scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Hyperspectral dataset of historical documents and mock-ups from 400 to 1700 nm (HYPERDOC)

Ana Belén López-Baldomero[1]✉, Juan Luis Nieves[1], Francisco Moronta-Montero[1], Miguel Ángel Martínez-Domingo[1], Ramón Fernández-Gualda[1], Javier Hernández-Andrés[1], Anna Sofía Reichert[2], Ana López-Montes[2], Teresa Espejo[2], Javier Romero[1] & Eva María Valero[1]

HYPERDOC is a hyperspectral imaging dataset of historical documents and mock-ups, designed to facilitate research in material identification in the cultural heritage domain. It contains mock-ups of historical inks (metallo-gallate, sepia, carbon-based, and mixtures) on various supports, including some artificially aged, and historical documents from the 15th to 17th centuries (manuscripts, illuminated manuscripts, and family trees). Hyperspectral reflectance images were acquired using line-scan cameras in the VNIR (400-1000 nm) and SWIR (900-1700 nm) ranges and were spatially registered. Small regions of interest, referred to as 'minicubes', were extracted from the full document images, and pixel-level ground truth material annotations were performed. False-color RGB images and metadata were included in both the full document and minicube captures. The HYPERDOC dataset has been successfully applied in various experimental studies, including ink classification using machine learning models, spectral unmixing, colorimetric analysis, and binarization. These applications highlight the dataset's potential, which is publicly available to promote interdisciplinary collaboration and advance the use of hyperspectral imaging in the conservation field.

Background & Summary

Historical documents serve as invaluable repositories of cultural and scientific heritage, capturing knowledge, art and events of human history through manuscripts, archives, and different written or printed works that are worthy of study. To ensure their preservation and enhance accessibility and understanding, over 60 datasets of digital images of historical documents have been developed[1]. These serve as resources for diverse image processing applications, including layout analysis[2], binarization (i.e., separating ink and support into binary values)[3,4], content analysis[5,6], author identification[7], and improved readability of degraded documents[8].

Most existing datasets comprise digital images acquired using conventional RGB cameras, which capture information in only three channels: red, green, and blue. However, in the late 1970s and early 1980s, hyperspectral cameras were developed. They are able to capture hundreds or even thousands of spectral channels, from ultraviolet to short-wave infrared[9]. Unlike RGB imaging, hyperspectral imaging captures spectral radiance, which can be converted into spectral reflectance or transmittance for each pixel. This represents the ratio of reflected or transmitted radiation to the incident radiation. The spectral fingerprint enables material identification and mapping by combining spectral and spatial information. Spectral images are recorded in a relatively fast and non-invasive way, which has led this analytical technique to gain prominence in the field of cultural heritage in recent years[10–12].

In the context of document analysis, hyperspectral and multispectral imaging have demonstrated significant advantages over conventional methods. For instance, binarization using hyperspectral[13] or multispectral[14–21] data achieves improved separation of ink and support compared to RGB imaging. In forensic analysis, spectral data have enabled the detection of ink mismatches, aiding in the identification of document alterations or forgeries[22,23]. Hyperspectral imaging has also been used for material identification, such as inks[24,25] and pigments[26,27].

[1]Color Imaging Laboratory, Department of Optics, Faculty of Sciences, University of Granada, Granada, 18071, Spain. [2]Department of Painting, Faculty of Fine Art, University of Granada, Granada, 18071, Spain. ✉e-mail: anabelenlb@ugr.es

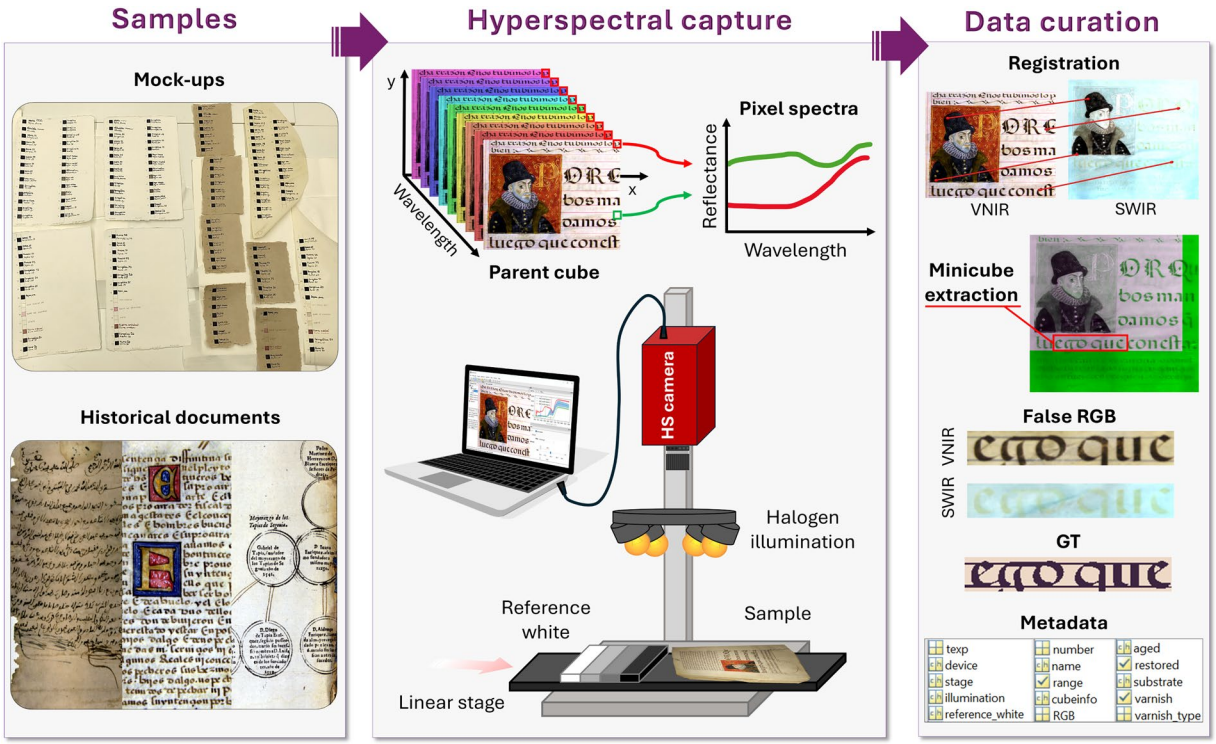

Fig. 1 Graphical abstract of the capture methodology and data curation.

Knowledge about the materials used provide insights into the provenance, authenticity, and historical context of documents[13], aiding in tasks like dating manuscripts[28], determining authorship, detecting falsifications or undocumented restorations, identifying causes of deterioration[12,29] and developing preservation and restoration strategies[12,30].

Spectral datasets can be very helpful for performing material identification. Some datasets of pigments have been proposed and used in the past for this purpose. For example, a multispectral dataset of pigments with information of 18 bandpass filters covering the 400-925 nm range was proposed by A. Cosentino (https://chsopensource.org/multispectral-imaging-pigments-checker-standard/). In this case, the Pigments Checker STANDARD v.5 of CHSOS was captured, including 5 blacks among 69 pigments applied on cardboard. From the data, it was shown that browns and blacks lack sufficient spectral features for identification. Therefore, multispectral imaging in that spectral range is not adequate. A more comprehensive hyperspectral pigment dataset was later developed by Deborah *et al.*[31], containing 195 pigment patches from Kremer color charts in the range of 400 to 1000 nm. In these, 14 black pigments were included, applied on acid-free 180 gram paper. Despite these advances, to our knowledge, no hyperspectral dataset exists with the purpose to assist in material identification for historical documents. To further advance the field, the availability of datasets with annotated ground truth is essential. Such datasets are critical for training and evaluating machine learning and deep learning models to perform different tasks in historical document analysis[32]. However, the creation of ground-truth data is often labor-intensive and costly. Additionally, datasets must encompass sufficient diversity to enable models to generalize effectively, including documents of varying origins, materials, and historical periods.

In this work, a hyperspectral dataset of historical documents and mock-ups captured in the 400-1700 nm spectral range is presented[33]. This dataset was developed within the framework of the HYPERDOC project (https://colorimaginglab.ugr.es/pages/hyperdoc/project), which focuses on using hyperspectral imaging to analyze historical and artistic documents for material identification and mapping.

The workflow for data collection, capture, and post-processing is summarized in Fig. 1. Ink mock-ups were created using historical recipes and materials[34,35], including metallo-gallate inks, sepia, carbon-based inks, and mixtures, which were applied on five different supports. The dataset also includes pencil mock-ups and historical inks subjected to artificial aging. Additionally, historical documents from the 15th to 17th centuries, sourced from the Archive of the Royal Chancellery and the Provincial Historical Archive of Granada (Spain), were captured within two spectral ranges: 400-1000 nm and 900-1700 nm, which were spatially registered. From these captures, the full hypercubes (referred to as parent cubes) were cropped into smaller, representative Regions of Interest, termed *minicubes,* to facilitate faster data processing. Ground Truth (GT) annotations were created to label the materials present at the pixel level. False-color RGB images were generated for both spectral ranges, and Metadata was also integrated into each parent cube and minicube to provide detailed information.

The HYPERDOC dataset stands as a unique and versatile resource, integrating mock-ups and historical documents to support a wide range of applications in hyperspectral imaging and historical document analysis. Its subsets have been utilized in studies addressing diverse tasks such as ink classification using machine learning

