweaking weight parameters. In the case of the optimizers, the responsibility of the analyst is bigger than in the selection of the activation functions, because optimizers need to be tuned. For example, convergence to an optimum depends on how fast or slow gradients operate. This is why selection of the learning rate is important. Regularization can also be implemented to avoid overfitting. Other parameters may also be selected. Gradients (which are partial derivatives connecting the loss function to the weights) show any minor or major impact that any modification of parameters does to the weights. The goal is to reduce the loss function through gradient descent. One of the most widely used optimizers is stochastic gradient descent (SGD). This optimizer uses only batches or random selection of cases in the original training sample, instead of all the cases contained therein. It is used in combination with another technique called "momentum", which consists of accumulating the gradient values of past steps along the descent trajectory in order to determine directionality. It is initiated at 0.5 and it increasingly must reach 0.9 over subsequent iterations. Here, we will use SGD as the baseline optimizer. The default parameters were: Learning rate = 1×10^{-3}; momentum = 0.9.

Adam (adaptive moment estimation) has been proposed as another of the most efficient optimizers [24]. It takes another optimizer (Adagrad, adaptive gradient algorithm) as a baseline and modifies it by scaling the learning rate instead of averaging it. It maintains an exponentially decreasing average of previous gradients. The algorithm implements an exponential modification of the average of the moving gradients and squares the gradient. We selected the VGG16 and Jason 2 models and tested them with alternative optimizers: SGD, Adam, Adagrad, Adadelta, and RMSProp. Given that the best results were obtained with SGD and Adam, we decided to use these two optimizers as competitors in combination with different activation functions. This created a network of combinations

including: four different DL architectures, three activations functions and two optimizer types. This resulted in 24 different combinations or models.

Model evaluation was based on the joint consideration of accuracy, loss and F1-scores (Table 1). The latter was determinant given that it provides a good indication on how good balanced accuracy is. We considered the two best combinations for each model. The first one was labeled "best" and the second one as "good", with the remainder being labeled as "normal". This was applied to the six function-optimizer combinations of each model. To properly assess the importance of each function and optimizer, a Bayesian network was used. This was based on score-searching structure learning algorithms, namely a hill-climbing algorithm with different scores [25]: the Bayesian information criterion (BIC) score, which is equivalent to the minimum description length (MDL), the multinomial log-likelihood (loglik), the Bayesian Dirichlet equivalent (BDE) score, the Bayesian Dirichlet sparse score (BDS), and the locally averaged Bayesian Dirichlet (BDLA) score [26]. The final network selected was the one showing the lowest score (i.e., Akaike–BIC). The "bnlearn" R library was used for this analysis (www.r-project.org, accessed on 2 February 2021).

Table 1. Accuracy and loss of the four selected sequential and parallel models, according to combination of activation function and optimizer.

Model	Function	Optimizer	Accuracy	Loss	F1-Score
VGG16					
	relu	SGD	95.83	0.214	0.73
	swish	SGD	95.83	0.286	0.67
	mish	SGD	95.31	0.245	0.67
	relu	Adam	96.35	0.247	0.65
	swish	Adam	95.31	0.246	0.77
	mish	Adam	96.35	0.258	0.69
Densenet 201					
	relu	SGD	94.27	0.2	0.67
	swish	SGD	93.75	0.206	0.63
	mish	SGD	95.31	0.185	0.67
	relu	Adam	95.31	0.162	0.71
	swish	Adam	93.75	0.224	0.64
	mish	Adam	89.06	0.395	0.65
Jason2					
	relu	SGD	96.35	0.133	0.69
	swish	SGD	95.38	0.16	0.7
	mish	SGD	96.35	0.126	0.68
	relu	Adam	94.79	0.194	0.76
	swish	Adam	93.75	0.224	0.69
	mish	Adam	95.83	0.212	0.71
Resnet 50					
	relu	SGD	97.92	0.112	0.8
	swish	SGD	98.44	0.058	0.77
	mish	SGD	97.92	0.104	0.69
	relu	Adam	95.83	0.124	0.76
	swish	Adam	97.4	0.074	0.71
	mish	Adam	97.92	0.147	0.67

2.2. Phase 2: GAN-Augmented Sampling and Model Testing

Subsequently to the parameter and model testing, the most successful combinations of model-function-optimizer were selected and used as the protocol for the complete model analysis with image augmentation. For this last stage, the original highly unbalanced sample was augmented with GAN. Given that, originally, they were between five and eight times smaller than the cut mark subsample, we artificially augmented only the tooth mark and trampling mark datasets by adding 500 images for each category.

In order to generate images for each class, two main requirements were to be fulfilled. First, the generated images should be large enough to be fed to the discriminator. On the other hand, in order to train the algorithm, few images were made available. Despite that modern image generation models complying with these two requisites do exist, as the recently published F2GAN [27], technical constraints made this and other approaches, such as DAGAN [28] non-viable.

As a result, in order to tackle the problem, a divide and conquer strategy was adopted: a simple GAN for creating high-fidelity low-resolution images and a super resolution GAN (SRGAN) that would upscale the generated images into a resolution optimum to train the model. Regarding the GAN, a simple approach was taken, using a simple GAN [11], but with small modifications. In the case of the generator, the initialization of the input vector followed a standard normal distribution. This vector was convolved into a squared array with 128 channels and width 64. After a LeakyReLU activation, the data were convolved and upsampled using a Conv2DTranspose layer, normalized using batch normalization, and then activated with a LeaykyReLU. This process was done twice, using a momentum of 0.8 in batch normalization and an alpha of 0.2 in LeakyReLU in both cases. As a result, the generation returned 64 × 64 grayscale images (Figure 1).

