Multimedia Systems (2023) 29:847–869
https://doi.org/10.1007/s00530-022-01025-2

REGULAR PAPER

# Multimodal metadata assignment for cultural heritage artifacts

**Luis Rei[1,2] · Dunja Mladenic[1] · Mareike Dorozynski[3] · Franz Rottensteiner[3] · Thomas Schleider[4] · Raphaël Troncy[4] · Jorge Sebastián Lozano[5] · Mar Gaitán Salvatella[5]**

Received: 30 May 2022 / Accepted: 11 November 2022 / Published online: 21 November 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

**Abstract**
We develop a multimodal classifier for the cultural heritage domain using a late fusion approach and introduce a novel dataset. The three modalities are Image, Text, and Tabular data. We based the image classifier on a ResNet convolutional neural network architecture and the text classifier on a multilingual transformer architecture (XML-Roberta). Both are trained as multitask classifiers. Tabular data and late fusion are handled by Gradient Tree Boosting. We also show how we leveraged a specific data model and taxonomy in a Knowledge Graph to create the dataset and to store classification results.

**Keywords** Cultural heritage · Multimodal · Deep learning · Text classification · Multilingual · Image classification · Transformer · Convolutional neural networks

## 1 Introduction

### 1.1 Motivation

Some domains within cultural heritage domains deal with knowledge that is not broadly known by the public, but only by domain experts. Despite many objects having been digitized, even those experts still struggle to find in online catalogs what they are looking for. Thus, they are forced to return to the cumbersome manual consultation of published catalogs, or even card files. If such is the situation for experts, the broader public is still more removed from access to that information. The European production of silk fabrics is an example of one such domain. It is witness to an essential field of European and global history, linked to world trade routes, the production of luxury goods of enormous symbolic importance, technological developments and the very advent of the Industrial Revolution. However, the material vulnerability of these objects and the institutional fragility of many local heritage organizations has rendered it relatively hidden to the public. As regards the information about that heritage, many descriptions, and images of objects exist within in-house databases that are only available as local files. In other cases, those records are uploaded by many museums across the globe, in siloed repositories and heterogeneous, often incompatible formats. A few of

✉ Luis Rei
luis.rei@ijs.si

Dunja Mladenic
dunja.mladenic@ijs.si

Mareike Dorozynski
dorozynski@ipi.uni-hannover.de

Franz Rottensteiner
rottensteiner@ipi.uni-hannover.de

Thomas Schleider
thomas.schleider@eurecom.fr

Raphaël Troncy
raphael.troncy@eurecom.fr

Jorge Sebastián Lozano
jorge.sebastian@uv.es

Mar Gaitán Salvatella
m.gaisal@uv.es

[1] Department for Artificial Intelligence, Jošef Stefan Institute, Jamova Cesta 39, 1000 Ljubljana, Slovenia

[2] Jošef Stefan Institute International Postgraduate School, Jamova Cesta 39, 1000 Ljubljana, Slovenia

[3] Institute of Photogrammetry and GeoInformation, Leibniz University Hannover, Nienburger Straße 1, D-30167 Hannover, Germany

[4] EURECOM, Campus SophiaTech, 450 Route des Chappes, 06410 Biot, France

[5] Departamento de Historia del Arte, Universitat de València, Spain, Av. Blasco Ibáñez 28, 46010 Valencia, Spain

them give public access to the images and metadata of such silk objects through APIs, many more through their websites, but harmonization and integration efforts have been very scarce. Therefore, it is very hard for general audiences, historical experts, and industry (e.g., fashion designers) to access this knowledge.

One possible application for a cultural heritage domain such as European silk fabrics are Exploratory search engines, which help users to explore a topic of interest [46]. They enable serendipitous discovery of items, and they are especially appropriate when these items come with rich structured metadata. ADASilk [1], named after Ada Lovelace, is such an exploratory search engine, based on a knowledge graph (KG), that enables to search and browse silk fabrics objects for both domain experts and users not necessarily familiar with this topic. Thus, not only historians or scholars, but also designers or simply fans of fashion can access such a significant and little-known part of our heritage.

Some records have essential information, like the production year or the weaving techniques, semantically annotated, others include it only in rich textual descriptions, and for some objects it is not available at all. These missing metadata can be considered as gaps that potentially could be filled in. Thanks to the progress in natural language processing, information extraction, and image processing, there are now techniques that can help to address such problems. Digitization of culturally significant assets is a time-consuming process that requires experts and funding. This often forces a cultural institution to make a trade-off between the number of objects digitized and the effort per object. Less effort per object often implies a smaller number of details captured, less strict guidelines, and sometimes mistakes. Nevertheless, this area could benefit from automated aids for collection caretakers, that often must catalog similar or identical objects scattered across the world. Obtaining predictions or suggestions for their description and possible matching pieces would be a great help for that task, taking also into account the many objects still waiting to be properly cataloged.

This paper presents methods that enable further annotation of these museum objects through a multimodal classification approach that trains models to predict such missing metadata from images, text descriptions and other (available) metadata. The outcome is then further used to enrich an underlying knowledge graph that feeds the ADASilk exploratory search engine. Domain experts can easily assess the quality of the automatically generated annotations through rich visualization and connections between the items.

### 1.2 Hypothesis

Our first hypothesis is that we can predict, fairly accurately, a set of domain-relevant properties of cultural heritage objects (silk fabrics) from images and text descriptions. Our second hypothesis is that a multimodal approach involving both images, text descriptions and additional knowledge about other properties than those to be predicted will produce better results than any method relying on a single modality. In this context, the term "better" refers to both, the quality of the results and the number of objects for which this information is inferred. That is, we expect the multimodal approach to result in more correct predictions and in predictions for a larger number of objects than the single modality methods. These hypotheses will be evaluated in the context of digitized metadata of silk fabric artifacts with data originating in multiple museums.

### 1.3 Contributions

The main scientific contributions of this paper are related to our research hypotheses. We introduce a multimodal machine learning approach, adapted to the cultural heritage domain, for predicting properties of digitized artifacts. We perform an in-depth analysis of the performance of our classification models, i.e., models based on individual modalities and the multimodal classifier. Additionally, we introduce a novel dataset[2] to the cultural heritage and multimodal analysis domains that includes data for four different tasks and three different modalities. It consists of harmonized text and image data from heterogeneous, multilingual sources that went through different stages of preprocessing, cleaning, and enrichment like domain expert-guided entity linking and grouping.

Finally, we show how our metadata predictions can be properly represented through classes and properties in our data model, which includes using information about a.o. their time stamp and or the used algorithm, and consequently integrated into existing Knowledge Graphs.

### 1.4 Challenges

The challenges faced in this work can be split broadly into those pertaining to the creation of the dataset and those related to the automated annotation. The latter ones can be further categorized according to the modality that is used for predicting the properties of the objects.

[1] https://ada.silknow.org/.

[2] https://zenodo.org/record/6590957.

#### 1.4.1 Data and labels

The data used in this work belongs to the cultural heritage domain. More specifically, it is related to silk textiles produced in Europe, primarily in the period between the 15th and the 19th centuries. In the domain of cultural heritage, we cannot expect all class labels to be equally likely or equally correlated. For example, in some locations, more silk fabric objects were produced than in others. Similarly, we know that the production of silk fabric objects in a given location likely started after a certain point in time and possibly subsided after a certain date. We also know that catalogs are curated by humans and often have strong thematic biases. For example, certain museums focus almost exclusively on objects created within one location. The data we use in this work was aggregated from different sources. That is, it was crawled from 12 different museum or collection websites. Each museum may have different standards for how it collected the underlying objects and how it digitized the information related to these objects. Importantly, this gives each museum its own standards for how to write text descriptions, how to create images, and how to annotate properties. Regarding these properties of digitized artifacts, accurately representing them requires adequate data modelling capabilities and considerable domain expert collaboration. This collaboration is also important in creating a dataset for machine learning. Labels need to be mapped from annotations made in different languages and grouped into domain relevant classes. Due to the partially automated nature required to create the dataset, challenges arise that are common in such processes: label text requires normalization such as correcting typos, unifying the styles of dates, and matching different locations to specific countries. Errors made in this process can often be systematic, for example, a failure to link a specific value of a property due to the form of writing it particular to that catalog will likely result in that value not being present in all records originating in that catalog.

#### 1.4.2 Image classification

In the context of this paper, the classification of images aims to predict abstract properties of the silk fabrics depicted in the images. Whereas it may be relatively straightforward to learn to classify the material of a depicted piece of fabric, the prediction of semantic information such as the production place of the fabric, the period of time in which the fabric was produced or the technique used to manufacture the fabric is assumed to be much more challenging. Furthermore, it is assumed that there are interdependencies between the these properties of silk fabrics, e.g. a certain production technique may only have been used in a certain period of time. This is why multi-task learning is investigated for image classification. However, standard multi-task classification frameworks require one reference label for every task to be learned during training for every training sample; The challenge we have to face is that in real world data, as they were collected for the dataset presented in this paper, there may be many training samples for which annotations are unavailable for some of the target variables to be predicted. Accordingly, this fact must be taken into account in the training of a multi-task classifier. Additionally, the available number of class labels constituting the class distribution of a variable is often imbalanced for real-world datasets. This constitutes a further challenge to supervised learning, which is addressed by utilizing a suitable training strategy for the image classification method.

#### 1.4.3 Text classification

Supervised approaches are often challenging to perform with data from the cultural heritage domain for several reasons. Text descriptions are not present for the majority of objects in an archive. Many of the text descriptions that are available, in most museums, tend to be short sentences, almost title-like. In specific domains, such as the cultural heritage of silk production, many of the terms used in the text are very domain specific. Each museum has their own standard of how and what to write in a text description: some may focus on the history of the objects and write very grammatical paragraph-length descriptions meant to be read by the public, others may focus on the properties of the object and write a single enumerating sentence, and others still, may focus solely on the depictions or visual patterns of an object. Finally, museums are spread geographically, and thus we can expect to deal with multiple languages, making our problem multilingual and cross-lingual. To summarize, we end up with a small collection of domain specific texts, written in different languages, with different content both semantically and syntactically, and wildly varying lengths. These texts are then associated with labels, based on the provided properties of the object. As already discussed, these labels are not all equally likely or correlated, and many of these accidental regularities are likely to interact with the language and the particularities of the text style of the museum.

#### 1.4.4 Multimodal classification

One of the challenges in this work is that we want to integrate predictions made from images and text. Most work done in the literature, is exclusive to depictions or type of object: the image shows a scene or object and the text describes it. In our case, there may be no scene depicted in an object, and we do not consider describing the object beyond certain properties. For example, if we have a fabric that shows a certain pattern, describing the visual shapes of the pattern (e.g., triangles) is not a goal. Rather, we

need to deduce, from the image, properties of how, when, where, and with what the object was made. Similarly, with text descriptions, there may be a good amount of words that describe visual patterns, scenes depicted, and historical facts associated with the object, but the goal is, again, to determine those same intrinsic properties of the object's making. Another challenge that is uncommon is the reduced and variable overlap between images and text descriptions. Not only is our work subject to a comparatively small dataset, restricted by historical reality and difficulties of data collection, but we must also deal with the fact that for most archives of culturally relevant objects, many objects that have been photographed have no corresponding textual description. In fact, we'll see that less than half of all objects have both these modalities. Another challenge, uncommon outside of retrieval scenarios, is that we can have multiple different images, with different angles and focus, per each individual object while it makes no sense to talk about multiple text descriptions per object. Yet another challenge we need to deal with, common to many real world applications but not to research datasets, is that we do not have all properties for all objects. For example, for a given object, we might know what material and techniques were used but not when or where it was made. Finally, our dataset, although drawn from several museums, contains under 30k objects and approximately 11k text descriptions. Effectively making it small compared to general datasets, but not uncommonly so for a dataset in the cultural heritage domain.

## 2 Related work

### 2.1 Cultural heritage domain

Since the development of the web, many Cultural Heritage (CH) organizations provide metadata on their items through some search engine, APIs or aggregators. Unfortunately, there has been little unity in the data formats, which makes data integration a complex task. One solution to this problem is in the case of museum data is the use of Semantic Web technology and more specifically the development of Knowledge Graphs based on ontologies that follow the open CIDOC Conceptual Reference Model (CRM). CIDOC-CRM was developed for this purpose, i.e., to facilitate inter-museum data integration. It provides many relevant classes and properties to represent domain-specific CH objects and is easily extendable. It is the outcome of more than 20 years of development by ICOM's International Committee for Documentation (CIDOC) [19]. CIDOC-CRM can, however, only be considered a starting point for ontologies that deal with museum data, such as in our case. The fact that it can be easily extended makes it, however, easy to add more domain-specific classes and properties as they are needed in projects such as ours.