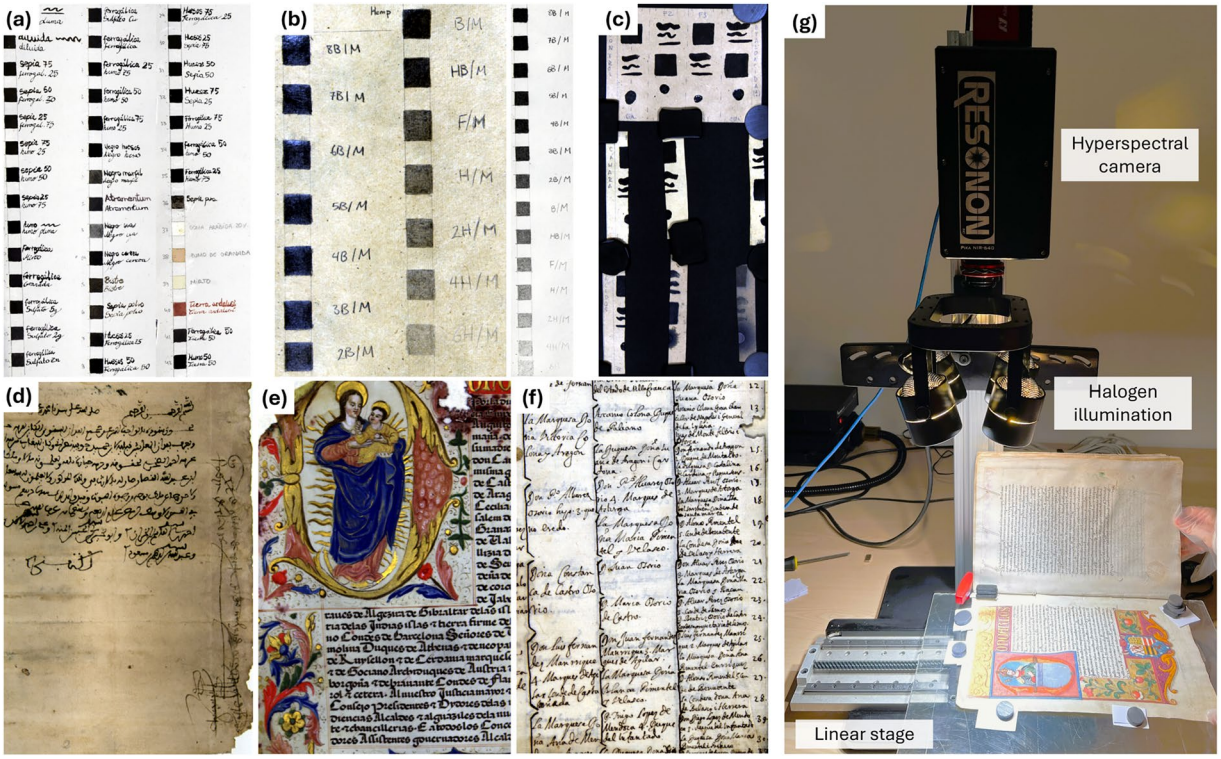

Fig. 2 Examples of samples from different subsets: (**a**) mock-ups of historical inks on cotton paper, (**b**) pencil mock-ups, (**c**) mock-ups of artificially aged metallo-gallate inks, (**d**) manuscript of the Provincial Historical Archive of Granada, (**e**) illuminated manuscript from the Archive of the Royal Chancellery of Granada, (**f**) family tree from the Archive of the Royal Chancellery of Granada; and (**g**) hyperspectral capture using the Pika IR+ camera.

and deep learning techniques[36,37], binarization to enhance text legibility[38], spectral unmixing to identify the materials[39], and colorimetric analysis of aging processes in mock-ups[40,41]. Thanks to the variety of recipes in mock-ups, they can be used to perform multivariate exploratory analysis in order to identify spectral features related to some kinds of inks or their components, and spectral changes related to the amount of ink deposited. It also facilitates comparative studies between mock-ups and historical documents, including analyses of artificial versus real aging, the state of conservation of the real documents in comparison to other samples, and the impact of different supports or writing instruments on spectral properties. The spectral data also allow simulations of document appearances under various illuminants.

Moreover, the HYPERDOC dataset fosters interdisciplinary collaboration between the image processing and restoration-conservation communities, encouraging the adoption of advanced techniques such as hyperspectral imaging, which remains underutilized in practical applications within archives and museums. This dataset holds significant potential to drive innovation in the restoration and preservation of our cultural heritage.

Methods

Samples description. *Mock-ups. Mock-ups of historical inks on different supports*. Mock-up samples of inks applied on different supports were prepared following historical recipes from the 13th to the 17th centuries and bound with Arabic gum[42]. Some samples from this subset of the dataset were introduced in a previous study[37]. The contour of a 1 x 1 *cm* square was drawn with pencil and filled with ink, and two lines of text were written with brush (top) and fountain pen (bottom) (see Fig. 2(a)). The exact quantities of ingredients used to prepare the inks are detailed in the column *'general info'* of the file *'HYPERDOC_dataset_info.ods'* included in the dataset, and the inks used are listed below:

- Metallo-gallate inks (or iron gall inks): Prominent from the medieval period onwards, particularly in Europe[43], these inks consist primarily of two key components: a metal (predominantly ferrous sulfate), and a vegetable tanning agent, typically derived from oak apples in the form of gallotannin extracts. Ancient recipes of iron gall ink were followed[42,44], resulting in eight variants created using different ratios of gallic acid and ferrous sulfate, including variants with copper and zinc sulfate. Detailed preparation can be found in reference[37]. Additionally, Andalusian ink recipes[44] were used to produce two variants incorporating pomegranate juice and myrtle leaf infusion. In addition, the pigment Atramentum (Kremer Pigmente GmbH) was used.
- Sepia inks: Derived from the ink sac of the cuttlefish *Sepia officinalis*[45]. Two types of sepia ink were used in this dataset: one extracted directly from the animal and another obtained in powder form from Kremer

Pigmente GmbH. For the natural extraction, three samples were prepared: pure ink diluted with water, and two others varying the concentration of Arabic gum binder.

- Carbon-based inks: Believed to be the earliest form of writing ink, carbon-based inks were traditionally made by burning oil or other materials (such as fruit stones, bones, or wood) in controlled conditions with limited air supply and mixing the resulting soot with a binder dissolved in a water-soluble medium[45]. The dataset includes several carbon-based inks from Kremer Pigmente GmbH: ivory black, lamp black, grape seed black, cherry black, and bistre.
- Mixed inks: Mixed inks, though recently gaining attention in scholarly and material studies, have been significant historically, especially in the Islamic world, as suggested by medieval Arabic recipes[46,47]. They have also been found in ancient contexts[30,45], but they still remain challenging to identify. The dataset includes 17 mixed ink formulations, combining sepia, iron gall ink, lamp black, bone black, and Andalusian red earth (from Kremer Pigmente GmbH) in various proportions.

All inks were applied to five types of support, selected for their historical relevance: three types of handcrafted paper from Paperlan® made of 100% cotton fiber, 100% linen fiber, and a linen/cotton mixture 50/50%, hemp paper from Wanderings®, and goatskin parchment from Forum Traiani®. These supports were selected based on those commonly found in historical documents[48].

The configuration of the mock-ups enables the study of spectral changes related to the amount of ink deposited or the recipe used for the same type of ink[37]. Additionally, the effects of different writing instruments on the spectral characteristics can also be evaluated.

Pencil mock-ups. This subset of mock-ups includes 14 pencil types from Faber Castell®with varying grades of hardness (8B, 7B, 6B, 5B, 4B, 3B, 2B, B, HB, F, H, 2H, 4H, and 6H) applied to 4 different supports: cotton-linen, cotton, linen, and hemp paper. Similarly to the ink mock-ups, a 1 x 1 *cm* square was filled with pencil and a line of text indicating its hardness was written (see Fig. 2(b)). In total, there are 56 samples.

Mock-ups of artificially aged metallo-gallate inks. Three variants of metallo-gallate inks were deposited on hemp paper, including pure iron gall ink, iron gall ink with copper sulfate, and a mixture of iron gall ink and lamp black. These inks were used to create squares, strokes, and drops, and subsequently subjected to artificial aging using two distinct methods (see Fig. 2(c)). In the first method, an aging chamber (Solarbox®3000 eRH, Neurtek) was used following the norm ISO 5630-3 (1996). The chamber operated at a temperature of 80 C, a relative humidity of 65%, and a radiation of 550 W/m^2. Samples were extracted and captured after 0, 72, 144, and 288 hours of aging, corresponding to 0, 3, 6 and 12 days, respectively. In the second method, aging under acidic conditions was studied by exposing samples to hydrochloric acid vapors for 72, 144, and 288 hours.

Historical documents. *Manuscripts of the Provincial Historical Archive of Granada*. This subset comprises five different documents preserved in the collection of Arabic documents at the Provincial Historical Archive of Granada[49]. Four of these are notarial documents dating from 1488 to 1494, while the fifth is an undated religious text. All five contain handwritten text (see example in Fig. 2(d)). Scanning Electron Microscopy with Energy Dispersive X-ray Spectroscopy (SEM-EDX) has identified various types of inks in these documents, including mixed iron gall ink with earth pigments, mixed carbon-based ink with earth pigments, pure carbon-based inks, and pure iron gall inks. The support used in all the documents has been confirmed to be linen paper, as determined through a combination of optical microscopy, Scanning Electron Microscopy (SEM), and Fourier Transform Infrared Spectroscopy (FTIR)[49,50].

Illuminated manuscripts from the Archive of the Royal Chancellery of Granada. This collection comprises seven documents on parchment, containing lawsuits of nobility dating from 1459 to 1608[51] (see example in Fig. 2(e)). Different pigments and dyes are present in certain areas of these documents; however, these regions were not included in the main focus of this dataset, which is primarily on inks. The inks used in the handwritten text were identified as iron gall ink with different sulfates, specifically zinc (Zn) and copper (Cu). Ink identification was performed using X-ray fluorescence (XRF), while the inorganic elements used in the preparation of supports were identified using a combined X-ray diffraction and X-ray fluorescence system[52].

Family tree book from the Archive of the Royal Chancellery of Granada. This series of eight documents from the 16th and 17th centuries comprises family trees, predominantly handwritten with some stamped sections (see example in Fig. 2(f)). All documents have cotton-linen paper as the support. Previous analyses identified two types of ink: a carbon-based ink and a mixture of sepia and iron gall ink. The documents were restored in 2005 through mechanical cleaning with non-greasy soft rubbers, washing in water, and drying under weight and blotters. The ink types were characterized using SEM by the conservators in charge.

Hyperspectral imaging capture. Two line-scan hyperspectral cameras from Resonon Ltd. (Bozeman, Montana, USA) were used, together with the associated software Spectronon Pro 3.5.5: the Pika L and the Pika IR+. Details about the spectral range covered by each camera, number of spectral channels, spectral resolution, number of pixels per line (spatial pixels), maximum frame rate, and F-number for each camera are provided in Table 1. These cameras operate on a push-broom technique, capturing images line by line, which requires either the movement of the object or the camera to scan the entire scene. For image acquisition, a linear translation stage from Resonon Ltd. was used along with 4 stabilized halogen lamps positioned to minimize specular reflections and placed at 30 cm from the documents. To ensure controlled lighting conditions, all other lights in the room were turned off. A video illustrating the capture process is available at the following link: https://www.youtube.com/watch?v=lIwNyjBeKmQ. The optimal exposure time was determined using the 90% reflectance patch from the Sphere Optics Zenith Lite Multistep of size 20 × 20 cm, or a Teflon bar with known reflectance, serving as the reference white. Then, the software automatically adjusted the scanning speed and camera data acquisition to ensure 1:1 vertical:horizontal aspect ratio. To maintain the reference white and the document at the same distance

Parameter	Pika L	Pika IR+
Spectral Range (nm)	400–1000	900–1700
Spectral Channels	300	368
Spectral Resolution - FWHM (nm)	2.7	5.6
Spatial Pixels	900	640
Max Frame Rate (fps)	249	240
f/#	2.4	1.8

Table 1. Specifications for the Pika L and Pika IR+ hyperspectral imaging systems.

from the camera, magnets and additional supports were used, as shown in Fig. 2(g). The distance between the camera and the samples was approximately 60 cm for the VNIR camera and 40 cm for the SWIR camera, resulting in linear fields of view (swath) of 13.5 cm and 14.5 cm, respectively. This setup yielded an estimated spatial resolution of 0.15 mm/pixel for the VNIR range and 0.227 mm/pixel for the SWIR range. Spectral binning was performed during capture to enhance the signal-to-noise ratio, resulting in 150 and 168 bands for the VNIR and SWIR captures, respectively. Before capturing the documents, reference images for calibration were acquired. To convert raw data into reflectance values, dark subtraction and flat-field correction were applied. These steps ensure that variations in illumination and sensor response across the field of view are accounted for, eliminating system-induced artifacts from the data. For this purpose, 30 lines of the reference white, the 90% reflectance patch from the Sphere Optics Zenith Lite Multistep or a Teflon bar, were captured. The mean value along the longitudinal axis was then used as the reference white to correct non-uniformities in illumination and determine the light incident on the sample. A dark reference image was also captured by blocking the light entering the camera, allowing the removal of intrinsic sensor noise caused by dark currents. All captures were saved in BIL format. Further details on the conversion of raw captures to reflectance data are provided in next subsection.

