This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N.693548

GRANT AGREEMENT NUMBER:	**693548**
PROJECT ACRONYM:	**ArchAIDE**
PROJECT TITLE:	**Archaeological Automatic Interpretation and Documentation of cEramics**
FUNDING SCHEME:	**H2020-REFLECTIVE-6-2015**
PRINCIPAL INVESTIGATOR	**Prof Maria Letizia Gualandi, UNIPI**
PROJECT COORDINATOR	**Dr Gabriele Gattiglia, UNIPI**
TEL:	**+39 05022 15817**
E-MAIL:	**maria.letizia.gualandi@unipi.it** **gabriele.gattiglia@for.unipi.it**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N.693548

D 8.1 Report of final recommendation for WP8

version: 1.1

Revision: Final

Work Package	**8**
Lead Author (Org)	**Eva Miguel Gascón (UB)**
Contributing Author(s) (Org)	**Gabriele Gattiglia (UNIPI), Llorenç Vila (ELEMENTS), Miguel Ángel Hervás (BARAKA), Nevio Dubbini (UNIPI), Luis Alejandro García (BARAKA)**
Due Date	**30.01.2019**
Date	**01.03.2019 (08.07.2019)**

Project co-funded by the European Commission within the ICT Policy Support Programme
Dissemination Level

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N.693548

P	**Public**	
C	**Confidential, only for members of the consortium and the Commission Services**	**X**

Revision History

Revision	Date	Author	Description
0.1	14.01.2019	Eva Miguel Gascón	First draft
0.2	14.01.2019	Miguel Angel Hervas	Content added
0.3	31.01.2019	Luis Alejandro García	Content added
0.4	01.02.2019	Eva Miguel Gascón	Content added
0.5	04.02.2019	Luis Alejandro García	Content added
0.6	12.02.2019	Gabriele Gattiglia	Content added
0.7	13.02.2019	Eva Miguel Gascón, Gabriele Gattiglia	Content added
0.8	26.02.2019	Nevio Dubbini, Gabriele Gattiglia	Content added
0.9	27.02.2019	Nevio Dubbini, Gabriele Gattiglia	Content added
1.0	28.02.2019	Eva Miguel Gascón, Gabriele Gattiglia, Nevio Dubbini	Final Revision
1.1	08.07.2019	Gabriele Gattiglia	Revision and re-submission

Disclaimer

Table of Contents

Abbreviations

WP: Work package

M: Month

UNIPI: Università di Pisa

UoY: University of York

UB: Universitat de Barcelona

UCO: Universitaet zu Koeln

TAU: Tel Aviv University

CNR: Centro Nazionale delle Ricerche

INERA: Inera srl

BARAKA: Baraka Arqueologos S.L.

ELEMENTS: Elements centro de gestio i difusio de patrimoni cultural

Executive summary

This document represents the first release of D8.1 *Report of final recommendation for WP8.* This deliverable evaluates the results obtained after tasks 8.1 and 8.2, which represents the first assessment of the overall system through the report on testbeds.

The testbeds needed to be done in different possible scenarios which involved both archaeological small and medium-sized enterprises (SME) and Higher Education Institutions (HEI) partners. Even if not immediately understandable, given the archaeological nature of the deliverable, the work went on through a continuous exchange of opinion between experts of different domains. Difficulties bore from the fact that the algorithm was constantly modified and improvements were on continuously. Thus, testbeds were done in different stages of the development of the neural system, being possible to testify the improvements on the accuracy and the quality of the results. Archaeologist, therefore, had a chance to test the system with both appearance-based and shape-based recognition, but not with all the types of ceramics that were taken into account in the project. On the contrary, the system was just tested with Montelupo pottery and Terra Sigillata Italica.

This document is about the assessment of the overall system (from the 2D acquisition to the automatically match and classification) and final recommendations for WP8. These testbeds carried out in WP8 brought to an improved model of the neural network.

Introduction

The core task of WP 8 is devoted to create two testbeds related to different application scenarios. On the one hand, archaeological small and medium-sized enterprises (SME) involved in contract archaeology. These potential users are heavily constrained by constant digging activity, handling great volumes of materials and having a short time for their study. Generally, this end users have also restrictions on facilities (including space) for carrying the study of the unearthed materials. On the other hand, the second type of end-user would be Higher Education Institutions (HEI) and research centres. These latter end-users may also face restrictions similar to the one of archaeological SMEs, especially during the fieldwork seasons abroad or in remote/isolated areas. However, in many circumstances, academic end-users have facilities in the own (research buildings) provided with suitable equipment for carrying the study of the unearthed materials. Moreover, this second user is not constrained to constant digging activity since fieldwork is scheduled in certain periods of the year.

Between M 22-32 different testbeds were created for assessing, on the one hand, the designed mobile tool and the Front-end Desktop Application for the 2D acquisition and the 2D drafting the profile of sherds; and on the other hand, the search and retrieval component for the automatically match and classification according to the digital pottery catalogues produced during the project. Continuous collaboration with ICT partners (TAU and INERA) during the WP8 has brought the suggestions coming from the testbeds to improve both the neural network as reported in D6.3 and the release of the final version of the mobile app and the desktop front-end (D7.2).

The testbeds were conducted on large numbers of specimens that were automatically classified according to typology (shape-based recognition) but also according to decoration (appearance-based recognition). All of them were performed during the last year of the project and were planned to be performed in different contexts of excavations and assemblages. All the results obtained have been evaluated through different statistical data treatments in order to have a clear analysis of the accuracy and the quality of the answers done by the system.

1 Archaeological testbeds: SMEs

In the case of the archaeological SMEs, tests were pretended to be developed straight in the field or during the subsequent phase of the study of potsherds, this means real-cases scenarios from contract archaeological interventions. The actual constraints of contract archaeology and the limited types available inside the ArchAIDE system made difficult to have a large number of sherds to be tested.

The archaeological sites that were planned to be part of the SME testbeds (Task 8.1) were:

-Urban excavations in Palma:

-Can Coll (Majolica pottery)

-Plaça de Cort (TSI, Amphora)

-Convent de Sant Bartomeu d'Inca (Majolica)

-La Misericòrida (Amphora)

-Urban excavation in Andratx:

-Tower of Sa Mola (Majolica pottery)

-Castell de Son Mas (Amphora, TSI)

-Urban excavation in Toledo:

-Real street

-Cuesta de los Portugueses street

-Roman city of Laminium (Amphora, TSI)

-Roman villa in Cabañas de la Sagra (Amphora, TSI)

The SMEs testbeds found difficulties to work with ceramic assemblages coming from excavations of other research teams and stored in different locations. In certain cases was complicated to have access to archaeological objects that are in the study phase by research projects outside ArchAIDE consortium.

Since the beginning of the WP8, SMEs started to search and request the study of pottery sherds in order to test the system. Access to the fragments began in M 31 and testing in M 32. This task was delayed in time due to the work commitments of the persons responsible for the archaeological material under study.

The tests were carried out with fragments already cleaned and selected, corresponding mostly to rims, which provide sufficient information about the typology of the pieces to which they belong. This is the reason why most of the tests were carried out in the laboratory and not directly in the field. In addition, this avoided the difficulty of finding archaeological excavations in process during the months of the project testing. The fragments worked with come both from emergency excavations and from systematic excavations, carried out in any case by SMEs.

The quality of the installations (space, type of lighting, cameras...) can introduce important differences. We worked with different types of natural light (both in the sun and shade) and artificial light (incandescent, led, fluorescent, etc).

Tests have been carried out with artificial light in the laboratories that archaeologists have to work in and in the warehouses of the museums where these archaeologists have deposited the ceramics. In many occasions, the conditions of light are not the most appropriate, but these conditions will be habitual also for the users of the app and, therefore, it was necessary to elaborate the tests with deficient light conditions. The app has the ability to correct the white balance, which reduces the influence of light quality. The first tests carried out seem to show that the app works in a similar way with any type of light, and even better with artificial light than with natural light.

Since the tests have been performed in controlled work environments, the operability of mobile phone and tablet is similar. In the case of using the application in the field, the mobile phone is more appropriate because of its easier operability.

Different mobile devices have been used for taking pictures of the sherds:

- Smartphone Xiaomi Redmi Note 4 (RAM 4 Gb, CPU Octa-Core Max 2,0 Ghz).
- Tablet Lenovo TB2-X30F (RAM 2Gb, CPU Qualcomm APQ8009 1,3 Gh).
- Smartphone Xiaomi A1 Android One (RAM 4Gb, CPU Qualcomm Snapdragon 625 Octa-Core 2,2 Ghz).
- Tablet Lenovo TB2-X30F (RAM 2Gb, CPU Qualcomm APQ8009 1,3 Gh).
- Smartphone Huawei P9 lite VNS-L31 (RAM 16 Gb, CPU Hisilicon Kirin 650 Octa-Core 2000 Mhz).
- Tablet Samsung Galaxy TAB A (2016) 10,1”, (RAM 2Gb, CPU Octa-Core 1.6 GHZ, 8 Mpx camera).
- Smartphone Samsung S7 SM-G930F (RAM 4Gb, CPU Octa-Core, 12 mpx camera).