There are more and more efforts of different CH organizations to integrate their data with Semantic Web technology and building knowledge graphs: CultureSampo is the result of integrating heterogeneous cultural content [28]. The challenges consisted amongst others of converting legacy data into linked data and making it heterogeneous. Getty ULAN was used as structured vocabulary to find connections between two referenced persons, for example. One similar example is ArchOnto [34], which specifically addresses the challenges of CH data from and for national archives. Both can be an inspiration for work such as ours in this paper, but given how fine-grained the vocabularies in Cultural Heritage domains, such as ours, can be, it is still necessary to deal with the languages and domain specific vocabulary differently in each case.

The training data used for the experiment in this paper is fully extracted from the SILKNOW Knowledge Graph that relies on classes and properties defined by CIDOC-CRM and its direct extensions CRMsci (Scientific Observation Model) and CRMdig (Model for provenance metadata). All our data is therefore part of the specific CH domain of "silk fabrics" and accordingly semantically annotated and enriched, for example through linking and normalization of properties, such as used materials and weaving techniques.

### 2.2 Knowledge graphs and culture AI

Knowledge graphs allow the representation of multi-source information about many entities and their relationships to each other. The data stored in a Knowledge graph can then be used for many other tasks, especially when structured knowledge of a specific domain is relevant, e.g. the development of product designs [37]. Other common domain-specific fields are Medicine, Cyber Security and Finance, but Knowledge graphs are also used a lot to aid product development and research for language-based tasks such as question answering systems, recommender systems and information retrieval [23, 64]. A knowledge graph can also help with textual metadata-aided visual pattern extraction and recognition [11], which is very relevant for this paper. Lastly, as we deal not only with images, but also textual metadata, the SemArt project can be considered related: it is a multi-modal dataset for semantic art understanding. Unlike in this study, they did, however, not work towards metadata completion, but focused only on retrieval [24] .

### 2.3 Image classification

Applying and adapting machine learning techniques to support solving tasks in the context of preserving the cultural heritage is a growing field of research. Many works address

image-based classification of artworks by training an image classifier on the basis of images with known class labels to make predictions for images with unknown properties [21]. First works investigate classical machine learning approaches aiming to predict characteristics of a depicted painting [3]. In [7], one-versus-all Support Vector Machines are trained based on HOG features (histograms of oriented gradients, [16]) of images showing paintings, with the goal to predict the artist of the painting.

Instead of training a classifier to predict variables based on handcrafted image features, Convolutional Neural Networks (CNNs) allow for simultaneously learning features from given input images as well as learning a mapping of these features to class scores based on labeled training images [35, 36]. Thus, a trained CNN can be used to predict a class label for an object with unknown properties from an image depicting that object. CNN-based classifiers are also applied in many works addressing attribute prediction for depicted objects in the context of cultural heritage, where the focus is on making predictions for images showing paintings [11, 52]. In [58], the *artist*, the *genre* as well as the *style* of a painting are learned by means of a variant of AlexNet [35], achieving on average 68.3% correctly classified images for the three variables using the WikiArt dataset (http://www.wikiart.org/). Investigating the prediction of a painting's *artist*, Sur and Blaine [57] obtain 82.5% overall accuracy on the Rijksmuseum dataset [44] utilizing a ResNet18 [26]. In both cases, there is one CNN per classification task, and network weights pre-trained on a variant of the ImageNet dataset [17] are used to improve the classification performance.

Instead of training a separate CNN per task to be learned, the concept of multitask learning aims to exploit interdependencies between related tasks by means of jointly learning them in one network and, thus, to improve the network's performance in solving the individual tasks [10]. Multi-task learning for CNNs is addressed in many recent works [15] investigating different strategies for combining the training of several tasks. In the domain of cultural heritage, the most frequently used strategy applies a feature extraction network producing a high-level image representation that is shared among all tasks and which is processed by additional task-specific layers designed to solve the individual classification tasks, e.g., [5, 25, 56]. These works do not only perform multitask learning for predicting characteristics of paintings on the basis of images, they also make use of pre-trained CNNs for the shared feature extraction network.

In contrast to all works cited so far, which are dedicated to the classification of paintings, we address the CNN-based classification of images of silk fabrics. Even though there are papers dealing with the CNN-based classification of images of textiles, e.g. [29, 43, 48, 61], distinguishing different textile patterns, no work could be found addressing the classification of images of fabrics in the context of cultural heritage except for our previous one. The image classification network presented in this paper can be seen as an expansion of [20], aiming to predict different properties of silk fabrics; the network takes images of silk fabrics as input, where a high-level image representation produced by a fine-tuned ResNet [27] is shared among all task-specific classification branches that deliver the predictions. In contrast to [20] as well as [5, 25, 56], we adapt the training of the network weights such that hard training examples get a higher impact on the weight updates. In this way we want to deal with the problem of class imbalance in the training data, aiming to improve the classification performance for underrepresented classes. For that purpose, we combine a variant of the focal loss [38] with the multi-task softmax cross entropy loss used in [20], leading to a training strategy that focuses on hard training examples in a multi-task scenario while allowing for missing class labels at training time for some of the tasks to be learned. Furthermore, we investigate the prediction of four variables instead of three like in [20] and evaluate our methodology on a much larger dataset consisting of images from several museum collections instead of one collection only.

### 2.4 Text classification

Much of the recent work in natural language processing has focused on fine-tuning large transformer neural networks [59] pretrained as language models such as BERT [18] and RoBERTA [41]. The most common approach is to add a task-specific head to the pretrained transformer to create the final model architecture. The full model is then trained on the task specific data. This is the process that is called fine-tuning. On most natural language processing (NLP) tasks, some variation of this approach provides the best results. Previously, many multilingual and cross-lingual artificial intelligence approaches to text used pretrained and aligned word-embeddings, such as the aligned fastText embeddings [8, 30]. But similarly to the overall trend in NLP, recent approaches have also moved towards using fine-tuning of pretrained transformers. Our work follows this trend, we fine-tune the pretrained XLM-R [14]. Multitask models are often very desirable from a practical perspective: a single model is easier to deploy and maintain, offers faster inference and occupies less space in memory when compared to multiple models. Further, multitask learning can often result in measurable improvements [9]. [40] showed that multitask training of BERT improved results across several tasks.

The use of Natural Language Processing, especially text classification, in the cultural heritage domain isn't very widespread. This is a consequence of the fact that the digitization of artifacts usually includes images and some labels or tags, but text descriptions are far less common. The highlight is the work of [51] on text descriptions of paintings

from the Rijksmuseum Amsterdam. They used an Information Extraction approach rather than classification. Their pipeline included Named Entity Recognition, Part-of-Speech tagging and dependency parsing to extract concepts from the text, those concepts were then matched to an ontology and finally classified according to a role. These roles included all the properties we use: Technique, Material, Date, Place, plus others such as "Creator" of the artwork, style, and the subject depicted. Their data, although limited to a total of 250 text descriptions, was manually annotated and each text contained a concept-role pairing. They reported an average $F1$ of 61.2% compared to non-expert human average of 65.1%. Our work differs from this in several key areas. First, our classification approach is more generalizable, as it does not necessarily require information to be directly present in the text and more resilient to misspellings and non-standard grammar. In fact, even correctly linking information known to represent a certain label (e.g. material) from tables can be challenging in the presence of spelling issues. Second, we work with multilingual data from multiple sources, which presents additional challenges. Finally, our dataset contains many more samples and uses automatic labelling based on information present in catalogs rather than externally annotated.

### 2.5 Tabular classification

Gradient Boosted Decision Trees (GBDT) [22] has long been the state of the art and the common choice for handling tabular data. Concurrently with our work, Neural Network based alternatives have been proposed which can outperform GBDT in certain situations [2, 31]. To the best of our knowledge, there has been no work published on tabular classification in the cultural heritage domain that we can provide an overview of.

### 2.6 Multimodal classification in cultural heritage

[4] presented a joint image-text neural network architecture for classifying images of paintings by artist and year. Their text input consisted of a limited set of labels (style, media, and genre) rather than text descriptions. Conceptually, this is similar to CLIP-Art [13], an application of CLIP [49] to the retrieval of artwork images. CLIP learns to associate a small text vocabulary akin to labels with images through joint contrastive pre-training. This was applied to The iMet Collection dataset [63], possibly the most similar dataset to our own, it includes images of artworks associated with labels (also called "tags") that describe what is visually depicted in the object (e.g. "Dragons"), its visible properties (dimensions, medium) as well as other culturally relevant properties (e.g., country of origin). Another very relevant dataset is in this context is Artpedia [55], a dataset of images of paintings associated with textual descriptions tagged as either "visual sentences" that describe the scene depicted in the painting or as "contextual sentences" that describe other aspects of the painting such as its historical context. The tasks for which this dataset was created consist of separating visual from contextual sentences and the retrieval of the correct image for a given text. The differences between the related work and our work are clear. We propose to handle images, multilingual text, and tabular data as equal modalities. We also propose to handle data from multiple collections.

## 3 Data

### 3.1 Knowledge graph

The SILKNOW Knowledge Graph[3] lies at the center of all efforts to create a unified representation of the metadata of European silk textiles, particularly from the 15th to the 19th century. All the data used in our experiments was downloaded from 16 sources, most of them are public online museum records, for which we built crawling and harvesting software. In addition to that, we have data from the SILKNOW[4] project partners Garin and the University of Palermo (Sicily Cultural Heritage). The dataset used in the experiments was created from a full export of all objects in the knowledge graph, which consists of the metadata of 38,873 unique silk objects before any preprocessing steps. This export includes in total 74,527 unique image files.

To model this heterogeneous data from so many sources, we chose and relied strongly on the CIDOC Conceptual Reference Model (CRM). We also developed our own SILKNOW ontology[5] to extend CIDOC-CRM with further classes and properties for cases where it did not cover some specifics of the silk textile domain and also for some extra information. For example, the confidence score for metadata predictions, once we started integrating the results of those predictions back to the KG.

To develop a converter[6] that could unify all the original data with all these classes and properties into one knowledge base, mappings have been created by domain experts. And on a technical level, all museum records had to be harvested and were first converted into a common JSON file format through our crawler software[7] but each array inside this format still had the original field labels from the museums before the final conversion. For example: the majority of

[3] https://zenodo.org/record/5743090.
[4] https://silknow.eu/.
[5] https://ontome.net/namespace/36.
[6] https://github.com/silknow/converter/.
[7] https://github.com/silknow/crawler/.

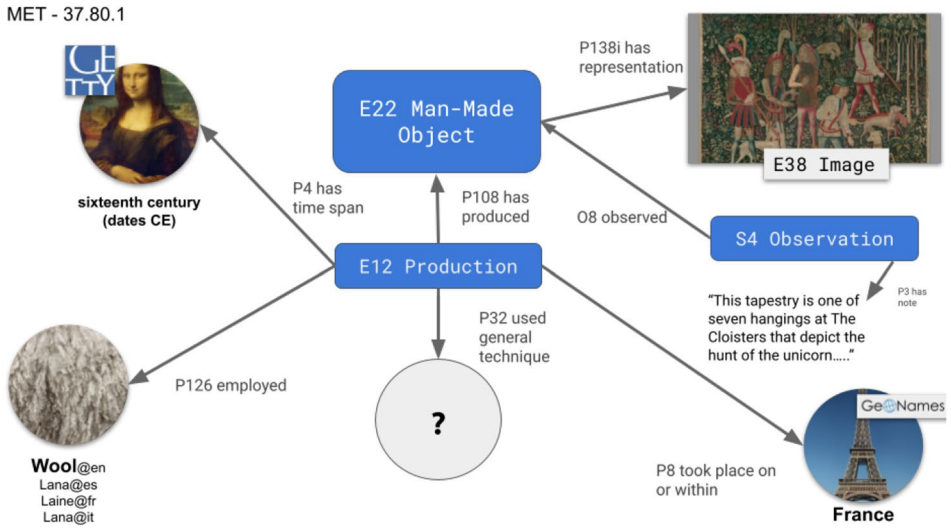


**Fig. 1** A record from the MET museum with a missing property represented in the knowledge graph using our ontology and controlled vocabularies

museums have a field for describing the production time of a silk object, but in most cases museums use different names for their field. Moreover, the museums are from all over the world and we are facing different languages for both the field names and their values. This is why we created a mapping for, e.g, a field named "Date" (Metropolitan Museum of Arts) and the class `E12_Production` with the property `P4_has_time-span` and another class `E52_Time-Span`. Likewise, a mapping rule will be written for the field named "date_text" (API of the Victoria and Albert Museum) and for the (Spanish) field named "Datación" (Red Digital de Colecciones de Museos de España).