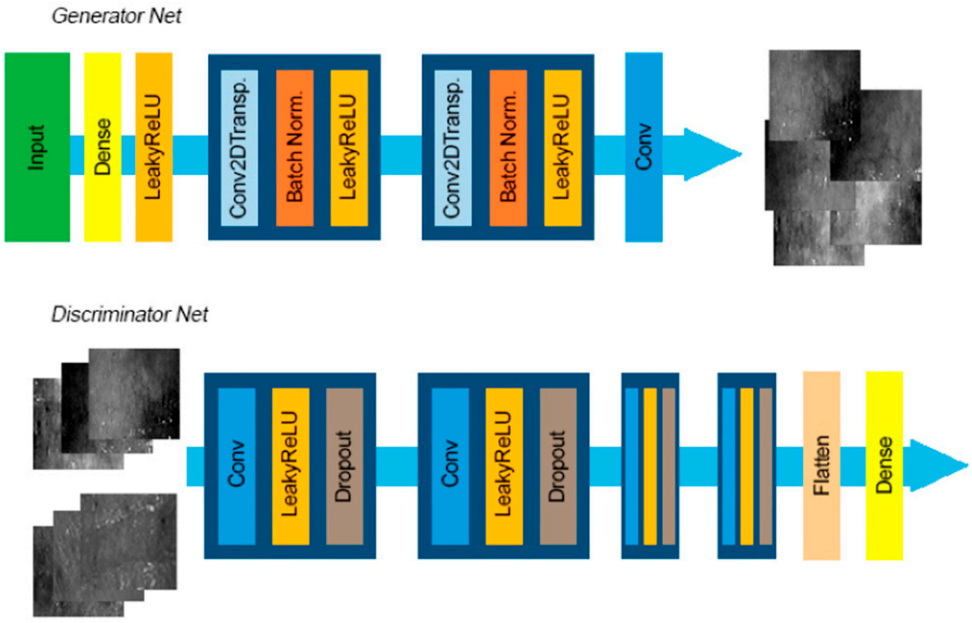

Figure 1. Graphic representation of the created GAN.

In order to train the discriminator, both real and fake images were used, applying a sequence of convolution, LeakyReLU, and a dropout at a rate of 0.5. This process was done four times. Including the dropout layer is a small change from the original implementation, which avoided the discriminator to learn the images and create overfitting or chances of collapse mode, improving the model's performance. For the generation of trampling and tooth marks, the GAN was run through 7000 epochs of training. After this, the model ended up generating realistic images (Figure 2).

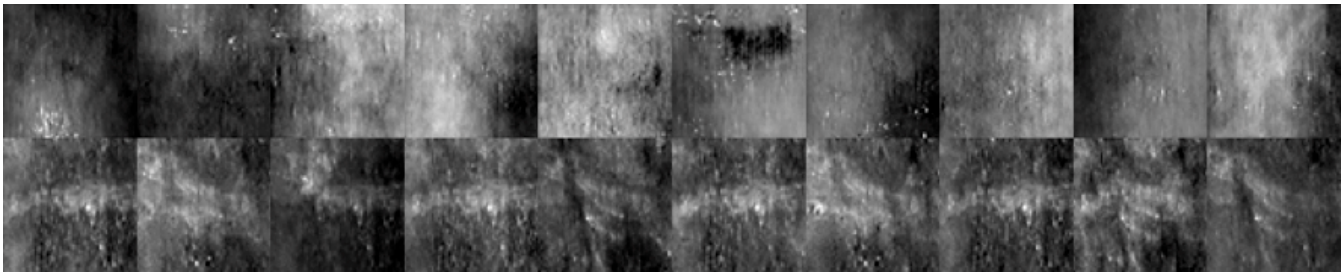

Figure 2. Sample of 10 generated images of tooth marks (**upper row**) and trampling (**lower row**).

Regarding the execution of the super resolution GAN, a literal implementation of the defined super-resolution GAN [29] was made. As a result, the generated images significantly increased their resolution, going from 64 × 64-pixel images to 256 × 256 pixel and, eventually, to 1024 × 1024-pixel images (Figure 3). Finally, it was decided that the 256 × 256 images were the ones to be added to the model. The code structure is added to this paper.

After completing the GAN-augmented dataset, the two most successful classification models obtained from the combination of activation function/optimizer were used to test their efficiency on the augmented data.

