International Journal of Digital Humanities (2023) 5:299–331
https://doi.org/10.1007/s42803-023-00070-1

ORIGINAL PAPER

Explainability and transparency in the realm of digital humanities: toward a historian XAI

Hassan El-Hajj[1,2] · Oliver Eberle[2,3] · Anika Merklein[1] · Anna Siebold[1,4] · Noga Shlomi[1,5] · Jochen Büttner[1] · Julius Martinetz[1,2,3] · Klaus-Robert Müller[2,3,6,7] · Grégoire Montavon[2,3,8] · Matteo Valleriani[1,2,5,9]

Received: 28 February 2023 / Accepted: 3 September 2023 / Published online: 2 October 2023

Abstract

The recent advancements in the field of Artificial Intelligence (AI) translated to an increased adoption of AI technology in the humanities, which is often challenged by the limited amount of annotated data, as well as its heterogeneity. Despite the scarcity of data it has become common practice to design increasingly complex AI models, usually at the expense of human readability, explainability, and trust. This in turn has led to an increased need for tools to help humanities scholars better explain and validate their models as well as their hypotheses. In this paper, we discuss the importance of employing Explainable AI (XAI) methods within the humanities to gain insights into historical processes as well as ensure model reproducibility and a trustworthy scientific result. To drive our point, we present several representative case studies from the *Sphaera* project where we analyze a large, well-curated corpus of early modern textbooks using an AI model, and rely on the XAI explanatory outputs to generate historical insights concerning their visual content. More specifically, we show that XAI can be used as a partner when investigating debated subjects in the history of science, such as what strategies were used in the early modern period to showcase mathematical instruments and machines.

Keywords Explainable AI · Digital humanities · Sphaera · Early modern printing · Computational humanities

Hassan El-Hajj and Oliver Eberle contributed equally to this work.

✉ Hassan El-Hajj
hhajj@mpiwg-berlin.mpg.de

✉ Matteo Valleriani
valleriani@mpiwg-berlin.mpg.de

Extended author information available on the last page of the article

1 Introduction

Recent years have witnessed a great amount of research articles dedicated to the use of Artificial Intelligence (AI) in the humanities, focusing on Natural Language Processing (NLP) and Computer Vision (CV) approaches dealing with historical texts and artifacts.

Such approaches tackled numerous problems such as historical document layout analysis and information extraction, printed and handwritten text recognition, as well as text reconstruction and restoration in historical documents.

Document Layout Analysis (DLA) is an active field of research with numerous competitions (Gao et al., 2017; Simistira et al., 2017; Clausner et al., 2019; Yepes et al., 2021) regularly evaluating new approaches. DLA has relied heavily on AI methods, with transformers[1] assuming control of the field (Huang et al., 2022). In *historical* document layout analysis in particular, e.g., in Xu et al. (2018), the authors relied on a Multi-Task Fully Convolutional Network (FCN) to segment highly unstructured manuscript and printed-text pages into multiple semantically relevant groups (e.g., marginalia, main text, and comments), while Ravichandra et al. (2022) opts for an object-detection based approach relying on the YOLO model (Redmon et al., 2015). Others have recognized the value of extracting images from historical documents due to their importance in transmitting the information and ideas contained in the texts, leading to approaches such as the FCN networks presented in Monnier and Aubry (2020) and the object detection-based methodologies applied to specific corpora adopted by Dutta et al. (2021); Büttner et al. (2022) from techniques like YOLO (Redmon et al., 2015), U-Net (Ronneberger et al., 2015), or Faster R-CNN (Ren et al., 2016). By getting closer to the textual content of these documents, numerous AI-based approaches for optical character recognition (OCR) and handwritten text recognition (HTR) have been proposed, with deep learning-based approaches (Jaderberg et al., 2016) setting new standards. More advanced deep learning techniques rely on Recurrent Neural Networks (RNN) (Tsochatzidis et al., 2021; Fischer, 2020; Puigcerver, 2017) and Gated-CNNS (Kang et al., 2020; de Sousa Neto et al., 2020; Bluche & Messina, 2017); most recently, transformer-based architectures have set new benchmarks (Wick et al., 2021; Li et al., 2021; Ströbel et al., 2022). Beyond OCR and HTR tasks, AI approaches are emerging as a leading method in text restoration and reconstruction, which is vital when working with often fragmentary historical data. In Assael et al. (2019), the authors focused on Greek inscriptions and proposed a sequence-to-sequence RNN model which they called Pythia, and which was later followed by a transformer-based architecture (Assael et al., 2022) called Ithaka which is able to restore, date, and attribute Greek inscriptions. Continuing with the subject of ancient languages, Latin was addressed in Bamman and Burns (2020), who proposed Latin-BERT, a pre-trained BERT model (Devlin et al., 2018) on a large corpus of Latin texts aimed at text restoration tasks. While numerous approaches for OCR/HTR and text reconstruction tackle different languages (e.g., Akkadian (Lazar et al., 2021;

[1] Transformers are a type of machine learning model that employs self-attention mechanisms to solve a wide array of tasks.

Fetaya et al., 2020), hieroglyphs (Barucci et al., 2021), etc.), the analysis of historical texts is heavily biased towards Ancient Greek and Latin (Sommerschield et al., 2023).

While the above only scratches the surface of AI approaches in the humanities, it offers a comprehensive overview of the different research tracks in the field which appear to be adopting, and adapting, AI approaches. However, the use of Explainable AI (XAI) for insight generation remains in its infancy despite calls for a closer integration of DH and XAI approaches (Berry, 2020; Díaz-Rodríguez & Pisoni, 2020; Huggett, 2021). The term XAI refers to the field of AI research that is dedicated to the generation of explanations with regard to the increasingly complex machine learning models (Montavon et al., 2018; Samek et al., 2019, 2021; Holzinger et al., 2022) and is crucial in numerous domains to ensure model safety, robustness, and resilience to data drift. They may also reveal useful correlations in the data as well as ensure that the model results are understood by domain experts (Samek et al., 2021; Lapuschkin et al., 2019; Kamath & Liu, 2021). This field opened the door to numerous impactful contributions across an array of knowledge areas. Such contributions have left a mark for instance on medical imaging (Holzinger et al., 2019; Binder et al., 2021; van der Velden et al., 2022). In Zhang et al. (2019), the authors proposed an explainable model proposing human-like pathological diagnostics, while Hofmann et al. (2022); Müller and Hofmann (2023) used XAI methods to highlight the most relevant brain features that contribute to "brain age" which is considered a brain health biomarker. XAI methods are also heavily used in meteorological studies where their scope is to better understand the radar images (Ebert-Uphoff & Hilburn, 2020), as well as validate and interpret models and generate insight into specific phenomena such as tornadoes (McGovern et al., 2019). In chemistry, XAI is often applied on graph structures (Schütt et al., 2019; Schnake et al., 2022; Jiménez-Luna et al., 2020), as in Preuer et al. (2019) who applied XAI methods to highlight the specific substructures of molecules relevant for novel drug discovery. The above represent a small subset of the fields where XAI has had an impact on model validation and insight generation; for a more detailed review of XAI methods and applications see Samek et al. (2021); Samek (2023).

In contrast to the extensive work on XAI in the above-mentioned fields, XAI applications in the humanities remain limited to a few isolated cases. In Pawlowicz and Downum (2021), the authors trained a classifier to distinguish between multiple pottery types of the Tusayan White Ware in use around northern Arizona between AD 825–1300. The different types of pottery feature similar designs, which prompted the authors to rely on Grad-CAM[2] Selvaraju et al. (2020) to generate interpretable explanatory outputs to investigate which areas of these images had the highest saliency in assigning a pottery type to a particular artifact (Pawlowicz & Downum, 2021). In the domain of art history, Offert (2018) highlights the benefits of using feature visualizations generated by a trained machine learning algorithm for digital art historians. Through the use of these features, the assessment of an artwork by an art historian

[2] Grad-CAM computes explanations by weighting and pooling activation maps at a selected layer using the gradient of the prediction score backpropagated to this layer. This results in coarse localization maps that highlight the important image regions.

would combine both the original artwork and its model representations, enriching the available data for the interpretative process (Offert, 2018). Similarly, Bell and Offert (2021) showcases a workflow that integrates explanations–using Grad-CAM–to help domain experts identify different painting styles under difficult conditions (e.g., similar painting styles). In this case, by training a convolutional neural network to distinguish between the different styles of painting, the authors were able to hone in on a specific region that was relevant for the classification based on the explanatory heatmap[3]. Such region of the painting is where the hands are displayed (Bell & Offert, 2021). In a similar manner, XAI methods are now being used in art forgery forensics (Ji et al., 2021).

The above applications of XAI in the humanities rely mostly on region activation maps, often on the application of Grad-CAM which returns a relatively broad region heatmap describing the model decision. While such approaches open the door to general model explanations, they often miss the mark in providing **detailed** explanations that can be interpreted by a **domain expert**. We argue that a domain expert needs a fine-grained level of explanation to formulate and test hypotheses and show that a pixel-level relevance scores generated by Layer-wise Relevance Propagation[4] (LRP) (Montavon et al., 2019) could be used for this end.

At this stage, a more theoretical definition of *explanation* is warranted. In this paper, we refer to explanation in a teleological, *post-hoc* mode that has become common in the computational realm (Berry, 2021). A teleological mode of explanation aims to understand the neural network model's behavior and provide us with clues to interpret the features and correlations within the input data that in turn lead to a certain output. This means that our explanation is directly dependent on the goal that the system (in this case the neural network) is trying to achieve and indirectly on the data used to train the model.

The explanatory outputs of such an explanation system need to satisfy important criteria in order to serve as the foundation for a sound human-machine interaction, which are: "explanations should be faithful and sufficient" and "explanations should be humanly interpretable" (Samek et al., 2021).

The first criterion is satisfied when the explanation accurately describes the model behavior (Lopes et al., 2023), which can be assessed by removing highly relevant features (highlighted by the XAI model) from the input and observing whether this leads to high decay in the network predictions. A rapid decay of the network prediction score after removing highly relevant features indicates that the explanation is a faithful representation of the model's processing (Samek et al., 2017). Satisfying the second criterion is often challenging, as a concise and clear definition of human interpretability is difficult to achieve. Human interpretations are often contrastive, selective, influenced by social contexts, and commonly do not rely on probabilities (Miller, 2019). Additionally, these interpretations might vary based on the downstream task, depending on the social sample, domain knowledge, and the output type (Subramanian et al., 1992; Huysmans et al., 2011; Miller, 2019). We pay special attention

[3] A heatmap is an image highlighting the areas of the input that most strongly influence the model's decision

[4] A technique used to understand the contribution of each neuron in the network by propagating relevance backwards through a trained network

to the type of produced explanatory outputs as they are the basis of domain expert-formulated historical hypotheses. As such, explanations intended to be analyzed by domain experts need to be more detailed, and thus more complex, than those provided to laypersons who only might be comfortable with simpler explanatory output. The complexity of an interpretation can be quantified by the file size of the information carried by its heatmap. In this case, a smaller heatmap file (providing image region scale heatmaps) is likely to be easily interpretable by the layperson (Narayanan et al., 2018), while a complex explanatory output (providing pixel scale heatmaps) is likely more adequate for a domain expert. A comparison of different explanatory output complexity is presented in Samek et al. (2021), where LRP (Monnier & Aubry, 2020) returns faithful, and complex explanatory outputs on a pixel level, highly suitable for a domain expert.