Another very central part of our knowledge representation is the SILKNOW Thesaurus[8], a controlled vocabulary which contains many explicit and multilingual concept definitions for materials, techniques, and motif depictions relevant for these silk textiles. Thanks to this thesaurus, a lot of information and entities from very explicit categorical fields of the original museum records could be linked, without any advanced machine learning techniques - the string literal could just be matched with the (multilingual) labels of the thesaurus and then replaced with a unique concept link. This explicit representation of knowledge forms the core of the dataset used to predict missing metadata. This includes cases where a categorical value is either not given at all or "hidden" in longer textual descriptions and not explicitly semantically annotated.

Once all the modelling, download, conversion and enrichment steps were taken, the final knowledge graph was uploaded onto a SPARQL endpoint from where all the data across languages and museums can be queried the same way. To make access easier, we also developed a RESTful API, so it is not necessary for web developers to write SPARQL queries, and an aforementioned exploratory search engine on top of this API, called ADASilk. It is aimed at users with only little technical background or little background knowledge about the domain of silk, to make them able to discover a lot of the data in the KG. ADASilk offers an advanced search with many filters, some topic suggestions, and in general a clean visual interface that shows all objects with their images and metadata.

### 3.2 Extracting and normalizing labels

The development of the SILKNOW Knowledge Graph is a combined effort of data processing that relies on a data modelling and annotation process created in collaboration with domain experts. This is especially true for the SILKNOW Thesaurus. The group labels used in the experiments in this paper are based on the hierarchy and relations of concepts of the silk textile domain described in this controlled vocabulary. As described in Sect. 3.1, a big part of categorical property values could be easily extracted, linked and through the string replacement indirectly automatically normalized thanks to the SILKNOW Thesaurus. This means that many concepts are accessible even though there were originally different strings, including typos in some cases, synonyms, or translations. An example would be a weaving technique like "Damask", which would be "Damas" in French and "Damasco" in Spanish and Italian: for all these, we replace the string literal with one link to the same concept. In addition to the SILKNOW Thesaurus, we also use linked open data like such as GeoNames[9] to normalize and link place names.

[8] https://skosmos.silknow.org/thesaurus/.

[9] https://www.geonames.org/.

Matching strings with such thesaurus or other controlled vocabularies was not without challenges. As will be also explained in more detail in Sect. 6, misspellings or unique punctuation could still cause the matching process not to work properly. To give an example: If the string value of a record was "silk; gold thread" the latter would not have been linked, due to a bug that did not properly consider a semicolon as a separator. Other such cases existed as well, as the development of the SILKNOW Knowledge Graph is an ongoing process and concurrent with this work. See Fig. 1 for an illustration of a museum record in the knowledge graph.

The aforementioned hierarchy defined in the SILKNOW Thesaurus can be used to select specific types or subtypes of properties. To refer back to the previous example, we could select only objects with the weaving technique "Damask", but also only objects made with "Two-coloured damask" which is even more specific. Based on the Thesaurus, we can also make sure that we only choose objects based on equivalent levels of this hierarchy.

Based on these enrichments and the linking process, we created a pipeline to extract the dataset based on pre-specified criteria. We first developed a comprehensive SPARQL query that outputs all museum objects described in the Knowledge Graph (KG) and includes, if available, the most relevant properties: the identifier of the object in the knowledge graph, the museum where the description comes from, the text description, and URL links to the images that illustrate the object. The results of this query were exported as a CSV file, which we then post-processed to make sure that we have a format of one row per object. In this final format, the CSV is used as the basis for all experiments.

### 3.3 Label grouping

In principle, the Knowledge Graph contents can be used to generate training and test samples for the classifiers described in Sect. 4. One would just have to associate the images and/or the text given for a record with the annotations in the categorical variables of interest. The available annotations can be easily converted into class labels. However, a statistical analysis of these annotations revealed that most of them occur very rarely in the data, while for all categorical variables there were one or a few classes which were dominant in the sense that many records belonged to them. Supervised classifiers have problems with imbalanced training data sets, and it would seem very difficult for a classifier to successfully differentiate classes for which it has seen only a very small number of training samples, if on the other hand there are thousands of samples for some other classes. To still be able to extract meaningful information from the available modalities using supervised methods while at the same time having the chance of achieving a reasonably good classification performance, a simplified class structure was defined. Domain experts analyzed the class distributions and aggregated classes corresponding to different categories into compound classes. Care was taken for the aggregated classes to be consistent with the Thesaurus, and aggregated only if they were considered to be related according to the domain experts. At the same time, the aggregation was guided by the frequency of occurrence of class labels so that the compound classes would occur frequently enough to be used for training the supervised classifiers described in Sect. 4.

The resultant simplified class structure was integrated into the Knowledge Graph in the form of so-called *group* fields, which were made available for all semantic properties of interest, principally, the ones corresponding to the different tasks in this work. Such grouping was applied to the following properties: Material, Technique, production place (with a country granularity), production time (with the century granularity) and the object type or object domain group, to be able to filter out non-textiles that use silk. Grouping was not an easy task, domain experts had to deal with more than 200 concepts that had to be grouped according to the aforementioned categories. Techniques were the most complex to group. To do so, domain experts grouped the concepts according to two fundamental criteria: (1) whether they belonged to the same hierarchy, for example, velvet and its types. In fact, there are many types of velvet, classified depending on the nature of the pile such as broderie velvet, ciselè velvet, cut velvet, pile-on-pile velvet, uncut velvet, etc. (2) If they were somehow related to a certain technique, for example, the effects obtained of applying differently warp and weft, that is, whenever a yarn is introduced into a fabric to produce an effect or pattern. On the other hand, materials were not complex as they were made in large groups according to their origin, that means according to the product obtained from the processing of one or more raw materials, in the course of which their structure has been chemically modified, e.g. animal fibres are distinguished from vegetable fibres. Using a conversion table for aggregation prepared by the domain experts, the contents of the *group* fields could be derived automatically from the original semantic annotations. Having thus expanded the Knowledge Graph, training, and test samples could be easily generated from it by appropriate SPARQL queries that would export the contents of the *group* fields associated with each record.

### 3.4 Dataset preparation and properties

The goal of the dataset preparation is the conversion of the knowledge graph data with normalized and grouped labels described, respectively, in Sects. 3.1, 3.2, and 3.3, into a dataset for the experiments in Sect. 6 using the classification methods described in Sect. 4.

**Table 1** Class structure and class distribution of the records

| Variable name | Class name | Total | Training | Validation | Test |
|---|---|---|---|---|---|
| Timespan | 19th century | 5849 | 3492 | 1180 | 1177 |
| | 18th century | 4397 | 2576 | 901 | 920 |
| | 20th century | 2483 | 1520 | 483 | 480 |
| | 17th century | 1134 | 689 | 231 | 214 |
| | 16th century | 880 | 542 | 180 | 158 |
| Place | FR | 5265 | 3156 | 1037 | 1072 |
| | IT | 3205 | 1853 | 687 | 665 |
| | GB | 2837 | 1721 | 562 | 554 |
| | ES | 2630 | 1605 | 521 | 504 |
| | IN | 1190 | 735 | 231 | 224 |
| | CN | 699 | 426 | 127 | 146 |
| | IR | 671 | 409 | 142 | 120 |
| | JP | 533 | 325 | 92 | 116 |
| | TR | 331 | 205 | 57 | 69 |
| Technique | Embroidery | 3123 | 1814 | 657 | 652 |
| | Velvet | 2193 | 1273 | 454 | 466 |
| | Damask | 1685 | 1004 | 333 | 348 |
| | Other technique | 1150 | 722 | 219 | 209 |
| Material | Animal fibre | 17,382 | 10,387 | 3445 | 3550 |
| | Vegetal fibre | 2051 | 1255 | 396 | 400 |
| | Metal thread | 2046 | 1223 | 422 | 401 |

**Table 2** Names of the museums contributing to the dataset with their identifiers (ID) used in this paper, and distribution of the 28,077 records over the museums for the training (train.), validation (val.) and test sets

| Museum name | ID | Total | train. | val. | Test |
|---|---|---|---|---|---|
| Metropolitan Museum of Arts | met | 6524 | 3835 | 1325 | 1364 |
| CDMT Terrassa | imatex | 6119 | 3690 | 1204 | 1225 |
| Victoria and Albert Museum | vam | 5527 | 3300 | 1133 | 1094 |
| Rhode Island school of design | risd | 3226 | 1913 | 634 | 679 |
| Boston Museum of fine arts | mfa | 2610 | 1579 | 517 | 514 |
| Garín 1820 | garin | 1558 | 972 | 300 | 286 |
| Collection du Mobilier National | mobilier | 1293 | 796 | 267 | 241 |
| Red Digital de Colecciones de Museos de España | cer | 781 | 490 | 142 | 149 |
| Joconde Database of French Museum collections | joconde | 375 | 224 | 78 | 73 |
| Smithsonian Museum | smithsonian | 38 | 29 | 14 | 14 |
| Versailles | versailles | 18 | 8 | 5 | 5 |
| Art Institute of Chicago | artic | 8 | 4 | 2 | 2 |

The first step was to select the records in the knowledge graph that were relevant to the domain.

The second step as to select only records that contained a value for one of the variables to be predicted, i.e., labeled samples. Uncommon labels, with a total frequency below 150, were discarded. The final step was to randomly split the records into disjoint sets:

- A training set consisting of 60% of the data for supervised learning;
- A validation (or development) set, consisting of 20% of the data for hyperparameter tuning and multimodal supervised learning;
- A test set, consisting of 20% of the data, for evaluation of the proposed method.

Given that the objective is to train and evaluate a multimodal multitask approach on records, that regularities exist within each collection (i.e., museum) that comprises the data, that the text modality is also multilingual, and that both modalities and task specific labels may be missing from a record, we believe the most reasonable way to split the dataset is a random split of records. The distribution of the data per each set and class label can be seen in Table 1.

**Table 3** Modality statistics of all records in the dataset that provide a class label for at least one of the variables

| Dataset | Total | With image | With text | With image and text | Without images and text |
|---|---|---|---|---|---|
| train. | 16,840 | 16,260 | 6717 | 6495 | 358 |
| val. | 5602 | 5419 | 2184 | 2101 | 100 |
| Test | 5635 | 5441 | 2133 | 2068 | 129 |
| Total | 28,077 | 27,120 | 11,034 | 10,664 | 587 |
| | 100.0% | 96.6% | 39.3% | 38.0% | 2.1% |

The values are given for the training (train.), validation (val.) and test sets as well as for the total dataset

The distribution of samples over the museums can be found in Table 2 and an overview of the modalities can be found in Table 3. We can see how 27,120 or 96.60% of the 28,077 records about annotated fabric objects contain at least one image, but only 11,034 or 39.29% of them contains a text description. The overlap consists of 10,664 or 37.98%. The proportion between training validation and test sets in each case corresponds roughly to the aforementioned 60-20-20 split.

Text data in our dataset consists of descriptions of fabrics or objects made mostly of fabrics. These descriptions range in length from a short sentences to multi-sentence

**Table 4** Examples of text descriptions present in our dataset

| Text description |
|---|
| White and silver striped fabric with supplementary weft of flat silver strips whose floats form vertical stripes with leaves at intervals. White floats of the weft form outlines for serpentine floral sprays spread over the striped areas. |
| Furnishing fabric, woven, British, c. 1895, Alexander Morton & Co., red/brown plain silk weave |
| Dibujo Palma en color azul grisáceo Urdimbre: Trama: 36 pasadas Rapport: 65 cm ancho y 104 cm alto (incompleto) |

**Table 5** Text length in characters and space delimited tokens

| | Min | Q1 | Median | Mean | Q3 | 95th percentile | Max |
|---|---|---|---|---|---|---|---|
| Characters | 60 | 173 | 343 | 693 | 856 | 2367 | 16,333 |
| Tokens | 7 | 28 | 56 | 115 | 142 | 392 | 2826 |

**Table 6** Language distribution of text descriptions based on language of the museum

| | English | Spanish | French | Catalan |
|---|---|---|---|---|
| Records | 7271 | 1975 | 1126 | 680 |

paragraphs to multi-paragraph texts with thousands of words. Some descriptions focus primarily on a single aspect, such as a scene depicted or the history of the object, while others focus on various properties of the object. Table 4 shows some examples of these descriptions. To eliminate some errors present in the data, we removed any text descriptions smaller than 60 characters. The resulting distribution of lengths is summarized in Table 5. These descriptions are in 4 different languages: English, Spanish, French, and Catalan. The counts for each are shown in Table 6.