Also digital cameras were used for taking pictures:

- Digital Camera Sony A5100 (24,4 Mpx APS-C sensor, lens Sony Lens G 18-55 mm)
- Digital Camera Sony Cybershot DSC HX300 (CMOS Exmor R 1-2/3 20,4 Mpx sensor, lens Zeiss Vario Sonnar T 4,3-215 mm).
- Digital Camera Canon Powershot SX729 HS (CMOS R 1-2/3 20,3 Mpx sensor, lens Canon 4,3-172 mm).

At the same time, the SME partners of the project were able to perform some tests on pottery assemblages from:

-Roman city of Pollentia (TSI, TSH, TSG, Amphora, Majolica pottery)

-Castle of Capdepera (Majolica, Amphora)

-Roman city of Ercávica (Amphora, TSI)

-Roman city of Libisosa (TSI)

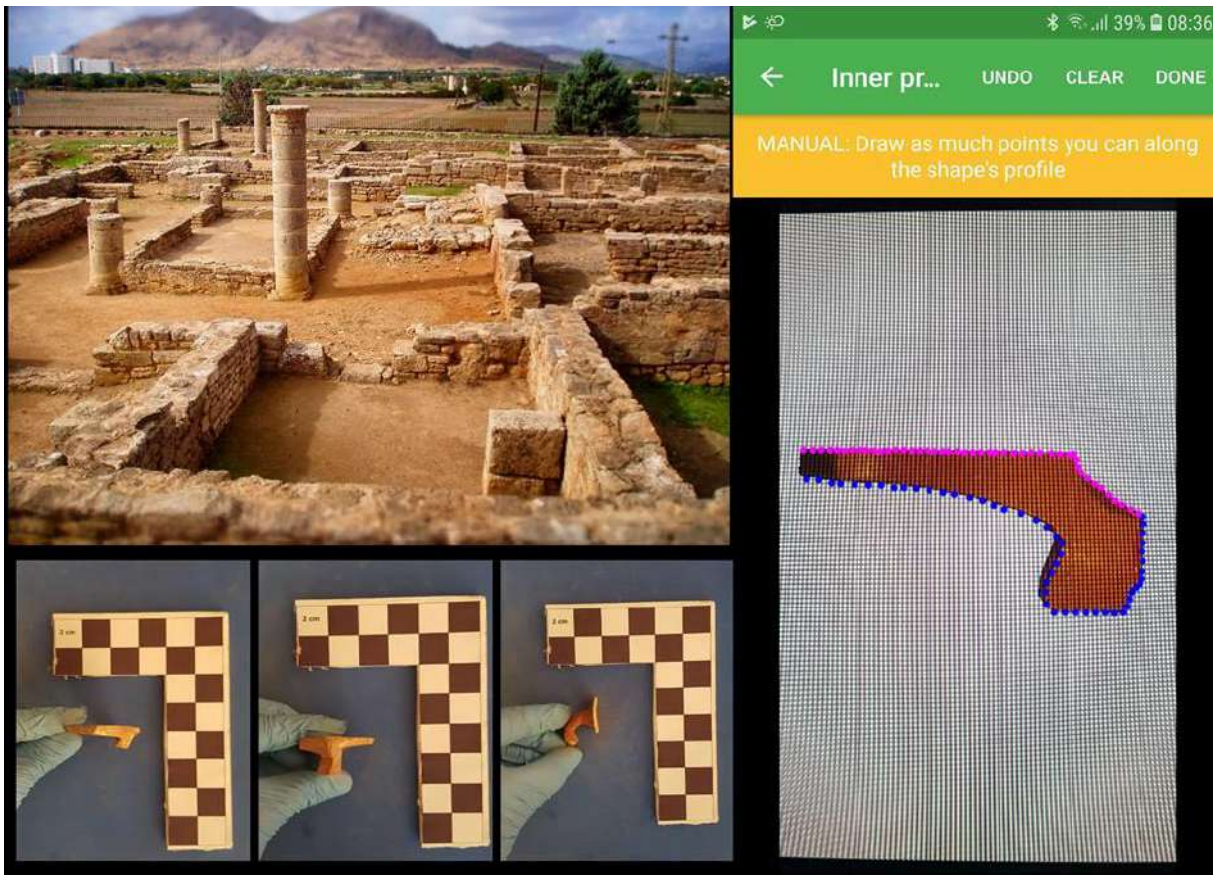

Fig. 1 - Terra Sigillata tests done by SME to the assemblage of the Roman city of Pollentia

2 Archaeological testbeds: HEI and Research Centers

In the case of HEI and Research Centres, the tests were carried out in the facilities of the centres with a selection of ceramic contexts already studied.

2.1. UB testbeds

From the UB it was possible to perform these testbeds not only at the UB's centre but also at the Archaeological Archive of the Museu d'Història de Barcelona. This means that the assemblages of the Roman pottery of the ancient Barcino were available for the testbeds (Amphora, TSI, TSH, TSG) as well as some examples of Maiolica of Montelupo.

For the appearance-based recognition tool, UB performed testbeds on 40 sherds of Montelupo pottery of the Museu d'Història de Barcelona from different excavations performed at the city of Barcelona. It is an important assemblage which testifies how Barcelona, as an important port of the Mediterranean sea, acted as a reception and redistribution centre where pottery of different provenances arrived. One of the types of ceramics that is possible to find is the Montelupo pottery. The identification of this type of majolica pottery is really recent, that is why is difficult to search it at the reports of the excavations performed before 2000. Although this limitation, it has been possible to test the image recognition tool with 40 sherds which belong to previous studies.

Fig. 2 - Montelupo pottery tested for the appearance-based recognition tool by UB

The study of the archaeological contexts of the 16th and 17th centuries in Barcelona (Beltrán and Miró 2010) demonstrated that a large amount of pottery from elsewhere was present in the city during the period. Valencian pottery, which until then had dominated the imported products, went into a severe decline during the course of the 16th century, and Italy became the main exporting country. Particularly pottery from Pisa and the Valdarno, Montelupo and Faenza, and especially from the region of Liguria are the most represented ones. This study did not classify the Montelupo pottery with the catalogue that ArchAIDE uses for the image recognition tool (Berti 1997). For this reason, testbeds also improved the classification of the museum for this assemblage that pretends to be a reference for SME working at the city of Barcelona.

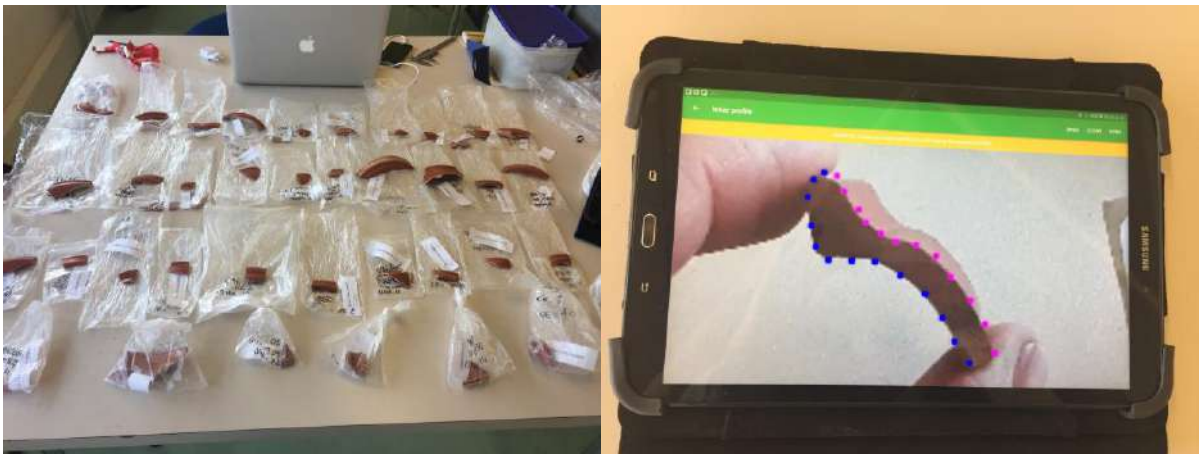

Fig. 3 - Terra Sigillata Italica prepared for the test of the shape-based recognition tool by UB with the mobile application

In the case of the shape-based recognition tool, it was possible to perform tests on the Terra Sigillata Italica founded at the ancient city of Barcino (Barcelona). The assemblage that was tested belongs to a huge deposit founded in the street Avignó of Barcelona. The materials had been previously studied by SME and stored at the Museu d'Història de Barcelona. The report already classified the materials but during the performance of the testbed, we were able to reclassify some of them. In some cases, the production was not the correct one (some sherds were labelled as Terra Sigillata Italica and they were Terra Sigillata South Gaulish, for instance). In terms of the typological classification, in some cases, the indications were not the correct ones and in most cases was not indicated. An aleatory sampling was performed in order to obtain up to 39 sherds for testing the system taking into account different sides of sherds and types.