Data curation. *Reflectance from raw.* Raw captures of the spectral cubes ($Raw(\lambda)$) were transformed into reflectance cubes ($\rho(\lambda)$) using the raw captures of the reference white ($Raw(\lambda)_{\text{white}}$) and the dark image ($Raw(\lambda)_{\text{dark}}$) according to the following equation:

$$\rho(\lambda) = \rho(\lambda)_{\text{white}} \cdot \frac{Raw(\lambda) - Raw(\lambda)_{\text{dark}}}{Raw(\lambda)_{\text{white}} - Raw(\lambda)_{\text{dark}}}, \quad (1)$$

where $\rho(\lambda)_{\text{white}}$ is the spectral reflectance of the reference white used for calibration. This transformation is performed pixel by pixel in the camera software before cube storage, except for the multiplication by $\rho(\lambda)_{\text{white}}$, as the software assumes it to be 100% at every wavelength.

Once the reflectance cubes were in BIL format, they were converted to MAT format using MATLAB (Release R2023a, The MathWorks, Inc., Natick, MA, USA). The code used for the transformation is available on GitHub (https://github.com/anabelenlb/HYPERDOC_Database_code). It is during this step that the reflectance of the object is multiplied by the reflectance of the reference white ($\rho(\lambda)_{\text{white}}$). During this process, linear interpolation was applied to ensure a consistent 5 nm sampling interval across both cameras, resulting in 121 bands between 400 and 1000 nm for the VNIR and 161 bands between 900 and 1700 nm for the SWIR.

Registration. Spatial registration is performed to align pixel-by-pixel captures of the VNIR and SWIR ranges and equalize the spatial resolution. In this case, the SWIR capture was used as the reference image, while the VNIR capture, with its higher spatial resolution, was transformed to minimize artifacts in the final registered image. The registration was performed using one band from the VNIR (700 nm) and one band from the SWIR hypercubes (1000 nm). These bands were selected based on preliminary trials, and their position below 1200 nm in the SWIR range. The latter condition was set to avoid proximity to the onset of the high reflectance region of metallo-gallate inks in the SWIR range, which could lead to a lack of key points necessary for proper registration. Feature-based image registration with SURF features[53] was used within the MATLAB Registration Estimator App (Release R2023a, The MathWorks, Inc., Natick, MA, USA), applying either an affine or projective spatial transform. The registration quality was assessed using overlay images and the Structural Similarity Index Measure (SSIM)[54], after testing different features and spatial transforms, to ensure satisfactory registration. The final registration transformation was then applied to all spectral bands within the VNIR cube. After the process, the parent cubes are obtained, that is, the hyperspectral images of the full pages from which the minicubes are extracted later on. These parent cubes are included in the folder 'ParentCubes' in the HYPERDOC dataset[33].

Minicube extraction. We define a *minicube* as a crop extracted from a full document or page, with sizes ranging from [34 x 33] to [181 x 508] pixels, selected from representative areas of the full documents. Spectral images delivered by the hyperspectral devices are stored as spectral cubes (i.e. hypercubes), usually of extremely large size of even gigabytes of data per capture. Thus, the extraction of minicubes facilitates faster processing of spectral information, as explained before. Each minicube contains data from one or two inks, the support, and sometimes pencil markings. Minicube extraction was performed on the registered VNIR and SWIR cubes using identical spatial coordinates, with Regions of Interest selected based on areas where different inks or materials were present in the document. These minicubes are included in the folder 'minicubes' in the HYPERDOC dataset[33].

Fig. 3 Steps involved in the creation of the Ground Truth (GT) images.

False RGB images and Ground Truth creation. For each parent cube and minicube, a false-color RGB image was generated by assigning specific spectral bands to the [R G B] channels. For the VNIR range, bands [50 34 9], corresponding to wavelengths of 645 nm, 565 nm, and 440 nm, respectively, were used, yielding a color appearance similar to that observed in the sample. Similarly, for the SWIR range, bands [141 61 21], representing wavelengths of 1700 nm, 1300 nm, and 1100 nm, were selected[33].

A Ground Truth image (GT) was also created for each minicube using a semi-automatic method. The process is illustrated in Fig. 3 and consists of two main steps: binarization, which separates the foreground (pigments and inks) from the background (support), and GT creation, which assigns labels to differentiate between materials. The binarization process involves four steps. First, a band with high contrast between the ink and the background was selected by searching for the minimum Signal-to-Noise Ratio (SNR) value. Second, the skeleton of the part of the image covered with ink was extracted using the MATLAB R2023a function bwskel, based on Lee *et al.*'s medial surface axis thinning algorithm[55]. Third, the skeleton width was adjusted until the intensity of surrounding pixels matched the average intensity of the borders of the ink covered region, extracted by the Canny edge detector. This is a variation of the method proposed by Ntirogiannis *et al.*[56], where the skeleton was manually corrected and then forced to grow until it met those borders. Fourth, manual correction using the open-source software GIMP was performed after obtaining the binarized image, by visually comparing the result with a false RGB image of the minicube. A different annotator then evaluated the revised GTs, and any discrepancies were resolved through consensus. After completing these steps, a binary image was generated with two labels: 0 (background) and 1 (foreground). The GT was then created by assigning different labels to distinguish materials in the binary image. The mapping between labels, GT colors, and materials is provided in the file *'Materials_label_and_colormap_assignation.ods'*. The final indexed images, including the index map and associated colormap, were saved as PNG files in the folder 'GT'[33]. GT images were not created for the parent cubes due to difficulties in providing accurate pixel-level annotations for these larger and more complex regions.

Metadata info. Detailed information about each sample is included as metadata within the minicubes. In total, 24 attributes were included, which can be divided into three main categories according to the type of information provided: sample information, capture information, and other relevant data.

The sample information group contains 11 attributes: identifier number, name, general information about the sample (ink components, origin or recipe), type of support, height in pixels, width in pixels, number of bands, wavelengths captured, date of production of the document, aging status (either naturally or artificially aged, or not aged at all), and restoration status (whether restored or not).

The capture information group contains 6 attributes: device used, range captured, stage, exposure time, type of illumination, and reference white used.

The other relevant information group contains 7 attributes: colormap of the GT, GT labels, parent cube name, and pixel coordinates used to extract the minicube within the parent cube. Using the GT from the PNG files and GT labels, spectra of pixels belonging to each class are averaged and the mean and standard deviation is stored, along with the number of pixels used in the average.

For the parent cubes, metadata were also included. In this case, as GT images are not available, only 16 attributes were included, excluding the identifier number and all attributes in the 'other relevant information' group.

Data Records

The HYPERDOC dataset[33], comprising hyperspectral images of historical documents and mock-ups, is publicly available in the *Figshare* repository. The data is structured as shown in Fig. 4 and includes the following folders and files:

- Folder *minicubes* - Hypercube files with metadata (in HDF5 format): This folder contains hyperspectral datacubes for each minicube, captured in the VNIR and SWIR spectral ranges, along with associated metadata. The datacubes are stored as a 64-bits double-precision floating-point matrices with dimensions M x N x λ, where M and N correspond to the spatial dimensions of the image, and λ represents the number of spectral bands or wavelengths. Each datacube includes metadata with 24 attributes, describing key information such as acquisition settings and sample details. A detailed description of these attributes, including their data types and possible values, is provided in Table 2. For the spectral range, three options

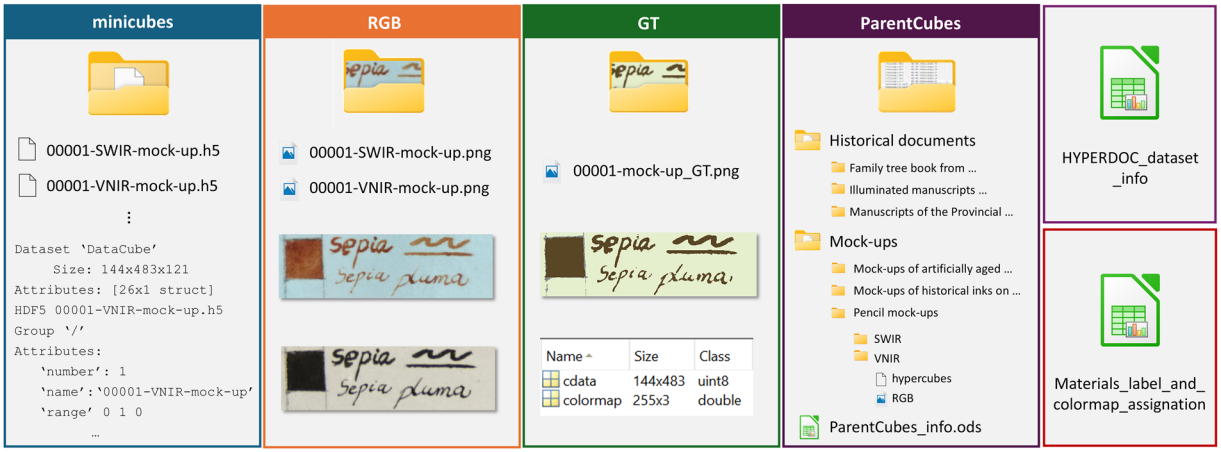

Fig. 4 Folder and file structure in the dataset.

are defined, including ultraviolet and visible (UVIS), even though no samples within this range are currently present in the dataset. This option has been included to accommodate future additions of UVIS samples to the dataset on the research group's website (https://colorimaginglab.ugr.es/pages/hyperdoc/project).

