Machine Learning for Environmental Sciences

Classifying fossil tracks with neural networks

Yvan Martinet

Université de Lausanne

Professor: Dr Tom Beucler

Source of Cover Image: Microsoft Bing AI

Machine Learning for Earth and Environmental Sciences (2023), **xx**:xx 2–8

Final Report

UNIL | Université de Lausanne

Classifying fossil tracks with neural networks

Yvan Martinet

Faculty of geosciences and environment, Université de Lausanne, Lausanne, 1015, Vaud, Switzerland.

Keywords: Quantitative analysis, Machine learning, Paleontology, Fossil footprints

Abstract
Fossil tracks are pivotal to understanding ancient life, yet their quantitative analysis faces challenges due to variability and undefined features. This study introduces a machine learning approach, employing convolutional neural networks (CNNs) to decode fossil track intricacies. Overcoming past statistical limitations, our CNN differentiates theropod and ornithischian dinosaur tracks using 1216 outline silhouettes. Remarkably, it correctly classifies 87% of the data. Nonetheless, this approach depends heavily on previous expert conducted classifications. Still, it proves to be a useful tool to discriminate between theropod and ornithischian tracks when coupled with human expertise.

Data and code for this study are available at: https://github.com/Yviflex/2023_ML_EES/tree/main/Dinosaur_tracks_CNN

1. Introduction

The persistent challenge of differentiating between herbivorous ornithischians and predominantly carnivorous theropods has captivated the attention of researchers. Conventional distinctions hinge upon the expectation that ornithischian tracks manifest wider and more symmetrical characteristics, encompassing unique traits such as differing digit impressions. However, the exclusivity of these features remains inconsistent, varying across distinct track types. Quantitative methods, encounter limitations, notably sample size constraints and measurement scheme issues, despite widespread application (Lallensack *et al.*, 2022).

In response to these challenges and to mitigate subjectivity, a machine learning approach emerges, employing convolutional neural networks (CNN). This type of machine learning algorithm is well known for its ability to classify images. Inspired by the human brain's structure, the CNN undergoes multiple training iterations, optimizing interconnected nodes for superior accuracy in categorizing images. Prior methodologies, such as linear and angular measurements, though useful, fall short in capturing all relevant details, such as the shape of claw marks, and struggle with consistent application across tracks of varying shapes. To overcome these challenges, black-and-white silhouettes of interpretive track outlines as input images are employed. By utilizing these silhouettes, the need for measuring specific anatomical features is circumvented, allowing for the inclusion of diverse track shapes. Despite criticisms of subjectivity and simplification, these silhouettes are considered effective in minimizing irrelevant information while encapsulating crucial details for track identification, such as digit terminations and the number and shape of phalangeal pads.This departure from traditional approaches may offer a comprehensive and robust solution to the intricate problem of classifying ancient footprints.

2. Dataset

The complexity inherent in published outline drawings, often reflective of the individual ichnologist's artistic signature, is simplified into black-and-white silhouettes to enhance the model's performance. This simplification involves converging complex drawings into continuous outlines or sets of separate outlines, providing a standardized format for analysis.

The dataset encompasses tracks of functionally tridactyl ornithischian and non-avian theropod dinosaurs spanning various geological periods. Notably, the inclusion criteria prioritize tracks with sufficient anatomical detail, while tracks displaying ambiguity or misleading features due to unfavorable substrate conditions or post-formational alterations are hoped to be addressed by training the model on a sufficiently large sample size. The collection of outlines was drawn from various literature and supplemented by Jens N. Lallensack, Anthony Romilio and Peter L. Falkingham with new outlines from three-dimensional models (figure 1). This forms a dataset of 1587 outlines of 100 pixels in height and width, with an imbalance between ornithischian and theropod tracks (Lallensack *et al.*, 2022). To address this, resampling is employed, randomly removing theropod examples to achieve a balanced dataset for training and testing purposes. As a result, 1216 outlines remain. Finally, the data has been split in training, validation and test according to a 0.8/0.1/0.1 pattern. This data preparation lays the foundation for the subsequent application of machine learning algorithms in the classification of tridactyl dinosaur tracks.

