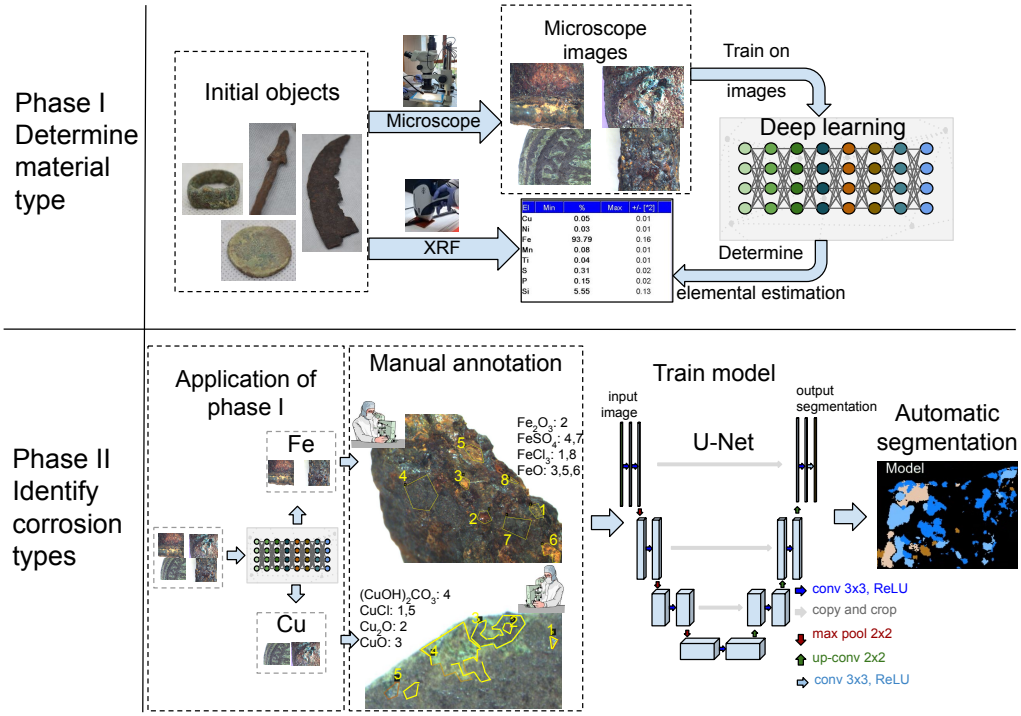
Graphical Abstract

Computational Framework for the Evaluation of the Composition and Degradation State of Metal Heritage Assets by Deep Learning

Ruxandra Stoean, Nebojsa Bacanin, Catalin Stoean, Leonard Ionescu, Miguel Atencia, Gonzalo Joya



Highlights

Computational Framework for the Evaluation of the Composition and Degradation State of Metal Heritage Assets by Deep Learning

Ruxandra Stoean, Nebojsa Bacanin, Catalin Stoean, Leonard Ionescu, Miguel Atencia, Gonzalo Joya

- Deep learning shows corrosion compounds of unrestored metal heritage assets.
- The model learns from microscopic images, based on the expert manual delineation.
- A preceding regression model approximates the chemical composition of the artefact.
- A semantic segmentation model recognizes and delineates the specific compounds.
- The framework is a cost- and time-effective alternative for asset on-site inspection.

Computational Framework for the Evaluation of the Composition and Degradation State of Metal Heritage Assets by Deep Learning

Ruxandra Stoean[a,b], Nebojsa Bacanin[c], Catalin Stoean[a,b], Leonard Ionescu[a,d], Miguel Atencia[e], Gonzalo Joya[e]

[a] *Artificial Intelligence and Machine Learning, Romanian Institute of Science and Technology, Saturn 24-26, Craiova, 400504, Romania*
[b] *Department of Computer Science, University of Craiova, A. I. Cuza 13, Craiova, 200585, Romania*
[c] *Faculty of Informatics and Computing, Singidunum University, Danijelova 32, Belgrade, 11000, Serbia*
[d] *Restoration and Conservation Laboratory, Oltenia Museum , Madona Dudu 14, Craiova, 200410, Romania*
[e] *Universidad de Málaga, Blvr. Louis Pasteur, 35, Malaga, 29071, Spain*

Abstract

The accurate assessment of the material constitution and degradation in newly discovered archaeological artefacts is paramount for the decisions surrounding a thorough treatment of the object during the restoration and conservation stages. The laboratories possess the competent experts and complex devices to perform this analysis properly. Nevertheless, a timely hint of an artificial intelligence assistant regarding the chemical composition and corrosion compound localization of a metal asset could save additional time and resources. The present paper proposes such a computational framework based on deep learning techniques that, on the base of its automatic determination of the chemical concentration of the predominant metal from a microscope image, can subsequently independently also recognize and delineate the corrosion spots of the products specific to that metal. The experiments have been performed on iron and copper heritage items from the Oltenia Museum, Romania. The results suggest that, even with an economic training information in terms of microscope images and annotations, the artificial intelligence framework can provide on-site support for an early examination of metal heritage assets.

Keywords: metal heritage assets, restoration, chemical composition, corrosion compounds, deep learning, semantic segmentation

1. Introduction

The factors underpinning the degradation of archaeological heritage assets are numerous, ranging from the environmental conditions to the material composition, the structural design and the manufacturing means. The complexity of the corrosion phenomenon of metal artefacts has been studied extensively [1, 2].

