FeatSet+: Visual Features Extracted from Public Image Datasets

Mirela T. Cazzolato[1,2], Lucas C. Scabora[1], Guilherme F. Zabot[1],
Marco A. Gutierrez[2], Caetano Traina Jr.[1], Agma J. M. Traina[1]

[1] Institute of Mathematics and Computer Sciences
University of São Paulo (ICMC-USP), São Carlos, Brazil
{mirelac, lucascsb, zabot}@usp.br, {agma, caetano}@icmc.usp.br
[2] The Heart Institute (InCor) – Clinical Hospital of Faculty of Medicine
University of São Paulo (HC-FMUSP), São Paulo, Brazil
marco.gutierrez@incor.usp.br

Abstract. Real-world applications generate large amounts of images every day. With the generalized use of social media, users frequently share images acquired by smartphones. Also, hospitals, clinics, exhibits, factories, and other facilities generate images with potential use for many applications. Processing the generated images usually requires feature extraction, which can be time-consuming and laborious. In this paper, we present FeatSet+, a compilation of color, texture and shape visual features extracted from 17 open image datasets reported in the literature. FeatSet+ provides a collection of 11 distinct visual features, extracted by well-known Feature Extraction Methods (FEMs) such as LBP, Haralick, and Color Layout. We organized the available features in a standard collection, including the metadata and labels, when available. Eleven of the datasets also contain classes, which aid the evaluation of supervised methods such as classifiers and clustering tasks. FeatSet+ is available for download in a public repository as *sql* scripts and *csv* files. Additionally, FeatSet+ provides a description of the domain of each dataset, including the reference to the original work and link. We show the potential applicability of FeatSet+ in four computational tasks: multi-attribute analysis and retrieval, visual analysis using Multidimensional Scaling (MDS) and Principal Components Analysis (PCA), global feature classification, and dimensionality reduction. FeatSet+ can be employed to evaluate supervised and non-supervised learning tasks, also widely supporting Content-Based Image Retrieval (CBIR) applications and complex data indexing using Metric Access Methods (MAMs).

Categories and Subject Descriptors: H.2 [**Database Management**]: Miscellaneous; H.3 [**Information Storage and Retrieval**]: Miscellaneous

Keywords: Dataset, image, visual features, color, texture, shape, CBIR, retrieval, analysis

1. INTRODUCTION

Images acquired from different application scenarios have been the focus of numerous studies for decades now. Users and applications generate large amounts of images, sharing them over the internet, mainly due to the widespread use of social networks, blogs, public repositories, and cooperative research. Regular and surveillance cameras, mobile devices, microscopes, Magnetic Resonance Imaging (MRI) machines, and X-Rays are just a few examples of simple and specialized acquisition equipment capable of generating images in different contexts and resolutions.

Data management systems usually compare and query over scalar data, such as small strings, numbers, and dates. Order and equality operators (*i.e.*, $<$, $\leq$, $=$, $\geq$, and $>$) are adequate for scalar data. On the other hand, images are considered complex data that order operators cannot compare since they do not possess the order property. Also, checking if a pair of images is equal or different brings too little semantics to the analysis. Accordingly, applications compare complex objects according to their content, employing similarity operators such as the *Range* and kNN queries for Content-Based Retrieval (CBR).

Feature Extraction Methods (FEMs) generate feature vectors as low-level representations of images according to their visual content. The generated features can describe, among others, the color distribution of an image, the grayscale variation, the texture patterns, and the objects' salient edges. Similarity operators compare pairs of complex objects by employing distance functions to the corresponding feature vectors, measuring how dissimilar they are. Also, machine learning algorithms employ feature vectors to train their model in supervised scenarios (for instance, when we have a label for every image in the dataset) and in exploratory analyses. All of these analyses depend on extracting the features from the image datasets, which can be both time-consuming and laborious due to the necessity of implementing and/or setting up FEMs.

Many works from the literature have employed visual features for image analysis in different contexts. For instance, in the work [de Sousa Fogaça and Bueno 2020], the authors mapped color-based features into the multidimensional space to estimate the trajectory of objects by simulating their evolution over time. In [Pereira and Ribeiro 2021], the authors explored visual features extracted from mammograms for the semantic annotation and classification of images using an ontology. Low-level features have been widely applied to validate the indexing capabilities of Metric Access Methods [Zabot et al. 2019; Moriyama et al. 2021]. Also, in [Maheshwari et al. 2021] the authors exploited several visual features to identify COVID-19 in images. However, few existing studies make the employed visual features available for download, *e.g.*, [Chino et al. 2015; Rodrigues et al. 2020]. Such studies usually limit the provided data to specific sets of images (*e.g.*, [Oliveira et al. 2017; Cazzolato et al. 2017]), or provide extracted visual representations for a particular computational task or data domain related to the images [Cazzolato et al. 2016]. However, most image- and feature-related studies provide the code or the reference for the specific FEM employed so that readers can extract the visual features of the desired images on their own.

Motivated by the potential applicability of image features and the difficulties of employing FEMs, we propose the `FeatSet+` dataset in this work. `FeatSet+` is a compilation of widely-used visual features extracted from public image datasets of different application scenarios. The contributions of `FeatSet+` are two-fold:

—The curation of 17 open image databases, organizing their main information in a single repository;
—Making it readily available visual features based on color, texture, and shape, extracted from the images using 11 distinct FEMs, widely employed in the literature, including those from the MPEG7 Standard [Manjunath et al. 2002]. The feature vectors are organized into a standard model and openly available.

We cast the using possibilities for `FeatSet+`, and show four examples of analysis of the available data:

(1) Multi-attribute analysis and retrieval: we take advantage of subclasses available in the original datasets to show how we can improve similarity queries with multi-attributes.
(2) Visual analysis: We show that the employed features present different distribution dispersion even for a single dataset, regarding classes' dispersion and two visual tools (Multidimensional Scaling and Principal Component Analysis).
(3) Global feature classification: we employed off-the-shelf classification approaches to classify the 11 labeled datasets, showing opportunities of improvement in future work employing `FeatSet+`.
(4) Principal Component Analysis for dimensionality reduction: we show an example of approach to reduce the dimensionality of the feature vectors and a combination of them (with color and texture features).

Previous use of data. A small part of FeatSet+ has been employed in the previous studies [Zabot et al. 2019; Zabot et al. 2019]. In those studies, the authors explored different visual features to validate a novel Multi-Metric Access Method, aimed at indexing complex objects based on images' visual characteristics and the correlation among the distance spaces. In a previous, shorter work [Cazzolato et al. 2021] we presented FeatSet, a compilation of visual features extracted from 13 public image datasets. In this work, we present an extended, and complete version of the data used in [Zabot et al. 2019; Zabot et al. 2019; Cazzolato et al. 2021]. FeatSet+ is a superset of FeatSet with four new datasets, composed of diverse visual features extracted from various public image datasets of different application scenarios. It allows analysts to deeper evaluating machine learning approaches, CBIR strategies, and related techniques. We further extend the presentation and discussion with two new application scenarios for FeatSet+, including examples for multi-attribute queries, classification, feature selection, and data visualization.

Paper outline. The remaining sections of this paper are organized as follows. Section 2 describes FeatSet+. Section 3 discusses application scenarios and challenges for FeatSet+. Section 4 details the steps to download FeatSet+, and describes the data organization and description of the dataset's public repository. Finally, Section 5 concludes this work.

2. FEATSET+: A COLLECTION OF VISUAL FEATURES FROM IMAGE DATASETS

In this section, we detail the process of acquiring the original images, extracting the visual features, and composing FeatSet+, as Figure 1 illustrates. First (Step i), we looked for open image datasets to extract features. We focused on literature papers proposing or using image data from open repositories and websites for this task. We did not systematically search for existing image collections in the literature. In turn, we included datasets focused on diverse applications (*e.g.*, medicine, emergency, object recognition) and of different sizes. We started the composition of FeatSet+ by including image collections employed in previous works of our group (*e.g.*, [Chino et al. 2015; Bedo et al. 2015; Cazzolato et al. 2017; Oliveira et al. 2017]), and extracting their visual features. Then, we searched complementary collections in the literature. The datasets had to be openly available, free of use, and include the original image files to be considered. As a result, FeatSet+ is of general use as it provides various visual representations of images from many applications.