To drive our point, we present a novel approach to historical image analysis that harnesses the explanatory outputs provided by our XAI model (see Section 3), and thus enables correlations to be revealed within a curated dataset of early modern printed illustrations (see Fig. 1). Such correlations enable domain experts, in this case historians, to better understand and analyze the content of the dataset at a pixel level. By looking at the insights generated by XAI method, we instrumentally choose a typical question in the frame of material history of science and technology of the early modern period, namely to provide a definition of early modern mathematical instruments. Such a research question, which is developed along the lines of three case studies, is instrumentally used to display the effectiveness of our approach. In this process, we treat the AI model as a helping hand or research companion–in line with similar approaches in medical AI (Klauschen et al., 2018; Ratti, 2022) and cyber

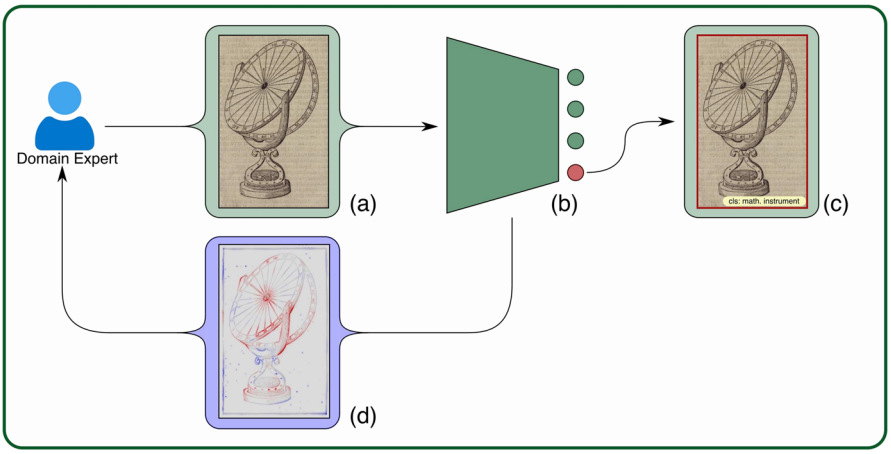

Fig. 1 Diagram representing a simplified overview of our proposed workflow for a historian XAI research companion (see the Appendices A and B for a technical rundown of the workflow). The workflow starts by the collection of curated and annotated data by the domain expert (a), which is then used to train a neural network (b), in this case, a VGG-16 network. Once trained, the model is able to provide an accurate prediction (c). We rely on this trained network to generate pixel level heatmaps (d) showing which pixels contributed to the classification prediction result in (b) and (c). This heatmap (d) is then read by the domain expert in order to generate, validate, and investigate historical hypotheses based on the curated data (a)

analysis (Holder & Wang, 2021)–that provides suggestions and reveals interesting data correlations that might have been overlooked by the domain expert.

This approach breaks with a general trend in the humanities to "simply" use AI to classify, and generally speaking, to automatically assign labels to humanities dataset elements, and fill a much needed research gap by elevating the human-machine interaction from one that is mainly operationally driven–such as the use of the machine as a supporting tool for domain experts–to one in which machine and human-expert interact via a common visual interpretation channel to produce interpretative historical results.

2 Dataset

The robust application of machine learning approaches typically requires a large amount of labeled data to ensure that our explanations are trustworthy, as well as generalizable. To meet this criterion, we rely on the *Sphaera* dataset, created in the frame of the project "The Sphere: Knowledge System Evolution and the Shared Identity of Europe" (https://sphaera.mpiwg-berlin.mpg.de), which contains data and metadata on over 350 early modern textbooks based on the *Tractatus de Sphaera* by Johannes de Sacrobosco (–1256). Electronic copies of these books are available via the project's database, comprising over 70,000 pages, 23,000 of which contain visual elements. These visual elements were collected both manually and with the help of neural networks (Büttner et al., 2022). The *Sphaera* dataset is stored in a large knowledge graph modeled according to the CIDOC-CRM standards (Bekiari et al., 2021), where information about the editions, as well as fine-grained information about their content is stored (Kräutli & Valleriani, 2018; El-Hajj et al., 2022).

For the purpose of this paper, and to accomplish our aim to train a neural network capable of correctly classifying pages containing illustrations displaying a mathematical instrument, we initially collected a total of 2,879 pages containing such illustrations, whose labels were carefully studied as part of a PhD dissertation within the Sphere project by Shlomi (2023). In addition to the pages containing illustrations of mathematical instruments, we collected 3,000 pages that do not contain any illustrations and which serve as a contrast signal to guide the model to learn class-specific instrument features. With such a binary dataset, our neural network, which is based on VGG-16 (Simonyan & Zisserman, 2015) and further described in Section 3.1, could successfully distinguish pages containing illustrations of a mathematical instrument from those with no illustrations. We applied the same method in Section 3 on this model, however, the generated explanations we received were not constructive in helping us understand the underlying features that represent a mathematical instrument because the model was simply learning the feature associated with the presence of an illustration vs. non-illustration. To gain further insights into what really defines a mathematical instrument, we created a richer dataset that encouraged our model to learn more discriminative features within these illustrations, which consequently led to more specific insights.

To accomplish this, we added two additional classes to our dataset. The first includes images of pages containing scientific illustrations that do not directly denote any

material object, such as an instrument, but rather have a descriptive or explanatory function in reference to the subject matter, in this case astronomical phenomena. These scientific illustrations, similarly to those representing mathematical instruments, were recovered from the *Sphaera* dataset. The second added class refers to illustrations of material objects, namely those which one can refer to as machines. This data was collected from Branca (1629); Zonca (1607); Ramelli (1588); Besson (1595) using Cor*Deep* (https://cordeep.mpiwg-berlin.mpg.de), a web service designed to extract and classify visual elements from historical documents (Büttner et al., 2022). In total, the dataset contains 5,879 pages distributed across the four classes, as shown in Table 1; each class is represented by a single sample in Fig. 2.

3 Methods

In this section, we introduce the methods used for data processing and model training. Further assuming a successfully trained model, we describe how the layer-wise relevance propagation (LRP) (Bach et al., 2015; Montavon et al., 2019) approach is used to extract explanations for each data point. A detailed explanation of the LRP rules is provided in Appendix A.

3.1 Neural network training

Given the comparably limited number of annotated training data available, as well as the large hetereogenity of the historical pages, we use the pretrained VGG-16 (Simonyan & Zisserman, 2015) convolutional neural network architecture to extract feature representations. This encoder consists of five convolutional blocks, each followed by a max-pooling layer with kernel size 2×2. Convolutional layers use 3×3 filter kernels sizes and ReLU activation functions. The resulting representations from the pretrained VGG-16 model are then used for the classification block, which consists of fully connected convolutional layers and a final softmax layer that predicts class probability for each of the four classes.

To address the hetereogenity of the input data, we standardize the pages using min-max normalization, apply thresholding using the 10% and 90% quantiles of the pixel value distribution and scale each image in proportion to a reference height or width of 800 pixels depending on its orientation using bilinear interpolation. These steps ensure that sufficiently high image quality is maintained while reducing variation in

Table 1 Distribution of image samples used across the four different classes

Class	Count
Mathematical Instruments	657
Scientific Illustrations	1870
Machines	352
Other	3000

Fig. 2 For each of the four classes - mathematical instruments, scientific illustrations, machines, and other - one example has been chosen to describe the general features of each respective class. Figure a) displays a typical early modern machine, in this case a structure that holds a gear-wheel, activated by a vertical gear-drum that turns the big wheel on the left. Figure b) is a scientific illustration that demonstrates the sphericity of the planet Earth by showing the reader its empirical proof, namely the fact that two observers on a ship–one at the top of the mast and one below on the gangway of the hull–would discover the castle, toward which they are navigating at different times, the one on the mast earlier than the other. Figure c) displays a typical page with no illustration that still contains other graphical layout features of an early modern textbook. Figure d) displays a common mathematical instrument, namely an armillary sphere which is a mechanical miniaturized reproduction of the geocentric cosmos. Figure a) from Branca (1629, p. 30), courtesy of the Library of the Max Planck Institute for the History of Science, Berlin. Figure b) from Sacrobosco (1547, Sign. B-2), Bayerische Staatsbibliothek, urn:nbn:de:bvb:12-bsb10173470-0. Figure c) from Piccolomini (1553, p. 17), courtesy of the Library of the Max Planck Institute for the History of Science, Berlin. Figure d) from Barozzi (1607, p. 104), Biblioteca Digital Hispánica, PID bdh0000001287

scan resolution, background texture, as well as colorization, brightness, and contrast (see Appendix B for details).

During optimization, we use 80% of the data for training the classification head parameters using the Adam optimizer (Kingma & Ba, 2015) and an initial learning rate of 0.001, which decays every seven epochs with gamma set to 0.1 and a batch size of 1 to allow for different page sizes and orientations. The resulting test set accuracy is 0.96 with class-wise F1 scores ranging from 0.89 for instruments to 1.0 for machines.

3.2 Layer-wise relevance propagation

We apply LRP (Bach et al., 2015) to attribute the predictions of our trained neural network to the input features (i.e., pixels). More specifically, by feeding a given input image to the network and denoting by y the resulting value of the output neuron for a given class, say instrument, LRP generates a collection of scores $R_1, R_2, \ldots, R_{\#\text{pixels}}$ identifying the contribution of each pixel to the output value y. This collection of scores can be represented as a heatmap in which the blue color indicates pixels that contribute *negatively* to the given class (i.e. are in contradiction with it), and red color

indicates pixels that contribute *positively* (i.e. support it). Because there are four classes in our dataset (other, mathematical instruments, machines, scientific illustrations), we generate four heatmaps for each image indicating pixels that support/contradict the respective class (see Fig. 4).

Technically, LRP pixel-wise scores are computed in an iterative fashion aiming to "invert" the nonlinear function implemented by the deep neural network. LRP starts at the output of the network with the predicted value y. The value y is then redistributed backwards in the network, layer after layer, by means of propagation rules (see Appendix A). LRP propagation rules are designed so that (1) neurons that are locally relevant (contribute strongly to the next layer) must receive more relevance than locally irrelevant neurons, and (2) quantities being redistributed must be conserved locally in the network, similar to water flowing through a network of pipes or a current traversing an electrical circuit. A variety of LRP rules implementing these two requirements have been proposed and they address different types of neurons and or levels of nonlinearity at each layer (Montavon et al., 2019). In practice, these rules are selected in a way that the resulting explanation faithfully reflects the true decision strategy of the neural network model and remains at the same time easily readable for the human-expert. In our experiments, we apply at each layer the same LRP rules as in Eberle et al. (2022).

4 Three case studies and the quest for early modern mathematical instruments

By applying the method from Section 3 on the dataset described in Section 2, we obtain a treasure trove of information about the image pixels with the highest contributions to the learning task. In the following section, the model explanations are discussed in reference to the predefined classes in Section 2 (excluding the category "other"), and therefore subdividing the argument into three case studies. Due to the complexity of the LRP explanations, interpretations are usually formulated by domain experts, who in our case are historians of early modern science.

4.1 The historical research question

On an abstract level, mathematical instruments of the early modern period are defined as measuring, computational, or demonstrating objects that embody mathematical knowledge. A historical overview can be found in Bennett (1987) and Bennett (2011). In the framework of the *Sphaera* corpus, mathematical instruments are those ordinarily used to measure time, as well as terrestrial or celestial angular distances. However, this abstract definition does not help us generate a concrete definition of a mathematical object that aims to answer how these objects were built and operated in the early modern period. The vagueness of the abstract definition becomes very evident when acknowledging that all of the illustrations in Fig. 3 show devices belonging to the same broad class of mathematical instruments.