- Folder *RGB* - False RGB images (in PNG format): This folder contains false-color RGB images for both the VNIR and SWIR ranges, generated using the methods described in *'Methods'* section. These images provide a convenient way to quickly visualize the minicube content.
- Folder *GT* - Ground Truth images (in PNG format): This folder includes ground truth (GT) images, where pixel values are directly mapped to colormap indices that relate to RGB data specific for each material and assigned according to a pre-existing list of materials present in the dataset. Each file contains an indexed image stored in the variable *'cdata'*, along with the associated colormap. The correspondence between material types and their respective indices and RGB values in the colormap is documented in the file *'Materials_label_and_colormap_assignation.ods'*.
- Folder *ParentCubes*: This folder is organized in two subfolders, one for each data set: 'Historical documents' and 'Mock-ups'; and an OpenDocument Spreadsheet (in ODS format) named *'ParentCubes_info.ods'*, which contains parent cubes information for quick reference, extracted from the metadata. Each folder has its own subfolders for each data subset. Within those subfolders, there are the 'VNIR' and 'SWIR' folders, which contain the hypercubes in .h5 format, including the datacube and metadata with 18 attributes, as well as the 'RGB' folder with false RGB images in PNG format.
- File *'Materials_label_and_colormap_assignation.ods'* - OpenDocument Spreadsheet with material labels and colormap assignments (in ODS format): This file contains the mapping between material types (e.g., inks, pencils, or supports) and their corresponding indexed values in the GT images. The colormap assigns RGB values (0-255) to each material index. For each minicube, the associated materials can also be found in the *'GTLabels'* attribute within the metadata.
- File *'HYPERDOC_dataset_info.ods'* - OpenDocument Spreadsheet with hypercube information (in ODS format): The spreadsheet provides essential information about the minicubes for quick consult extracted from the metadata, which includes the set they belong to (mock-ups or historical documents), subset, name of the minicubes, name of the parent cubes, coordinates within the parent cube used for minicube extraction, general information about the minicube, materials found, support, date, information about if it is aged, restored, or not, and finally, exposure time and reference white used during capture.

Table 3 summarizes the number of minicubes and the total pixel counts associated with each material type across the different subsets. For a visual representation, Fig. 5 presents the distribution of minicubes (top) and the distribution of pixels on a logarithmic scale (bottom) across subsets and classes in the dataset using horizontal bar graphs. To the right of 0, the minicubes correspond to historical documents, while to the left, they belong to the mock-ups set. A higher number of minicubes is included in the mock-up category compared to historical documents, representing 73% of the total dataset. In terms of pixel count, mock-ups account for nearly 94% of the total.

The high number of minicubes containing pencil is due to the first two subsets of mock-ups (historical inks on different supports and pencil samples), where all samples include pencil. However, this material is completely absent in the set of historical documents. Other materials or classes not present in the historical document subsets include cotton, hemp, pure andalusian red earth, pure sepia, mixture of carbon and sepia, and a combination of metallo-gallate ink with carbon-based ink. Among historical documents, the most represented ink class is pure metallo-gallate ink. For the supports, linen, cotton-linen, and parchment are equally distributed in terms of minicube count; however, in pixel count, linen is less represented compared to the other two. In mock-ups, red earth and its mixtures are among the least represented materials, reflecting their limited historical use.

Group	Metadata	Description	Data type
1	number	Identifying number, from 00001 onwards.	String
1	name	Name of the minicube.	String
1	cubeinfo	Information about the type of sample (mock-up or historical) and relevant details such as materials or recipes used.	String
1	substrate	Type of substrate or support: parchment or paper, and fiber type.	String
1	height	Height of the minicube in pixels.	32-bits unsigned integer
1	width	Width of the minicube in pixels.	32-bits unsigned integer
1	bands	Number of spectral bands in the minicube.	32-bits unsigned integer
1	wl	Wavelengths or spectral bands captured.	32-bits unsigned integer
1	date	Date of creation of the mock-ups or historical documents (year or century, depending on the information available).	String
1	aged	Indicates whether the document is aged. Possible values: 'No' (not aged), 'Art.' (artificial aging, with details of the method and hours), and 'Nat.' (naturally aged).	String
1	restored	Indicates whether the document has been restored. Logical value.	8-bits unsigned integer
2	device	Hyperspectral camera used in the capture.	String
2	range	3 × 1 logical indicating the capture range: 1 0 0 for UVIS; 0 1 0 for VNIR; 0 0 1 for SWIR.	8-bits unsigned integer
2	stage	Translation stage used to perform the capture. All captures were done using the 'linear' stage.	String
2	texp	Exposure time during capture in milliseconds.	64-bits double precision floating point
2	illumination	Illumination used during the capture. Halogen lamps were used in all cases.	String
2	reference_white	Reference white used to calibrate reflectance measurements. Possible values: 'Teflon' (for the teflon bar) and 'Multi_90' (the 90% reflectance patch from the Sphere Optics Zenith Lite Multistep).	String
3	GT_cmap	RGB values from 0 to 1 associated with each index in the GTs. Size 16 x 3.	64-bits double precision floating point
3	GTLabels	Materials associated to indexes used in GT.	String
3	parent_cube	Hypercube from which the minicube was extracted.	String
3	position	Coordinates used to extract the minicube from the parent cube: [xmin xmax ymin ymax].	32-bits unsigned integer
3	spectra_mean	Mean spectra of all pixels associated with the same index in the GT. Size: λ x number of indexes in the GT.	32-bits single precision floating point
3	spectra_std	Standard deviation of the mean spectra. Size: λ x number of indexes in the GT.	32-bits single precision floating point
3	pixels_averaged	Number of pixels associated with the same indexes in the GT and used to calculate the mean. Size: 1 x number of indexes in the GT.	32-bits unsigned integer

Table 2. Description and data types of attributes in the Metadata field within each minicube, categorized into three groups: (1) sample information, (2) capture information, and (3) other relevant data.

Technical Validation

Some subsets or samples of this dataset have been used in previous studies to perform tasks such as classification[36,37], binarization[38], spectral unmixing[39], and colorimetric analysis[40,41]. A schematic representation of the results obtained in these studies is provided in Fig. 6.

In a recent study, three categories of inks were classified using machine learning techniques: pure metallo-gallate inks, carbon-containing inks, and non-carbon-containing inks[36]. Five traditional machine learning algorithms—Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Random Forest (RF), and Partial Least Squares Discriminant Analysis (PLS-DA)—as well as a Deep Learning-based model, were trained and evaluated on mock-ups of historical inks applied to different supports and on all the minicubes extracted from historical documents in the present dataset. Data from both the VNIR and SWIR ranges were combined using data fusion techniques, achieving a micro-averaged accuracy exceeding 90%. Examples of classification maps are shown in Fig. 6. The carbon-containing ink category included pure carbon-based inks and mixtures of these with metallo-gallate or sepia inks. Similarly, the non-carbon-containing group included sepia inks and mixtures of sepia with metallo-gallate ink[36]. Such spectral identification of inks would directly inform the selection of appropriate restoration materials and conservation protocols for historical documents.

A part of the subset of mock-ups of historical inks on different supports was previously presented[37] and a classification task was performed using VNIR data from samples applied to parchment and cotton-linen paper. A Bilayered Neural Network was trained and tested to distinguish four classes: iron gall ink, non-iron gall ink, support, and pencil. The first group contained all the pure and mixed inks with some amount of iron gall ink. The model achieved an overall accuracy of 67%, with notably higher precision in support classification (97%) compared to ink classification, particularly when carbon-based inks were mixed with iron gall inks.

In another study, spectral unmixing techniques have also been applied to identify pure components in mixtures of historical inks, using both mock-ups and selected parent cubes from historical documents included in this dataset[39]. Mixtures of iron gall, sepia, and carbon-based inks were analyzed by merging the VNIR and SWIR ranges. Pencil and support materials, including parchment and cotton-linen fibers in paper, were also incorporated into the unmixing process. Challenges were encountered in accurately identifying components when carbon was present in the mixtures. This is due to the very low spectral reflectance of carbon-based inks across

all wavelengths. When mixed with other components, the spectrum of the mixture remains uniformly low due to the subtractive nature of the mixture, and this hinders the identification of non carbon-based components. Examples of concentration maps and error maps derived from the spectral reconstruction are shown in Fig. 6.

In addition to classification, a binarization task was also explored[38], where several binarization algorithms including Otsu[57], Niblack[58], Wolf[59], Bradley[60] and a Deep Learning-based method, were evaluated using a subset of 16th- and 17th-century family trees from the Archive of the Royal Chancellery of Granada. When comparing results from the VNIR and SWIR ranges, the Bradley algorithm consistently produced the best results. These binarization techniques offer strong potential for conservation, improving manuscript readability and enhancing Optical Character Recognition (OCR) which supports both digital preservation and content accessibility.

For the subset of artificially aged metallo-gallate inks, both colorimetric and spectral analyses were performed to investigate their aging processes. Spectral differences were examined using metrics such as Root Mean Square Error (RMSE) and the complement of the Goodness-of-Fit coefficient (cGFC), while color differences were assessed using the CIEDE00 color difference formula[40,41]. These results support conservation strategies by linking spectral features to degradation processes, allowing conservators to base decisions on spectral evidence rather than visual inspection alone.

Beyond these specific applications, the mean reflectance spectra and standard deviation for each class and subset were calculated to preliminary explore the spectral features in the dataset (see Figs. 7 and 8 and the next section for details). Additionally, Principal Component Analysis (PCA) was used as a dimensionality reduction technique for visualization purposes (see Fig. 9).

Average spectra and standard deviation per class and subset. The mean spectra and standard deviation were computed for each foreground material, including pencil, pigments and inks (left plots), and for each support material (right plots), with the results separated by subset (different rows in Fig. 7 for the mock-ups and Fig. 8 for the historical documents). These calculations were performed by averaging all the pixels in the minicubes belonging to the same class and subset, using the ground truth (GT) information (example code for performing this calculation in MATLAB and Python is available on GitHub https://github.com/anabelenlb/HYPERDOC_Database_code). VNIR and SWIR spectral ranges are presented in the same plot. A noticeable difference in the reflectance spectra is observed in the overlapping region between 900 and 1000 nm, which is a common artifact when data is captured using different sensors. This discrepancy arises from various factors, including differences in spectral bandwidths, low signal-to-noise ratios due to low sensor responsivity in the extremes of the spectra, and slight misalignments in the image acquisition setup, all of which can impact the Bidirectional Reflectance Distribution Function (BRDF)[61]. To mitigate this issue, various strategies can be employed to ensure a smooth connection between the spectra. One such approach, proposed by Grillini *et al.*[62], involves a logistic splicing correction.

