




Peer Community In Archaeology

A significant contribution to the problem of unbalanced data in machine learning research in archaeology

Alex Brandsen  based on peer reviews by **Simon Carrignon, Joel Santos** and 1 anonymous reviewer

Sebastian Gampe, Karsten Tolle (2024) Creating an Additional Class Layer with Machine Learning to counter Overfitting in an Unbalanced Ancient Coin Dataset. Zenodo, ver. 4, peer-reviewed and recommended by Peer Community in Archaeology.

<https://doi.org/10.5281/zenodo.8298077>

Submitted: 29 August 2023, Recommended: 16 April 2024

Cite this recommendation as:

Brandsen, A. (2024) A significant contribution to the problem of unbalanced data in machine learning research in archaeology. *Peer Community in Archaeology*, 100395. [10.24072/pci.archaeo.100395](https://doi.org/10.24072/pci.archaeo.100395)

Published: 16 April 2024

Copyright: This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

This paper [1] presents an innovative approach to address the prevalent challenge of unbalanced datasets in coin type recognition, shifting the focus from coin class type recognition to coin mint recognition. Despite this shift, the issue of unbalanced data persists. To mitigate this, the authors introduce a method to split larger classes into smaller ones, integrating them into an 'additional class layer'.

Three distinct machine learning (ML) methodologies were employed to identify new possible classes, with one approach utilising unsupervised clustering alongside manual intervention, while the others leverage object detection, and Natural Language Processing (NLP) techniques. However, despite these efforts, overfitting remained a persistent issue, prompting the authors to explore alternative methods such as dataset improvement and Generative Adversarial Networks (GANs).

The paper contributes significantly to the intersection of ML techniques and archaeology, particularly in addressing overfitting challenges. Furthermore, the authors' candid acknowledgment of the limitations of their approaches serves as a valuable resource for researchers encountering similar obstacles.

This study stems from the D4N4 project, aimed at developing a machine learning-based coin recognition model for the extensive "Corpus Nummorum" dataset, comprising over 19,600 coin types and 49,000 coins from various ancient landscapes. Despite encountering challenges with overfitting due to the dataset's imbalance, the authors' exploration of multiple methodologies and transparent documentation of their limitations enriches the academic discourse and provides a foundation for future research in this field.

References:

[1] Gampe, S. and Tolle, K. (2024). Creating an Additional Class Layer with Machine Learning to counter Overfitting in an Unbalanced Ancient Coin Dataset. Zenodo, 8298077, ver. 4 peer-reviewed and recommended by Peer Community in Archaeology. <https://doi.org/10.5281/zenodo.8298077>

Reviews

Evaluation round #2

DOI or URL of the preprint: <https://doi.org/10.5281/zenodo.10424274>

Version of the preprint: 2

Authors' reply, 28 March 2024

Dear Alex,

thank you for your feedback. We now have included a related work section in our paper. We think that it has improved the paper even further.

With kind regards

Sebastian Gampe and Karsten Tolle

Decision by **Alex Brandsen** , posted 10 January 2024, validated 10 January 2024

Major improvements made, just needs a related work section to finalise

Dear Authors,

thank you for making these revisions, this has significantly improved the paper, and the updated structure makes a lot more sense.

I understand you were/are under time constraints, but unfortunately I believe the paper should include a related work / literature review section listing relevant work related to coin classification and/or countering overfitting. Discussing this related work would enhance the understanding of why the presented methods were chosen.

It's also a requirement from PCI, see https://archaeo.peercommunityin.org/help/guide_for_authors#h_8540551119281613314275613, section 3.2.2.: "The introduction should build on relevant recent and past research performed in the field."

Other than that, I think you addressed all the reviewer's comments, and it should be ready to publish once you add the related work section in the introduction.

I believe the PCI system will automatically set a deadline for these edits, but let me know if you need more time, or if you have any other questions.

Kind regards,

Alex

Evaluation round #1

DOI or URL of the preprint: <https://doi.org/10.5281/zenodo.8298078>

Version of the preprint: 1

Authors' reply, 22 December 2023

Dear reviewers,

thank you for your helpful feedback. We have tried to incorporate as many of your suggestions as possible in the new version. Unfortunately, we were not able to implement everything due to time constraints. Nevertheless, we believe that the paper has been significantly improved thanks to your help.