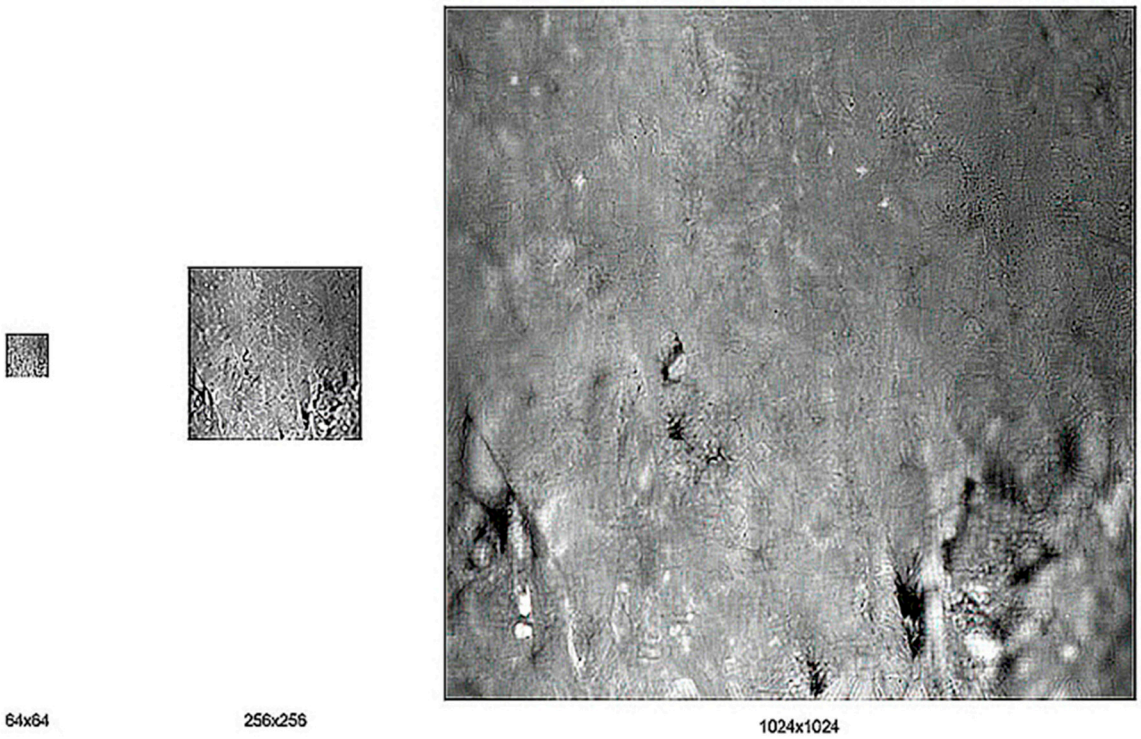

Figure 3. Image resolution improvement by applying SRGAN twice.

3. Results

3.1. Phase 1: Parameter Selection and Model Protocol

There was a major improvement over previous modeling of the same dataset [1], with most models providing an accuracy close to 95% (Table 1). In this case, the differences among models were rather minor. Sequential models like VGG16 and Jason 2 performed very well and at least one of the parallel architectures (ResNet50) also performed very efficiently. As a matter of fact, the latter model provided the highest accuracy, the lowest loss

and the best-balanced accuracy, as reported by the F1-scores. Within these architectures, the choice of function had a minor impact in model performance (Table 1). The optimizer was more relevant to the final results. In most models, if considering the best two combinations for each model, SGD produced better performance of the network resulting in higher accuracy and lower loss (Table 1). SGD displayed the best performance of the "relu" (37.5%), the "swish" (12.5%) and the "mish" (12.5%) functions (adding results qualified as "best" and "good"). In contrast, Adam only managed to get a lower performance of the "relu" (25%) and the "swish" (12.5%) functions (Table 2). In total, 62.5% of the best performing combinations were achieved by SGD and only 37.5% was obtained with Adam. This, despite that three of the four best single combinations were obtained with Adam: Adam–relu (2), Adam–swish (1) and SGD–relu (1) (Table 2).

Table 2. Proportion distribution of each activation function according to the two optimizers in the three categories: best (first best model), good (second best model) and normal.

Result = BEST			
optimizer		**function**	
	mish	**relu**	**swish**
Adam	0	0.25	0.125
SGD	0	0.125	0
Result = GOOD			
optimizer		**function**	
	mish	**relu**	**swish**
Adam	0	0	0
SGD	0.125	0.25	0.125
Result = NORMAL			
optimizer		**function**	
	mish	**relu**	**swish**
Adam	0.5	0.25	0.375
SGD	0.375	0.125	0.375

The Bayesian networks supported this interpretation and expanded it. The "BIC" (score = −71.52577), the "loglik" (score = −56.68172), the "bde" (score = −72.85762), the "bda" (score = −72.85762), and the "bdla" (score = −72.71031) hill-climbing score factors indicated by a majority of four to one, that the differences among the activation functions were not that relevant (Figure 4). It was the optimizer that made the difference.

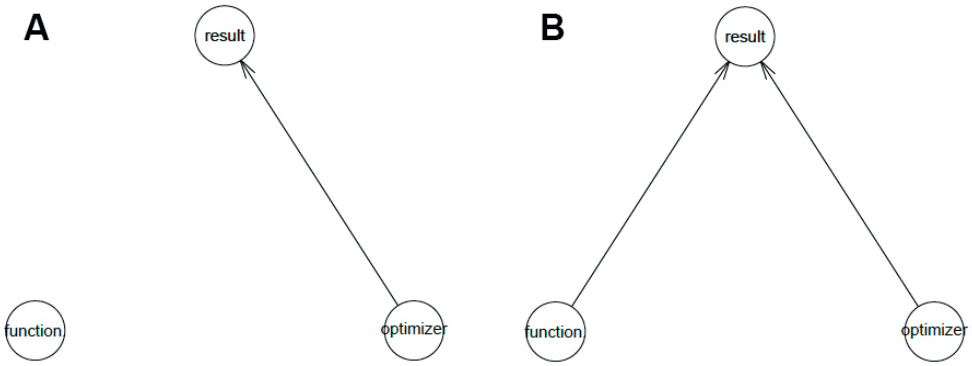

Figure 4. Bayesian network selection by the hill-climbing algorithm. Model (**A**), voted by majority of score functions; model (**B**), voted by the "loglik" function.