The restoration process is accordingly very complex and necessitates a highly trained multidisciplinary human resource and specialized X-ray and spectroscopy devices. This presumes effort and time in order to most rigorously assess the state of the artefact and decide on the appropriate approaches to restoration and conservation.

In this context, a preliminary fast inspection of the archaeological piece can be nevertheless made as early as at the excavation site, with the help of a portable microscope and an artificial intelligence assistant. The aim of the present paper is to put forward a deep learning (DL) framework for an initial computational assessment of the chemical composition and degradation of an unearthed metal artefact. The approximation of the concentration of the predominant metal and the subsequent recognition and representation of its corrosion compounds present on the surface of the object can be automatically performed by the proposed DL models on the microscope image and give a first impression to the restoration specialist, before more detailed analysis is performed in the specialized lab.

The pieces studied in this paper come from the Museum of Oltenia, Craiova, Romania and are represented by Fe and Cu artefacts, which are prevalent in archaeological sites. The concentration estimation of the first stage of the DL framework is close to the XRF approximation taken at a middle point of the object and is thus reliable as the starting point of the evaluation. The automatic corrosion product segmentation of the current stage of the computational framework comes similar to the manual delineation of the human experts, even if the training was based on rough, timewise annotation masks. To the best of our knowledge, this is the first DL framework for the multimodal analysis of material composition and degradation of metal heritage artefacts.

The paper is structured as follows. The research aim is detailed in section 2. The formulation of the problem treated, the underlying data and the methodology proposed beyond the state of the art are outlined in section 3. Section 4 presents the measures elected for the performance of the framework in assessing the degradation and the class balance consideration between the compounds. The results of the models are given in section 5, followed by their discussion in section 6. The conclusions of the study and some steps for future work are shown in section 7.

2. Research aim

The aim of the current work is to put forward a computational support tool for an on-the-spot objective assessment of the degradation state of a metal artefact from microscope images. DL techniques are endowed with a structural architecture of layers that mimic the human brain and the visual animal cortex. These two natural characteristics make the networks able to learn patterns found in images and connect them to taught outputs. In the present context, the DL approaches learn from training samples manually annotated by human experts how to segment the corrosion compounds found on the surface of metal archaeological objects made of Fe and Cu and semantically recognize each of them. This is the secondary component of a two-step complete framework of a computational assistant for the estimation of the elemental composition at the surface of a metal artefact and the subsequent automated delineation of its corrosion compounds. The system can be used for first inspection immediately at the excavation site, needing as little as a portable microscope. The framework is part of the larger project OPERA[1] that targets the analysis of heritage assets for material characterization and content reconstruction through artificial intelligence.

3. Material and methods

The time and contact with the soil, also enhanced by water, minerals, salts, temperature and living organisms, affect the metal artefact by strongly inducing corrosion. Moreover, once the object gets in contact with the air, corrosion is accelerated. Its structure, the presence of defects, as well as its chemical composition, add to the degradation of the piece.

[1] https://sites.google.com/view/pce-opera/

3.1. Problem formulation

In the restoration laboratory, the surface of a metal artefact and sometimes its metallic core (visible when the piece is fragmented and the fractures are transversal) are subject to a primary analysis. Either only some of the corrosion compounds specific to that metal or even all the possible corrosion states (oxides, sulfates, chlorides or carbonates) may be present.

In the case of Fe assets, the oxidation takes place in steps [3]: in the first stage FeO is formed and remains stable only in the absence of O_2; once O_2 is present, it is transformed to $Fe(OH)_3$ and the core changes to magnetite (FeO and Fe_2O_3). The other common forms of corrosion are represented by sulfates ($FeSO_4$) and chlorides ($FeCl_3$). The temperature variations, the minerals and salts in the soil, the exposure to the sulfur in the air induce the appearance of sulfate and carbonate compounds. The presence of chlorides is confirmed by the immersion of the object in a precipitate made of one drop of $AgNO_3$ for every 1cm^3 of distilled water or through the use of a humid room, where the piece is introduced in a desiccator together with a recipient with distilled water and the indicative presence of liquid drops on the metal is checked after 48 hours.

For Cu objects, the alloy of manufacture is usually very reactive and oxidizes rapidly when it gets in contact with O_2, hence facilitating the formation of the black CuO and the red Cu_2O. The green corrosion layers indicate the presence of the $(CuOH)_2CO_3$ and $CuCl$, which are generated by the hydration and the salts in the soil.

Once the corrosion compounds have been identified, the piece usually undergoes a combined chemical and mechanical treatment. Chemical treatments can be made for total or partial corrosion removal, dechlorination, corrosion inhibition or preservation. In some cases, methods of converting corrosion compounds into other chemically stable ones can be applied. The processes are diverse, being able to be applied to the parts by total immersion in chemical solutions or point-wise. There can also be electrochemical treatments, also applied by immersion in an electrolyte or point by point with an electrochemical cleaning device. Chemical treatments can be interspersed with other methods, using devices (ultrasonic immersion baths) followed by manual brushing with suitable hand brushes. Mechanical treatments can be applied individually or in tandem with chemical ones, as is the method of cleaning metal surfaces using ultrasound machines, microblasting or laser technology. The corrosion compounds produced by the reaction of the metal with sulfur can therefore be removed by chemical and ion exchange

solutions. Dechlorination, performed to remove chlorides, is very important for the chemical stabilization of metals. Chlorinated corrosion compounds are also called metal cancer. They act destructively not only on the surface of metals but also in their crystalline structure. Applied to pieces that need patina preservation, dechlorination can take place over a long period of time, months or even years. Other corrosion products can be removed by chemical and mechanical methods (listed above). Corrosion inhibition treatments play an important role in the preservation of metals and are usually done by immersion in the inhibiting solution, in the ambient environment or under vacuum, using suitable devices. Applying a protective layer to the metal surface is fundamental. This film creates a barrier between the metal and the surrounding environment, stopping oxygen and harmful substances from reacting chemically with the metal [4].