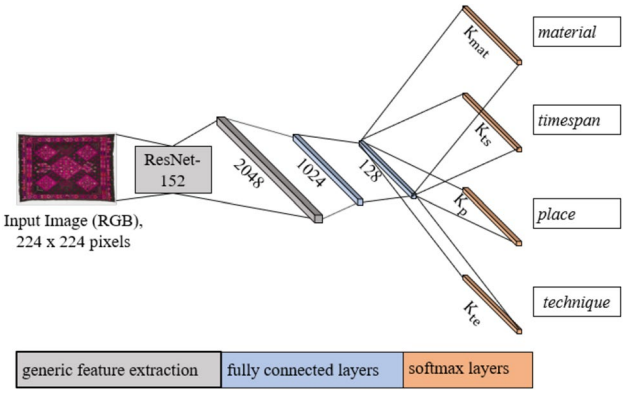


**Fig. 2** Network architecture of the CNN for multitask image classification. The input image scaled to 224 x 224 pixels is presented to a pre-trained ResNet-152 (grey) to extract generic features. The resulting 2048-dimensional feature vector is mapped to a domain-specific joint representation of 128 dimensions by two fully connected layers (blue). The task-specific classification branches consist of one softmax layer each (orange) that delivers the class scores for the corresponding variable. $K_{mat}$, $K_{ts}$, $K_p$, and $K_{te}$ denote the number of class labels for the tasks *material*, *timespan*, *place*, and *technique*, respectively

## 4 Methods

### 4.1 Image classification

The goal of the image classification is to predict one class label per classification task, i.e., the prediction of a class label for each of the target variables *technique*, *timespan*, *material* and *place*, for an image that illustrates an object. For that purpose, an image classifier is trained using all images of all records contributing to the dataset described in Sect. 3.4. We propose to use a convolutional neural networks (CNN) for that purpose, motivated by the success of CNN in image classification. As there are many records with annotations for more than one of these variables, we propose to train the classifier to predict all classes simultaneously in a multitask framework, exploiting the inherent relations between the variables to learn a joint representation that is used by task-specific classification heads. A detailed description of the chosen network architectures can be found in Sect. 4.1.1, whereas the strategies used for training are presented in Sect. 4.1.2.

#### 4.1.1 Network architecture

Figure 2 shows the structure of the CNN for multitask learning for the prediction of the four target variables. Its input consists of an RGB image scaled to a size of 224 x 224 pixels. This image is presented to the ResNet 152 network of [27] pre-trained on ImageNet [17], which serves as a generic feature extractor for the image [53] and produces a feature vector of 2048 dimensions. We apply dropout with a probability of 10% after this layer [54]. This is followed by $L_{fc} = 2$ fully connected layers, the first one having 1024 and the second one having 128 nodes, which are shared by all tasks.

Rectified linear units [45] are used as nonlinearities in both of these joint layers. They produce a joint representation of the image of $N_r = 128$ dimensions. This representation is processed by four task-specific classification branches, each consisting of one additional softmax layer only, which delivers the class scores $y_{km}(\mathbf{x}, \mathbf{w})$ for the input image $\mathbf{x}$ to belong to class $k$ for variable $m$. The number of nodes of the softmax layer corresponds to the number of classes to be differentiated for a specific task. The CNN architecture is shown in Fig. 2.

The CNN predicts one class label per task for every image. In case of multiple images per record, one such class label is predicted for each one of the images and the prediction with the highest softmax score is chosen to be the prediction for the record.

#### 4.1.2 Training

In training, the parameters $\mathbf{w}$ of the CNN described in Sect. 4.1.1 are learned by minimizing a loss function $E(\mathbf{w})$. The parameters of our network consist of the parameters $\mathbf{w}_R$ of ResNet-152, which are initialized from a pre-trained model published by [27], and the parameters $\mathbf{w}_{FC}$ of the fully connected and softmax layers, which are initialized randomly by a variant of the Xavier initialization also described in [27]. In the training procedure, we determine the parameters $\mathbf{w}_{Rt}$ of the last $NL_{RT}$ layers of ResNet-152 considering exclusively entire residual blocks and the parameters $\mathbf{w}_{FC}$ of the fully connected layers, whereas the parameters $\mathbf{w}_{Rf}$ of the first $152 - NL_{RT}$ ResNet-152 layers are frozen [62]. Thus, the parameter vector consists of three subsets: $\mathbf{w} = \left(\mathbf{w}_{Rf}^T, \mathbf{w}_{Rt}^T, \mathbf{w}_{FC}^T\right)^T$. $NL_{RT}$ is a hyperparameter to be tuned.

Two loss functions can be used for training the network. The first one, originally proposed in [20], is an extension of the standard softmax cross-entropy loss with weight decay [6]:

$$E_{SCE}(\mathbf{w}) = -\sum_{n=1}^{N}\left(\sum_{m\in M_n}\sum_{k=1}^{K_m} t_{nmk}\cdot ln\left(y_{km}\left(\mathbf{x}_n, \mathbf{w}\right)\right)\right) + \omega_R \cdot R\left(\mathbf{w}_{Rt}, \mathbf{w}_{FC}\right) \tag{1}$$

In Eq. 1, $y_{km}\left(\mathbf{x}_n, \mathbf{w}\right)$ is the softmax score for the $n$th training image $\mathbf{x}_n$ to belong to class $k$ for variable $m$. The indicator variable $t_{nmk}$ is one if the class label of sample $n$ for variable $m$ is $k$ and zero otherwise. The sum is taken over all $N$ training samples and $K_m$ classes for task $m$. $M_n$ is the set of tasks for which the true class label is known for the training sample $n$, so that the loss in Eq. 1 considers exclusively samples $x_n$ with $t_{nmk} = 1$ for learning task $m$. In this way, the fact that the annotations for most samples are incomplete, i.e. that annotations are only available for a subset of the variables to be predicted, can be considered. If multiple annotations are available, the corresponding classification losses will be backpropagated to the joint layers from multiple classification branches, thus supporting the learning of a joint representation for all variables. The outputs for variables for which the true class label is unknown will not contribute to the loss and to the parameter update. Finally, the term $R\left(\mathbf{w}_{Rt}, \mathbf{w}_{FC}\right)$ corresponds to regularization by weight decay, which is only applied to the parameters to be updated in training; $\omega_R$ is a hyperparameter defining the influence of this term on the result.

One problem of the data described in Sect. 3.4 is its imbalanced class distribution. In this case, minimizing the cross-entropy loss in Eq. 1 will favor the dominant classes, resulting in a poor performance for the underrepresented ones. To mitigate these problems, a multi-class extension of the focal loss [38, 39] with regularization is utilized for training:

$$E_F(\mathbf{w}) = -\sum_{m\in M_n}\left(\sum_{n=1}^{N}\sum_{k=1}^{K_m}\left(1 - y_{km}\left(x_n, \mathbf{w}\right)\right)^{\gamma}\cdot t_{nmk}\cdot ln\left(y_{km}\left(x_n, \mathbf{w}\right)\right)\right) + \omega_R \cdot R\left(\mathbf{w}_{Rt}, \mathbf{w}_{FC}\right) \tag{2}$$

The only difference between the loss functions in Eqs. 1 and 2 is the penalty term $\left(1 - y_{km}\left(x_n, \mathbf{w}\right)\right)^{\gamma}$, where $\gamma$ is a hyperparameter modulating the influence of this term on the result. This penalty term forces the loss to put more emphasis on samples that are difficult to classify (having a small score $y_{km}$ for the correct class). Assuming the samples of underrepresented classes to be hard to classify by the CNN, this loss is expected to improve the results for these classes.

Starting from initial values derived in the way described earlier, stochastic minibatch gradient descent based on the ADAM optimizer [32] is applied to determine the CNN parameters, using the default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$) and a minibatch size of 300. The base learning rate $\eta$ is another hyperparameter to be tuned. We use early stopping and use the model parameters leading to the lowest loss on the validation set.

### 4.2 Text classification

Our problem is defined as value prediction for certain properties of an object, a silk fabric, given its text description, which can be written in any one of the four languages listed in Table 6. We have 4 tasks, each denominated according to the property of the underlying fabric object we want to predict: the *technique* and *material* used to create it, the *timespan* or time period when it created, and the *place* where it was created. While some descriptions directly contain some of this information, as seen in Table 4, this is sufficiently uncommon to prevent a purely extractive approach

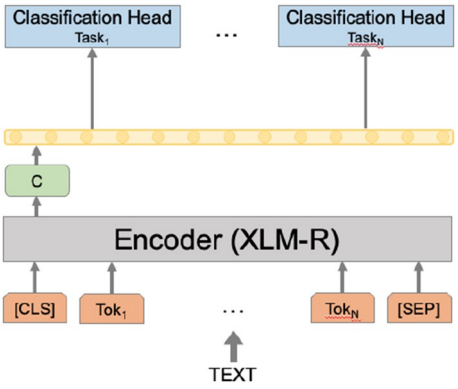


**Fig. 3** Multitask architecture: a shared XLM-R based encoder followed by task specific classification heads. The input to each classification head is the ouput of the transformer "C" corresponding to the input token "[CLS]"

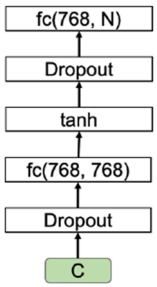


**Fig. 4** Task specific classification head: a fully connected (FC) layer followed by a tanh activation, followed by the output projection FC layer. Dropout is applied before both FC layers

from yielding good results. For example, of the 3 texts we showed, only one gives any indication as to where it was produced (*"British"*). We instead rely on regularities present in the text descriptions to make informed guesses. More technically, we frame our problem as a multiclass, multitask, multilingual text classification problem. That is, given a text description of a fabric, written in any language, we want to assign exactly one label out of a set of mutually exclusive class labels for each of the properties we wish to predict, i.e., the tasks.

The text classifier uses a hard parameter sharing based multitask architecture [50], shown in Fig. 3. It consists of a shared encoder followed by task-specific classification heads. The encoder is the multilingual large pretrained transformer, XML-R [14]. Following the method outlined in [18], a special classification token, CLS, is prepended to all inputs. The final hidden state corresponding to this token is used as the aggregate sequence representation. It is the only transformer output forwarded to the classification heads. All classification heads are identical except for the output dimension of the last layer, the output projection layer, which equals the number of classes of the task. A diagram of a classification head is shown in Fig. 4. A softmax function can convert the output logits of the last layer to normalized probabilities.

To finetune our transformer-based classifier, at each step, a task is randomly selected using proportional sampling. A batch of examples for this task is then created and fed to the classifier. The cross entropy loss is then calculated and weights adjusted through backpropagation. Adam [33] is used as the optimizer with weight decay [42].

### 4.3 Tabular classification

When considering Knowledge Graph records of objects, we can represent them as structured data. That is, a table where each row represents an object and each column a property. We use four separate task-specific classifiers to perform tabular classification. These all use the same learning algorithm, Gradient Boosted Decision Trees (GBDT) [22], implemented in XGBoost [12]. The input to the tabular classifier consists of the categorical values of non-target variables plus the identifier for the museum, as shown in Table 7. We replace missing values for a feature with a predefined value, represented by the symbol "[NA]" ("Not Available") in the table. The output of the classifier, for each example, consists of $N$-dimensional logit vector. It is used with the softmax function to predict a target class out of $N$ possible classes. This classifier trained by gradient descent to minimize the cross-entropy loss.