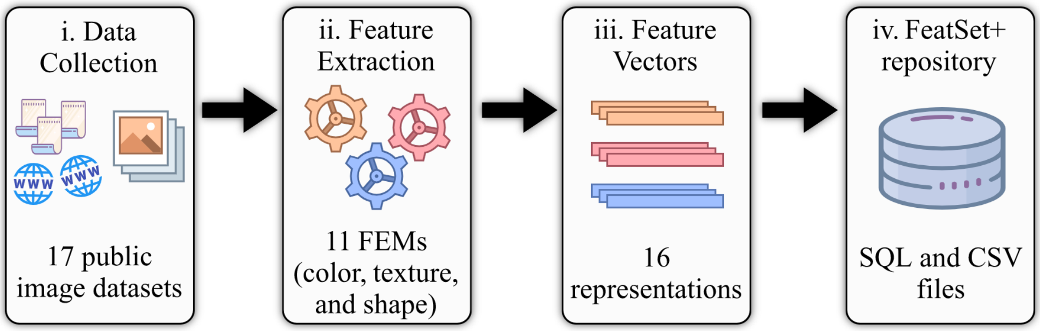


Fig. 1. Steps carried to compose FeatSet+.

2.1 Data Collection and Preprocessing

Table I lists the datasets collected, as well as the specific reference, the number of available images, and a brief description of each. Figure 2 shows examples of images acquired from each dataset. All 17 datasets were manually collected from their original sources, which are referenced in the repository, as will be described in Section 4, with a "*read me*" file containing the original source, URL, date of collection, license, a brief description, and the corresponding reference. The Preprocessing step ensures that every image will be in an image file format (*i.e.*, *.png*, *.jpeg*, and *.jpg*), converting each file using Python scripts and libraries (*e.g.* OpenCV) if necessary. For instance, the *ds-MNIST* dataset provides the images as multidimensional matrices. We converted each image file from *ds-MNIST* to *.png*, which is one of the image formats accepted by the implementation of FEMs employed in this work. The script to convert the images from *ds-MNIST* is also available in our repository.

Table I. List of datasets composing `FeatSet+`, with the corresponding reference, the number of available images, and a brief description of the application scenario.

Dataset	#Images	Application Scenario
ds-BoWFire [Chino et al. 2015]	226	Images of fire incidents in emergency situations, labeled fire and non-fire.
ds-Flickr-Fire [Bedo et al. 2015]	1,984	Images acquired from Flickr, using tags related to fire to filter the information.
ds-Mammoset [Oliveira et al. 2017]	3,457	Regions of Interest obtained from mammograms, with tissue labels such as benign and malignant.
ds-LibraGestures [Bastos et al. 2015]	4,800	Images of hand gestures representing the Brazilian Sign Language (Libras).
ds-Food5k [Singla et al. 2016]	5,000	Images of food (2,500) and non-food (2,500).
ds-FlickrFireSmoke [Cazzolato et al. 2017]	5,556	Images acquired from Flickr, using tags related to fire and smoke to filter the information.
ds-Covid19 [Cohen et al. 2020]	5,933	Chest X-Rays and Computed Tomographies, taken from patients which are positive or suspected of COVID-19 or other viral and bacterial pneumonias.
ds-COIL100 [Nene et al. 2020]	7,200	Images of objects depicted at angles in a 360 rotation, at every 5 degrees.
ds-CUB-200-2011 [Wah et al. 2011]	11,788	Images of 200 bird species acquired from Flickr, where each species is associated with a Wikipedia article and organized by a scientific classification among order, family, genus, and species.
ds-Letters [Hajder 2020]	15,340	Images of standard fonts from Windows, where each letter is organized in classes by typeface.
ds-Cars [Krause et al. 2013]	16,185	Images of cars from 196 classes, including annotations.
ds-Food-11 [Singla et al. 2016]	16,643	Images of food, grouped into 11 categories.
ds-Dogs [Khosla et al. 2011]	20,580	Images of dogs of 120 breeds from around the world.
ds-DeepLesion [Yan et al. 2017]	33,334	Image Slices extracted from Computed Tomographies.
ds-AwA [Xian et al. 2019]	37,322	Images of 50 animals acquired from Flickr, also containing 80 attributes of predicates (*e.g.*, domestic, forest, and arctic).
ds-MNIST [Lecun et al. 1998]	70,000	Images of 10 handwritten digits (0 to 9).
ds-CompCars [Yang et al. 2015]	164,344	Depicts images of cars, taken from two scenarios: web-nature and auto parts.

2.2 Feature Extraction and Vectorial Representation

After curating every dataset, organizing the files and available metadata, we employed feature extraction methods (FEMs) to obtain the visual features from the acquired images. Figure 1 (Steps ii and iii) illustrates this task. Each FEM receives as input an image file and generates a d-dimensional vector, where the features are represented as an array of floats.

We employed the FEMs listed in Table II with the corresponding acronyms, number of dimensions, and types. Notice that every employed FEM generates a specific number of features (dimensions), and corresponds to a specific type T of visual feature, where $T \in$ *{Color, Texture, Shape}*. In the case of NCH, we generated six histogram variations, with 8, 16, 32, 64, 128, and 256 features. As a result, we have 11 distinct FEMs, which can generate 16 different feature configurations.