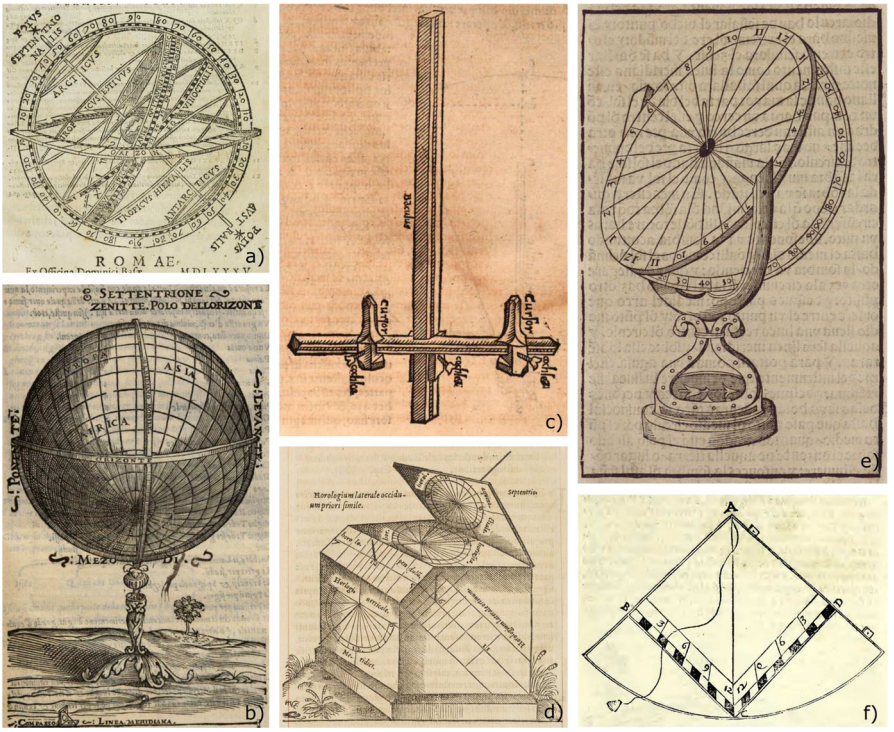

Fig. 3 Paradigmatic selection of early modern mathematical instruments. a) armillary sphere, b) globe, c) Jacob's Staff, d) universal sundial, e) universal meridian, f) quadrant. a) from Sacrobosco and Clavius (1585, Title page), courtesy of the Library of the Max Planck Institute for the History of Science, Berlin. b) from da Firenze (1537, Sign. H-II-7), Österreichische Nationalbibliothek, http://data.onb.ac.at/rep/10B4373E. c) from Schreckenfuchs et al. (1569, p. 285), Bayerische Staatsbibliothek, urn:nbn:de:bvb:12-bsb10141204-0. d) from Finé (1551, p. 18), courtesy of the Library of the Max Planck Institute for the History of Science, Berlin. e) from Cortés (1556, fol. XLVII v), Biblioteca Digital Hispánica, PID bdh0000254979. f) from Finé (1587, p. 32), courtesy of the Library of the Max Planck Institute for the History of Science, Berlin

Some of the instruments in Fig. 3, like the armillary sphere (Fig. 3a) and the globe (Fig. 3b), are mechanical reproductions of objects or layouts of objects that existed (or were believed to exist) in reality. The armillary sphere represents the geocentric cosmos, while the globe represents the Earth with additional features. In the first case, the armillary sphere includes a series of scales on its movable elements, thus allowing its user to determine the length of the solar day at any latitude throughout the year. Armillary spheres can also be more complex, for instance, when a planetary model is built-in so as to allow the determination of the position of all main celestial bodies at any given day and time. Armillary spheres are multipurpose mathematical instruments and, in addition, they also have a high pedagogical value as they intend to mechanically represent and visualize the constitution of the cosmos. The globe has very similar features. It is usually equipped with scales (though this example does not show them); it has movable parts; and, in this case, it clearly displays a geographic coordinate system, which at that time was considered the projection of the cosmos

coordinate system onto the planet (cosmography). This last feature finally enabled the calculation of terrestrial distances. Adding the distribution of landmass and water surfaces allowed the globe to become a representation of Earth, and because of this, this instrument also had high pedagogical value. Figure 3b shows a relatively poor example of such representation, as the continents are simply shown by placing their names at the appropriate locations.

The Jacob's staff is a distinct instrument (Fig. 3c). It does not represent anything found in nature, but is rather a purely functional device. Its function is to measure terrestrial distances from the point at which the observer (instrument's user) is positioned. It works on the basis of simple triangulation techniques where the two pivots at its base are movable in order to change the length of the triangle's base. The user would look through the center of that base and direct the staff toward the point whose distance from the observer is then measured. No scales and numerical values are added to this instrument. The length of the base between the movable pivots could be measured after observation, for instance by means of a ruler. It is hard to state what the qualitative difference is between the Jacob's staff on one hand and the globe and the armillary sphere on the other, but both groups of objects are considered mathematical instruments.

When it comes to time measurement, the universal sundial (Fig. 3d) and the universal meridian (Fig. 3e), which is also a sundial, are considered to be purely time measurement instruments. The diffusion of the mechanical clock and the corresponding division of the day into 24 equal hours began in the thirteenth century. However, the traditional division of the day according to the variable length of solar day and night, and therefore to the variable length of hours during the year as depending on the latitude, was still highly relevant in the early modern period for multiple reasons. The increased mobility during the early modern period increased the demand for sundials that could be operated at different latitudes. While the second case (Fig. 3e) is visually closer to our current understanding of a clock since the values are displayed in circular form, the first case (Fig. 3d) has a different purpose. In fact, this illustration does not really represent an instrument, but rather a display cabinet for different types of sundials.

Finally, the quadrant shown in Fig. 3f is common, especially within the *Sphaera* corpus, due to its function in the frame of astronomic observations. This type of instrument was mostly used to measure the angular height of celestial bodies over the horizon by placing it close to one's eye and pointing it towards the celestial body. The plumb line would then show the angle in degrees or hours and therefore the angular distance between the celestial body and the horizon. It shows scales with values and therefore is a kind of mathematical instrument that, qualitatively, belongs to the group of measuring instruments to which also the universal meridian belongs, and to a lesser extent, the globe and the armillary sphere (Fig. 3e).

It is clear from these short descriptions that a global and unique definition of mathematical instruments able to describe the different functions that such instruments perform and embody is difficult to establish. This situation in turn highlights the need to take a step back and think of an overarching definition, if only within the field of early modern astronomy.

In order to reach this new definition of mathematical instrument, we propose a novel approach that combines a) the insights revealed by using XAI methods applied to an AI model trained with these illustrations and b) the knowledge of expert historians.

4.2 Case study 1: mathematical instruments

Looking at the correctly classified instances of mathematical instruments (95%), we are immediately drawn by the LRP results to the finely graduated elements (scales and their values) of the mathematical instruments whose pixels appear to be most relevant for the correct classification. These finely graduated elements appear to be key elements in guiding our model to identify this class: they dominate the majority of the classified image results and are invariant to the numerous shapes, designs, and representations of these instruments within the pages of the different editions of textbooks. Examples of this finding are numerous. The quadrants displayed above (Fig. 3f) and the universal meridian (Fig. 3e), for instance, show this feature very clearly (Figs. 4 and 5).

In the cases where our model fails to recognize a mathematical instrument or falsely classifies something else as a mathematical instrument, the LRP result reveals insights into the causes of this misclassification. For example, the false positive in Fig. 6 shows the classification of the image as a mathematical instrument, while in fact this is a scientific illustration. This particular scientific illustration is designed to explain that the earth has a round shape, which is shown by the fact that sunrise is earlier in eastern locations on earth than in western locations. While this scientific illustration reproduces natural phenomena, it is also enriched by a densely graduated ring along the orbit of the sun in order to show the hourly divisions. In this case, it is clear that the model focused on the finely graduated orbital rings using this as the basis for its classification of this illustration as a mathematical instrument.

Considering the already discussed Jakob's staff in Fig. 3c, it becomes clear why this instrument is also misclassified as a machine. The heatmap shows that pixels representing mechanical elements (e.g. the pivots) are highly relevant for the classification results (Fig. 7) and, by comparison, it is possible to infer that this misclassification is due to the absence of detectable scales and/or numerical values, i.e. the omission of the most relevant pixels for the mathematical instrument class. As mentioned above, it is known that a graduated ruler is needed to measure the distance between the two pivots at the base of the instrument. Had this ruler been present in this illustration,

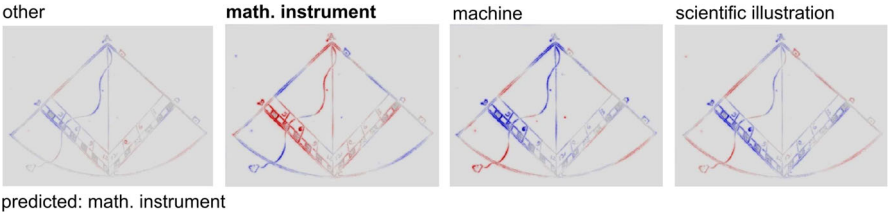

Fig. 4 Heatmaps of Fig. 3f detecting the relevant elements of a quadrant (red pixels are those that contribute positively to the class in question; blue pixels are those that contribute negatively to the class in question) and classifying it as a mathematical instrument

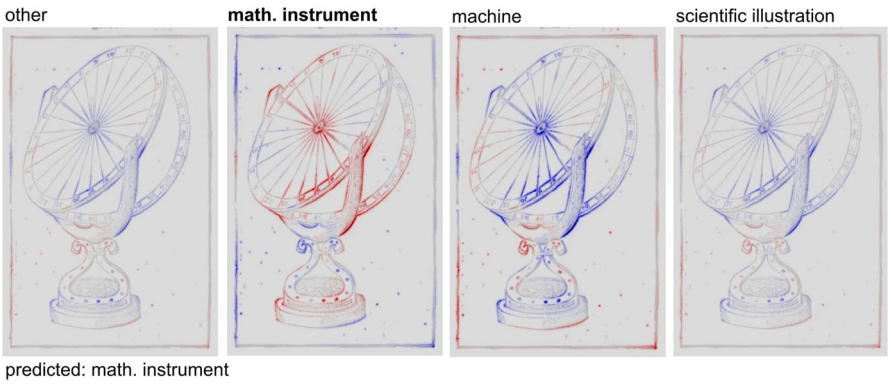

Fig. 5 Heatmap of Fig. 3e detecting the relevant elements of a universal meridian (red pixels are those that contribute positively to the class in question; blue pixels are those that contribute negatively to the class in question) and classifying it as a mathematical instrument

perhaps the classification result would have been correct (i.e., as part of the mathematical instrument class). The lack of graduated element has led to a false negative for this mathematical instrument. This result emphasizes that within the *Sphaera* corpus, the representation of an instrument, at least from the perspective of the model, is highly related to it having a finely graduated element.

In addition to the graduated scales, some elements of the object's materiality and morphology, such as the bases on which these instruments stand, played an important role in guiding the model toward a correct classification result as shown in Fig. 5. The displayed materiality is not a feature that concerns solely mathematical instruments and its meaning is best understood after considering the machine class in the following section.

4.3 Case study 2: machines

The Jacob's staff case demonstrated that the absence of the most relevant features, in this case, graduated scales, can quickly lead to a misclassification. This is one of

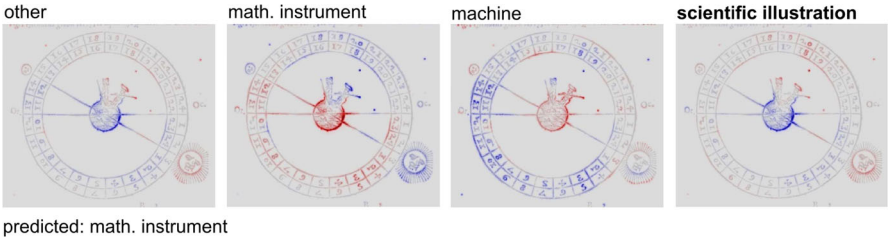

Fig. 6 Scientific illustration showing the dependence between the position of the observer on the spherical Earth and the times of sunset and sunrise. The illustration is misclassified as an instrument and the heatmap shows that this is due to the presence of a scale around the illustration which displays the hours (red pixels are those that contribute positively to the class in question; blue pixels are those that contribute negatively to the class in question)

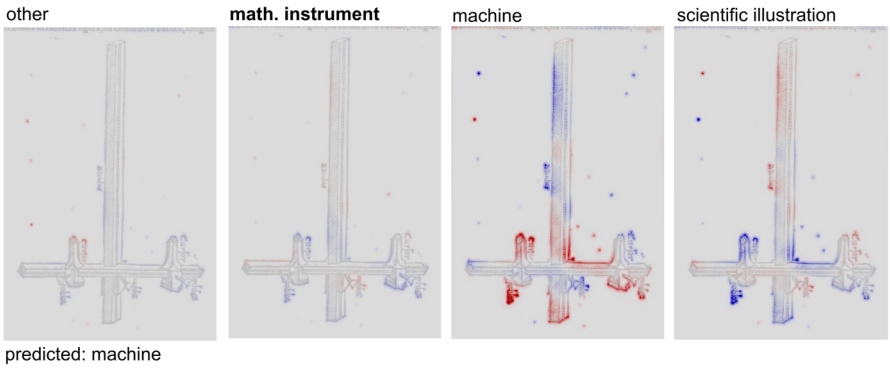

Fig. 7 Heatmap of Fig. 3c in which the Jacob's staff is misclassified as a machine. The heatmap highlights the mechanical feature of the instrument related to its movable parts and, by comparing it with other cases, it is possible to infer that the misclassification is also due to the lack of a scale with values (red pixels are those that contribute positively to the class in question; blue pixels are those that contribute negatively to the class in question)

the main reasons why the training set was enriched to include objects (machines) that are functionally and semantically distinct from mathematical instruments and clearly represent a distinct category.