In order to control all the modifications between the different versions of the app, every time changes were made, we verified the possible improvements with the same exact sherds, both for appearance-based and shape-based recognition tools. All pictures were taken with mobile devices (Smartphone Samsung Galaxy S6, Smartphone Samsung Galaxy J5, Smartphone Huawei Mate RNE-L21, Tablet Samsung Galaxy TabA, iPhone 6) and also digital camera (Panasonic DMC-LZ40).

2.2. UNIPI testbeds

From the UNIPI side, it was possible to perform testbeds on appearance-based and shape-based recognition. In the case of appearance-based recognition, it was possible to test the ArchAIDE system on Maiolica of Montelupo's materials stored into the museum of Montelupo warehouse at Montelupo Fiorentino. This gave us the possibility to check the performance of the application on a wide variety of decorative genres which allowed more robust considerations on how the application works into the field. The possibility to have easy access to the warehouse allowed UNIPI to perform the testbeds in three different periods: in March and November 2018, and in February 2019. This also allows to monitor the improvement of the recognition model through time. Moreover, in order to evaluate differences in the performance of the neural network, we take pictures both with mobile devices (Smartphone Samsung Galaxy S9, Tablet Samsung Galaxy TabS2) and a digital camera (Canon EOS 500d). The recognition process was directly implemented through the mobile application when the picture was taken with mobile devices and, indirectly, through the desktop application uploading in a personal computer the images taken with the digital camera (for the functionalities related to mobile and desktop application see D7.2.

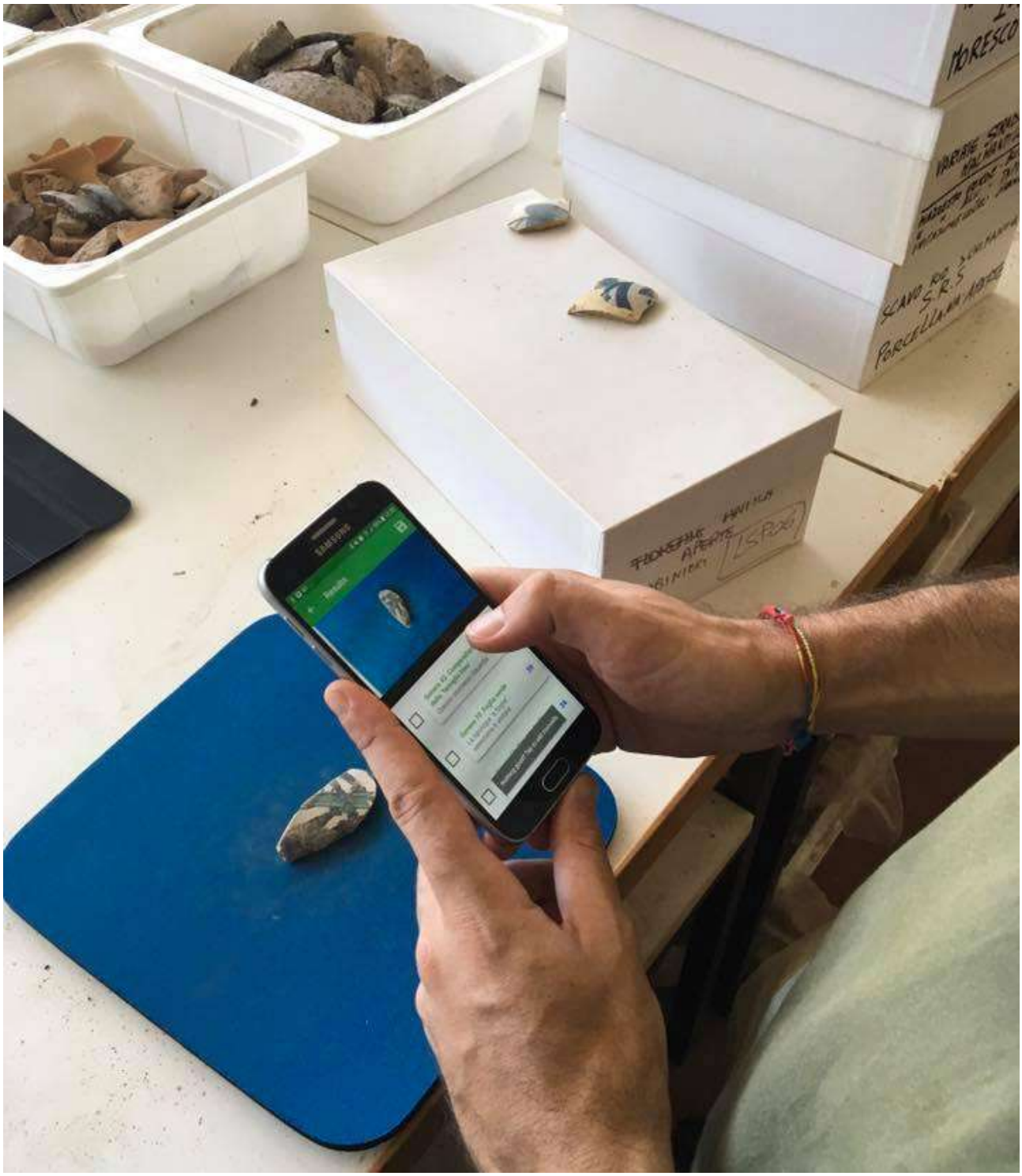

Fig. 4 - Testbeds on appearance-based recognition carried out in Montelupo Fiorentino by UNIPI with the mobile application

In the case of Terra Sigillata Italica, UNIPI worked with a set of materials coming from archaeological excavations carried on in Pisa and in the nearby, which are currently stored into the University's warehouse. Pisa was an important production centre of Terra Sigillata, for this reason, this class of pottery is very common to be found in urban archaeological excavation. Given the fact that the project partners which have seat in Spain tested the performances of the mobile application, UNIPI carried out its testbeds on the desktop application. All the potsherds were photographed with the camera incorporated in a Tablet Samsung TabS2, uploaded on a personal computer and processed through the desktop application (D7.2)

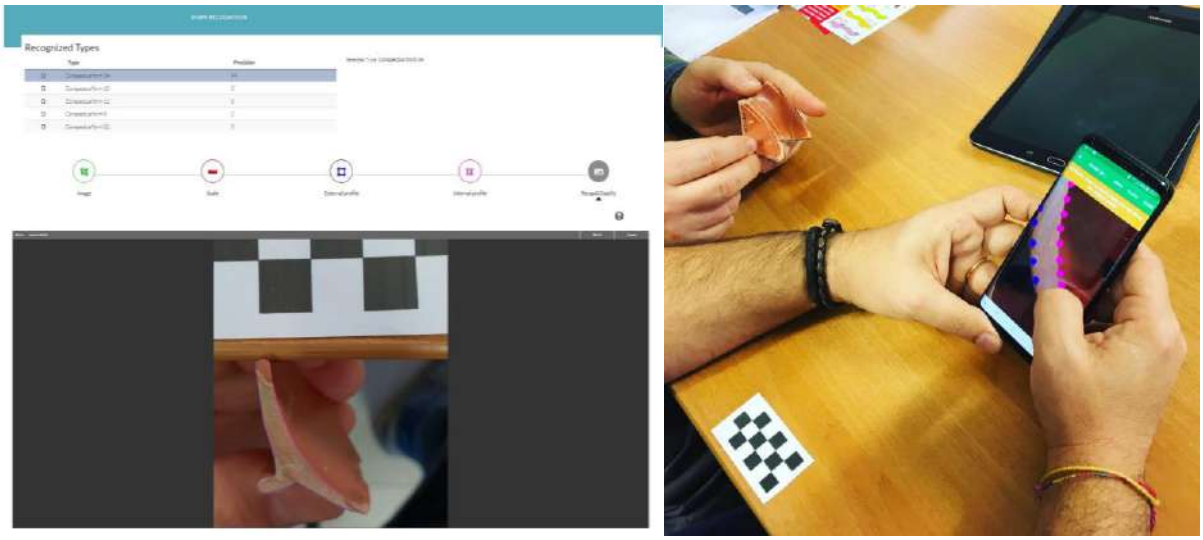

Fig. 5 - Testbeds on shape-based recognition carried out at UNIPI with the desktop application (on the left) and with the mobile application (on the right)

3. Methodology

Before starting the testbeds, an in-depth analysis of the main parameters that would have might influence the results was performed. Being aware of the different way in which the appearance-based model and the shape-based model work (D6.3), different parameters have been chosen and recorded during the testbeds.
For the appearance-based recognition, the system has to work with an image. For this reason, the parameters were chosen between the factors that could influence a good image and consequently a valid result:

- **the condition of the light** (artificial/natural); we wanted to test if the classification model was able to work into the most common light conditions that can be met in archaeological practice, i.e. in the sunlight, on the field and in artificial light, in a storehouse;
- **the quality of the camera**; we wanted to test if the model was efficient enough in working with devices that are more likely to be available in archaeological practice: high and medium quality mobile devices' cameras; medium quality digital camera sensor;
- **the dimension of the potsherds**; we wanted to test if the model was able to reach good results with small, medium and big size potsherds, given the fact that unearthed archaeological materials range from a few to tens centimetres;

For the shape-based recognition, the system has to work with the internal and external profiles extracted from an image. This mean that the process is not only influenced by the quality of the picture taken, but also by the quality of the vectorial file (.svg) produced tracing the profiles with the tools available into the system.
For this reason, the parameters were chosen between the factors that could influence a good tracing and consequently a valid output:

- **the size of the potsherds**; we wanted to test if the model was able to reach good results with small, medium and big size potsherds, given the fact that unearthed archaeological materials range from a few to tens centimetres;
- **the device used**; we want to understand if the quality of tracing is sufficient using either finger on a small touchscreen (i.e. smartphone), a medium touchscreen (i.e. tablet), or a mouse on a PC screen (D7.2);
- **the part of the vessel**; even taking into account that the rim is the most diagnostic part of a vessel, we want in any case to test if the model was robust enough with all the parts in which a vessel can be broken (Rim, Foot, Body or Handle);

- **the fracture produced**; the way in which a vessel can be broken produces fractures that can be more or less vertical, this feature can influence the quality of the extracted profile, and we want to test how it affects the performance of the recognition system.