Ink spectra in the visible range are similar, showing low reflectance values and flat shapes, consistent with the black or brownish appearance of these inks. However, in the near-infrared range, metallo-gallate inks, both pure and mixed with red earth, begin to diverge from other inks, exhibiting a reflectance spectrum increasingly similar to that of the support. This trend is especially prominent beyond 1200 nm, where metallo-gallate inks become nearly transparent. In contrast, carbon-based inks strongly absorb infrared radiation, maintaining low reflectance values. Sepia ink and its mixtures with metallo-gallate allow more infrared transmission but do not reach the near-total transparency observed in metallo-gallate inks. Mixtures of carbon-based inks with other pigments, such as sepia or red earth, significantly reduce reflectance, resulting in spectra resembling those of pure carbon-based inks. Pencil spectra generally exhibit flat shapes with lower reflectance values in the pencil mock-ups subset due to the inclusion of pencils with varying hardness, including very dark grades. In the subset of mock-ups with historical inks, only HB pencils were used. Andalusian red earth displays a spectrum characteristic of red hues, while it becomes transparent in the infrared.

Support materials such as cotton-linen, linen, and cotton exhibit similar reflectance spectra. Parchment shares a similar shape but demonstrates lower reflectance values, indicative of its darker tone. Additionally, parchment tends to show greater heterogeneity, particularly in historical documents, as modern parchment is generally more uniform in composition and appearance. Hemp has a different shape from the others, making it easily distinguishable from other supports.

For the historical documents, notable differences emerge when comparing spectra to those of mock-ups. For instance, carbon-based inks in historical manuscripts, particularly those from the Provincial Historical Archive of Granada (first row, Fig. 8), do not exhibit the nearly complete absorption of infrared radiation seen in mock-ups. Instead, these inks also become partially transparent in the infrared range. This discrepancy could be attributed to aging processes, such as surface wear due to rubbing, which may reduce the ink layer thickness, causing the spectrum to resemble that of the underlying support. Similarly, minor variations in support spectra compared to mock-ups are likely due to aging effects.

Diversity in standard deviation is observed across subsets, derived from factors such as the use of different supports or variability within ink classes. For example, the pure metallo-gallate ink class includes inks with additives like pomegranate juice, myrtle infusion, or varying amounts of Cu, Zn, and Fe sulfates. Likewise, the pencil subset includes a range of hardness grades. This variability is intentional, as it enhances the robustness of classification models by capturing a wide range of potential conditions, preparing them for real-world applications rather than highly controlled scenarios.

Comparing spectral libraries of black pigments or inks is challenging, as they consist of different samples prepared using diffrent techniques, binders, supports, and different data acquisition procedures or instruments. While it can be found only one hyperspectral pigment dataset published in the range of 400 to 1000 nm[31], in general, the already existing libraries contain a single homogeneized measurement per sample made by

Set	Subset	Class			Number of minicubes	Number of pixels
Mock-ups	Historical inks on different supports	Metallo-gallate ink	pure		40	612,949
		Metallo-gallate ink	mixture	carbon	30	452,478
		Metallo-gallate ink	mixture	sepia	15	235,795
		Metallo-gallate ink	mixture	earth	5	82,876
		Carbon	pure		30	429,217
		Carbon	mixture	earth	5	64,782
		Carbon	mixture	sepia	30	423,009
		Sepia	pure		20	300,988
		Pencil			180	183,024
		Andalusian red earth pigment			5	89,075
		Parchment			36	1,204,442
		Cotton-linen			36	1,217,969
		Linen			36	1,456,972
		Hemp			36	1,229,443
		Cotton			36	1,433,623
	Pencil	Pencil			56	153,920
		Cotton-linen			14	121,414
		Linen			14	106,705
		Hemp			14	99,962
		Cotton			14	119,084
	Artificially aged metallo-gallate inks	Metallo-gallate ink	pure		18	95,255
		Metallo-gallate ink	mixture	carbon	10	51,403
		Hemp			28	203,886
Historical documents	Manuscripts Provincial Historical Archive	Metallo-gallate ink	pure		20	13,930
		Metallo-gallate ink	mixture	earth	5	3,501
		Carbon	pure		4	1,942
		Carbon	mixture	earth	3	2,434
		Linen			36	82,664
		Metallo-gallate ink	mixture	unknown	4	1,896
	Illuminated manuscripts Royal Chancellery Archive	Metallo-gallate ink	pure		29	94,981
		Parchment			29	251,723
	Family tree book Royal Chancellery Archive	Metallo-gallate ink	mixture	sepia	24	29,109
		Carbon	pure		23	16,654
		Cotton-linen			31	196,182

Table 3. Classes and total number of minicubes and pixels in each set and subset (commas used as thousands separators).

using spectrorradiometers or Fiber Optics Reflectance Spectroscopy (FORS)[63] (https://chsopensource.org/pigments-checker/). As the dataset presented in this paper contains hundreds or thousands of datapoints, the average spectra of mock-up inks were used for comparison with existing ink spectral libraries.

The hyperspectral pigment dataset[31] includes seven pigments or inks that are also present in our dataset: grape black, ivory black, cherry black, bistre, atramentum, Andalusian red earth (or red ochre), and sepia. In our classification, the first four inks are grouped together as carbon-based inks due to the same origin and similar spectral characteristics. This grouping was validated by comparing the mean spectra provided by the authors of the dataset (https://hyppigments.streamlit.app/) with our data, confirming that carbon-based inks exhibit a consistently flat, low reflectance in the VNIR range. Similarly, the spectra for Andalusian red ochre, atramentum, and sepia show comparable shapes across both datasets.

Another publicly available dataset includes reflectance spectra obtained with spectrometers such as the GorgiasUV (200-1000 nm) and InGaAs (900-1700 nm) spectrometers (https://chsopensource.org/pigments-checker/). This dataset contains seven black pigments or inks also present in our dataset: ivory black, vine black, bone black, lamp black, iron gall ink, Andalusian red ochre, and sepia. Again, the first four were grouped as carbon-based inks, exhibiting flat spectra between 400 and 1700 nm. The iron gall ink becomes transparent in the infrared, and it is interesting to see how the shape of the spectra in the 900-1700 nm range is completely influenced by the support, since in this case it was deposited on cardboard. Sepia ink becomes transparent beyond 1500 nm, displaying a common feature with the sepia spectra in our dataset, as shown in Fig. 7 upper left plot.

Overall, our spectral data aligns well with existing datasets of inks and black pigments, supporting the reliability of this dataset for further analysis and applications.

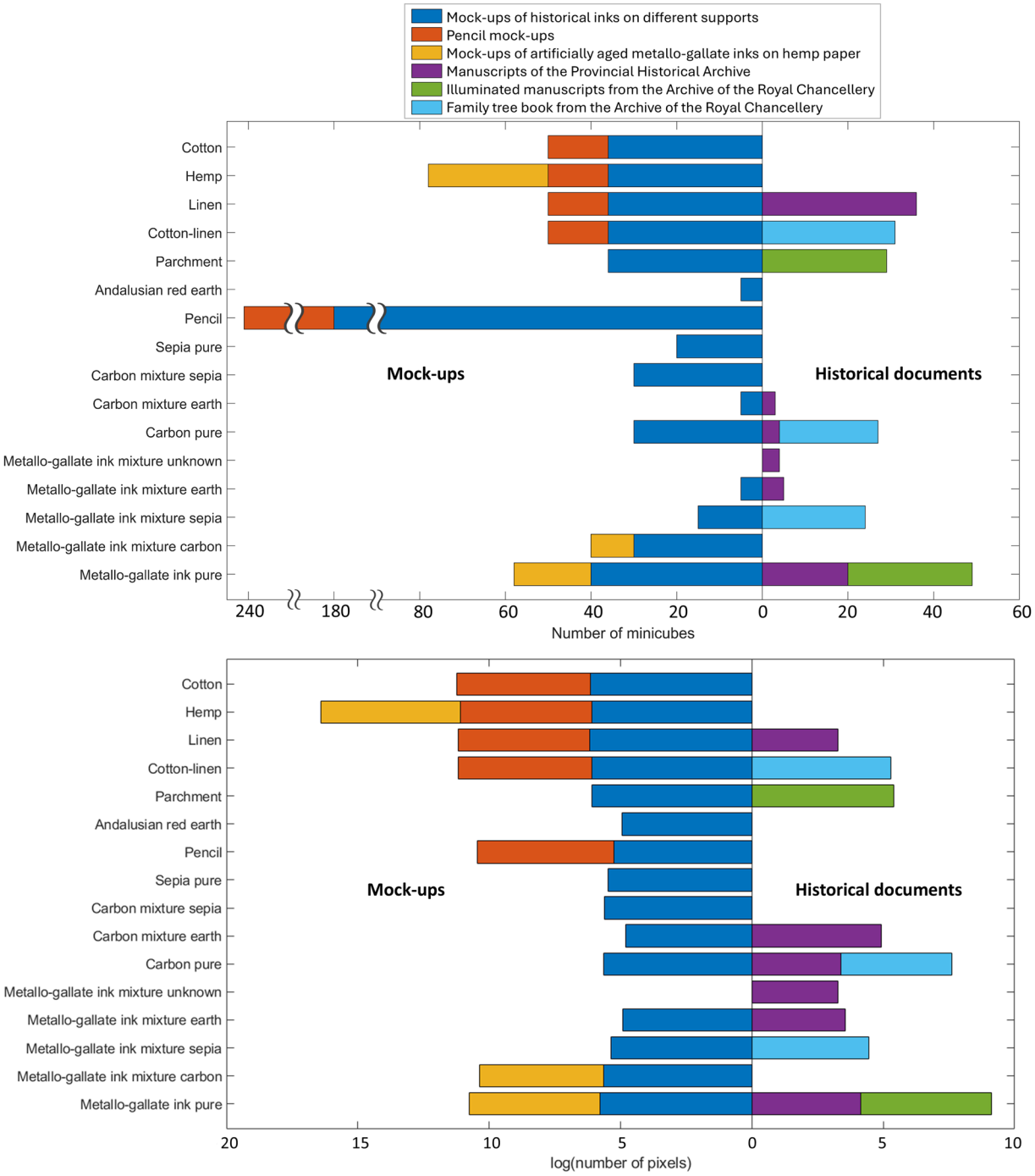

Fig. 5 Horizontal bar graphs showing the distribution of data: number of minicubes (top) and number of pixels on a logarithmic scale (bottom) by class and subset.

Principal Component Analysis (PCA). PCA is a widely used dimensionality reduction technique in spectral data analysis, often employed for assessing the separability of datasets. For each minicube, the average spectrum was calculated for each class, and PCA was performed using these averaged spectra. Two principal components (PCs) were selected based on the Variance Accounted For (VAF) metric, identifying the inflection point where the VAF versus PCs curve flattens. PC1 explains 85.0% of the total variance, and PC2 explains 11.0%, resulting in a combined VAF of 96% with just two components. Fig. 9 presents the score plots for PCs 1 and 2, showing inks and pencil in the left plot and supports in the right plot. Each class is represented by a unique color, and subsets are distinguished by different symbols.