This analysis of parameter combinations suggests that, although several combinations must be tried for every dataset, the most parsimonious baseline model should be a combination of "relu" and SGD. Tuning is mandatory, since the best model obtained in this study

was the combination of "relu–SGD" (accuracy = 97.9; F1-score = 0.8) and "swish-SGD" (accuracy = 98.4; F1-score = 0.77) under the ResNet50 architecture.

Regarding the average performance of each architecture using all combinations, the ranking (based on accuracy) is as follows: ResNet50 (97.4%), VGG16 (95.8%), Jason2 (95.3%), and Densenet 201 (93.5%) (Figure 5).

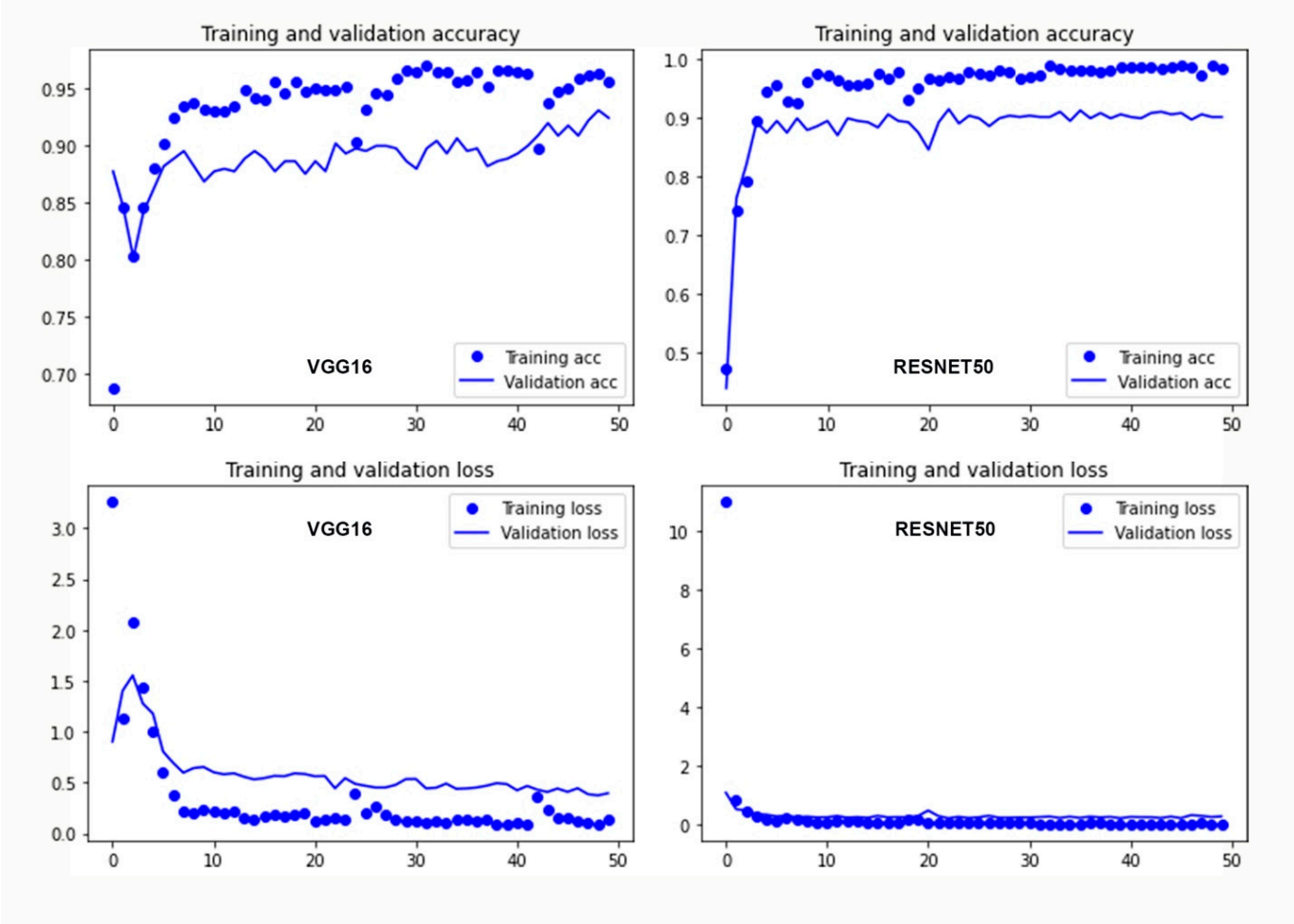

Figure 5. Accuracy and loss values for the VGG16 and ResNet50 models on the GAN-augmented dataset displaying two different combinations of activation function and optimizer: swish–SGD (VGG16) and swish–SGD (ResNet50). Horizontal axis displays the number of epochs.

