2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)

IArch : An AI Tool for Digging Deeper into Archaeological Data

1st Chahrazed Labba
University of Lorraine
CNRS, LORIA
Nancy, France
chahrazed.labba@loria.fr

2nd Ameline Alcouffe
Paul Sabatier University
CNRS, CAGT
Toulouse, France
ameline.alcouffe@univ-tlse3.fr

3rd Eric Crubézy
Paul Sabatier University
CNRS, CAGT, IUF
Toulouse, France
eric.crubezy@univ-tlse3.fr

4th Anne Boyer
University of Lorraine
CNRS, LORIA
Nancy, France
anne.boyer@loria.fr

***Abstract*—The use of Artificial Intelligence (AI), notably Machine Learning (ML), is gaining momentum in archaeology, opening up new possibilities such as artifact classification, site location prediction, and remains analysis. One of the major challenges in this regard is the lack of qualified archaeologists who are experts in machine learning. In this study, we introduce IArch, a tool that enables eXplainable Artificial Intelligence (XAI) data analytics for archaeologists without requiring specific programming skills. It specifically allows data analysis performed to either validate existing data-supported hypotheses or generate new ones. The tool covers the entire workflow for applying ML, from data processing to explaining the final results. The tool allows the use of supervised and unsupervised ML algorithms, as well as the SHapley Additive exPlanations (SHAP) technique to provide archaeologists with global and individual explanations for the predictions. We demonstrate its use on data from a Xiongnu cemetery (100 BC/AD 100) in the Mongolian steppes.**

***Index Terms*—Machine Learning, supervised, unsupervised, Explainability, Archaeology, classification, clustering**

I. Introduction

Artificial Intelligence (AI), particularly Machine Learning (ML), is gaining prominence in a variety of domains, including archaeology. According to the Oxford dictionary [1], archaeology is defined as ”the study of human history and prehistory through the excavation of sites and the analysis of artifacts and other physical remains”. This inclusive definition also includes funerary archaeology or anthropology, a speciality to study funerary rites, necropolises or graveyard from the past more or less ancient – like predynastic necropolis or 17th century graveyard -, human remains deposited in a grave, clothing appearance or jewellery of the deceased one, genetic data or epidemiological data and so on. Therefore, data collection can take years, if not decades and is delivered in enormous chunks of complex, heterogeneous, and uncertain data, resulting in scattered and messy datasets. ML makes it possible to process this type of data and create models from it, which can then be used to interpret additional data. Indeed, we distinguish various types of ML applications to archaeology, such as the analysis of archaeological remains [1], [2], the prediction of archaeological sites [3], [4] and artifact classification [5], [6].

However, according to [7], while the benefits of machine learning for archaeology are obvious, the contribution of an archaeologist to machine learning applications is not. Further, the most significant barrier is the requirement for a skilled archaeologist who is also a specialist in machine learning. To overcome this challenge, some research works [7], [8] focused on presenting ML applications in archaeology and describing the applied ML models. Others [9], [10] focused on introducing AI-based technologies to assess archaeologists with image analysis (e.g. recognition of pottery artifacts,restoring ancient texts). As far as we know, current tools only support images as a data type. Further, an important element has been neglected in the literature, which is the lack of tools designed around explainable ML approaches that do not require archaeologists to be ML experts in order to use and understand them. Indeed, explainability is essential for establishing trust in AI and delivering more in-depth data insights. In ML, explainability allows to learn more about the underlying data that was used to create the model by describing why an AI model produces a particular prediction. Using an explainable ML methodology to analyze archaeological data necessitates the usage of many libraries published in various programming languages. As the use of these libraries frequently necessitates varying amounts of coding, they are largely designed for data science developers rather than domain experts.

The contribution of this work is twofold : **i)** the IArch tool that allows archaeologists to accomplish eXplainable Artificial Intelligence (XAI) data analyses without having specific programming expertise. The tool covers the complete ML workflow, from data processing and feature selection to applying the ML models and explaining the predictions using the SHapley Additive exPlanations (SHAP) [11]. **ii)** The validation of existing hypothesis and the generation of new ones using data from a Xiongnu cemetery (100 BC/AD 100) in the Mongolian steppes. In order to do this, the first experiment supports hypotheses about cultural and socio-economic networks. The second experiment focuses on developing new hypothesis about cemetery access, such as allowing or limiting acceptance to the family necropolis.

The rest of the paper is organized as follows : Section.II presents the related work. Section.III presents the SHAP library used to achieve XAI. The IArch tool and its use is

[1] https://www.oxfordlearnersdictionaries.com/definition/english/archeology?q=archeology

2375-0197/23/$31.00 ©2023 IEEE
DOI 10.1109/ICTAI59109.2023.00012

presented in Section.IV. In section.V, we present a scenario to validate its use. Before concluding, Section.VI presents the discussion and the threats to validity.

