**How many specimens make a sufficient training set for automated 3D feature extraction?**

James M. Mulqueeney1, 2, Alex Searle-Barnes1, Anieke Brombacher1, Marisa Sweeney1, Anjali Goswami2 & Thomas H. G. Ezard1

1Department of Ocean & Earth Science, National Oceanography Centre Southampton, University of Southampton Waterfront Campus, Southampton, UK

2Department of Life Sciences, Natural History Museum, London, UK

**[Abstract]**

*Deep learning has emerged as a robust tool for automating feature extraction from 3D images, offering an efficient alternative to labour-intensive and potentially biased manual image segmentation methods. However, there has been limited exploration into the optimal training set sizes, including assessing whether artificial expansion by data augmentation can achieve consistent results in less time and how consistent these benefits are across different types of traits. In this study, we manually segmented 50 planktonic foraminifera specimens from the genus Menardella to determine the minimum number of training images required to produce accurate volumetric and shape data from internal and external structures. The results reveal unsurprisingly that deep learning models improve with a larger number of training images with eight specimens being required to achieve 95% accuracy. Furthermore, data augmentation can enhance network accuracy by up to 8.0%. Notably, predicting both volumetric and shape measurements for the internal structure poses a greater challenge compared to the external structure, due to low contrast between different materials and increased geometric complexity. These results provide novel insight into optimal training set sizes for precise image segmentation of diverse traits and highlight the potential of data augmentation for enhancing multivariate feature extraction from 3D images.*

**Subject Areas:** computational biology, Artificial Intelligence

**Keywords:** deep learning, data augmentation, image segmentation, planktonic foraminifera, feature extraction

**Author of Correspondence:** James M. Mulqueeney; email: j.m.mulqueeeney@soton.ac.uk; james.mulqueeney1@nhm.ac.uk

**1. Introduction**

Three-dimensional imaging techniques, such as x-ray micro-computed tomography (micro-CT), have revolutionised the characterisation of both internal and external structures of diverse objects. With the ability to generate high-resolution images, researchers can visualise and quantify intricate 3D features with wide-ranging applications [1, 2]. However, objective extraction of these features remains a major challenge.

Currently, the prevailing approach to extracting 3D features from image data involves manual segmentation, which is a labour-intensive [3] and subjective process that lacks reproducibility [4]. This manual approach limits the study of a large number of samples and the exploration of complex hypotheses. As the acquisition of high-resolution scans has increased steadily [5, 6], there is a pressing need to enhance the efficiency of this important processing step.

Machine learning techniques, particularly deep learning, and convolutional neural networks (CNNs) offer a promising solution for automating image segmentation. These methods can potentially accelerate this processing step to deliver accurate and repeatable results, whilst being accessible to various research fields [7, 8]. However, despite their merits, a fundamental trade-off exists between the quantity of samples necessary for generating accurate neural networks and the time-consuming, subjective nature of manual segmentation for evaluation. The specific number of required samples is likely to vary across datasets and depends on the traits being extracted. Hence, there is an imperative to establish the minimum number of manually segmented specimens needed to train these neural networks and to understand how they vary in different trait extraction scenarios.

Strategies to generate training sets that reduce manual processing while maintaining performance are required. Smart interpolation, whereby pre-segmented slices and the volumetric image data are used to predict segmentation across the entire specimen, is one suggested approach [9, 10], but this remains time-consuming for creating large training sets. Data augmentation, which artificially expands the size of the training set without collecting new data, can mitigate overfitting and bolster the accuracy of CNNs during training [11, 12]. Consequently, data augmentation techniques may allow smaller training sets to achieve equivalent accuracy levels of their larger counterparts.

In this study, we aim to determine the minimum number of images needed to train a neural network to produce segmentation data that is statistically indistinguishable from manually generated data. To do so, we use a dataset of computed tomography (CT) scans of planktonic foraminifera from the genus *Menardella.* We assess the efficacy of each training set in extracting volumetric and shape data for the external calcite and internal chamber space of selected specimens (see figure 1). Additionally, we introduce a novel 3D data augmentation technique to bolster training sets by generating six different orientations of each specimen through rotation. This comparison serves to assess how data augmentation strategies can improve training sets to achieve accurate and efficient 3D feature extraction.

**2. Methods and Materials**

**2.1 Data collection**

50 planktonic foraminifera, comprising 4 *Menardella menardii*, 17 *Menardella limbata*, 18 *Menardella exilis*, and 11 *Menardella pertenuis* specimens, were used in our analyses (electronic supplementary material, figures S1 and S2). The taxonomic classification of these species was established based on the analysis of morphological characteristics observed in their shells. In this context, all species are characterised by lenticular, low trochosprial tests with a prominent keel [13]. Discrimination among these species is achievable, as *M. limbata* can be distinguished from its ancestor, *M. menardii*, by having a greater number of chambers and a smaller umbilicus. Moreover, *M. exilis* and *M. pertenuis* can be discerned from *M. limbata* by their thinner, more polished tests and reduced trochospirality. Furthermore, *M. pertenuis* is identifiable by a thin plate extending over the umbilicus and possessing a greater number of chambers in the final whorl compared to *M. exilis* [13].

The samples containing these individuals and species spanned 5.65 million years ago (Ma) to 2.85 Ma [14] and were collected from the Ceara Rise in the Equatorial Atlantic region at Ocean Drilling Program (ODP) Site 925, which comprised Hole 925B (4°12.248'N, 43°29.349'W), Hole 925C 20 (4°12.256'N, 43°29.349'W), and Hole 925D (4°12.260'N, 43°29.363'W). See Curry et al., [15] for more details. This group was chosen to provide inter- and intraspecific species variation, and to provide contemporary data to test how morphological distinctiveness maps to taxonomic hypotheses [16].

The non-destructive imaging of both internal and external structures of the foraminifera was conducted at the µ-VIS X-ray Imaging Centre, University of Southampton, UK, using a Zeiss Xradia 510 Versa X-ray tomography scanner. Employing a rotational target system, the scanner operated at a voltage of 110 kV and a power of 10 W. Projections were reconstructed using Zeiss Xradia software, resulting in 16-bit greyscale .tiff stacks characterised by a voxel size of 1.75 μm and an average dimension of 992 x 1015 pixels for each 2D slice.

**2.2 Generation of training sets**

We extracted the external calcite and internal cavity spaces from the micro-CT scans of the 50 individuals using manual segmentation within Dragonfly v. 2021.3 (Object Research Systems, Canada). This step took approximately 480 minutes per specimen (24,000 minutes total) and involved the manual labelling of 11,947 2D images. Segmentation data for each specimen were exported as multi-label (3 labels: external, internal, and background) 8-bit multipage .tiff stacks and paired with the original CT image data to allow for training (see figure 2).

The 50 specimens were categorised into three distinct groups (electronic supplementary material, table S1): 20 training image stacks, 10 validation image stacks, and 20 test image stacks. From the training image category, we generated six distinct training sets, varying in size from 1 to 20 specimens (see table 1). These were used to assess the impact of training set size on segmentation accuracy, as determined through a comparative analysis against the validation set (see Section 2.3).