3.2. Phase 2: GAN-Augmented Sampling and Model Testing

The GAN-augmented data yielded slightly lower accuracy rates for the testing sets than the raw data (Table 3). It produced similar estimates to previous testing using high-resolution images [1]. In contrast, it also yielded higher balanced accuracy estimates (Table 3). The greater increase in F1-scores is the most modified outcome. In this case, the results are highly variable depending on the combination and not a clear strategy is visible. With "SGD", "swish" produced better results with some models but not others, and the same is observed for "Adam" and "relu". Overall, VGG16 and ResNet50 were the best models. In contrast, the Jason2 model was less successful and its accuracy stagnated several points below, ranging from 57.1 to 88.9 (Table 3). It also yielded lower estimates of balanced accuracy; however, it still outperforms its efficiency when using the non-augmented dataset, by having improved the balanced accuracy by almost ten points in some of the combinations [1]. Within each model, the activation function plays a minor

role in the outcome, and the optimizer seems more relevant, but less so than with the raw dataset. It is the model architecture that makes a difference.

Table 3. Accuracy loss and F1-score of the models using the augmented dataset, according to combination of activation function and optimizer.

	Activation Function	Optimizer	Accuracy	Loss	F1-Score
VGG16	relu	SGD	87.02	0.49	0.84
	swish	SGD	90.82	0.37	0.84
	mish	SGD	84.6	0.56	0.84
	relu	Adam	89.18	0.40	0.80
	swish	Adam	83.41	0.77	0.74
	mish	Adam	88.2	0.64	0.81
Densenet201	relu	SGD	81.73	0.57	0.81
	swish	SGD	83.04	0.53	0.81
	mish	SGD	82.37	0.47	0.80
	relu	Adam	83.26	0.52	0.81
	swish	Adam	81.03	0.66	0.81
	mish	Adam	82.37	0.48	0.81
Jason 2	relu	SGD	88.99	0.49	0.83
	swish	SGD	80.29	1.07	0.71
	mish	SGD	80.5	0.54	0.72
	relu	Adam	86.5	0.36	0.84
	swish	Adam	69.1	1.16	0.64
	mish	Adam	57.1	2.26	0.55
ResNet50	relu	SGD	88.17	0.42	0.84
	swish	SGD	89.96	0.29	0.84
	mish	SGD	89.06	0.32	0.84
	relu	Adam	91.18	0.24	0.83
	swish	Adam	91.29	0.29	0.87
	mish	Adam	87.28	0.50	0.84

The Bayesian networks supported these interpretations. Several algorithms provide different insights into the variable relationship network. The "BIC" (score = −217.1542), the "loglik" (score = −76.27329), the "bde" (score = −202.0743), and the "bdla" (score = −203.2623) hill-climbing score factors indicated different options. The "loglik" factor suggests that the model architecture has the major impact on all the other elements (Figure 6); however, this model has the largest AIC (Akaike information criterion) score. The "bde" score factor suggests a more complex picture, in which the parameters affect the model performance. This is more specifically nuanced by the "bdla" factor, which indicates that although the model architecture has a major impact on the performance of the activation function and the resulting balanced accuracy, it is the optimizer that has a major role in model performance (Figure 6). If we select the optimal network, resulting from the BIC score factor, it can be concluded that all these relationships are of interest, but in general, the only meaningful relationship is that between the accuracy and its impact on loss and the F1-score, underscoring that with the GAN-augmented dataset, the activation function and the optimizer do not create significant differences within each model architecture performance.

It could be argued, based on these results, that the augmented dataset contributed to building a more reliable model, because even if the general accuracy did not improve or became slightly lower, the balanced accuracy was significantly higher.