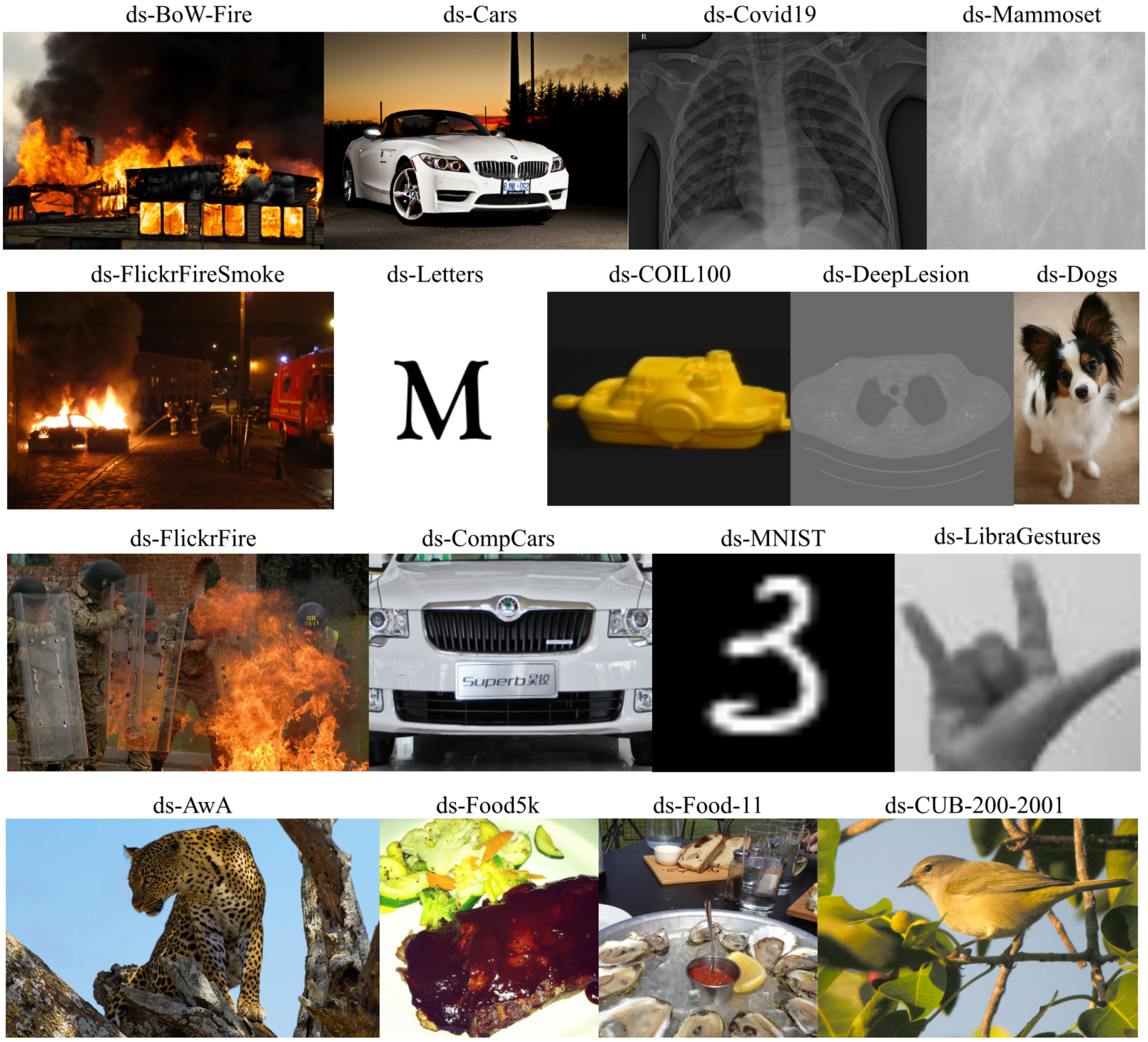


Fig. 2. Examples of images from the public datasets contained in `FeatSet+`.

Table II. Feature Extraction Methods (FEMs) employed, the corresponding acronyms, number of dimensions, and feature type.

FEM	Acronym	#Dimensions	Type
Color Temperature	CT	3	Color
Texture Spectrum	TS	8	Texture
Color Layout	CL	16	Color
Haralick	Hr	24	Texture
Color Structure	CS	128	Color
Edge Histogram	EH	150	Shape
Local Binary Pattern	LBP	177	Texture
Scalable Color	SC	256	Color
BIC Histogram	BIC	512	Color
Total Color Histogram	TCH	768	Color
Normalized Histogram	NCH	8, 16, 32, 64, 128, 256	Color

The implementations of FEMs used in this work are from the Arboretum library, available at the Database and Image Group (GBdI) website[1]. Each FEM receives as the input an image file collected and/or converted during the Preprocessing step. The output is the vectorial representation of the extracted features, composed of real-valued attributes, each one representing a feature of the specific FEM. Most of the Arboretum's available FEMs are from the MPEG-7 Standard [Manjunath et al. 2002], proposed by ISO/IEC JTC1, which aims at building an efficient way to search, filter, and identify multimedia content, defining the expected representations for the images in terms of color, texture, and shape. In this article, we employ the following MPEG-7 extractors: Color Layout, Color Structure, Scalable Color, Color Temperature, Edge Histogram, and Texture Browsing. Briefly, they work as the following:

- **Color Layout**: Describes the image color distribution considering spatial location [Kasutani and Yamada 2001]. It splits the image into squared sub-regions and labels each square with a few nonlinear quantized DCT coefficients of grid-based average colors.
- **Scalable Color**: A color histogram in the HSV color space, which is encoded by a Haar transform and is intended to capture the prominent color distribution [Manjunath et al. 2001].
- **Color Structure**: Aims at capturing both the color content and information about the spatial arrangement of that color content [Sikora 2001]. It works like a histogram that counts how many times a color is present in structures with fixed-size windows. Each fixed-size window selects equally spaced pixels to represent the local color structure. The window size and the number of local structures are parameters of the Color Structure descriptor.
- **Edge Histogram**: Represents the spatial distribution of five types of edges, Vertical, Horizontal, 45 degree, 135 degree and non-directional, regarding $N \times N$ blocks, where N is an extractor parameter [Park et al. 2000], which we employed as $N = 4$ (*i.e.*, 16 blocks). Each block is constructed by partitioning the original image into squared regions and consists of local histograms of these edge directions, which may optionally be aggregated into global or semi-global histograms.
- **Texture Browsing**: This extractor is obtained from the same parameters used in the Gabor filters applied to the images [Lee and Chen 2005]. Its vector consists of 12 positions: 2 for regularity, 6 for directionality, and 4 for coarseness.
- **Color Temperature**: Is based on the hypothesis that there is a correlation between the illumination properties of the image and its "feeling of temperature". CT represents the feature vector as the linearized pixels in the XYZ color space, discarding the luminance of Y channel that is above a given threshold parameter. CT averages the color coordinate in XYZ and converts it to UCS. Finally, CT calculates the two-color isotemperature lines from the color diagrams [Bedo et al. 2015].

BIC (Border/Interior Pixel Classification) [Stehling et al. 2002], TCH, and Normalized histograms describe the grayscale color distribution of the pixels in the image. Haralick is a texture FEM that computes the image dimensions as variances and moments based on co-occurrence matrices. Texture Spectrum and LBP describe the local correlation among grayscale values within pixels.

2.3 Data Description

After curating the public datasets and extracting the visual features, we organized the `FeatSet+` repository (Step iv of Figure 1). Figure 3 shows the generic `FeatSet+` schema, which is similar for every one of the 17 datasets. Figure 3(a) is the metadata table, which includes the object identifier (OID) for each complex object in the dataset, employed as the primary key (PK), the filename of the original image, and the set of classes, if any. The image filename is the same used in the original dataset, allowing reproducibility. The reference to the original images also admits data scientists to download the original images and perform further analysis, such as noise reduction and segmentation.

[1] The Arboretum library is available at `https://gbdi.icmc.usp.br/`, under the "Projects" menu.

Table III details the existing set of classes on `FeatSet+`. The datasets *ds-BoWFire*, *ds-Flickr-Fire*, *ds-LibraGestures*, *ds-Food5k*, *ds-CUB-200-2011*, *ds-Letters*, *ds-Food-11*, *ds-Dogs*, *ds-AwA*, and *ds-MNIST* have a single set of classes, represented by the column `class` in the metadata table. Dataset *ds-FlickrFireSmoke* has two sets of classes, one to denote the presence or absence of fire in the image (column `class_0`), and the other to determine if the image has or not smoke in it (column `class_1`). The remaining datasets do not have classes. Figure 3(b) illustrates the set of FEM tables originated from Section 2.2. Each FEM table has the OID column as a foreign key (FK) to the respective metadata table, and every dimension of the feature vector is stored in a column named `feature_i`, for $0 \leq i \leq d$, where d is the number of dimensions of the current FEM, given by Table II. Figure 3(c) shows the feature reference table *FeatureEquivalence*, which contains the name of each of the d features generated by the employed FEMs. The table has the FEM name, feature ID, and the description of every feature.

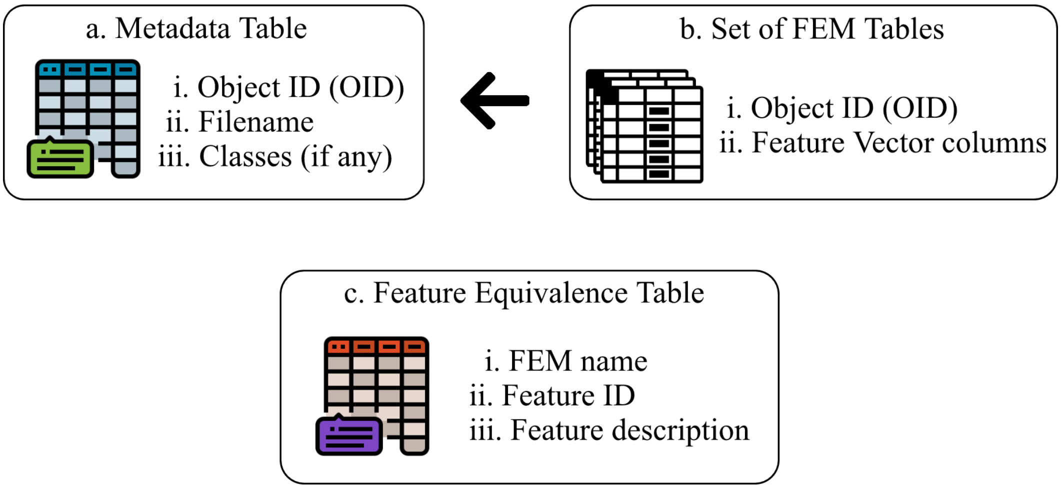


Fig. 3. Schema for `FeatSet+`.