The classic definition of an early modern machine is that of an object that enables the accomplishment of a task by means of the efficient performance of a mechanical device and thus lowering the need of human resources and/or reducing the time required for its performance. This definition is too general, and does not allow us to hone in on what constitutes a machine in our training set.

Our trained model was able to perfectly classify the machines in our dataset, and by looking at the LRP explanatory outputs, we were able to deduce that the most influential elements driving the correct classification of a machine-class within our dataset are the pixels representing aspects of the mechanical apparatus on the one hand, and the presence of a structured environment on the other. If a machine is analyzed in a representation that depicts it in its natural or semi-natural context, then the mechanical apparatus is activated. Figure 8a shows a machine to raise water from wells, operated by a man who interacts with a series of mechanical cranks and gear-wheels, which in turn drive the hydraulic apparatus with its dual pump. The LRP heatmaps (Figs. 8b and c) show that the pixels representing the underground section of this machine illustration are the most relevant for our model to generate a correct classification; more specifically, these pixels are those that represent the hydraulic apparatus.

Further investigation, however, shows that the proper mechanical apparatus (pulleys, gear wheels, handles, etc.) is less determinant than one would expect. For example, a machine for striking gold medals is examined in Fig. 9a. This machine is activated by a pneumatic device (labeled G and M on the right side of Fig. 9a) that channels hot air and fumes from a fire to activate a mechanical contrivance (K at the top right). This in turn moves three further mechanical elements, each comprising a gear-drum and a gear-wheel. This drivetrain finally activates a press made of two drums on whose surface the forms for the gold (in shape of medals) are engraved

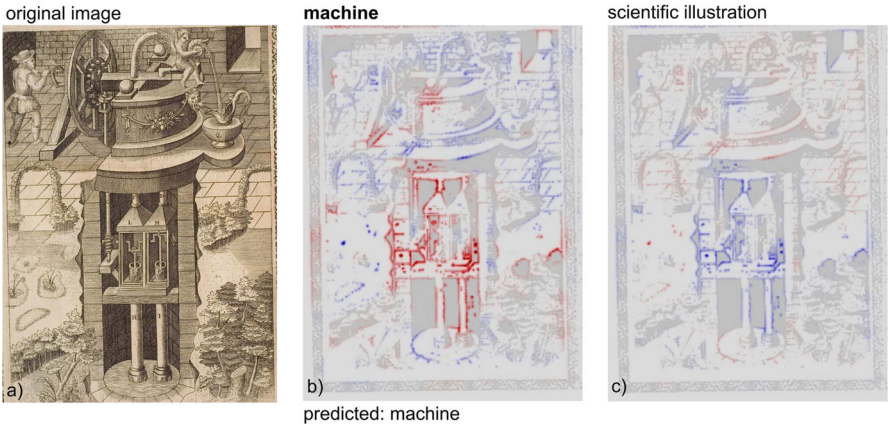

Fig. 8 a) Machine to raise water from wells. From Ramelli (1588, p. 9v), courtesy of the Library of the Max Planck Institute for the History of Science, Berlin; b) and c) heatmaps that highlight the hydraulic apparatus of the machine as the reason for its correct classification (red pixels are those that contribute positively to the class in question; blue pixels are those that contribute negatively to the class in question)

(A, E, D, and the operator V). This machine produces medals in series and is more efficient than the traditional method also represented here by the blacksmith (T at the bottom-center). Unlike Fig. 8, the most relevant pixels in Figs. 9b and c do not strictly refer to the mechanical elements that play a direct role in the transmission of forces, but rather to the scaffolding and poles that hold these mechanical elements together. This feature, which is often observed in the heatmaps, is difficult to interpret but a plausible reason is that most of the mechanical elements are circular, namely they have

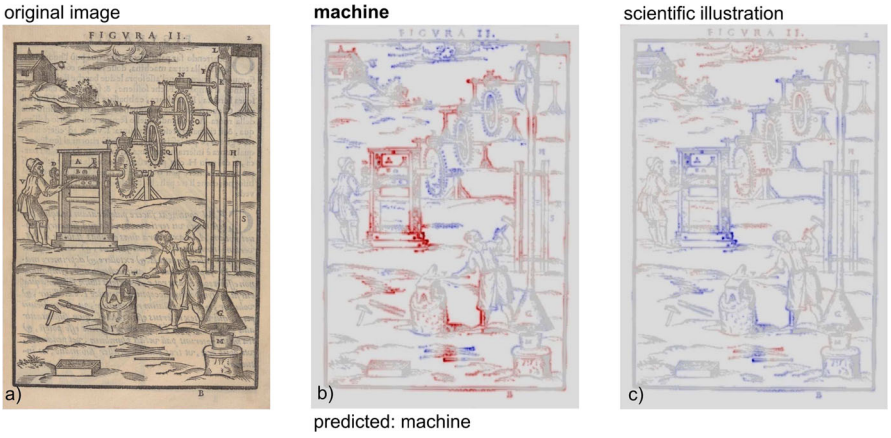

Fig. 9 a) Machine to strike gold medals. From Branca (1629, p. 2), courtesy of the Library of the Max Planck Institute for the History of Science, Berlin; b) and c) heatmaps of a) that highlight the scaffolding and the poles that hold the mechanical elements together, rather than the mechanical elements themselves (red pixels are those that contribute positively to the class in question; blue pixels are those that contribute negatively to the class in question)

a common characteristic with many further instances of other classes, be it instruments or scientific illustrations, as it will be discussed in detail in the next case study.

There are also misclassifications with respect to the early modern machine class such as the case presented in Fig. 10a. This is a lathe equipped with a semicircular blade which is rotated by means of a handle. Material (wood or stone) roughly spherical in shape is placed inside this machine, which then refines the object's shape (by cutting the excess) to produce an almost perfect spherical shape. The example of the lathe machine is inserted in the historical sources analyzed here because it was used to furnish an operational, almost material definition of what a sphere is: the fundamental geometric concept to understand the cosmos and the working of the *machina mundi*. For this reason, this machine is printed numerous times in the *Sphaera* corpus. The multiple instances of this machine appearing in our dataset were consistently classified as mathematical instruments. This misclassification is due to numerous reasons and can be explained as follows. While this machine shows strong "materiality" aspects, highlighted by the realistic footstands, it does not present any cogs, wheels, levers, or poles, which are typical of the machine class like in Fig. 9a. This leads the model to incorrectly classify this machine as a mathematical instrument. However, looking at Section 4.2, we can see that the main characteristic of mathematical instruments appears to be graduated elements, which this particular image lacks. We can conclude here that while graduated and mechanical elements played a major role in the classification of both mathematical instruments and machines respectively, the materiality of the represented object plays a secondary yet important role in defining these two abovementioned classes. As will be shown in Section 5, the issue at stake here is the concept of "materiality" as it needs further analytical qualifications.

4.4 Case study 3: scientific illustrations

The books used in this study feature many illustrations, but only a fraction of these represent a mathematical instrument or a machine; the other images vary between diagrammatic representations of mathematical concepts and schematic representations of the movements and geometrical constellations of celestial bodies, such as the orbit

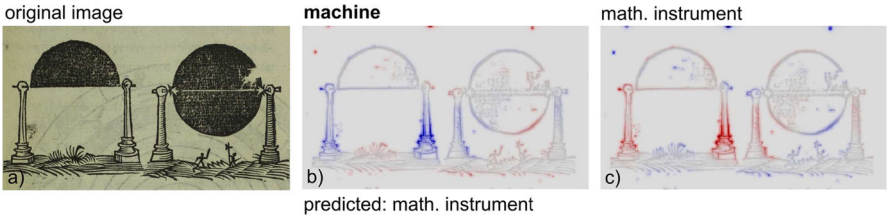

Fig. 10 a) Lathe machine. From Sacrobosco and Melanchthon (1545, Sign. B-ii-1), courtesy of the Libraries of the University of Oklahoma; b) and c) heatmaps of a) that misclassify the lathe as an instrument due to the presence of a footstand, which is also typical for instruments, and, upon further comparison, also because of the absence of an evident mechanical apparatus (red pixels are those that contribute positively to the class in question; blue pixels are those that contribute negatively to the class in question)

of the moon or the positions of the planets during a solar eclipse. Together, these images constitute the category referred to here as scientific illustrations. However, due to the heterogeneity of these illustrations, the LRP results rarely offer any insight into what constitutes a scientific illustration. By looking at the LRP results of scientific illustrations with respect to other classes, it is possible to gain insights into what does *not* constitute the illustration of a mathematical instrument or a machine.

In the example shown in Fig. 11a, the model correctly predicted that this is a scientific illustration. By looking at its prediction with respect to the machine class shown in Figs. 11b and c, it is possible to infer that the presence of multiple circles negatively contributes to the classification of a machine. This observation is validated by multiple other scientific illustrations and might be helpful in understanding why the often circular mechanical elements of the drivetrains of machines are not particularly relevant for the classification of machines and, above all, why graduated scales are relevant for the classification of instruments. The reason probably lies in the common feature of "circularity" of elements found in these images.

Finally, there are many misclassifications also among the scientific illustrations. A typical example is the illustration frequently used to demonstrate the sphericity of the Earth (Fig. 2b). This, as many other cases, shows that, when the illustration is particularly rich and moves toward artistic expression, the model misclassifies them as other pages, namely pages that are not supposed to contain visual elements (Fig. 12). The example shows that the lack of linearity in the drawing is associated with other features that are typical of a page with no visual elements, an issue for which there is currently no feasible explanation.

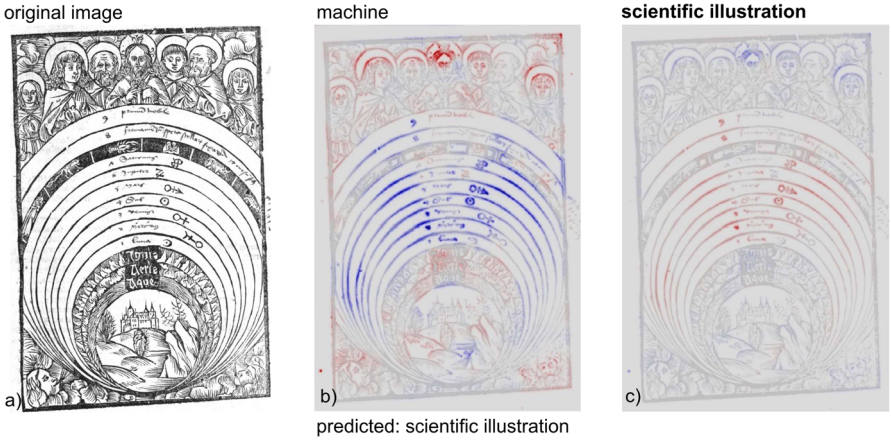

Fig. 11 a) Scientific illustration representing, from outside to inside, the empyrean, the spheres of the prime mobile, the firmament, the seven planets, the elements and the Earth at the center. From Sacrobosco and Glogów (1513, Sign. A-iiii-7), Regensburg, Staatliche Bibliothek, urn:nbn:de:bvb:12-bsb11110894-9; b) and c) heatmaps of a) that allow us to infer why circular shapes cannot be relevant for the correct classification of machines (red pixels are those that contribute positively to the class in question; blue pixels are those that contribute negatively to the class in question)

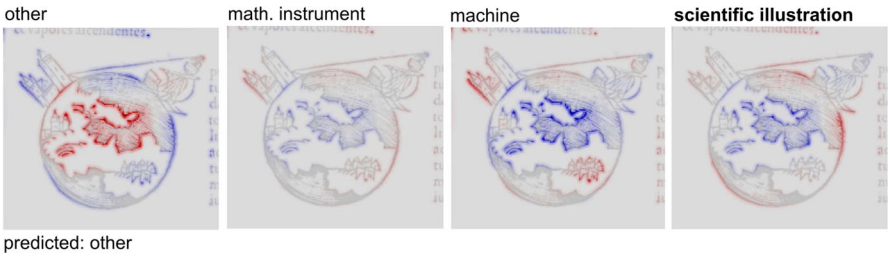

Fig. 12 Heatmap showing how particularly rich scientific illustrations, as in the case of Fig. 2b, are misclassified as other page (red pixels are those that contribute positively to the class in question; blue pixels are those that contribute negatively to the class in question)

5 Discussion

The take-home definitions so far regarding illustrations within the studied corpus read as follows: a mathematical instrument is an object with a graduated scale; a machine is an object with frames and scaffolding that hold together mechanical elements; scientific illustrations are characterized by circular elements. These XAI assisted definitions are heavily influenced by the subject of the collection of historical sources analyzed here, namely university textbooks used to teach geocentric astronomy.