#### 4.3.1 Hyperparameters

While a detailed explanation of each hyperparameter that control the resulting model and learning of GBTs is beyond the scope of this work, we believe some contextualization

**Table 7** Tabular Classification, one example input row per task

| Target variable | Target value | Feature | | | | |
|---|---|---|---|---|---|---|
| | | Museum | Place | Timpespan | Technique | Material |
| Place | FR | risd | – | [NA] | [NA] | Animal fibre |
| Timespan | XVIII | met | [NA] | – | Embroidery | Animal fibre |
| Technique | Other technique | garin | ES | XX | – | Vegetal fibre |
| Material | Vegetable fibre | vam | GB | XIX | Embroidery | – |

Note: time label format changed to roman numbers for ease of readability

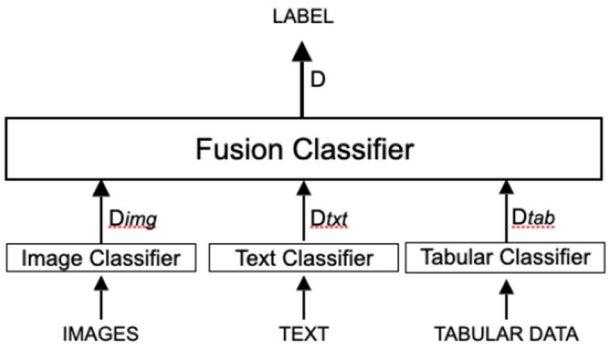


**Fig. 5** Architecture of the multimodal classifier. Each classifier based on a single modality takes its own independent decision, $D_c$, which serves as input to the multimodal classifier. The final decision $D$ is taken by the multimodal classifier, predicting a task-specific label and assigning it to the record

is required. This is due to the relatively larger number of hyperparameters tuned for GBTs in Sect. 5.3 compared to the Neural Network based methods used for the other modalities, and for the convenience of the reader.

The hyperparameters *max_depth* (maximum depth of a tree), *min_child_weight* (minimum weight for tree partitioning), and *gamma* (minimum loss reduction for tree partition) all directly control model complexity, which in turn can have significant consequences in terms of fitting. The hyperparameters *subsample* (the percentage of data sampled per iteration) and *colsample_bytree* (the ratio of features sampled per iteration) can reduce overfitting by adding random noise to the iterative tree building process. Finally, the *learning rate* and *number of rounds* control, respectively, the amount of learning per round and the total amount of learning (i.e., the total number of trees).

### 4.4 Multimodal classification

Our approach to multimodal classification, shown in Fig. 5, follows a decision level late fusion approach, in which the decision (prediction) from each of the 3 modalities serves as the input to a classifier that takes the final decision on which label to assign to the record. We choose the GBDT algorithm for the multimodal classifier. The input is just one column for each of the three modalities, each column containing the class labels predicted by the corresponding classifier for all of the tasks. If a modality is missing, the values in the corresponding column are set to the missing value indicator [NA], just in the way missing class labels are considered by the tabular classifier. Thus, the multimodal classifier can cope with incomplete records (i.e. records with missing modalities) by design. We created a separate multimodal classifier for each task, i.e. no multitask learning is applied in multimodal classification. .

**Table 8** Hyperparameters tuned (image classification)

| Hyperparameter | Range | Best |
|---|---|---|
| Learning rate $\eta$ | [1e-5, 1e-3] | 1e-4 |
| Weight decay $\omega_R$ | [0.0, 1e-5] | 1e-3 |
| Degree of fine-tuning $NL_{RT}$ | [0, 36] | 30 ($E_F$)<br>15 ($E_{SCE}$) |
| Loss $E(\mathbf{w})$ | {$E_{SCE}(\mathbf{w})$ (Eq. 1), $E_F(\mathbf{w})$ (Eq. 2)} | Focal |

An optimal variant is obtained with $\eta$=1e-4, $\omega_R$=1e-3, $NL_{RT}$=30 (i.e., 10 residual blocks), with the focal loss $E_F(\mathbf{w})$

There are several advantages to late fusion over early or intermediate level fusion in our case. Firstly, each record may have multiple images but a single text description. Effectively, the input dimensionality is different. With late fusion, we allow the image classifier to deal with it independently, e.g., by classifying multiple images for the same object and picking the decision with the highest confidence. Secondly, the decisions, represented by a one-hot class vector, have a smaller dimensionality than intermediate representations and thus are more appropriate for scenarios with few samples, which is a common problem in the context of our domain (cultural heritage).

## 5 Experiments and results

### 5.1 Image classification

For all experiments in the frame of image classification, we use the split of the dataset described in Sect. 3.4 in order to train the CNN for image classification presented in Sect. 4.1.1 by means of the training strategy described in Sect. 4.1.2. We use all images that are assigned to a record for training and classification, assigning the class labels of the corresponding records to all images associated with it. As pointed out in Sect. 4.1.1, for records associated with multiple images, all images are classified by the CNN at test time, and the image-based prediction having the highest class score is chosen to be the final result.

*Experimental setup* The workflow of our experiments is as follows: The training dataset is used to update the weights $\left(\mathbf{w}_{Rt}^T, \mathbf{w}_{FC}^T\right)^T$ of the CNN with early stopping. The model parametrization and hyperparameters leading to the lowest loss are calculated on the validation set. In this context, we tuned the hyperparameters listed in Table 8 and described in Sect. 4.1, choosing the values achieving the highest average $F1$ scores on the validation set. Table 8 also presents the selected hyperparameter values. Finally, all test set records for which at least one image is available are used for an

**Table 9** $F1$ scores ($F1$) and overall accuracies (OA) of the image classifier obtained by minimizing the Softmax loss (Eq. 1) and the focal loss (Eq. 2) both for the validation and the test sets (evaluated per record)

| | Variable | Validation set | | Test set | |
|---|---|---|---|---|---|
| | | $F1$ [%] | OA [%] | $F1$ [%] | OA [%] |
| Focal loss | Place | 49.2 | 62.5 | 47.0 | 63.1 |
| | Timespan | 58.4 | 63.8 | 57.5 | 64.5 |
| | Technique | 75.5 | 79.0 | 77.9 | 80.2 |
| | Material | 52.2 | 80.6 | 51.2 | 80.6 |
| | Average | 58.8 | 71.5 | 58.4 | 72.1 |
| Softmax loss | Place | 48.2 | 61.0 | 47.2 | 62.2 |
| | Timespan | 56.0 | 64.4 | 54.2 | 64.9 |
| | Technique | 72.2 | 75.8 | 74.0 | 76.8 |
| | Material | 45.0 | 79.4 | 43.4 | 80.7 |
| | Average | 55.4 | 70.2 | 54.7 | 71.2 |
| Δ | Average | 3.4 | 1.3 | 3.7 | 0.9 |

Δ gives the difference between the quality metrics achieved using the focal loss and the softmax loss

independent evaluation, using the hyperparameters values tuned on the validation set.

We will report the overall accuracies as well as the average $F1$ scores of the best CNN variant in terms of the average $F1$ score obtained on the test and validation sets for two variants: the first CNN variant is trained by minimizing the softmax cross-entropy loss (Eq. 1), whereas the second variant is trained by minimizing the focal loss (Eq. 2). The overall accuracy *OA* describes the percentage of correctly classified images, denoted as true positives *TP*, among all classified images. As the *OA* is biased towards classes with more examples in an imbalanced class distribution, the classification performance of underrepresented classes is not reflected by the *OA*. In contrast, the class-specific $F1$ scores, being the harmonic means of precision (i.e., the percentage of the images assigned to a certain class that actually corresponds to that class in the reference) and recall (i.e., the percentage of the samples of a class according to the reference which is also assigned to that class by the CNN) reflect the classifier's ability to predict a certain class. We report the average $F1$ scores (also referred to as macro-averaged $F1$ score) per variable, i.e., the average values of all class-specific $F1$ scores of the classes for that variable.

*Results*

The quality metrics obtained on the validation and test sets are listed in Table 9. These quality metrics are determined on the basis of the prediction results for records (i.e., not on the raw results for individual images in case of records having multiple images). In this section, some general observations and the conclusions drawn from them will be briefly described, where a more detailed analysis of the results can be found in Sect. 6.

Comparing the $F1$ scores as well as the OAs obtained on the validation and the test set, respectively, shows that the hyperparameter tuning on the validation set did not result in overfitting as the order of magnitude of the quality metrics on the validation and the test set are en par. Furthermore, the average $F1$ scores and the OAs are higher in case of minimizing the focal loss in training. Accordingly, it can be concluded that the classifier is able to better predict the classes of the four tasks by focusing on harder training examples, as is realized in the case of the focal loss. In particular, underrepresented classes benefit more from the use of the focal loss, which is indicated by the larger improvements in terms of the $F1$ scores compared to the improvements in terms of OA. The average $F1$ scores over all variables is 3.7% higher in the evaluation for minimizing the focal loss compared to minimizing the softmax cross-entropy, whereas the improvement in terms of OA amounts to 0.9% on average.

**Table 10** Hyperparameter tuning

| Hyperparameter | Range | Best |
|---|---|---|
| Batch size | 4, 8, 32 64 | 4 |
| Learning rate | 1e-6, 1e-5, 3e-5, 5e-5 1e-4 | 1e-5 |
| Weight decay coefficient | 0.0, 0.01, 0.02, 0.04 0.05 | 0.04 |
| Total epochs | 4, 8, 12, 16, 20 | 16 |

Hyperparameters, the investigated range, and the value chosen to be the best in 50 random trials according to macro-$F1$ evaluated on the validation set

## 5.2 Text classification

*Experimental setup* In the text classification experiment we use the method described in Sect. 4.2, implemented using PyTorch [47] and Transformers [60], and the data described in Sect. 3.4 split into training, validation, and test subsets as described in Sect. 3.4. We used the base XLM-R architecture ( 125M parameters) with 12-layers, 768-hidden-state and its respective provided weights. The layers in the classification heads are initialized using the normal distribution $\mathcal{N}(0.0, 0.02)$ with bias parameters set to zero.

First, we performed a 50-trial random search hyperparameter tuning implemented using Optuna [1]. During hyperparameter tuning, the text classifier is trained on the train set, and we chose the hyperparameters that resulted in the highest macro $F1$ score obtained by evaluating on the validation set. These hyperparameters are detailed in Table 10. We then train a model on the train set with the previously selected hyperparameters and evaluate it on both the validation and test sets.

**Table 11** *F*1 scores (*F*1) and overall accuracies (OA) obtained in the multitask experiment both for the validation set and the test set (text classification)

| Variable | Validation set | | Test set | |
|---|---|---|---|---|
| | *F*1 [%] | OA [%] | *F*1 [%] | OA [%] |
| Place | 92.2 | 91.3 | 93.1 | 92.6 |
| Timespan | 83.8 | 89.8 | 82.1 | 88.0 |
| Technique | 87.1 | 88.2 | 88.0 | 89.7 |
| Material | 78.5 | 85.0 | 78.7 | 85.4 |
| Average | 85.4 | 88.6 | 85.5 | 88.9 |

*Results* The results of text classification are shown in Table 11, which presents the overall accuracy and the average *F*1 scores achieved on all records containing text in the validation and test sets. In terms of *F*1 and overall accuracies, these are the best results for any single modality by a significant margin. This is offset by the fact that the text modality is only present in 39.3% of all records (Table 3).

## 5.3 Tabular classification

*Experimental setup* The experiments for tabular classification follow a similar protocol as those for the image and text classifiers, the main exception being that this classifier is not based on multitask learning. Thus, for each task we train an individual classifier with different parameters and hyperparameters, selected by task-specific hyperparameter tuning using grid search. We show the hyperparameters, the search space for tuning, and the selected values in Table 12. Note that the ranges selected were all within very reasonable intervals as an additional guard against overfitting.

*Results* We show the evaluation results in Table 13. Given that it essentially relies on co-occurrences of very coarse labels, the results seem reasonable. In fact, in terms of *F*1 and accuracy, they almost match the image classifier. We also show feature importance by gain in Table 14. For every task, the tabular classifier's most important feature is the museum. That could probably be expected, because museums are not random collections of objects.

**Table 13** *F*1 (*F*1) and overall accuracies (OA) obtained in the experiment both for the validation set and the test set (tabular classification)

| Variable | Validation set | | Test set | |
|---|---|---|---|---|
| | *F*1 [%] | OA [%] | *F*1 [%] | OA [%] |
| Place | 47.9 | 62.4 | 46.2 | 61.9 |
| Timespan | 57.4 | 65.1 | 58.6 | 67.6 |
| Technique | 68.6 | 74.2 | 68.3 | 73.0 |
| Material | 50.7 | 82.1 | 49.4 | 82.1 |
| Average | 55.4 | 70.0 | 55.6 | 71.2 |

**Table 14** Tabular classifier: feature importance per task (information gain)

| Target | Feature | | | | |
|---|---|---|---|---|---|
| Variable | Museum | Place | Timespan | Technique | Material |
| Place | 0.49 | – | 0.20 | 0.12 | 0.19 |
| Timespan | 0.41 | 0.31 | – | 0.16 | 0.12 |
| Technique | 0.40 | 0.29 | 0.17 | – | 0.14 |
| Material | 0.39 | 0.21 | 0.16 | 0.24 | – |

## 5.4 Multimodal classification

*Experimental setup* For the experiments involving multimodal classifiers, we started by training the three classifiers based on single modalities (images, text, tabular, respectively) on the training set independently of each other in the way described in Sects. 5.1 – 5.3. After that, these classifiers were used to classify the samples in the validation set. Finally, we used these predictions as inputs to train the multimodal classifier on the validation set. We used five-fold cross-validation on the validation set to perform hyperparameter tuning using grid search for the same hyperparameters that used in tuning the tabular classifier. The details of hyperparameter tuning are shown in Table 15.