In the left plot, the point clouds for pure metallo-gallate inks (dark blue) and their mixtures with earth (black) overlap, indicating that their spectra are highly similar. Similarly, pure carbon-based inks (pink) cluster closely with their mixtures with metallo-gallate ink (red), sepia (purple), or earth (yellow). These two groups,

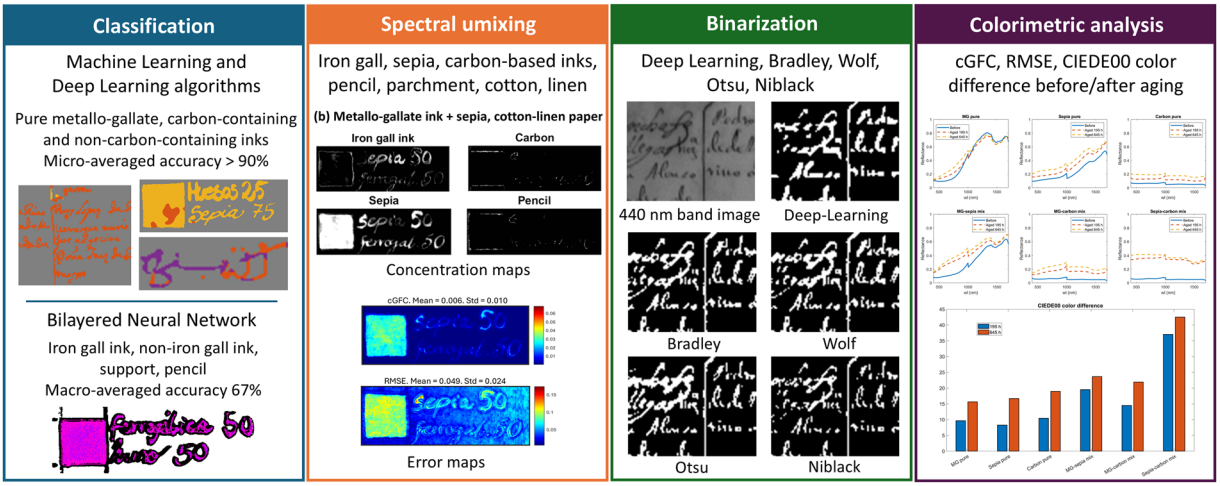

Fig. 6 Schematic representation of the results from previous studies utilizing the HYPERDOC dataset for classification[36,37], binarization[38], spectral unmixing[39], and colorimetric analysis[40,41].

metallo-gallate-based and carbon-based inks, form distinct, separable clusters. However, the clouds for pure sepia (teal) and its mixture with metallo-gallate ink (light green) lie between these two clusters, reflecting intermediate spectral characteristics. Pencil samples, in contrast, form a more or less distinct cluster in a separate region. As for the standard deviation in the mean spectra plot, a high heterogeneity is found in the PCA due to different factors, including differences in spectra between mock-ups and historical documents, the application of the same inks on different supports, variations in ink recipes, and the grouping diverse inks within the same class. For example, the minicubes from the Manuscripts of the Provincial Historical Archive of Granada (depicted as circles) exhibit minimal variability, forming a concentrated cluster. A similar pattern is observed for the artificially aged mock-ups: they form two distinct clusters based on the type of ink and are separated from the non-aged mock-ups. In this case, the support does not influence these clusters, as all inks were applied on hemp paper. However, samples from other subsets and classes are more dispersed, often overlapping in the PCA space.

In the right plot, hemp samples are clearly separated from other supports, forming a distinct cluster. However, notable differences can be observed between mock-ups and historical document samples of the same support types. For instance, cotton-linen samples (blue clouds) and linen samples (red clouds) from mock-ups and historical documents form separate clusters despite belonging to the same class. Parchment samples (green clouds) are distributed across a wider area, also showing a clear distinction between mock-ups and historical documents. In contrast, cotton samples are tightly grouped, reflecting the limited variability within this class, as there were no samples with pure cotton support in the historical documents. Overall, the results of the PCA analysis show that separability among classes is not enough for tackling material identification using PCA components as input. They also suggest that the dataset is wide enough to cover for a fair amount of the variability found in both mock-ups and historical document samples.

Usage Notes

The HYPERDOC dataset[33] presented here is part of the Hyperdoc project (https://colorimaginglab.ugr.es/pages/hyperdoc/project). To facilitate its use, example code, available on GitHub (https://github.com/anabelenlb/HYPERDOC_Database_code), is provided in both MATLAB and Python to perform the following tasks:

- Exploration of the general content of the minicube stored as an HDF5 file, including access to metadata stored as attributes.
- Extraction of the hyperspectral image data from the dataset named `DataCube`, along with relevant attributes as variables in the workspace or environment.
- Visualization of false-color RGB images derived from selected spectral bands.
- Extraction of the Ground Truth (GT) data.
- Calculation of the mean and standard deviation for each class in the GT using the hyperspectral data from the `DataCube`.
- Plotting of the mean reflectances and their standard deviations for both the VNIR and SWIR minicubes.

To execute the MATLAB code, a MATLAB version R2011a or later is required, along with the Image Processing Toolbox. For Python, the following packages are needed: `h5py`, `numpy`, `pillow`, and `matplotlib.pyplot`.

The resulting images based on the provided code are shown in Fig. 10, using the minicubes '00007-VNIR-mock-up.h5' and '00007-SWIR-mock-up.h5'.

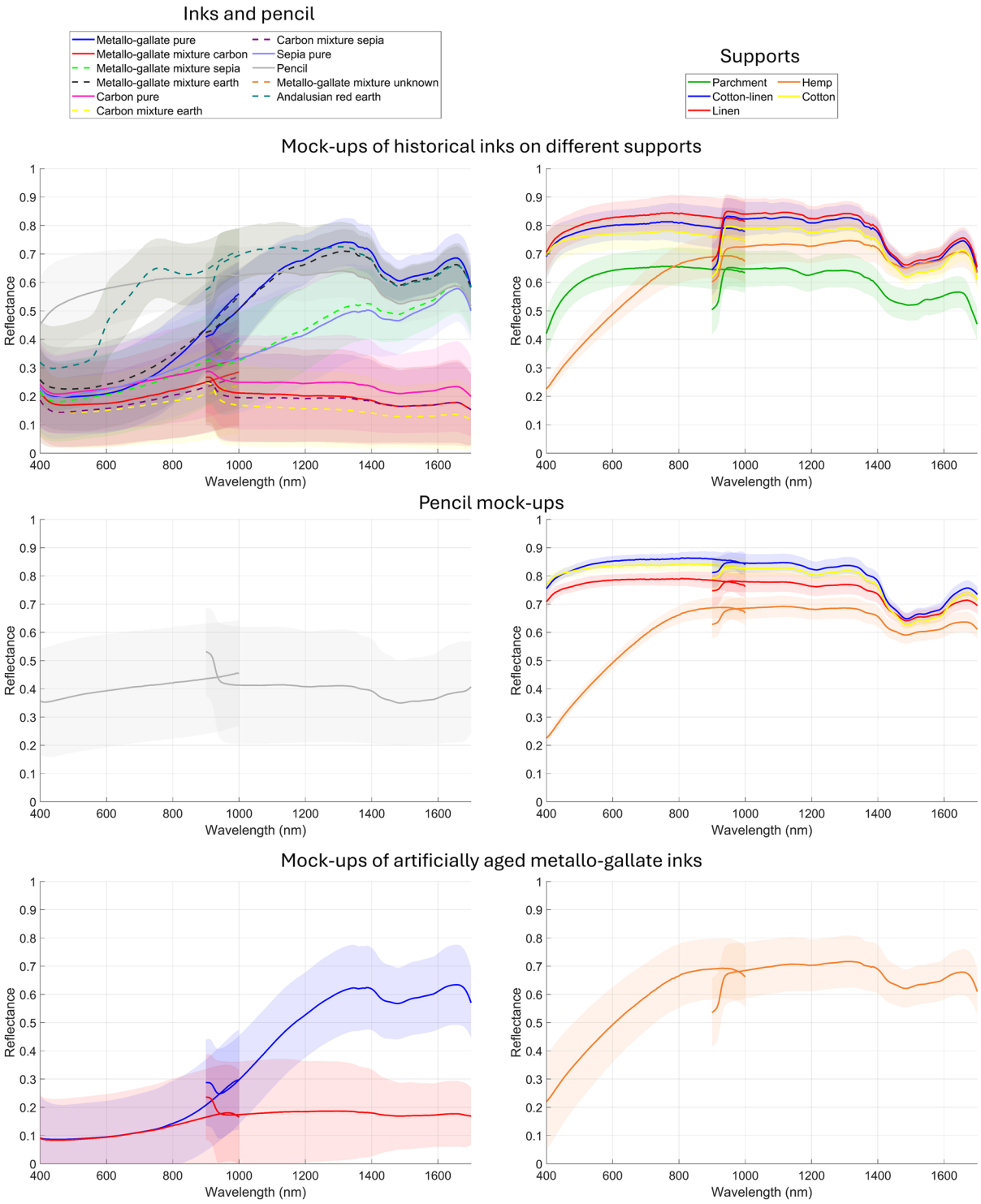

Fig. 7 Average spectra and standard deviation per class (each line in the plot) and subset (each row) in the mock-up set. Inks and pencil are represented in the left graphs, while supports are shown in the right graphs.