A number of cases shown in Figs. 7, 10 and 12, demonstrate that the results achieved are not entirely satisfactory. To express this positively, our interaction with the explanatory output of the XAI model enabled us to identify a different interpretative layer or, more precisely, it revealed that an additional interpretative layer which was initially ignored, is more relevant than initially expected, as described in the following.

The goal of the presented historical research was to provide a definition of early modern mathematical instruments. These illustrations within our corpus often represented real objects; however, only a small fraction of these objects survives today. Museum collections contain some of these objects, but their number is limited, no complete dataset has been created, and, most importantly, even if such a dataset existed, it would not cover the myriad of instruments that populated the early modern period. Faced with this situation, historians of science and technology rely, as in the case of this work, on representations of such objects. These are abundant in the numerous preserved historical textual sources. The massive digitization efforts of the last twenty years, moreover, have made them largely and easily accessible.

Historians of science and technology interested in research questions similar to the ones proposed in this paper (see Section 4.1) often encounter the same issue of ambiguous definitions. What are the defining features of a specific class of images within a corpus? Are general class definitions a sufficient criteria to distinguish between class instances? What are the most important features that describe a class?

As shown in Section 4, the explanatory output of our XAI model can act as an intermediate layer, distorting our classical perception of the images in question, and guiding us toward specific pixel groups within images to help redefine or reshape our definitions of the selected classes. Of course, explanatory output ambiguity is present in this case too, as demonstrated in Fig. 10, highlighting the need for a continuous

human-machine interaction to reach the desired objective. In other words, the heatmaps show what can be considered as the machine's "definition." Thus, the heatmaps enable us to compare the human definitions with those of the machine, and to accordingly reconsider and potentially define them.

In fact, the definitions we–and other researchers–often try to seek are not those of the machines and instruments themselves, but those of their representations: the conceptual ideas in the minds of the actors of the early modern period who designed, drew, and cut the woodblocks used for their printing. In other terms, what we are investigating here is the result of a reality abstraction exercise by the authors, publishers, printers, and woodcutters, which eventually resulted in these preserved illustrations.

This process offered a degree of freedom but also a restraining condition that the material object itself cannot offer and does not provide, respectively. First of all, the represented object, such as a machine or an instrument did not have to exist. It could be the illustrated design of a new instrument, namely just a mental exercise. From this perspective, the analysis looks at what elements made a printed illustration a mathematical instrument illustration. Second, even in the cases where the represented object existed, there may no longer be a physical counterpart. In other words, in many cases, there is no extant artifact that enables us to compare the real physical instrument with its representation in the *Sphaera* dataset, meaning that abstractions are often not verifiable. This condition, combined with the high variability of instrument types and the consequent low number of images per specimen, makes working with this kind of datasets quite challenging. The degree of abstraction used in the illustrations of the studied corpus often required using a specific visual language, or visual conventions, to represent the objects in question. In this case, the use of the XAI explanatory output allowed us to better investigate these abstractions and grasp some of the visual conventions, which revealed interesting historical insights about the thought process of the early modern actors.

In this respect, the analysis highlights the most relevant image features as viewed by our trained model while considering not only the technical knowledge of the early modern actors in representing machines, mathematical instruments and scientific illustrations, as well as diagrammatic and decorative illustrations, but also how they imagined them. We refer to this feature as "visual conventions."

The first discoverable convention is related to **materiality**. As shown by the heatmap of the Universal meridian (Fig. 5c), the foot-stand of the instrument is an important element for this classification. This aspect is not only demonstrated by many similar heatmaps but also by the misclassification of the lathe as an instrument (Fig. 10). The feature of materiality in these cases conveys the message that these instruments either existed or *could have* existed.

The concept of "materiality" as used until now needs to be further distinguished into a closely related category, namely a second relevant visual convention, here called **environment**. This term refers to elements inserted in the image that show the surrounding context in which an instrument or machine could be found. Instruments such as the globe (Fig. 3b) or the cabinet (Fig. 3d) display a naturalistic environment. Heatmaps show an activation of the environment especially in reference to the representations of machines. These are often represented within a rich environment, which

often includes individuals operating these machines, as well as animals (e.g. oxen) and surrounding architecture.

A close look at the heatmap of the hydraulic machine (Fig. 8), for instance, shows that some tiles on the left play a relevant role. This feature is particularly evident in most of the machine representations, as shown also in the mill displayed in Fig. 13. The heatmap shows that, besides those machine, components that display the actual mechanical elements, stairs, windows, and, partially, even human beings act as determinants for its correct classification (Fig. 13).

The numerous cases displaying this feature show that the model is guided toward a machine class prediction based on the presence of regularly spaced lines, often denoting floor or roof tiles or blocks of stone in wall construction, among others. This aspect reveals important insights into the thought process of the historical actors involved, who clearly chose to represent machines, not only in a rich environment, but also in one where the reader of these books could estimate the size of the machine by easily comparing it to familiar objects such as tiles or stone blocks.

The final convention, which in this case is not highlighted by the activation displayed in the heatmaps, could be referred to as **proportion**. Many images, especially of the scientific illustrations class, and to some extent those of the instruments and machines, such as the globe (Fig. 3b) and the lathe (Fig. 10a), show a lack of proportionality among the represented elements. For instance, the lack of proportion between the globe and the tree at the bottom of the image or the lathe and soldiers below it. Such over-proportionality of elements is probably due to pedagogical intentions of the historical actors.

Misclassifications and the examination of their LRP heatmaps can shed further light on visual language and its evolution. For example, Fig. 6 shows a scientific illustration

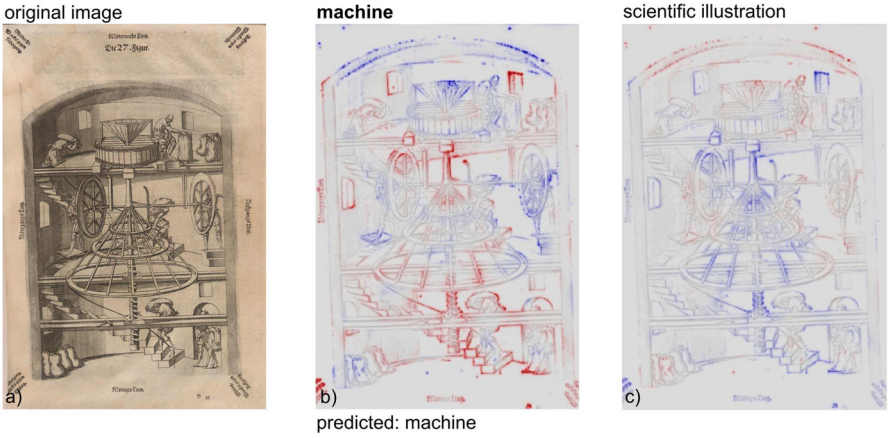

Fig. 13 a) Mechanical mill represented in its architectural context. From Besson (1595, Sign. H-iii-1), courtesy of the Library of the Max Planck Institute for the History of Science, Berlin; b) and c): heatmaps of a) showing that besides those machine components that display the actual mechanical elements, stairs, windows, and, partially, even human beings act as determinants for its classification as a machine (red pixels are those that contribute positively to the class in question; blue pixels are those that contribute negatively to the class in question)

that was misclassified as an instrument because it included a scale. The visual structure of this image, without the scale, is typical for explaining the sphericity of Earth with different sunrise and sunset times at multiple locations. The "addition" of a scale to the typical illustration shows a conscious employment of the visual convention of mathematical instruments in a scientific illustration. This addition makes the illustration look more like an instrument, and thus relates the represented content with the practice of exact measurements. It can be understood as a visual statement reflecting larger processes of mathematization of scientific and cosmological knowledge as well as changes in the practical and pedagogical orientation of the content (Oosterhoff, 2018).

5.1 Definitions, visual language, and training set

Taking into account the additional layer of visual language conventions of the historical actors encourages a reconsideration of the initial definitions. The examination of the LRP heatmaps showed that the full image was taken into consideration by the model and the results were based on different, distinct, visual aspects within the image. We were thus faced with the need to strongly consider the similarity between instruments and their contemporary illustrations. We understood that we could benefit from art-historical knowledge about visual language and in turn contribute to that field by revealing and defining visual conventions through large corpora.

The examination of the heatmaps points to elements that identify the different kinds of illustrations. The illustration of a mathematical instrument can be considered to be an object with a graduated scale that is represented by elements indicating its materiality and denoting that they exist or could exist. A machine illustration depicts an object with a structure that holds together mechanical elements and is represented within a realistic environment to convey additional information, such as its size or how it works. Unlike scientific illustrations, the representations of machines do not contain simple symmetric or concentric shapes, which implies that machines were portrayed as complex systems, without resorting to abstraction and simplification.

If we consider the results of this project to be the "discovery" of visual language conventions, and not the definition of what a mathematical instrument is, we can use LRP heatmaps to study both the evolution of visual language itself as well as the history of mathematical instruments. Visual conventions are structures or symbols that are widely used, and can thus be better "exposed" with the use of large corpora rather than a more traditional and manual examination of single sources. Although they do not necessarily represent objects that in fact existed, such conventions may reflect common knowledge used in printed books of the period and is thus meaningful in the study of the knowledge tradition.

Beyond the study of visual language for itself, the analysis of the XAI explanatory output for the study of mathematical instruments highlighted that the analysis of historical sources requires the consideration of all intrepretational layers or, in other terms, that both the form and substance of the images need to be considered *at the same time* for this kind of research. Considering the visual language can be a tool for better defining and studying mathematical instruments through the database.

The initial training set conceived for this investigation was not designed to cover all aspects the XAI analysis can highlight. Our research has shown that the introduction of XAI in the frame of historical research requires the generation of a transparent workflow dictating the principles to create training sets able to cover all interpretational layers and all their specific aspects and potentialities. In this sense, working with the explanatory output of an AI model can require multiple iterations, refining the data selection criteria and research questions at every step.

Attempts have been made to discuss the different aspects of "data modeling" in the field of computational history. This term can be used in a very broad and general sense to address ways in which historical information is organized in order to apply systematic, computational, or quantitative methods. Some of these attempts emphasize a semiotic approach that highlights the variety of types of similarities between objects (historical source or idea) and their representations (either representations of single objects or of relations between historical objects or ideas) (Kralemann & Lattmann, 2023; Ciula & Eide, 2016; Flanders & Jannidis, 2023). In the context of this research, we have seen that it is crucial to consider and strictly define whether the object falls under the definition of mathematical instruments or of scientific illustrations and the visual conventions they include. Furthermore, we had to also consider the nature of the relationship and similarity between the two. Thus, this research demonstrates the importance of a semiotic discussion in the process of building a training set, database, and research questions based on them. In other terms, the construction of the dataset itself should have included a discussion of the semiotic relation between the instrument and image and, therefore, an interpretation of the statistical results of such classification of images must include a consideration of visual language factors that inherently influence the activation of statistical tools.

6 Conclusion

We applied a XAI approach, more precisely an LRP approach to our classification model in order to explain how and why it classifies the illustrations of our corpus. We intentionally trained our model based on a specific, curated, dataset with carefully chosen classes in order to help us gain insights concerning our initial research question, namely what is an early modern mathematical instrument and, as our research shows, what differentiates it from an early modern machine and an early modern abstract scientific illustration. While we were able to achieve interpretable and useful results, we were also faced with unexpected outputs that obliged us to reevaluate our class definitions. As we are using illustrations of instruments and machines, XAI has shown us that we need to consider not only the proper content of those illustrations but also the visual conventions used by the historical actors to produce them. We have learned, through the interaction between the domain-experts' analysis and the explanatory model, which visual conventions were actually used. To put it emphatically, the XAI companion, our new team member, assumed the role of an art historian.