	A	B	C	D	E	F	G	H	I	J	K
1	ID INPUT	SETTINGS	DEVICE	DEVICE	LIGHT	LIGHT TYPE	GEN_INPUT	SIZE	ORDER_OUT	GEN_OUTPUT	SCORE
42									1	23	95
43									2	53.1	1
44	3	1	Smartphone	samsung S6	Natural	N	23	S	3	33	0
45									4	19	0
46									5	30	0
47									1	23	98
48									2	50	0
49	3	2	Smartphone	samsung S6	Artificial	F	23	S	3	19	0
50									4	20	0
51									5	33	0
52									1	23	99
53									2	19	0
54	3	3	Tablet	Samsung TabS2	Natural	N	23	S	3	50	0
55									4	34	0
56									5	28	0
57									1	23	76
58									2	62	12
59	3	4	Tablet	Samsung TabS2	Artificial	F	23	S	3	72	1
60									4	6	1
61									5	29	1
62									1	63	61
63									2	55	12
64	4	1	Smartphone	samsung S6	Natural	N	55	M	3	54	8
65									4	72	5
66									5	68	1

Fig. 6 - The shared google spreadsheet used by partners for recording the results of appearance-based recognition achieved with the mobile application

Data collection has been carried out following a randomisation procedure in order to have approximately the same number of photographs for each possible setting. Specifically, data collection for appearance-based testing randomised the following parameters, since they could have been sensible parameters to consider:

- Natural light VS artificial light;
- Tablet, smartphone and camera devices;
- Size of sherds, by measuring the biggest dimension and classifying the sherd as big if the biggest dimension is greater than 10 cm, small if less than 5 cm, medium if between 5 and 10 cm.

Data collection shape-based testing randomized the following parameters:

- Natural light VS artificial light;
- Tablet, smartphone and camera devices;
- Size of sherds, by measuring the biggest dimension and classifying the sherd as big if the biggest dimension is greater than 10 cm, small if less than 5 cm, medium if between 5 and 10 cm;
- The part of the sherd distinguished as Rim, Body, Foot or Handle;
- Fracture, distinguished as Vertical or Oblique.

The automatic classification has been noted down taking into account which were the conditions (natural/artificial light) when the photo was taken, the type (smartphone/tablet) and model of the device that was used and in which position within the five possible answers appeared the correct classification. Statistical analysis has been performed to all of these data in order to be able to explore and evaluate the results obtained. In addition, we have individually tested different fragments of the same vessel to see if the tool provides in all cases the same five possible answers and in the same order.

4. Analysis of the results

We tested the ArchAIDE app appearance-based recognition performances both on the desktop and mobile devices.

4.1 Appearance-based recognition

The following analysis has been conducted on the basis of 274 different pictures of sherds, taken from 49 different genres out of 84 genres. While the sample cannot be considered representative of the whole population of ceramic genres, it includes the most common genres found in archaeological excavations. In order to study the possible dependence of app performance to practical parameters, we took note also of the following data, which served as analysis dimensions.

Device:

- 123 pictures have been taken from a **camera**;
- 620 have been taken from a **smartphone**;
- 135 have been taken from a **tablet**.

We took also notes of the device models, in the case in the future it can be useful for comparing different hardware performances. Settings of the photograph:

- 310 have been taken at **artificial light**
- 305 with **natural light**.

Size of sherds:

- 180 were sherds of **big size**, i.e. their biggest dimension was greater than 10 cm;
- 590 were sherds of **medium size**, i.e. their biggest dimension was greater between 5 cm and 10 cm;
- 600 were sherds of **small size**, i.e. their biggest dimension was smaller than 5 cm.

The first step was to establish an average accuracy of the app in recognising the genre of the sherd's decoration. We computed two different accuracies, denominated top-1 and top-5. As for the top-5 accuracy, a result is considered right if the right answer is (anywhere) in the 5 output by the app. As for top-1 accuracy, a result is right if it is the first answer.
However, genres are recognised with different ease (both by archaeologists and by the app), so in averaging the proportion of cases where right genre (i.e. the same as input) appeared in 5 outputs returned by the app, we have considered each input genre with the same weight. Hence the average accuracy has been computed by weighting each genre by the inverse of its frequency in the test.
Some tests on the app performances have been run for desktop and mobile separately, in order to highlight possible differences. In the case of scores asserting the confidence of outputs, instead, we aggregated the results because they show the same pattern.

4.1.1 Desktop platform performances

The average desktop app top-5 accuracy is 77.2%, while average (top-1) accuracy is 51%. This accuracy, which seems to us a good result, is the process of improvement both of the neural network and in the general workflow. In the following table you can find different accuracy as measured along time.

DATE	DESKTOP TOP-5 ACCURACY	DESKTOP TOP-1 ACCURACY
March 2018	27.9%	9.9%

November 2018	68.2%	44.6%
February 2019	77.2%	51%

Table 1 - Top-5 and top-1 accuracy of the desktop application in the different testing procedures across time

We computed the top-5 accuracy with respect to the dimension of fragments, and we found that it is 56.9 % for small fragments, 20.3 % for medium-sized fragments and 2.4% for big fragments. These results would tell us that smaller fragments are better recognised than bigger ones. However, such percentages are not significant *per se* and are somewhat reverted in case of mobile devices. Indeed they are more related to the genres than to the size of sherds. In order to highlight the relationship of accuracy with size and genres together many more data would be needed: this is a matter of future investigation we are carrying on.

Accuracy is not related to the light type, being approximately equal with artificial and natural light. We noted the same for mobile devices, so it is going to be no more a parameter of interest.
While accuracy does not show a statistically significant correlation with the number of photos available for each genre, we noted that some genres are output more frequently than others, independently on the input genre. As it can be seen by the following table, genre 10_3, 25 and 29 appears together about 100 times over the total of 615 output shown by the app.

Freq	GENRE
39	10_3
31	25
31	29
28	50
27	20
25	13
24	38
23	18
23	40
22	62

Table 2 - Highest frequencies of genres output, independently on the input, shown by desktop app. Frequencies are on the first column and ceramic genres on the second

If we consider the output genres shown by the desktop app, without taking into account the correspondent input, there is a statistically significant association between the frequencies of outputs and the number of photos available for each genre. The same association is also present on the mobile side, though weaker. In the following figure, we show Number of pics available for each genre (on y-axis) VS frequency of genres in the output. The clear trend is confirmed by a statistically significant linear regression test (p-value < 2.1e-06), whose estimated slope is 7.55.

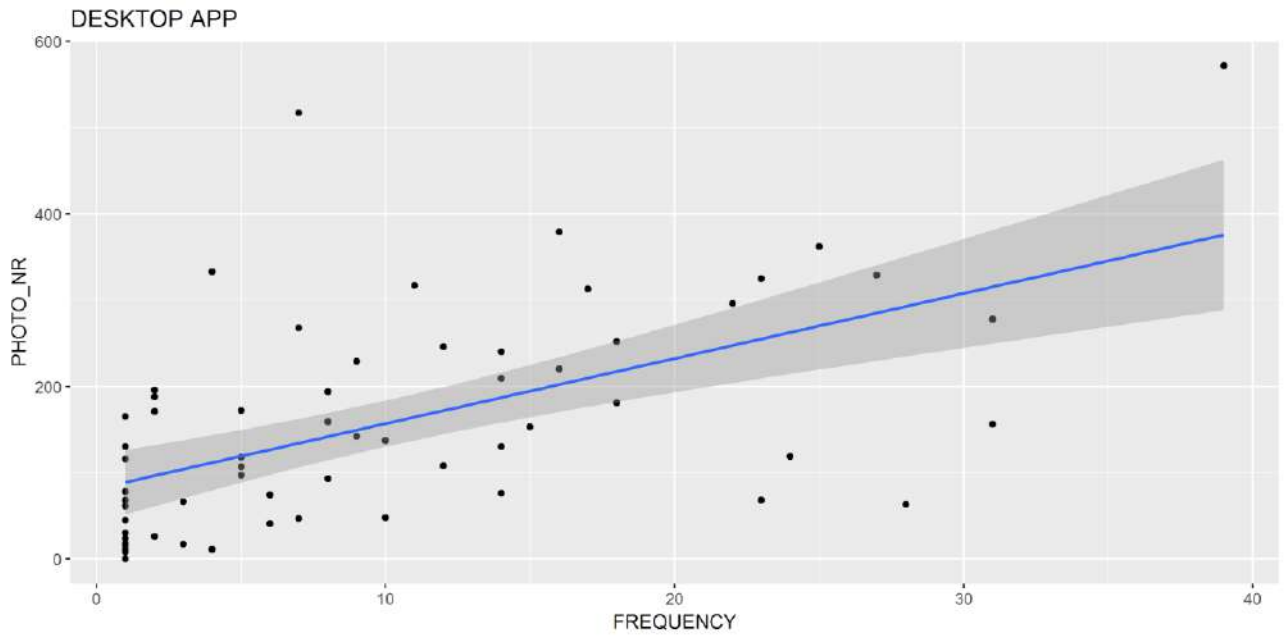

Fig. 7 - Number of pics available for each genre (on y-axis) VS frequency of genres in the output (x-axis), for the desktop app testing procedure.