Limitations and further considerations for dataset usability. One limitation of the HYPERDOC dataset is the imbalance between the mock-up samples and those derived from historical documents. Mock-ups offer a controlled and reproducible framework ideal for training robust algorithms and investigating specific interactions between inks and supports. Simultaneously, the inclusion of historical samples provides valuable insights into real-world conditions of aging and degradation, broadening the applicability of the dataset to practical scenarios. Efforts have been made to include a broader range of historical materials; however, access to such documents remains a challenge due to their fragile nature and restricted availability. Moreover, the aging processes of inks and supports, along with their interactions and the uncertainty about the recipes used to prepare

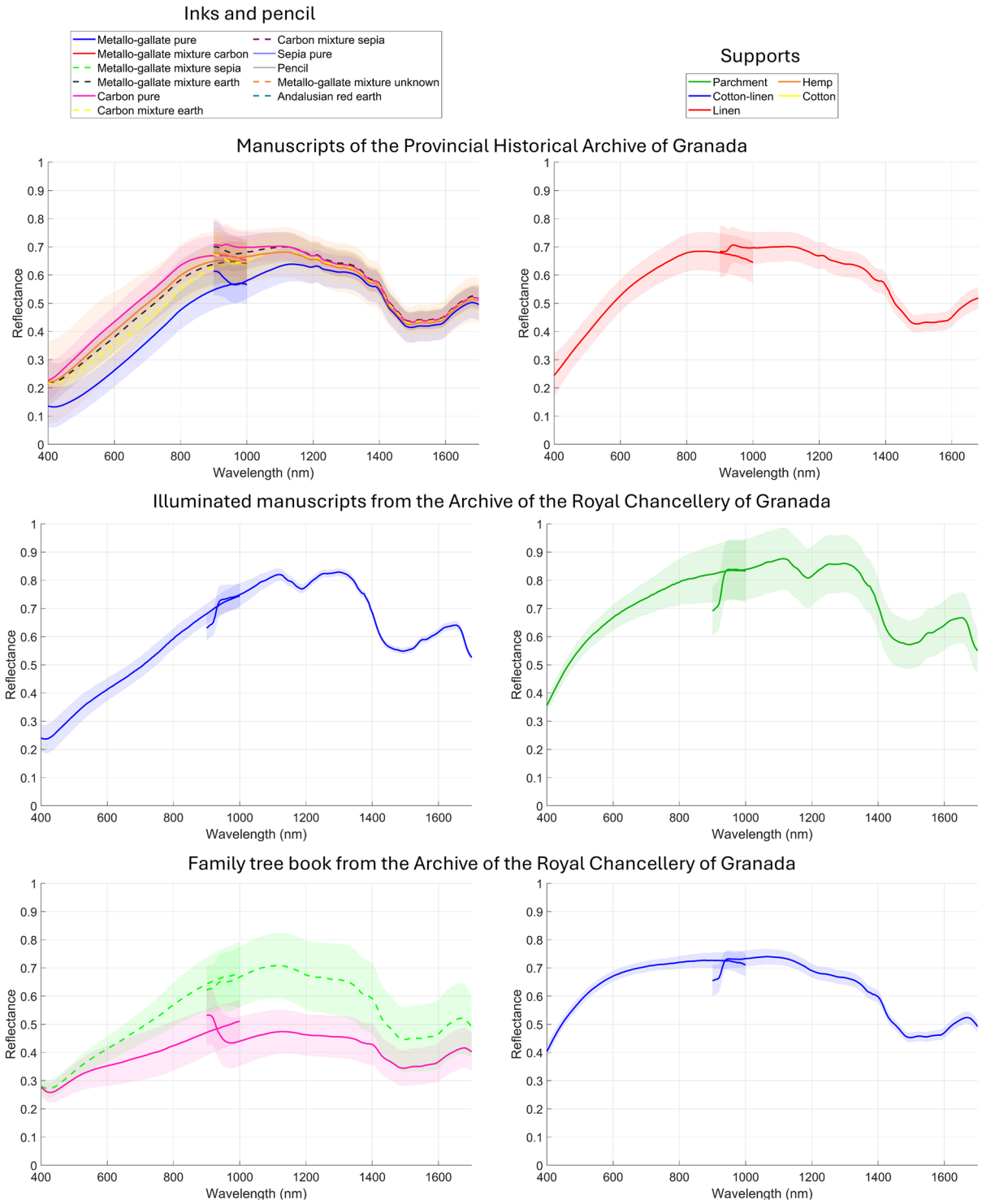

Fig. 8 Average spectra and standard deviation per class (each line in the plot) and subset (each row) in the historical documents set. Inks are represented in the left graphs, while supports are shown in the right graphs.

inks and supports in historical documents introduce further variability in the spectral properties of these materials, which are challenging to model. These processes may result in spectral features that diverge from those of mock-ups, potentially reducing the dataset's generalizability to other historical samples of different periods and origins.

Additionally, identifying the materials present in historical samples often requires complementary analytical techniques. Without these methods, it is difficult to precisely determine the composition of certain inks, pigments, or supports, which can affect the accuracy of the ground truth (GT) annotations.

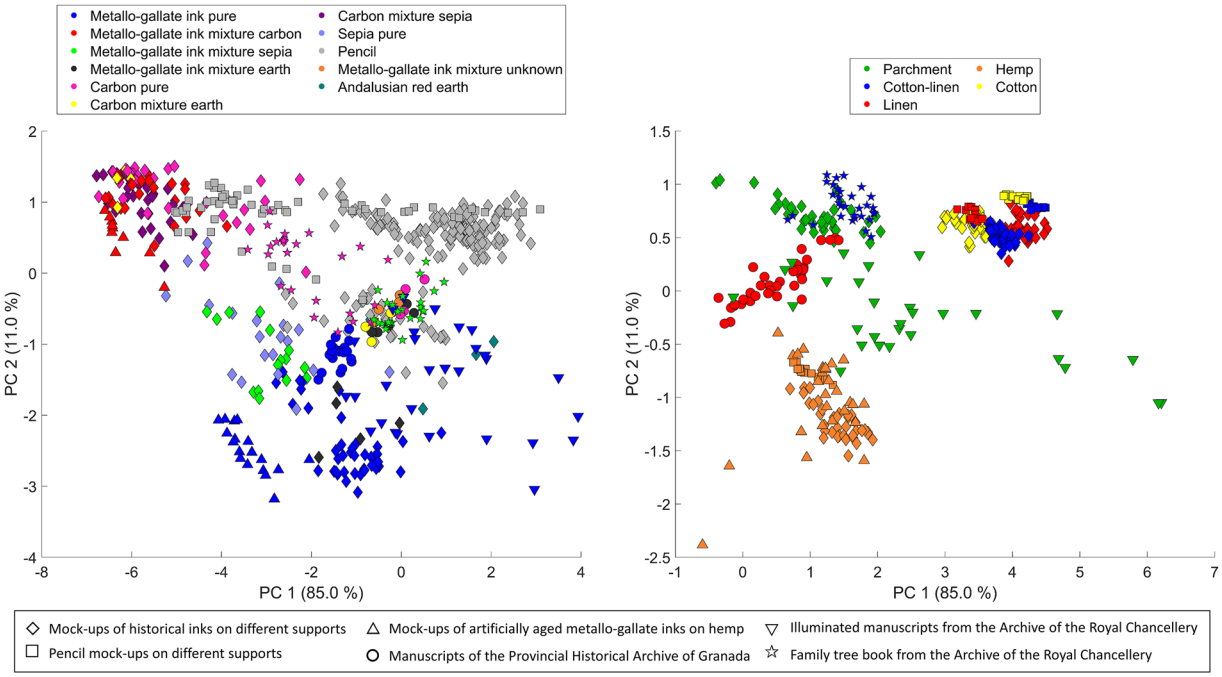

Fig. 9 Score plots of PCA for inks and pencil (left) and supports (right). Colors indicate different classes, and symbols denote different subsets.

Fig. 10 False-color RGB images (left) created using bands [645, 565, 440] nm (top) and [1700, 1300, 1100] nm (bottom), Ground Truth (GT) image (center), and the mean spectral reflectances with standard deviation for the labels present in the GT, extracted from the minicubes '00007-VNIR-mock-up.h5' and '00007-SWIR-mock-up.h5' (right).

The GT annotations in this dataset were generated using a semi-automatic approach, including manual corrections and verification by different annotators, which ensures a reasonable degree of accuracy. However, human involvement may still introduce occasional inconsistencies, particularly when defining boundaries between inks and supports. Interactions between inks and supports, such as penetration or blending at the interface, create transitional zones where both materials are mixed, complicating precise annotation. To address this challenge, algorithms could be employed to erode or expand the annotated ink regions in the GT.

Despite its limitations, the HYPERDOC dataset is a unique and comprehensive resource that integrates mock-ups and historical documents, supporting diverse applications as previously demonstrated. It bridges the gap between the scientific and conservation communities, promoting the adoption of advanced techniques such as hyperspectral imaging, which remain relatively novel in the field. By fostering interdisciplinary collaboration and enabling the development of innovative methodologies, HYPERDOC contributes to the analysis of historical documents, ensuring its ongoing relevance and advancing the safeguarding of our cultural heritage.

Code Availability

Hyperspectral data capture and reflectance correction were performed using Spectronon Pro 3.5.5. MATLAB code for converting from BIL to MAT format, from MAT to HDF5, and for extracting Ground Truth data, along with additional MATLAB and Python code for the visualization and analysis of minicubes, is available on GitHub: https://github.com/anabelenlb/HYPERDOC_Database_code.