With kind regards

Sebastian Gampe and Karsten Tolle

Decision by [Alex Brandsen](#) , posted 24 October 2023, validated 24 October 2023

Good paper, needs some revisions to be accepted

Dear Authors,

thank you again for submitting to PCI. I read your paper with great interest, and would like to see it published in the CAA proceedings.

However, the 3 reviewers all recommend revisions before accepting the paper. Please check the individual reviews and make revisions where needed, or address the comments of the reviewers. One common theme across the reviews is the structure/clarity of the paper, I would recommend paying specific attention to this particular issue.

If you have any questions, please do not hesitate to contact me.

Kind regards,

Alex Brandsen.

Reviewed by [Joel Santos](#), 15 October 2023

This paper comes from a project, D4N4 (Data quality for Numismatics based on Natural language processing and Neural Networks), aimed to develop a machine learning-based coin type recognition model to cover as many coin types as possible from the "Corpus Nummorum" (CN) dataset. This dataset comprises approximately 19,600 coin types and over 49,000 coins from four ancient landscapes (Thrace, Moesia Inferior, Troas, and Mysia).

This paper addresses the challenges they encountered while dealing with a highly unbalanced dataset, where some coin classes had very few images while others had hundreds of photos. Their main focus was not on improving machine learning algorithms but on the setup to overcome overfitting situations.

The authors aimed to enhance their mint recognition model by subdividing larger classes that impact the overfitting into smaller ones. They employed three different methods for this purpose:

Deep Clustering: An unsupervised clustering method. Although they created new classes based on clusters, this approach resulted in inhomogeneous clusters and lower accuracy.

Object Detection: Based on a Regional Convolutional Neural Network (R-CNN) to predict objects on coins and build new classes based on frequent combinations of subjects. However, this approach did not reduce overfitting effectively.

Natural Language Processing (NLP): This method showed the most promise in reducing overfitting by utilizing an NLP pipeline and creating new classes based on the entities found in textual descriptions of coins. Still, the confusion between new and old classes remained an issue.

The authors explored several approaches to mitigate overfitting in their coin recognition model, but none provided a perfect solution. They continue to investigate other methods to address the problem, such as

improving the dataset, using Generative Adversarial Networks (GANs) to create virtual coin images, and eventually making their dataset available for other researchers to apply their machine learning methods.

One of the most admirable things in the text is its admission that the three hypotheses to solve the overfitting problem did not work. We should have more academic texts unafraid of showing their difficulties in achieving the intended results. This could be very helpful to projects following similar steps, avoiding previously attempted dead-ends.

While the research is commendable, several areas in the text require criticism or improvement:

1. **Clarity and Structure:** The text is somewhat convoluted, making it challenging for readers, mainly less technical ones, to follow the main points. It would benefit from a clearer and more structured presentation. For example, right at the beginning of the text (lines 45-46), it presents the paper's goal before stating the issue they are trying to tackle. A minor issue is the English review, which must be done before the publication (e.g., line 41, "an very unbalanced" or line 46, "For this pupose")

2. **Literature Review and Relevance of Methods:** The text lacks some literature review. What has been done so far in this or other fields using the chosen methods? Discussing related work would enhance understanding of why the presented methods were chosen. The rationale for selecting each method should be clarified.

3. **Data Description:** The text mentions the dataset, but it could be more informative about its origins, sources, and potential biases. This information is crucial for assessing the dataset's quality and comparing similar works trying to achieve the same goals.

4. **Evaluation Metrics:** While the text presents a table with results, it doesn't elaborate on the specific evaluation metrics used. A brief explanation of the metrics (e.g., Top-1 Accuracy) would help readers interpret the results. However, this depends on the target readers of the journal that will publish this article.

5. **Overfitting measurement and class choice:** The overfitting measurement is done visually (at least in the text). Since this is a technical paper, it would benefit from a more measurable approach. Why the Pergamon and Perinthos were chosen? The justification, the size of the sample for Pergamon, and the fact that the Perinthos collection is very different from the Pergamon one falls short. The reasons for this choice should be connected with the initial problem, the overfitting situation. Are there other classes with fewer samples but with high overfitting problems? Are there classes with a high number of samples but with low overfitting problems?

6. **Discussion of Results:** The text provides results but lacks an in-depth discussion. A discussion of what the results imply and their significance would be valuable. The text mentions that the three methods were deemed unsuitable. Was that decision only based on the final accuracy? Was it based on the visual overfitting check regarding the sub-classes of Pergamon and Perinthos? Does the overall overfitting of those two classes increase or diminish (being only big inside those same two classes)? If it decreases, how does the initial hypothesis of reducing the overfitting effect to increase accuracy remain? A deeper discussion would be appreciated, with more measurable results on the three hypotheses result presentation.