The computational analysis by artificial intelligence considers FeO, Fe_2O_3, $FeSO_4$ and $FeCl_3$ corrosion compounds for Fe and CuO, Cu_2O, $CuCl$ and $(CuOH)_2CO_3$ for Cu, as the most usually found products in the heritage assets.

3.2. Data

Several microscopical images were captured for the same archaeological metal object with an Olympus SZX-7 optical stereo microscope, having a Quick Photo Micro 2.2 digital camera. The images were taken from different parts of the piece. Still, for objectivity reasons, we separated the training, validation and test set based on the initial objects. Accordingly, each object has all microscopical images in either of the three data sets, i.e. training, validation, or test. The balance between the data sets concerning the number of images targeted having about 70% in the training set and the rest split in half between the validation and the test set.

The Fe data set contains 29 objects for which a total of 228 microscopical images are made. Out of these, 20 objects represented by 157 images form the training set, while the validation set comprises 4 objects totaling 35 images and the test set contains 5 objects and 36 images, respectively.

The Cu data set has 27 objects counting a total of 153 images. The training set has 20 objects, while the validation and test sets consist of 4 and 3 objects respectively. On the number of images, the training has 110 images, the validation has 21 and the test set has 22 items.

The corrosion compounds for both metals were visually identified and confirmed by a portable Bruker Titan S1 XRF spectrometer, having the

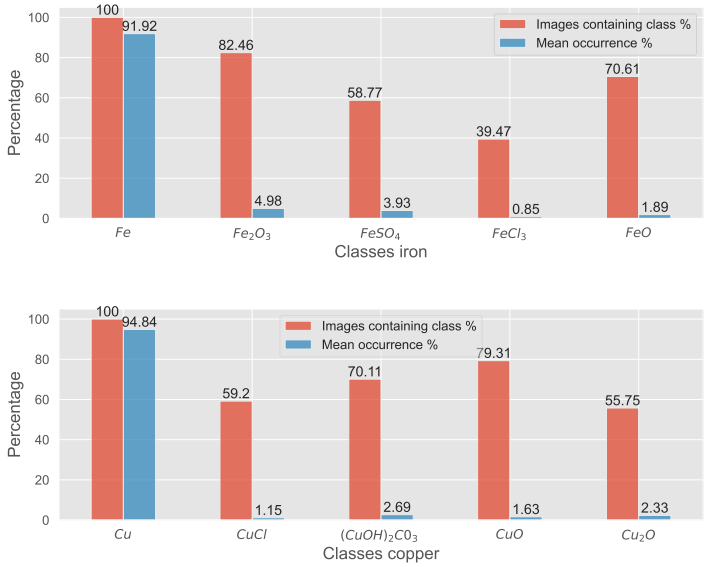


Figure 1: Occurrence in percentage for each class in turn in the entire data set for iron (top plot) and copper (bottom plot). *Images containing class* indicate how many pictures in the data set contain any pixel of the corresponding class, while *mean occurrence* shows the average percentage of pixels of that class calculated only for images where the class actually appears.

Artax software. The restoration specialists annotated the areas by labelled polygons of the corresponding corrosion products. The software used for the manual segmentation was the VGG Image Annotator [5].

Figure 1 shows an overview of the number of images that contain the various corrosion classes for Fe and Cu, respectively. Each couple of bars per class illustrates the percentage of images from the entire data set that contain that specific class (first bar) and the pixel coverage, also in percentage, for the same class in the selected images that possess regions of the class (second bar). It can be thus observed that some classes are almost absent in the two data sets. For the iron data, only 39.47% of the images in the data set contain any pixels of class $FeCl_3$ and, moreover, in the small subsection of images that possess regions with $FeCl_3$, the average percentage of image coverage is only 0.85%. Similarly, within the copper data, class $CuCl$ appears in only 59.2% of the images and in these the coverage in percentage is only 1.15%.

3.3. State of the art

The general study of corrosion and its impact on materials by machine learning has begun to be intensive. Corrosion due to the atmospheric conditions and leading to the destabilization of structures made of different materials has been subject to prediction through various opaque (neural networks, support vector machines) and transparent (decision trees, random forests) methods [6]. DL was also employed for the classification of clean slides versus electrochemically-induced uninhibited and inhibited corrosion [7, 8].

From the practical perspective, the computational investigation of the corrosion state of a structure has targeted specifically the sector of civil engineering. Also, corrosion has been seen as represented only by rust, which computationally reduces the task to solely a binary discrimination between regions. Several DL representative approaches for semantic segmentation, like U-Net and Mask R-CNN, have been employed for metal constructions [9, 10].

Hence, the current paper attempts to fill the gap in three aspects. First of all, there appear to be no papers related to the computational analysis of the distinct degradation process of heritage assets, which happens over centuries, in the presence of many environmental, chemical and biological factors of the archaeological site and where the objects have different, historical age-related technological properties. Secondly, the consideration of corrosion should be made with respect to all its various compounds for a metal. Moreover, thirdly, the methodology should prove generalization ability in dealing with corrosion products of other metals.