Table III. Existing set of labels (classes) in `FeatSet+`.

Dataset	**Set of Labels (Classes)**
ds-BoWFire [Chino et al. 2015]	Presence/Abscence of Fire
ds-Flickr-Fire [Bedo et al. 2015]	Presence/Abscence of Flame
ds-LibraGestures [Bastos et al. 2015]	Gesture translation
ds-Food5k [Singla et al. 2016]	Food or Non-Food
ds-FlickrFireSmoke [Cazzolato et al. 2017]	Presence of Fire and/or Smoke
ds-CUB-200-2011 [Wah et al. 2011]	200 classes of birds (*e.g.*, Blue Grosbeak, Sayornis, and Bohemian Waxwing)
ds-Letters [Hajder 2020]	Letter's Font
ds-Food-11 [Singla et al. 2016]	11 categories of food: Bread, Dairy product, Dessert, Egg, Fried food, Meat, Noodles/Pasta, Rice, Seafood, Soup, and Vegetable/Fruit
ds-Dogs [Khosla et al. 2011]	Dog's breed
ds-AwA [Xian et al. 2019]	11 animals (*e.g.*, dalmatian, fox, bobcat, and dolphin)
ds-MNIST [Lecun et al. 1998]	A digit from 0 to 9

Both metadata and FEM tables in `FeatSet+` are organized in separated Structured Query Language (SQL) scripts to create and populate those tables. We decided to maintain the separated tables for each dataset within `FeatSet+` because this organization allows users to select only the scripts from the datasets that they want to work with. We also provide Comma-Separated Values (CSV) files for

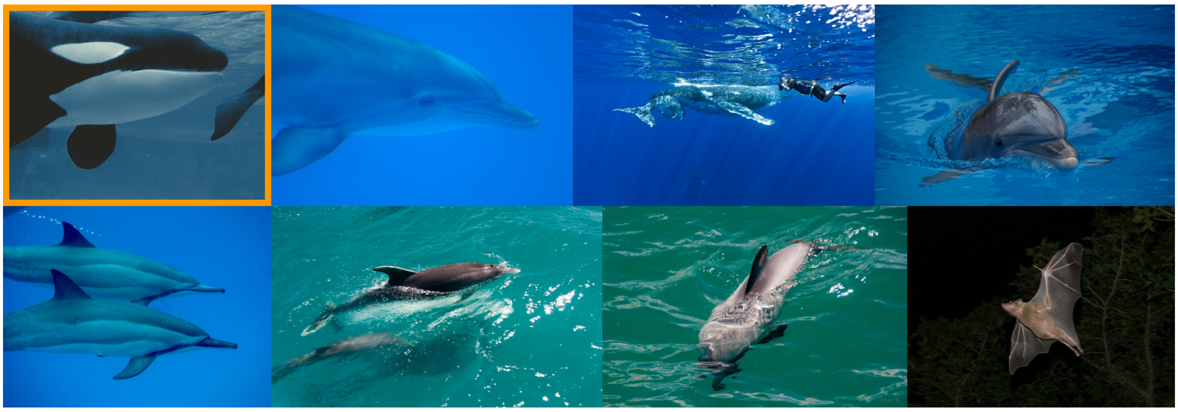


Fig. 4. Multi-attribute Analysis: among the animals containing the predicate "bipedal", we posed a similarity query to check the most similar images. The first image of the sequence was the object used as the query center.

every dataset and FEM, allowing users to input the data into machine learning libraries, manage the features outside the Database Management System (DBMS), and also concatenate the desired files whenever necessary.

3. APPLICABILITY AND CHALLENGES FOR FEATSET+

FeatSet+ opens research opportunities regarding various Computer Science tasks, such as evaluating Content-Based Image Retrieval (CBIR) approaches, machine learning methods, data visualization, and information prediction. As FeatSet+ constitutes a compilation of public datasets acquired from various application contexts, the visual features are well-suited for multidisciplinary studies. In this section, we discuss potential research opportunities. We make the scripts used to generate all analyses described in this section available in the FeatSet+ repository (as detailed in Section 4). Table II describes the acronyms of FEMs mentioned in this section.

3.1 Multi-Attribute Analysis and Retrieval

Several of the available datasets covered by FeatSet+ provide additional information that can be employed as subclasses of the images. For instance, *ds-Mammoset* has, for a set of images acquired in the subset of DDSM repository [Oliveira et al. 2017; Oliveira et al. 2019], four categories based on the view of the breast image, which are: LCC (Left CranioCaudal), RCC (Right CranioCaudal), (iii) LMLO (Left MedioLateral Oblique), and (iv) RMLO (Right MedioLateral Oblique) [Oliveira et al. 2017]. In a transactional database, we can organize the working data with scalar and complex attributes, allowing users to pose queries over them and obtain complementary information. Accordingly, users can filter the view of mammograms from *ds-Mammoset* that they want to analyze (*e.g.*, LMLO), and perform the similarity-based analysis over the filtered set. Another example is the *ds-AwA* dataset, which provided 80 additional attributes, that are predicates regarding the 11 animal categories. For instance, we can filter the objects of the dataset to select only images depicting animals that are "bipedal", resulting in 3,706 tuples. In the next step, we can query over the filtered objects considering the similarity between the images. Figure 4 shows an example of a query posed over images of animals filtered with the predicate "bipedal". The first image was selected as the query center of a kNN search, with $k = 8$ and using the CL FEM.

3.2 Visual Analysis

The Multidimensional Scaling (MDS) method represents the (dis)similarity among objects onto a projection of the data in a low-dimensional space [Borg and Groenen 2005]. Here, we employed MDS to show the advantage of representing the image datasets using diverse visual representations. The color represents the label of the images ("Fire" and "Not Fire"). Figure 5(a) shows the distance space distribution formed by an image sample from *ds-BoWFire*, using the different visual features provided by `FeatSet+`. We observe major differences in the data dispersion from the generated distance spaces. For instance, CL shows objects dispersed almost homogeneously, while CT depicts the objects in a "line-shaped" dispersion. In contrast, Figure 5(b) shows the 1st and 2nd Principal Components (PCs) obtained by applying the Principal Component Analysis (PCA) over the dataset. The two PCs capture most of the data variance, and the visualization shows the data dispersion on a lower-dimensional projection. Both plots (a) and (b) consider the same image sample, and we observe distinct class dispersions. This result highlights that further opportunities exist for further analysis targeting to find better separations between the classes of data. Next, we further analyze the first eight representations of *ds-BoWFire* in the classification task.