Received: 31 January 2025; Accepted: 9 July 2025;
Published online: 16 July 2025

References

1. Nikolaidou, K., Seuret, M., Mokayed, H. & Liwicki, M. A survey of historical document image datasets. *International Journal on Document Analysis and Recognition (IJDAR)* **25**, 305–338 (2022).
2. Granell, E. *et al*. Processing a large collection of historical tabular images. *Pattern Recognition Letters* **170**, 9–16 (2023).
3. Pratikakis, I. *et al*. Competition on document image binarization (DIBCO 2019). In *2019 15th International Conference on Document Analysis and Recognition*, 1547–1556 (IEEE, 2019).
4. Pratikakis, I., Zagoris, K., Barlas, G. & Gatos, B. Icfhr2018 competition on handwritten document image binarization contest (H-DIBCO 2018). In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 1–5 (IEEE, 2018).
5. Serrano, N., Castro, F. & Juan, A. The RODRIGO database. In *LREC*, 19–21 (2010).
6. Alaei, A., Nagabhushan, P. & Pal, U. A new dataset of persian handwritten documents and its segmentation. In *2011 7th Iranian conference on machine vision and image processing*, 1–5 (IEEE, 2011).
7. Abdelhaleem, A. *et al*. Wahd: a database for writer identification of arabic historical documents. In *2017 1st International workshop on arabic script analysis and recognition (ASAR)*, 64–68 (IEEE, 2017).
8. Shahkolaei, A., Beghdadi, A., Al-Máadeed, S. & Cheriet, M. MHDID: a multi-distortion historical document image database. In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, 156–160 (IEEE, 2018).
9. Goetz, A. F. Three decades of hyperspectral remote sensing of the earth: A personal view. *Remote sensing of environment* **113**, S5–S16 (2009).
10. Wei, C.-a, Li, J. & Liu, S. Applications of visible spectral imaging technology for pigment identification of colored relics. *Heritage Science* **12**, 321 (2024).
11. Picollo, M., Cucci, C., Casini, A. & Stefani, L. Hyper-spectral imaging technique in the cultural heritage field: New possible scenarios. *Sensors* **20**, 2843 (2020).
12. Cucci, C. & Casini, A. Hyperspectral imaging for artworks investigation. In *Data handling in science and technology*, vol. 32, 583–604 (Elsevier, 2019).
13. Ciortan, I., Deborah, H., George, S. & Hardeberg, J. Y. Color and hyperspectral image segmentation for historical documents. In *2015 Digital Heritage*, vol. 1, 199–206 (IEEE, 2015).
14. Lettner, M. & Sablatnig, R. Higher order mrf for foreground-background separation in multi-spectral images of historical manuscripts. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, 317–324 (2010).
15. Salehani, Y. E., Arabnejad, E., Rahiche, A., Bakhta, A. & Cheriet, M. Msdb-nmf: Multispectral document image binarization framework via non-negative matrix factorization approach. *IEEE Transactions on Image Processing* **29**, 9099–9112 (2020).
16. Hedjam, R. & Cheriet, M. Novel data representation for text extraction from multispectral historical document images. In *2011 International Conference on Document Analysis and Recognition*, 172–176 (IEEE, 2011).
17. Hedjam, R., Nafchi, H. Z., Moghaddam, R. F., Kalacska, M. & Cheriet, M. Icdar 2015 contest on multispectral text extraction (ms-tex 2015). In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 1181–1185 (IEEE, 2015).
18. Mitianoudis, N. & Papamarkos, N. Multi-spectral document image binarization using image fusion and background subtraction techniques. In *2014 IEEE International Conference on Image Processing (ICIP)*, 5172–5176 (IEEE, 2014).
19. Diem, M., Hollaus, F. & Sablatnig, R. Msio: Multispectral document image binarization. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 84–89 (IEEE, 2016).
20. Moghaddam, R. F. & Cheriet, M. A multi-scale framework for adaptive binarization of degraded document images. *Pattern Recognition* **43**, 2186–2198 (2010).
21. Hollaus, F., Diem, M. & Sablatnig, R. Binarization of multispectral document images. In *Computer Analysis of Images and Patterns: 16th International Conference, CAIP 2015, Valletta, Malta, September 2-4, 2015, Proceedings, Part II 16*, 109–120 (Springer, 2015).
22. Khan, Z., Shafait, F. & Mian, A. Automatic ink mismatch detection for forensic document analysis. *Pattern Recognition* **48**, 3615–3626 (2015).
23. Islam, A. U., Khan, M. J., Asad, M., Khan, H. A. & Khurshid, K. ivision hhid: Handwritten hyperspectral images dataset for benchmarking hyperspectral imaging-based document forensic analysis. *Data in Brief* **41**, 107964 (2022).
24. Morales, A., Ferrer, M. A., Diaz-Cabrera, M., Carmona, C. & Thomas, G. L. The use of hyperspectral analysis for ink identification in handwritten documents. In *2014 International Carnahan Conference on Security Technology (ICCST)*, 1–5 (IEEE, 2014).
25. López-Baldomero, A. B. *et al*. Selection of optimal spectral metrics for classification of inks in historical documents using hyperspectral imaging data. In *Optics for Arts, Architecture, and Archaeology (O3A) IX*, vol. 12620, 99–111 (SPIE, 2023).
26. Grabowski, B., Masarczyk, W., Głomb, P. & Mendys, A. Automatic pigment identification from hyperspectral data. *Journal of Cultural Heritage* **31**, 1–12 (2018).
27. Cucci, C., Delaney, J. K. & Picollo, M. Reflectance hyperspectral imaging for investigation of works of art: old master paintings and illuminated manuscripts. *Accounts of chemical research* **49**, 2070–2079 (2016).
28. Omayio, E. O., Indu, S. & Panda, J. Historical manuscript dating: traditional and current trends. *Multimedia Tools and Applications* **81**, 31573–31602 (2022).
29. Melo, M. J. *et al*. Iron-gall inks: a review of their degradation mechanisms and conservation treatments. *Heritage Science* **10**, 145 (2022).
30. González-García, S., López-Montes, A. & Espejo-Arias, T. The use of writing inks in 12th–19th century arabic manuscripts: A study of the collection of the school of arabic studies, granada (spain). In *Science, Technology and Cultural Heritage*, 121–126 (CRC Press, 2014).
31. Deborah, H. Hyperspectral pigment dataset. In *2022 12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 1–5 (IEEE, 2022).
32. Lombardi, F. & Marinai, S. Deep learning for historical document analysis and recognition—a survey. Journal of Imaging 6, 110 (2020).
33. López-Baldomero, A. B. *et al*. Hyperspectral dataset of historical documents and mock-ups from 400 to 1700 nm (HYPERDOC). *Figshare*. https://doi.org/10.6084/m9.figshare.28319165 (2025).
34. Goltz, D. M. A review of instrumental approaches for studying historical inks. *Analytical letters* **45**, 314–329 (2012).
35. Zamorano, G. M. C. The presence of iron in inks used in valencian manuscripts from the 13th to 17th century. *Microchemical Journal* **143**, 484–492 (2018).
36. López-Baldomero, A. B., Buzzelli, M., Moronta-Montero, F., Martínez-Domingo, M. Á. & Valero, E. M. Ink classification in historical documents using hyperspectral imaging and machine learning methods. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **335**, 125916 (2025).
37. López-Baldomero, A. B. *et al*. Hyperspectral database of synthetic historical inks. In *Archiving Conference*, vol. 21, 11–16 (Society for Imaging Science and Technology, 2024).
38. Moronta-Montero, F. *et al*. Evaluation of binarization methods for hyperspectral samples of 16th and 17th century family trees. In *Archiving Conference*, vol. 21, 94–100 (Society for Imaging Science and Technology, 2024).

39. López-Baldomero, A. B., Valero, E. M., Martínez-Domingo, M. A., Reichert, A. S. & López-Montes, A. Spectral unmixing to identify historical inks (2024). Colour and Visual Computing Symposium (CVCS), Gjøvik.
40. Reichert, A. S., López-Baldomero, A. B., Collado-Montero, F. J. & López-Montes, A. Study of traditional writing inks: creation of standard samples and colorimetric analysis. In *TechnoHeritage2024. Book of Abstracts*, 118 (Universidade de Vigo, 2024).
41. Valero, E. M., López-Baldomero, A. B., Reichert, A. S. & López-Montes, A. Hyperspectral insights into iron gall ink aging. In *TechnoHeritage2024. Book of Abstracts*, 182 (Universidade de Vigo, 2024).
42. Díaz Hidalgo, R. J. *et al*. The making of black inks in an Arabic treatise by al-qalalūs\T1\ī dated from the 13th c.: reproduction and characterisation of iron-gall ink recipes. *Heritage Science* **11**, 7 (2023).
43. Nehring, G., Bonnerot, O., Gerhardt, M., Krutzsch, M. & Rabin, I. Looking for the missing link in the evolution of black inks. *Archaeological and Anthropological Sciences* **13**, 1–10 (2021).
44. Contreras Zamorano, G. M.*La tinta de escritura en los manuscritos de archivo valencianos, 1250-1600. Análisis, identificación de componentes y valoración de su estado de conservación*. Ph.D. thesis, Universitat de València (2015).
45. Mitchell, C. A. Inks. Lecture I. *Journal of the Royal Society of Arts* **70**, 647–660 (1922).
46. Colini, C. *et al*. The quest for the mixed inks. *manuscript cultures* **2018**, 41–48 (2018).
47. Rabin, I. Material studies of historic inks: transition from carbon to iron-gall inks. In *Traces of ink*, 70–78 (Brill, 2021).
48. Conn, D. & Walus, D. Paper objects and books. *The Preservation Management Handbook: A 21st-Century Guide for Libraries, Archives, and Museums* 147–181 (2014).
49. Espejo, T., Lazarova, I. & Cano, M. La colección de manuscritos árabes del archivo histórico provincial de granada. primeros apuntes sobre su caracterización. In *VIII Congreso Nacional de Historia del Papel en España: Actas*, 33–44 (2008).
50. Espejo, T., Lazarova Stoytcheva, I., Campillo García, D., Durán Benito, A. & Jiménez de Haro, A. Caracterización material y proceso de conservación de la colección de documentos árabes manuscritos del Archivo Histórico Provincial de Granada. *Al-Qantara* **32**, 519–532 (2011).
51. de Guevara, M. L. *et al*. "Pleitos de hidalguía. Extracto de sus expedientes que se conservan en el Archivo de la Real Chancillería de Granada correspondiente a la primera parte del reinado de Felipe II (1556-1570)": en cuatro volúmenes. *Hidalgos: la revista de la Real Asociación de Hidalgos de España* 94–95 (2021).
52. Duran, A., López-Montes, A., Castaing, J. & Espejo, T. Analysis of a royal 15th century illuminated parchment using a portable XRF–XRD system and micro-invasive techniques. *Journal of archaeological science* **45**, 52–58 (2014).
53. Bay, H., Tuytelaars, T. & Van Gool, L. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, 404–417 (Springer, 2006).
54. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**, 600–612 (2004).
55. Lee, T.-C., Kashyap, R. L. & Chu, C.-N. Building skeleton models via 3-D medial surface axis thinning algorithms. *CVGIP: graphical models and image processing* **56**, 462–478 (1994).
56. Ntirogiannis, K., Gatos, B. & Pratikakis, I. An objective evaluation methodology for document image binarization techniques. In *2008 The Eighth IAPR International Workshop on Document Analysis Systems*, 217–224 (IEEE, 2008).
57. Otsu, N. *et al*. A threshold selection method from gray-level histograms. *Automatica* **11**, 23–27 (1975).
58. Niblack, W.*An introduction to digital image processing* (Strandberg Publishing Company, 1985).
59. Wolf, C., Jolion, J.-M. & Chassaing, F. Text localization, enhancement and binarization in multimedia documents. In *2002 International Conference on Pattern Recognition*, vol. 2, 1037–1040 (IEEE, 2002).
60. Bradley, D. & Roth, G. Adaptive thresholding using the integral image. *Journal of graphics tools* **12**, 13–21 (2007).
61. Grillini, F. *et al*. Relationship between reflectance and degree of polarization in the VNIR-SWIR: A case study on art paintings with polarimetric reflectance imaging spectroscopy. *Plos one* **19**, e0303018 (2024).
62. Grillini, F., Thomas, J.-B. & George, S. Logistic splicing correction for VNIR–SWIR reflectance imaging spectroscopy. *Optics Letters* **48**, 403–406 (2023).
63. Cosentino, A. FORS spectral database of historical pigments in different binders. *E Conserv. J* **2**, 54–65 (2014).

Acknowledgements

This work was supported by Grant PID2021-124446NB-I00 funded by MICIU/AEI /10.13039/501100011033 and by ERDF, EU; Ministry of Universities (Spain) [grant number FPU2020-05532]; and "ESF Investing in your future" [grant number PRE2022-101352]. We would like to acknowledge David Torres Ibáñez, Director of the Archive of the Royal Chancellery of Granada, and Eva Martín López, Director of the Provincial Historical Archive of Granada, for their assistance.

Author contributions

A.B.L.B.: data collection, data annotation, preparation, structuring, creating the figures, writing the manuscript; J.L.N.: data annotation, data processing, critical revision of the manuscript. F.M.M.: data annotation, data collection, samples set selection; M.A.M.D.: data collection, data annotation, structuring; R.F.G.: data annotation; J.H.A.: data annotation; A.S.R.: preparation, sample set selection; A.L.M.: preparation, sample set selection, critical revision of the manuscript; T.E.: conceptualization, sample set selection; J.R.: conceptualization, critical revision of the manuscript; E.M.V.: management, data collection, sample set selection, data annotation, critical revision of the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.B.L.-B.