3.4. Methodology

The approach proposed in the current study is the second phase of a bipartite framework for a compositional and deterioration assessment of metal heritage artefacts. The first stage of the framework consisted of a DL model for the approximation of the elemental composition of an artefact as an immediate alternative to XRF [11], [12].

Phase I of the computational investigation of the initial state of an excavated archaeological object implied training the deep neural network with microscope images and the attached XRF estimation of the percentage of selected chemical elements. The DL architecture was trained on objects of Fe and Cu, as those most commonly found in excavation sites. The pattern recognition task was formulated as a multi-output regression problem, with

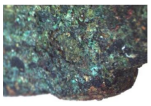
Macroscopical	Microscopical	Material	XRF	DL
		Fe	1.85	3.03
		Cu	61.73	64.91
		Fe	96.26	93.89
		Cu	0.03	2.64
		Fe	4.19	3.40
		Cu	38.06	41.55
		Fe	94.30	90.61
		Cu	0.12	0.16

Figure 2: Four samples shown at the macroscopical and microscopical levels, together with the XRF estimations (ground truth) for iron and copper and the values determined by the DL model (also in percentage) in the last column.

the input given by the microscope images and the output by the approximated percentage of the two appointed chemical elements. Once the model learnt the association, it approximated by itself the elemental nature of new artefacts. Additionally, a slack window of error between the DL estimation and the XRF ground truth was allowed in computing the test accuracy, since the XRF value is itself an approximation that is sensitive to the place of radiation. Some examples of the closeness between the percentage values estimated by the DL and the XRF are given in Figure 2.

As phase I of the framework is finished, the predominant metal has already been computationally determined in phase I. In the case of Fe, a threshold of 60% is indicative of the nature of the piece, while, for Cu, a value above 30% can already designate an alloy of this metal.

In phase II, the corrosion products of the defining metal will be determined on the surface of the object, as seen from a microscope image. Two models had been trained to recognize the corrosion compounds specific to each of the two metals. Recall that FeO, Fe_2O_3, $FeSO_4$ and $FeCl_3$ are searched for in the case of Fe and CuO, Cu_2O, $CuCl$ and $(CuOH)_2CO_3$ will be identified for Cu.

The training of the models for corrosion product detection assumes that

there is a data set of microscope images where the area of every compound is manually delineated and labeled accordingly by human experts. These annotated training samples are given to a DL model for semantic segmentation that learns the shapes of interest and their characteristics. On a new microscope image, the trained architecture is then able to outline the location and type of the corrosion products.

There are several DL architectures for semantic segmentation. In the present framework, we opted for a U-Net [13], which is acknowledged for its performance in such tasks. Its construction has a spatial contraction path and an expansion turn. In the contraction stage, the network uses convolution and max pooling for downsampling, while doubling the kernel depth. In the expansion stage, there is upsampling and convolution, plus concatenation with features from the contraction stage, while halving the kernel depth. Within our particular implementation, we opted for ResNet50 as encoder and decoder, since this proved to be convenient both with respect to the quality of results and running time.

A depiction of the entire framework is given in Figure 3. In phase I, the DL architecture had been trained on microscope images with the XRF estimation on a central point of radiation of the objects and is able to differentiate between Fe and Cu pieces on its own. The framework therefore firstly gives the metal of the asset and its concentration. The U-Net model that had been trained for the semantic segmentation of the corrosion compounds specific to the resulting metal is appointed accordingly. The framework thus secondly outputs the corrosion spots and the corresponding products.

4. Calculation

Once the semantic segmentation tasks were formulated for each of the two metals under study, the chosen DL (U-Net) architecture was tailored for their specificity. However, before running the model, several different implications related to the performance measures and class balance had to be considered.

4.1. Quality measures

The quality of the semantic segmentation results is measured based on several facets. For both types of material, there are 5 different classes, starting with the clean material (Fe or Cu), which is largely represented in both cases, meaning that more than 90% of the surfaces within the microscopical