7. **Future Work:** The text mentions future steps but could provide a more precise outline of what is planned next. This would give readers a sense of the research's ongoing and prospective significance. Is the approach taken in this paper (reducing the size of certain classes) abandoned?

In summary, the text discusses a study on image recognition and machine learning. However, there is potential to enhance the content's clarity, structure, justifications, and context. By offering more detailed information and explaining the methodology, results, and implications, the text could become more accessible and informative for readers, particularly those not experts in the field. The tackling of these situations would make this article ready for publication. My advice is to review and resubmit.

Reviewed by [Simon Carrignon](#), 22 October 2023

The paper provides an excellent example of the overfitting problem often encountered in machine learning and discusses in details the challenges in resolving this issue. I think it's a valuable contribution to the literature

on applying Machine Learning techniques in Archaeology and would be an ideal publication for PCI Archaeology. The online tool provided with the paper, along with the detailed explanation of their limitations presented in the paper, stand as a significant resource on itself, that many other researchers will find useful.

However, I have a few comment on the form of the paper's structure that merit consideration before publication.

While I recognize that emphasizing the 'unsuccessful' nature of the outcomes may not seem the most gratifying approach, I really hope researchers will become more comfortable with this transparency in the near future. This case study serves as a good instance of such reporting, where various methods are explored but ultimately fall short of resolving a particular issue. However, I think the paper's title and abstract could still mislead reader that the authors will provide a method that successfully addresses the challenges of overfitting. Clarifying that this isn't the case (particularly with the use of 'creating' in the title, which I find somewhat misleading) would be beneficial. The paper's most important value in my option is its detailed account of the rigorous attempts to combat overfitting through three ML classification methods, none of which fully succeeded due to uneven sampling in the dataset. This insight could prevent others from spending time and effort on similar approaches. I think this point should be explicitly stated in the abstract and introduction. The way it reads for know feels to me that it is still not clear if one of the methods will solve or not the problem.

Still in the introduction, the initial sentences are a bit confusing; it feels like the real introduction doesn't start until line 55. Everything before that (about the goal and focus, from lines 40 - 45) becomes much clearer after the subject and the D4N project are properly presented. A minor reorganization of the introduction, including the aspects I mentioned earlier, could significantly clarify the paper's goals and interests.

Finally, I think that a few words could be added about the general limitation this paper expose on the use of Machine Learning in archaeology. Classification is a huge part of archaeology and isn't much discussed in the paper at all ; while an interesting aspect of this research is it's illustration of how some archaeological problem are unlikely to be solved by machine learning due to the nature of the archaeological record and how machine learning algorithm works. No matter how sophisticated are the neural networks, the archaeological record will always be heavily biased, uneven and uncertain and other statistical methods need to be used to asses this uncertainty. The paper could make an more general and interesting point by addressing some of these issues.

Regarding the online app, the code on Google Colab appears functional; though, I didn't have the opportunity to extensively test it by uploading different images of coins myself.

I would be very happy to read a revised version of the paper.

Reviewed by anonymous reviewer 1, 24 September 2023

Summary of the content

This paper presents a method to counter a problem which is well known in the case of coin type recognition: to have an unbalanced dataset for which models will tend to classify the most represented class in the dataset. The authors tried to tackle this problem, by shifting the problem from coin class type recognition to coin mint recognition. This led to more samples per class, though the problem of an unbalanced dataset is still present. They decided to split the biggest classes into smaller ones to obtain a balanced dataset. These newly introduced classes have been incorporated in an 'additional class layer'. They used three different ML approaches to find new possible classes for the two mint classes with the majority of samples. The first approach is based on an unsupervised clustering method with additional manual work, the other approaches take into account the motifs of the coins themselves. The first relies on an object detection model that predicts trained entities and the second is based on Natural Language Processing (NLP) to find entities in textual descriptions of the coins. Based on the combination of obverse and reverse results the new additional class layer has been defined for each of these two approaches.

Considerations of the work

The motivation of the work is well explained, and properly tackling the problem of unbalanced datasets is fundamental to defining robust models through ML approaches.

However, sometimes the work is difficult to follow and methods and procedures have to be described in a clearer manner. Moreover, the paper presents different ways to refer to figures (e.g., fig. or Figure) and when the authors introduce an acronym they should use it consistently throughout the paper.