From the initial six training sets, we created six additional training sets through data augmentation using the NumPy library [17] in Python. This augmentation method was chosen for its simplicity and accessibility to researchers with limited computational expertise, as it can be easily implemented using a straightforward batch code. This augmentation process entailed rotating the original images five times (the maximum amount permitted using this method), effectively producing six distinct 3D orientations per specimen for each of the original training sets (see figure 3). The augmented training sets comprised between 6 and 120 .tiff stacks (see table 1).

**2.3 Training the neural networks**

CNNs were trained using the offline version of Biomedisa, which utilises a 3D U-Net architecture [18] – the primary model employed for image segmentation [19], and is optimised using Keras with a TensorFlow backend. We used patches of size 64 x 64 x 64 voxels, which were then scaled to a size of 256 x 256 x 256 voxels. This scaling was performed to improve the network’s ability to capture spatial features and mitigate potential information loss during training. We trained 3 networks for each of the training sets to check the extent of stochastic variation on the results [20].

To train our models in Biomedisa, we used a stochastic gradient descent with a learning rate of 0.01, a decay of 1 × 10-6, momentum of 0.9, and Nesterov momentum enabled. A stride size of 32 pixels and a batch size of 24 samples per epoch were used alongside an automated cropping feature, which has been demonstrated to enhance accuracy [21]. The training of each network was performed on a Tesla V100S-PCIE-32GB graphics card with 30989 MB of available memory. All the analyses and training procedures were conducted on the High-Performance Computing (HPC) system at the Natural History Museum, London.

To measure network accuracy, we used the Dice similarity coefficient (Dice score), a metric commonly used in used in biomedical image segmentation studies [22, 23]. The Dice score quantifies the level of overlap between two segmentations, providing a value between 0 (no overlap) and 1 (perfect match). For two segmentations, and consisting of n labels, the Dice score is defined as:

Where and are the total number of voxels of each segmentation, respectively, and is the subset of voxels of with label *i*.

We conducted experiments to evaluate the potential efficiency gains of using an early stopping mechanism within Biomedisa. After testing a variety of epoch limits, we opted for an early stopping criterion set at 25 epochs, which was found to be the lowest value as to which all models trained correctly for every training set. By “trained correctly” we mean if there is no increase in Dice score within a 25-epoch window, the optimal network is selected, and training is terminated. To gauge its impact of early stopping on network accuracy, we compared the results obtained from the original six training sets under early stopping to those obtained on a full run of 200 epochs.

**2.4 Evaluation of feature extraction**

We used the median accuracy network from each of the 12 training sets to produce segmentation data for the external and internal structures of the 20 test specimens. The median accuracy was selected as it provides a more robust estimate of performance by ensuring that outliers had less impact on the overall result. We then compared the volumetric and shape measurements from the manual data to those from each training set. The volumetric measurements were total volume (comprising both external and internal volumes) and percentage calcite (calculated as the ratio of external volume to internal volume, multiplied by 100).

To compare shape, mesh data for the external and internal structures was generated from the segmentation data of the 12 training sets and the manual data. Meshes were decimated to 50,000 faces and smoothed before being scaled and aligned using Python and Generalized Procrustes Surface Analysis (GPSA) [24], respectively. Shape was then analysed using the landmark-free morphometry pipeline, as outlined by Toussaint et al., [25]. We used a kernel width of 0.1mm and noise parameter of 1.0 for both the analysis of shape for both the external and internal data, using a Keops kernel (PyKeops; <https://pypi.org/project/pykeops/>) as it performs better with large data [25]. The analyses were run for 150 iterations, using an initial step size of 0.01. The manually generated mesh for the individual st049\_bl1\_fo2 was used as the atlas for both the external and internal shape comparisons.

**3. Results**

**3.1 Early stopping vs. 200 epochs**

A likelihood ratio test found no detectable difference between the observed correlation of the six original training sets under early stopping within 25 epochs and the full run of 200 epochs (Analysis of Variance, F4, 5 = 0.0424, p > 0.05), with a strong correlation between the early stopping and full run values (see figure 4; R2 = 0.9997, p < 0.001). Consequently, we report results using the early stopping criterion for all subsequent tests.

**3.2 Network accuracy**

The Dice scores of both the original and augmented datasets increased with the number of training images (figure 5; electronic supplementary material, table S2). The improvement in accuracy from 1 specimen to 20 specimens was 12.34% and 4.67% for the original and augmented datasets, respectively. Most of this improvement occurred between training sets of 1 to 8 specimens. Dice scores varied across all training sets (Analysis of Variance, F11, 24 = 92.84, p < 0.001). The augmented training data resulted in higher mean Dice scores compared to their original counterparts, especially at lower specimen numbers (figure 5). Increasing specimen numbers increased the Dice score for the augmented data (*β* = 0.314 on the scale of the logit link function from a generalised linear model with a quasibinomial error structure, s.e. = 0.027, p < 0.001), but did so faster for the original data (*β* = 0.211 on the scale of the logit link function, s.e. = 0.031, p < 0.001) albeit from a much lower initial baseline (*β* =- 0.633 on the scale of the logit link function, s.e. = 0.0059, p < 0.001). The model explains 96% of the variation in the data.

**3.3 Volumetric comparison**

We identified a significant overall correlation between the manually generated and network-generated total volumes for the 20 test images (F 1, 216 = 51223, p < 0.001; figure 6; electronic supplementary material, table S3), but uncovered significant differences in the degree of these correlations across training sets (F 11, 216 = 19.660, p < 0.001). Here, the R2 values, serving as indicators of goodness of fit, spanned from 0.9505 for the training set with 1 specimen to 0.9998 for that containing 20 specimens with data augmentation - a difference of 0.0493.

For percentage calcite, the correlation between the manually generated and network-generated values was also significant (F1, 216 = 1093.2, p < 0.001), and again significant differences were noted in the degree of correlation across training sets (F11, 216 = 21.367, p < 0.001). Notably, the deviation in the gradients from a perfect fit were more pronounced for the lower training for percentage calcite than for the total volumes. This difference was reflected in the R2 values, which ranged from 0.0608 for the training set with 1 specimen to 0.9607 for the 20-specimen set with image augmentation – a much greater difference of 0.8999.

**3.4 Shape comparison**

We expanded the analysis to compare shape estimates between the manually derived data and the six training sets for both external and internal structures. In total, 16,530 control points were generated for the external structure, while 17,325 control points were generated for the internal structure. These data points were subsequently reduced to five principal axes which display 100% of the total variation through non-linear Kernel principal component analysis (kPCA) [26], using 1000 iterations. This dimensionality reduction technique works by allowing separation of nonlinear data by making use of kernels and projecting data into higher dimensional space where it becomes linearly separable.