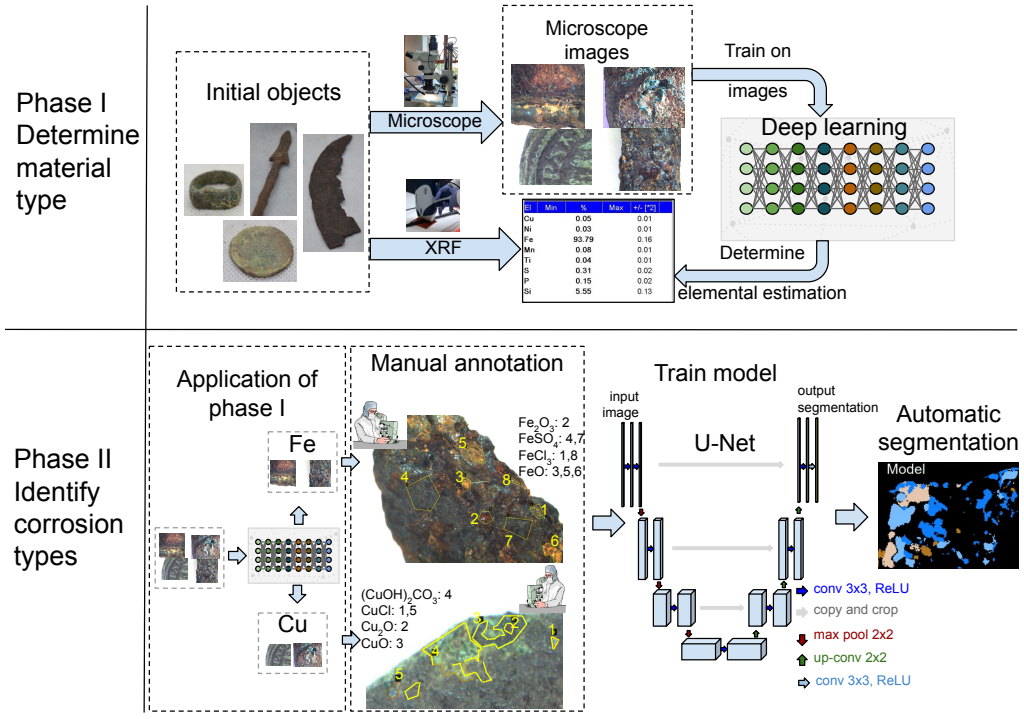


Figure 3: Overview of the proposed DL framework for material recognition and degradation in metal heritage assets.

Data class	Classified as C (pos)	Classified otherwise (neg)
C (pos)	tp	fn
Rest of classes (neg)	fp	tn

Table 1: The meaning of true positive (tp), false negative (fn), true negative (tn) and false positive (fp) for a class C in a multi-class problem.

images represent the clean material, as seen in Figure 1 for Fe and Cu. In both situations, there are 4 other classes that have a low overall appearance in the data set. Some classes are not represented at all in a large part of images and, in the images where they are present, they have a very small number of pixels, i.e. between 0.85% and 4.98%. Consequently, it is important to evaluate the quality of the outputs through various means [14]. We will further describe the used measures and the formulas used to calculate them.

We deal with a multi-class problem, hence we will refer to each class in turn versus the rest when computing measures like precision, recall, F1-score or Jaccard index. Accordingly, for a class C, assessments like true positive (tp), false negative (fn), true negative (tn) and false positive (fp) will refer to the specific class C against the other classes altogether. The manner in which they are defined for class C in a multi-class problem is expressed in Table 1. Considering we have a class C_i, $i \in \{1, 2, ..., n\}$, where there are n classes (the corrosion compounds plus no corrosion), we will further refer its corresponding measures by tp_i, fn_i, tn_i and fp_i, respectively.

Within the image segmentation task, we will verify the correctness of the results at the level of the pixels in the images. Thus, every pixel is seen as a sample that may be assigned to a class correctly or not. The ground truth is given by the annotations of human experts. We then compute each measure at the level of an image and then we average the obtained results for all the images in the validation or test sets accordingly.

The precision for a class C_i represents the number of correctly classified samples (pixels, in our case) divided by the number of samples labeled by the models as C_i. Equation (1) shows how it is computed in the binary classification. In our specific image segmentation problem, i.e. as we deal with a multi-class problem, the precision for one class C_i is computed similarly, that is fp_i refers to all the pixels that are assigned by the model to class C_i, while they correspond in reality to another class. The precision value for a class C_i deteriorates when the model wrongly assigned class C_i to pixels

from a larger area than it should.

$$P_i = \frac{tp_i}{tp_i + fp_i} \tag{1}$$

Recall of a certain class C_i reflects the effectiveness of the model to identify the pixels that belong to that class. Its formula is given in (2) and computes the number of correctly classified pixels in class C_i (as found by the classifier) divided by the number of all pixels that belong in reality to class C_i. The recall measures the amount of relevant pixels that are retrieved for the class, that is how much of the targeted pixels are actually correctly identified.

$$R_i = \frac{tp_i}{tp_i + fn_i} \tag{2}$$

The F1-score achieves a combination between the precision and recall of a model into a single metric by calculating their harmonic mean. Its formula is given in equation (3).

$$F1_i = \frac{2P_iR_i}{P_i + R_i} \tag{3}$$

The Jaccard index is described by equation (4) and it is generally defined as the intersection over union when it is used for evaluating the similarity and diversity of sample sets. For image segmentation, this is better described as the ratio between the overlapping area of the target and the predicted zones, and the area of reunion of the two.

$$J_i = \frac{tp_i}{tp_i + fp_i + fn_i} \tag{4}$$

All the above quality measures are generally defined for binary classification, but they can be formulated for multi-class as well if the score is computed for each individual class in turn and the tp, tn, fp and fn are defined as in Table 1. Nevertheless, measures can also evaluate the model overall [14], e.g. by considering weights for the different classes that are proportional to the quantity of items in each class. Equation (5) illustrates the manner of calculating the weighted precision for the multi-class problem: n denotes the number of classes of the problem, $\mid C_i \mid$ represents the number of samples from class C_i and P_i is the precision from class C_i defined as in Eq. (1). Similarly to the weighted precision, measures like weighted recall, F1-score or Jaccard index can be defined. The overall accuracy is computed

by dividing the samples that are correctly classified over the total amount of samples. These weighted values will be the ones that will be reported in the experiments of section 5.

$$P = \frac{\sum_{i=1}^{n} \mid C_i \mid \cdot P_i}{\sum_{i=1}^{n} \mid C_i \mid} \tag{5}$$

4.2. Calculation of weights for classes

In the standard U-Net approach for image segmentation, each class is treated equally. However, when classes are highly imbalanced, it may happen that the model often misses the items of the less represented ones and favors the ones that are more extensively present. Since class imbalance is a trait of the current task, as seen from Figure 1, besides the standard U-Net implementation, we also experimented with using specifically tailored weights for all the classes as described in [15] and using the implementation in [16].

Within the image segmentation experiments, we split each of the two data sets into training, validation and test sets. The weights for the classes are in each case computed based on the distribution of the classes from the training data.