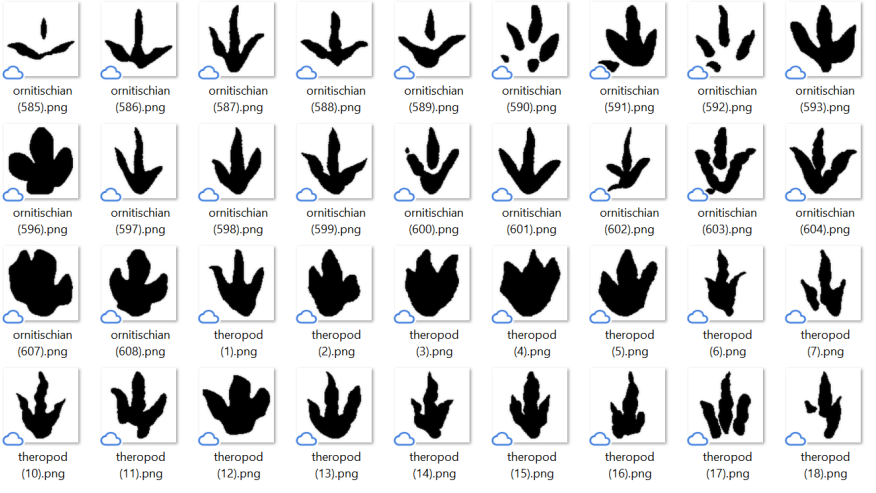

Figure 1. *This is a small sample of the 1216 images used in this research.*

3. Methodology

3.1. Baseline and KNN classification

The model training was done using open-source machinelearning libraries TensorFlow (www.tensorflow.org; version 2.14.0), scikit-learn (www.scikit-learn.org; version 1.3.2) and keras (https://keras.io; version 3.0.1) which were controlled through the Google Colab interface (https://colab.google) using Python scripting. Also, the Shapley Additive Explanations (SHAP) library was used in the end to explain the CNN outputs (https://shap.readthedocs.io/en/latest/).

As mentioned before, convolutional neural networks are models of choice for this research, as they are notorious for their efficiency in image classification. Nonetheless, it is fitting to examine more simple algorithms before jumping to CNNs to make sure they do not achieve better performance. Thus, a logistic regression with a C value of 1 and 100 maximum iterations was first used as a baseline to explore simple

machine learning algorithm performance on our data. Following this step, the data was submitted to a more complex k-nearest neighbor (KNN) classification to assert performance before delving into deep learning. Here, GridSearchCV was used to determine the best hyperparameters with 6 cross-validation folds to evaluate its generalization capacity. Distance based weights and 4 neighbors constituted the best hyperparameters. Also, data augmentation with random shifts was applied to provide our model with more robustness.

3.2. CNN classification

The performance of a model is significantly influenced by both its architecture and the parameters chosen. Achieving precise optimization by finding the best parameters can be difficult due to the multitude of possibilities available. To navigate this challenge, different models are used to conduct tests with different architectures, exploring different numbers of epochs (training iterations where the model sees all the data) and batch sizes (the number of silhouettes the model processes simultaneously). Also, data augmentation is applied on each images in order to provide the model with enhanced robustness. In the model selection process, the use of loss was prioritized over accuracy, discarding models showing overfitting. Following testing, the most promising structure happened to be an architecture featuring four convolutional layers with respectively 8, 16, 32, and 32 neurons (figure 2). Pooling layers were incorporated between these layers, and a dense layer with 256 neurons was employed for flattening. Spatial dropout 2D of 0.1 was applied after each pooling layer and a simple dropout after the 256-neuron dense layer. The 'Adam' optimizer and a batch size of 32 were chosen. Binary Cross Entropy was used as the loss function as it fits our binary classification task.

3.3. Results

In the case of our selected model, this process resulted in a training period of 200 epochs, concluding with a validation loss of 0.28 and a validation accuracy of 0.88. Interestingly, validation loss and accuracy kept constantly decreasing and increasing respectively (figure 3 and 4), indicating that the model struggles to converge. Such behavior led to try adjusting the learning rate, by increasing or decreasing it. However, the outcomes were unsatisfactory. Also, high variance between each epoch can be observed. This is most likely due to the rather limited amount of images used.

For each silhouette under examination, the neural network provided a numerical output ranging from 0 to 1, indicating the confidence level in track affiliations as ornithischian or theropod. A value close to 0 indicates a classification towards the ornithischian class. On the contrary, a value close to 1 shows preference toward the theropod class. Values neighboring 0.5 denote an ambiguous outcome, indicating no clear tendency toward either category. Figure 5 shows clear evidence of our model struggling to classify ambiguous predicted values. With a predicted value of 0.5065, it wrongly classifies an ornithischian into the theropod class.