3.3 Global Feature Classification

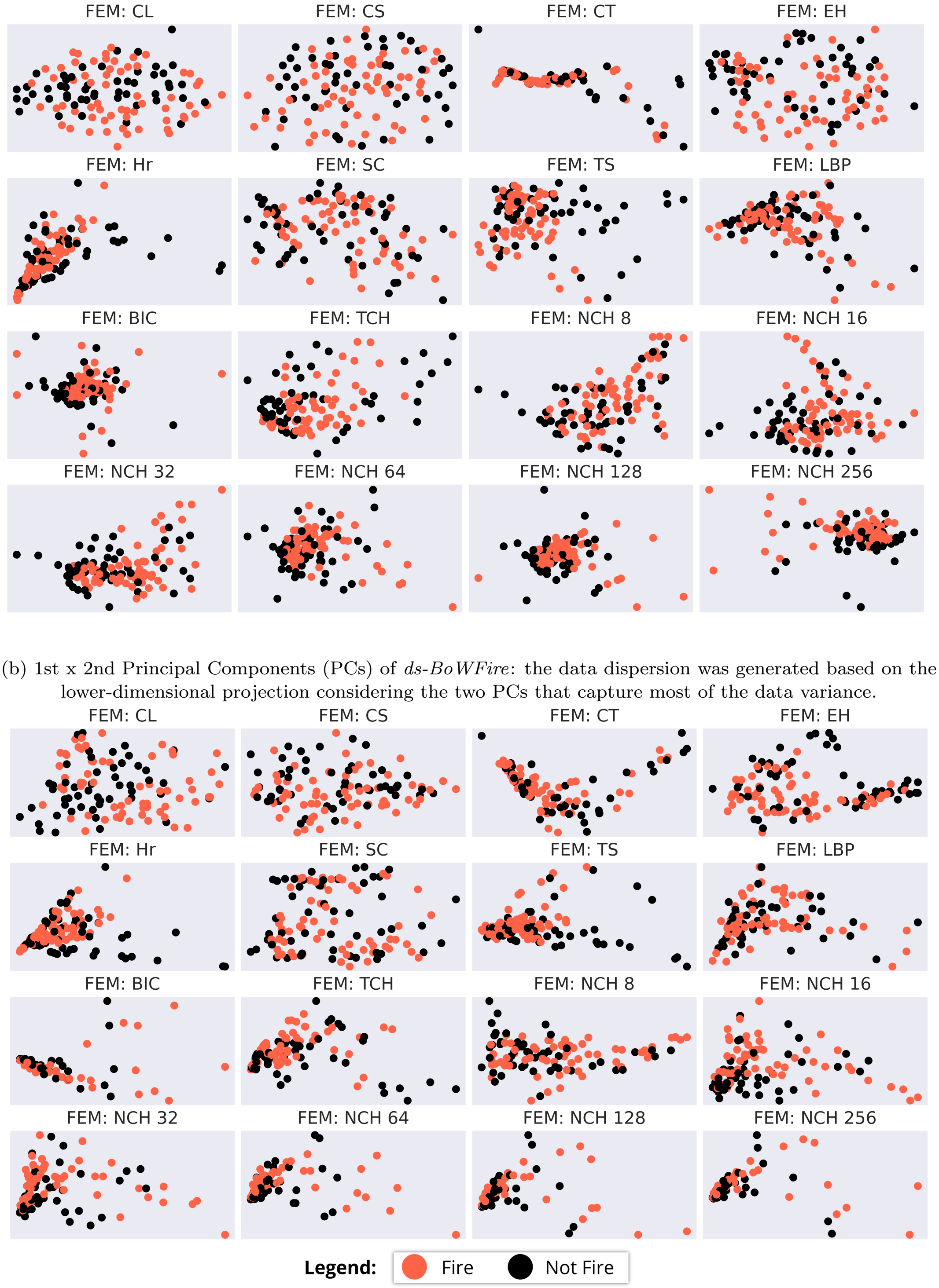
We selected a set of off-the-shelf classifiers to evaluate the impact of different feature representations in the classification precision. Figure 6 shows the precision results for 14 conventional classifiers. We observe that the features CS, Hr, TS, TCH, and LBP achieved the highest classification results among the employed FEMs. We also observe cases with high variation results for different classifiers in the same FEM representation with SC.

Figure 7 shows the precision results when classifying the different datasets covered by `FeatSet+`, when employing the CL FEM. In (a), we can observe the variation of the precision obtained by each classifier regarding every dataset. In (b), we identified that the datasets presenting large sets of labels (such as *ds-AwA*, *ds-CUB-200-2011* and *ds-Dogs*) presented poor precision results. The results indicate the need of further analysis, such as feature selection, to improve the classification quality.

3.4 Principal Components Analysis

Many of the employed FEMs, such as BIC, TCH, and NCH, produce high-dimensional feature vectors (see Table II). Also, in many application scenarios, the analyst may opt to combine features of different visual characteristics to improve the semantics of the image representation. For example, if we consider *ds-Mammoset*, microcalcifications can show different color and texture patterns that, when combined, allow a more profound pattern recognition. Feature concatenation can improve data representation but has the cost of increasing the data dimensionality, which can be approached in different manners. One example is the application of the Principal Component Analysis (PCA) to reduce data dimensionality.

Figure 8 shows the proportion of explained variances according to the principal components generated by PCA. In the examples, we selected four datasets from `FeatSet+`, and explored three single feature representations (a, b, and c) and the combination of LBP with CL (d). To improve the visualization, we plotted a maximum of 30 principal components. The dashed horizontal lines represent curve elbows, visually observed in the plots. The curve elbow can be used as a heuristic to select the number of principal components to employ in data analysis, selecting the position where the curve stops decreasing and flattens out. Using this criterion, in (a) we could use 7 or 10 principal components, in (b) 6 or 13, in (c) 7, and in (d) 8 principal components. One can also consider the dimensions whose sum is at least a threshold, for example, 70% of the entire variance. In this case, in plot (a), we would select the first 18 principal components, in (b) the first 8, in (c) the first 6, and in (d) the first five principal components. Regardless of the employed heuristic, the selected visual


(a) MDS distribution of *ds-BoWFire*: the spatial distance among objects is preserved.

(b) 1st x 2nd Principal Components (PCs) of *ds-BoWFire*: the data dispersion was generated based on the lower-dimensional projection considering the two PCs that capture most of the data variance.

Fig. 5. Comparing the visual analysis of the class distributions in *ds-BoWFire*: (a) The MDS plot depicts the dispersion of objects within the various (original) distance space distributions, generated with different visual features. (b) Two-dimensional plots of PCA showing different distance dispersion obtained over the same set of images.