II. Related work

In the field of archaeology, ML has been applied to a variety of data types, including numerical and/or categorical data, textual data, images and geospatial data. To analyze these data, ML approaches such as unsupervised and supervised learning, as well as neural networks, were applied to accomplish different goals. For example, in [1] [5], ML algorithms are applied to numerical data to classify patterns in distinct pottery types. When it comes to analyzing textual archaeological material, ML has been utilized to automatically translate ancient languages [14] and extract significant information from various archaeological texts [15] [16]. With regard to image processing in the archaeology filed, we distinguish different applications such as obtaining information on artistic practices and cultural influences along the Silk Road [17], the detection of ancient rock carvings in natural environments [18], the analysis of pottery images and their classification based on various features [9], the analysis of shell midden [19] and human remains [20]. Furthermore, ML has been frequently used to analyze geospatial data in order to find archaeological sites [21] [3] [22].

Some of this research has supplied the archaeological community with data analysis tools, most of the time guided by a specific goal (e.g. text reconstruction, pottery recognition). The table I describes these tools by presenting the main focus, the type of the data (e.g. image, numerical), the used AI models, and the used XAI techniques to provide explanations. In [9], the authors presented an open system developed for the automatic collecting and recognition of pottery fragments using neural network methods. The system allows gathering images of pottery from numerous sources, including archaeological databases, museums, and private collections. To obtain robust and accurate recognition performance, the authors employed deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). In [12] [10], two tools respectively, Pythia and Ithica, harness the power of artificial intelligence and deep learning to reconstruct and identify, from images, fragmented or damaged texts, thus providing valuable insights into historical documents that were previously unavailable or incomplete. In [13], the authors presented, Arch-I-Scan, a tool that combines imaging technology and machine learning algorithms to provide a viable solution for the efficient and accurate identification of pottery in archaeological contexts. Arch-I-Scan analyzes visual aspects of pottery such as shape, color, and surface patterns using deep neural networks (CNN, ResNet). In [14], the author presented the CUNAT software package for automated translation of ancient inscriptions, documents, and writings in multiple languages. The toolkit is built on the utilization of many technologies such as neural networks, AI, phone-based VR, machine learning, and computer vision.

There is no doubt that the tools provided are effective in fulfilling the objectives for which they were designed; yet, the following observations are made : i) image processing is the most widely employed type of data in archaeology, which was also confirmed by [8]. Thus, altering the type of input to, say, numerical/categorical data (e.g. pottery classification based on numerical data), does not reveal if the tools maintain their performance. ii) In most cases, the model's efficacy is measured in terms of time savings and scalability when compared to traditional techniques based on human effort performed by experts in the field. Models are black boxes employed to achieve a certain goal, but none of the existing tools provide explanations and interpretations of the predictions.

Based on the available literature and to the best of our knowledge, we believe the archaeological community is lacking tools based on explainable ML that do not require particular expertise to fully understand and evaluate predictions. In the rest of this paper we present the IArch tool that is based on explainable ML and dedicated to analyze numerical and categorical archaeological data to validate hypothesis and generate new ones.

III. Explainability in AI

Before introducing the IARch tool, this section presents the eXplainable AI (XAI) algorithm used in this work. The need for XAI has increased significantly in recent years. As AI technologies evolve and play an increasingly important role in a variety of domains, there is an increasing demand for transparency, interpretability and trust in AI systems. Various algorithms, such as SHAP [2], LIME [3](Local Interpretable Model-Agnostic Explanations), Anchor [4], and InterpretML [5], have been designed to meet these requirements. The selection of the appropriate tool depends on the requirements, the type of model and the desired level of interpretability. In the current work, SHAP library [11] is used to interpret and explain the predictions of the ML models. This library is based on the principle of Shapley values, which is derived from cooperative game theory and aims at assigning a contribution value to each player participating in a collaborative game. In ML, Shapley values are used to quantify the contribution of each feature in the prediction of a model in order to provide both global and individual explanations. Shapley values are formally computed by examining all possible permutations of features and evaluating the marginal contribution of each feature to the final prediction. The calculation of Shapley values can be formulated as follows:

Let N be a set of features and v a valuation function. The Shapley value ϕ_i for a feature i is computed as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \left[\frac{(|S|!(|N| - |S| - 1)!)}{|N|!} \right] \cdot (v(S \cup \{i\}) - v(S))$$

[2]https://shap.readthedocs.io/en/latest/
[3]https://github.com/marcotcr/lime
[4]https://github.com/marcotcr/anchor
[5]https://interpret.ml/docs/getting-started

TABLE I
ARCHAEOLOGICAL TOOLS BASED ON AI

Tool	Focus	Data Type	AI models	XAI method
ArchAIDE [9]	Collection and recognition of pottery artifacts	Images	-CNN -RNN	None
Pythia [12]	Restoration of ancient Greek epigraphy	Images	-CNN -RNN	None
Ithaca [10]	Restoring and attributing ancient texts	Images	-RNN	None
Arch-I-Scan [13]	Automated pottery identification	Images	-CNN -ResNet	None
CUNAT [14]	Automated translation of archaeological documents and texts	Images	-CNN	None

where $|S|$ is the size of the set S and v(S) is the model's prediction value when the features of S are included.