When providing our model with the 96 outlines from the test set, it reached an accuracy of 0.87, an F1 score of 0.88 and a recall of 0.89. In comparison, our logistic regression resulted in an accuracy of 076, an F1 score of 0.76 and a recall of 0.75. The KNN classification had an accuracy of 0.82, an F1 score of 0.83 and a recall of 0.83. Thus, the use of a CNN to tackle this task proves to be worthwhile. The three confusion matrix of these models help better visualising their differences in performance (figures 6, 7 and 8).

If predicted values ranging between 0.4 and 0.6 are considered to be ambiguous, our model had difficulties to classify 9% of the test set. It correctly classifies 50% of the ambiguous instances. If the rest of the values indicate a high level of confidence of our model, it correctly classifies 90% of the outlines it is confident about.As a benchmark to these results, Lallensack's own study provides interesting results to better assert our model's capabilities (Lallensack *et al.*, 2022). Despite not being evaluated on the same exact test set, the results from five different ichnologists (experts) and Lallensack's model can be used to compare with the results obtained with the present model (Table 1). In this case, human experts

```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 sequential_2 (Sequential)   (None, 100, 100, 3)       0

 conv2d_4 (Conv2D)           (None, 98, 98, 8)         224

 max_pooling2d_4 (MaxPoolin  (None, 49, 49, 8)         0
 g2D)

 dropout_5 (Dropout)         (None, 49, 49, 8)         0

 conv2d_5 (Conv2D)           (None, 47, 47, 16)        1168

 max_pooling2d_5 (MaxPoolin  (None, 23, 23, 16)        0
 g2D)

 dropout_6 (Dropout)         (None, 23, 23, 16)        0

 conv2d_6 (Conv2D)           (None, 21, 21, 32)        4640

 max_pooling2d_6 (MaxPoolin  (None, 10, 10, 32)        0
 g2D)

 dropout_7 (Dropout)         (None, 10, 10, 32)        0

 conv2d_7 (Conv2D)           (None, 8, 8, 64)          18496

 max_pooling2d_7 (MaxPoolin  (None, 4, 4, 64)          0
 g2D)

 dropout_8 (Dropout)         (None, 4, 4, 64)          0

 flatten_1 (Flatten)         (None, 1024)              0

 dense_2 (Dense)             (None, 256)               262400

 dropout_9 (Dropout)         (None, 256)               0

 dense_3 (Dense)             (None, 1)                 257

=================================================================
Total params: 287185 (1.10 MB)
Trainable params: 287185 (1.10 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 2. *This is the architecture used for the convolutional neural network.*

are less accurate in classifying the tracks. Lallensack's and the present CNN achieve very similar results when considering there are no ambiguous cases. The present model however, performs significantly better when considering ambiguous cases. As a result it can be seen as a more trustworthy model.

3.4. XAI clarifications