Are there any recurrent or more frequent mistakes in the desktop app? Yes. The following table shows them ordered by error proportion over the number of times input is presented. For example, the genre 10_3 (column GEN_INPUT) has been presented 12 times (column N_INPUT): 8 times (column FREQUENCY) the outputs include the genre 13 (column GEN_OUTPUT), corresponding to a percentage of 66.7% (column PERCENTAGE). For understanding this behaviour, we have to consider that the classification of the genres has been made using history of art methods based on pattern discontinuity and chronology. In the case of genres 10_3 and 13, the recurrence of mistakes could be explained with the similarity between them, especially taking into account that genre 10_3 is characterised by wide differences between patterns that belong to the same genre. In the case of genres 26 and 18, it could be explained with the recurrence of the rhombus pattern, while in other cases, such as 25/29, the recurrence is more difficult to explain in archaeological terms and should be further investigated.

GEN_INPUT	GEN_OUTPUT	FREQUENCY	PERCENTAGE	N_INPUT
10_3	13	8	66.7	12
10_3	20	7	58.3	12
25	29	6	85.7	7
10_3	9	5	41.7	12
13	10_3	4	100.0	4
25	13	4	57.1	7
26	18	4	100.0	4
18	25	4	100.0	4
10_3	29	4	33.3	12
45	40	4	100.0	4

Table 3 - Frequent mistakes shown by the desktop app: input genre is in the column GEN_INPUT; output genre is in the column GEN_OUTPUT, the number of times the input genre has been presented is in the column N_INPUT; the number of times the input genre has been (erroneously) recognised as output genre is in the column FREQUENCY; the proportion of FREQUENCY over N_INPUT is in the column PERCENTAGE.

4.1.2 Mobile App performances

The average mobile app top-5 accuracy is 83.8% and top-1 accuracy is 55.2%. This accuracy, again, resulted as a process of improvement both of the neural network and in the general workflow. In the following table, you can find the different accuracy as measured along time

DATE	MOBILE TOP-5 ACCURACY	MOBILE TOP-1 ACCURACY
March 2018	33.3%	12.1%
November 2018	70.3%	38.9%
February 2019	83.8%	55.2%

Table 4 - Top-5 and top-1 accuracy of the mobile application in the different testing procedures across time

We computed the top-5 accuracy with respect the dimension of fragments, and we found that it is 20.5 % for small fragments, 45 % for medium-sized fragments and 19.4% for big fragments. Again, such percentages are not significant *per se*, and are somewhat reverted in case of mobile devices. Indeed they are more related to the genres than to the size of sherds. In order to highlight the relationship of accuracy with size and genres together many more data would be needed: this is a matter of future investigation we are carrying on. The accuracy of appearance-based recognition on mobile devices is not related to the light type, being approximately equal with artificial and natural light.

Accuracy is not related to the light type, being approximately equal with artificial and natural light. As in the case of the desktop app, does not show a statistically significant correlation with the number of photos available for each genre. However, some genres appear as output (independently of the related input) more often than others. The following table is the equivalent of the one in the paragraph describing the desktop performances:

Freq	GENRE
52	13
48	20
41	50
40	25
37	38
36	26
36	53_1
35	21
29	29
27	10_3

Table 5 - Highest frequencies of genres output, independently on the input, shown by mobile app. Frequencies are on the first column and ceramic genres on the second.

Also in the case of the mobile app, there is a statistically significant association between the frequencies of outputs (without considering corresponding inputs) and the number of photos available for each genre. In the following figure, it is shown the number of pictures available for each genre (on y-axis) VS frequency of genres in the output. A linear regression results in a statistically significant test (p-value < 0.0006), with estimated slope 4.25.

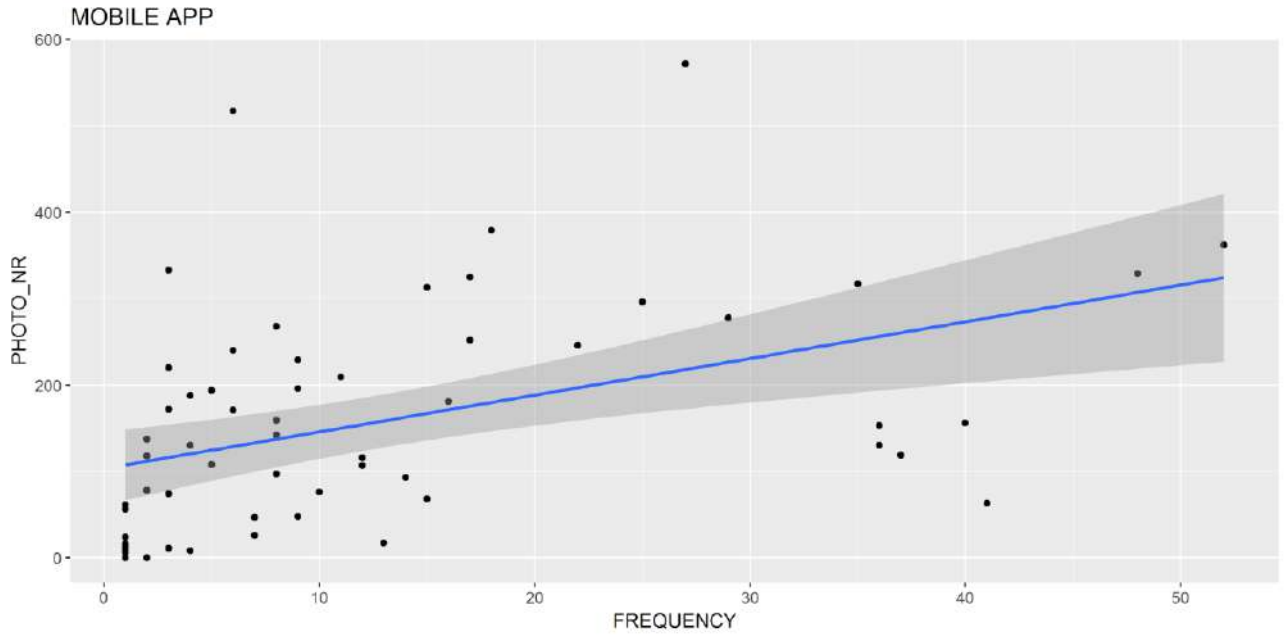

Fig. 8 - Number of pics available for each genre (on y-axis) VS frequency of genres in the output (x-axis), for the mobile app testing procedure.

In the following table, more frequent mistakes are shown, ordered by error proportion over the number of times input is presented. Looking for example at the first row of the table, the genre 18 (column GEN_INPUT) has been presented 7 times (column N_INPUT): 6 times (column FREQUENCY) the outputs include the genre 13 (column GEN_OUTPUT), corresponding to a percentage of 85.7% (column PERCENTAGE).

GEN_INPUT	GEN_OUTPUT	FREQUENCY	PERCENTAGE	N_INPUT
18	13	6	85.7	7
18	26	6	85.7	7
6	1	4	66.7	6
53_1	13	4	80.0	5
18	16	4	57.1	7
10_3	20	4	80.0	5
20	21	4	100.0	4
33	21	4	100.0	4
10_3	29	4	80.0	5
68	29	4	80.0	5

Table 6 - Frequent mistakes shown by the the mobile app: input genre is in the column GEN_INPUT; output genre is in the column GEN_OUTPUT, the number of times the input genre has been presented is in the column N_INPUT; the number of times the input genre has been (erroneously) recognised as output genre is in the column FREQUENCY; the proportion of FREQUENCY over N_INPUT is in the column PERCENTAGE.

4.1.3 Confidence in classification

The app shows, both in the mobile and in the desktop version, 5 results together with 5 scores assessing the confidence of the suggested classification. In order to test whether and how much a higher score is associated with better predictions, we have computed the differences between consecutive scores shown by the app results at each trial, and take the maximum. The maximum of the score differences represents a measure of the strength specific results are pushed against the others.

In the following, we show the distribution (histogram) of the maximum of the score differences. Maximum of the score differences (maxdiffscore) is on the x-axis, percentages are on the y-axis.