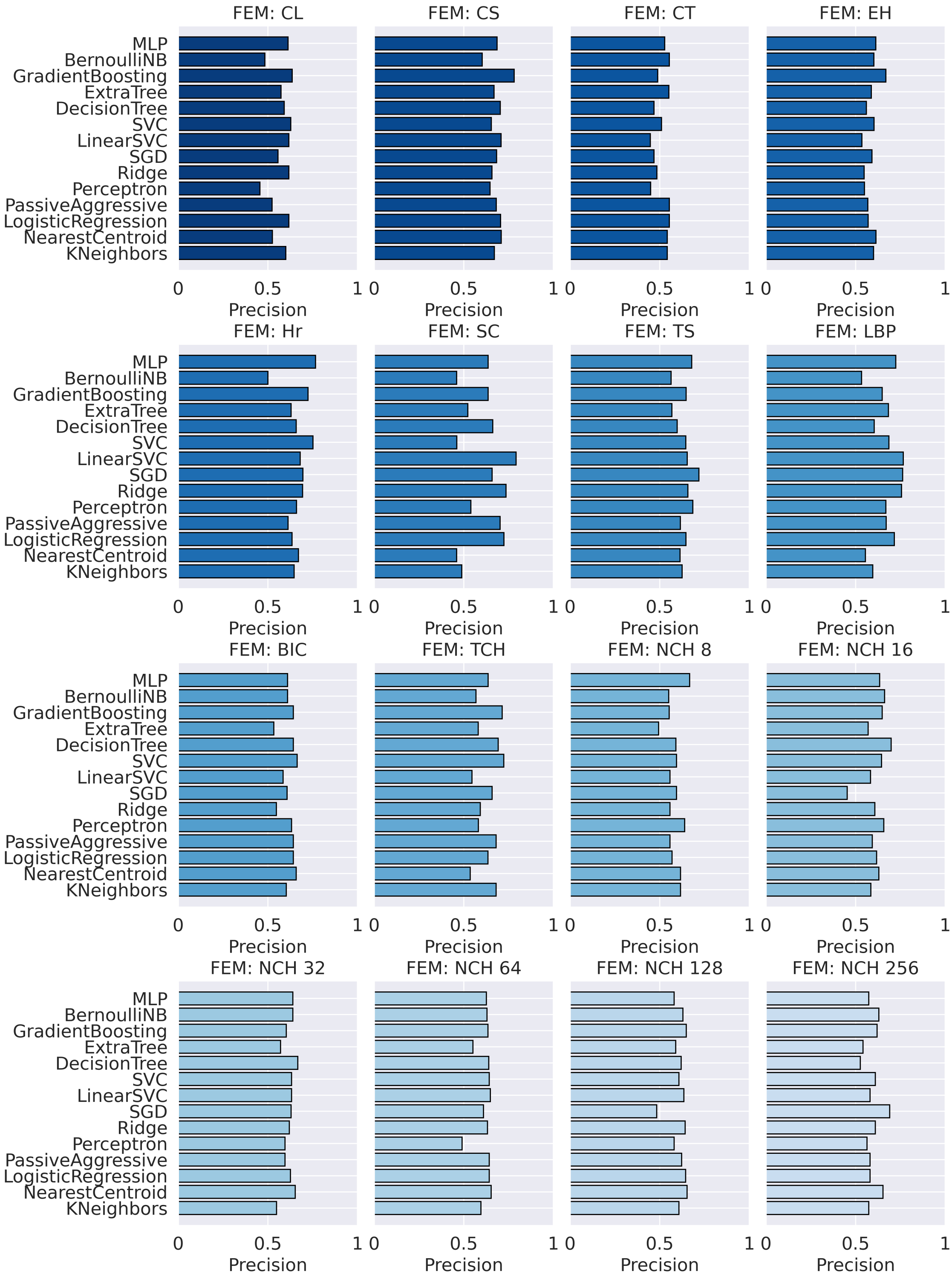


Fig. 6. Comparison of the precision results for 14 different approaches of classification techniques, considering the *ds-BoWFire* dataset, for all available FEMs.

Symbols and colors represent the dataset as shown in (b).

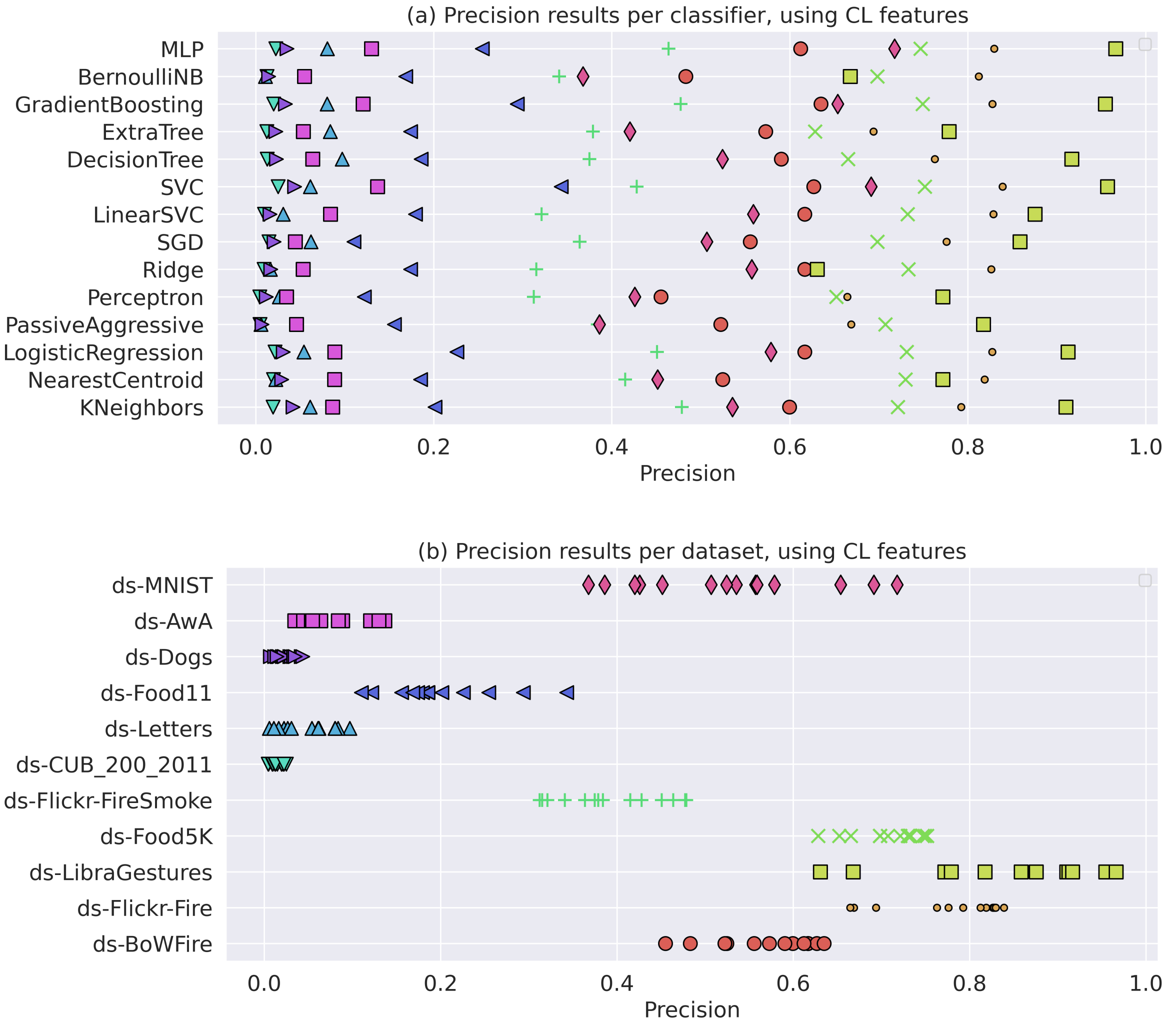


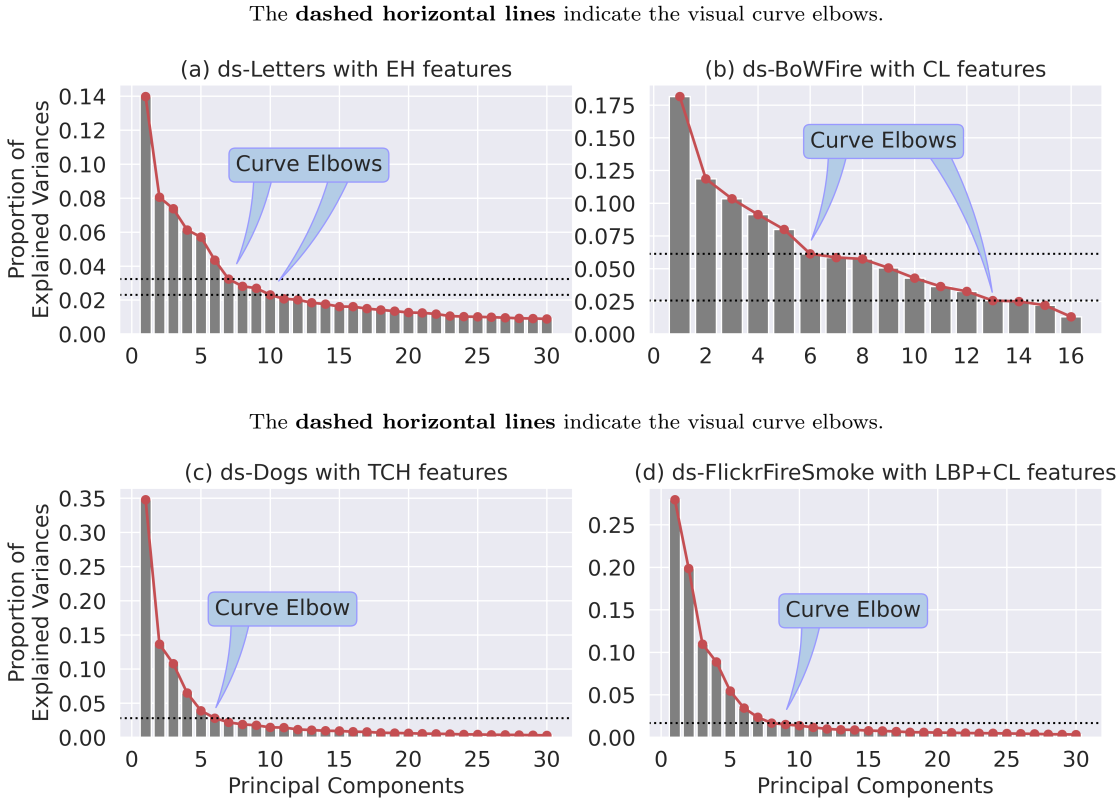
Fig. 7. Comparison of the precision for 14 different approaches of classification techniques, considering the set of 11 labeled datasets within FeatSet+ (see Table III), using Color Layout visual features.

features can be used to further analyze the complex objects by employing CBIR or machine learning methods. Although PCA was used here as a dimensionality reduction technique when concatenating feature vectors, it is also an example of how feature analysis layers can be stacked to develop deep learning models for image feature engineering.