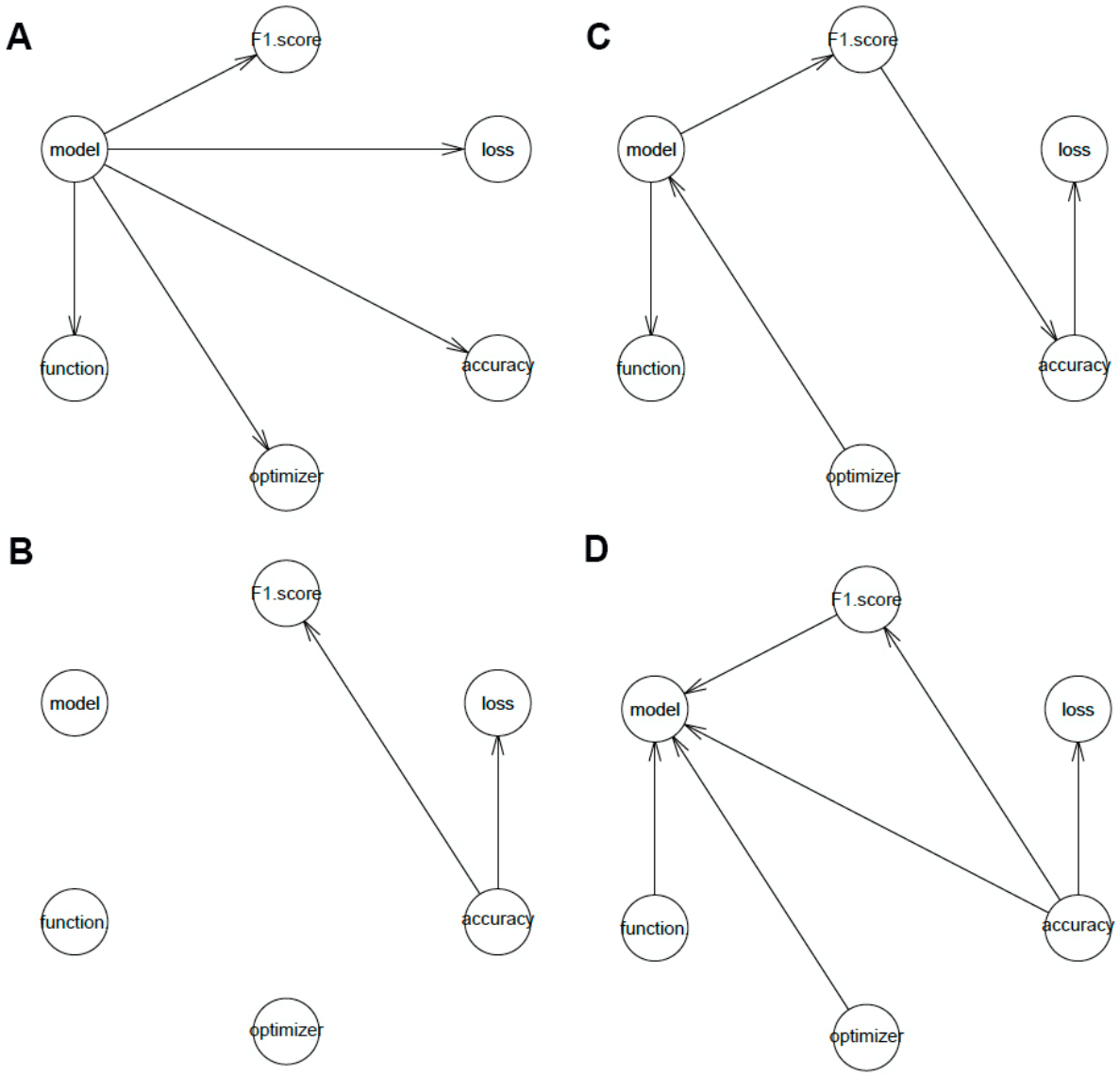

Figure 6. Bayesian network selection by the hill-climbing algorithm on the GAN-augmented image data bank. Model (**A**), voted by the "loglik" score function; model (**B**), voted by the "BIC" function; model (**C**), voted by the "bdla" function; model (**D**), voted by the "bde" function.

4. Discussion

The augmented dataset contained 1657 images, in which the three classes appeared balanced: 606 tooth marks, 488 cut marks and 563 trampling marks. The original imbalanced model, based on a smaller raw data sample, classified correctly 92% of the testing set [1]; however, its F1-score (0.73) was substantially lower. Split by class, cut marks were the best classified (F1 = 0.97), followed by tooth marks (F1 = 0.80), but trampling marks were frequently misclassified (F1 = 0.42), most of the time as tooth marks. This creates uncertainty as to the power of the classifications because of the imbalanced dataset. The original number of cut marks in the sample was eight times bigger than the trampling dataset. Likewise, it was almost five times bigger than the tooth mark subsample. Data augmentation using the generation of new images is essential to balance each class dataset. This usually results in higher accuracy in classification and more reliability in classification probabilities. In the present work, the GAN-augmented sample and models have yielded a

slightly lower global accuracy than when using just the raw data, but the balanced accuracy was systematically higher. The swish-Adam combination in the ResNet50 model yielded the most accurate (91.29%) architecture and the F1.score yielded the highest balanced accuracy (87%) (Table 3); this latter was significantly higher than when using the raw data. One biasing feature that we observed in our augmented sample is that the generator created preferentially trampling images from the original sample which showed the least resemblance to either cut marks or tooth marks. This created an artificial sample where these originally minority marks became predominant in the augmented sample; hence the higher balanced accuracy.

Previously, the VGG16 model was used to preliminarily classify some controversial bone surface modifications from the archaeological record [1]. These marked bone specimens are of extraordinary importance because they could potentially attest some of the "first" traces of human butchery in the locations where they were found. These involved specimens from Bluefish Caves (18,000 B.P., Yukon, Canada), purportedly belonging to the earliest presence of humans in the American continent [30], Anjohibe (1400–2000 B.C.) [31], Itampolo (1100–1800 B.P.) [32] and Christmas River (>10,000 B.P.) [33] (Madagascar)—interpreted as some of the earliest evidence of human presence on the island-, Dikika (3.2 Ma, Ethiopia)—potentially the first evidence of tone tool use- and Barranco León and Fuentenueva 3 (1.4 Ma, Orce, Spain)—presented as the oldest cut marks in Europe [34]. In order to assess architecture-model variability, here we used the most successful model (ResNet50) to tentatively classify some of these marks, bearing in mind that since these BSM were taken from published photographs not following the protocol applied to the experimental BSM, the conclusions are of limited value. Instead of using the swish–SGD combination, which yielded the highest accuracy (98.44%), we used the relu–SGD combination because its balanced accuracy was higher (80%) and then, less prone to misclassify classes. We only used a few of the archaeological marks that were interpreted by human experts as cut marks and that with the VGG16 model had previously been classified as non-anthropogenic [1]. The resulting classification did not vary from that obtained previously with the VGG16 model, using higher resolution images (Table 4). However, the probabilities became smaller (and so did the variance) because we were dealing with lower information images that also imparted some deformation over the original photographs.

Table 4. Probability of classification of a selection of archaeological BSM images for each BSM type using the ResNet50 model.