Both the internal and external shape data exhibited strong and statistically significant correlations between the manually generated and network-generated shape values across all three axes, indicating a consistency in their shape estimates (figure 7; table 2; electronic supplementary material, table S4 and S5). In the analysis of the external structure the high correlation between the manually and network-generated across all training sets was maintained (table 2). When performing kPCA on the external shape results we found that PC1 controlled 36.4% of the total variation, followed by PC2 with 19.2%, and PC3 with 17.9%, resulting in the first three axes controlling 73.5% of the total variation (electronic supplementary material, table S6). In contrast, the internal structures displayed substantial variations across the training sets as the number of images used increased from 1 to 20 across all three axes (table 2), indicating differences amongst the accuracy of different training sets. In this internal analysis, PC1 accounted for 37.6% of the total variation, followed by PC2 with 19.7%, and PC3 with 18.1%, resulting in the first three axes controlling 75.4% of the total variation (electronic supplementary material, table S7). Due to the high proportion of variance controlled by these PC axes we use these for our primary comparisons.

**4. Discussion**

In this study, we observed a positive correlation between network accuracy and the quantity of training images, with data augmentation proving an effective tool for enhancing performance, especially for smaller training sets. External structures were extracted relatively accurately in the smaller training sets, unlike the internal structures, and size was more straightforward to extract than shape. Noting that some traits will always be more difficult to extract accurately than others, we discuss key concepts for the ongoing implementation of automated feature extraction in the biological sciences.

**4.1 Impact of early stopping**

The early stopping feature substantially reduced training time whilst effectively preserving network accuracy (electronic supplementary material, table S2). These qualities made the early stopping feature particularly valuable for larger training sets, and when applying data augmentation techniques.

Beyond improving time efficiency, the integration of early stopping features also serves to mitigate generalisation errors and overfitting [27, 28]. As a result, the application of early stopping features is widely applied when training deep learning models [29, 30]. Our findings strongly endorse the use of these early stopping features in future research using 3D image segmentation.

**4.2 Impacts of training set size and data augmentation**

Our findings reaffirm the principle that expanding the training set leads to the production of better deep learning models [31, 32], albeit with diminishing returns as accuracy approaches 100% [33]. The expansion of available training data plays a crucial role in reducing generalisation error and thus, in facilitating the achievement of optimal accuracy levels [34]. This is reflected in the stability of accuracy observed across epochs in larger training datasets, which are less affected by noise (electronic supplementary material, figures S3-14).

Importantly, the minimum number of training images to achieve accurate results and the degree to which increasing the training set size enhances network accuracy remains task-specific and context-dependent [35]. The choice of the network architecture may vary in its suitability for segmenting specific material or structures [36] meaning training sets must carefully be selected. Notably, in this study, only a small number of individuals were required to achieve high accuracy scores, with eight specimens being required to achieve the goal of 95% accuracy (with and without augmentation) when comparing the manually and network generated data. This achievement can be attributed to the wealth of 2D slices per specimen and the ready availability of high-quality segmentation data. Without these attributes, a larger number of individuals for training would likely be necessary, resulting in delayed increases in accuracy. Consequently, the generation of high-quality training data emerges as a factor more crucial than its sheer quantity [37], with any errors or inconsistencies within the training set likely to manifest in the deep learning models.

Data augmentation is a valuable tool for enhancing network accuracy, addressing limitations stemming from insufficient training data availability. Our results strongly support that data augmentation can effectively boost model accuracy [38, 39]. Although variable by feature, we have demonstrated that it is possible to train a network with as few as one or two specimens to extract segmentation data almost as accurately as 8 specimens using this augmentation approach. As a result, fewer images are needed for training and thus we can significantly reduce the time spent manually segmenting. Moreover, its application in this context is straightforward and universally applicable to 3D image datasets.

**4.3 Extracting phenotypic traits**

The methodology presented in this study represents a significant advancement in the extraction of phenotypic traits from organisms such as planktonic foraminifera, offering notable improvements in processing times and repeatability compared to other established methods [40, 41]. Our approach exhibits remarkable efficiency, culminating in the generation of results for a single specimen in approximately a minute, thus facilitating the execution of large-scale studies with unprecedented speed. Moreover, our methodology excels in repeatability, as it mitigates the impact of fluctuations in thresholding or grayscale values, ensuring more consistent and reliable data extraction. Crucially, our approach extends the scope of trait extraction beyond percentage calcite, encompassing additional parameters such as overall size (internal + external volumes) and shape, which remain elusive with alternative methodologies. This multifaceted advancement underscores the versatility and efficacy of our approach in unravelling the intricacies of foraminiferal morphology.

Our results suggest extracting the internal structure of planktonic foraminifera poses a greater challenge compared to extraction of the external structure, evidenced in the measures of percentage calcite and the shape of the internal structure (see figures 5b, 6). This discrepancy in difficulty can likely be attributed to material homogeneity, which is reflected in low contrast differences within CT scans [42, 43]. This challenge is particularly prominent in the case of planktonic foraminifera, where the internal structure contains sedimentary infill and nannofossil ooze with densities similar to external calcite [44, 45]. The quantity of this infill is likely to vary among specimens, ranging from abundant to absent, while the presence of external calcite remains consistently detectable. Consequently, to alleviate the impact of irrelevant features, such as sediment infill, on trait extraction, we need training sets that encompass a broader array of images, representing a fuller sample of the entire image population compared to datasets without these elements. This can be achieved by either increasing the size of the training set or employing data augmentation techniques, thereby facilitating more accurate predictions.

Further enhancements in the accuracy of selected networks and mitigations of irrelevant features can be achieved through the application of post-processing tools. These tools facilitate the removal of connected components and the application of smoothing methods to eliminate noise from deep learning outputs [46, 47], both of which hold the potential to yield more precise segmentation data.

**4.4 Implications for future studies**

There is a growing demand in the field of phenomics for the integration of machine learning techniques to perform image segmentation, driven by the increasing abundance of available data [48]. Unlike manual segmentation, which is inherently prone to observer bias [49], deep learning methods offer a robust means to ensure greater consistency across measurements [50], facilitating efficient and scalable comparative studies. The complexity and size of phylogenies [51] and birth-death models [52] are only getting larger, with the use of continuous traits [53] and geometric morphometric approaches [54] now commonplace. Thus, there is a need to develop morphometric approaches that can keep pace with these advances.

The application of the automated segmentation and the data augmentation techniques applied here can be used universally to improve data acquisition. They integrate seamlessly with widely used segmentation software like Dragonfly, Avizo/Amira, ImageJ, and MITK, so can be incorporated into existing workflows. They also accommodate various input sources, encompassing images obtained through X-ray synchrotron microscopy [55], soft microscopy [56], and magnetic resonance imaging (MRI) [57]. This versatility empowers the workflow to conduct large-scale analyses efficiently and with a high level of accuracy within reasonable timeframes. With the provision of suitable input data [58], these trained networks become accessible to anyone for segmentation, regardless of their expertise in the morphology of the selected specimen.

The ability to extract different 3D features using deep learning remains likely to be context specific, with a variety of features such as the presence of artefacts and irregularities [59], broken or incomplete specimens [60], and low contrast images [47] each requiring individual solutions to overcome. Improvement in the quality of image data acquisition and selection of specimens can aid in mitigating some of these factors [6], however, these must coincide with high-quality training sets.

The degree of tolerable error is also influenced by the quantification required. Here, our results demonstrate that extracting the total volume was easier than percentage calcite and shape. The variation in total volume is much greater and is less complicated to measure, and thus is less confounded by inaccurate measurements than more plastic traits. The landmark-free method applied here, for instance, requires the generation of meshes, and scaling and alignment of the data which further produces error. Researchers should thus consider their intended measurements when building training data and then during the actual training of their neural networks.