5. Results

One question that emerges from the image segmentation application is whether the calculated weights would lead to better results. In order to provide an answer to this, each setup is used for both data sets and all the measurements described in subsection 4.1 are used. This way, the benefits and the downsides of using class weights will be outlined.

As described in subsection 3.2, the data sets are split into training, validation and test sets. Figure 4 shows the manner in which the images are split between the validation and the test set. The top row shows the iron data set and the bottom row points to the copper one. The left plots show in percentages how many of the images of the corresponding split actually contain any pixel of that class. The right plots indicate the mean occurrences, also in percentages, of the actual pixels of these classes, as they appear only in the images where the class is present, i.e. in the images that are considered for

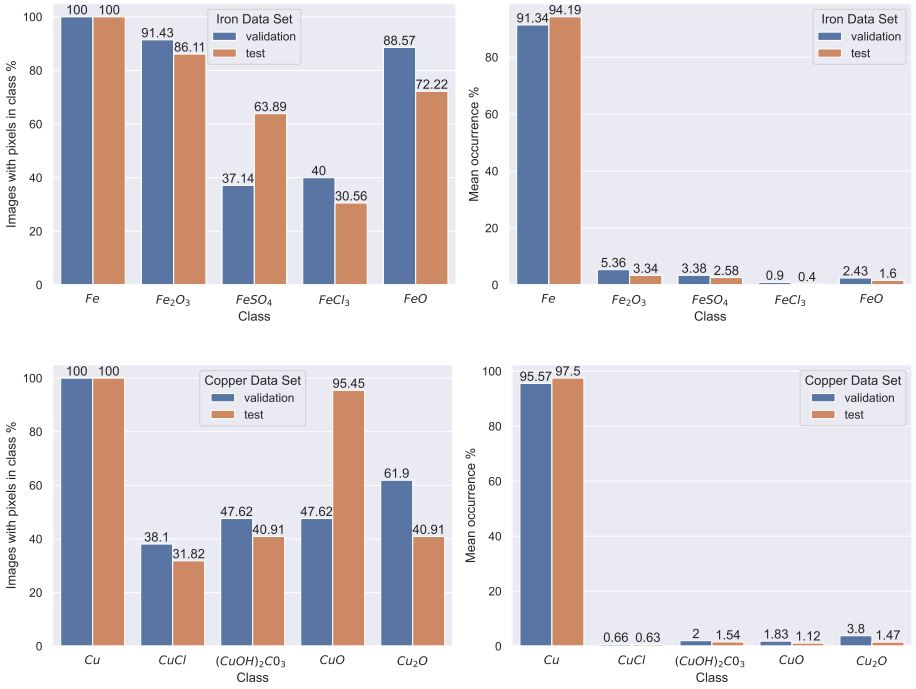


Figure 4: Occurrence of pixels in the validation and test data sets for iron (top row) and copper (bottom row). Left plots indicate the percentage of images in which pixels of the specific class appear. Right plots show the average percentage of the pixels of that class computed only for images where the classes are represented.

creating the plots on the left side of the figure. These plots can be compared with the overall spread in the data set in Figure 1 to observe that the class distribution was maintained to a proper extent within the splits.

Table 2 indicates the validation and test results for the weighted measures, while table 3 dissects the results for each class in turn. The results are shown both for the case when no specific weights are calculated for the classes and when these are used. The findings are further discussed in the subsequent section.

Figure 5 illustrates a copper test sample in which the original image can be observed together with the human expert manual annotation and with the image segmentation determined by the model with and without using calculated weights for the classes.

Material	Set	Weights	Precision	Recall	F1	Jaccard	Accuracy
Iron	validation	yes	0.929	0.606	0.703	0.563	0.606
		no	0.914	0.933	0.906	0.865	0.919
	test	yes	0.945	0.596	0.696	0.557	0.596
		no	0.934	0.941	0.929	0.894	0.929
Copper	validation	yes	0.960	0.830	0.886	0.812	0.835
		no	0.950	0.962	0.949	0.929	0.959
	test	yes	0.975	0.872	0.912	0.855	0.871
		no	0.973	0.987	0.977	0.965	0.976

Table 2: Weighted overall results presented as ratios for the two materials for validation and test sets, with or without using weights for the classes.

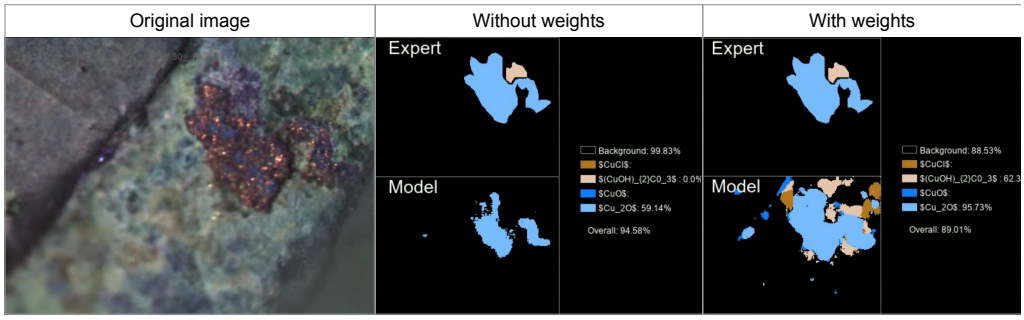


Figure 5: Copper sample from the test set and the results of the model without and with weights for the classes.