Site	Tooth Mark	Cut Mark	Trampling Mark	Classification
Bluefish Caves	0.21	0.302	0.486	trampling
Dik 53-3-D	0.215	0.309	0.474	trampling
Dik 53-3-E	0.19	0.338	0.468	trampling
Dik 53-3-H	0.213	0.295	0.491	trampling
Dik 53-3-I	0.211	0.298	0.495	trampling
FuenteNueva 3	0.481	0.215	0.303	tooth mark
FuenteNueva 3	0.215	0.306	0.477	trampling

5. Conclusions

The present study shows that when using image data augmentation, even if the resolution of the images is substantially reduced (which enhances computation), the accuracy can be balanced. For BSM, the augmented samples can be biasing if expanding the least common types of marks only because they are the ones that avoid confusion with the other categories. The application of the protocols described in the comparison of combinations of activation functions and optimizers to the artificially-augmented data also shows that protocols should be taken only as a baseline procedure, since what worked best in the same architectures with pre-augmented data, does not necessarily work best with augmented datasets. Although the augmented data enabled that the function/optimizer combination was virtually irrelevant in the final results, impacting them only in decimal modification, it

also showed that the best hyper-parameter combination is contingent on the characteristics of each dataset. This also expands Wolpert's "no-free lunch" theorem [35] from model selection to hyper-parameter selection.

Author Contributions: Formal analysis M.D.-R., G.C.-A. and E.B.; software, M.D.-R.; G.C.-A. and A.F.-J.; writing—review and editing, M.D.-R.; G.C.-A. and E.B. All authors have read and agreed to the published version of the manuscript.

Funding: Ministry of Education, Science and Universities, Spain (grant: HAR2017-82463-C4-1-P).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original image data set is available at https://doi.org/10.7910/DVN/62BRBP (accessed on 3 June 2021). The code can be found at https://github.com/anderfernandez/GAN-Fosiles (accessed on 3 June 2021).

Acknowledgments: We thank the Spanish Ministry of Education, Science and Universities for funding this research (HAR2017-82463-C4-1-P). We also appreciate the constructive comments made by three reviewers. We would like to express our thanks to M. A. Maté-González for having invited us to participate in this Special Issue.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Domínguez-Rodrigo, M.; Cifuentes-Alcobendas, G.; Jiménez-García, B.; Abellán, N.; Pizarro-Monzo, M.; Organista, E.; Baquedano, E. Artificial intelligence provides greater accuracy in the classification of modern and ancient bone surface modifications. *Sci. Rep.* **2020**, *10*, 18862. [CrossRef] [PubMed]
2. Cifuentes-Alcobendas, G.; Domínguez-Rodrigo, M. Deep learning and taphonomy: High accuracy in the classification of cut marks made on fleshed and defleshed bones using convolutional neural networks. *Sci. Rep.* **2019**, *9*, 18933. [CrossRef] [PubMed]
3. Pizarro-Monzo, M.; Domínguez-Rodrigo, M. Dynamic modification of cut marks by trampling: Temporal assessment through the use of mixed-effect regressions and deep learning methods. *Archaeol. Anthropol. Sci.* **2020**, *12*, 4. [CrossRef]
4. Abellán, N.; Jiménez-García, B.; Aznarte, J.; Baquedano, E.; Domínguez-Rodrigo, M. Deep learning classification of tooth scores made by different carnivores: Achieving high accuracy when comparing African carnivore taxa and testing the hominin shift in the balance of power. *Archaeol. Anthropol. Sci.* **2021**, *13*, 31. [CrossRef]
5. Jiménez-García, B.; Aznarte, J.; Abellán, N.; Baquedano, E.; Domínguez-Rodrigo, M. Deep learning improves taphonomic resolution: High accuracy in differentiating tooth marks made by lions and jaguars. *J. R. Soc. Interface* **2020**, *17*, 20200446. [CrossRef]
6. Jiménez-García, B.; Abellán, N.; Baquedano, E.; Cifuentes-Alcobendas, G.; Domínguez-Rodrigo, M. Corrigendum to "Deep learning improves taphonomic resolution: High accuracy in differentiating tooth marks made by lions and jaguars". *J. R. Soc. Interface* **2020**, *17*, 20200782. [CrossRef]
7. Chollet, F. *Deep Learning with Python*; Manning Publications Company: New York, NY, USA, 2017; p. 361. ISBN 9781617294433.
8. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
9. Mikolajczyk, A.; Grochowski, M. Data augmentation for improving deep learning in image classification problem. In Proceedings of the 2018 International Interdisciplinary PhD Workshop (IIPhDW), Swinoujście, Poland, 9–12 May 2018; pp. 117–122.
10. Zhang, W.; Kinoshita, Y.; Kiya, H. Image-Enhancement-Based Data Augmentation for Improving Deep Learning in Image Classification Problem. In Proceedings of the 2020 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan), Taoyuan, Taiwan, 28–30 September 2020; pp. 1–2.
11. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27, pp. 2672–2680.
12. Langr, J.; Bok, V. *GANs in Action: Deep learning with Generative Adversarial Networks*; Manning Publications Company: New York, NY, USA, 2019; ISBN 9781617295560.
13. Yi, X.; Walia, E.; Babyn, P. Generative adversarial network in medical imaging: A review. *Med. Image Anal.* **2019**, *58*, 101552. [CrossRef]
14. Sun, Y.; Yuan, P.; Sun, Y. MM-GAN: 3D MRI Data Augmentation for Medical Image Segmentation via Generative Adversarial Networks. In Proceedings of the 2020 IEEE International Conference on Knowledge Graph (ICKG), Nanjing, China, 9–11 August 2020; pp. 227–234.
15. Lan, L.; You, L.; Zhang, Z.; Fan, Z.; Zhao, W.; Zeng, N.; Chen, Y.; Zhou, X. Generative Adversarial Networks and Its Applications in Biomedical Informatics. *Front Public Health* **2020**, *8*, 164. [CrossRef] [PubMed]