3.5 Challenges and Opportunities of Analysis

FeatSet+ comprises small and large datasets (ranging from 226 to 164,344 objects), and FEMs of low and high dimensionality (from 3 to 768 dimensions). The variation in size and dimensionality can support the validation of techniques focused on content-based retrieval and diversity, complex data indexing with metric access methods, and similar methods. Similarity-based comparisons of complex objects can also support the identification of near-duplicate images through feature-matching.

The dataset organization allows straightforwardly employing machine learning methods since they


The **dashed horizontal lines** indicate the visual curve elbows.

The **dashed horizontal lines** indicate the visual curve elbows.

Fig. 8. The scree plots show the proportion of explainable variances according to the principal components generated by PCA.

already are in the input format of many existing analysis libraries, such as Scikit Learn[2] for Python. Users can perform classification and clustering methods, compute correlations among the different data representations, perform object recognition, among others. `FeatSet+`'s schema also allows users to work with the visual features inside the Database Management System (DBMS) by loading the available database files provided in the Git repository.

`FeatSet+` can be further extended by extracting new visual features with other FEMs reported in the literature. Finally, users can include new image datasets into `FeatSet+`, extracting their features using the FEMs reported in this work, which are openly available at the Arboretum library.

4. PUBLIC REPOSITORY AND CITATION REQUEST

`FeatSet+` is publicly available for research use under the Creative Commons license. All data and additional information are organized in the Git repository *https://github.com/mtcazzolato/featset/*. The repository is organized as follows:

—`FeatSet+`/
  - —*README.md*: Read me file with `FeatSet+` description, citation instructions for every part of the dataset, and other relevant information.
  - —*SQL-Scripts-Link*: Link to download the SQL scripts used to load the data.
  - —*CSV-File-Link*: Link to download the CSV files with the data.

[2] The Scikit Learn Python library: https://scikit-learn.org/stable/

—*python-scripts/:* Folder with the three python scripts that generate:
  —The application analyses of Section 3; the plots presented in Figures 4, 5, 6, 7, and 8; and the images of the *ds-MNIST* dataset.
—*schema.png*: The database schema.

Each one of the 17 datasets presented in Table I follows the schema illustrated in Figure 3. The metadata of each dataset is stored in an SQL file with the dataset name (*e.g.*, ***ds-CompCars.sql***), every FEM is inside of a file with the same name plus its respective acronym as a suffix (*e.g.*, ***ds-CompCars_CL.sql*** for Color Layout). The file *FeatureEquivalence* is the reference table to describe the features of the available FEMs, containing the columns *FEM_ name*, *Feature_ ID*, and *Feature_ description.* All SQL scripts start with the `CREATE TABLE` statement, followed by the `INSERT INTO` statements to populate those tables.

Additionally, the `FeatSet+` repository also provides a Comma-Separated Values (CSV) file alternative for every single table from the 17 datasets and the feature reference table *FeatureEquivalence*. `FeatSet+` is available for researchers and data scientists under the Creative Commons License. In case of publication of any work derived from any of the datasets from `FeatSet+`, we ask the authors to acknowledge the original image dataset owners and us by citing both the image dataset reference and this paper, following the instructions of the provided `README.md` file of our repository.

5. CONCLUSION

In this work, we proposed the `FeatSet+` dataset, a compilation of visual features extracted from public image datasets of different application scenarios. `FeatSet+` is composed of 17 datasets and has 11 visual features representing the images. We provided four examples of how the feature vectors can be explored for different computational tasks. Also, the public datasets inside `FeatSet+` are from diverse application domains, which can aid analysts in the evaluation of their techniques in a wide range of examples. `FeatSet+` is organized in a public repository and is available in SQL scripts to load the database in a DBMS and in CSV files to be used directly in existing data analysis libraries.

As future work, we intend to (i) perform hyperparameter tuning over off-the-shelf classifiers explored in Figure 6 to improve the overall precision results. The obtained results can work as starting points for users of `FeatSet+`, in case they want to improve classification results. Also, since `FeatSet+` is limited to the provided 17 datasets and 11 visual features, we intend to (ii) create an automated process to include novel open image databases into `FeatSet+` while reducing manual effort. At the same time, we intend to (iii) add other embedding approaches employing Deep-Learning techniques for transfer learning, such as VGG-16 [Simonyan and Zisserman 2015]. Finally, it is worth mentioning that our Preprocessing step does not enhance image quality (*e.g.*, brightness, contrast, noise removal). To overcome this limitation, users must preprocess the original images and then extract the wanted visual features.

Acknowledgment

This research was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, by the São Paulo Research Foundation (FAPESP, grants No. 2016/17078-0, 2020/07200-9, 2020/11258-2, 2021/11403-5), and by the National Council for Scientific and Technological Development (CNPq).

REFERENCES

Bastos, I. L. O., Angelo, M. F., and Loula, A. C. Recognition of static gestures applied to brazilian sign language (Libras). In *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*. pp. 305–312, 2015. DOI: 10.1109/SIBGRAPI.2015.26.

Bedo, M. V. N., Blanco, G., Oliveira, W. D., Cazzolato, M. T., Costa, A. F., Rodrigues-Jr., J. F., Traina, A. J. M., and Traina Jr., C. Techniques for effective and efficient fire detection from social media images. In *ICEIS 2015 - Proceedings of the 17th International Conference on Enterprise Information Systems, Volume 1, Barcelona, Spain, 27-30 April, 2015*, S. Hammoudi, L. A. Maciaszek, and E. Teniente (Eds.). SciTePress, pp. 34–45, 2015. DOI: 10.5220/0005341500340045.

Borg, I. and Groenen, P. *Modern Multidimensional Scaling: Theory and Applications.* Springer Series in Statistics. Springer New York, 2005. ISBN: 9780387251509.

Cazzolato, M., Scabora, L. C., Zabot, G. F., Gutierrez, M. A., Traina-Jr., C., and Traina, A. J. M. FeatSet: A compilation of visual features extracted from public image datasets. In *Anais do III Dataset Showcase Workshop.* SBC, Porto Alegre, RS, Brasil, pp. 89–100, 2021. DOI: 10.5753/dsw.2021.17417.

Cazzolato, M. T., Avalhais, L. P. S., Chino, D. Y. T., Ramos, J. S., Souza, J. A., Rodrigues-Jr, J. F., and Traina, A. J. M. FiSmo: A compilation of datasets from emergency situations for fire and smoke analysis. In *SBBD2017 - SBBD Proceedings of Satellite Events of the 32nd Brazilian Symposium on Databases - DSW (Dataset Showcase Workshop).* SBC, Uberlandia, Brazil, pp. 213–223, 2017. ISBN: 978-85-7669-399-4.URL: sbbd.org.br/2017/wp-content/uploads/sites/3/2017/10/proceedings-satellite-events-sbbd-2017.pdf.

Cazzolato, M. T., Bedo, M. V. N., Costa, A. F., de Souza, J. A., Jr., C. T., Jr., J. F. R., and Traina, A. J. M. Unveiling smoke in social images with the smokeblock approach. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing, Pisa, Italy, April 4-8, 2016*, S. Ossowski (Ed.). ACM, pp. 49–54, 2016. DOI: 10.1145/2851613.2851634.