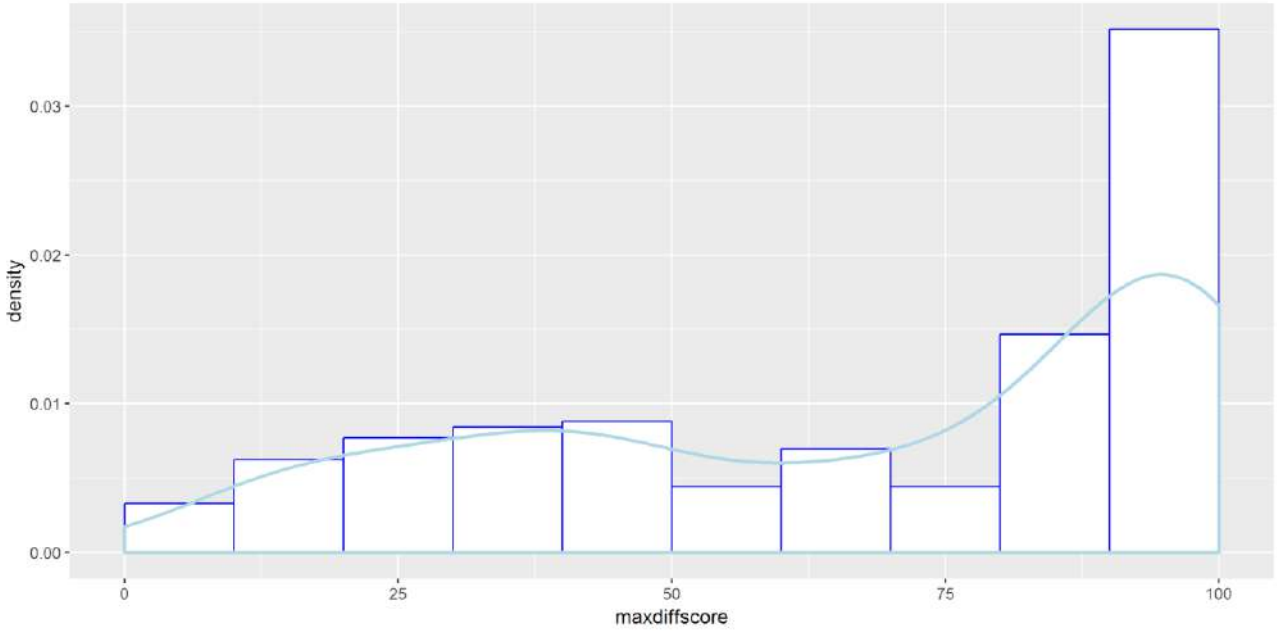

Fig. 9 - Distribution (histogram) of the maximum of the score differences.

Maximum of the score differences have been then tested against the top-5 accuracy by building an accessory binary variable valued 1 when the right answer was present in the 5 outputs and 0 otherwise. Maximum of the score differences are here plotted against **top-5** presence of the right answer. T-test for comparing the means of the two groups and Mann-Whitney test for comparing the medians of the two groups are significant. Both are standard tests in statistics. Means are 64.3 (on top-5 presence equal to 0) and 79.3 (on top-5 presence equal to 1), while respective medians are 68.5 and 89.5.

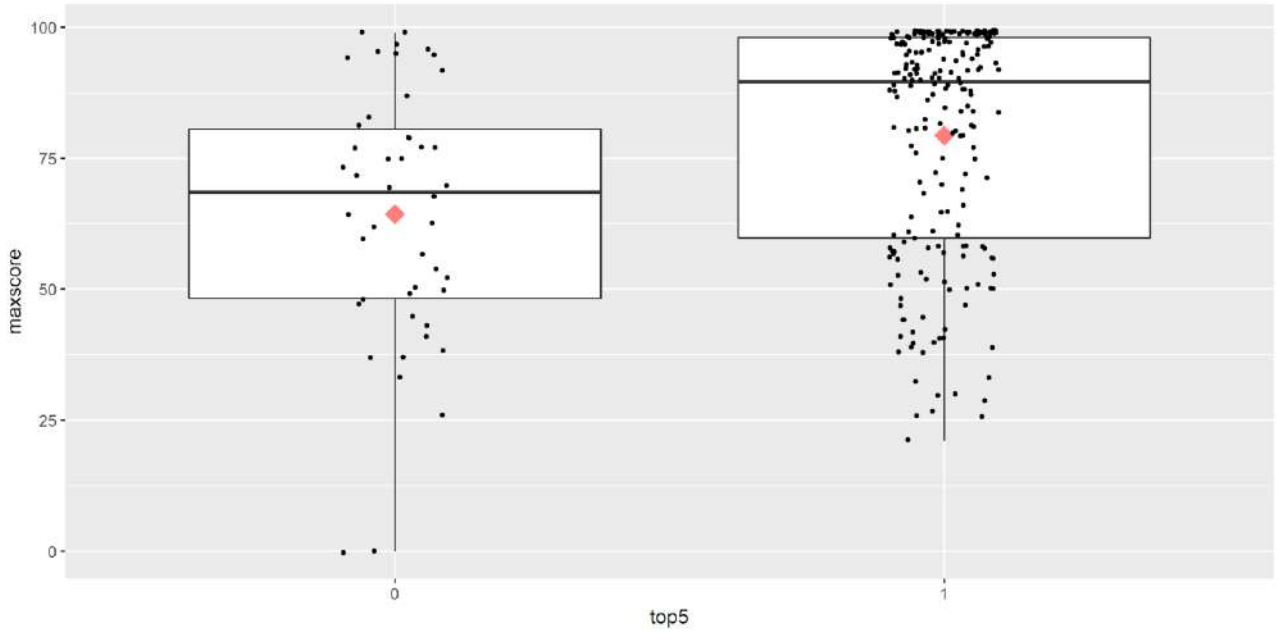

Fig. x - Maximum of the score differences VS top-5 presence of the right answer.

In order to get similar results for top-1 accuracy, we built an accessory binary variable valued 1 when the right answer was present in the first output, and 0 otherwise. Maximum of the score differences are here plotted against **top-1** presence of the right answer. T-test test for the mean of the two groups and Mann-Whitney test for the median of the two groups are significant. Means are 67.3 (on top-1 presence equal to 0) and 84.8 (on top-1 presence equal to 1), while respective medians are 68 and 93.

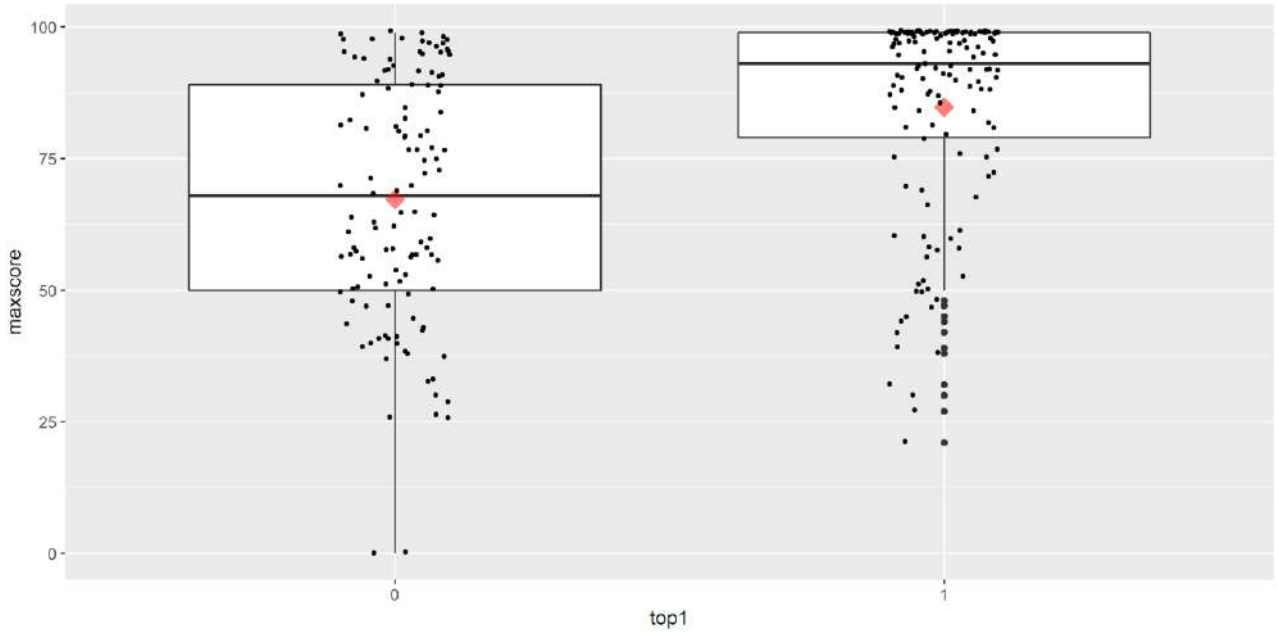

Fig. 10 - Maximum of the score differences VS top-1 presence of the right answer.