16. Chang, Q.; Qu, H.; Zhang, Y.; Sabuncu, M.; Chen, C.; Zhang, T.; Metaxas, D.N. Synthetic learning: Learn from distributed asynchronized discriminator gan without sharing medical image data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, DC, USA, 13–19 June 2020; pp. 13856–13866.
17. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; ISBN 9780262337373.
18. Domínguez-Rodrigo, M.; de Juana, S.; Galán, A.B.; Rodríguez, M. A new protocol to differentiate trampling marks from butchery cut marks. *J. Archaeol. Sci.* **2009**, *36*, 2643–2654. [CrossRef]
19. Brownlee, J. *Deep Learning with Python: Develop Deep Learning Models on Theano and TensorFlow Using Keras*; Machine Learning Mastery. 2017. Available online: https://books.google.rs/books/about/Deep_Learning_With_Python.html?id=K-ipDwAAQBAJ&printsec=frontcover&source=kp_read_button&redir_esc=y#v=onepage&q&f=false (accessed on 3 June 2021).
20. Brownlee, J. *Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions*; Machine Learning Mastery. 2018. Available online: https://books.google.rs/books/about/Better_Deep_Learning.html?id=T1-nDwAAQBAJ&printsec=frontcover&source=kp_read_button&redir_esc=y#v=onepage&q&f=false (accessed on 3 June 2021).
21. Eger, S.; Youssef, P.; Gurevych, I. Is it Time to Swish? Comparing Deep Learning Activation Functions Across NLP tasks. *arXiv* **2019**, arXiv:1901.02671.
22. Jinsakul, N.; Tsai, C.-F.; Tsai, C.-E.; Wu, P. Enhancement of Deep Learning in Image Classification Performance Using Xception with the Swish Activation Function for Colorectal Polyp Preliminary Screening. *Sci. China Ser. A Math.* **2019**, *7*, 1170. [CrossRef]
23. Misra, D. Mish: A Self Regularized Non-Monotonic Activation Function. *arXiv* **2019**, arXiv:1908.08681.
24. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
25. Nagarajan, R.; Scutari, M.; Lèbre, S. *Bayesian Networks in R*; Springer: New York, NY, USA, 2013; pp. 122, 125–127.
26. Scutari, M.; Denis, J.-B. *Bayesian Networks: With Examples in R*; CRC Press: Boca Raton, FL, USA, 2014; ISBN 9781482225587.
27. Hong, Y.; Niu, L.; Zhang, J.; Zhao, W.; Fu, C.; Zhang, L. F2GAN: Fusing-and-Filling GAN for Few-shot Image Generation. In *Proceedings of the 28th ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 2535–2543, ISBN 9781450379885.
28. Antoniou, A.; Storkey, A.; Edwards, H. Data Augmentation Generative Adversarial Networks. *arXiv* **2017**, arXiv:1711.04340.
29. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
30. Bourgeon, L.; Burke, A.; Higham, T. Earliest Human Presence in North America Dated to the Last Glacial Maximum: New Radiocarbon Dates from Bluefish Caves, Canada. *PLoS ONE* **2017**, *12*, e0169486. [CrossRef]
31. Gommery, D.; Ramanivosoa, B.; Faure, M.; Guérin, C.; Kerloc'h, P.; Sénégas, F.; Randrianantenaina, H. Les plus anciennes traces d'activités anthropiques de Madagascar sur des ossements d'hippopotames subfossiles d'Anjohibe (Province de Mahajanga). *Comptes Rendus Palevol* **2011**, *10*, 271–278. [CrossRef]
32. Anderson, A.; Clark, G.; Haberle, S.; Higham, T.; Nowak-Kemp, M.; Prendergast, A.; Radimilahy, C.; Rakotozafy, L.M.; Ramilisonina, L.M.; Schwenninger, J.-L.; et al. New evidence of megafaunal bone damage indicates late colonization of Madagascar. *PLoS ONE* **2018**, *13*, e0204368. [CrossRef] [PubMed]
33. Hansford, J.; Wright, P.C.; Rasoamiaramanana, A.; Pérez, V.R.; Godfrey, L.R.; Errickson, D.; Thompson, T.; Turvey, S.T. Early Holocene human presence in Madagascar evidenced by exploitation of avian megafauna. *Sci. Adv.* **2018**, *4*, eaat6925. [CrossRef] [PubMed]
34. Espigares, M.P.; Patrocinio Espigares, M.; Palmqvist, P.; Guerra-Merchán, A.; Ros-Montoya, S.; García-Aguilar, J.M.; Rodríguez-Gómez, G.; Serrano, F.J.; Martínez-Navarro, B. The earliest cut marks of Europe: A discussion on hominin subsistence patterns in the Orce sites (Baza basin, SE Spain). *Sci. Rep.* **2019**, *9*, 1–13. [CrossRef]
35. Wolpert, D.H. The Existence of A Priori Distinctions Between Learning Algorithms. *Neural Comput.* **1996**, *8*, 1391–1420. [CrossRef]