Many features of this research are characteristic of approaches that are common in the field of digital humanities, or more specifically digital history. The contribution is clearly the result of a group effort that relies on combining historical and machine

learning expertise. It provides a case study in an open and exploratory manner, describing the research undertaken in terms of a journey (including steps that did not yield the intended results), rather than presenting definitive findings. Moreover, the paper is informed by the idea of advancing a relatively new computational method and contributing to an evolving field. In terms of the historical research, it constitutes a proposal for a novel approach for analyzing historical sources, in particular scientific images.

Less characteristic, perhaps, is the fact that the focus is neither on the results of the applied computational method (images classified in a certain way), nor on a discussion of the applicability of the method, as is often the case in the digital humanities and digital history respectively. Instead, XAI–the method for analyzing the corpus' images–is used to create an interactive dialogue: between algorithmic classification and the historians involved. It is this dialogue that finally provides the basis for a nuanced discussion of the different types of scientific images contained in the corpus.

As shown, XAI makes transparent what specific features the ML model based its classification decisions on; the black box is thus, at least partially, opened. On the one hand, this means that the classification results are not taken at face value (in order to move on to the next step of the process). Rather, their very purpose is to be examined, reviewed, and questioned. Which part of the image led the model to choose the classification "instrument" and not "machine"? From the very beginning, the researchers' attitudes toward the model's decisions are therefore interrogative, perhaps even critical. On the other hand, the close observation made possible by applying XAI may provide further insights into the historical material. The larger question behind the endeavor is thus what historians can learn from engaging with the logic behind the model's decisions. In many cases, following and studying this logic will mean confronting a *different kind* of logic. In this particular case, historians using a XAI companion (who were trained in a certain way) were led to areas or elements of the images that they would not necessarily have considered relevant for classifying beforehand. The model's classification mechanisms thus become an element of discovery and–in combination with the historians' knowledge–a potential source for knowledge production. One could thus argue that XAI has the potential to alter the conventions of observation and in doing so may provide a new readability of the images, a readability that constitutes an interplay between machine logic and the historian's rationale.

However, XAI's potential for discovery is limited and remains, as pointed out earlier, teleological. It will only ever provide explanations for decisions it has been instructed to make. Data remain *capta*, taken not given, to use Johanna Drucker's definition, whether XAI is used to make its decisions transparent or not (Drucker, 2011). As always in data-driven research, the data must therefore be chosen carefully, and its selection and effect on the application of XAI must be taken into consideration when studying its decisions.

Nevertheless, this contribution shows that engaging with the interrogative approach engendered by XAI and taking its classification decisions as a productive "derivation of the eye," can produce a different way of seeing and thus provide a new perspective on the classification of historical sources. In fact, the presented approach is not limited to early modern illustrations but can provide relevant insights (depending on the quality of

the dataset used) in numerous humanities fields. One can envision that such an approach may help identify minute stylistic changes in artifacts such as statues or pottery, identify variations in motifs or ornaments transmitted over centuries, or precisely highlight artist styles as initially proposed by Bell and Offert (2021). Finally, the interrogative approach to machine-generated results may shift the overall perception of decisions made by AI, or, on a broader scale, change the way digital humanities scholars think about computational methods and infrastructures in general. This holds great potential for the discipline and would enable a criticality that is already visible in the many self-reflexive, theory-driven digital humanities approaches developed in recent years.

Appendix A: Explainable AI

The field of Explainable AI (XAI) aims to develop techniques that reveal what data patterns contribute the most to the prediction of a given machine learning model. This is necessary since most widely used modern ML models are typically composed of several, complex processing steps (or layers) that result in a highly nonlinear prediction. In contrast to linear models, for which the importance of a specific feature can be directly observed from the value of its activation score, for nonlinear methods specific methods to analyze the models processing are needed (Samek et al., 2019, 2021; Holzinger et al., 2022).

Most commonly used models, e.g. convolutional neural networks, graph neural networks and recurrent neural networks, are not interpretable by design and hence *post-hoc* explanations are used to compute explanations and gain insight into the inner model processing.

Layer-wise relevance propagation

To decompose the prediction of a typically complex deep neural network, the framework of Layer-wise relevance propagation (LRP) offers methods to compute explanations by highlighting relevant features in the data (Bach et al., 2015; Montavon et al., 2018, 2017).

Given some set of features x_p in input $\mathbf{x}$ (x_p in this work here denotes pixels), we aim to identify relevance scores R_p that reveal which features have contributed most positively ($R_p > 0$) or negatively ($R_p < 0$) to the model prediction $f(\mathbf{x})$. The resulting heatmap R thus provides a human-interpretable, intuitive way to better understand which features the model focuses on to make a specific prediction.
We denote the relevance from neurons k at layer $l+1$ as $R_k^{(l+1)}$.

Then, the lower level relevance at neuron j can be computed by summing over received messages $R_{j\leftarrow k}^{(l,l+1)}$ from neurons k in the higher layer $l+1$:

$$R_j^{(l)} = \sum_k R_{j\leftarrow k}^{(l,l+1)}. \tag{1}$$

The relevance message is generally proportional to the ratio defined by quantities q_{jk}, which is the contribution of neuron j to the activation of neuron k and the relevance observed $R_k^{(l+1)}$:

$$R_{j\leftarrow k}^{(l)} = \frac{q_{jk}}{\sum_j q_{jk}} \cdot R_k^{(l+1)}. \tag{2}$$

Finally, the full relevance of neuron j is computed by pooling over all incoming messages $R_j = \sum_k R_{j\leftarrow k}$.

Depending on the network type, layer index, and neuron type, different propagation rules have been proposed to compute q_{jk} (a summary of different rules can be found in Montavon et al. (2019)). Here, we focus on the LRP-γ rule that favors positive over negative contributions, and which is given by:

$$R_{j\leftarrow k}^{(l)} = \frac{a_j \cdot (w_{jk} + \gamma w_{jk}^+)}{\sum_j a_j \cdot (w_{jk} + \gamma w_{jk}^+)} \cdot R_k^{(l+1)}, \tag{3}$$

with lower-level neuron activation a_j, the weight w_{jk} between neuron j and k and parameter γ, which controls the preference of positive contributions using the rectified weight w_{jk}^+. This rule has been shown to offer a robust way to compute relevance redistribution by reducing gradient noise and generally more stable gradients (Montavon et al., 2019).

Appendix B: Experimental details

Data

For our case studies, we consider three classes of interest "mathematical instrument", "machine" and "scientific illustration". These are selected by domain experts and labeled accordingly. To facilitate this process, illustrations are extracted from full book pages using the automated image segmentation pipeline Cor*Deep* (Büttner et al., 2022). A web service to extract visual elements from various input types including PDF and common image file formats is accessible via https://cordeep.mpiwg-berlin.mpg.de/.

In addition, we include a contrast class ("other") that serves as an additional training signal to guide the model in focusing on class-specific features that describe our classes of interest instead of identifying spurious correlations, e.g., always predicting the class "mathematical instrument" with high confidence when there is a specific symbol in the top right corner that by chance occurs many times on our training samples for 'mathematical instrument' but which is overall not specific to the depicted object.

Data processing

To allow the ML model to focus on the extraction of task-related features, in a first step, we standardize the considered source material via binarization of the images. We normalize each image using min-max normalization and apply a percentile filter

at 0.8, utilizing the 10th and 90th percentiles of the pixel value distribution as cutoff values. This binarization process addresses various issues such as color heterogeneity, different page background textures, as well as contrast and brightness variations among the images.

We further use a reference value of 800 pixels to scale images in proportion to their original width and height using bilinear interpolation. The resulting data is finally separated into train/test splits (80/20%) for the subsequent model training and evaluation.

Model and optimization

We utilize a standard pretrained VGG-16 model that was originally trained to separate between 1,000 classes of natural images of objects (Simonyan & Zisserman, 2015). We replace the last linear layer with one that has as many output neurons as there are classes, which corresponds to four in our case. To predict these four distinct classes with high accuracy, we finetune the parameters of the final classification layer using the training dataset. During optimization, we minimize the cross-entropy loss between true and predicted labels using the Adam optimizer (Kingma & Ba, 2015) with a learning rate of $1e^{-3}$, stepwise learning rate decay every sevens epochs by a factor of 0.1 for a maximal number of 25 training epochs. We use a batch size of 1 to allow for different page sizes and orientations, and measure a test set accuracy of 0.96 with class-wise F1 scores ranging from 0.89 for 'mathematical instruments' to 1.0 for 'machines'.

To gain insight into which features the model uses to make its predictions, we will next compute LRP explanations.

Explanations

We use the LRP-γ propagation rule (Montavon et al., 2019), set $\gamma = 0$ for the classification layers and $\gamma = [0.5, 0.25, 0.1, 0.0]$ for layers 2-10, 11-17, 18-24 and 25-31, respectively. To handle the input image domain appropriately, we apply the $z^{\mathcal{B}}$-rule (Montavon et al., 2017) at the first layer. We perform the LRP explanation procedure for each of the four classes by explaining the predicted evidence of each class.

For visualization of the relevance heatmaps, we assign blue for negative relevance scores and red for positive scores. Color intensity is controlled by an opacity parameter α in the range between 0 and 1. We set α to be proportional to the absolute maximal relevance value of any of the four heatmaps. Further resources including demonstrations and tutorials regarding the implementation of LRP can be accessed via http://heatmapping.org/.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Oliver Eberle, Hassan El-Hajj, Grégoire Montavon, and Matteo Valleriani. The first draft of the manuscript was written by Hassan El-Hajj, Oliver Eberle, Grégoire Montavon, Matteo Valleriani, Anika Merklein, Anna Siebold, and Noga Shlomi, while Klaus-Robert Müller read and reviewed the draft and provided valuable comments. All authors commented on previous versions of the manuscript. All authors have read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the German Ministry for Education and Research as BIFOLD – Berlin Institute for the Foundations of

Learning and Data (grants 01IS18025A and 01IS18037A), the Rotenstreich fund, and by the Max Planck Institute for the History of Science. Furthermore Klaus-Robert Müller was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea Government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University and No. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation).

Data and Code Availability Code for the reproduction of our results can be found here: https://github.com/oeberle/xai_digital_humanities.

Declarations

Ethical Approval Not applicable

Informed Consent Not applicable.

Competing Interests The authors declare no competing interests.

References

Assael, Y., Sommerschield, T., & Prag, J. (2019). Restoring ancient text using deep learning: a case study on Greek epigraphy. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP), pp. 6368–6375. Association for Computational Linguistics, Hong Kong, China. https://doi.org/10.18653/v1/D19-1668. https://aclanthology.org/D19-1668

Assael, Y., Sommerschield, T., Schillingford, B., Bodbar, M., Pavlopoulos, J., Chatzipanangiotou, M., Androutsopulos, I., Prag, J., & de Freitas, N. (2022). Restoring and attributing ancient texts using deep neural networks. *Nature, 603*, 280–283. https://doi.org/10.1038/s41586-022-04448-z

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE, 10*(7), 1–46. https://doi.org/10.1371/journal.pone.0130140

Bamman, D., & Burns, P.J. (2020). Latin BERT: A contextual language model for classical philology. CoRR abs/2009.10053 arXiv:2009.10053

Barozzi, F. (1607). Cosmografia in Quattro Libri Divisa, la Quale Con Sommo Ordine, e Maravigliosa Facilitá, e Brevitá Introduce Alla Grande Mathematica Construttione di Tolomeo, & á Tutta l'Astrologia. Composta da Francesco Barozzi Gentil'huomo Venetiano. Con la Prefatione di Esso Autore, nella Quale si Ha Una Perfetta Divisione dell'Astrologia, & Una Narratione de Gli Autori Illustri, e De' Volumi da Loro in Tutte Le Parti di Essa Composti: & si Mostrano 84 Errori di Gio. de Sacrobosco, & Molt'altri De' Suoi Espositori, & Settatori, & Con Ragione si Riprendono. Preciede Ancho Alcuni Communi Mathematici, Arithmetici, & Geometrici Principij, Con Alcune Cose di Nuovo dall'Autore Ritrovate: & Alquante Propositioni, delle Quai per Tutta l'Opera si Fá Mentione: & Finalmente Un Indice Ricchissimo delle Cose in Essa Cosmografia Contenute. Grazioso Percacino, Venice. https://hdl.handle.net/21.11103/sphaera.100531

Barucci, A., Cucci, C., Franci, M., Loschiavo, M., & Argenti, F. (2021). A deep learning approach to ancient egyptian hieroglyphs classification. *IEEE Access, 9*, 123438–123447. https://doi.org/10.1109/ACCESS.2021.3110082

Bekiari, C., Bruseke, G., Doerr, M., Ore, C.-E., Stead, S., & Velios, A. (2021). Definition of the cidoc conceptual reference model v7.1.1. *The CIDOC Conceptual Reference Model Special Interest Group*. https://doi.org/10.26225/FDZH-X261

Bell, P., & Offert, F. (2021). Reflections on connoisseurship and computer vision. *Journal of Art Historiography,24*

Bennett, J. (2011). Early modern mathematical instruments. *Isis,102*(4), 697–705. Accessed 11 July 2023

Bennett, J. A. (1987). *The Divided Circle: A History of Instruments for Astronomy Navigation and Surveying*. Christie's collectors library: Phaidon Press, Michigan, USA.