Using SHAP DeepExplainer allows to better understand how the model decides upon classifying the dinosaur tracks. In figures 9 and 10, blue and red pixels show where low, respectively high, values tend to influence the model's decision towards the actual class. Regarding ornithischian track outlines, the model seems to mostly rely on pixels inside of the silhouette. Thick heels and short toes tend to describe ornithischian footprints. On the other hand, the model seems to rely more on the contours of the track outlines when classifying theropods. Narrow heels and long toes seem to reveal theropod characteristics. Nonetheless, interpretations of such results is highly subjective and the observer may be pushed to see what he wants to see.

4. Limitations

4.1. Absence of context

The convolutional neural network was trained to classify shapes without having access to any context information regarding time, footprint size, environmental conditions, geological composition, etc.. In the assignment of a track, an ichnologist considers vital contextual information, including factors such

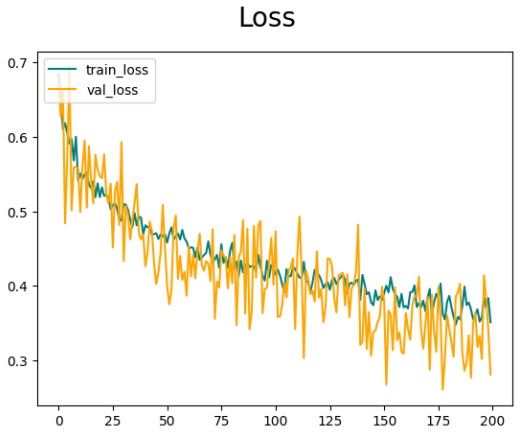

Figure 3. *Training and validation loss over each epoch.*

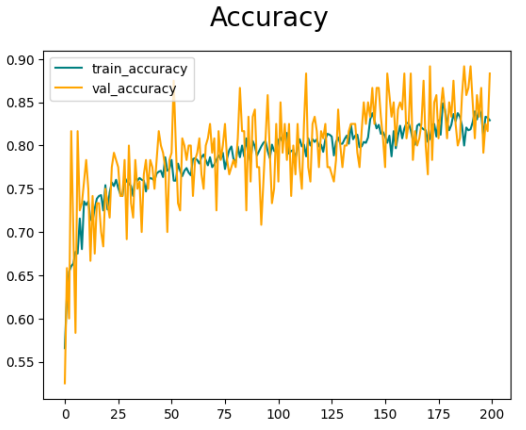

Figure 4. *Training and validation accuracy over each epoch.*

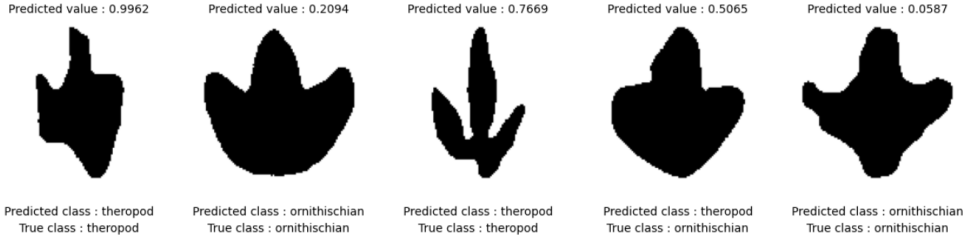

Figure 5. *Examples of dinosaur track CNN classification outputs.*

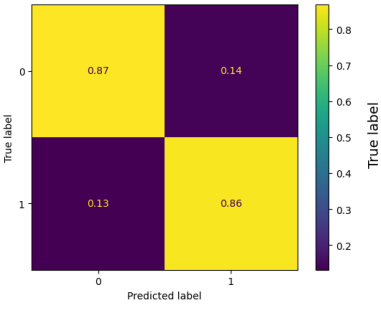

Figure 6. *CNN classification confusion matrix.*

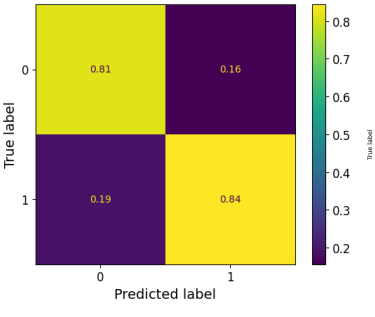

Figure 7. *KNN classification confusion matrix.*

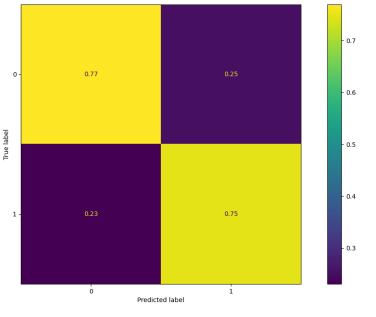

Figure 8. *Logistic regression confusion matrix.*

as size, stratigraphy, and the shape of other tracks within the same trackway. Interestingly, what might seem like a limitation in our neural network could actually be its main strength. It provides unbiased assessments of shape irrespective of context. Therefore, it becomes the responsibility of the ichnologist to integrate the neural network's shape evaluation with all relevant contextual information for a meaningful interpretation of the track (Lallensack *et al.*, 2022).

Identifier	% correct	% unsure	%false
Expert 1	67	3	31
Expert 2	58	25	17
Expert 3	58	25	17
Expert 4	42	44	14
Expert 5	58	22	19
Lallensack CNN (no ambiguity)	86	0	14
Lallensack CNN (ambiguity between 0.4 and 0.6)	67	22	11
Present CNN (no ambiguity)	88	0	12
Present CNN (ambiguity between 0.4 and 0.6)	82	9	9

Table 1. *Results of the experts surveyed by Lallensack et al., his algorithm and ours.*

4.2. Previous scientific classification

An unavoidable constraint in this approach is dependence on prior classifications of tracks as either 'theropod' or 'ornithischian' to train the models. Operating under the assumption that the majority of these initial identifications are likely accurate may be risky. Thus, possibility of misidentifications has to be acknowledged (Lallensack *et al.*, 2022).

4.3. Subjectivity of silhouette outlines