According to the data we have collected, we have tried to identify a score better distinguishing with right from wrong predictions. in order to do that, we have related the maximum score of each trial to the

proportion of object rightly identified, in the following way. Consider the top-5 prediction. For each score threshold x, 0<x<100, we have

1. The accessory binary variable valued 1 when the right answer was present in the 5 outputs and 0 otherwise. Call it top-5 classification.

2. The prediction on the top-5 classification made by classifying as 0 all the trials having maximum score less than x, and as 1 all the trials having maximum scores bigger or equal than x. Call it top-5 prediction.

In this way, we can build a ROC curve[1] by plotting, as in the following figure, for each possible threshold x, the true positive fraction VS the false positive fraction, i.e. for each threshold x

- the true positive fraction (y-axis) is the proportion of top-5 classification correctly addressed as 1 by top-5 prediction;
- the false positive fraction (x-axis) is the proportion of top-5 classification erroneously addressed as 1 by top-5 prediction.

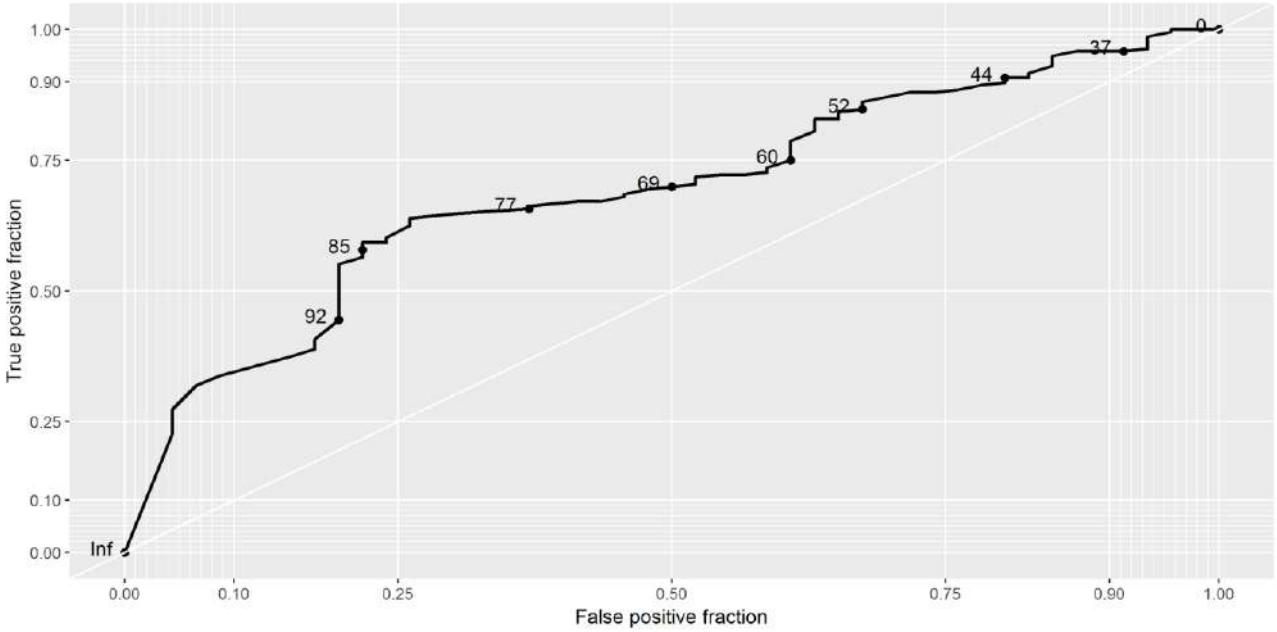

Fig. 11 - ROC curve of top-5 classification VS top-5 prediction.

Based on the ROC curve, the optimal threshold is identified as score=80, giving a specificity of 74% and sensitivity of 64%.

In the case of top-1 prediction, we followed a similar approach. For each score threshold x, 0<x<100, we have

[1] https://www.sciencedirect.com/science/article/abs/pii/S0031320396001422

3. The accessory binary variable valued 1 when the right answer was the first output and 0 otherwise. Call it top-1 classification.

4. The prediction on the top-1 classification made by classifying as 0 all the trials having maximum score less than x, and as 1 all the trials having maximum scores bigger or equal than x. Call it top-1 prediction.

In this way, we get a ROC curve as in the following figure.

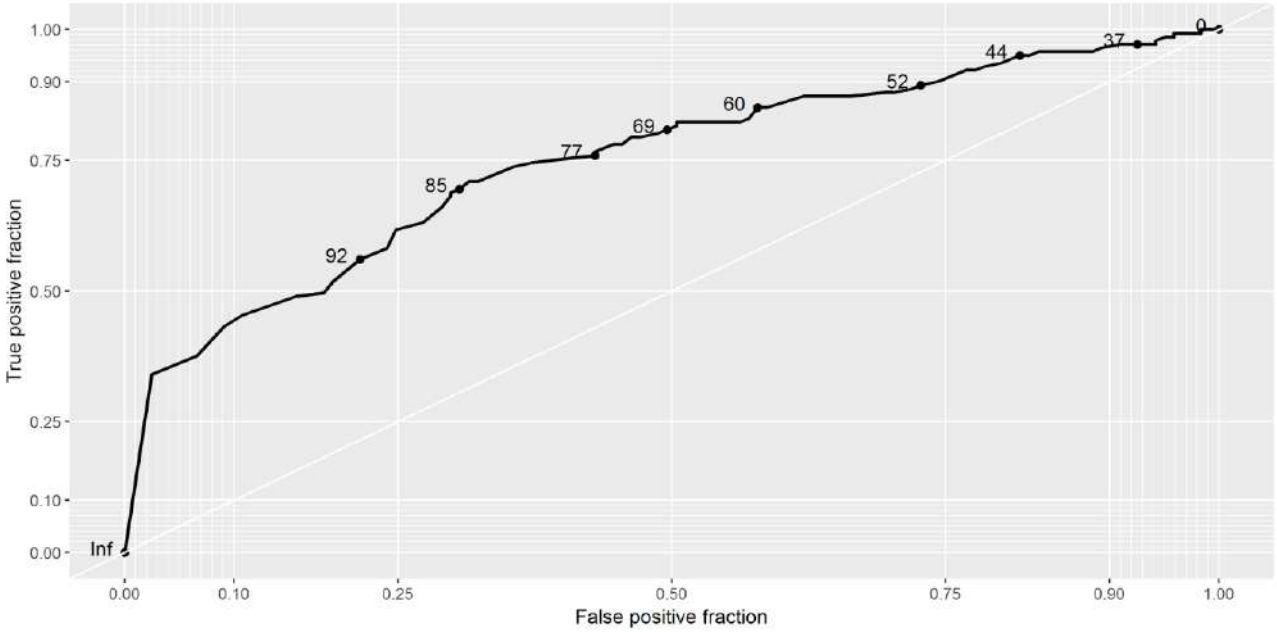

Fig. 12 - ROC curve of top-1 classification VS top-1 prediction.

Based on the ROC curve, the optimal threshold is identified as score=84, giving a specificity of 69% and sensitivity of 70%.

The above results clearly suggest that higher scores are associated with better predictions. This could seem quite obvious information, but it brings **high value in terms of the improvement of app performances**. Indeed, there is a variability of outputs when the app is tested on the same sherd different times. This variability is perfectly normal and is given for instance by

- inherent random component of the neural network processing;
- different condition of object position, such as rotation or centring;
- different application of crop tools, zooming in or out, focusing or not on some details.

Since higher scores are associated with better performances, a sensitive improvement can be achieved by taking more pics of the same object by randomly applying various rotation, cropping and zooming conditions, and then keep the (first) five outputs based on their average or consensus score. **This can boost the app recognition performances significantly, without acting on the underlying neural network.** Moreover, it can be also a way to limit the inherent variability due to photograph settings.

4.2 Shape-based recognition

The tests conducted on the ArchAIDE app shape-based recognition performances are analogous to those on shape-based recognition.

The accuracy of the shape-based recognition is not as high as the appearance based recognition, and exhibits a big difference between desktop and mobile devices. This prevented us to segment the analysis in a more detailed way, i.e. by using further analysis dimensions. The results would not have been interpretable in the right terms, since we haven't got enough high accuracy in order to understand how it is related to these analysis dimensions. So, for instance, we haven't analysed how higher scores are related to better performances, because we do not have enough high score in order to get a meaningful result. Another important reason for not considering further segmentation is the fact that desktop app performances are extremely lower, as compared to mobile app ones.

However, we took notes of different dimensional analysis, i.e. size of sherds, fractures, part of the vessels the sherds belong to, and the device models. These other dimensions of analysis are not expressly present in this deliverable, since prediction levels show no adequate accuracy. This further information will be useful in future investigations for segmenting accuracy levels in a more detailed way.

The analysis has been conducted on the basis of 381 different pictures of sherds, taken from 42 different types. As for the device:

- 25 pictures have been taken from a **camera**;
- 210 have been taken from a **smartphone**;
- 146 have been taken from a **tablet**.

As in the case of appearance-based performances, the average accuracy has been computed by weighting each genre by the inverse of its frequency in the test.