The SHAP library automates the computation of the Shapley values for ML models and provides various visualization choices and charts for interpreting and analyzing the predictions. This enables domain experts to obtain understandable explanations of ML model predictions.

The selection of the SHAP library was not arbitrary. On one hand, SHAP enables acquiring explanations at many levels of granularity. Indeed, it provides both global explanations that reveal the relative relevance of features in the overall model, and local explanations that explain individual predictions. On the other hand, SHAP is an agnostic approach that could be used to a variety of ML models, including decision trees, neural networks, and support vector machines. It is not limited to a certain model, as other tools may need to be adapted to explain different types of models.

IV. THE IARCH TOOL

This section begins with a description of the IArch tool (Section IV-A), followed by the workflow pipelines for the core services, including the validation of hypothesis (Section IV-B) and the generation of new ones (Section IV-C).

A. Overview

The IArch tool provides users with a dashboard that enables them to process their data, including data profiling, validating hypothesis using explainable classification and generating new ones using explainable clustering. The dashboard does not require any special coding skills. However, in order to design experiments and comprehend visualizations, we provide users with fundamental theoretical knowledge about data processing and machine learning as part of the IArch user handbook. For each of the offered functionalities a set of actions are required. One important step is to upload the data to be analyzed. IArch's current version allows for the analysis of numerical and categorical data. The data to be uploaded must consequently be in one of the formats allowed by the tool (e.g. csv, excel). Once the data is uploaded, a profiling phase is triggered in order to obtain information about its structure, quality and characteristics. This step gives users insights into their data by providing visuals and statistics such as data distribution charts per feature, missing values, and a correlation matrix. After examining the data, the user has the option of selecting one of the main services including classification and clustering.

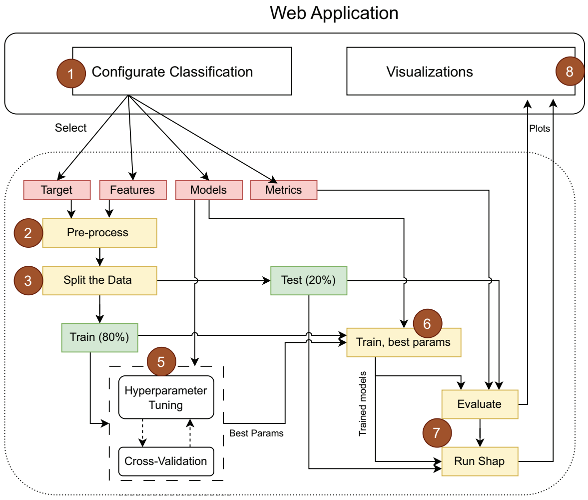

Fig. 1. Workflow Pipeline for classification

B. Explainable Classification to Validate Hypothesis

Classification is a fundamental process in ML that involves categorizing data into predetermined class labels. In classification, we typically start with a labeled dataset in which each data instance is assigned a class label. The goal is to build a ML model that can understand patterns and relationships in data and effectively predict class labels for new, unseen data. In archaeology, classification has mostly been used to classify images (such as pottery). This does not exclude it from being used for other kinds of data types such as numerical or categorical data.

In archaeology, classification helps to validate existing hypotheses, refine interpretations and discover new insights in the past. However, it is critical to approach classification in archaeology with domain expertise, because archaeologists must also be ML specialists. Thus, through the classification service, we assist them in validating their hypotheses, without requiring specific ML skills while ensuring robust and reliable results. The Fig.1 presents the workflow pipeline for applying classification to uploaded data. By using the web application's graphical interface, the user begins by configuring the classification experiment. A dataset must first be uploaded for this step to be successful. Configuration entails selecting the target variable, features, models and precision measures. The target and features are displayed according to the uploaded data. Models include tree-based models (e.g. random forest, XGboost, CatBoost...) and neural networks (e.g. artificial neural network...). Whereas the metrics include accuracy, F1 score, precision and recall. When the user finishes the configuration and pushes the start button, an automated classification process begins. The process is divided into various stages. The first stage is to pre-process the data corresponding to the selected features and target by encoding categorical variables and using scaling techniques. The data is then separated into two sets: training (80%) and test (20%). The training set is initially utilized in a cross-validation method, which is combined with hyper-parameters tuning, to establish the optimal parameters for the selected models and assess their stability on different data folds. In order to use the best parameters for our models, we used a technique known as GridSearch. It is a hyperparameter adjustment strategy that consists of defining a range of potential hyperparameter values and exhaustively searching for all possible combinations of these values to get the best configuration.

Once the best parameters are determined, each model with its corresponding parameters is trained on the full training set and then assessed on the test set by calculating the selected metrics. Furthermore, the SHAP lib (See Section III) is used to explain the results of the classification of the selected models. The SHAP visualisations include several kinds of diagrams, such as: i) the decision diagram, which shows how predictions vary in response to a set of features values. This method is useful for exposing model behavior; ii) the beeswarm plot that displays a dense information summary of how the main data features, based on their values, affect model results; iii) the waterfall diagram that displays explanations for individual predictions. It can be used to better understand misclassified samples. The different types of diagrams can help to determine the appropriate set of characteristics and eliminate those that are less important for validating the hypothesis in question.

C. *Explainable Clustering to Generate new Hypothesis*

In archaeology, as in many other fields, when data are not massive, their labeling is carried out manually by experts of the field. However, manual labeling of data can be subjective and introduce bias or omit essential information, especially when the data is heterogeneous, uncertain, and has multiple features. In this case, human subjectivity, inconsistencies, and restrictions can all have an effect on the quality and accuracy of data labeling. IArch's service for generating new hypotheses enables for the generation of explainable new data labeling by combining several approaches such as clustering, classification, and the usage of the SHAP library. Indeed, clustering is the process of grouping data points with similar proprieties together using a clustering model (e.g. k-means). However, clustering models, as an unsupervised learning approach, are

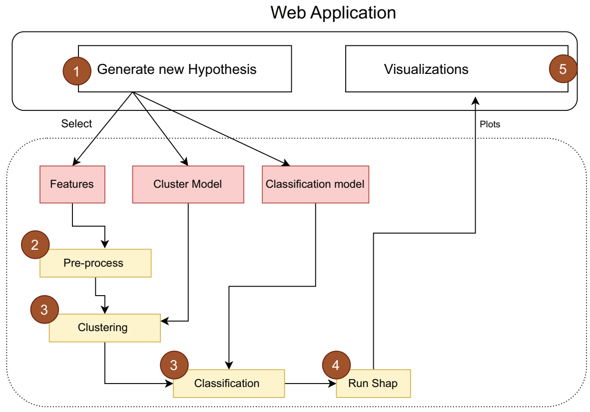

Fig. 2. Workflow Pipeline for generating new hypothesis

considered as a black box since they do not give clear explanations for the clusters they form. Applying a classification model to the clustered data and considering the clusters as the target variable is one technique to explain clustering findings. The classification results are then explained using the SHAP library. Thus, generating explainable labels may allow the archaeologist to better comprehend the data and consider novel hypotheses that may have been overlooked due to manual labeling. The Fig. 2 presents the workflow pipeline for generating and explaining new hypothesis. By using the web application's graphical interface, the user begins by configuring the experiment to generate new hypothesis. A dataset must first be uploaded for this step to be successful. Configuration entails selecting the features, a clustering model and a classification model. The features are displayed according to the uploaded data. Clustering models include K-means and hierarchical clustering. Whereas the classification models include tree-based models (e.g. random forest). When the user finishes the configuration and pushes the start button, an automated hypothesis generation process begins. The process is composed of different phases. The first stage is to preprocess the data corresponding to the selected features by encoding categorical variables and using scaling techniques. The selected classification model is then applied to generate new class labels. Once this step is complete, a classification phase using the selected classification model is triggered. The model is trained using the labels generated in the previous phase as the target variable. To explain the results, the SHAP library is utilized. Beeswarm plots are used to show the set of features that were critical and decisive in the establishment of each cluster. In order to ensure this, the beeswarm plots are given for each of the created clusters, highlighting the most relevant features and their contribution to the model' output based on their values (large or small).

V. Experiments

In this section we present how IArch tool can be used to visualize, analyze and get new insights (see Section V-B and

Section V-C) into data coming from a Mangolian graveyard (see Section V-A). The full implementation of IArch is shared on github [6].

A. Case Study

Funerary archaeology deals with necropolises or graveyards from the past. The case studied here is indeed a Mongolian graveyard from archaeological Xiongnu culture operating during the Iron Age (first century BC to first century AD). In this graveyard and this culture, just only few people are inhumed in a very specific way that is the signification of a certain elite. Indeed, the Xiongnu are nomadic people who travel with their herds and horses through the season and bury their ancestors in a single sacred place, often on mounds, under stone circles, buried several metres below the surface sometimes into two coffins with furniture and food offerings, and animal sacrifices. Fifty-four people are buried in the graveyard, but only 47 of them are studied with rigour. For each of the 47, the grave 'architecture (as depth, double coffin, etc.) and cultural item (as furniture, offerings, etc.) are precisely described defining specific patterns. As explained before, archaeology is also genetics data. Indeed, 18 buried people are linked to each other by genetic relationship on five generation genealogy which allowed a definition of a genealogical cultural pattern and a social cultural pattern dissociated from the first one.

Before starting data analysis, the user must upload the data to the IArch web application. Once uploaded, a statistical description of the data is provided. The user is alerted of missing values, as well as the number of categorical and numerical variables. Graphs depicting the distribution of values are presented for each feature. In addition, a heatmap is displayed. This first description allows to gain a first insight into the data.

B. Validate Hypothesis

Through a complete anthropological and archaeological analysis of the graveyard, the experts, with whom we conceived the tool, made hypotheses that are corroborated by simple and advanced statistical analytics. Indeed, thanks to the geographical position of the graveyard, next to the Chinese Empire of Western Han, and genetic profile of certain people, they supposed a strategy of trade and matrimonial alliances that have led to the genealogy of five generations studied here and conditioned the differences in wealth between people with a great wealth for the genealogy people. There are tow hypothesis to validate:

- H1: verify if we can predict genealogical affiliation (family membership) based on cultural features.
- H2: verify if we can predict the wealth index of the buried people by considering cultural features.

For H1, the target variable is the family, which has two different values: IF (In family) and OF (Outside the Family). Among the 47 subjects, 28 belong to class OF, 19 are labeled IF. Only features that are related to the cultural context are used to predict the family. These features include for example: grave volume, grave depth, wealth index, and Chinese imports.

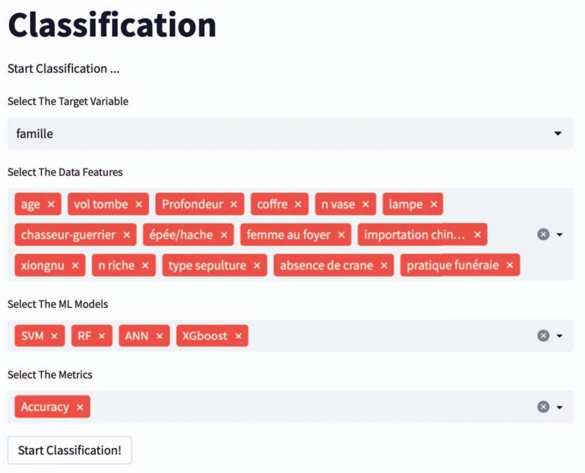

Fig. 3. Graphical interface to configure H1

The Fig. 3 presents the graphical interface corresponding to the configuration of H1. The user selects the target variable, the features, the models and the metrics. It is important to note that the default model selection is set to all existing models proposed by the tool, similarly for the metrics. If the user wishes to test with a limited set of models and/or metrics, he can remove them from the default selection. For this experiment, four algorithms were selected, namely SVM, RF, ANN and XGboost, and accuracy was chosen as the performance measure. Once classification begins, IArch determines the appropriate parameters for each model using cross-validation (5 folds) and hyperparameter tuning, then evaluates the models on the same set of test data against the selected performance measures.

TABLE II
ACCURACY RESULTS FOR H1 AND H2

	Hypothesis	RF	ANN	SVM	XGboost
Accuracy	H1	60%	54.54%	60%	70%
	H2	80%	80%	90%	100%

As shown in the Table II, for H1, XGboost provides the best accuracy of 70% with only three subjects miss-classified over 10 subjects in the test data.

IArch provides for each of the selected models, summary plots to explain the predictions.The Fig.4 shows, the features importance for the XGboost model. Indeed, the volume of the grave, the depth of the grave and the inhumation into two coffins are the three top features considered by XGboost to predict the genealogy. On one hand, high values (red dots) of the three preceding features contribute positively to the prediction of the IF class label. On the other hand, for the

[6]https://github.com/Chahrazed-Labba/IArch

fourth feature ”age over 30”, the lower the values (blue dots), the more the feature contributes to the prediction of the IF class. Whereas, certain characteristics such as gender and being a housewife are neutral and do not contribute to the result of the model. For the archaeologists who used this tool, XGboost presents an excellent selection of features to predict the family target variable.

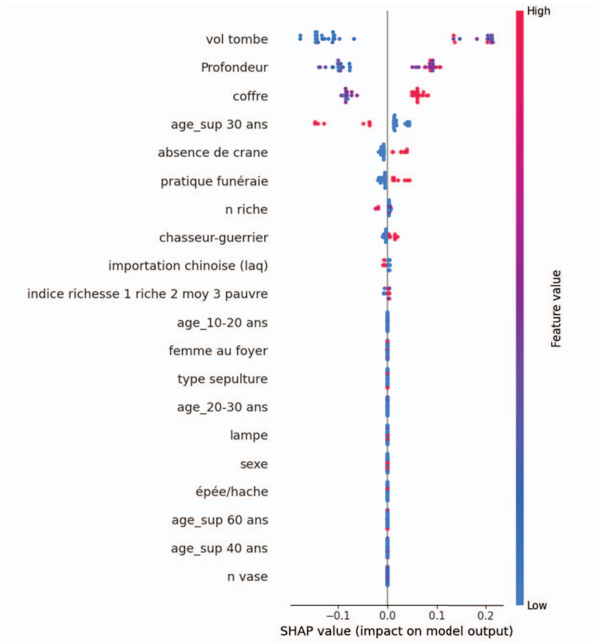

Fig. 4. Shap: Features importance for XGboost applied to verify H1

For H2, the target variable is the wealth index, which has three different values: poor, medium and rich. Among the 47 subjects, 27 are poor, 13 are rich and the rest are medium. We used the same models and precision metric (Accuracy), as H1, for the configuration of the experiment to verify H2.