The primary limitations of the method presented stem from the inherent subjectivity involved in interpreting outline silhouettes based off three dimensional objects. Outline drawings done by different researchers may exhibit significant shape variations. The trade-off between simplifying information for a coherent understanding of foot shape and the potential exclusion of valuable extramorphological features poses both an advantage and a drawback. Concerns arise due to the varying degrees of 'idealization' in outlines among different experts, with some asserting that features like displacement rims and collapse structures, often omitted, are integral to the track and carry valuable information about the trackmaker (Lallensack *et al.*, 2022).

5. Opening and conclusion

Fossil tracks are invaluable sources of palaeontological data, but their interpretation poses challenges due to the influence of various factors on their shape. Traditional ichnology relies on interpretive outline drawings, which, though problematic for quantitative analysis, still hold value when executed meticulously. The presented approach advocates for the use of neural networks, demonstrating their potential to surpass human experts in tasks like classification. However, the subjective nature of outline drawings, influenced by assumptions about the trackmaker, introduces risks of circular arguments. Three-dimensional data, collected through methods like photogrammetry, could represent interesting advancement of this method. It could achieve long-term digital preservation and unlocking the full potential of fossil tracks for quantitative analysis (Rejcek, 2022). According to Lallensack, the current shift towards standardized three-dimensional data collection is deemed essential for the future application of advanced quantitative methods, like neural networks, in the field of ichnology (Lallensack *et al.*, 2022).

References

[1] **Géron, A (2010)**Hands-On Machine Learning with Scikit-Learn, Keras, andTensorFlow, *O'Reilly Media, Inc. ISBN: 9781492032649.*

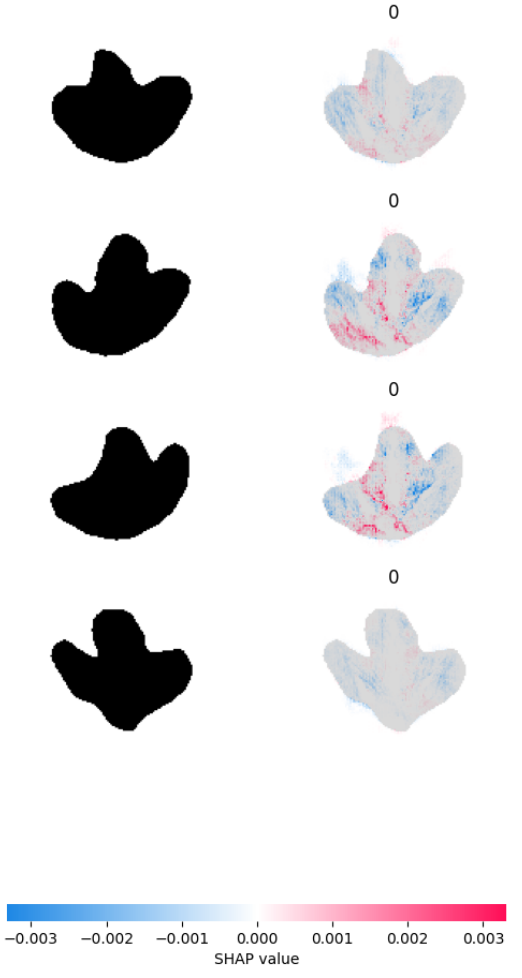

Figure 9. *Computed SHAP values for the ornithischian class.*

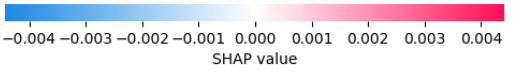

Figure 10. *Computed SHAP values for the theropod class.*

[2] **Lallensack J.N., Romilio A. and Falkingham P.L.(2022)** A machine learning approach for thediscrimination of theropod andornithischian dinosaur tracks, *J.R Soc. Interface. 19:20220588.*

[3] **Falkingham, P.L., Marty, D., Richter, A. (2016)** Dinosaur Tracks: The next steps, *Indiana University Press, Bloomington URL:http://www.jstor.org/stable/j.ctt1c5ckcb.*

[4] **Paulston Park (2023)**Dinosaur footprints: how they were fossilised?, *blog.paultonspark.co.uk URL:https://blog.paultonspark.co.uk/dinosaur-footprints-how-they-were-fossilised/.*

[5] **Rejcek, P. (2022)**AI breakthrough could revolutionize how we research dinosaur fossils, *Frontiers URL: https://www.frontiersin.org/news/2022/01/27/frontiers-earth-science-ai-reconstruct-dinosaur-fossils/*