Contemporary Mathematics
https://ojs.wiserpub.com/index.php/CM/

Research Article

A Non Negative Matrix Factorization-Based Data Augmentation Procedure for Two-Dimensional Data

Serena Crisci[1], Valentina De Simone[1,2], Ferdinando Zullo[1*]

[1] Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Viale Lincoln 5, Caserta, 81012, Italy
[2] Institute for High-Performance Computing and Networking, CNR, Via Pietro Castellino 111, Naples, 80131, Italy
E-mail: ferdinando.zullog@unicampania.it

Received: 29 July 2025; **Revised:** 11 November 2025; **Accepted:** 11 November 2025

Abstract: In this paper, we investigate how geometric transformations—such as translations and rotations—affect matrix factorizations, with a particular focus on Non negative Matrix Factorization (NMF) in the context of supervised learning. We describe a novel feature extraction and data augmentation framework that leverages the invariance properties of matrix decompositions under linear transformations. Specifically, we show how applying such transformations in the input space induces systematic variations in the factorization structure, which we exploit to generate new feature vectors from the left factors of NMF applied to transformed data. This provides both augmented training examples and their interpretable non negative representations. Our approach thus enhances feature interpretability while preserving nonnegativity and structure. Preliminary numerical experiments on a binary image classification task related to archaeological data demonstrate the effectiveness of the method.

***Keywords*:** non negative matrix factorization, data augmentation, geometric transformation, glyph image

MSC: 65D18, 68U10, 94A08

1. Introduction

The emergence of machine learning and deep learning techniques opens a huge opportunity for their implementation in real-world scenarios. One of the tasks for which these techniques have the most significant potential is visual inspection since both of these techniques can determine relationships between large volumes of data. However, large volumes of images are often not available. As a solution to this, data augmentation techniques are applied as a powerful tool to expand the diversity and volume of training data, thereby enhancing the robustness and generalization capabilities of machine learning models. Data augmentation involves generating synthetic data by applying various transformations to existing data samples while preserving their semantic meaning.

Transformations play a crucial role in data augmentation by enriching the training data with various and realistic variations that preserve the semantic meaning of the data. For example, [1] presents a comprehensive survey of data augmentation techniques across multiple data modalities. It introduces a unified taxonomy that highlights how the intrinsic relationships between and within data instances can be exploited, providing a broader perspective that goes beyond modality-specific or operation-centric approaches. [2] provides a comprehensive review of image augmentation

DOI: https://doi.org/10.37256/cm.7220268042

techniques for deep learning, organized into a novel taxonomy. It classifies methods into model-free, model-based, and optimizing policy-based approaches, offering insights into their underlying principles and helping guide the choice or design of suitable augmentation strategies for computer vision tasks. [3] explores data augmentation specifically for text categorization by combining semantic-enriched representations with augmented training data. The study demonstrates how transformations that preserve semantic meaning can improve model performance, particularly in domains with limited labeled text data. [4] reviews a wide range of data augmentation techniques across images, text, and signal data, emphasizing how these methods increase dataset size and diversity to improve model generalization and reduce overfitting. The study also discusses the use of Generative Adversarial Networks (GAN)-based augmentation, highlighting practical strategies to enhance the robustness and accuracy of machine learning models in settings with limited or imbalanced data. Therefore, in image data augmentation, rotating or flipping an image should not distort its content to the extent that it becomes unrecognizable. Generally, the choice of transformations depends on the characteristics of the data, the machine learning task, and the desired augmentation strategy to improve model performance and generalization. In particular, we focus on image data augmentation. Traditional approaches are based on basic image manipulations of the images directly and are easy to implement. Following [5], they are typically divided into:

• Geometric transformations: right or left rotation on an axis between 1° and 359°, horizontal or vertical axis shearing, translation, scaling or resizing, mirroring, reflection, horizontal or vertical axis flipping.

• Color space transformations: changing Red, Green, and Blue (RGB) color values with matrix operations to increase or decrease brightness, contrast, or lighting.

• Kernel filters: changing the sharpness or blur of the image by using sliding window matrices.

• Random erasing: randomly selecting a patch of an image and masking it with fixed, mean or random pixel values.

• Mixing images: blending and mixing multiple images.