As shown in Table II, XGboost provides a 100% of accuracy. The social differences as quality of cultural items or quantity of cultural items are so pronounced that models predicted the level of wealth without any error. For the archaeologists, the models are much better at predicting a social pattern like the level of wealth than a genetic link as genealogy. It is very interesting because it could lead to a new interpretation: Xiongnu people could be outside of an influent genealogy but could be as rich as them because of their social status for example or their ancestry (from Chinese origin for example). The Xiongnu seemed to consider kinship relationships when identifying each other and granting or denying access to the family necropolis, but when it came to the appearance of the deceased, they tended to favour a social consideration.

C. Generate Hypothesis

Experts in archaeology and anthropology are interested in the possibility of generating hypothesis with AI models. Indeed, the human mind is by nature influenced by his ideas and knowledge and directing his research based on them. For example, the family target variable is determined by the experts only based on the genetic data. The service for generating new hypothesis helps experts to explore new directions not biased by potential *a priori* and results could lead to a new anthropological hypothesis.

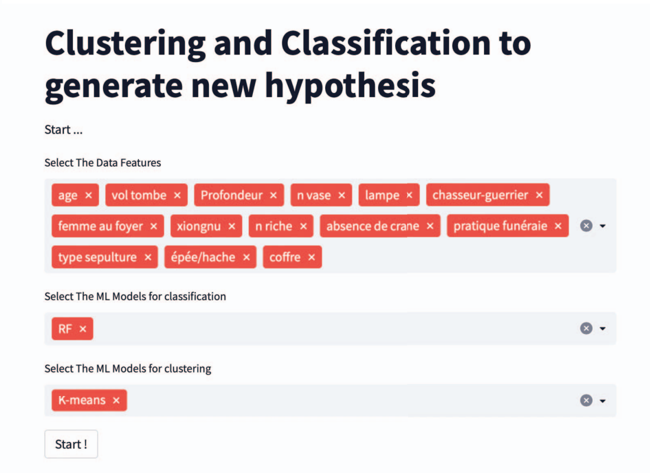

Fig. 5. Graphical interface to configure experiment for generating new hypothesis

The Fig. 5 presents the graphical interface corresponding to the configuration of generation of hypothesis. The user selects the features. Whereas, the default clustering and classification models are set to k-means and RF respectively. If the user wishes to change this, he can provide another configuration. For this experiment, both k-means and RF were used to respectively cluster and classify the data using different combinations of the features. The first experiment consists in using cultural data, GPS position and genetic data. K-means algorithm determines an optimal number of two clusters which dissociate men and women using the genetic data (men have two haplogroups for mitochondrial DNA and Y chromosome, while women have only one haplogroup for mitochondrial DNA). This classification is useless for the experts since no new hypothesis is generated.

As a second experiment, the genetic data is removed, the number of optimal cluster is then equal to five. To explain the K-means clustering findings, RF is trained on the clustered data using the cluster id as the target variable. Then, Shap library is used to explain the results of RF. For each cluster, the beeswarm diagram (Fig. 6) shows the decisive characteristics that were used to build the cluster, and the impact of their values on the model output (red dots for high values, blue dots for lower values). As shown in the Fig. 6, clusters are created