Berry, D. (2020). The explainability turn: Critical digital humanities and explanation. In: L. Estill, J. Guiliano, & C. Crompton (Eds.), *DH2020 Book of Abstracts* (pp. 459–461). ADHO, Ottawa

Berry, D. (2021). Explanatory publics: Explainability and democratic thought. In: B. Balaskas & C. Rito (Eds.), *Fabricating Publics: The Dissemination of Culture in Post-truth Era* (pp. 211–232). Open Humanities Press, Bristol

Besson, J. (1595). Theatrum Oder Schawbuch Allerley Werckzeug und Rüstungen. Foillet, Mümbelgart

Binder, A., Bockmayr, M., Hägele, M., Wienert, S., Heim, D., Hellweg, K., Ishii, M., Stenzinger, A., Hocke, A., Denkert, C., Müller, K.-R., & Klauschen, F. (2021). Morphological and molecular breast cancer profiling through explainable machine learning. *Nature Machine Intelligence, 3*, 1–12. https://doi.org/10.1038/s42256-021-00303-4

Bluche, T., & Messina, R. (2017). Gated convolutional recurrent neural networks for multilingual handwriting recognition. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Vol. 01, pp. 646–651. https://doi.org/10.1109/ICDAR.2017.111

Branca, G. (1629). Le Machine : Volume Nuovo et di Molto Artificio da Fare Effeta Maravigliosi Tanto Spiritali Quanto di Animale Operatione Arichito di Bellissime Figure Con Le Dichiarationi a Ciascuna di Esse in Lingua Volgare et Latina. Mascardi, Roma

Büttner, J., Martinetz, J., El-Hajj, H., & Valleriani, M. (2022). Cordeep and the sacrobosco dataset: Detection of visual elements in historical documents. *Journal of Imaging,8*(285). https://doi.org/10.3390/jimaging8100285

Ciula, A., & Eide, O. (2016). Modelling in digital humanities: Signs in context *32*. https://doi.org/10.1093/llc/fqw045

Clausner, C., Antonacopoulos, A., & Pletschacher, S. (2019). Icdar2019 competition on recognition of documents with complex layouts - rdcl2019. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1521–1526. https://doi.org/10.1109/ICDAR.2019.00245

Cortés, M. (1556). Breve Compendio de la Sphera Y de la Arte de Navegar, Con Nueuos Instrumentos Y Reglas, Exemplificado Con Muy Subtiles Demonstraciones: Compuesto Por Martin Cortes Natural de Burjalaros en el Reyno de Aragon Y de Presente Vezino de la Ciudad de Cadiz: Dirigido al Invictissimo Monarcha Carlo Quinto Rey de las Hespanas Etc. Senor Nuestro. Antônio Alvares, Seville. https://hdl.handle.net/21.11103/sphaera.101394

da Firenze, M. (1537). Sphera Volgare Novamente Tradotta Con Molte Notande Additioni di Geometria, Cosmographia, Arte Navigatoria, et Stereometria, Proportioni, et Quantita Delli Elementi, Distanze, Grandeze, et Movimenti di Tutti Li Corpi Celesti, Cose Certamente Rade et Maravigliose. Autore M. Mauro Fiorentino Phonasco et Philopanareto. A Messer Giovan'Orthega Di Carion Burgense Hispano, et Dino Compagni Patritio Fiorentino, Mathematici. Bartolomeo Zanetti, Florence. https://hdl.handle.net/21.11103/sphaera.101009

de Sousa Neto, A.F., Bezerra, B.L.D., Toselli, A.H., & Lima, E.B. (2020). Htr-flor: A deep learning system for offline handwritten text recognition. In: 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 54–61. https://doi.org/10.1109/SIBGRAPI51738.2020.00016

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 arXiv:1810.04805

Díaz-Rodríguez, N., & Pisoni, G. (2020) Accessible cultural heritage through explainable artificial intelligence. In: Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization. UMAP '20 Adjunct, pp. 317–324. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3386392.3399276

Drucker, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly,5*(1)

Dutta, A., Bergel, G., & Zisserman, A. (2021). Visual analysis of chapbooks printed in scotland. In: The 6th International Workshop on Historical Document Imaging and Processing. HIP '21, pp. 67–72. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3476887.3476893

Eberle, O., Büttner, J., Kräutli, F., Müller, K.-R., Valleriani, M., & Montavon, G. (2022). Building and interpreting deep similarity models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(3), 1149–1161.

Ebert-Uphoff, I., & Hilburn, K. (2020). Evaluation, tuning, and interpretation of neural networks for working with images in meteorological applications. *Bulletin of the American Meteorological Society, 101*(12), 2149–2170. https://doi.org/10.1175/BAMS-D-20-0097.1

El-Hajj, H., Zamani, M., Büttner, J., Martinetz, J., Eberle, O., Shlomi, N., Siebold, A., Montavon, G., Müller, K.-R., Kantz, H., & Valleriani, M. (2022). An ever-expanding humanities knowledge graph: The sphaera corpus at the intersection of humanities, data management, and machine learning. *Datenbank-Spektrum: Zeitschrift für Datenbanktechnologien und Information Retrieval*. https://doi.org/10.1007/s13222-022-00414-1

Fetaya, E., Lifshitz, Y., Aaron, E., & Gordin, S. (2020). Restoration of fragmentary babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences, 117*(37), 22743–22751. https://doi.org/10.1073/pnas.2003794117

Finé, O. (1551). Sphaera Mundi, Sive Cosmographia Quinque Recèns Auctis & Emendatis Absoluta: in Qua Tum Prima Astronomiae Pars, Tum Geographie, Ac Hydrographie Rudimenta Pertractantur. Authore Orontio Finaeo Delphinate, Regio Mathematicarum Lutetiae Professore. Michel Vascosan, Paris. https://hdl.handle.net/21.11103/sphaera.101206

Finé, O. (1587). Opere di Orontio Fineo del Delfinato: Divise in Cinque Parti; Arimetica, Geometria, Cosmografia, et Orivoli, Tradotte da Cosimo Bartoli, Gentilhuomo, et Academico Fiorentino: Et Gli Specchi, Tradotti Dal Cavalier Ercole Bottrigaro, Gentilhuomo Bolognese. Nuovamente Poste in Luce. Francesco de Franceschi, Venice. https://hdl.handle.net/21.11103/sphaera.101202

Fischer, A. (2020). Automatic handwriting recognition in historical documents. In: A. Fischer, M. Liwicki, & R. Ingold (Eds.), *Handwritten Historical Document Analysis, Recognition, and Retrieval - State of the Art and Future Trends* (pp. 67–80). World Scientific, Friboug, Switzerland

Flanders, J., & Jannidis, F. (2023). Knowledge organization and data modeling in the humanities

Gao, L., Yi, X., Jiang, Z., Hao, L., & Tang, Z. (2017). Icdar2017 competition on page object detection. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Vol. 01, pp. 1417–1422. https://doi.org/10.1109/ICDAR.2017.231

Hofmann S., Beyer F., Lapuschkin S., Goltermann O., Loeffler M., Robert-Müller K., Villringer A., Samek W., & Witter A. (2022). Towards the interpretability of deep learning models for multi-modal neuroimaging: Finding structural changes of the ageing brain. *NeuroImage,261*, 119504. https://doi.org/10.1016/j.neuroimage.2022.119504

Holder, E., & Wang, N. (2021). Explainable artificial intelligence (xai) interactively working with humans as a junior cyber analyst. *Human-Intelligent Systems Integration, 3*, 139–153. https://doi.org/10.1007/s42454-020-00021-z

Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., & Samek, W. (2022). xxai - beyond explainable artificial intelligence. In: A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020,* July 18, 2020, Vienna, Austria, Revised and Extended Papers, pp. 3–10. Springer, Cham. https://doi.org/10.1007/978-3-031-04083-2_1

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery, 9*(4), 1312. https://doi.org/10.1002/widm.1312

Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022). Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM international conference on multimedia. MM '22, pp. 4083–4091. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3503161.3548112

Huggett, J. (2021). Algorithmic agency and autonomy in archaeological practice. *Open Archaeology, 7*(1), 417–434. https://doi.org/10.1515/opar-2020-0136

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems, 51*(1), 141–154. https://doi.org/10.1016/j.dss.2010.12.003

Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2016). Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision, 116*(1), 1–20. https://doi.org/10.1007/s11263-015-0823-z

Ji, F., McMaster, M. S., Schwab, S., Singh, G., Smith, L. N., Adhikari, S., O'Dwyer, M., Sayed, F., Ingrisano, A., Yoder, D., Bolman, E. S., Martin, I. T., Hinczewski, M., & Singer, K. D. (2021). Discerning the painter's hand: machine learning on surface topography. *Heritage Science, 9*(1), 152. https://doi.org/10.1186/s40494-021-00618-w

Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-020-00236-4

Kamath, U., & Liu, J. (2021). Introduction to interpretability and explainability. In: Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning, pp. 1–26. Springer, Cham. https://doi.org/10.1007/978-3-030-83356-5_1

Kang, L., Riba, P., Rusiñol, M., Fornés, A., & Villegas, M. (2020). Pay Attention to What You Read: Non-recurrent Handwritten Text-Line Recognition

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In: International Conference on Learning Representations (Poster)

Klauschen, F., Müller, K.-R., Binder, A., Bockmayr, M., Hägele, M., Seegerer, P., Wienert, S., Pruneri, G., de Maria, S., Badve, S., Michiels, S., Nielsen, T.O., Adams, S., Savas, P., Symmans, F., Willis, S., Gruosso, T., Park, M., Haibe-Kains, B., Gallas, B., Thompson, A.M., Cree, I., Sotiriou, C., Solinas, C., Preusser, M., Hewitt, S.M., Rimm, D., Viale, G., Loi, S., Loibl, S., Salgado, R., & Denkert, C. (2018). Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning. *Seminars in Cancer Biology,52*, 151–157. https://doi.org/10.1016/j.semcancer.2018.07.001. Immuno-oncological biomarkers

Kralemann, B., & Lattmann, C. (2023). Models as icons: Modeling models in the semiotic framework of peirce's theory of signs *190*(16), 3397–3420. https://doi.org/10.1007/s11229-012-0176-x. Accessed 15 Feb 2023

Kräutli, F., & Valleriani, M. (2018). CorpusTracer: A CIDOC database for tracing knowledge networks. *DSH, 33*(2), 336–346.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications, 10*, 1096.