*Results* The results of the experiments are shown in Table 16. In this table, we compare the results of the multimodal classifier with and without using the raw tabular data as an additional input. As expected, the variant of

**Table 12** Hyperparameter tuning for the tabular classifier: hyperparameters, the investigated range of values (Range) and interval of the search, and best values for each task, chosen by grid search according to macro-*F*1 evaluated on the validation set

| Hyperparameter | Range | Interval | Place | Timespan | Technique | Material |
|---|---|---|---|---|---|---|
| colsample_bytree | [0.6, 1.0] | 0.2 | 0.8 | 0.8 | 0.6 | 0.8 |
| Gamma | [0.0, 0.4] | 0.2 | 0.4 | 0.2 | 0.2 | 0.0 |
| learning_rate | [0.1, 0.3] | 0.1 | 0.3 | 0.3 | 0.3 | 0.2 |
| max_depth | [2, 8] | 2 | 4 | 4 | 4 | 8 |
| min_child_weight | [1, 4] | 1 | 1 | 2 | 4 | 2 |
| n_round | [100, 500] | 100 | 100 | 100 | 100 | 500 |
| subsample | [0.6, 1.0] | 0.2 | 0.6 | 1.0 | 0.6 | 0.8 |

**Table 15** Hyperparameter tuning for the multimodal classifier: hyperparameters, the investigated range of values (Range) and interval of the search, and best values chosen by grid search according to macro-*F*1 evaluated on the validation set

| Hyperparameter | Range | Interval | Place | Timespan | Technique | Material |
|---|---|---|---|---|---|---|
| colsample_bytree | [0.6, 1.0] | 0.2 | 0.6 | 0.6 | 0.8 | 0.6 |
| gamma | [0.0, 0.4] | 0.2 | 0.2 | 0.4 | 0.4 | 0.4 |
| learning_rate | [0.1, 0.3] | 0.1 | 0.1 | 0.1 | 0.3 | 0.3 |
| max_depth | [2, 8] | 2 | 4 | 4 | 6 | 4 |
| min_child_weight | [1, 4] | 1 | 1 | 1 | 4 | 4 |
| n_round | [100, 200] | 100 | 100 | 100 | 100 | 100 |
| subsample | [0.6, 1.0] | 0.2 | 0.6 | 0.6 | 0.6 | 0.6 |

The hyperparameter space is the same for all experiments reported in this section. The selected hyperparameters apply to the multimodal classifier using the complete set of input modalities, as shown in Fig. 5 and correspond to the results shown in Table 16

**Table 16** *F*1 scores (*F*1) and overall accuracies (OA) on the test set of the multimodal classifier

| Variable | *F*1 [%] | OA [%] |
|---|---|---|
| Place | 76.7 | 79.0 |
| Timespan | 73.1 | 80.1 |
| Technique | 83.8 | 85.4 |
| Material | 61.3 | 85.5 |
| Average | 73.7 | 82.5 |

**Table 17** Feature importance, measured by information gain, for the multimodal classifier per modality for all tasks

| Variable | Text | Image | Tabular |
|---|---|---|---|
| Place | 0.47 | 0.21 | 0.31 |
| Timespan | 0.40 | 0.23 | 0.37 |
| Technique | 0.20 | 0.36 | 0.43 |
| Material | 0.46 | 0.23 | 0.31 |

the classifier using these additional input features produces slightly better results than the one without these features. Table 17 gives the feature importance, measured by information gain, of each individual modalitiy. We also performed an ablation study to assess the importance of the individual modalities for the classification results The ablation study was performed by removing one of the modalities from the input of the fusion classifier, leaving only the other two modalities (Table 19).

The overall accuracy and mean *F*1 score achieved by the multimodal classifier are better than those for the image classifier (Table 9) and slightly worse than those reported for the text classifier (Table 11), but this comparison is inconclusive because the results in Table 11 only consider records for which text is available, which is only about 39% of the test set, whereas the evaluation of the image classifier is based on about 96% of the test set and multimodal classification is based on the complete test set.

For the majority of the samples, only images and / or tabular information are available, and thus the prediction would be based on these modalities. To be able to allow for a comparison of the results of all modalities, we carried out an evaluation of all modality-specific classifier and the multimodal classifier on the entire test set. In this evaluation, a record for which a modality was missing was considered a wrong prediction for that modality-specific classifier. For instance, a record without images

**Table 19** Average *F*1 scores of the multimodal classifier using input modalities (average over all tasks)

| Input modality | *F*1 score [%] |
|---|---|
| Image + text | 70.5 |
| Image + tabular | 59.8 |
| Text + tabular | 69.9 |
| All | 73.7 |

**Table 18** Mean *F*1 scores (*F*1) and overall accuracies (OA) of the different classifiers evaluated on the entire test set

| Classifier | Image | | Text | | Tabular | | Multimodal | |
|---|---|---|---|---|---|---|---|---|
| Variable | *F*1 | OA | *F*1 | OA | *F*1 | OA | *F*1 | OA |
| Place | 38.0 | 46.8 | 64.6 | 42.3 | 46.2 | 61.9 | 76.7 | 79.0 |
| Timespan | 49.2 | 45.6 | 54.0 | 44.7 | 58.6 | 67.6 | 73.1 | 80.1 |
| Technique | 73.5 | 70.5 | 40.9 | 26.1 | 68.3 | 73.0 | 83.8 | 85.4 |
| Material | 46.5 | 67.5 | 37.4 | 21.6 | 49.4 | 82.1 | 61.3 | 85.5 |
| Average | 51.8 | 57.5 | 49.2 | 33.7 | 55.6 | 71.2 | 73.7 | 82.5 |

Samples for which a modality was missing are considered as errors for the corresponding modality-specific classifier

was considered to be an incorrect prediction for the image classifier. The resultant overall accuracy values and mean $F1$ scores are shown in Table 18. In this comparison, the multimodal classifier outperforms all the classifiers based on a single modality only.

The results of the modality ablation study, shown in Table 19, and comparing with the individual modality results show in Table 18 confirms the assumption that each modality provides meaningful contribution. The results of a multimodal classifier that combines any two modalities are superior to those of any individual modality. Further, combining all three produces the best results.

## 6 Analysis

### 6.1 Image classification

Here, we will provide a detailed analysis of the results of the CNN-based image classifier in Table 9. The table shows that the classification performance strongly varies between tasks. Comparing the OAs, one can see that the variable *material* achieves the highest OAs, followed by *technique* and *timespan*; the worst OA is achieved for the variable *place*. Taking the class structure shown in Table 1 into account, a connection can be made to the number of classes constituting a task's class structure. The larger the number of classes to be distinguished, the lower the achieved percentage of correctly classified images in the softmax experiment, where a similar behavior can be observed for the focal experiment; *material* having three classes has the highest OA of 80.7%, followed by *technique* having four classes with 76.8% correct predictions and *timespan* with five classes with a OA of 64.0%, whereas *place* with nine classes has the lowest OA of 62.2%.

An analysis of the task-specific $F1$ scores in connection with the class distributions of the respective task indicates a dependency of the $F1$ score on the degree of class imbalance. Taking the ratio of the number of image examples for the majority class, i.e., the class with the most labeled examples in the dataset, in relation to the number of image examples for the minority class, i.e., the class with the fewest examples, a negative correlation between this ratio and the achieved task-specific $F1$ score can be observed for the focal loss experiment, where a similar behavior can be observed for the softmax experiment. The majority class of *technique* has 2.5 times as many examples as the minority class and *technique* has the highest $F1$ score of 77.9%, followed by *timespan* with a ratio of 4.9 and a score of 57.5% and *material* with a ratio of 7.7 having a score of 51.2%. The lowest $F1$ score of 47.0% is obtained for *place* with a ratio of 8.3. We attempted to overcome this dependency of the $F1$ scores on the class distributions through focusing on hard training examples by means of the presented variant of the focal loss in Eq. 2. Analyzing the improvements of the $F1$ scores by utilizing the focal loss instead of the softmax cross-entropy loss shows that the focal loss indeed reduces this dependency: except for the variable *place*, there is an improvement of the task-specific $F1$ scores, and in these cases it is larger for tasks with a high class imbalance (indicated by a high ratio between the number of examples for the majority class and the minority class, respectively). The $F1$ score of *material* (ratio of 7.7) is improved by 7.8%, whereas the $F1$ score of *technique* (ratio of 2.5) is improved by 3.9%. The variable *place* with a ratio of 8.3 should have received the largest improvement in $F1$ score according to the general trend, but it actually is slightly worse (-0.2%). We assume this to be related to the large number of classes to be distinguished for *place*, which might make a correct prediction more complicated for this variable than for the other ones.

In summary, the utilization of the focal loss improves the performance of the trained classifier in correctly predicting the properties of silk based on images. Even though the variable-specific $F1$ score still seems to depend on the degree of imbalance of a task's class distribution, focusing on hard examples during training primarily improves the task-specific $F1$ scores of tasks with large class imbalances, as long as the number of classes to be differentiated is not too large. Solving the remaining challenge of predicting all classes of a task equally well may require more data, as not all aspects of all silk properties are equally well represented in the available images.

### 6.2 Text classification

The results for the text modality, shown in Table 11, are better, in terms of $F1$ and OA, than the results for the image modality (Table 9) or the tabular modality (Table 13). This is not surprising, since the properties we are predicting are important to domain experts, they, or their taxonomy subclasses, are often included in the text descriptions of the cultural heritage objects (altough not necessarily using the same words). Even when they are not, we can intuitively expect some degree of similarity between text descriptions of objects with similar underlying values for these properties, either globally or at least within the same museum.

The biggest disadvantage of the text classifier is that text descriptions are present in the dataset far less often than images, as shown in Table 3. Once adjusted for missing modality, given that more than 60% of the records are missing this modality, the text classifier actually performs the worst of any modality, as shown in Table 18.

We analyzed about 20 misclassified English language test set examples for each task. In around half of the cases, there was no direct information that could've allowed an accurate classification. E.g., no location mentioned when attempting to classify place or no year mentioned when attempting

**Table 20** Examples of misleading text descriptions

| Text description (Snippet) | Predicted | True |
|---|---|---|
| Motifs found on **seventeenth-century** coverlets but must have been made in the early eighteenth century (...) Embroidery of Gujerat [sic] in the **Seventeenth Century** | XVII | XVIII |
| Derived from engravings after Maarten de Vos which first appeared in Gerard de Jode's **1579** illustrated bible | XVI | XVII |
| Center text reads **"Vole vole mon coeur!** | FR | GB |
| Depiction from the **Italian** poem | IT | GB |

Emphasis added to highlight the misleading text snippets

to classify timespan. This forces the classifier to rely on other statistical regularities present in the text to provide a classification.

The *material* task is particular. Its most common class, "animal fibre", is a de facto background class. All records in the dataset should be of silk fabrics, which means the material they are made of is an "animal fibre". Some have other materials too. These other materials can correspond to a vegetable fibre (e.g., cotton) or a thread with some metal (e.g., gold thread). While the problem of not having label specific information in the text is common in the examples we analyzed (6/20), obviously incorrectly labeled examples were even more common (9/20). This occurs when either the original record was missing the correct label or when the automatic extraction and linking of the label failed. The high prevalence of this type of error within this task in the examples we analyzed, combined with its absence in other tasks, leads us to suspect that this is the main cause of the relatively lower accuracy and $F1$ scores for this task.

The *technique* task is also particular in terms of the examples we analyzed. A significant number of examples (5/19) contain information that would imply multiple labels, where usually a small part of the object was produced using a different technique from the main part of the object. A similar type of error occurs in the *timespan* task within a similar proportion of examples (5/10). In the *timespan* task, this can occur when an object was produced at a certain date but later altered or when the estimated date of production within the text crossed centuries.

We hypothesize that the somewhat better results for the place task are connected to regularities between the museum and an object's place of production. This connection is suggested in Table 14. Text descriptions are very indicative of the museum, not just in the language but usually also in style, length, and topics.