Material	Set	Weights	Class	Precision	Recall	F1	Jaccard
Iron	validation	yes	Fe	0.99	0.60	0.73	0.59
			Fe_2O_3	0.28	0.72	0.36	0.25
			$FeSO_4$	0.22	0.67	0.29	0.20
			$FeCl_3$	0.08	0.71	0.15	0.08
			FeO	0.19	0.54	0.24	0.14
		no	Fe	0.92	0.99	0.96	0.92
			Fe_2O_3	0.68	0.21	0.28	0.19
			$FeSO_4$	0.65	0.13	0.17	0.11
			$FeCl_3$	0.50	0.04	0.06	0.03
			FeO	0.61	0.05	0.08	0.05
	test	yes	Fe	0.99	0.58	0.72	0.58
			Fe_2O_3	0.13	0.84	0.20	0.12
			$FeSO_4$	0.13	0.77	0.21	0.12
			$FeCl_3$	0.12	0.49	0.18	0.11
			FeO	0.16	0.52	0.21	0.12
		no	Fe	0.95	0.97	0.96	0.93
			Fe_2O_3	0.43	0.35	0.24	0.14
			$FeSO_4$	0.20	0.03	0.05	0.03
			$FeCl_3$	0.56	0.08	0.14	0.08
			FeO	0.75	0.07	0.12	0.07
Copper	validation	yes	Cu	0.99	0.84	0.90	0.83
			$CuCl$	0.13	0.67	0.19	0.12
			$(CuOH)_2CO_3$	0.16	0.69	0.24	0.15
			CuO	0.15	0.73	0.23	0.14
			Cu_2O	0.36	0.64	0.39	0.27
		no	Cu	0.96	0.99	0.98	0.96
			$CuCl$	0.45	0.03	0.05	0.03
			$(CuOH)_2CO_3$	0.23	0.01	0.01	0.01
			CuO	0.53	0.12	0.17	0.11
			Cu_2O	0.57	0.23	0.28	0.20
	test	yes	Cu	0.99	0.88	0.93	0.87
			$CuCl$	0.19	0.26	0.12	0.07
			$(CuOH)_2CO_3$	0.35	0.54	0.33	0.21
			CuO	0.13	0.59	0.14	0.08
			Cu_2O	0.40	0.58	0.35	0.23
		no	Cu	0.98	1.00	0.99	0.98
			$CuCl$	0.00	0.00	0.00	0.00
			$(CuOH)_2CO_3$	0.84	0.16	0.20	0.12
			CuO	0.45	0.11	0.16	0.10
			Cu_2O	0.83	0.20	0.29	0.18

Table 3: Results for the two materials, for each class in turn, when weights are computed for the classes (*yes* for the specific column) and without using special attention for the class imbalance ($Weights$ with *no*). Results are shown for both, validation and test sets.

6. Discussion

By comparing the results from the validation set with the ones on the test set, on corresponding circumstances, a positive observation is that the model does not overfit. The remark is based on the fact that the results on the test sets generally have better results than the ones from the validation or if otherwise, they have very close values. For this purpose, it is straightforward to compare the results from Table 2. The trend is kept for all the used measures. The only case where the test outputs are not better than the validation ones is for the iron data set, when using weights but, in this case, the results are almost identical.

A direct comparison between the results obtained when using or not weights for the classes can be seen in the same Table 2. Of course, the values should be compared separately for each metal data set and even separately for each split in turn. Besides precision, all the other measures also indicate the fact that the quality of the outputs clearly deteriorates when weights are employed.

For a better grasp of the meaning of these outputs, we switch to Table 3 where results are presented per class. We observe here that the precision for the clean material, either Fe or Cu, is better when the weights are used. On the other hand, for the other classes that are less represented in the data sets, vice versa happens. Moreover, similar to the results in Table 2, recall, F1 and Jaccard illustrate the opposite of precision: the results for Fe and Cu improve when weights are not used and the ones for the other classes worsen in this same case.

Figure 5 illustrates what actually happens for one image when class weights are used or not. Obviously, the expert annotations are the same for both cases, with and without weights. The background is clearly better determined when no weights are used for the classes (a pixel-level accuracy of 99.83%) than when they are employed (88.53%). On the other hand, the main class delineated by the human expert, Cu_2O, represented in light blue, is better identified when weights are used (95.73% as opposed to 59.14%). Moreover, when the calculated weights are considered, 62.3% of the $(CuOH)_2CO_3$ is also identified, while in the other case this is not found at all. Nevertheless, the overall accuracy of this image is 94.58% when weights are not used and 89.01% when they are used.

The conclusion regarding the usage of the class weights or not is that the importance of the less represented classes increases indeed when they are

used, leading to a better recall for them. The model-predicted regions for these classes are enlarged, and it often happens that some regions that are not identified at all by the model without weights are found when the weights are considered. This is proved not only for the sample in Figure 5, but also by comparing the recall values for the same material, set and class, when weights are used or not, e.g. Fe_2O_3 of 0.72 vs 0.21 for validation or of 0.84 vs 0.35 for the test set or, if we take the least represented class from copper, $CuCl$, 0.67 vs 0.03 for validation and 0.26 vs 0 for the test set. However, the regions are overestimated and the background is obviously affected, which leads to a significant loss in its recall (for classes Fe and Cu). The precision, on the other hand, evaluates for a class the proportion of the correctly classified pixels from the number of samples labeled by the models as belonging to that class, and here the overestimation is naturally penalized, as, for the less represented classes, this measure indicates that weights should not be used. F1 score evaluates a combination of precision and recall and is points overall that, without the specific class weights, the background is better identified and the corrosion classes are slightly better found; the differences are not as obvious as they are for precision and recall, respectively. The Jaccard index shows a similar trend to the F1 score.