**Conclusion**

We provide compelling evidence that leveraging deep learning for automated segmentation yields equivalent results to manual segmentation when extracting 3D features from images and also drastically reduces the time required. Data augmentation emerged as a powerful tool, elevating network accuracy and enhancing the efficacy of smaller training sets. However, it is imperative to acknowledge that the accuracy of feature extraction is contingent on the specific traits targeted, with eight specimens being required here to extract all of the desired traits within levels of 95% accuracy in this case study. The observed differences for internal and external traits extracted underscores the crucial need for thoughtful consideration when training deep learning networks for specific applications.

**Data accessibility.** All computer code and final raw data used for this research work are stored in GitHub: <https://github.com/JamesMulqueeney/Automated-3D-Feature-Extraction> and have been archived within the Zenodo repository: <https://doi.org/10.5281/zenodo.11109348>. All other data is available in the Dryad Repository: <https://doi.org/10.5061/dryad.1rn8pk12f>.

**Authors’ contributions.** J.M.M, T.H.G.E and A.G designed the concept of the study. A.B performed all the taxonomic classifications and picking of the foraminiferal specimens. A.S.B performed the computer tomography scanning of all specimens. J.M.M, A.S.B and M.S performed manual image segmentation across the 50 specimens. J.M.M trained the AI models and performed the volumetric and shape comparisons. J.M.M wrote the initial draft manuscript, which was iterated with A.G. and T.H.G.E. All authors provided ideas and discussion to the paper and have read and approved the final version.

**Competing interests.** We declare we have no competing interests.

**Funding.** J.M.M was supported by the Natural Environmental Research Council [grant number NE/S007210/1]. J.M.M, A.S.B, A.B, M.S and T.H.G.E were also supported through the Natural Environmental Research Council [grant number NE/P019269/1]. A.G was supported by Leverhulme Trust grant RPG-2021-424 and European Research Council grant STG-2014-637171.

**Acknowledgments.** The authors would like to thank Philipp D. Lösel whose advice on performing the AI network training was invaluable, and all the other members of the Goswami and Ezard Labs who generously provided ideas for the concept of the paper. The authors acknowledge the µ-VIS X-ray Imaging Centre at the University of Southampton for the provision of the X-ray tomographic imaging facilities. We are also grateful for the thoughtful comments from the two anonymous reviewers and the editors.

**References**

[1] Speijer, R.P., Van Loo, D., Masschaele, B., Vlassenbroeck, J., Cnudde, V. & Jacobs, P. 2008 Quantifying foraminiferal growth with high-resolution X-ray computed tomography: New opportunities in foraminiferal ontogeny, phylogeny, and paleoceanographic applications. *Geosphere* **4**, 760-763. (doi:<https://doi.org/10.1130/GES00176.1>).

[2] Cnudde, V. & Boone, M.N. 2013 High-resolution X-ray computed tomography in geosciences: A review of the current technology and applications. *Earth-Science Reviews* **123**, 1-17. (doi:<https://doi.org/10.1016/j.earscirev.2013.04.003>).

[3] Hughes, E.C., Edwards, D.P., Bright, J.A., Capp, E.J., Cooney, C.R., Varley, Z.K. & Thomas, G.H. 2022 Global biogeographic patterns of avian morphological diversity. *Ecology Letters* **25**, 598-610. (doi:<https://doi.org/10.1111/ele.13905>).

[4] Schmelzle, S., Heethoff, M., Heuveline, V., Lösel, P., Becker, J., Beckmann, F., Schluenzen, F., Hammel, J.U., Kopmann, A. & Mexner, W. 2017 The NOVA project: maximizing beam time efficiency through synergistic analyses of SRμCT data. In *Developments in X-Ray Tomography XI* (pp. 100-116, SPIE.

[5] Maire, E. & Withers, P.J. 2014 Quantitative X-ray tomography. *International materials reviews* **59**, 1-43. (doi:<https://doi.org/10.1179/1743280413Y.0000000023>).

[6] Withers, P.J., Bouman, C., Carmignato, S., Cnudde, V., Grimaldi, D., Hagen, C.K., Maire, E., Manley, M., Du Plessis, A. & Stock, S.R. 2021 X-ray computed tomography. *Nature Reviews Methods Primers* **1**, 18. (doi:<https://doi.org/10.1038/s43586-021-00015-4>).

[7] Yang, X., Yu, L., Li, S., Wang, X., Wang, N., Qin, J., Ni, D. & Heng, P.-A. 2017 Towards automatic semantic segmentation in volumetric ultrasound. (pp. 711-719, Springer.

[8] Yan, K., Bagheri, M. & Summers, R.M. 2018 3D context enhanced region-based convolutional neural network for end-to-end lesion detection. (pp. 511-519, Springer.

[9] Van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E. & Yu, T. 2014 scikit-image: image processing in Python. *PeerJ* **2**, e453. (doi:<https://doi.org/10.7717/peerj.453>).

[10] Lösel, P.D., van de Kamp, T., Jayme, A., Ershov, A., Faragó, T., Pichler, O., Jerome, N.T., Aadepu, N., Bremer, S. & Chilingaryan, S.A. 2020 Introducing Biomedisa as an open-source online platform for biomedical image segmentation. *Nature communications* **11**, 1-14. (doi:<https://doi.org/10.1038/s41467-020-19303-w>).

[11] Krizhevsky, A., Sutskever, I. & Hinton, G.E. 2017 Imagenet classification with deep convolutional neural networks. *Communications of the ACM* **60**, 84-90. (doi:<https://doi.org/10.1145/3065386>).

[12] Taqi, A.M., Awad, A., Al-Azzo, F. & Milanova, M. 2018 The impact of multi-optimizers and data augmentation on TensorFlow convolutional neural network performance. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (pp. 140-145, IEEE.

[13] Kennett, J. & Srinivasan, M. 1983 Neogene Planktonic Foraminifera, A Phylogenetic Atlas: Hutchinson Ross Pub. *Co., Stroudsburg, Pennsylvania* **265**.

[14] Wilkens, R.H., Westerhold, T., Drury, A.J., Lyle, M., Gorgas, T. & Tian, J. 2017 Revisiting the Ceara Rise, equatorial Atlantic Ocean: isotope stratigraphy of ODP Leg 154 from 0 to 5 Ma. *Climate of the Past* **13**, 779-793. (doi:<https://doi.org/10.5194/cp-13-779-2017>).

[15] Curry, W.B., Shackleton, N.J. & Richter, C. 1995 Leg 154. *Synthesis. Proceedings ODP, Initial Reports* **154**, 421-442.

[16] Pearson, P.N. & Ezard, T.H.G. 2014 Evolution and speciation in the Eocene planktonic foraminifer Turborotalia. *Paleobiology* **40**, 130-143. (doi:<https://doi.org/10.1666/13004>).

[17] Van Der Walt, S., Colbert, S.C. & Varoquaux, G. 2011 The NumPy array: a structure for efficient numerical computation. *Computing in science & engineering* **13**, 22-30. (doi:<https://doi.org/10.1109/MCSE.2011.37>).

[18] Ronneberger, O., Fischer, P. & Brox, T. 2015 U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18* (pp. 234-241, Springer.

[19] Siddique, N., Paheding, S., Elkin, C.P. & Devabhaktuni, V. 2021 U-net and its variants for medical image segmentation: A review of theory and applications. *Ieee Access* **9**, 82031-82057. (doi:<https://doi.org/10.1109/ACCESS.2021.3086020>).

[20] Desai, A.D., Gold, G.E., Hargreaves, B.A. & Chaudhari, A.S. 2019 Technical considerations for semantic segmentation in MRI using convolutional neural networks. *arXiv preprint arXiv:1902.01977*. (doi:

<https://doi.org/10.48550/arXiv.1902.01977>).

[21] Lösel, P.D., Monchanin, C., Lebrun, R., Jayme, A., Relle, J., Devaud, J.-M., Heuveline, V. & Lihoreau, M. 2022 Natural variability in bee brain size and symmetry revealed by micro-CT imaging and deep learning. *bioRxiv*, 2022.2010. 2012.511944. (doi:<https://doi.org/10.1371/journal.pcbi.1011529>).

[22] Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P. & Carass, A. 2018 Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature communications* **9**, 5217. (doi:<https://doi.org/10.1038/s41467-018-07619-7>).

[23] Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R. & Blaschko, M.B. 2019 Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22* (pp. 92-100, Springer.

[24] Pomidor, B.J., Makedonska, J. & Slice, D.E. 2016 A landmark-free method for three-dimensional shape analysis. *PloS one* **11**, e0150368. (doi:<https://doi.org/10.1371/journal.pone.0150368>).

[25] Toussaint, N., Redhead, Y., Vidal-García, M., Lo Vercio, L., Liu, W., Fisher, E.M.C., Hallgrímsson, B., Tybulewicz, V.L.J., Schnabel, J.A. & Green, J.B.A. 2021 A landmark-free morphometrics pipeline for high-resolution phenotyping: application to a mouse model of Down syndrome. *Development* **148**, dev188631. (doi:<https://doi.org/10.1242/dev.188631>).

[26] Schölkopf, B., Smola, A. & Müller, K.-R. 1997 Kernel principal component analysis. In *International conference on artificial neural networks* (pp. 583-588, Springer.

[27] Caruana, R., Lawrence, S. & Giles, C. 2000 Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Advances in neural information processing systems* **13**.

[28] Zhang, T. & Yu, B. 2005 Boosting with early stopping: Convergence and consistency. (doi:<https://doi.org/10.1214/009053605000000255>).

[29] Zhang, Y.-D., Hou, X.-X., Chen, Y., Chen, H., Yang, M., Yang, J. & Wang, S.-H. 2018 Voxelwise detection of cerebral microbleed in CADASIL patients by leaky rectified linear unit and early stopping. *Multimedia Tools and Applications* **77**, 21825-21845. (doi:<https://doi.org/10.1007/s11042-017-4383-9>).

[30] Amyar, A., Modzelewski, R., Li, H. & Ruan, S. 2020 Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation. *Computers in biology and medicine* **126**, 104037. (doi:<https://doi.org/10.1016/j.compbiomed.2020.104037>).

[31] Bardis, M., Houshyar, R., Chantaduly, C., Ushinsky, A., Glavis-Bloom, J., Shaver, M., Chow, D., Uchio, E. & Chang, P. 2020 Deep learning with limited data: Organ segmentation performance by U-Net. *Electronics* **9**, 1199. (doi: <https://doi.org/10.3390/electronics9081199>).

[32] Narayana, P.A., Coronado, I., Sujit, S.J., Wolinsky, J.S., Lublin, F.D. & Gabr, R.E. 2020 Deep‐learning‐based neural tissue segmentation of MRI in multiple sclerosis: effect of training set size. *Journal of Magnetic Resonance Imaging* **51**, 1487-1496. (doi:<https://doi.org/10.1002/jmri.26959>).

[33] Kavzoglu, T. 2009 Increasing the accuracy of neural network classification using refined training data. *Environmental Modelling & Software* **24**, 850-858. (doi:<https://doi.org/10.1016/j.envsoft.2008.11.012>).

[34] Ying, X. 2019 An overview of overfitting and its solutions. In *Journal of physics: Conference series* (p. 022022, IOP Publishing.

[35] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. 2014 Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **15**, 1929-1958.

[36] Fang, Y., Wang, J., Ou, X., Ying, H., Hu, C., Zhang, Z. & Hu, W. 2021 The impact of training sample size on deep learning-based organ auto-segmentation for head-and-neck patients. *Physics in Medicine & Biology* **66**, 185012. (doi:<https://doi.org/10.1088/1361-6560/ac2206>).

[37] Nikolados, E.-M., Wongprommoon, A., Aodha, O.M., Cambray, G. & Oyarzún, D.A. 2022 Accuracy and data efficiency in deep learning models of protein expression. *Nature Communications* **13**, 7755. (doi:<https://doi.org/10.1038/s41467-022-34902-5>).

[38] Pereira, S., Pinto, A., Alves, V. & Silva, C.A. 2016 Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE transactions on medical imaging* **35**, 1240-1251. (doi:<https://doi.org/10.1109/TMI.2016.2538465>).

[39] Wang, G., Li, W., Ourselin, S. & Vercauteren, T. 2019 Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4* (pp. 61-72, Springer.

[40] Todd, C.L., Schmidt, D.N., Robinson, M.M. & De Schepper, S. 2020 Planktic foraminiferal test size and weight response to the late Pliocene environment. *Paleoceanography and Paleoclimatology* **35**, e2019PA003738. (doi:<https://doi.org/10.1029/2019PA003738>).

[41] Kimoto, K., Horiuchi, R., Sasaki, O. & Iwashita, T. 2023 Precise bulk density measurement of planktic foraminiferal test by X-ray microcomputed tomography. *Frontiers in Earth Science* **11**, 1184671. (doi:<https://doi.org/10.3389/feart.2023.1184671>).

[42] Furat, O., Wang, M., Neumann, M., Petrich, L., Weber, M., Krill III, C.E. & Schmidt, V. 2019 Machine learning techniques for the segmentation of tomographic image data of functional materials. *Frontiers in Materials* **6**, 145. (doi: <https://doi.org/10.3389/fmats.2019.00145>).

[43] Alathari, T. 2015 Feature extraction in volumetric images, University of Southampton.

[44] Zarkogiannis, S.D., Kontakiotis, G., Gkaniatsa, G., Kuppili, V.S., Marathe, S., Wanelik, K., Lianou, V., Besiou, E., Makri, P. & Antonarakou, A. 2020 An improved cleaning protocol for foraminiferal calcite from unconsolidated core sediments: HyPerCal—a new practice for micropaleontological and paleoclimatic proxies. *Journal of Marine Science and Engineering* **8**, 998. (doi:<https://doi.org/10.3390/jmse8120998>).

[45] Zarkogiannis, S.D., Antonarakou, A., Fernandez, V., Mortyn, P.G., Kontakiotis, G., Drinia, H. & Greaves, M. 2020 Evidence of stable foraminifera biomineralization during the last two climate cycles in the tropical Atlantic Ocean. *Journal of Marine Science and Engineering* **8**, 737. (doi:<https://doi.org/10.3390/jmse8100737>).