Finally, we would like to point out at the existing of misleading text examples such as in Table 20 where information in the text can correspond to an incorret label. We do this to give the reader a better understanding of the challenges faced by an algorithm.

## 6.3 Tabular classification

When compare to the other modalities, the tabular classifier performs (Table 13) slightly worse than the image classifier (Table 9) in terms of average $F1$ score, respectively, 55.6% versus 58.4%. However, this situation is reversed when accounting for missing modality (Table 18), with the tabular classifier outperforming the image classifier with $F1$ scores of 55.6% versus 51.8%. Intuitively, from a domain perspective, we can expect that these variables to be associated. For example, a certain country is more active in the textile industries during a certain timespan than during others. Further, museums are typically curated and not random collections. However, given the limited number of features and the coarseness of the labels, we should not overestimate the strength of the association between variables, which we calculated as Cramer's V in Fig. 6. Cramers' V is a symmetric measure that gives a value between 0 and 1 for the association between two nominal variables. We see that the values are, by themselves, relatively low, with

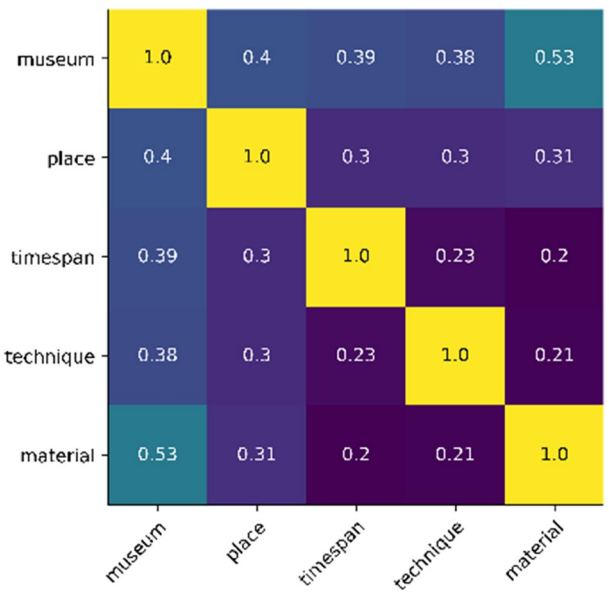


**Fig. 6** Association between features of the tabular classifier (Cramer's V)

only material and museum having a value above 0.5. The association between museum and the other properties was the highest (first column or row) which again reinforces the belief that the curated nature of museum collections and its impact on the other properties is learnable from this dataset. The strength of the association does not directly translate into feature importance as measured by information gain (Table 14). Rather, it seems, a particular combination of these associations is being learned by the tree boosting algorithm in such a way that results in a relatively effective classifier.

### 6.4 Multimodal classification

As pointed out in Sect. 5.4, the comparison of the classification accuracies indicates that the text classifier achieves the best performance of all modalities (Table 11), but of course it is only applicable when text is available, which is only the case for a relatively low number of records, (cf. Table 3). This confirms our hypothesis that multimodal classification results in a better classification performance if one of the aspects under consideration is to obtain correct predictions for a number of records that is as large as possible. When evaluated on samples having text, the text classifier might achieve higher accuracy metrics; however, a considerable percentage of samples cannot be classified in that way, and the total number of correct classifications is largest when using multimodal classification (cf. Table 18).

As Tables 17, 18 and 19 imply, each modality contributes significantly to the multimodal classifier. Looking at the feature importance of the multimodal classifier in Table 17, we can see that the output of the text classifier is the most important feature, except when it comes to predicting technique, almost certainly due to the relatively small number of records with an annotation for *technique* in the validation set for which text is available (487 records as opposed to about 1100-1600 for the other tasks). Most errors in the timespan task occur between chronologically similar dates. Most errors in the place task occur between countries that are geographically close to each other, e.g., Italy (IT) and France (FR). As far as material is concerned, errors occur primarily between animal fiber and the other labels, because all objects are made of silk and due to the label imbalance. No clear trend can be observed for the prediction of technique.

## 7 Integrating and visualizing the predictions

To model the predictions as part of the SILKNOW Knowledge Graph ontology we use classes and properties of the Provenance Data Model (Prov-DM), more specifically the PROV ontology (PROV-O)[10], an OWL2 ontology. It makes it possible to map PROV-DM to RDF. It allows the expression of important elements of the predictions for each modality. These different predictions can be represented using different `prov:activity` classes each. The image, text description, or categories each prediction is based on are represented with the property `prov:used`. The exact date of the prediction is represented with prov:atTime and prov:wasAssociatedWith connects the activity class to the `prov:SoftwareAgent` class, which is used to describe the particular algorithm and model used. The predicted metadata value is represented with rdf:Statement, connected to prov:activity via a `prov:wasGeneratedBy` property. The confidence score of the prediction is expressed through the property L18 ("has confidence score") from our own SILKNOW Ontology. The predicted value is expressed in form of a URI with `rdf:object`, the type of the predicted property through rdf:predicate and its fitting CIDOC-CRM property type. The property `rdf:subject` connects the statement to the production class (E12) of the object in the Knowledge Graph. Every prediction is inserted in the appropriate part of the existing KG. For example, if a material value gets predicted, it gets inserted with the CIDOC-CRM property `P126_employed` at the production class of the object. See Fig. 7 for an illustration of the data model.

The prediction models were only trained on group labels (Sect. 3.3), and thus can only predict those. It is sometimes necessary to map them back to a more concrete concept of the SILKNOW Thesaurus. If, for example, "Damask" is predicted in form of its facet link http://data.silknow.org/vocabulary/facet/damask it will be automatically converted into http://data.silknow.org/vocabulary/168, as facet links are too general for concrete category values. All predictions are converted one after another using the described data model and saved as a turtle file format which is uploaded and stored as its own graph identified by http://data.silknow.org/predictions. This makes it possible to always identify and eventually separate predictions from original values obtained from the museums. In total, 98,379 predictions were made for 19,248 distinct objects. These were uploaded into the SILKNOW Knowledge Graph.

In our exploratory search engine ADASilk, the predictions are displayed differently from values that come originally from the museums: They are shown in blue together with their confidence score as a percentage next to them. A tooltip is available to explain how this value was predicted, including the modality, algorithm, model identifier, etc. To display predictions like this on ADASilk, the respective SPARQL query was updated and new subqueries are used to take into account the aforementioned new properties. See Fig. 8 for a screenshot.

[10] https://www.w3.org/TR/prov-dm/.

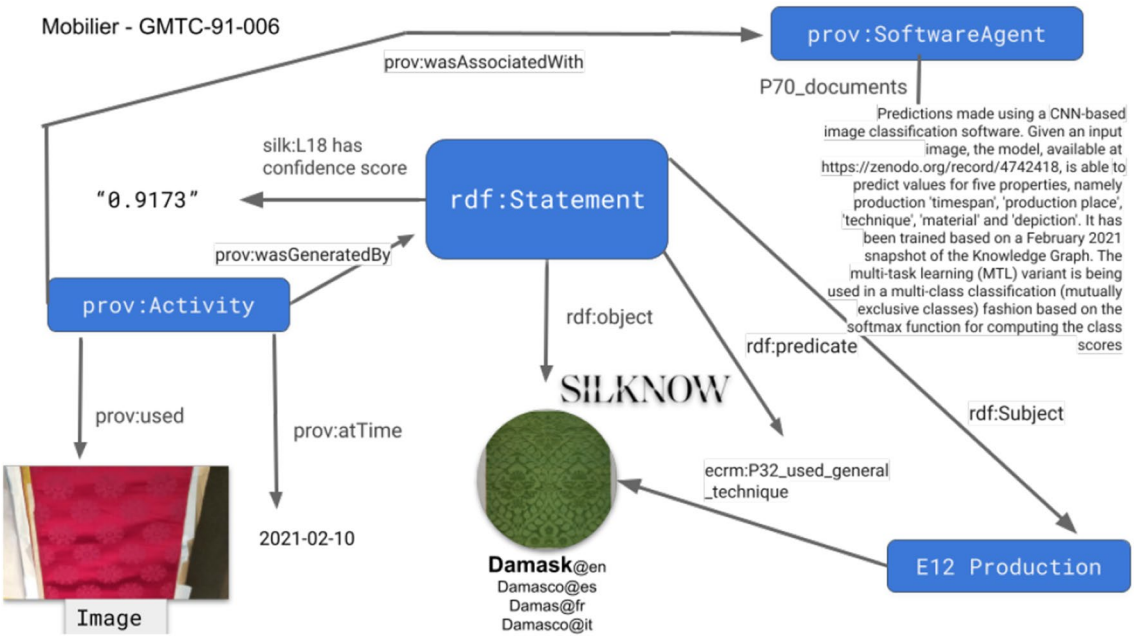


**Fig. 7** Graph showing the prediction of the production technique (damask) with a high confidence score (0.9173) using the textual analysis software

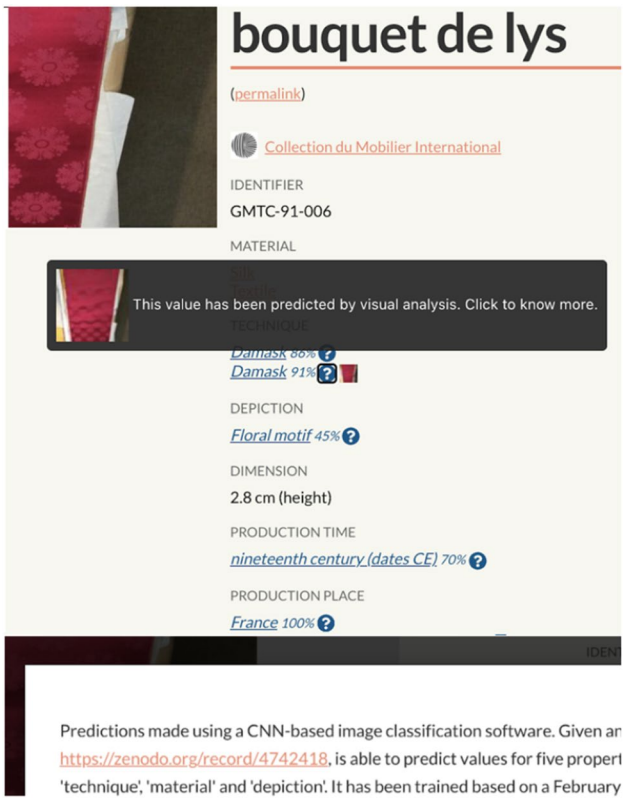


**Fig. 8** UI Screenshot showing the prediction of the production technique (damask) with a high confidence score (0.9173) using the image analysis software

## 8 Conclusions

We presented results for three individual modality classifiers, as well as multimodal results. In terms of our original hypothesis, presented in Sect. 1.2, we showed that we were in fact able to accurately predict missing properties in the digitized silk fabric artifacts that made up our dataset. While the quality of predictions varied between individual modalities, we showed that the multimodal approach provided the best results. To recapitulate, our contributions included the already mentioned multimodal approach tailored specifically to the challenges we faced in the multimodal scenario, including the incomplete overlap of data across modalities. The individual approaches of each modality-specific classifier also provide a useful contribution to the automated classification of cultural heritage objects. The image and text classifier offer the possibility of being applied to data outside a Knowledge Graph (KG) or database, possibly even directly submitted by the user of a system. The tabular classifier, on the other hand, offers the possibility of classifying data in a KG or database when no text descriptions or images are present by relying on other properties. It is also important to remember that in most practical situations, including inside a KG or other knowledge bases, images are more common than text descriptions of objects in the cultural heritage domain.

The data we used in our work originally comes from many museum sources and is from a very specific Cultural Heritage domain: historical, European silk fabrics. We

applied common methods to process such data and developed an ontology and a Knowledge Graph out of the original museum texts and images. Such an effort comes typically with challenges, which in our case consisted mostly of a small amount of (training) data, domain specificity, different styles of writing texts and capturing images of objects, different languages (in the case of texts) and finally simply annotation errors, typos, and other errors that happened during the original digitization. Not all of these challenges can be completely overcome and some of them, like the metadata gaps, even constitute part of the motivation to conduct this research work. As some data imperfection could still not be totally excluded, some removal of data was necessary to ensure sufficiently clean and class balanced data for our supervised approaches. This could, however, be very much alleviated through the grouping of certain labels, which was also possible through our domain expert-designed thesaurus about silk fabric concepts. In the end we can present a cultural heritage dataset that can be used for automated classification or even multimodal approaches. In this work, we also provide the data modelling of how metadata predictions for data such as our can be represented within knowledge graphs or other knowledge bases.