In detail

Starting from the abstract, the authors write "One is an unsupervised clustering method without additional manual work. The other two are supervised approaches taking into account the motifs of the coins themselves:". We suggest authors rewrite it as "One is an unsupervised clustering method without additional manual work. The other two are supervised approaches which explicitly take into account the motifs of the coins themselves:". This is because we do not know what the unsupervised method is taking into account to cluster the samples, it could be the case it is taking into account the motifs, too.

Moreover, we suggest rewriting "Based on the combination of obverse and reverse results from these two approaches the new additional class layer were defined." to make clearer the fact that the creation of the new class layer has been defined independently for the method a) and for the method b).

We suggest authors name Regional Convolutional Neural Network as Region Based Convolution Neural Network.

In lines 47 and 50 authors use 'object detection' and 'Object Detection', we suggest choosing a standard and continuing to use it during the paper.

In lines 53-54, 'All of the above methods run on Jupyter Notebook and Python programming language.' to be rewritten as 'All of the above methods run on Jupyter Notebook and are written in Python programming language.'

In lines 58-59 the authors use two times 'to improve', we encourage them to use synonyms in sentences that are one after the other. For example, they could change "The goal is to use it to improve and verify the data quality of existing data, and also use it to improve the process of entering new coins." to "The goal is to use it to improve and verify the data quality of existing data, and also use it to help the process of entering new coins."

In lines 63-64 authors write 'we merged the obverse and reverse images of a coin into a single image showing both (as can be seen below in Figure 3)', and to see the image the reader has to scroll down the paper. It is better to add a single image near this paragraph. We suggest to put it above.

We suggest the authors use a consistent notation: there are both Figure and fig.. Moreover, we suggest using a similar notation to refer to tables (if referring to figures as fig., let's refer to tables as tab.).

To help the readability of the images we suggest putting a slope in the text below the figure: the text regarding the names of the columns. For example, let's use 45 degrees of slope to help the readability.

In line 82 authors say "Most mint classes consist of several different coin types which differ more or less from each other.". It would be appreciated a figure with some examples of high and low difference in the same mint class.

In line 84, the authors say that in the case of mint recognition "we encounter another problem when training mint recognition: the unbalanced dataset.", this problem is present also in the case of class type classification, so we suggest making a connection to the previous case "we still encounter the problem of having an unbalanced dataset when training mint recognition, though the number of instances per class has been increased to a manageable number to train DL models".

In line 85, the authors say "to the different output of the mint", we suggest explaining better this sentence. What is the output? In the same line, we suggest saying "in the area of interest" and not "in our area".

In lines 97-98, the authors say that a confusion matrix has the diagonal deep red when the model is 100% correct. The colour depends on the settings set for plotting, we suggest using a more correct language, saying that the matrix would be non-zero only on the diagonal.

In the case of the method Deepclustering, for the fact that Pergamon and Perinthos have 3600 images and 1800 images, respectively, could be the case of considering 15 clusters for Pergamon and half or a lower

number for Perinthos? Is there a reason for considering the same number of clusters for the two mints? To consider a number of clusters proportional to the cardinality of the samples of the two mints could possibly lead to a more balanced result.

We suggest explaining better the sentence "This means that our original classes do not live on as a remnant collection with images that could not be merged with others to form a new class." It seems that the clustering model is using all the images from Pergamon and Perinthos and creating clusters using all the images. This leads to the definition of clusters that comprise both images from the two mints. How do the authors derive the 15 clusters for Pergamon and Perinthos, respectively? This has to be explained better.

In lines 183-184 authors say "which produces a set of bounding boxes to predict the region and the class of an object in an image". We suggest writing say "which produces a set of region proposals that are likely to contain objects, and uses a CNN to extract features from each region proposal to classifying objects within these regions."

Authors say "We trained the R-CNN model on frequently occurring subjects on the coins like "head" or "sitting person".". It would be interesting for the reader to know exactly what are the frequently occurring subjects, though authors give the reference of a thesis.

We suggest adding the images of the coins whose description is in Fig. 6.

We suggest removing the citations in 'Summary and Conclusions'.

Line 254 presents the sentence "Neither approach produced appropriate results for our problem", what approach? We suppose it is the one based on R-CNN.

Errors

- Lines 29-30 'new additional class layer were defined' to 'new additional class layers were defined'
- There are many n.d. present in the paper to be removed. Maybe there is some problem with the insertion of the citations.