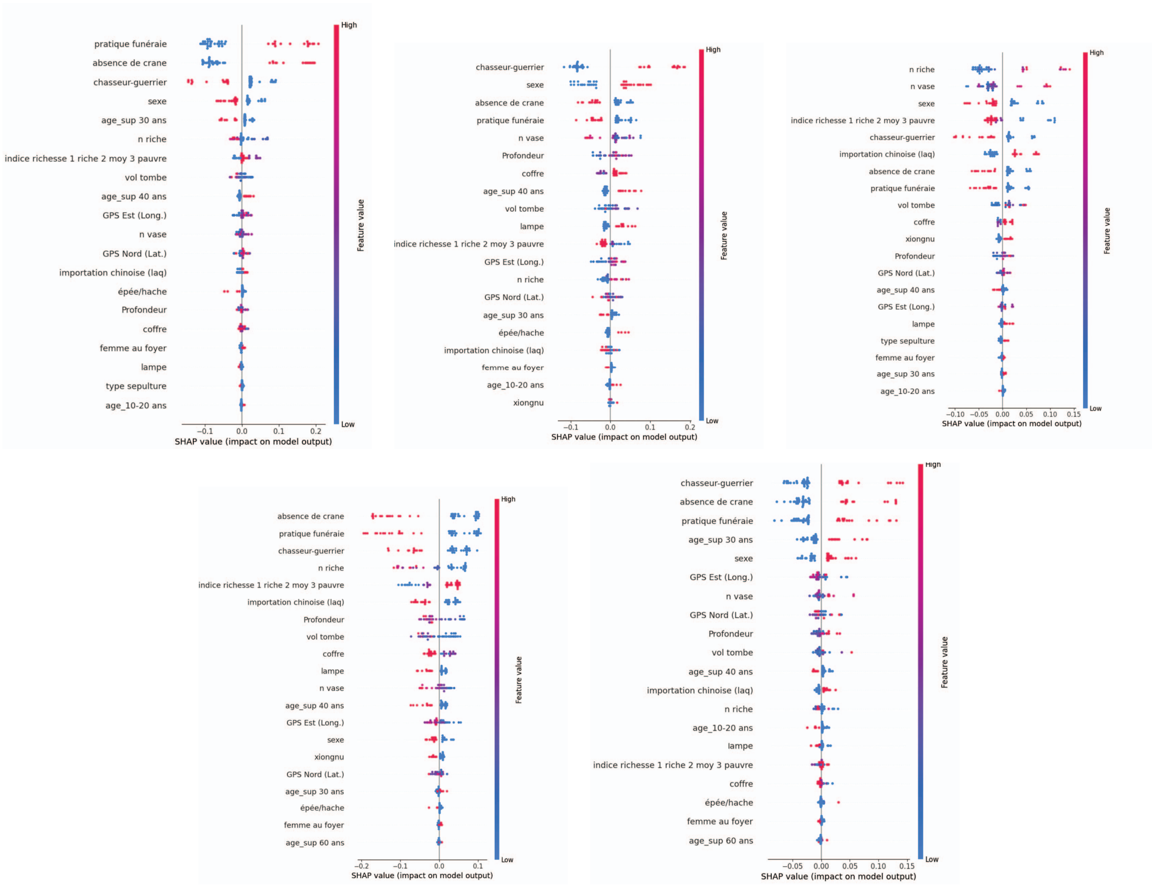

Fig. 6. SHAP Features Importance for each cluster

using different features that are very interesting to the experts. The first cluster groups mid and poor people whose skulls have been removed. The second cluster includes rich men up to 40 years old who are hunter warriors (with bows and arrows) and inhumed inside two coffins. The third cluster is constituted by rich women with furniture from Chinese import. Then, every other poor people (men and women) are grouped in the fourth cluster. And finally, the last cluster integrates hunter-warrior men up to 30 years old and whose skulls have been removed.

This classification showed that people from the genealogy are groups for the majority in the second and third clusters, the richest cluster. In addition, three people from the genealogy are grouped with the poorest people (fourth cluster). These three people are three women that should never be inhumed in this sacred place. Indeed, in the past, in majority of society patrilinear and patrilocal system was the rule, that means women have to leave their birth and growth place to follow their husbands. Among the three, one, from the first generation, had children inhumed in the necropolis but is poor despite the greatest economic time of the Xiongnu Empire. *Is she poor because she shouldn't be there?* The two others are mother and daughter and from the same lineage as the first one. They are inhumed in the necropolis during the collapse of the Empire, a collapse a great economic time. *Are they poor because of the end of the Empire or because they are coming from a 'bad' lineage?* Finally, two other couples from the 'bad' lineage are interesting: the first one is in a different cluster but have their skull removed. Furthermore, they are excluded from the genealogy position with an extreme north position in the graveyard. *Were they allowed to access to the sacred graveyard despite their 'bad' lineage, and was their skull, the greatest symbol of ancestrality, removed to another point more suited to their 'situation'?* The second couple is the opposite. Indeed, the man is from the 'bad' lineage but his wife is from Chinese ancestry and she is extremely rich. He's not as rich as she is. So, *was it the wife who was very rich and who led her husband to acquire this wealth by marriage and regain power and recognition?*

To conclude, the IArch tool through its service generating new hypothesis allowed experts to make new hypotheses to understand their archaeological site or try to understand it. The use of an explainable clustering allowed grouping together people according to different features and lead to time saving and generating new unexpected hypotheses.

VI. Threats to Validity

The current work presents some limitations that we tried to mitigate when possible: (i) The IArch tool was designed in collaboration with and tested by a limited number of archaeologists. We believe, however, that the tool should be disseminated to a larger number of domain experts; ii) In its current version, the IArch tool supports only categorical/numerical data. We intend to add more data types such as images; (iii) For the service to validate existing hypothesis and/or generated new ones, only some IA models are provided. We plan to extend the current version of the tool to provide a wide range of IA models (iv). In terms of explicability, only the Shap library is used to explain predictions. We intend to introduce new explanatory tools and provide the user with the option of selecting the tool of his choice.

VII. Conclusion and Future work

In this paper, we presented IArch, a tool that enables XAI data analytics for archaeologists without requiring specific programming skills. It allows data processing and analysis to either validate existing data-supported hypotheses or generate new ones. The tool covers the entire workflow for the application of ML, from data processing to the explanation of final results. The tool enables the use of supervised and unsupervised ML models, as well as the SHAP library, to provide archaeologists with explanations of the predictions. We demonstrate its use on data from a Xiongnu cemetery (100 BC/AD 100) in the Mongolian steppes. The archaeologists found that the IA models predict wealth better than genetics and family affiliation. Further, the use of the explainable clustering has lead to the generation of new unexpected hypotheses. As a future work, we plan to test the tool with a broader group of archaeologists . Further, we intend to enhance the tool with new features such as support for different data formats such as images, as well as more ML algorithms and explanatory tools.