[46] Larrazabal, A.J., Martinez, C. & Ferrante, E. 2019 Anatomical priors for image segmentation via post-processing with denoising autoencoders. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22* (pp. 585-593, Springer.

[47] Tuladhar, A., Schimert, S., Rajashekar, D., Kniep, H.C., Fiehler, J. & Forkert, N.D. 2020 Automatic segmentation of stroke lesions in non-contrast computed tomography datasets with convolutional neural networks. *IEEE Access* **8**, 94871-94879. (doi:<https://doi.org/10.1109/ACCESS.2020.2995632>).

[48] Lürig, M.D., Donoughe, S., Svensson, E.I., Porto, A. & Tsuboi, M. 2021 Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology. *Frontiers in Ecology and Evolution* **9**, 148. (doi:<https://doi.org/10.3389/fevo.2021.642774>).

[49] Joskowicz, L., Cohen, D., Caplan, N. & Sosna, J. 2019 Inter-observer variability of manual contour delineation of structures in CT. *European radiology* **29**, 1391-1399. (doi:<https://doi.org/10.1007/s00330-018-5695-5>).

[50] Willers, C., Bauman, G., Andermatt, S., Santini, F., Sandkühler, R., Ramsey, K.A., Cattin, P.C., Bieri, O., Pusterla, O. & Latzin, P. 2021 The impact of segmentation on whole‐lung functional MRI quantification: repeatability and reproducibility from multiple human observers and an artificial neural network. *Magnetic resonance in medicine* **85**, 1079-1092. (doi:<https://doi.org/10.1002/mrm.28476>).

[51] Vernygora, O.V., Simões, T.R. & Campbell, E.O. 2020 Evaluating the performance of probabilistic algorithms for phylogenetic analysis of big morphological datasets: a simulation study. *Systematic Biology* **69**, 1088-1105. (doi:<https://doi.org/10.1093/sysbio/syaa020>).

[52] Šmíd, J. & Tolley, K.A. 2019 Calibrating the tree of vipers under the fossilized birth-death model. *Scientific Reports* **9**, 5510. (doi:<https://doi.org/10.1038/s41598-019-41290-2>).

[53] Parins-Fukuchi, C. 2018 Use of continuous traits can improve morphological phylogenetics. *Systematic Biology* **67**, 328-339. (doi:<https://doi.org/10.1093/sysbio/syx072>).

[54] Smith, U.E. & Hendricks, J.R. 2013 Geometric morphometric character suites as phylogenetic data: extracting phylogenetic signal from gastropod shells. *Systematic biology* **62**, 366-385. (doi:<https://doi.org/10.1093/sysbio/syt002>).

[55] van de Kamp, T., Mikó, I., Staniczek, A.H., Eggs, B., Bajerlein, D., Faragó, T., Hagelstein, L., Hamann, E., Spiecker, R. & Baumbach, T. 2022 Evolution of flexible biting in hyperdiverse parasitoid wasps. *Proceedings of the Royal Society B* **289**, 20212086. (doi:<https://doi.org/10.1098/rspb.2021.2086>).

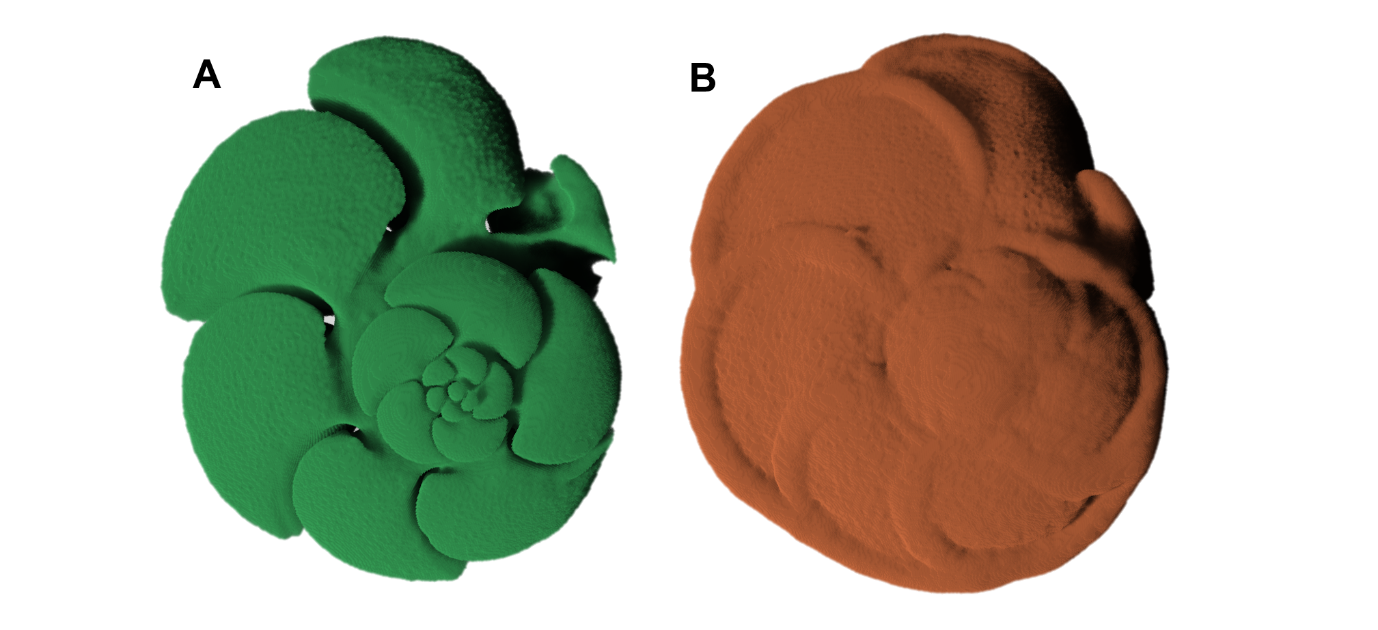
[56] Erozan, A., Lösel, P., Heuveline, V. & Weinhardt, V. 2023 ACSeg: Automated 3D Cytoplasm Segmentation in Soft X-Ray Tomography. *Preprint Series of the Engineering Mathematics and Computing Lab*. (doi:<https://doi.org/10.11588/emclpp.2023.2.94947>).

[57] Lösel, P. & Heuveline, V. 2017 A GPU based diffusion method for whole-heart and great vessel segmentation. In *Reconstruction, Segmentation, and Analysis of Medical Images: First International Workshops, RAMBO 2016 and HVSMR 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 1* (pp. 121-128, Springer.