Chino, D. Y. T., Avalhais, L. P. S., Rodrigues-Jr., J. F., and Traina, A. J. M. BoWFire: Detection of fire in still images by integrating pixel color and texture analysis. In *28th SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2015, Salvador, Bahia, Brazil, August 26-29, 2015.* IEEE Computer Society, pp. 95–102, 2015. DOI: 10.1109/SIBGRAPI.2015.19.

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., and Ghassemi, M. Covid-19 image data collection: Prospective predictions are the future. *CoRR* vol. abs/2006.11988, 2020. URL: https://arxiv.org/abs/2006.11988.

de Sousa Fogaça, I. C. O. and Bueno, R. Temporal evolution of complex data. In *Anais do XXXV Simpósio Brasileiro de Bancos de Dados, SBBD 2020, online, September 28 - October 1, 2020.* SBC, pp. 25–36, 2020. DOI: 10.5753/sbbd.2020.13622.

Hajder, S. *Letters organized by typefaces*, 2020. Last accessed in October, 2020. URL: https://www.kaggle.com/killen/bw-font-typefaces?select=BRLNSR.

Kasutani, E. and Yamada, A. The mpeg-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205).* Vol. 1. pp. 674–677 vol.1, 2001. DOI: 10.1109/ICIP.2001.959135.

Khosla, A., Jayadevaprakash, N., Yao, B., and Fei-Fei, L. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition.* Colorado Springs, CO, 2011. URL: https://people.csail.mit.edu/khosla/papers/fgvc2011.pdf.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013.* IEEE Computer Society, pp. 554–561, 2013. DOI: 10.1109/ICCVW.2013.77.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86 (11): 2278–2324, 1998. DOI: 10.1109/5.726791.

Lee, K.-L. and Chen, L.-H. An efficient computation method for the texture browsing descriptor of mpeg-7. *Image and Vision Computing* vol. 23, pp. 479–489, 05, 2005. DOI: 10.1016/j.imavis.2004.12.002.

Maheshwari, S., Sharma, R. R., and Kumar, M. LBP-based information assisted intelligent system for COVID-19 identification. *Comput. Biol. Medicine* vol. 134, pp. 104453, 2021. DOI: 10.1016/j.compbiomed.2021.104453.

Manjunath, B., Ohm, J., Vasudevan, V., and Yamada, A. Color and texture descriptors. *Circuits and Systems for Video Technology, IEEE Transactions on* vol. 11, pp. 703 – 715, 07, 2001. DOI: 10.1109/76.927424.

Manjunath, B. S., Salembier, P., and Sikora, T. *Introduction to MPEG-7: multimedia content description interface.* John Wiley & Sons, 2002. ISBN: 978-0-471-48678-7.

Moriyama, A., Rodrigues, L. S., Scabora, L. C., Cazzolato, M. T., Traina, A. J. M., and Traina, C. Vd-tree: how to build an efficient and fit metric access method using voronoi diagrams. In *SAC '21: The 36th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, Republic of Korea, March 22-26, 2021*, C. Hung, J. Hong, A. Bechini, and E. Song (Eds.). ACM, pp. 327–335, 2021. DOI: 10.1145/3412841.3441915.

Nene, S. A., Nayar, S. K., and Murase, H. Columbia object image library (COIL-100). Tech. rep., Technical Report CUCS-006-96, 2020. Last accessed in October, 2020.

Oliveira, P. H., Scabora, L. C., Cazzolato, M. T., Bedo, M. V. N., Traina, A. J. M., and Traina-Jr., C. MAMMOSET: An Enhanced Dataset of Mammograms. In *Proceedings of the Satellite Events of the 32nd Brazilian Symposium on Databases.* SBC, pp. 256–266, 2017. URL: https://sbbd.org.br/2017/wp-content/uploads/sites/3/2017/10/proceedings-satellite-events-sbbd-2017.pdf.

Oliveira, P. H., Scabora, L. C., Cazzolato, M. T., Oliveira, W. D., Paixão, R. S., Traina, A. J. M., and Traina, C. Employing domain indexes to efficiently query medical data from multiple repositories. *IEEE J. Biomed. Health Informatics* 23 (6): 2220–2229, 2019. DOI: 10.1109/JBHI.2018.2881381.

Park, D. K., Jeon, Y. S., and Won, C. S. Efficient use of local edge histogram descriptor. In *Proceedings of the ACM Multimedia 2000 Workshops, Los Angeles, CA, USA, October 30 - November 3, 2000*, S. Ghandeharizadeh, S. Chang, S. Fischer, J. A. Konstan, and K. Nahrstedt (Eds.). ACM Press, pp. 51–54, 2000. DOI: 10.1145/357744.357758.

Pereira, J. W. and Ribeiro, M. X. Semantic annotation and classification of mammography images using ontologies. In *34th IEEE International Symposium on Computer-Based Medical Systems, CBMS 2021, Aveiro, Portugal, June 7-9, 2021*, J. R. Almeida, A. R. González, L. Shen, B. Kane, A. Traina, P. Soda, and J. L. Oliveira (Eds.). IEEE, pp. 378–383, 2021. DOI: 10.1109/CBMS52027.2021.00043.

Rodrigues, L. S., Cazzolato, M. T., Traina, A. J. M., and Traina, C. Taking advantage of highly-correlated attributes in similarity queries with missing values. In *Similarity Search and Applications - 13th International Conference, SISAP 2020, Copenhagen, Denmark, September 30 - October 2, 2020, Proceedings*, S. Satoh, L. Vadicamo, A. Zimek, F. Carrara, I. Bartolini, M. Aumüller, B. Þ. Jónsson, and R. Pagh (Eds.). Lecture Notes in Computer Science, vol. 12440. Springer, pp. 168–176, 2020. DOI: 10.1007/978-3-030-60936-8_13.

Sikora, T. The mpeg-7 visual standard for content description-an overview. *IEEE Transactions on Circuits and Systems for Video Technology* 11 (6): 696–702, 2001. DOI: 10.1109/76.927422.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun (Eds.), 2015.

Singla, A., Yuan, L., and Ebrahimi, T. Food/non-food image classification and food categorization using pre-trained googlenet model. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. MADiMa '16. Association for Computing Machinery, New York, NY, USA, pp. 3–11, 2016. DOI: 10.1145/2986035.2986039.

Stehling, R. O., Nascimento, M. A., and Falcão, A. X. A compact and efficient image retrieval approach based on border/interior pixel classification. In *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 4-9, 2002*. ACM, pp. 102–109, 2002. DOI: 10.1145/584792.584812.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* 41 (9): 2251–2265, 2019. DOI: 10.1109/TPAMI.2018.2857768.

Yan, K., Wang, X., Lu, L., and Summers, R. M. DeepLesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations. *CoRR* vol. abs/1710.01766, 2017. URL: http://arxiv.org/abs/1710.01766.

Yang, L., Luo, P., Loy, C. C., and Tang, X. A large-scale car dataset for fine-grained categorization and verification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, pp. 3973–3981, 2015. DOI: 10.1109/CVPR.2015.7299023.

Zabot, G. F., Cazzolato, M. T., Scabora, L. C., Faiçal, B. S., Traina, A. J. M., and Traina Jr., C. UCORM: indexing uncorrelated metric spaces for concise content-based retrieval of medical images. In *32nd IEEE International Symposium on Computer-Based Medical Systems, CBMS 2019, Cordoba, Spain, June 5-7, 2019*. IEEE, pp. 306–311, 2019. DOI: 10.1109/CBMS.2019.00070.

Zabot, G. F., Cazzolato, M. T., Scabora, L. C., Traina, A. J. M., and Traina Jr., C. Efficient indexing of multiple metric spaces with spectra. In *IEEE International Symposium on Multimedia, ISM 2019, San Diego, CA, USA, December 9-11, 2019*. IEEE, pp. 169–176, 2019. DOI: 10.1109/ISM46123.2019.00038.