We have shown that properties of silk fabrics can be predicted from images of these fabrics. In this context, we proposed to use the focal loss for training to compensate for the effects of class imbalance in the training set, a problem that is quite common in the cultural heritage domain. Our results indicate that the proposed strategy can mitigate this problem to a certain degree, in particular improving the classification performance for the underrepresented classes in terms of the $F1$ score. Image classification performs particularly well for the task to predict the technique used for producing a fabric. Nevertheless, there is still room for improvement, as indicated by the performance metrics for all variables.

When text descriptions are present, the text classifier provides the best results of any single modality. It seems, thus, that the text classifier was able to overcome the primary challenges it faced: small dataset, domain specificity, cross-linguality, and museum specific text styles. This was primarily achieved by the choice of XLM-R as the basis of the text classifier. Misleading text descriptions stand out as a challenge for text classification.

When all data is considered, we have shown that the multimodal approach is the best according to the macro $F1$ metric. While most records contain images, not all do (3.4%) and a smaller number of records contain neither text nor images (2.1%). On the other hand, if we had tried to implement a classifier using the text modality alone, we could only classify 40% of the records. While we can say that a multimodal approach does allow us to classify a greater number of records than using images alone, the primary practical benefit of the multimodal approach over performing just image classification is probably the qualitative improvement in classification results demonstrated.

In terms of the dataset, a perhaps a better approach could be found for dealing with noisy labels, as well as finding better ways to deal with fine-grained labels and label ontology mismatches.

Future work on image classification could concentrate on improving the performance for underrepresented classes even more, e.g., by using methods for few-shot learning. Furthermore, as some experimental results indicated that some training labels might be incorrect, training methods that are robust against such errors ("label noise") could be investigated.

The code and data used to perform the experiments reported in this work is available online at https://github.com/silknow/multimodal_cultural_heritage and https://zenodo.org/record/6590957, respectively.

**Acknowledgements** This work was supported by the Slovenian Research Agency and the European Union's Horizon 2020 research and innovation program under SILKNOW grant agreement No. 769504.

## References

1. Akiba, T., Sano, S., Yanase, T., et al.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International conference on knowledge discovery and data mining (2019)
2. Arik, S.O., Pfister, T.: Tabnet: Attentive interpretable tabular learning. In: Proceedings of the AAAI conference on artificial intelligence 35(8):6679–6687. (2021) https://ojs.aaai.org/index.php/AAAI/article/view/16826
3. Arora, R.S., Elgammal, A.M.: Towards automated classification of fine-art painting style: a comparative study. In: International conference on pattern recognition, pp. 3541–3544 (2012)
4. Belhi, A., Bouras, A., Foufou, S.: Leveraging known data for missing label prediction in cultural heritage context. Appl. Sci. (2018). https://doi.org/10.3390/app8101768
5. Belhi, A., Bouras, A., Foufou, S.: Towards a hierarchical multitask classification framework for cultural heritage. In: 2018 IEEE/ACS 15th international conference on computer systems and applications (AICCSA), IEEE, pp. 1–7 (2018b)
6. Bishop, C.M.: Pattern Recognition and Machine Learning, 1st edn. Springer, New York (NY), USA (2006)
7. Blessing, A., Wen, K.: Using machine learning for identification of art paintings. Technical report. Stanford University, USA (2010)
8. Bojanowski, P., Grave, E., Joulin, A., et al.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
9. Caruana, R.: Multitask learning. Mach. Learn. **28**(1), 41–75 (1997)
10. Caruana, R.A.: Multitask learning: a knowledge-based source of inductive bias. In: International conference on machine learning, pp. 41–48 (1993)
11. Castellano, G., Vessio, G.: Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview. Neural Comput. Appl. **33**(19), 12263–12282 (2021)

12. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. Association for computing machinery, New York, NY, USA, pp. 785–794, https://doi.org/10.1145/2939672.2939785(2016)
13. Conde, M.V., Turgutlu, K.: Clip-art: contrastive pre-training for fine-grained art classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops, pp. 3956–3960 (2021)
14. Conneau, A., Khandelwal, K., Goyal, N., et al.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for computational linguistics, Online, pp. 8440–8451, https://doi.org/10.18653/v1/2020.acl-main.747 (2020)
15. Crawshaw, M.: Multi-task learning with deep neural networks: a survey. arXiv preprint arXiv:2009.09796 (2020)
16. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), IEEE, pp 886–893 (2005)
17. Deng, J., Dong, W., Socher, R., et al.: Imagenet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition, pp. 248–255 (2009)
18. Devlin, J., Chang, M.W., Lee, K., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol. 1 (Long and Short Papers). Association for computational linguistics, Minneapolis, Minnesota, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423 (2019)
19. Doerr, M.: The CIDOC CRM, an ontological approach to schema heterogeneity. In: Semantic interoperability and integration (2005)
20. Dorozynski, M., Clermont, D., Rottensteiner, F.: Multi-task deep learning with incomplete training samples for the image-based prediction of variables describing silk fabrics. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. **4**(2/W6), 47–54 (2019)
21. Fiorucci, M., Khoroshiltseva, M., Pontil, M., et al.: Machine learning for cultural heritage: a survey. Pattern Recogn. Lett. **133**, 102–108 (2020)
22. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)
23. Gao, Y., Li, Y., Lin, Y., et al.: Deep learning on knowledge graph for recommender system: a survey. CoRR abs/2004.00387. arXiv:2004.00387 (2020)
24. Garcia, N., Vogiatzis, G.: How to read paintings: semantic art understanding with multi-modal retrieval. In: Proceedings of the European conference in computer vision workshops (2018)
25. Garcia, N., Renoust, B., Nakashima, Y.: Contextnet: representation and exploration for painting classification and retrieval in context. Int. J. Multimed. Inf. Ret. **9**(1), 17–30 (2020)
26. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016a)
27. He, K., Zhang, X., Ren, S., et al.: Identity mappings in deep residual networks. In: European conference on computer vision, pp. 630–645 (2016b)
28. Hyvönen, E., Mäkelä, E., Kauppinen, T., et al.: Culturesampo: a national publication system of cultural heritage on the semantic web 2.0. In: Aroyo, L., Traverso, P., Ciravegna, F., et al. (eds.) The Semantic Web: Research and Applications, pp. 851–856. Springer, Berlin Heidelberg (2009)
29. Iqbal Hussain, M.A., Khan, B., Wang, Z., et al.: Woven fabric pattern recognition and classification based on deep convolutional neural networks. Electronics **9**(6), 1048 (2020)
30. Joulin, A., Bojanowski, P., Mikolov, T., et al.: Loss in translation: learning bilingual word mapping with a retrieval criterion. In: Proceedings of the 2018 conference on empirical methods in natural language processing (2018)
31. Kadra, A., Lindauer, M., Hutter, F., et al.: Well-tuned simple nets excel on tabular datasets. In: Beygelzimer A., Dauphin Y., Liang P., et al. (eds) Advances in Neural Information Processing Systems, https://openreview.net/forum?id=d3k38LTDCyO (2021)
32. Kingma, DP., Ba, J. Adam: A method for stochastic optimization. In: 3rd International conference on learning representations (ICLR 2015) (2015a)
33. Kingma, DP., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (Poster), arXiv:1412.6980 (2015b)
34. Koch, I., Ribeiro, C., Lopes, C.: ArchOnto, a CIDOC-CRM-based linked data model for the Portuguese archives, Springer, pp 133–146. https://doi.org/10.1007/978-3-030-54956-5_10(2020)
35. Krizhevsky, A., Sutskever, I., Hinton, GE.: ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems 25 (NIPS'12), pp 1097–1105 (2012)
36. LeCun, Y., Boser, B., Denker, J.S., et al.: Backpropagation applied to handwritten ZIP code recognition. Neural Comput. **1**(4), 541–551 (1989)
37. Li, X., Chen, C.H., Zheng, P., et al.: A knowledge graph-aided concept-knowledge approach for evolutionary smart product-service system development. J. Mech. Des. **142**(101), 403 (2020). https://doi.org/10.1115/1.4046807
38. Lin, T.Y., Goyal, P., Girshick, R., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988 (2017)
39. Liu, W., Chen, L., Chen, Y.: Age classification using convolutional neural networks with the multi-class focal loss. IOP Conf. Ser.: Mater. Sci. Eng. **428**(012), 043 (2018). https://doi.org/10.1088/1757-899x/428/1/012043
40. Liu, X., He, P., Chen, W., et al.: Multi-task deep neural networks for natural language understanding. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for computational linguistics, Florence, Italy, pp. 4487–4496. https://doi.org/10.18653/v1/P19-1441 (2019a)
41. Liu, Y., Ott, M., Goyal, N., et al.: Roberta: a robustly optimized Bert pretraining approach. 1907.11692 (2019b)
42. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9 (2019)
43. Meng, S., Pan, R., Gao, W., et al.: A multi-task and multi-scale convolutional neural network for automatic recognition of woven fabric pattern. J. Intell. Manuf. **32**(4), 1147–1161 (2021)
44. Mensink, T., Van Gemert, J.: The Rijksmuseum challenge: museum-centered visual recognition. In: Proceedings of international conference on multimedia retrieval, pp. 451–454 (2014)
45. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 807–814 (2010)
46. Palagi, E.: Evaluating exploratory search engines: designing a set of user-centered methods based on a modeling of the exploratory search process. PhD thesis, Université Côte d'Azur (2018)
47. Paszke, A., Gross, S., Massa, F., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach H., Larochelle H., Beygelzimer A., et al (eds) Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035 (2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
48. Puarungroj, W., Boonsirisumpun, N.: Recognizing hand-woven fabric pattern designs based on deep learning. In: Advances in Computer Communication and Computational Sciences, pp. 325–336. Springer, Singapore (2019)

49. Radford, A., Kim, JW., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: Meila M., Zhang T. (eds) Proceedings of the 38th international conference on machine learning, proceedings of machine learning research, vol. 139. PMLR, pp. 8748–8763 (2021)
50. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
51. Ruotsalo, T., Aroyo, L., Schreiber, G., et al.: Knowledge-based linguistic annotation of digital cultural heritage collections. IEEE Intell. Syst. **24**(2), 64 (2009)
52. Santos, I., Castro, L., Rodriguez-Fernandez, N., et al.: Artificial neural networks and deep learning in the visual arts: a review. Neural Comput. Appl. **33**(1), 121–157 (2021)
53. Sharif Razavian, A., Azizpour, H., Sullivan, J., et al.: CNN features off-the-shelf: an astounding baseline for recognition. In: IEEE Conference on computer vision and pattern recognition workshops, pp 806–813 (2014)
54. Srivastava, N., Hinton, G., Krizhevsky, A., et al.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(2014), 1929–1958 (2014)
55. Stefanini, M., Cornia, M., Baraldi, L., et al.: Artpedia: a new visual-semantic dataset with visual and contextual sentences. In: Proceedings of the international conference on image analysis and processing (2019)
56. Strezoski, G., Worring, M.: Omniart: multi-task deep learning for artistic data analysis. arXiv preprint arXiv:1708.00684 (2017)
57. Sur, D., Blaine, E.: Cross-depiction transfer learning for art classification. Tech. Rep. CS 231A and CS 231N, Stanford University, USA (2017)
58. Tan, WR., Chan, C.S., Aguirre, HE., et al.: Ceci n'est pas une pipe: a deep convolutional network for fine-art paintings classification. In: IEEE International conference on image processing, pp. 3703–3707 (2016)
59. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Guyon I., Luxburg U.V., Bengio S., et al (eds) Advances in Neural Information Processing Systems, vol 30. Curran Associates, Inc., https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (2017)
60. Wolf, T., Debut, L., Sanh, V., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. Association for computational linguistics, Online, pp 38–45, https://www.aclweb.org/anthology/2020.emnlp-demos.6 (2020)
61. Xiao, Z., Liu, X., Wu, J., et al.: Knitted fabric structure recognition based on deep learning. J. Text. Inst. **109**(9), 1–7 (2018)
62. Yosinski, J., Clune, J., Bengio, Y., et al.: How transferable are features in deep neural networks? Adv. Neural Inf. Process. Syst. **27**, 3320–3328 (2014)
63. Zhang, C., Kaeser-Chen, C., Vesom, G., et al.: The imet collection 2019 challenge dataset. arXiv preprint arXiv:1906.00901 (2019)
64. Zou, X.: A survey on application of knowledge graph. J. Phys: Conf. Ser. **1487**(012), 016 (2020). https://doi.org/10.1088/1742-6596/1487/1/012016


**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.