References

[1] S. H. Bickler, "Machine learning identification and classification of historic ceramics," *Archaeology*, vol. 20, pp. 20–32, 2018.

[2] M. Grove and J. Blinkhorn, "Neural networks differentiate between middle and later stone age lithic assemblages in eastern africa," *PloS one*, vol. 15, no. 8, p. e0237528, 2020.

[3] G. Caspari and P. Crespo, "Convolutional neural networks for archaeological site detection–finding "princely" tombs," *Journal of Archaeological Science*, vol. 110, p. 104998, 2019.

[4] S. Klassen, J. Weed, and D. Evans, "Semi-supervised machine learning approaches for predicting the chronology of archaeological sites: A case study of temples from medieval angkor, cambodia," *PloS one*, vol. 13, no. 11, p. e0205649, 2018.

[5] A. Chetouani, S. Treuillet, M. Exbrayat, and S. Jesset, "Classification of engraved pottery sherds mixing deep-learning features by compact bilinear pooling," *Pattern Recognition Letters*, vol. 131, pp. 1–7, 2020.

[6] C. Cintas, M. Lucena, J. M. Fuertes, C. Delrieux, P. Navarro, R. González-José, and M. Molinos, "Automatic feature extraction and classification of iberian ceramics based on deep convolutional networks," *Journal of Cultural Heritage*, vol. 41, pp. 106–112, 2020.

[7] G. F. Cacciari, I.and Pocobelli, *Machine Learning: A Novel Tool for Archaeology*. Cham: Springer International Publishing, 2022, pp. 961–1002.

[8] S. H. Bickler, "Machine learning arrives in archaeology," *Advances in Archaeological Practice*, vol. 9, no. 2, p. 186–191, 2021.

[9] M. L. Gualandi, G. Gattiglia, and F. Anichini, "An open system for collection and automatic recognition of pottery through neural network algorithms," *Heritage*, vol. 4, no. 1, pp. 140–159, 2021.

[10] Y. Assael, T. Sommerschield, B. Shillingford, M. Bordbar, J. Pavlopoulos, M. Chatzipanagiotou, I. Androutsopoulos, J. Prag, and N. de Freitas, "Restoring and attributing ancient texts using deep neural networks," *Nature*, vol. 603, no. 7900, pp. 280–283, 2022.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.

[12] Y. Assael, T. Sommerschield, and J. Prag, "Restoring ancient text using deep learning: a case study on greek epigraphy," *arXiv preprint arXiv:1910.06262*, 2019.

[13] D. van Helden, E. Mirkes, I. Tyukin, and P. Allison, "The arch-i-scan project: Artificial intelligence and 3d simulation for developing new approaches to roman foodways," *Journal of Computer Applications in Archaeology*, Aug 2022.

[14] D. H. Sanders, "Neural networks, ai, phone-based vr, machine learning, computer vision and the cunat automated translation app–not your father's archaeological toolkit," in *2018 3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018)*. IEEE, 2018, pp. 1–5.

[15] A. Felicetti, "Teaching archaeology to machines: Extracting semantic knowledge from free text excavation reports," *ERCIM News*, vol. 111, pp. 9–10, 2017.

[16] A. Brandsen, S. Verberne, K. Lambers, M. Wansleeben, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck *et al.*, "Creating a dataset for named entity recognition in the archaeology domain," in *Conference Proceedings LREC 2020*. The European Language Resources Association, 2020, pp. 4573–4577.

[17] S. Kogou, G. Shahtahmassebi, A. Lucian, H. Liang, B. Shui, W. Zhang, B. Su, and S. van Schaik, "From remote sensing and machine learning to the history of the silk road: large scale material identification on wall paintings," *Scientific reports*, vol. 10, no. 1, p. 19312, 2020.

[18] G. Tsigkas, G. Sfikas, A. Pasialis, A. Vlachopoulos, and C. Nikou, "Markerless detection of ancient rock carvings in the wild: rock art in vathy, astypalaia," *Pattern Recognition Letters*, vol. 135, pp. 337–345, 2020.

[19] S. H. Bickler, "Prospects for machine learning for shell midden analysis," *Archaeology*, vol. 48, 2018.

[20] D. Huffer and S. Graham, "Fleshing out the bones: Studying the human remains trade with tensorflow and inception," *Journal of Computer Applications in Archaeology*, vol. 1, no. 1, 2018.

[21] A. Bonhage, M. Eltaher, T. Raab, M. Breuß, A. Raab, and A. Schneider, "A modified mask region-based convolutional neural network approach for the automated detection of archaeological sites on high-resolution light detection and ranging-derived digital elevation models in the north german lowland," *Archaeological Prospection*, vol. 28, no. 2, pp. 177–186, 2021.

[22] A. Guyot, M. Lennon, T. Lorho, and L. Hubert-Moy, "Combined detection and segmentation of archeological structures from lidar data using a deep learning approach," *Journal of Computer Applications in Archaeology*, vol. 4, no. 1, p. 1, 2021.