Lazar, K., Saret, B., Yehudai, A., Horowitz, W., Wasserman, N., & Stanovsky, G. (2021). Filling the gaps in Ancient Akkadian texts: A masked language modelling approach. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 4682–4691. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. https://doi.org/10.18653/v1/2021.emnlp-main.384. https://aclanthology.org/2021.emnlp-main.384

Li, M., Lv, T., Cui, L., Lu, Y., Florêncio, D.A.F., Zhang, C., Li, Z., & Wei, F. (2021). Trocr: Transformer-based optical character recognition with pre-trained models. CoRR abs/2109.10282 arXiv:2109.10282

Lopes, P., Silve, E., Barga, C., Oliveira, T., & Rosado, L. (2023). Xai systems evaluation: A review of human nad computer-centered methods. *Applied Science,12*(19)

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society, 100*(11), 2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Monnier, T., & Aubry, M. (2020). docExtractor: An off-the-shelf historical document element extraction. In: ICFHR

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2019). Layer-wise relevance propagation: An overview. In: Explainable AI, pp. 193–209. Springer, Cham. https://doi.org/10.1007/978-3-030-28954-6_10

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition, 65*, 211–222. https://doi.org/10.1016/j.patcog.2016.11.008

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing, 73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Müller, K.-R., & Hofmann, S. M. (2023). Interpreting deep learning models for multi-modal neuroimaging. In: 2023 11th International Winter Conference on Brain-Computer Interface (BCI), pp. 1–4. https://doi.org/10.1109/BCI57258.2023.10078502

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2018). How do humans understand explanations from machine learning systems? *An evaluation of the human-interpretability of explanation*

Offert, F. (2018). Images of image machines. visual interpretability in computer vision for art. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops

Oosterhoff, R. J. (2018). *Making Mathematical Culture: University and Print in the Circle of Lefèvre D'Étaples*. Oxford-Warburg Studies: Oxford University Press, Oxford, UK.

Pawlowicz, L. M., & Downum, C. E. (2021). Applications of deep learning to decorated ceramic typology and classification: A case study using tusayan white ware from northeast arizona. *Journal of Archaeological Science, 130*, 105375. https://doi.org/10.1016/j.jas.2021.105375

Piccolomini, A. (1553). Editione Tertia. Della Sfera del Mondo di M. Alisandro Piccolomini, Divisa in Libri Quattro, i Quali Non per Via di Traduttione, Ne à Qual si Voglia Particolare Scrittore Obligati, Ma Parte da Migliori Raccogliendo, e Parte di Nuovo Producendo, Contengano in Se Tutto Quel Ch'intorno à Tal Materia si Possa Desiderare, Ridotti a Tanta Agevolezza, et à Cosi Facil Modo di Dimostrare, Che Qual si Voglia Poco Essercitato Ne Gli Studij di Mathematica PotràAgevolißimamente, et Con Prestezza Incenderne Il Tutto. Di Nuovo Ricorretta, et Ampliata. Delle Stelle Fisse. Libro Uno Con Le sue Figure et Con Le sue Tavole, Dove Con Maravigliosa Agevolezza Potrà Ciascuno Conoscere Qualunque Stella delle Quarantaotto Imagini del Cielo Stellato, et Le Favole Loro Integramente, et Sapere in Ogni Tempo Del'anno, à Qual si Voglia Hora di Notte, in Che Parte del Cielo si Truovino, Non Solo Le Dette Imagini, Ma Qualunque Stella di Quelle. Bartolomeo Cesano, Venice. https://hdl.handle.net/21.11103/sphaera.101047

Preuer, K., Klambauer, G., Rippmann, F., Hochreiter, S., & Unterthiner, T. (2019). Interpretable deep learning in drug discovery. In: W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 331–345). Springer, Cham. https://doi.org/10.1007/978-3-030-28954-6_18

Puigcerver, J. (2017). Are multidimensional recurrent layers really necessary for handwritten text recognition? In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Vol. 01, pp. 67–72. https://doi.org/10.1109/ICDAR.2017.20

Ramelli, A. (1588). Le Diverse et Artificiose Machine del Capitano Agostino Ramelli Dal Ponte Della Tresia Ingegniero del Christianissimo Re di Francia et di Pollonia: Nelle Quali si Contengono Uarij et Industriosi Mouimenti, Degni Digrandißima Speculatione, per Cauarne Beneficio Infinito in Ogni Sorte D'operatione. In casa del Autore, Parigi

Ratti, E. (2022). Integrating artificial intelligence in scientific practice: Explicable AI as an interface. *Philosophy and Technology,35*(3). https://doi.org/10.1007/s13347-022-00558-8

Ravichandra, S., Siva Sathya, S., & Lourdu Marie Sophie, S. (2022). Deep learning based document layout analysis on historical documents. In R. R. Rout, S. K. Ghosh, P. K. Jana, A. K. Tripathy, J. P. Sahoo, & K.-C. Li (Eds.), *Advances in Distributed Computing and Machine Learning* (pp. 271–281). Springer, Singapore

Redmon, J., Divvala, S.K., Girshick, R.B., & Farhadi, A. (2015). You only look once: Unified, real-time object detection. CoRR abs/1506.02640 arXiv:1506.02640

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation

Sacrobosco, J. d. (1547). Sphaera Iohannis de Sacrobosco. Martinus Nutius I for Jan Waen, Leuven. https://hdl.handle.net/21.11103/sphaera.100125

Sacrobosco, J. d., & Clavius, C. (1585). Christophori Clavii Bambergensis Ex Societate Iesu in Sphaeram Ioannis de Sacro Bosco Commentarius Nunc Tertio Ab Ipso Auctore Recognitus, & Plerisque in Locis Locupletatus. Permissu Superiorem. Domenico Basa, Rome. https://hdl.handle.net/21.11103/sphaera.101120

Sacrobosco, J. d., & Glogów, J. o. (1513). Introductorium Compendiosum in Tractatum Spere Materialis Magistri Joannis de Sacrobusto Quem Abbreviavit Ex Almagesti Sapientis Ptholomei Claudi Philosophi Alexandrini Ex Pheludio Progeniti per Magistrum Joannem Glogoviensem Foeliciter Recollectum. Florian Ungler for Jan Haller, Kraków. https://hdl.handle.net/21.11103/sphaera.100913

Sacrobosco, J. d., & Melanchthon, P. (1545). Sphaera Ioannis de Sacrobosco Typis Auctior, Quam Antehac, Atque Ex Diligenti Manu Scriptorum Impressorumque Codicum Collatione Castigatior, Praemissa Philippi Melanchthonis Doctiss. Praefatione, Qua Utilitatem Astrologicae Scientiae, & Christiano Homini Non Negligendam Scite Probat. Jean Loys for Guillaume Richard, Paris. https://hdl.handle.net/21.11103/sphaera.101054

Samek, W. (2023). Chapter 2 - explainable deep learning: concepts, methods, and new developments. In: J. Benois-Pineau, R. Bourqui, D. Petkovic, & G. Quénot (Eds.), *Explainable Deep Learning AI* (pp. 7–33). Academic Press, London. https://doi.org/10.1016/B978-0-32-396098-4.00008-9

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.) (2019). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning Lecture Notes in Computer Science, Vol. 11700. Springer, Cham

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K.-R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems, 28*(11), 2660–2673. https://doi.org/10.1109/TNNLS.2016.2599820

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE, 109*(3), 247–278.

Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schütt, K. T., Müller, K.-R., & Montavon, G. (2022). Higher-order explanations of graph neural networks via relevant walks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(11), 7581–7596. https://doi.org/10.1109/TPAMI.2021.3115452

Schreckenfuchs, E.O., Regiomontanus, J., Sacrobosco, J.d. (1569). Erasmi Osvvaldi Schreckenfuchsii Commentaria, in Sphaeram Ioannis de Sacrobusto, Accuratissima, Quibus Non Solumea Quae in Autoris Contextu Sunt, Sed Alia Etiam Ad Sphaericam Doctrinam Necessaria, Explicantur: Tabularum Atque Constructio, Ex Suis Principiis per Demonstrationum Seriem Clarem Dilucide Atque Docetur. His Adiecti Sunt Eiusdem Autoris Canones, Quibus Usus Tabularum, Quae Operi Ex Libro Directionum Ioannis Regiomontani, Passim Inseruntur, Ad Pulcherrimas Inquisitiones Astronomicas, Luculentissime Continetur. Reliqua Ad Consummatam Doctrinam Hanc Pertinentia, Ex Illum Primo Mobili, Eadem Forma Edito, Petes. Heinrich Petri, Basel. https://hdl.handle.net/21.11103/sphaera.101080

Schütt, K. T., Gastegger, M., Tkatchenko, A., & Müller, K.-R. (2019). Quantum-chemical insights from interpretable atomistic neural networks. In: W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 311–330). Springer, Cham. https://doi.org/10.1007/978-3-030-28954-6_17

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-cam: Visual explanations from deep networks via gradient-based localization *128*, 336–359. https://doi.org/10.1007/s11263-019-01228-7

Shlomi, N. (2023). The Evolution of Visual Language in Early Modern Astronomy. Tel-Aviv University, The Cohn Institute for the History and Philosophy of Science and Ideas (in progress)

Simistira, F., Bouillon, M., Seuret, M., Würsch, M., Alberti, M., Ingold, R., & Liwicki, M. (2017). Icdar2017 competition on layout analysis for challenging medieval manuscripts. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Vol. 01, pp. 1361–1370. https://doi.org/10.1109/ICDAR.2017.223

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In: ICLR

Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I., & Freitas, N.d. (2023). Machine Learning for Ancient Languages: A Survey. *Computational Linguistics,* 1–45. https://doi.org/10.1162/coli_a_00481

Ströbel, P., Clematide, S., Hodel, T., & Volk, M. (2022). Transformer-based htr for historical documents. In: Workshop on Computational Methods in the Humanities 2022

Subramanian, G. H., Nosek, J., Raghunathan, S. P., & Kanitkar, S. S. (1992). A comparison of the decision table and tree. *Commun. ACM, 35*(1), 89–94. https://doi.org/10.1145/129617.129621

Tsochatzidis, L., Symeonidis, S., Papazoglou, A., & Pratikakis, I. (2021). Htr for greek historical handwritten documents. *Journal of Imaging7*(12). https://doi.org/10.3390/jimaging7120260

van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022). Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis, 79*, 102470. https://doi.org/10.1016/j.media.2022.102470

Wick, C., Zöllner, J., & Grüning, T. (2021). Transformer for handwritten text recognition using bidirectional post-decoding. In: J. Lladós, D. Lopresti, & S. Uchida (Eds.), *Document Analysis and Recognition - ICDAR 2021* (pp. 112–126). Springer, Cham

Xu, Y., Yin, F., Zhang, Z., Liu, C.-L. (2018). Multi-task layout analysis for historical handwritten documents using fully convolutional networks. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pp. 1057–1063. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden. https://doi.org/10.24963/ijcai.2018/147

Yepes, A.J., Zhong, X., & Burdick, D. (2021). ICDAR 2021 Competition on Scientific Literature Parsing

Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M. M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., Ahmad, N., Khalil, F. K., Dickinson, S. I., Shi, X., Liu, F., Su, H., Cai, J., & Yang, L. (2019). Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence, 1*, 236–245.

Zonca, V. (1607). Novo Teatro di Machine et Edificii. Appresso Pietro Bertelli, Padova

Authors and Affiliations

Hassan El-Hajj[1,2] · Oliver Eberle[2,3] · Anika Merklein[1] · Anna Siebold[1,4] · Noga Shlomi[1,5] · Jochen Büttner[1] · Julius Martinetz[1,2,3] · Klaus-Robert Müller[2,3,6,7] · Grégoire Montavon[2,3,8] · Matteo Valleriani[1,2,5,9]

1 Max Planck Institute for the History of Science, Boltzmannstr. 22, 14195 Berlin, Germany

2 BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany

3 Machine Learning Group, Technische Universität Berlin, Marchstr. 23, 10587 Berlin, Germany

4 German Center for Art History Paris (DFK Paris), Rue des Petits Champs, 45, Paris 75001, France

5 The Cohn Institute for the History and Philosophy of Science and Ideas, Faculty of Humanities, Tel-Aviv University, Tel-Aviv 6997801, Israel

6 Department of Artificial Intelligence, Korea University, Seoul 136-713, South Korea

7 Max Planck Institute for Informatics, Stuhlsatzenhausweg 4, 66123 Saarbrüken, Germany

8 Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 7, 14195 Berlin, Germany

9 Institute of History and Philosophy of Science, Technology, and Literature, Faculty I, Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany