Soft Computing
https://doi.org/10.1007/s00500-023-08258-x

An art painting style explainable classifier grounded on logical and commonsense reasoning

Vicent Costa[1] · Jose M. Alonso-Moral[3] · Zoe Falomir[4] · Pilar Dellunde[1,2,5]

Accepted: 11 April 2023

Abstract

This paper presents the art painting style explainable classifier named ANYXI. The classifier is based on art specialists' knowledge of art styles and human-understandable color traits. ANYXI overcomes the principal flaws in the few art painting style classifiers in the literature. In this way, we first propose, using the art specialists' studies, categorizations of the Baroque, Impressionism, and Post-Impressionism. Second, we carry out a human survey with the aim of validating the appropriateness of the color features used in the categorizations for human understanding. Then, we analyze and discuss the accuracy and interpretability of the ANYXI classifier. The study ends with an evaluation of the rationality of explanations automatically generated by ANYXI. We enrich the discussion and empirical validation of ANYXI by considering a quantitative and qualitative comparison versus other explainable classifiers. The reported results show how ANYXI is outstanding from the point of view of interpretability while keeping high accuracy (comparable to non-explainable classifiers). Moreover, automated generations are endowed with a good level of rationality.

Keywords Knowledge representation and reasoning · Explainable artificial intelligence · Fuzzy logic · Human-centered artificial intelligence

1 Introduction

The worth of visual arts, linked to culture and even economics, in society is out of doubt (Toll 2018). It is, therefore, hardly surprising that in the last years, with the advent of digitalization, virtual art encyclopedias and virtual museum tours have increased the number of online images of art paintings. In this way, we can find in the literature several visual databases from which we highlight, considering their amount of information and accessibility, the following ones: Art500k, WikiArt, and the Metropolitan Museum of Art Collections Database. First, Art500k[1] (Mao et al. 2017, 2019) is a large-scale visual arts dataset displaying more than 500,000 images enriched with metadata (e.g., the painting's style, the event represented, or the painting place). Second, WikiArt[2] is a non-profit dataset and constantly growing project which includes more than 250,000 artworks by 3,000 artists. This dataset also incorporates relevant data such as the painting's style or genre (if applicable). Third, the Metropolitan Museum of Art

Vicent Costa, Jose M. Alonso-Moral and Zoe Falomir have contributed equally to this work.

✉ Vicent Costa
vicent@iiia.csic.es

Jose M. Alonso-Moral
josemaria.alonso.moral@usc.es

Zoe Falomir
zfalomir@uji.es

Pilar Dellunde
pilar.dellunde@uab.cat

[1] Artificial Intelligence Research Institute (IIIA-CSIC), Consejo Superior de Investigaciones Científicas, Campus UAB, 08193 Bellaterra, Spain

[2] Department of Philosophy and Institut d'Història de la Ciència, Universitat Autònoma de Barcelona, Campus UAB, 08193 Bellaterra, Spain

[3] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Rua de Jenaro de la Fuente Dominguez, 15782 Santiago de Compostela, Spain

[4] Departament d'Enginyeria i Ciència dels Computadors, Universitat Jaume I, Av. Vicent Sos Baynat s/n, E-12071 Castelló, Spain

[5] Barcelona Graduate School of Mathematics (BGSMath), Gran Via de les Corts Catalanes, 585, 08007 Barcelona, Spain

[1] The database is available at https://deepart.ust.hk/ART500K/art500k.html.

[2] The database is available at https://www.wikiart.org/.

Published online: 17 May 2023

Collections Database[3] covers more than 470,000 images of artworks in its collection available under the Creative Commons open access license and includes data such as title, culture (e.g., French) or period.[4]

The design of these databases is intended to perform retrieval and classification activities. However, there are other art visual databases with different purposes. For example, the WikiArt Emotions dataset (Mohammad and Kiritchenko 2018) displays 4,105 art pieces (paintings mostly) with annotations for emotions evoked in the observer. The general purpose of this dataset is emotional recognition.

All the datasets enumerated above are just examples of how more and more data come around us every day. We are living in a digital era in which society demands data scientists to process and extract value from available data. Accordingly, artificial intelligence (AI) was pointed out as one of the most disruptive and strategic technologies of this century (European 2018).

AI is a multidisciplinary research field that is becoming pervasive in modern society. It is devoted to designing physical and virtual machines intending to perform intelligent activities and others involving other human faculties (e.g., association, motor control, or perception). As part of these intelligent activities, data analysis tasks in general and classification tasks, in particular, have been of great interest in AI. In addition, explainability is a fundamental topic for AI since its lack decreases the trust in the outcomes of AI systems, reduces their fairness (without explainability, claiming responsibility becomes hard) and usability (Ribeiro et al. 2016), and makes it more probable to overlook whether algorithms have been trained using a biased dataset (Hagras 2018; Samek et al. 2019). Research on explainable AI (XAI) involves not only technical but also ethical and legal issues (Arrieta et al. 2020; Gunning et al. 2021; Alonso-Moral et al. 2022). XAI is a fruitful research field with outstanding applications, for example, in medicine (Ma et al. 2021; El-Sappagh et al. 2021). Furthermore, with XAI explanations aimed at humans, both quantitative and qualitative evaluations are of paramount importance (Vilone and Longo 2021).

All in all, in the last years, there is a growing interest in the challenge of applying AI for art painting style categorization, as we will see in the next section. Unfortunately, only a few authors have addressed the challenge of designing XAI for art painting style categorization.

The main contribution of this paper is the design and validation of a new art painting style explainable classifier, named ANYXI, which provides continuity to the research presented by Costa et al. (2018, 2021). ANYXI uses human-understandable color features and categorizations based on the art experts' definitions of the styles, yielding human-friendly explanations of its classification results. Although it is not the only explainable algorithm in the literature for art painting styles (Costa et al. 2021; Costa 2020), it has many additional advantages over previous work. Indeed, the new explanations are closer to those produced by art specialists. ANYXI automatically parameterizes the logical formulas used to represent the categorizations and incorporates a method to resolve tied classification results. Furthermore, in this paper, we present an exhaustive empirical validation of the ANYXI classifier and conclude that it shows a higher classification accuracy compared to other painting style explainable classifiers existing in the literature. Moreover, its classification performance is similar to that achieved by non-explainable painting classifiers supported by machine learning methods.

As another contribution of this paper, we enrich the analysis of the balance between performance and explainability exhibited by the ANYXI classifier. First, in the context of a survey study, we determine whether the color traits associated with each art style and used to design the classifiers are human-understandable and suitable to classify paintings into the art styles under consideration. In addition, we go deeper into the experimental analysis of ANYXI and compare it versus four different classifiers in terms of accuracy and interpretability metrics. Namely, we will consider a well-known black-box classifier such as random forest (RF) (Breiman 2001) and the white-box decision trees proposed by Quinlan (1986), but also two fuzzy gray-box classifiers. On the one hand, a classifier generated by the Fuzzy Unordered Rule Induction Algorithm (FURIA) (Hühn and Hüllermeier 2009) which is recognized because of being very accurate at the cost of lacking linguistic interpretability. On the other hand, a classifier generated by the GUAJE[5] open source software (Pancho et al. 2013) which is recognized because of facilitating the design of inherently interpretable fuzzy classifiers enriched with global linguistic semantics. Considering that all the classifiers are trained and tested on the same dataset, we extend the discussion by using the rationality criteria introduced by Falomir and Costa (2021) for qualitative comparison of the explanations associated with the different XAI classifiers under consideration in this work.

The rest of the manuscript is organized as follows: Section 2 provides readers with a brief review of the state of

[3] The database is available at https://www.metmuseum.org/art/collection/search#!?searchField=All&showOnly=openAccess&sortBy=relevance&offset=0&pageSize=0.

[4] The Metropolitan Museum of Art Open Access CSV is available at https://github.com/metmuseum/openaccess.

[5] GUAJE stands for Generating Understandable and Accurate Fuzzy Classifiers in a Java Environment and is freely available at https://gitlab.citius.usc.es/jose.alonso/guaje/.

the art. Section 3 introduces the preliminary concepts (i.e., the color model used) and the meaningful features (i.e., the color traits extracted) to be used in the design of the ANYXI classifier. Section 4 presents a survey study to test whether the color traits determining the art styles can be considered human-understandable. Section 5 shows in detail the design of the new explainable classifier, ANYXI, which is validated in Sect. 6. The manuscript concludes with a critical discussion and final remarks in Sect. 7.

2 Related work

In the literature, we find a great diversity of publications dealing with the challenge of categorizing art styles using AI. Jiang et al. (2006) classified traditional Chinese paintings using colors and support vector machines (SVMs). They proposed a scheme to detect traditional Chinese paintings from general images and categorize them into Xieyi (freehand style) and Gongbi (traditional Chinese realistic painting) schools. The related dataset was made up of 9,515 images. In addition, Shen (2009) classified paintings using a radial basis function (RBF) neural network classifier. The related dataset comprised 1,080 digital paintings from 25 different artists. Gatys et al. (2016) presented an artificial neural system that separated image content from style using deep neural networks and allowed recasting one image's content in another image's style. Karayev et al. (2014) applied deep neural networks which were previously trained on object detection for style recognition to classify artworks according to their period. In this way, the authors proved the convenience of transfer learning from traditional photographic domains. As in this paper, they included the Baroque, Impressionism, and Post-Impressionism, and the results showed a general accuracy of around 79%.

Condorovici et al. (2015) presented a fusion scheme based on combining a multilayer perceptron classifier with SVMs. They examined eight art painting styles (Baroque, Cubism, Renaissance, Byzantine Icons, Impressionism, Greek Pottery Paintings, Rococo, and Romanticism) in a dataset with more than 4,000 paintings. Shamir and Tarakhovsky (2012) automated the recognition of nine painters (Monet, Renoir, van Gogh, Rothko, Pollock, Kandinsky, Dalí, Ernst, and de Chirico) and three art styles (Impressionism, Surrealism, and Abstract Expressionism). In the same vein, Shamir (2015) and Burcoff and Shamir (2017) used computational methods to analyze particular painters, specifically Pollock and Picasso, respectively. Siddiquie et al. (2009) used Boost-based SVM, an alternate method to select training data instances using AdaBoost for each of the base kernels, to classify painting styles. In particular, they classified images from the Abstract Expressionist, Baroque, Cubism, Graffiti, Impressionism, and Renaissance. Mensink and Gemert (2014) classified paintings from the Rijksmuseum using Fisher vectors on a dataset with 6,629 artists. Falomir et al. (2015) used qualitative color descriptors and machine learning techniques (k-Nearest Neighbors and SVMs) to categorize painting styles. As in this paper, they considered Baroque, Impressionist, and Post-Impressionist paintings and obtained an accuracy of 75% for a dataset of 70 images. More recently, Falomir et al. (2018) added global features to the qualitative color descriptors and obtained an accuracy of 65% for a dataset containing 252 images from the three mentioned styles. Notice that for further details, Castellano and Vessio (2021) provide readers with an overview of some of the most notable papers investigating the application of deep learning-based approaches to pattern extraction and recognition in visual artworks.

In addition, the literature has already outlined and discussed the main strengths and weaknesses of machine learning methods in the context of art style classification (Anguita et al. 2010). These methods often achieve high accuracy but generally need considerable amounts of training data. All in all, previous work, grounded on machine learning, has proved to be effective in designing accurate classifiers in the context of visual arts. However, intending to integrate such classifiers with decision-support systems, they need to be enriched with reasoning capabilities. Notice that, logical representation and reasoning have already been associated with image interpretation but not with visual art classification. Reiter (1980) applied non-monotonic reasoning for image description. Falomir et al. (2011) used description logic to interpret digital images by describing each object in terms of its color and qualitative shape but also regarding its main spatial features (location, relative orientation, and topology). This way, it is possible the inference of new object categories (e.g., doors) by reasoning. For example, Dasiopoulou et al. (2010) presented a fuzzy description logic-based reasoning framework for reasoning over an extracted description of an outdoor image and handling the underlying vagueness formally. González et al. (2017) and Rubio et al. (2017) applied a general type-2 fuzzy logic method for edge detection to color format images.

It is worth noting that as far as we know none of the previously mentioned research work paid attention to providing users with human-friendly explanations of the reported classification results. Regarding art painting style categorization in the context of XAI, Costa et al. (2018, 2021) presented the ℓ-SHE classifier which integrates qualitative color descriptors and t-norm-based logics for generating explanations associated with art painting style categorization. In this way, ℓ-SHE classifies paintings into the Baroque, Impressionism, and Post-Impressionism. The authors designed three versions of the ℓ-SHE classifier, depending on the logic selected: rational Pavelka logic as well as expansions of Gödel logic and product logic with rational constants. Even if the ℓ-SHE classifier is pioneering in explaining art style classifications, we can remark four general disadvantages: (i) it gets an accuracy rate lower than other art painting style classifiers in the literature (the three versions of the classifier, corresponding to rational Pavelka logic and expansions of Gödel logic and product logic with rational constants, show general accuracy, for the 337 cases under consideration, of around 64%, 53%, and 60%, respectively—a similar approach based on logic aggregators presented by Costa (2020) showed only a bit higher accuracy rates); (ii) the art styles' categorizations are not complete (for example, the definition of Post-Impressionism includes only two color traits and ignores the characteristic contrast between some colors), and hence, in some cases, the explanations provided may be deemed as poor; (iii) it does not integrate a method for breaking tied classification; and (iv) the method for parameterizing the logical formulas used to categorize the styles, which is not defined in an automated fashion, makes it harder to use the algorithm with other datasets.

We will see in Sect. 5 how ANYXI, a new art painting style explainable classifier, overcomes all the four mentioned limitations in the ℓ-SHE classifier. Let us first introduce some preliminary concepts needed to understand later the design and implementation of the new classifier.

3 Preliminary concepts

In this section, we introduce the concepts and dataset to be used later in the design and evaluation of the ANYXI classifier. Section 3.1 describes the three selected art styles (Baroque, Impressionism, and Post-Impressionism), highlighting the colors used in their paintings. Section 3.2 presents an automatic method to extract color names and frequencies from images, in this case, the paintings corresponding to the selected art styles.

3.1 Analyzing color composition of paintings in art styles

The colors in the Baroque, Impressionism, and Post-Impressionism were analyzed as a baseline for this paper. Figure 1 presents some art pieces, their styles, and authors.

The Baroque style started around 1600 in Rome, Italy, and spread to most of Europe. Important Baroque painters are, for example, Rembrandt and Velázquez. Baroque painting style exaggerated lighting, created by contrasting *dark* colors to *light-pale* colors (Rzepińska and Malcharek 1986). Furthermore, the paintings correspond mainly to indoor scenes. They are also characterized by including ferroxide-based *yellows*, *oranges* and *reds* (Hill 1980; Grygar et al. 2003).

The Impressionism started with an exhibition organized in Paris, France, in 1874 by, among others, Claude Monet, Edgar Degas, and Camille Pissarro. Impressionist painters captured the effects of sunlight by painting *en plein air* (outdoors) and produced grays and dark tones by mixing complementary colors. Rather than neutral *white*, *grays*, and *blacks*, Impressionists painters often rendered shadows and highlights in color. Moreover, the development of synthetic pigments provided the artists with vibrant shades of *blue*, *green*, and *yellow* (Mamassian 2008; Powell-Jones 1979). Furthermore, in paintings made *en plein air*, *bright* and *light* colors appear, and shadows are boldly painted with the blue and the grey of the sky as it is reflected onto surfaces (Dewhurst 1908; Berson 1996), giving a sense of freshness previously not represented in painting (blue shadows on snow inspired this technique). In this way, landscapes were also brought up to date with innovative compositions, light effects, and use of color (Samu 2004).

The Post-Impressionism, started in 1910, broke the conventions of the Impressionism to reproduce naturalistic light and color. Significant Post-Impressionist painters are, for example, van Gogh and Seurat. Post-Impressionist painters continued using vivid colors, but they were more inclined to emphasize geometric forms and use unnatural colors. Post-Impressionist paintings are mostly influenced by color contrast, specially *red* vs. *green* and *blue* vs. *yellow* colors (Mamassian 2008). Post-Impressionist painters also used complementary colors to create vibrant contrast and mutual enhancement when juxtaposed and to paint shadows in adjacent objects (using the right hue, the grey color was obtained).[6]

Note that these art style categorizations use intuitive and human-understandable color features. These color traits come from experts who categorize paintings into art styles. Accordingly, in the literature, several datasets are available to study paintings corresponding to these art styles. Here, we select two datasets as benchmarks, the QArt-Dataset (Falomir et al. 2015, 2018) and the Painting-91 (Khan et al. 2014).

On the one hand, the QArt-Dataset contains 90 images: 30 Baroque paintings (15 by Diego Velázquez and 15 by Johannes Vermeer), 30 Impressionist paintings (15 by Pierre-Auguste Renoir and 15 by Claude Monet), and 30 Post-Impressionist paintings (15 by Vincent van Gogh and 15 by Paul Gauguin). See Fig. 2 for an extract from the QArt-Dataset.

On the other hand, the Painting-91 dataset contains 4266 images from 91 different painters, and these authors are cat-

[6] More about complementary colors: https://www.nationalgallery.org.uk/paintings/glossary/complementary-colours (National Gallery).

(a) (b) (c)

The Baroque style: (a) The Maids of Honour; (b) Equestrian Portrait of Prince Balthasar Charles; (c) Girl reading a Letter at an Open Window.

(d) (e) (f)

The Impressionism style: (d) The Basin at Argenteuil; (e) The Saint-Lazare Station; (f) Luncheon of the Boating Party.

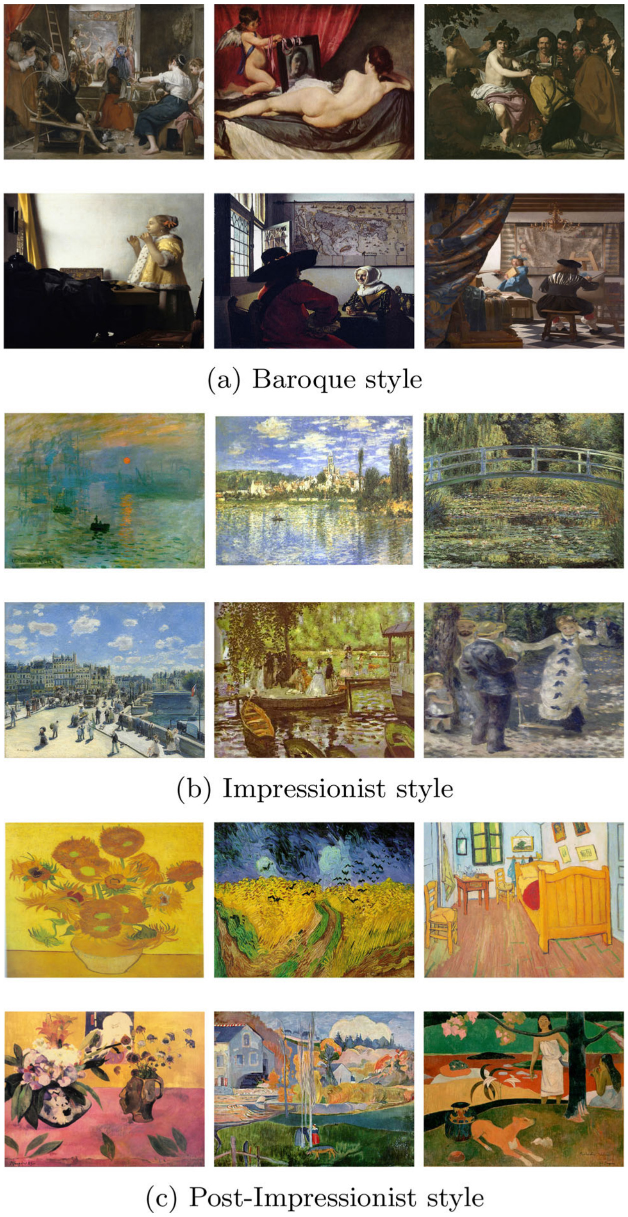
(g) (h) (i)

The Post-Impressionism style: (g) Café Terrace at Night; (h) Thatched Cottages in the Sunshine; (i) Les Alyscamps.

Fig. 1 Paintings corresponding to the Baroque style (authored by Velázquez and Vermeer), Impressionism style (by Monet and Renoir), and Post-Impressionism style (by van Gogh and Gaugain). All rights by Wikimedia Commons, public domain. The color version of this figure is available on the online version of this paper

egorized into 13 art styles.[7] From this dataset, Falomir et al. (2018) selected the paintings of six authors, which resulted in 247 images: 74 for Baroque style (39 by Velázquez and 35 by Vermeer), 82 for Impressionist paintings (46 by Renoir and 36 by Monet) and 91 for the Post-Impressionism (40 by van Gogh and 51 by Gauguin). This smaller dataset is called Painting-91-BIP (Costa et al. 2021). See some illustrative examples of this dataset in Fig. 3.

[7] Painting-91's art styles: Abstract expressionism, Baroque, Constructivism, Cubism, Impressionism, Neoclassical, Pop art, Post-Impressionism, Realism, Renaissance, Romanticism, Surrealism, and Symbolism. For more information, see Khan et al. (2014).



(a) Baroque style

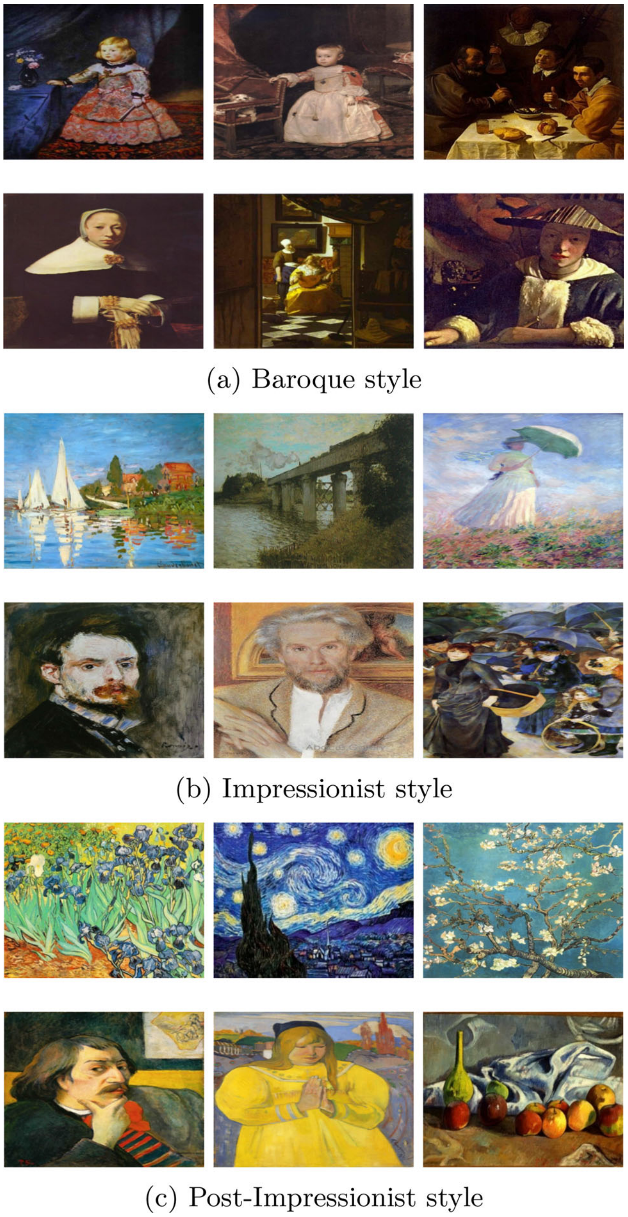
(b) Impressionist style

(c) Post-Impressionist style

Fig. 2 Examples of paintings from the QArt-Dataset. All rights by Wikimedia Commons, public domain. The color version of this figure is available on the online version of this paper

3.2 Automatically extracting color names and frequencies from any painting

The method used for extracting color names and frequencies from the selected paintings is based on the qualitative color descriptor (QCD) introduced by Falomir et al. (2015, 2013). The rest of this section is organized as follows. First, Sect. 3.2.1 presents an overview of the QCD model. Then, Sect. 3.2.2 describes how the color frequencies are extracted from the images in the QArt-Dataset.



(a) Baroque style

(b) Impressionist style

(c) Post-Impressionist style

Fig. 3 Examples of paintings from the Painting-91-BIP dataset. All rights by Wikimedia Commons, public domain. The color version of this figure is available on the online version of this paper

3.2.1 The qualitative color descriptor

The QCD model (Falomir et al. 2015, 2013) defines the Qualitative Color Reference System ($QCRS$), a reference system in the Hue, Saturation and Lightness (HSL) color space for qualitative color description built according to Fig. 4 and defined as:

$$QCRS = \{uH, uS, uL, \mathrm{QC}_{NAME_{1,\ldots,5}}, QC_{INT_{1,\ldots,5}}\},$$

where uH is the unit of Hue, uS is the unit of Saturation, uL is the unit of Lightness; $\mathrm{QC}_{NAME_{1,\ldots,5}}$ refers to the color names,

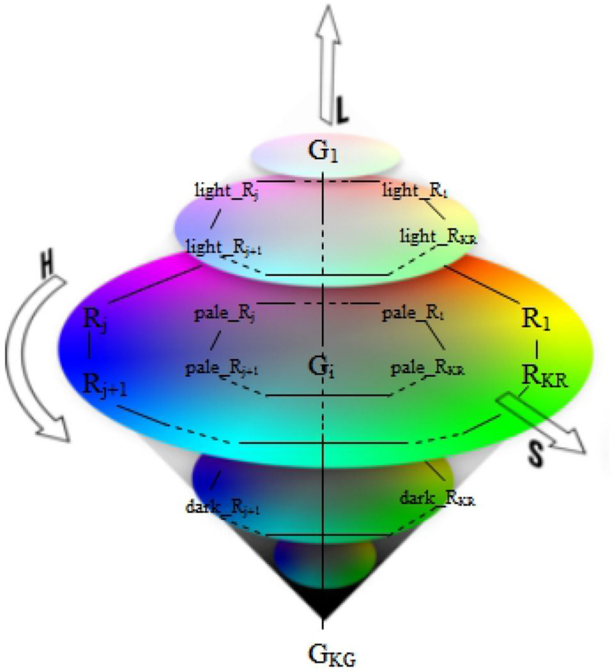


Fig. 4 Diagram for describing QCD: discretization of the HSL color space. The color version of this figure is available on the online version of this paper

and $\mathrm{QC}_{INT_{1,\dots,5}}$ refers to the intervals of HSL coordinates associated with each color. The chosen QC_{NAME} are: QC_{NAME_1} = {*black, dark_grey, grey, light_grey, white*}, QC_{NAME_2} = {*red, orange, yellow, green, turquoise, blue, purple, pink*}, QC_{NAME_3} = {*pale_i*| $i \in \mathrm{QC}_{NAME_2}$}, QC_{NAME_4} = {*light_i*| $i \in \mathrm{QC}_{NAME_2}$}, and QC_{NAME_5} = {*dark_i*| $i \in \mathrm{QC}_{NAME_2}$}.

As a baseline, the $QCRS$ was calibrated according to the vision system used; the chosen QC_{INT} are given in Fig. 4 and Table 1 showing the color HSL values assigned to each color name.

3.2.2 Extracting color frequencies

The QCD model extracts the color names and its corresponding frequencies of any digital image. These color frequencies are associated with the color traits that characterize art styles.

For each image Img, its color histogram is obtained (by applying computer vision techniques in the same vein as proposed by Falomir et al. (2018)), as:

$$f_1(Img), f_2(Img), \dots, f_{37}(Img) \in \mathbb{N}^{37},$$

where $f_i(Img)$ corresponds to the number of pixels labeled as QC_i in Img, where $i \in NAME$. Let $T(Img)$ be the number of pixels in Img, we define the frequency of the color QC_i, $F_i(Img)$, as $f_i(Img)/T(Img)$. Observe that the total number of colors QC_i is 37. Note also that for any image Img, $f_i(Img), F_i(Img) \geq 0$ for $1 \leq i \leq 37$. Then, we transform the color traits in each painting into expressions with the following syntax: $color\_painting(P, QC_i, F_i)$, where P corresponds to the digital image identifier (provided by the chosen dataset), $QC_i \in QC_{NAME1,\dots,5}$, F_i is defined as indicated above and $1 \leq i \leq 37$.

Figure 5 shows an example of a digital image of a painting by Velázquez (it corresponds to the second painting in Baroque style, the *Equestrian Portrait of Prince Balthasar Charles*, in the Fig. 1), which is described by the logical facts containing the image identifier ($v10$ in this case), the qualitative color names in the painting and their corresponding frequencies.

3.3 Art styles defined by qualitative distinctive color traits: logical definition

Section 3.3.1 shows how the logical facts containing color names and frequencies (extracted automatically as explained in the previous section) can be used for defining new expert knowledge about color trait relations by analyzing the darkness level, the hue of the colors, etc.

These color traits allow us to characterize each Baroque, Impressionism and Post-Impressionism art styles (as described in Sect. 3.3.2).

3.3.1 Defining distinctive color traits

In (Def.1 Costa et al. (2021)), the authors extended the QCD model to add the following semantics related to the predicates used by experts when describing the color features of a painting:

$\underline{dark\_colors}$= $\{black, dark\_red, dark\_orange, dark\_yellow, dark\_green, dark\_turquoise, dark\_blue, dark\_purple, dark\_pink, dark\_grey\}$

$\underline{pale\_colors}$= $\{pale\_red, orange, yellow, green, turquoise, blue, purple, pink, grey\}$

$\underline{light\_colors}$ = $\{white, light\_red, orange, yellow, green, turquoise, blue, purple, pink\}$

$\underline{grey\_hue}$= $\{grey, pale\_grey, light\_grey, dark\_grey\}$ (analogously for $\{red\_hue, orange\_hue, yellow\_hue, green\_hue\}$ and $\{turquoise\_hue, blue\_hue, purple\_hue, pink\_hue\}$

$\underline{warm\_hue}$= $\{red\_hue, orange\_hue, yellow\_hue\}$

$\underline{vivid\_colors}$= $\{red, orange, yellow, green, turquoise, blue, purple, pink\}$

$\underline{\text{red\_colors}}$= $\{red, orange, pale\_red, light\_red, dark\_red, pale\_orange, dark\_orange, light\_orange\}$

Table 1 The Hue Saturation and Lightness (HSL) intervals corresponding to color names (QC_{INT}) defined as by Sanz et al. (2015)

	Color	UH	US	UL
QC_{LAB_1}	*black*	[0, 360]	[0, min{20, $2UL$, $200 - 2UL$}]	(0, 20]
	dark_grey			(20, 40]
	grey			(40, 60]
	light_grey			(60, 80]
	white			(80, 100]
QC_{LAB_2}	*red*	(335, 360] ∧ [0, 20]	(50, min{100, $2UL$, $200 - 2UL$}]	(40, 60]
	orange	(20, 50]		
	yellow	(50, 80]		
	green	(80, 160]		
	turquoise/cyan	(160, 200]		
	blue	(200, 239]		
	purple	(239, 297]		
	pink/magenta	(297, 335]		
QC_{LAB_3}	$pale\_QC_{LAB_2}$	Idem	(20, 50]	(40, 60]
QC_{LAB_3}	$light\_QC_{LAB_2}$	Idem	(20, $200 - 2UL$]	(60, 90]
QC_{LAB_3}	$dark\_QC_{LAB_2}$	Idem	(20, $2UL$]	(10, 40]

```
colour_painting(v10, black, 0.362).
colour_painting(v10, dark_blue, 0.0002).
colour_painting(v10, dark_green, 0.025).
colour_painting(v10, dark_grey, 0.117).
colour_painting(v10, dark_orange, 0.022).
colour_painting(v10, dark_pink, 0.001).
colour_painting(v10, dark_red, 0.078).
colour_painting(v10, dark_turquoise, 0.056).
colour_painting(v10, dark_yellow, 0.006).
colour_painting(v10, grey, 0.060).
colour_painting(v10, light_green, 0.014).
colour_painting(v10, light_grey, 0.054).
colour_painting(v10, light_orange, 0.010).
```

```
colour_painting(v10, light_red, 0.003).
colour_painting(v10, light_turquoise, 0.006).
colour_painting(v10, light_yellow, 0.007).
colour_painting(v10, orange, 0.009).
colour_painting(v10, pale_green, 0.046).
colour_painting(v10, pale_orange, 0.017).
colour_painting(v10, pale_red, 0.007).
colour_painting(v10, pale_turquoise, 0.058).
colour_painting(v10, pale_yellow, 0.0128).
colour_painting(v10, red, 0.007).
colour_painting(v10, turquoise, 0.0004).
colour_painting(v10, white, 0.021).
```

Fig. 5 This is an example of the logical facts extracted with color names and color frequencies corresponding to the painting by Velázquez the *Equestrian Portrait of Prince Balthasar Charles*. This image identifier is $v10$ in the QArt-Dataset. Note that the use of Britain English is limited to the notation of the logical facts to be coherent with the original notation presented in Costa et al. (2021). The color version of this image is available on the online version of this paper

3.3.2 Characterizing art styles using color traits

A representation of the characteristic color traits of the Baroque, Impressionism, and Post-Impressionism styles using fuzzy sets is proposed. In this paper, we extend the proposal in (Costa et al. 2021) and add three color traits to balance the number of features of the categorizations and make them closer to art experts' opinions. As a result, in this section, twelve propositional variables are introduced to represent some of the color traits used by art experts to describe the color of a painting. The process of assigning each group of four variables to a class is determined by the art styles' definitions proposed by the specialists and presented in Sect. 3.1: for each distinctive color feature of each art style, a propositional variable is defined and three disjoint sets of for variables each one are obtained. These sets of variables are disjoint since the art styles selected do not share these main characteristic color traits,

First, considering the color traits outlined by the art specialists shown in Sect. 3.1 and the extension of the QCD model presented above, the following distinctive color features for the Baroque style are proposed:

$\underline{darkness\_level}$: the accumulative sum of the frequencies of $dark\_colors$.
$\underline{no\_paleness\_level}$: the total frequency of colors that are not $pale\_colors$.
$\underline{contrast\_level}$: the total frequency of $dark$ and $pale$ colors bounded to 1.
$\underline{red\_colors}$: the relation between the amount of red_colors in a painting, and the total number of qualitative colors (QCs) in the painting.

Regarding the Impressionism style, we follow again the definition proposed by the art experts (shown in Sect. 3.1), and the following characteristic color features are suggested:

$\underline{bluish\_level}$: the total frequency of the QCs extracted as having $bluehue$.
$\underline{greyish\_level}$: the total frequency of the QCs extracted as having $greyhue$.
$\underline{diversity\_of\_hues}$: all the QCs in a painting are grouped according to their hues and they are related to the total number of hues in QCD, which is 11 ($card(vivid\_colors \cup \{black, white\}) = 11$).
$\underline{diversity\_of\_qcds}$: the relation between the amount of qualitative colors (including all their pale-, light-, and dark- variants) in a painting, and the total number of QCs possible (i.e., 37).

Finally, we use proceed analogously, and these four color traits for Post-Impressionism are proposed:

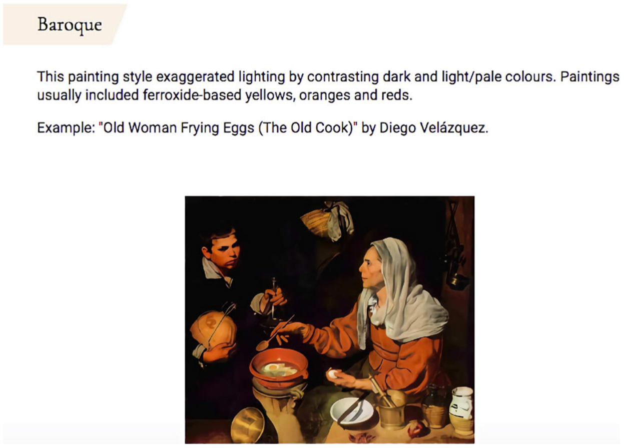
$\underline{vividness\_level}$: the total frequency of the QCs extracted as having pure hue.
$\underline{warm\_colors\_level}$: the total frequency of the QCs extracted as having $warm\ hue$.
$\underline{contrast\_blue\_yellow\_level}$: the total frequency of $bluish$ and $yellowish$ (i.e., the total frequency of $yellow$, $pale\_yellow$ and $dark\_yellow$ colors bounded to 1). 1
$\underline{contrast\_red\_green\_level}$: the total frequency of $reddish$ and $green$ colors bounded to 1.

From each painting in the dataset considered in this paper, we obtain the corresponding color names and frequencies appearing in the painting plus the values corresponding to the twelve color traits defined above.

4 A survey on art style painting classification based on color traits

In order to test whether the above-mentioned color traits are human-understandable and sufficient to distinguish paintings belonging to Baroque, Impressionism, and Post -Impressionism, a survey was designed which provides a reading definition describing colors in each style.

This survey was conducted online, using the Google Forms platform. First, participants were given instructions about how to fulfill the survey. They were provided with an explanation of the different color traits for each art style together with a displayed example (see Fig. 6; for simplifying, the images in this section focus on the Baroque style). Throughout the survey, the narrative art style color descriptions were available to participants. The participants' task consisted of looking at an image of a painting and deciding which one of the three art style descriptions corresponds to the painting (see Fig. 7). After each question, participants were asked whether they had seen the enquired painting before or not. In case of a positive answer, this response was not considered in our study (note that a participant may know the painter and even the style in advance, while the purpose of this study was to identify the style only by the color traits, not taking into account previous knowledge).



Baroque

This painting style exaggerated lighting by contrasting dark and light/pale colours. Paintings usually included ferroxide-based yellows, oranges and reds.

Example: "Old Woman Frying Eggs (The Old Cook)" by Diego Velázquez.

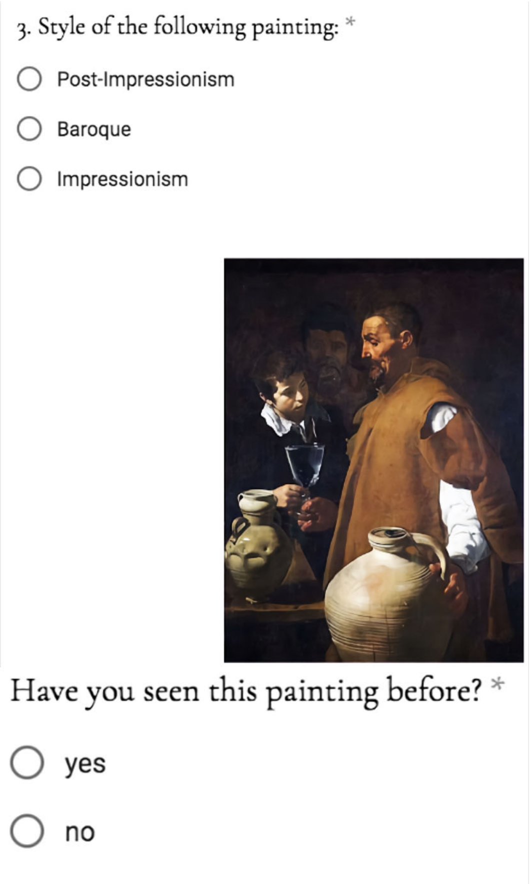
Fig. 6 Descriptions of the Baroque style and example yielded to participants

We randomly assigned participants into 6 groups. Each group consisted of 15 questions so that no survey took much longer than 15 min. In total, we gathered responses corresponding to the 90 paintings in the QArt-Dataset (Falomir et al. 2015, 2018). And we also gathered participants' age, art interests and knowledge, and language. There was no economic retribution for answering the survey.

A total of 150 participants took the survey (53.3% women and 46.7% men). The study aimed to decide the art style of the paintings based on the colors, not on previous knowledge. We discarded those participants who knew in advance four or more paintings in the survey. 87.8% of the final selected participants had seen at most one of the paintings in the survey before. Their answers corresponding to the known paintings were discarded.

The distribution corresponding to participants' age and native language is shown in Fig. 8. The survey was deployed in English and Spanish so participants could choose the more advisable language for them. All of the participants had a good command of at least one of the two languages.

In addition, participants were asked about their knowledge in art from which we defined four categories: no academic knowledge of art (*None*), they studied art at high school (*High school*), at university (*University*) or as a part of a Ph.D. (*Ph.D.*). The results obtained revealed that 40% of them have some relevant background in art. Participants were also asked



3. Style of the following painting: *

○ Post-Impressionism

○ Baroque

○ Impressionism

Have you seen this painting before? *

○ yes

○ no

Fig. 7 A question in the survey. The color version of this figure is available on the online version of this paper

about their interest in art: *art lover* (two or more hobbies related to art), *interested in art* (one hobby related to art), and *disregarding art* (any hobby related to art). Although 60% of participants did not have any academic knowledge of art, 75% of participants were interested in art.

Details are depicted in Fig. 9.

After filtering out those participants who had seen some paintings before, a total of 60 participants were taken for simulating a tenfold cross-validation experiment. We took 10 participants for each of the six groups (45% women and 55% men).

This tenfold cross-validation will allow us to compare the survey outcomes to the results obtained later by the automatic classification algorithms considered in this paper. In this way, we considered 10 surveys of the paintings and computed all the corresponding scores (see Table 2 versus Table 3, which will be introduced later in Sect. 6). On average, the ratio of correctly classified instances by the participants in the survey (RCCI in Table 2) is 78.01%, and the average of the F-measure for the three styles considered is 80%, similar to the RCCI. The F-Measure formula is indicated in

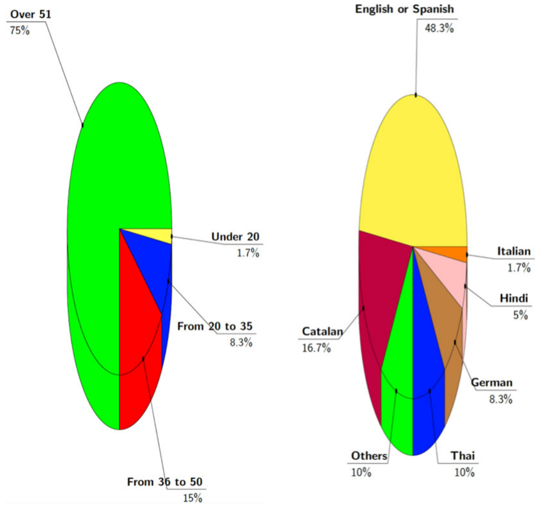


Fig. 8 Participation in the survey, in %, regarding participants' age (pie chart on the left) and native language (pie chart on the right). The expression *Others* refers to those languages with a representation of less than 1.2%, such as Serbian, Polish, Greek, or Papiamento

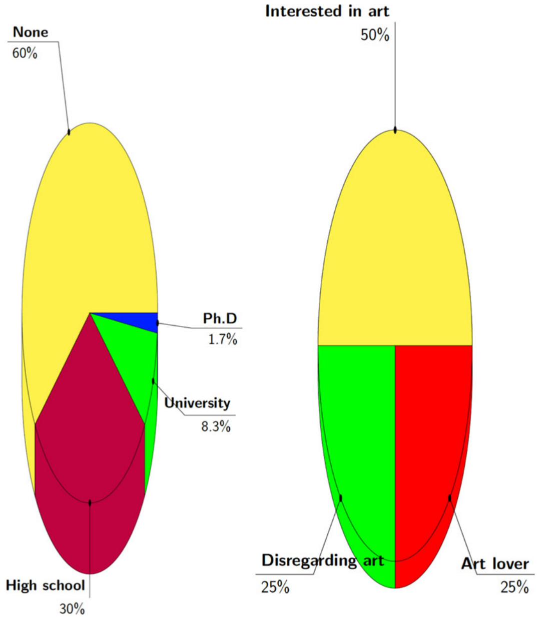


Fig. 9 Participation in the survey, in %, regarding participants' academic knowledge of art (pie chart on the left) and interest in art (pie chart on the right)

Sect. 6.1. In Table 6.1, F_{Bar} refers to the F-Measure obtained for the Baroque paintings of each Test, F_{Imp} to the F-Measure obtained for the Impressionist artworks of each Test, and F_{Pos} to the F-Measure obtained for the Post-Impressionist paintings of each Test. In the survey, the Baroque art style emerges as the simplest style to classify since the metrics associated, F_{Bar} in Table 2, shows the most accurate results. The Impressionist style, in contrast, is the most challenging style to classify (pay attention to the lower values of F_{Imp} in the table).

5 The ANYXI classifier: a human-centered design

This section presents the logic art painting style explainable classifier named ANYXI. There are two general motivations for introducing this algorithm. On the one hand, as stated in the introduction and the related work sections, ANYXI overcomes the flaws in ℓ-SHE regarding the classification rate, the parameterization of the evaluated Horn clauses used to categorize the art painting styles, the tied classification results, and the difference in the explanations related to each style. On the other hand, the painting style categorizations used by ANYXI are more exhaustive and similar to art experts' definitions than those presented by the ℓ-SHE classifier. In addition, the definitions of art styles utilized by ANYXI are very close to those used in the survey introduced in Sect. 4, so we could hypothesize that they are a plausible way to explain to people the color traits which are characteristic of the art styles under consideration.

5.1 A propositional fuzzy language for knowledge representation in art style categorization

In this section, we introduce a propositional language extended with rational truth constants, and use it for categorizing the three art styles considered in this paper.

The twelve color traits introduced in Sect. 3.3.1 are commonly used by the experts to explain paintings, and thus can help to foster human-computer interaction, as validated in the survey presented in Sect. 4, and for yielding explanations of the classification results. Furthermore, we can naturally view those distinctive color features as fuzzy notions. In this way, following (Costa et al. 2021), we introduce a propositional variable for each of these color traits and, using a propositional fuzzy language expanded with truth-constants, propose one evaluated Horn clause for categorizing each painting style considered. Before introducing these evaluated propositional variables, let us recall the syntax and semantics of this formal language.

Syntax and semantics of a continuous t-norm-based propositional fuzzy logic (Chapter I, Definition 1.1.13 Cintula et al. (2011)). A language of continuous t-norm based propositional fuzzy logic contains a set of propositional variables Var, the binary connectives in the set $\{\rightarrow, \&, \wedge, \vee, \leftrightarrow\}$, the unary connective $\neg$, and the truth-constants $\overline{0}, \overline{1}$. Let $[0, 1] \subseteq \mathbb{R}$, where $\mathbb{R}$ denotes the set of real numbers, a $[0, 1]$-*evaluation* e is a mapping $e : Var \rightarrow [0, 1]$. Given a continuous t-norm $*$, an evaluation e extends uniquely to an evaluation e^* of the set of well-formed formulas as usual. For each rational number $r \in [0, 1]$, we consider the truth-constant $\overline{r}$ so that $e^*(\overline{r}) = r$.

Let φ, ψ be two formulas, we recall the interpretation of & and → in rational Pavelka logic (RPL for short) and product logic expanded with rational constants ($\sqcap(\mathbb{Q})$ for short), respectively:

$$(\underline{\text{RPL}})\ e(\varphi\&\psi) = max\{0, e(\varphi) + e(\psi) - 1\}, \text{ and } e(\varphi \rightarrow \psi) = min\{1 - e(\varphi) + e(\psi), 1\}.$$

$$(\underline{\sqcap(\mathbb{Q})})\ e(\varphi\&\psi) = e(\varphi) \cdot e(\psi), \text{ and}$$

$$e(\varphi \rightarrow \psi) = \begin{cases} 1 & \text{if } e(\varphi) \leq e(\psi) \\ \frac{e(\psi)}{e(\varphi)}, & \text{otherwise.} \end{cases}$$

At this point, we consider the following propositional variables referring to the color traits defined in Sect. 3.3.2:

$$darkness\_level, no\_paleness\_level, \\ contrast\_level, red\_colors, bluish\_level, \\ greyish\_level, diversity\_of\_hues, \\ diversity\_of\_qcds, vividness\_level, \\ warm\_colors\_level, \\ contrast\_blue\_yellow\_level, \quad contrast\_red \\ {}_green\_level;$$

and consider also the following propositional variables referring to the styles Baroque, Impressionism, and Post-Impressionism, respectively: $baroque, impressionism,$ $postimpressionism$. Then, we define the language for knowledge representation in art style categorization, as the language of the logics RPL and $\sqcap(\mathbb{Q})$, with exactly these fifteen propositional variables, the connectives & and →, and one rational truth constant for each rational number r in $[0, 1]$.

The evaluated Horn clauses we consider for representing the categorizations of the styles are a generalization of the notion of the RPL∀-Horn clause introduced in Costa and Dellunde (2017).

Definition 1 (Evaluated Horn clause (Definition 10 Costa and Dellunde 2017)) An *atomic evaluated formula* (φ, r) is defined as $\overline{r} \rightarrow \varphi$, where $r \in [0, 1]$ is a rational number and φ is an atomic formula without truth constants apart from $\overline{0}$

Table 2 Results of the preliminary survey. RCCI is the ratio of correctly classified instances and F is the F-Measure. RCCI and F are formally defined in Sect. 6.1. Std stands for standard deviation

Test	RCCI		F		F_{Bar}		F_{Imp}		F_{Pos}	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Test 0	80.00	12.61	0.80	0.13	0.94	0.11	0.68	0.04	0.75	0.03
Test 1	86.67	12.61	0.87	0.13	0.99	0.05	0.79	0.04	0.80	0.04
Test 2	77.78	7.41	0.78	0.08	0.84	0.13	0.65	0.03	0.81	0.01
Test 3	71.11	15.89	0.82	0.16	0.94	0.11	0.74	0.06	0.74	0.05
Test 4	74.44	11.77	0.85	0.12	0.87	0.14	0.82	0.15	0.84	0.17
Test 5	76.67	15.23	0.77	0.15	0.85	0.19	0.68	0.04	0.72	0.19
Test 6	66.67	10.48	0.67	0.11	0.72	0.22	0.59	0.14	0.68	0.09
Test 7	**87.78**	15.23	**0.88**	0.15	**0.96**	0.01	0.82	0.06	0.85	0.14
Test 8	83.33	7.86	0.83	0.07	0.83	0.10	0.77	0.01	0.89	0.11
Test 9	75.56	10.21	0.76	0.10	0.68	0.05	0.71	0.02	0.83	0.15
Mean	**78.01**	**11.93**	**0.80**	**0.12**	**0.86**	**0.11**	**0.73**	**0.06**	**0.79**	**0.10**
Std	**6.63**	**2.99**	**0.06**	**0.03**	**0.10**	**0.06**	**0.08**	**0.05**	**0.07**	**0.06**

Table 3 Reported mean and standard deviation (Std) for all considered accuracy metrics over tenfold cross-validation for the QArt-337 dataset

Algorithm	RCCI		F		F_{Bar}		F_{Imp}		F_{Pos}	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
RF	**71.21**	7.92	**0.71**	0.08	0.77	0.09	**0.64**	0.12	**0.72**	0.09
J48	61.99	8.28	0.62	0.09	0.71	0.10	0.51	0.13	0.63	0.09
FURIA	67.05	6.19	0.66	0.06	0.77	0.08	0.55	0.08	0.68	0.06
GUAJE	66.18	6.56	0.65	0.07	**0.78**	0.06	0.55	0.10	0.64	0.13
ANYXI-1-RPL	**70.58**	7.89	**0.71**	0.08	**0.82**	0.10	0.63	0.11	**0.68**	0.09
ANYXI-2-RPL	69.66	8.45	0.70	0.09	0.78	0.09	**0.67**	0.17	0.68	0.11
ANYXI-3-RPL	66.72	7.37	0.68	0.08	0.77	0.09	0.60	0.10	0.67	0.11
ANYXI-1-$\sqcap(\mathbb{Q})$	62.60	6.17	0.64	0.01	0.70	0.08	0.56	0.13	0.63	0.13
ANYXI-2-$\sqcap(\mathbb{Q})$	62.90	6.65	0.64	0.09	0.71	0.09	0.57	0.14	0.64	0.14
ANYXI-3-$\sqcap(\mathbb{Q})$	67.61	9.10	0.68	0.10	0.75	0.12	0.62	0.09	0.64	0.12

In the case of F-Measure, we report the value for each single class (F_{ci}) and the averaged one (F)

and $\overline{1}$. An *evaluated Horn clause* has the form

$$(\varphi_1, r_1) \& \ldots \& (\varphi_n, r_n) \to (\varphi, s),$$

where $(\varphi_1, r_1), \ldots, (\varphi_n, r_n)$ and (φ, s) are atomic evaluated formulas.

We use the analysis of the color compositions of the art styles presented in Sect. 3.1 and propose the following evaluated Horn clauses to categorize the three art styles considered. The evaluated Horn clause H_1 categorizes the Baroque style:

$$(darkness\_level, r_1)\&(no\_paleness\_level, r_2)\& \\ (contrast\_level, r_3)\&(red\_colors, r_4) \to (baroque, 1).$$

The evaluated Horn clause H_2 represents the Impressionist style:

$$(diversity\_of\_qcds, r_5)\&(diversity\_of\_hues, r_6)\& \\ (bluish\_level, r_7)\&(greyish\_level, r_8) \to \\ (impressionism, 1).$$

And the evaluated Horn clause H_3 categorizes the Post-Impressionist style:

$$(vividness\_level, r_9)\&(warm\_colors\_level, r_{10}) \\ \&(contrast\_blue\_yellow\_level, r_{11}) \\ \&(contrast\_red\_green\_level, r_{12}) \to \\ (postimpressionism, 1).$$

These evaluated Horn clauses express the knowledge from art specialists we have about distinctive color traits of each style under consideration and differ from those clauses presented in (Costa et al. 2021) in the number of color traits considered and their parameterization (i.e., the obtaining of the parameters $r_1, r_2, \ldots, r_{12}$, which is not fixed now and depends on the training dataset to be considered by ANYXI).

Fig. 10 Paintings from the QArt-Dataset used to exemplify a $Training$ set. All rights by Wikimedia Commons, public domain. The color version of this figure is available on the online version of this paper

5.2 Parameterizing the evaluated Horn clauses

We consider three parameterizations of the evaluated Horn clauses categorizing the art styles, using three different measures: the arithmetic mean, median, and geometric mean. In this section, we briefly explain the method only for the arithmetic mean since the remaining are performed analogously.

Let f_{mean} be the function that yields the arithmetic mean of at least two real numbers and $f_{mean} = id$ whether the function has a single argument.

For each experimentation, let $Training$ be the set of paintings randomly selected for training the algorithm. An example of $Training$ could be $\{v10, rn12, m5, gg9\}$), i.e., the set formed by the Baroque painting *Equestrian Portrait of Prince Balthasar Charles* ($v10$), the Impressionist paintings *Doge's Palace, Venice* ($rm10$) and *Le Pont routier, Argenteuil* ($m5$), and the Post-Impressionist painting *Petit breton à l'oie* ($gg9$).

Although several conditions might be added to the set $Training$ (e.g., a balanced appearance of representative elements), we only set one condition: $Training$ must contain at least one element of each art style (so, $card(Training) \geq 3$). Then, the rational truth constants of the evaluated Horn clauses are obtained as follows:

$r_1 = f_{mean}(e_{p_i}(darkness\_level))$, for each i such that $p_i \in Training$ and p_i is a Baroque painting.
$r_2 = f_{mean}(e_{p_i}(no\_paleness\_level))$, for each i such that $p_i \in Training$ and p_i is a Baroque painting.
$(\ldots)$
$r_5 = f_{mean}(e_{p_i}(diversity\_of\_qcds))$, for each i such that $p_i \in Training$ and p_i is an Impressionist painting.
$(\ldots)$
$r_{12} = f_{mean}(e_{p_i}(contrast\_red\_green\_level))$, for each i such that $p_i \in Training$ and p_i is a Post-Impressionist painting.

Whenever it is necessary, we indicate the selected measure in the parameter's subindex (e.g., $r_{1:median}, r_{7:gmean}, r_{10:mean}$). Furthermore, whenever the parameters $r_1, r_2, \ldots, r_{12}$ are obtained using the arithmetic mean, we will refer to ANYXI as ANYXI-1; as ANYXI-2 if the metric used is the median, and as ANYXI-3 whenever the parameters are obtained using the geometric mean.

5.3 Classifying the paintings

In this section, we use the semantics of the two fuzzy logics considered, RPL and $\sqcap(\mathbb{Q})$, to define the classification function of ANYXI.

First, note that for any digital painting p, we can associate an evaluation e_p of the twelve variables in the antecedent of the Horn clauses H_1, H_2, H_3 in the following way. Let us consider, for example, painting $v10$ and $\sqcap(\mathbb{Q})$, then we obtain that:

$$e_{v10}(contrast\_level, r_3) = min\left\{\frac{e_{v10}(contrast\_level)}{r_3}, 1\right\}$$

(observe that the value $e_{v10}(contrast\_level)$ is determined by the color frequencies in $v10$ and the definition of the $contrast\_level$ presented in Sect. 3.3.1).

We define now the membership degrees to the Baroque, Impressionism, and Post-Impressionism.

Definition 2 Given a painting p, the membership degree of p for the Baroque, denoted by $B(p)$, is defined as the evaluation, by e_p, of the antecedent of the evaluated Horn clause H_1, that is:

$$e_p((darkness\_level, r_1)\&(no\_paleness\_level, r_2)\&(contrast\_level, r_3)\&(red\_colors, r_4)).$$

The membership degrees for Impressionism and Post-Impressionism ($I(p$ and $PI(p)$, respectively) are defined analogously.

In this way, we remark that the evaluated Horn clauses proposed in Sect. 5.1 are not only used to represent the knowledge about the art styles but also to obtain a membership degree for each painting style considered and, thus, to classify the paintings.

Furthermore, note that the membership degrees for the three art styles (i.e., the interpretation of the antecedent of the evaluated Horn clauses) depends on the semantics of the logic selected. In this paper, we consider two different logics, RPL and $\sqcap(\mathbb{Q})$, and then, we obtain two versions of the

ANYXI classifier according to the chosen logic. In this way, we consider, in total, six versions of ANYXI:
ANYXI-1-RPL, ANYXI-2-RPL, ANYXI-3-RPL,
ANYXI-1-$\sqcap(\mathbb{Q})$, ANYXI-2-$\sqcap(\mathbb{Q})$, and ANYXI-3-$\sqcap(\mathbb{Q})$.

To finish this section, let us exemplify with more details how to obtain the membership degree for an art style, given a painting p. For the sake of clarity, let us first introduce some notation:

$$\begin{aligned}
&B_1(p) = e_p(darkness\_level, r_1),\\
&B_2(p) = e_p(no\_paleness\_level, r_2),\\
&B_3(p) = e_p(contrast\_level, r_3),\\
&B_4(p) = e_p(red\_colors, r_4),\\
&I_1(p) = e_p(diversity\_of\_qcds, r_5),\\
&I_2(p) = e_p(diversity\_of\_hues, r_6),\\
&I_3(p) = e_p(bluish\_level, r_7),\\
&I_4(p) = e_p(greyish\_level, r_8),\\
&PI_1(p) = e_p(vividness\_level, r_9),\\
&PI_2(p) = e_p(warm\_colors\_level, r_{10}),\\
&PI_3(p) = e_p(contrast\_blue\_yellow\_level, r_{11}),\\
&PI_4(p) = e_p(contrast\_red\_green\_level, r_{12}).
\end{aligned}$$

And let us focus on the ANYXI-1-RPL version. Then,

$$B(p) = max\{0, B_1(p) + B_2(p) + B_3(p) + B_4(p) - 3\}$$

(i.e., $B(p) = e_p((darkness\_level, r_1)\&$
$(no\_paleness\_level, r_2)\&(contrast\_level, r_3)\&$
$(red\_colors, r_4))$,

$$\begin{aligned}
&I(p) = max\{0, I_1(p) + I_2(p) + I_3(p) + I_4(p) - 3\}, \text{ and}\\
&PI = max\{0, PI_1(p) + PI_2(p) + PI_3(p) + PI_4(p) - 3\}.
\end{aligned}$$

In contrast, observe that for ANYXI-1-$\sqcap(\mathbb{Q})$, we would obtain:

$$\begin{aligned}
B(p) &= B_1(p) \cdot B_2(p) \cdot B_3(p) \cdot B_4(p)\\
I(p) &= I_1(p) \cdot I_2(p) \cdot I_3(p) \cdot I_4(p), \text{ and}\\
PI &= PI_1(p) \cdot PI_2(p) \cdot PI_3(p) \cdot PI_4(p).
\end{aligned}$$

Finally, let us remark that a painting p could present an event of a tied membership degree in the following forms: $B(p) \leq I(p) = PI(p)$, $I(p) \leq B(p) = PI(p)$, $PI(p) \leq B(p) = I(p)$ or even $B(p) = I(p) = PI(p)$. Section 5.5 explains how the ANYXI classifier tackles this problem. But, before, we need to consider a method for generating explanations.

5.4 Generating explanations of the classification results

The use of fuzzy notions helps us interpret the classifier designed, ANYXI, so that explaining its classifications becomes attainable. So, in this section, we define the functions for generating the explanations of the classification results yielded by ANYXI.

The ANYXI classifier yields explanations in the following way. For a painting p, $baroque\_explanations(p)$ is defined as:

If $B_1(p) \geq f_{mean}(B_1(p_i))$, where i is such that $p_i \in Train$ and p_i is a Baroque painting, then "The darkness evidences the Baroque style".
If $B_2(p) \geq f_{mean}(B_2(p_i))$, where i is such that $p_i \in Train$ and p_i is a Baroque painting, then "The contrast of dark and pale colors evidences the Baroque style".
If $B_3(p) \geq f_{mean}(B_3(p_i))$, where i is such that $p_i \in Train$ and p_i is a Baroque painting, then "The lack of pale colors evidences the Baroque style".
If $B_4(p) \geq f_{mean}(B_4(p_i))$, where i is such that $p_i \in Train$ and p_i is a Baroque painting, then "The level of reddish evidences the Baroque style".

The $impressionism\_explanations(p)$ is defined as:

If $I_1(p) \geq f_{mean}(I_1(p_i))$, where i is such that $p_i \in Train$ and p_i is an Impressionist painting, then "The diversity of qualitative colors evidences the Impressionist style".
If $I_2(p) \geq f_{mean}(I_2(p_i))$, where i is such that $p_i \in Train$ and p_i is an Impressionist painting, then "The variety of hues evidences the Impressionist style".
If $I_3(p) \geq f_{mean}(I_3(p_i))$, where i is such that $p_i \in Train$ and p_i is an Impressionist painting, then "The amount of bluish evidences the Impressionist style".
If $I_4(p) \geq f_{mean}(I_4(p_i))$, where i is such that $p_i \in Train$ and p_i is an Impressionist painting, then "The amount of grey evidences the Impressionist style".

And the $postimpressionism\_explanations(p)$ is defined as:

If $PI_1(p) \geq f_{mean}(PI_1(p_i))$, where i is such that $p_i \in Train$ and p_i is a Post-Impressionist painting, then "The presence of vivid colors evidences the Post-Impressionist style".
If $PI_2(p) \geq f_{mean}(PI_2(p_i))$, where i is such that $p_i \in Train$ and p_i is a Post-Impressionist painting, then "The high level of warm colors evidences the Post-Impressionist style".
If $PI_3(p) \geq f_{mean}(PI_3(p_i))$, where i is such that $p_i \in Train$ and p_i is a Post-Impressionist painting, then "The contrast between blue and yellow evidences the Post-Impressionist style".
If $PI_4(p) \geq f_{mean}(PI_4(p_i))$, where i is such that $p_i \in Train$ and p_i is a Post-Impressionist painting, then

"The contrast between red and green evidences the Post-Impressionist style".

As shown in this section, the explanations generated by ANYXI are fairly simple: they only regard color information and concern the influence of each distinctive color feature of the selected art style. Other approaches for classification tasks, as, for instance, generated by SOTA deep learning methods, also include visual and contextual traits making the explanations richer. The motivations and advantages of choosing this simple strategy, however, are twofold, the first being the most relevant. On the one hand, this approach follows the cognitive hypothesis that color features are enough for classifying paintings from the art styles chosen. Indeed, as shown in Sect. 4, people correctly classify artworks from these styles using these color-based categorizations without needing additional features. In this way, we believe the human-centered design of ANYXI does not need to add any additional information. On the other hand, this simple method avoids hard computation. Nevertheless, it is worth mentioning that this method's simplicity and cognitive hypothesis are not guaranteed in the case of adding new art styles, which could be not completely categorized in terms of color traits.

5.5 Solving tied classification results

An important limitation of the ℓ-SHE classifier (Costa et al. 2021) is the lack of a protocol for solving tied classification results. Indeed, main visual databases on art painting (e.g., WikiArt or Art500k) do not present *double* classifications of the type *The painting belongs to the Baroque and the Impressionism*. The ANYXI classifier, human-centered designed, solves this situation. However, there may be doubts about the classification result, or perhaps a painting may join traits from different styles, and, in these cases, it would be convenient advising users of the borderline classification. This improvement of the ANYXI design is left for future work.

The ANYXI method solves the case of tied membership degrees by considering the number of explanations relevant to the classification, the quantitative difference between the evaluation of the color traits in the painting and those color features related to the corresponding style, and the levels of warm and red colors. More specifically, in the event of a tied membership degree, the following functions are used to determine the classification. For a painting p, we define the following three functions, related to each art style considered.

Regarding the Baroque style,

$$\begin{aligned} tiesolver_b(p) = & \frac{7}{20} explanations_b(p) \\ & + \frac{2}{10} warmth(p) + \frac{9}{20} difference_b(p), \end{aligned} \tag{1}$$

where $explanations_b(p)$ indicates the number of explanations that would be yielded in p was classified into the Baroque; the functions $warmcolorlevel(p)$ and $difference_b(p)$ are defined as follows:

$$\begin{aligned} & warmth(p) \\ & \quad = \begin{cases} 1 & \text{if } warm\_colors\_level(p) \leq 0.32 \text{ or} \\ & (red\_colors(p) \geq 0.4 \text{ and} \\ & warm\_colors\_level(p) \geq 0.32) \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{2}$$

and

$$\begin{aligned} difference_b(p) = \sum_{i=1, j \in J}^{12} = (3 \cdot e_p(j) - r_{i:mean} \\ -r_{i:median} - r_{i:gmean}), \end{aligned} \tag{3}$$

where J is the set containing the twelve propositional variables related to the color traits introduced in Sect. 3.3.2, that is, $J = \{darkness\_level, \ldots, contrast\_red\_green\_level\}$.

The thresholds 0.32 and 0.4 from $warmth(p)$ were obtained from experimental analysis of the paintings in the QArt-Dataset. Indeed, the experimental analysis showed these thresholds related to the two color features $warm\_colors\_level$ and red_colors were efficient for classifying paintings from the three art styles selected.

With respect to the Impressionist and Post-Impressionist styles, we define

$$\begin{aligned} tiesolver_i(p) = & \frac{7}{20} explanations_i(p) \\ & + \frac{2}{10} warmth(p) + \frac{9}{20} difference_i(p), \text{ and} \\ tiesolver_{pi}(p) = & \frac{7}{20} explanations_{pi}(p) \\ & + \frac{2}{10} warmth(p) + \frac{9}{20} difference_{pi}(p), \end{aligned} \tag{4}$$

where $explanations_i(p)$, $explanations_{pi}(p)$, and $difference_i(p)$, $difference_{pi}(p)$ are defined analogously.

In this way, in the event of a tied membership degree, ANYXI classifies according to the maximum value of the $tiesolver$ functions of the styles whose membership degrees are tied.

Considering the definitions of the membership degrees and the *tiesolver* functions, classifying a painting becomes trivial. Let us define max so that the codomain is a multiset in the case some values in the argument are repeated. Let 1 represent the Baroque, 2 the Impressionism, 3 the Post-Impressionism and, for any p, let $C_1 = B(p)$, $C_2 = I(p)$, $C_3 = PI(p)$ and let us denote $max_{j \in \{1,2,3\}}\{C_j\}$ by A. The rules for the events of a tied membership degree (i.e., A is not a singleton set) are shown next:

- If the cardinality of A is 2, say for instance $A = \{C_1, C_2\}$, $max\{tiesolver_1(p), tiesolver_2(p)\}$ determines the classification.
- If the cardinality of A is 3, $max\{tiesolver_1(p), tiesolver_2(p), tiesolver_3(p)\}$ determines the classification.

6 Experiments

In this section, we go in depth with the empirical validation of the ANYXI classifier. Section 6.1 introduces the dataset under consideration, the quality metrics as well as the algorithms that were selected for comparison purposes. Section 6.2 summarizes the reported results.

6.1 Experimental settings

In the experimental analysis, we have considered the QArt-337 dataset which includes 337 samples of paintings. Each sample is described in terms of the 12 meaningful features (*darkness_level*, *no_paleness_level*, and so on) previously described in Sect. 3.3.2. The classification tasks consists of identifying one out of three art styles (Baroque, Impressionism, Post-Impressionism) in terms of the values given for all the 12 selected features.[8]

On the one hand, for comparison purpose, we have selected the following classifiers:

- **Random forest** (RF). This is a black-box classifier which usually achieves high accuracy in most classification problems, and therefore, it is commonly taken as baseline from the point of view of accuracy. Indeed, Delgado et al. (2014) carried out an exhaustive empirical study, which verifies the previous claim. It is worth noting that RF combines with an ensemble learning method a pool of complementary decision trees, which are generated with the C4.5 algorithm first introduced by Quinlan (Quinlan 1986, 1993).
- **The J48 algorithm** is the implementation in Weka (Witten et al. 2011) of the Quinlan's C4.5 algorithm (Quinlan 1986, 1993). It generates pruned trees which are commonly deemed as white-box classifiers because their behavior can be understood by traversing the trees from root to leaves.
- **The fuzzy unordered rule induction algorithm** (FURIA) was proposed by Hühn and Hüllermeier (2009). It generates fuzzy IF-THEN classification rules with fuzzy sets of trapezoidal shape in the antecedent of each rule. It is worth noting that FURIA rules lack of linguistic interpretability because they deal with fuzzy sets which have local semantics. Accordingly, FURIA classifiers are deemed as gray-box classifiers because they are made up of a set of rules which can be interpreted (at a certain degree).
- **GUAJE** stands for Generating Understandable and Accurate Fuzzy Systems in a Java Environment (Pancho et al. 2013). This toolbox is aimed for building explainable fuzzy classifiers which are considered gray-box classifiers but linguistically interpretable by design (Alonso et al. 2021). In contrast to FURIA rules, the GUAJE rules extracted from data are linguistically grounded in agreement with strong fuzzy partitions defined by experts. As a result, induced rules can be naturally integrated with expert rules. Among the algorithms provided by GUAJE for rule induction, in this paper we will consider only pruned fuzzy decision trees with post hoc linguistic simplification, which have previously proved their ability to produce a good interpretability-accuracy trade-off. Moreover, as we will show in Sect. 6.3 with some illustrative examples, given a data sample GUAJE provides us with its classification along with both factual and counterfactual linguistic explanations which can be directly compared to those generated by ANYXI.

On the other hand, the goodness of an explainable classifier has to be evaluated in terms of the balance between accuracy and interpretability. More precisely, we have applied tenfold cross-validation and reported the mean (and standard deviation) values of the following quality metrics, which are commonly used in the literature (Alonso et al. 2021):

- for **Accuracy**: the ratio of correctly classified instances (*RCCI*), and F-Measure (*F*) is computed as follows:

$$RCCI = 100 \cdot \left(1 - \frac{EC}{N}\right) (\%)$$

$$F = 2 \cdot \left(\frac{Precision \cdot Recall}{Precision + Recall}\right)$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Being N the number of samples (cases) in the dataset, EC the number of misclassified cases (i.e., predicted class is different from the target one), TP accounts for the number of true positive cases, FP is the number of false-positive cases, and FN stands for false negatives. It is worth noting that true positives as well as false positives and negatives are computed for each single class versus the others. For example, if the target class were

[8] The QArt-337 dataset includes images from QArt-Dataset and Painting-91-BIP (which were introduced in Sect. 3.2). It is available at https://gitlab.citius.usc.es/jose.alonso/xai4art/.

Table 4 Reported mean and standard deviation (Std) for all considered interpretability metrics over tenfold cross-validation for the QArt-337 dataset. In the case of the black-box classifier RF, these metrics cannot be computed

	NR		TRL		NC	
Algorithm	Mean	Std	Mean	Std	Mean	Std
RF	–	–	–	–	–	–
J48	31.90	4.01	196.80	31.18	61.80	6.39
FURIA	9.30	3.16	35.00	14.14	28.60	10.95
GUAJE	79.10	4.84	418.40	34.39	59.30	1.06
ANYXI	**3.00**	0.00	**15.00**	0.00	**15.00**	0.00

Baroque, then the F-Measure would be named as F_{Bar}, with TP counting for those cases in the dataset which are correctly classified as Baroque, FP counts for cases wrongly classified as Baroque, while they actually correspond to either Impressionism or Post-impressionism, and FN counts for those cases in the dataset which are not classified as Baroque when they actually correspond to either Impressionism or Post-impressionism.

- for **Interpretability**: the number of leaves/rules (NR), the total rule length (TRL), and the number of concepts (NC). In case of decision trees, we first translate the tree branches into IF-THEN rules and then we compute the interpretability metrics previously enumerated. TRL accounts for the total number of conditions in all the rules (including consequent). NC computes the number of distinct conditions which appear in the rule base, i.e., we assume each condition to represent a concept and we count the number of different concepts in the rule base.

6.2 Experimental results

We applied tenfold cross-validation to evaluate the goodness of all the versions of the ANYXI classifier presented in Sect. 5 in comparison with all the alternative classifiers which we briefly introduced in the previous section. The testing platform was JupyterLab,[9] and Tables 3 and 4 summarize the results obtained.

As expected, the black-box classifier RF achieved the highest accuracy (see values highlighted in bold in the Table 3). However, it is worth noting that GUAJE and ANYXI-1-RPL are even more accurate than RF for the case of Baroque (see column F_{Bar}) in the table. In addition, all in all (regarding the aggregated metrics for the three classes, i.e., regarding columns RCCI and F in the table), ANYXI-1-RPL emerges as the second most accurate algorithm, with RCCI only slightly lower than RF. Moreover, the ANYXI classifier stands out as the best algorithm when paying attention to interpretability metrics (see Table 4). It is worth noting that the reported values for NR, TRL and NC are the same for all the versions of ANYXI under evaluation since the rule base is always the same and only the related operators change from one version to another (accordingly, standard deviation is zero). As described in Sect. 5, the ANYXI classifier deals with three expert rules (one per art style), so NR=3. Each rule handles evaluated horn clauses which are combined with connectives & and →. From the point of view of knowledge representation, each rule comprises four premises and one conclusion, so TRL=3·(4+1)=15. All the three rules relate the twelve color traits introduced in Sect. 3.3.1 and the three art styles, so NC=15.

6.3 Illustrative examples

In this section, we present and discuss two illustrative examples concerning the explanations automatically generated by the best ANYXI model (specifically, ANYXI-1-RPL) and GUAJE.

The first example we consider is *The Saint-Lazare Station* by Monet (Fig. 1(e)), identified as $m31$ in the QArt-337 dataset. We could consider this painting as a clear case of Impressionism.

Following the notation stated in Sect. 5.3, the ANYXI-1-RPL classifier first obtains the parameterization of the evaluated Horn clauses H_1, H_2, H_3. As a result, the rational truth constants of the Horn Clauses are shown in Table 5. Using these parameters, the related membership degrees are obtained (see Table 5). Finally, from all these data, the rule firing degrees are obtained and shown in Table 5.

As it can be easily appreciated, $I(m31)$ takes the highest value, so the painting is classified as Impressionist. In addition, the firing degree of the winner rule is much higher than the others; what confirms that it is a non-ambiguous case. This result is explained by ANYXI-1-RPL as follows: *The painting $m31$ belongs to the Impressionism. The diversity of qualitative colors evidences the Impressionist style. The variety of hues evidences the Impressionist style. The amount of bluish evidences the Impressionist style. The amount of grey evidences the Impressionist style.* Notice that, the previous explanation is intuitive and in agreement with Fig. 1(e).

As an alternative, just for comparison purpose, we can analyze in detail the classification (and the related explanation), which is provided by GUAJE for $m31$: *We have medium confidence in the classification result because activation degree is between 0.375 and 0.625. The classification is probably Impressionism. There is also a small chance that it is Baroque. On balance, Impressionism is more likely, because in accordance with rule 13, classification is Impressionism in case that* $darkness\_level$ *is average and*

[9] For the details on the experimentation, see https://gitlab.citius.usc.es/jose.alonso/xai4art/.

Table 5 Data related to the illustrative example *The Saint-Lazare Station* ($m31$) . The values of B, I, and PI are, respectively, 0.29, 1.00, and 0.09

r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}	r_{12}
0.79	0.89	0.90	0.47	0.79	0.48	0.08	0.46	0.12	0.48	0.18	0.40
B_1	B_2	B_3	B_4	I_5	I_6	I_7	I_8	PI_9	PI_{10}	PI_{11}	PI_{12}
0.75	0.86	0.89	0.80	1.00	1.00	1.00	1.00	0.88	0.56	1.00	0.64

Table 6 Data related to the illustrative example *Thatched Cottages in the Sunshine* ($vg28$) . The values of B, I, and PI are, respectively, 0.00, 0.75, and 0.86

r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}	r_{12}
0.79	0.89	0.90	0.47	0.67	0.49	0.08	0.45	0.12	0.49	0.17	0.40
B_1	B_2	B_3	B_4	I_5	I_6	I_7	I_8	PI_9	PI_{10}	PI_{11}	PI_{12}
0.34	1.00	0.34	0.95	0.97	1.00	0.92	0.87	1.00	0.99	0.87	1.00

$diversity\_of\_qcds$ *is low or average and* $greyish\_level$ *is high and vividness is very low.*

It is worth noting that the winner rule in the case of GUAJE, i.e., rule 13, has a firing degree of 0.492 (computed with the min-max inference mechanism) in this example and it is as follows: **IF** $darkness\_level$ is average AND $diversity\_of\_qcds$ is [low OR average] AND $greyish\_level$ is high AND *vividness* is very low **THEN** Art style is Impressionism. The interpretation of propositions in the rule is done as follows:

- The $darkness\_level$ is average (with membership degree equals 0.833) because it equals 0.541. This feature can take values in the range [0.024, 0.978], and it has associated linguistic terms in the set [very low, low, average, high, very high].
- The $diversity\_of\_qcds$ is [low OR average] (with membership degree equals 1.0) because it equals 0.514. This feature can take values in the range [0.162, 0.946], and it has associated linguistic terms in the set [very low, low, average, high].
- The $greyish\_level$ is high (with membership degree equals 0.492) because it equals 0.614. This feature can take values in the range [0.007, 0.982], and it has associated linguistic terms in the set [very low, low, average, high, very high].
- The $vividness\_level$ is very low (with membership degree equals 0.989) because it equals 0.002. This feature can take values in the range [0.0, 0.743], and it has associated linguistic terms in the set [very low, low, average, high].

Thus, even if the GUAJE rule base includes 80 rules, only two rules (i.e., the winner rule and the rule with the highest firing degree among all rules associated with a different output class) are taken into account when elaborating the factual explanation in natural language. Indeed, such an explanation is much more technical than the one given by ANYXI-1-RPL. In practice, GUAJE just verbalizes the information embedded in the winner rule because in this example the alternative rule (69) points at Baroque but with a much smaller firing degree (0.204).

Interestingly, in addition to the previous factual explanation, GUAJE also produces counterfactual explanations: *Classification would be Baroque if* $diversity\_of\_qcds$ *were smaller (0.454). Classification would be Post-Impressionism if vividness were bigger (0.094).* These counterfactual explanations look for the minimal changes in the input values that would produce a winner rule pointing at a different alternative class. In this case, only changing slightly the value of *vividness* (from 0.002 to 0.094) the painting may be classified as Post-Impressionism instead of Impressionism. This is due to the fact that both art styles are somehow related. However, it is harder to pass from Impressionism to Baroque. In this case, GUAJE suggests decreasing the value of $diversity\_of\_qcds$ (from 0.514 to 0.454).

Let us now consider a second illustrative example. Namely, we consider the painting *Thatched Cottages in the Sunshine* by van Gogh (Fig. 1(h)), identified as $vg28$ in the QArt-337 dataset. Here, the classification task is harder and more challenging than in the first example because colors in the painting might be confused with those characteristics of Impressionism even if it belongs to Post-Impressionism.

We proceed like in the first example. Once again, following the notation stated in Sect. 5.3, the ANYXI-1-RPL classifier first obtains the parameterization of the evaluated Horn clauses H_1, H_2, H_3. As a result, the rational truth constants of the evaluated Horn clauses are shown in Table 6. Then, the related membership degrees are obtained (see Table 6). Finally, using all these data, the rule firing degrees are obtained and shown in Table 6.

As it can be easily appreciated, $PI(vg28)$ takes the highest value, so the painting is classified as Post-Impressionist. However, in contrast to the first illustrative example, here the firing degree of the winner rule is not so far from the firing degree of the alternative rule associated to $I(vg28)$; what confirms that it is a hard ambiguous case. This result is explained by ANYXI-1-RPL as follows: *The painting* $vg28$ *belongs to the Post-Impressionism style. The presence of vivid colors evidences the Post-Impressionism style. The high level of warm colors evidences the Post-Impressionism style. The contrast between red and green evidences the Post-Impressionism style.*

As an alternative, GUAJE produces the following explanation: *We have medium confidence in the classification result because activation degree is between 0.375 and 0.625. The classification is probably Post-Impressionism or Impressionism. On balance, Post-impressionism is more likely, because in accordance with rule 33, classification is Post-impressionism in case that* $darkness\_level$ *is low and* $greyish\_level$ *and vividness are very low and* $diversity\_of\_hues$ *is average.*

It is worth noting that the winner rule in this case, i.e., rule 33, has a firing degree of 0.414 and it is as follows: **IF** $darkness\_level$ is low AND $diversity\_of\_hues$ is average AND $greyish\_level$ is very low AND *vividness* is very low **THEN** Art style is Post-Impressionism. The interpretation of propositions in the rule is as follows:

- The $darkness\_level$ is low (with membership degree equals 0.609) because it equals 0.356. This feature can take values in the range [0.024, 0.978], and it has associated linguistic terms in the set [very low, low, average, high, very high].
- The $diversity\_of\_hues$ is average (with membership degree equals 1) because it equals 0.636. This feature can take values in the range [0.273, 1.0], and it has associated linguistic terms in the set [very low, low, average, high, very high].
- The $greyish\_level$ is very low (with membership degree equals 0.414) because it equals 0.15. This feature can take values in the range [0.007, 0.982], and it has associated linguistic terms in the set [very low, low, average, high, very high].
- The $vividness$ is very low (with membership degree equals 0.833) because it equals 0.031. This feature can take values in the range [0.0, 0.743], and it has associated linguistic terms in the set [very low, low, average, high].

In this case, the alternative rule (9) which points at Impressionism instead of Post-Impressionism has a firing degree of 0.347. Since the difference between firing degrees of the two most relevant rules (9 and 33) is smaller than 0.1, this case is considered as ambiguous; what is highlighted by GUAJE when stating *The classification is probably Post-Impressionism or Impressionism*, i.e., the two classes are almost equally possible but *On balance, Post-Impressionism is more likely* because firing degree of rule 33 is slightly higher than firing degree of rule 9 in this case. Once again, this explanation is in agreement with the one given by ANYXI-1-RPL, even if GUAJE provides further technical details and ANYXI-1-RPL resembles more understandable for lay users.

In addition, GUAJE generates the following counterfactual explanations: *Classification would be Baroque if* $diversity\_of\_qcds$ *were smaller and* $darkness\_level$ *and* $bluish\_level$ *were bigger* ($darkness\_level$ = 0.384; $diversity\_of\_qcds = 0.456$; $bluish\_level = 0.157$). *Classification would be Impressionism if* $diversity\_of\_qcds$ *and* $greyish\_level$ *were smaller and* $darkness\_level$ *were bigger*
($darkness\_level$ = 0.382; $diversity\_of\_qcds$ = 0.453; $greyish\_level = 0.127$).

6.4 Discussion

All in all, classification made by ANYXI-1-RPL is consistent and in agreement with GUAJE. Moreover, both explainable classifiers generate complementary explanations which rely on different fuzzy reasoning approaches. Therefore, each classifier highlights different relevant features when explaining the classification result. In addition, explanations provided by ANYXI-1-RPL are intuitive and aimed for end-users, while explanations provided by GUAJE are a bit more technical. It is also worth noting that explanations provided by ANYXI-1-RPL are based on static templates and therefore somehow repetitive, while explanations provided by GUAJE come out of a dynamic combination of templates with the assistance of the SimpleNLG library (Gatt and Reiter 2009), which takes care of enhancing textual realization and avoiding repetitions.

As described by Falomir and Costa (2021), there are seven key criteria that rational explanations in classification algorithms should meet. Accordingly, automated explanations are expected to be:

1. **Human-understandable**: People can read and understand the general meaning of the given explanation. This is a necessary condition to evaluate the rationality of explanations.
2. **Conceptual**: The classifier employs concepts to categorize the groups and create the propositions explaining the result. These concepts used to classify an item into a group must be related to the characteristic and pertinent traits of the corresponding class.
3. **Coherent with human perception**: The classifiers must align their perception (sensor data) to concepts that people can understand and usually use to intercommunicate, so the notions used in the explanation must be coherent and aligned with human perception.

4. **Context adequate**: In linguistics, it is essential to find a set of attributes (minimal or psychologically plausible) for the intended referent, but not all true for any distractor. A rational explanation should be aware of the context and use a collection of attributes as minimal as possible.
5. **Personalized according to users' background**: Since the rationality of some explanations depends on the user's background, the more adapted to it, the more rational the explanation is.
6. **Coherent with observable human reasoning**: Explanations regarding classification results should be coherent with observable and rational patterns of human reasoning (note that these patterns depend on the classification problem considered). Otherwise, explanations may become strange, sometimes even misleading, and unexpected by users.
7. **Contrastive and counterfactual**: The literature (Poyiadzi et al. 2020; Miller 2019; Byrne 1998; Stepin et al. 2021) has shown the importance of contrastive and counterfactual explanations. Therefore, rational classifiers should include these sorts of explanations.

Let us discuss below about the rationality of the explanations yielded by ANYXI-1-RPL and GUAJE regarding the two illustrative examples previously depicted. To do so, we follow the guidelines introduced by Falomir and Costa (2021) with the aim of evaluating if the criteria enumerated above are met here.

Concerning the explanations provided by ANYXI-1-RPL, we might conclude that their rationality is medium since they meet only four out of the seven criteria. First, the generated explanations are human-understandable (people can comprehend the text outcome easily). Furthermore, they meet the conceptual criteria (the classifier takes twelve color traits proposed from the art experts' studies, and these concepts are related to each distinctive style's characteristics). In addition, since the color traits are defined using the frequencies extracted with the QCD model, the explanations yielded by ANYXI-1-RPL are coherent with observable human perception. Last but not least, the explanations highlight the more relevant attributes that characterize the identified style (four attributes in the first example and three in the second one), so ANYXI-1-RPL also satisfies the context adequate criterion. Regarding the explanations generated by GUAJE, we might infer that their rationality is medium-high because they fulfill five of the seven criteria considered. More specifically, the outcomes depicted by GUAJE are human-understandable, and, for the same reasons as ANYXI-1-RPL explanations, they meet the criteria 2-4. In addition, in both examples, GUAJE yields counterfactual explanations that help to understand better the classification results.

However, as stated above, explanations provided by both classifiers fail to fulfill some of the desired criteria of rationality. First, neither ANYXI-1-RPL nor GUAJE personalize the explanations according to the user's background (even if explanations given by ANYXI-1-RPL are better adapted to lay users, while technical users are likely to appreciate more GUAJE's explanations, none of these explanations are customized on demand to the background of single individuals). Furthermore, automated explanations should be coherent with observable and rational patterns of human reasoning associated with art painting style classification. Nevertheless, the use of t-norm based logics in the ANYXI-1-RPL design makes this goal harder to achieve. For example, the logic $\sqcap(\mathbb{Q})$ admits annihilators, although the characteristic patterns of human aggregative reasoning related to this classification task do not often include them. And a similar case applies to GUAJE, whose underlying min-max inference mechanism and conditional formal system admits annihilators too. Finally, we note that the explanations given by ANYXI-1-RPL are not contrastive or counterfactual, and this is a drawback to be addressed in the near future. As we have empirically observed in this paper, users could demand contrastive or counterfactual explanations especially in the case of ambiguous borderline tough cases or for clarifying misclassification.

7 Final remarks and future lines of research

In summary, we presented the ANYXI classifier, an AI system based on art specialists' knowledge of art styles and human-understandable color traits defined from a qualitative color model. In addition, we analyzed the appropriateness of the art style definitions used by the classifier for human understanding and examined the accuracy, interpretability, and rationality of automated explanations. We improved the discussion of these aspects by considering other explainable classifiers designed using the same dataset as ANYXI.

More in detail, we first provided readers with a brief overview of research in art painting style categorization using AI, from which we highlighted three relevant points. Indeed, most of the related publications are beyond the scope of XAI, which is the sort of AI to strive for if the intended goal is acquiring a more trustworthy and fair discipline. Furthermore, the only few previous art painting style explainable classifiers in the literature present some disadvantages: on the one hand, from a computational point of view, the low accuracy, in comparison with other machine learning-based methods; on the other hand, its art styles categorizations could be more exhaustive and similar to art experts' categorizations. Last but not least, comparing the research work in the literature can be a bit futile since reported designs and experiments are hard to replicate since they usually come from different datasets, and authors do not often share all the experimental settings needed to evaluate the goodness of their performance.

The ANYXI classifier tackles these issues to improve. First, it proposes complete color-based categorizations of the art style under consideration. At the same time, the definitions suggested handle four notions per style, which is an affordable quantity of traits, from a cognitive point of view. In addition, we studied using a survey the validity of these categorizations and showed their appropriateness for including them in a human-centered design. All this together leads to a depiction of more human-understandable and complete explanations of the classification results. Furthermore, ANYXI emerges as the best classifier regarding interpretability, and it is almost as accurate as the black-box classifier considered as baseline in this work. Moreover, the reported results and drawn conclusions are sound because we built all the classifiers from the same dataset and tested them using the same 10-fold cross-validation experimental setting.

These contributions derive relevant research questions and lines for future research. Regarding the improvement of ANYXI. On the one hand, the explanations of the illustrative examples presented in Sect. 6.3 show that the texts are a bit repetitive. That might be reasonable when reading a few, but it can turn into an obstacle when considering high amounts of outcomes. Future studies could solve this issue further by using Natural Language Generation techniques. On the other hand, the discussion about the rationality of the explanations provided by ANYXI drives future research to focus on improving this aspect. First, the GUAJE outcomes of the illustrative examples warrant further investigation by combining both classifiers. In this way, we will supply a new version of ANYXI with the generation of counterfactual explanations so the classifier will meet other rationality criteria. And second, concerning the rationality criteria of explanations' personalization, we intend to use the work by Alonso and Bugarín (2019) and design a similar approach to distinguish between beginner and expert users. In addition, the main weakness of the method presented is the lack of scalability determined by using only color traits for categorizing the art styles. Considering an additional style not characterized by color would imply using evaluated Horn clauses with propositional variables representing other kinds of features (e.g., the geometric shape) and, considering the design of the ANYXI method, would also lead to different types of explanations, depending on the art style selected. Furthermore, the mentioned additional styles would give rise to redesigning the survey in Sect. 4. However, the study of the method's scalability can be addressed in two ways, which are left to future work. On the one hand, adding art styles well categorized in terms of color to the ANYXI method with an experimental analysis would shed light on its scalability. On the other hand, experimental analysis using new paintings from art styles different from the three ones selected in this paper, to know whether a random classification of these external paintings is obtained, would help to determine the scalability of the method.

Finally, we believe that systems such as ANYXI might enable collaborative intelligence between humans and machines. This way, future research includes showing the ANYXI outcomes to art specialists, seeking feedback from them, and establishing a collaboration to improve the knowledge about art painting styles.

Acknowledgements We acknowledge the reviewers for their useful comments.

Author Contributions Vicent Costa, Jose M. Alonso-Moral, and Zoe Falomir have contributed equally to this work regarding the conceptualization of the new algorithm, the design of experiments, data analysis, and writing of the final manuscript. Pilar Dellunde contributed to the conceptualization and formalization of the new algorithm (ANYXI) as well as the writing of this manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work is supported by MCIN/AEI /10.13039/501100011033 under grants PID2021-123152OB-C21, PID 2021-123152OB-C22, and RED2022-134315-T and the Galician Ministry of Culture, Education, Professional Training and University under grants ED431G2019/04 and ED431C2022/19. All grants were cofunded by the European Union and FEDER/ERDF (European Regional Development Funds). This work is also partially funded by the H2020-MSCARISE- 2020 project MOSAIC (Grant Agreement 101007627). Z. Falomir acknowledges funding by the Ramon-y-Cajal fellowship (RYC2019-027177-I / AEI / 10.13039/501100011033) awarded by the Spanish Ministry of Science, Innovation and Universities. V. Costa received funding from the Juan de la Cierva fellowship (FJC2021-047274-I) awarded by the Spanish Ministry of Science, Innovation and Universities, and the European Union (NextGenerationEU), and acknowledges the GEHUCT research group (2021 SGR 00517).

Data Availability The QArt-337 dataset, images, scripts needed for reproducibility of experiments as well as logs with reported results are available at https://gitlab.citius.usc.es/jose.alonso/xai4art/.

Declarations

Ethical approval The methods were carried out in accordance with the relevant guidelines and regulations.

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article. The authors have no relevant financial or non-financial interests to disclose.

Informed consent In the survey carried out in this paper, all the subjects participated voluntarily and received no compensation. The survey's instructions clearly informed participants of the goal of the data recollected and asked their consent. Answers were collected anonymously only if participants agreed to participate in our study, if not, the survey ended automatically asking no data.

References

Alonso JM, Bugarín A (2019) Expliclas: automatic generation of explanations in natural language for weka classifiers. In: International conference on fuzzy systems (FUZZ-IEEE), New Orleans, USA (pp. 1- 6). IEEE

Alonso JM, Castiello C, Magdalena L, Mencar C (2021) Explainable fuzzy systems: paving the way from interpretable fuzzy systems to explainable AI systems. Springer, Cham

Alonso-Moral JM, Mencar C, Ishibuchi H (2022) Explainable and trustworthy artificial intelligence. IEEE Comput Intell Mag 1:15. https://doi.org/10.1109/MCI.2021.3129953

Anguita, D., Ghio, A., Greco, N., Oneto, L., Ridella, S. (2010). Model selection for support vector machines: advantages and disadvantages of the machine learning theory. In: International joint conference on neural networks (IJCNN), Barcelona, Spain, 18-23 july, 2010. IEEE. https://doi.org/10.1109/IJCNN.2010.5596450

Arrieta AB, Rodríguez ND, Ser JD, Bennetot A, Tabik S, Barbado A, Herrera F (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Berson R (1996) The new painting: Impressionism, 1874-1886: documentation. Fine Arts Museums of San Francisco

Breiman L (2001) Random forest. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Burcoff A, Shamir L (2017) Computer analysis of Pablo Picasso's artistic style. Int J Art Cult Des Technol 6(1):1–18. https://doi.org/10.4018/IJACDT.2017010101

Byrne RMJ (1998) Spatial mental models in counterfactual thinking about what might have been. Kognit Wiss 7:19–26. https://doi.org/10.1007/BF03354959

Castellano G, Vessio G (2021) Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview. Neural Comput Appl. https://doi.org/10.1007/s00521-021-05893-z

Cintula P, Hájek P, Noguera C (2011) Handbook of mathematical fuzzy logic - vol. 1. College Publications

Condorovici RG, Florea C, Vertan C (2015) Automatically classifying paintings with perceptual inspired descriptors. J Vis Commun Image Represent 26:222–230. https://doi.org/10.1016/j.jvcir.2014.11.016

Costa V (2020) The art painting style classifier based on logic aggregators and qualitative colour descriptors (C-LAD). In: Rudolph S, Marreiros G (Eds.), Proceedings of the 9th European starting AI researchers' symposium (STAIRS), co-located with 24th European conference on artificial intelligence (ECAI), Santiago de Compostela, Spain (Vol. 2655). CEUR-WS.org. http://ceur-ws.org/Vol-2655/paper17.pdf

Costa V, Dellunde P (2017) Term models of horn clauses over rational Pavelka predicate logic. In: 47th international symposium on multiple-valued logic (ISMVL), Novi Sad, Serbia (pp. 112-117). IEEE Computer Society. https://doi.org/10.1109/ISMVL.2017.26

Costa V, Dellunde P, Falomir Z (2018) Style painting classifier based on horn clauses and explanations (SHE). In: Falomir Z, Gibert K, Plaza E (Eds.), Artificial intelligence research and development: current challenges, new trends and applications, 21st International conference of the catalan association for artificial intelligence (CCIA), Alt Empordà, Catalonia, Spain (Vol. 308, pp. 37-46). IOS Press. https://doi.org/10.3233/978-1-61499-918-8-37

Costa V, Dellunde P, Falomir Z (2021) The logical style painting classifier based on horn clauses and explanations (.-SHE). Log J IGPL, 29(1): 96.119. https://doi.org/10.1093/jigpal/jzz029

Dasiopoulou S, Kompatsiaris I, Strintzis MG (2010) Investigating fuzzy DLs based reasoning in semantic image analysis. Multimed Tools Appl 49:167–194. https://doi.org/10.1007/s11042-009-0393-6

Delgado MF, Cernadas E, Barro S, Amorim DG (2014) Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 15(1):3133–3181

Dewhurst W (1908) Impressionist painting: its genesis and development. J R Soc Arts 56(2887):475–489

El-Sappagh S, Alonso JM, Islam SMR, Sultan AM, Kwak KS (2021) A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Sci Rep. https://doi.org/10.1038/s41598-021-82098-3

European Commission (2018) Artificial Intelligence for Europe (SWD(2018) 137 final) (Tech. Rep.). Brussels, Belgium: Author. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018DC0237&from=EN

Falomir Z, Cabedo LM, Abril LG (2015) A model for colour naming and comparing based on conceptual neighbourhood. An application for comparing art compositions. Knowl Based Syst 81:1–21. https://doi.org/10.1016/j.knosys.2014.12.013

Falomir Z, Cabedo LM, Sanz I, Abril LG (2015) Guessing art styles using qualitative colour descriptors, SVMs and logics. In: Armengol E, Boixader D, Grimaldo F (Eds.), Artificial intelligence research and development: proceedings of the 18th international conference of the catalan association for artificial intelligence (CCIA), Valencia, Spain (Vol. 277, pp. 227-236). IOS Press. https://doi.org/10.3233/978-1-61499-578-4-227

Falomir Z, Cabedo LM, Sanz I, Abril LG (2018) Categorizing paintings in art styles based on qualitative color descriptors, quantitative global features and machine learning (qart-learn). Expert Syst Appl 97:83–94. https://doi.org/10.1016/j.eswa.2017.11.056

Falomir Z, Costa V (2021) On the rationality of explanations in classification algorithms. In: Villaret M, Alsinet T, Fernàndez C, Valls A (Eds.), Artificial intelligence research and development: proceedings of the 23rd international conference of the catalan association for artificial intelligence (CCIA) (Vol. 339, pp. 445-454). IOS Press. https://doi.org/10.3233/FAIA210165

Falomir Z, Jiménez-Ruiz E, Escrig MT, Cabedo LM (2011) Describing images using qualitative models and description logics. Spatial Cogn Comput 11(1):45–74. https://doi.org/10.1080/13875868.2010.545611

Falomir Z, Museros L, González-Abril L, Sanz I (2013) A model for qualitative colour comparison using interval distances. Displays 34(4):250–257. https://doi.org/10.1016/j.displa.2013.07.004

Gatt A, Reiter E (2009) SimpleNLG: a realisation engine for practical applications. European workshop on natural language generation (enlg) (pp. 90-93). Athens, Greece. https://aclanthology.org/W09-0613/

Gatys LA, Ecker AS, Bethge M (2016) Image style transfer using convolutional neural networks. In: Conference on computer vision and pattern recognition (CVPR), Las Vegas, USA (pp. 2414-2423). IEEE Computer Society. https://doi.org/10.1109/CVPR.2016.265

González CI, Melin P, Castillo O (2017) Edge detection method based on general type-2 fuzzy logic applied to color images. Information 8(3):104. https://doi.org/10.3390/info8030104

Grygar T, Hradilová J, Hradil D, Bezdička P, Bakardjieva S (2003) Analysis of earthy pigments in grounds of Baroque paintings. Anal Bioanal Chem 375:1154–1160. https://doi.org/10.1007/s00216-002-1708-x

Gunning D, Vorm E, Wang JY, Turek M (2021) DARPA's explainable AI (XAI) program: a retrospective. Appl AI Lett 2(4):e61. https://doi.org/10.1002/ail2.61

Hagras H (2018) Toward human understandable, explainable AI. Computer 51(9):28–36. https://doi.org/10.1109/MC.2018.3620965

Hill IB (1980) Impressionist painting. Galley Press

Hühn JC, Hüllermeier E (2009) FURIA: an algorithm for unordered fuzzy rule induction. Data Min Knowl Discov 19(3):293–319. https://doi.org/10.1007/s10618-009-0131-8

Jiang S, Huang Q, Ye Q, Gao W (2006) An effective method to detect and categorize digitized traditional Chinese paintings. Pattern Recognit Lett 27(7):734–746. https://doi.org/10.1016/j.patrec.2005.10.017

Karayev S, Trentacoste M, Han H, Agarwala A, Darrell T, Hertzmann A, Winnemoeller H (2014) Recognizing image style. In: Valstar MF, French AP, Pridmore TP (Eds.), British machine vision conference (BMVC), Nottingham, UK. BMVA Press. https://doi.org/10.5244/C.28.122

Khan FS, Beigpour S, van de Weijer J, Felsberg M (2014) Painting-91: a large scale database for computational painting categorization. Mach Vis Appl 25:1385–1397. https://doi.org/10.1007/s00138-014-0621-6

Ma Y-W, Chen J-L, Chen Y-J, Lai Y-H (2021) Explainable deep learning architecture for early diagnosis of Parkinson's disease. Soft Comput. https://doi.org/10.1007/s00500-021-06170-w

Mamassian P (2008) Ambiguities and conventions in the perception of visual art. Vis Res 48(20):2143–2153. https://doi.org/10.1016/j.visres.2008.06.010

Mao H, Cheung M, She J (2017) DeepArt: learning joint representations of visual arts. In: Proceedings of the 25th ACM international conference on Multimedia (pp. 1183-1191). https://doi.org/10.1145/3123266.3123405

Mao H, She J, Cheung M (2019) Visual arts search on mobile devices. ACM Trans Multimed Comput Commun Appl 60:1–23. https://doi.org/10.1145/3326336

Mensink T, van Gemert JC (2014) The rijksmuseum challenge: Museum-centered visual recognition. In: Kankanhalli MS, Rueger S, Manmatha R, Jose JM, van Rijsbergen K (Eds.), International conference on multimedia retrieval (ICMR), Glasgow, UK (p. 451). ACM. https://doi.org/10.1145/2578726.2578791

Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mohammad SM, Kiritchenko S (2018) WikiArt emotions: an annotated dataset of emotions evoked by art. In: Proceedings of the 11th edition of the language resources and evaluation conference (LREC). Miyazaki, Japan. http://www.lrec-conf.org/proceedings/lrec2018/summaries/966.html

Pancho DP, Alonso JM, Magdalena L (2013) Quest for interpretability-accuracy trade-off supported by fingrams into the fuzzy modeling tool GUAJE. Int J Comput Intell 6:46–60. https://doi.org/10.1080/18756891.2013.818189

Powell-Jones M (1979) Impressionist painting. Mayflower Books

Poyiadzi R, Sokol K, Santos-Rodríguez R, de Bie T, Flach PA (2020) FACE: feasible and actionable counterfactual explanations. In: Markham AN, Powles J, Walsh T, Washington AL (Eds.), Conference on AI, ethics, and society (AIES), New York, USA (pp. 344-350). ACM. https://doi.org/10.1145/3375627.3375850

Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106. https://doi.org/10.1023/A:1022643204877

Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann Publishers, San Mateo, CA

Reiter R (1980) A logic for default reasoning. Artif Intell 13(1–2):81–132. https://doi.org/10.1016/0004-3702(80)90014-4

Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you?: explaining the predictions of any classifier. In: Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D, Rastogi R (Eds.), 22nd international conference on knowledge discovery and data mining (SIGKDD), San Francisco, USA (pp. 1135-1144). ACM. https://doi.org/10.1145/2939672.2939778

Rubio E, Castillo O, Valdez F, Melin P, González CI, Martinez G (2017) An extension of the fuzzy possibilistic clustering algorithm using type-2 fuzzy logic techniques. Adv Fuzzy Syst 7094046:1–23. https://doi.org/10.1155/2017/7094046

Rzepińska M, Malcharek K (1986) Tenebrism in Baroque painting and its ideological background. Artibus Hist 7(13):91–112. https://doi.org/10.2307/1483250

Samek W, Müüller K (2019) Towards explainable artificial intelligence. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K (Eds.), Explainable AI: interpreting, explaining and visualizing deep learning (Vol. 11700, pp. 5-22). Springer. https://doi.org/10.1007/978-3-030-28954-61

Samu M (2004) Impressionism: art and modernity. In Heilbrunn Timeline of Art History. New York: The Metropolitan Museum of Art. Retrieved 2022-04-29, from http://www.metmuseum.org/toah/hd/imml/hdimml.htm

Sanz I, Museros L, Falomir Z, Gonzalez-Abril L (2015) Customising a qualitative colour description for adaptability and usability. Pattern Recognit Lett 67:2–10. https://doi.org/10.1016/j.patrec.2015.06.014

Shamir L (2015) What makes a Pollock Pollock: a machine vision approach. Int J Arts Technol 8(1):1–10. https://doi.org/10.1504/IJART.2015.067389

Shamir L, Tarakhovsky JA (2012) Computer analysis of art. ACM J Comput Cult Herit 5(2):1–11. https://doi.org/10.1145/2307723.2307726

Shen J (2009) Stochastic modeling western paintings for effective classification. Pattern Recognit 42(2):293–301. https://doi.org/10.1016/j.patcog.2008.04.016

Siddiquie B, Vitaladevuni SNP, Davis LS (2009) Combining multiple Kernels for efficient image classification. In: Workshop on applications of computer vision (WACV), Snowbird, USA (pp. 1-8). IEEE Computer Society. https://doi.org/10.1109/WACV.2009.5403040

Stepin I, Alonso JM, Catala A, Pereira- Fariña M (2021) A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access 9:11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

Toll H (2018) Handbook of arts-based research (manuel sur la recherche axée sur l'art), edited by Patricia Leavy. Can Art Ther Assoc J 31(2):105–107. https://doi.org/10.1080/08322473.2018.1520030

Vilone G, Longo L (2021) Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf Fusion 76:89–106. https://doi.org/10.1016/j.inffus.2021.05

Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques, 3rd edition. Morgan Kaufmann, Elsevier. https://www.worldcat.org/oclc/262433473

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