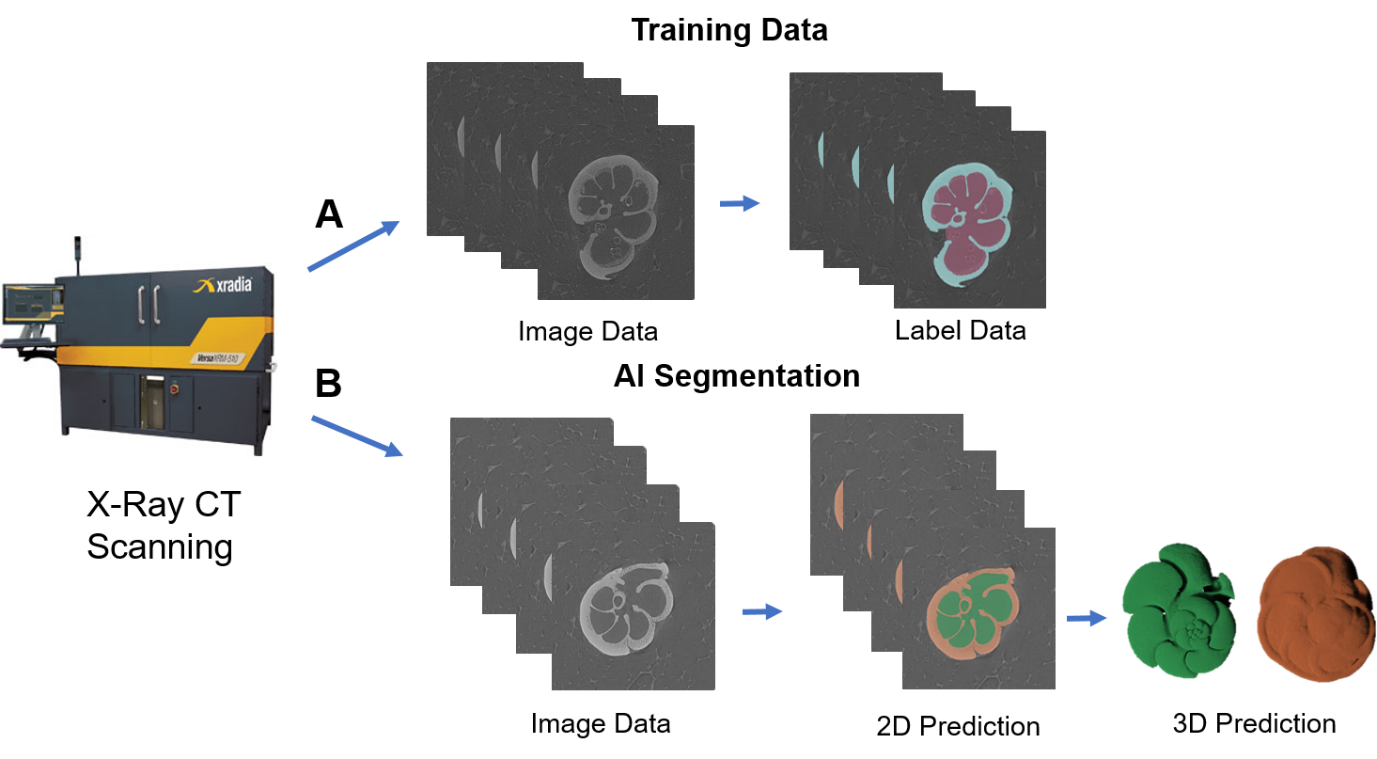
[58] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B. & Sánchez, C.I. 2017 A survey on deep learning in medical image analysis. *Medical image analysis* **42**, 60-88. (doi:<https://doi.org/10.1016/j.media.2017.07.005>).

[59] Das, S., Nayak, G.K., Saba, L., Kalra, M., Suri, J.S. & Saxena, S. 2022 An artificial intelligence framework and its bias for brain tumor segmentation: A narrative review. *Computers in biology and medicine* **143**, 105273. (doi:<https://doi.org/10.1016/j.compbiomed.2022.105273>).

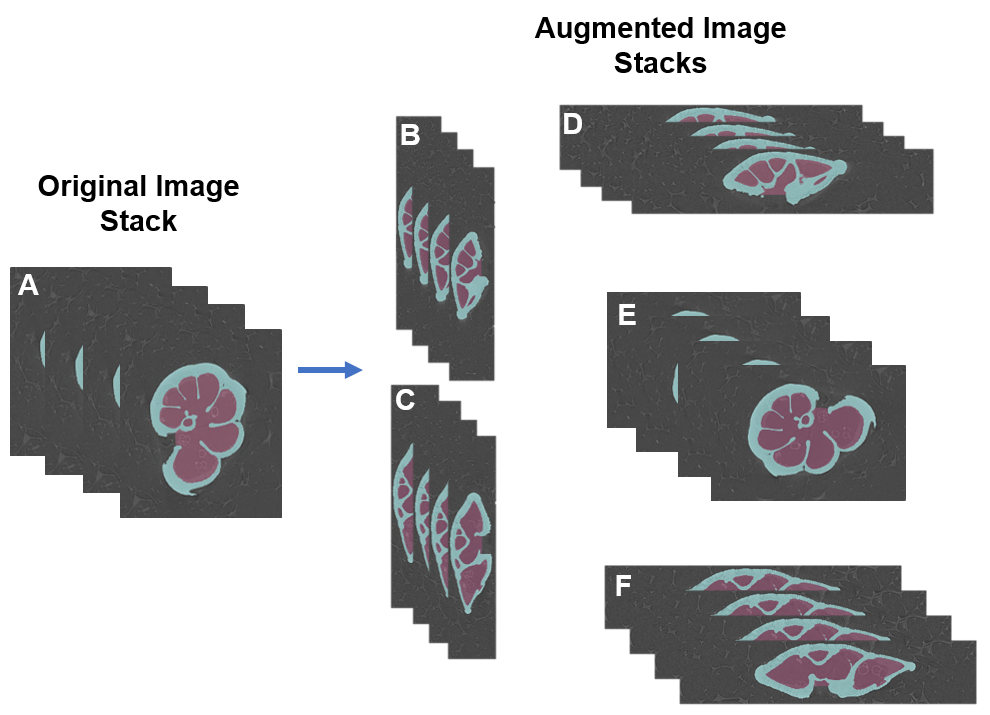
[60] Zhang, W., Ezard, T., Searle-Barnes, A., Brombacher, A., Katsamenis, O. & Nixon, M. 2020 Towards understanding speciation by automated extraction and description of 3D foraminifera stacks. (pp. 30-33, IEEE.

**Figures**

**Figure 1.** 3D models of the (a) internal and (b) external structure of planktonic foraminifera generated using manual segmentation in Dragonfly v. 2021.3 (Object Research Systems, Canada).

****

**Figure 2.** Workflow for producing training data and applying a deep convolutional neural network (CNN) to perform automated image segmentation. The workflow includes (a) the creation of training data for the input into Biomedisa and (b) an example application of the trained CNN to automate the process of generating segmentation (label) data.

****

**Figure 3.** Rotation of the original image to form the new augmentation training data. The original image in an (a) xyz orientation is rotated into 5 other three-dimensional planes: (b) yzx, (c) zyx, (d) xzy, (e) yxz, and (f) zxy orientations. These are then all paired together and used in training.

A graph of a line graph

Description automatically generated with medium confidence

**Figure 4.** Comparison of Dice scores obtained from early stopping within 25 epochs and a full run of 200 epochs for each of the 6 original image training sets shows no statistical difference. The correlation coefficient (R2) between the two sets of scores displayed is 0.9998.