The classes that are least identified by the U-Net model are $FeCl_3$ and FeO for the iron data set. The classes actually represent the compounds that are also least represented in the data set, not only in the number of images where such regions occur but especially in the size of the regions, as observed in Figures 1 and 4. For the copper data set, the most problematic class is $CuCl$, which is indeed the least represented both in terms of presence in images and with respect to the size of the regions. This indicates that the model did not have enough training examples to accurately learn the characteristics of these classes and, in order to overcome this problem, the data set needs to be enlarged.

7. Conclusions

The target of the current study was to analyze the potential of a deep learning computational assistant for the chemical assessment of archaeological artefacts before restoration. Microscope images of metal objects of iron and copper constitution were presented to the models. The framework first determines the main metal composition and subsequently delineates and names the corrosion compounds present at the surface of the object. This

automatic tool can serve as a rapid, effective support for an on-site inspection of the degradation state of an excavated object.

Although the results are already promising, a larger representation in pixel amount of the corrosion products throughout an additional higher number of images will lead to better recognition. The next steps will also use uncertainty quantification approaches [17, 18] to the semantic segmentation process to increase confidence in the outcome of the artificial intelligence models. Another path that will be followed is that not just one, but two experts delineate the same images, such as to compare the inter-observer versus observer-computer variability.

Acknowledgements

This work was supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI – UEFISCDI, project number 178PCE/2021, PN-III-P4-ID-PCE-2020-0788, *Object PErception and Reconstruction with deep neural Architectures* (*OPERA*), within PNCDI III.

References

[1] D. Dwivedi, J. Mata, Archaeometallurgical investigation of ancient artefacts' degradation phenomenon, npj materials degradation 3 (35) (2019).

[2] J. I. Iribarren, F. Liesa, Álvaro Meneguzzi, C. Alemán, E. Armelin, Spectroscopy investigations reveal unprecedented details in the corrosion of aisi 1012 upn profiles installed in a modernist building of beginning of 20th century, Journal of Cultural Heritage 42 (2020) 240–248.

[3] L. Ionescu, Z. K. Pinter, The recovery, restoration and conservation of a "spatha" found in capidava village, constanţa county, romania, Studia Universitatis Cibiniensis. Series Historica XVI (2019) 291–304.

[4] L. Ionescu, Restoration of a roman statuary group, representing the god jupiter, Studia Universitatis Cibiniensis. Series Historica XVI (2019) 277–289.

[5] A. Dutta, A. Zisserman, The VIA annotation software for images, audio and video, in: 27th ACM International Conference on Multimedia, ACM, New York, USA, 2019, p. 4.

[6] L. B. Coelho, D. Zhang, Y. V. Ingelgem, D. Steckelmacher, A. Nowe, H. Terryn, Reviewing machine learning of corrosion prediction in a data-oriented perspective, npj Materials Degradation 6 (2022) 8.

[7] A. Samide, C. Stoean, R. Stoean, Surface study of inhibitor films formed by polyvinyl alcohol and silver nanoparticles on stainless steel in hydrochloric acid solution using convolutional neural networks, Applied Surface Science 475 (2019) 1 – 5.

[8] A. Samide, R. Stoean, C. Stoean, B. Tutunaru, R. Grecu, Investigation of polymer coatings formed by polyvinyl alcohol and silver nanoparticles on copper surface in acid medium by means of deep convolutional neural networks, Coatings 9 (2019) 105.

[9] A. R. M. Forkan, Y.-B. Kang, P. P. Jayaraman, K. Liao, R. Kaul, G. Morgan, R. Ranjan, S. Sinha, Corrdetector: A framework for structural corrosion detection from drone images using ensemble deep learning, Expert Systems with Applications 193 (2022) 116461.

[10] I. Katsamenis, E. Protopapadakis, A. Doulamis, N. Doulamis, A. Voulodimos, Pixel-level corrosion detection on metal constructions by fusion of deep learning semantic and contour segmentation, in: Advances in Visual Computing, Springer, 2020, pp. 160–169.

[11] C. Stoean, L. Ionescu, R. Stoean, M. Boicea, M. Atencia, G. Joya, A convolutional neural network as a proxy for the XRF approximation of the chemical composition of archaeological artefacts in the presence of inter-microscope variability, in: 16th International Work-Conference on Artificial Neural Networks (IWANN), Advances in Computational Intelligence, Vol. 12862, 2021, pp. 260–271.

[12] R. Stoean, N. Bacanin, L. Ionescu, C. Stoean, M. Boicea, A.-M. Garau, C.-C. Ghitescu, Deep learning for a swift non-invasive recognition and delineation of corrosive iron compounds present on the surface of unrestored archaeological artefacts, Procedia Computer Science 207 (2022) 1303–1311.

[13] O. Ronneberger, P.Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention, Vol. 9351 of LNCS, 2015, pp. 234–241.

[14] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management 45 (4) (2009) 427–437.

[15] G. King, L. Zeng, Logistic regression in rare events data, Political Analysis 9 (2001) 137–163.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, Journal of machine learning research 12 (Oct) (2011) 2825–2830.

[17] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, S. Nahavandi, A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information Fusion 76 (2021) 243–297.

[18] W. Nash, L. Zheng, N. Birbilis, Deep learning corrosion detection with confidence, npj Materials Degradation 6 (26) (2022).