4.2.1 Desktop application performance

The average desktop app top-5 accuracy is 17.5%, while average (top-1) accuracy is 4.8%. Desktop recognition has been limited to 25 trials because of this very limited accuracy. Before proceeding, in the future, with further analysis better specifying the role of photograph settings or object feature in the recognition performances, we must understand the reasons for this low accuracy level.

4.2.2 Mobile application performance

As for mobile shape-based recognition performances, the average mobile app top-5 accuracy is 50.8% and top-1 accuracy is 18.9%.

The types that appear most frequently in the outputs are also the types that are most present in the input, so we had no reason to investigate this aspect further.

In the following histogram, we show the distribution of the maximum of the score differences. Maximum of the score differences (maxdiffscore) is on the x-axis, percentages are on the y-axis. Unlike the appearance-based case, we see that scores indicated low confidence on the results since low maximum of the score differences is mostly present in the outputs. This has been done with desktop and mobile app together, so it expresses a real improvement on the performances needed. Indeed, also the analysis of the impact of higher scores on the accuracy of prediction is not statistically significant.

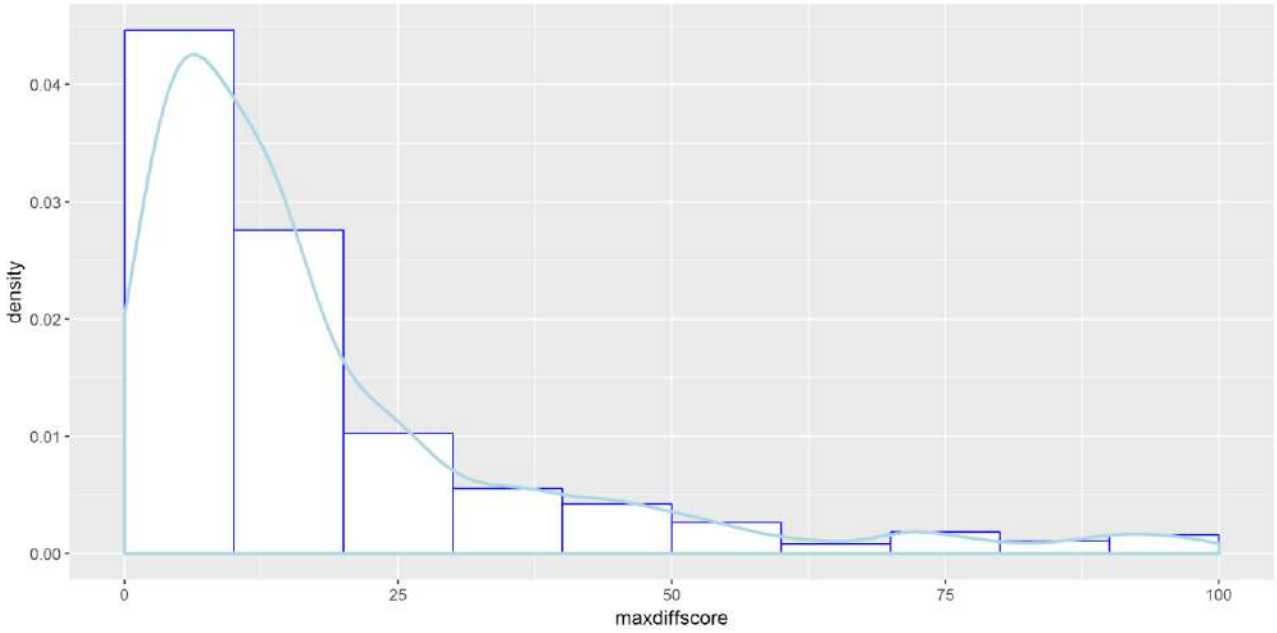

Fig. 13 - Fig. 9 - Distribution (histogram) of the maximum of the score differences.

5. Final recommendation

After performing the testbeds, the aim of this deliverable is to enumerate final recommendations both for optimising the App and for better user experience.

In the case of appearance-based recognition, feedback from the users and data analysis enlighten the following

pros

- the user experience is positive both with mobile devices and desktop application. In the case of using the application in the field, the mobile phone is more appropriate because of its easier operability;
- the pipeline is really simple and very fast. Each step is processed immediately after pushing the button, with any delay in the answer;
- it is very easy to take a picture of a potsherd and to modify it (cropping, flipping, rotating) with fingers. The experience is very similar to other well-known application;
- the recognition model is not influenced by the device/application used. Results are robust with all the devices used;
- particular skills by the users are not necessary, in order to obtain good results;
- a higher score is associated with correct output;
- given the easiness and fastness of use, appearance-based recognition tool can easily be used in-the-field both in excavation and post-excavation practice.

and cons:

- mistakes could happen in the last screen where it is not enough clear that the big ‘Process” button is for launching the GrabCut algorithm for eventually excluding the foreground from the image, while the button for launching the recognition system is the “done” button on the top-right of the screen;
- the desktop application has not been implemented with the GrabCut Algorithm;
- the use of GrabCut algorithms produced lower scores in the outputs.

Finally,

(1) for improving the user-experience we suggest to:

- modify the last screen in order to avoid mistakes in launching the GrabCut algorithm instead of the recognition process;

(2) for improving the results we suggest to:

- simply crop the image, instead of using the GrabCut algorithm;
- crop the image in order to centre as much as possible the decoration (small details give better results than a larger scene);
- rotate the image in order to orientate the potsherd as to have the rim towards the top of the screen. This is the way archaeologists generally take potsherds' pictures, moreover, this is the way the pictures for training the neural network were taken;
- implement an automatic tool inside the recognition process for randomly applying various rotation, cropping and zooming conditions, and then keep the (first) five outputs based on their average or consensus score. **This can boost the app recognition performances significantly, without acting on the underlying neural network.** Moreover, it can be also a way to limit the inherent variability due to photograph settings.

In the case of shape-based recognition, the skills on tracing of the user are really relevant if we want to have a correct answer from the system. In order to obtain good results, it is better to take into account the following aspects:

- the user needs to take care of the orientation of the sherd. This is important thus it really can make the difference in the answer of the system;
- sometimes it can be hard to take a photo taking into consideration the fact that the ruler should be at the same level of the fracture, in order to avoid problems in scaling, and the correct orientation of the potsherd. This problem is more evident with larger devices like tablets. The user needs to be able to correct the orientation with the edit tools. In this sense, a minimum of skills on the knowledge of how ceramic should be orientated is needed;
- it is necessary to perform a precise tracing. This means that the process is not so automatic as in the appearance-based recognition tool: it needs to be accurate and not random;
- tracing in smartphones is harder and much more difficult than in tablets, thus this might be the reason why the results are not as good as in tablets;
- for the scale, is a good option to just measure a part of the sherd (the distance between the top part of the rim and the wall, for instance) than using a scale while the picture is taken.

pros

- the shape-based recognition tool outputs are not influenced by the mobile device used;
- it is really useful to be able to flip the image thus not always the good and clean cut of the sherd is on the left side of the profile.
- it is very easy to rotate the image in order to correct the orientation of the sherd;
- the app has the ability to correct the white balance, which reduces the influence of light quality;
- the tracing tool allows to delete (on all the applications) or move (only in the desktop application) wrong points;
- the tracing tool allows the zooming of the image;
- although many of the sherds present oblique cuts, it is easy to focus the image in order to perform the tracing.

and cons:

- most of the responsibility on the final result belongs to the user. That means that if the classification is not the correct one it is not possible to know if it was the users or system's fault;
- the desktop application's outputs are not as reliable as the mobile application's ones;
- only the manual tracing tool works properly, the automatic and semi-automatic tools for the tracings are not working properly. They should be removed in the next version of the application;
- every time a point of the tracing is deleted, the screen goes back to the original zoom of the image
- the desktop application presents an easier tool for tracing which should provide better results for the shape-based classification, but it does not. This tool is close to the way archaeologist vectorise their drawings and it might be much more understandable for the user.
- every time a point of the tracing is deleted, the screen goes back to the original zoom of the image;
- the external profile is more visible when a point of the internal tracing has been deleted;
- for professional archaeologists, the tool seems more useful during the post-excavation study phase than during the excavation activity, when the work priorities are different.

Finally,

(1) for improving the user-experience we suggest:

- to improve the tracing for avoiding the cons enlightened;
- to have a preview the photo of the sherd we want to classify at the same time we look at the drawings contained in the knowledge.base and related to the five outputs.

(2) for improving the results we suggest to:

- create a filter in order to discriminate and tell the system which part of a vessel the user is going to trace (Rim, Foot, etc.);
- explore the possibility to include the vectorise drawings of archaeologist as a possible file to be classified by ArchAIDE.

References

Beltrán de Heredia Bercero, J. and Miró Alaix, N. (2010): El comerç de ceràmica a Barcelona als segles XVI-XVII: Itàlia, França, Portugal, els tallers del Rin i Xina, QUARHIS, 6: 14-91.

Berti, F. (1997): Storia della Ceramica di Montelupo, vol. 1 and 2, Cinisello Balsamo 1997.