**A graph of a number of specimens

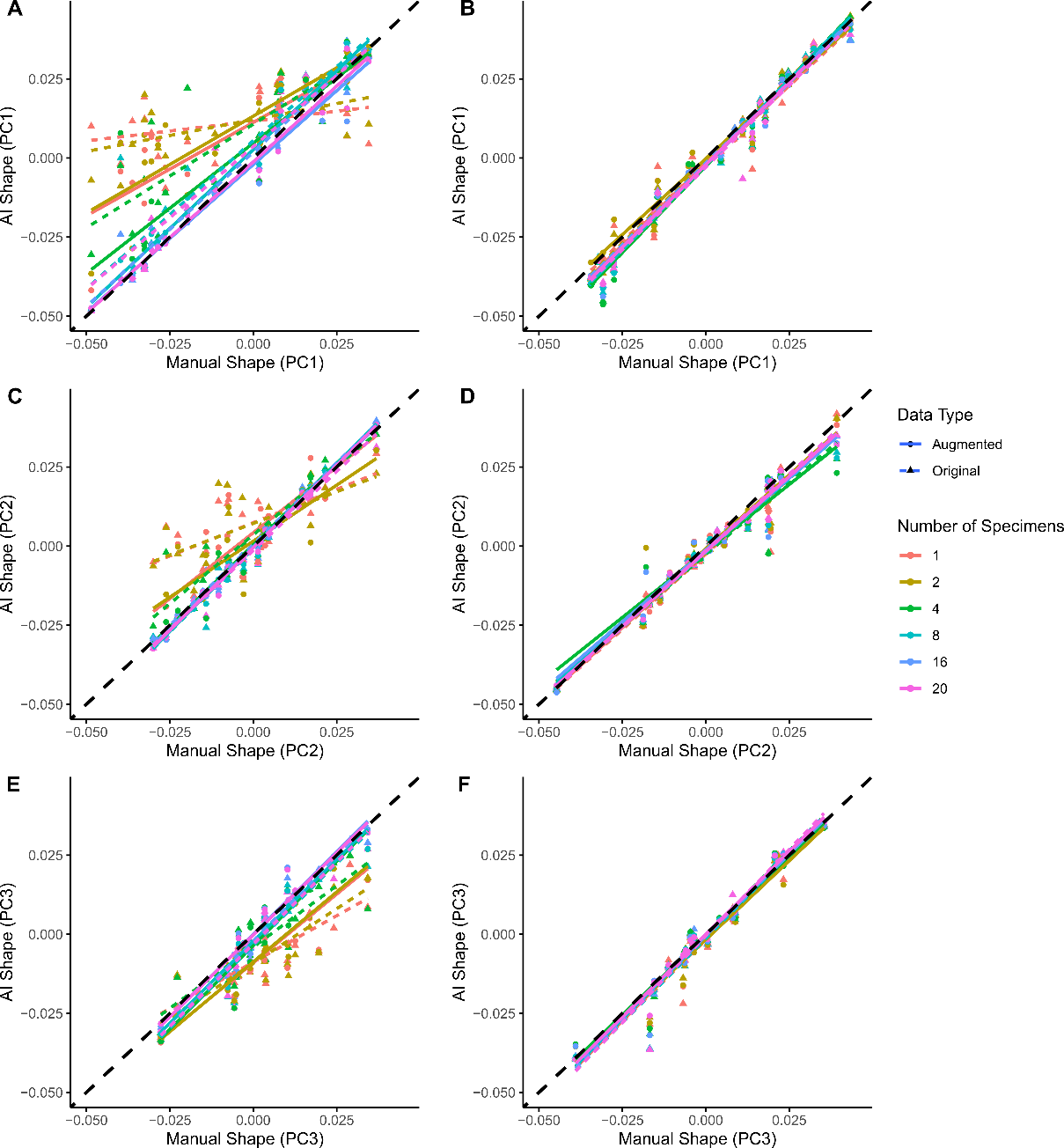
Description automatically generated**

**Figure 5.** Comparison of segmentation accuracy for Biomedisa automated segmentation using average Dice scores calculated for validation data generated from 10 specimens. The plot shows increasing the training set size and implementing data augmentation improves network accuracy. All models were trained using early stopping within 25 epochs.

**A graph of different colored lines

Description automatically generated with medium confidence**

**Figure 6.** Comparison of (a) total volumes (internal and external combined) and (b) percentage calcite obtained from manual segmentation compared to those obtained using AI segmentation using 6 different training sets consisting of different numbers of training images. The predictions of total volume are generally more accurate than percentage calcite.

****

**Figure 7.** Comparison of shape for the internal (a, c and e) and external (b, d and f) structures across PC axes 1 (a, b), 2 (c, d) and 3 (e, f) obtained from manual segmentation compared to those obtained using AI segmentation using 6 different training sets consisting of different numbers of training images. The figure demonstrates that external calcite is easier to predict than the internal chamber space.

**Tables**

**Table 1.** Design of the training sets detailing the name of the training set, the data type (original or new augmentation), the number of specimens in each training set, the number of 2D images available for training and the approximate time to perform the manual segmentation to create the training set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name of Training Set** | **Data Type** | **Number of Specimens** | **No. of foreground Images (XY)** | **Approx. Duration of Segmentation (Mins)** |
| T0\_AI\_1\_Images | Original | 1 | 241 | 480 |
| T1\_AI\_2\_Images | Original | 2 | 390 | 960 |
| T2\_AI\_4\_Images | Original | 4 | 1028 | 1920 |
| T3\_AI\_8\_Images | Original | 8 | 2187 | 3840 |
| T4\_AI\_16\_Images | Original | 16 | 3571 | 7680 |
| T5\_AI\_20\_Images | Original | 20 | 4251 | 9600 |
| Aug\_T0\_AI\_1\_Images | Augmentation | 1 | 4480 | 480 |
| Aug\_T1\_AI\_2\_Images | Augmentation | 2 | 8812 | 960 |
| Aug\_T2\_AI\_4\_Images | Augmentation | 4 | 18120 | 1920 |
| Aug\_T3\_AI\_8\_Images | Augmentation | 8 | 36398 | 3840 |
| Aug\_T4\_AI\_16\_Images | Augmentation | 16 | 71100 | 7680 |
| Aug\_T5\_AI\_20\_Images | Augmentation | 20 | 88432 | 9600 |

**Table 2.** There was no quantitative difference between manually generated and network-generated shape estimates for external and internal structures (left as F statistics on 1 and 216 degrees of freedom), but clear differences in predictive ability between external and internal shape estimates as the numbers of training images increases (right as F statistics on 11 and 216 degrees of freedom). F statistics are presented to 4 significant figures; values in bold indicate p<0.05 whereas non-bold is p>0.05.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | F-statistic between manual and network predictions | | F-statistic among different numbers of training images | |
| Axis | External | Internal | External | Internal |
| PC1 | **7918** | **548.4** | 0.5695 | **5.442** |
| PC2 | **3474** | **1002** | 0.2491 | **3.893** |
| PC3 | **3886** | **1208** | 0.3501 | **6.997** |