Among these, geometric transformations stand out as one of the most prevalent approaches. They aim to modify the geometric arrangement of images, displacing image pixels from their initial positions to new locations while preserving the original pixel values. These transformations enhance the training data by simulating real-world alterations in appearance, including variations in viewpoint, non-rigid deformations, perspective shifts, and changes in scale. Recently, many studies have focused on improving training data by increasing the diversity of extracted feature maps. A typical approach to achieving diverse feature representations is directly manipulating feature vectors within intermediate layers of deep neural networks (see, e.g., [6]).

Data augmentation techniques can be particularly useful to support the work of archaeologists in the context of the preservation and analysis of cultural heritage. This branch of research poses complex problems that are deeply and intrinsically interdisciplinary Indeed, in the last few years, new approaches and technologies have been increasingly utilized to support the art historians, architects and restorers. Machine learning and data-driven approaches can aid in analyzing cultural information and determining the necessary measures to ensure its interpretation and preservation for the future. However, the lack of large datasets is one of the main hurdles to face in cultural heritage and archaeological applications.

In this context, assuming that some data can be represented as a 2D array, we develop a data augmentation procedure based on Non negative Matrix Factorization (NMF) of the starting data. NMF has been widely considered for analyzing high-dimensional data, as it automatically extracts sparse and meaningful features from non negative data vectors while maintaining the interpretability and non-negativity of the components. Initially introduced by Paatero and Tapper in 1994 [7], NMF achieved popularity following the article [8] by Lee and Seung in 1999. It has gained prominence because of its ability to uncover latent structures, patterns, and features in data, mainly when the data inherently represent additive combinations, such as images, text, or gene expression data. These features have made it the powerful method in the machine learning area [9–11]. Furthermore, numerical approaches for solving matrix equations, provide a theoretical foundation for the manipulation and transformation of covariance structures, a conceptually related process to the feature extraction and noise reduction steps performed in NMF-based data augmentation.

In this paper, we exploit NMF to extract the main features of the data, which are given by the columns of the left factors of the decompositions, and then we apply linear transformations to these factors in order to obtain a new example

and its non negative factorization, simultaneously. This is achievable by exploiting the invariant properties of matrix decomposition under the action of translations and rotations.

Differently from conventional geometric transformation-based augmentation techniques, the proposed method provides two key benefits:

(1) Starting from a single factorization, it generates both new images and their corresponding decompositions, encoding meaningful features that can be exploited in neural network architectures, thereby increasing feature diversity and reducing the risk of overfitting [12, 13].

(2) It enables a more compact representation of the dataset, optimizing storage and computational efficiency while maintaining a rich variety of training examples that improve model generalization.

2. Some geometric transformations and factorization of matrices

This section is devoted to investigating the invariant properties of matrix decomposition under the action of affine transformations, namely translations and rotations. The analysis is quite general in the sense that we are not assuming symmetry properties or other prior information on the geometrical structure of the input data. We will then analyze in Section 3 some consequences on the NMF.

We will now see how translations or a rotation of 90, 180, and, 270 degrees affect a factorization of $M \in \mathbb{R}^{n\times m}$.

We start by describing the translations. Let $i \in \{0, \ldots, n\}$ and $j \in \{0, \ldots, m\}$. The **translation** of i rows and j columns is the map $\tau_{i,j}$ that acts on the matrices in $\mathbb{R}^{n\times m}$ in such a way that a matrix M associates the matrix M' having in the position (a', b') the element $M_{a,b}$, where

$$a' = \begin{cases} a+i & \text{if } a+i \le n \\ \\ a+i-n & \text{otherwise,} \end{cases} \quad \text{and} \quad b' = \begin{cases} b+j & \text{if } b+j \le m \\ \\ b+j-m & \text{otherwise.} \end{cases}$$

Clearly, the map $\tau_{i,j}$ satisfies the following property.

Proposition 2.1 For any $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$, the map $\tau_{i,j} : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ is linear.

Given a matrix M, we denote by $M^{<i,j>}$ (respectively $M^{<-i,-j>}$) the matrix obtained by applying a circular shift to the first column and row of M moving them according to the horizontal index i (resp. $-i$), and vertical index j, (resp. $-j$), that is, $\tau_{i,j}(M) = M^{<i,j>}$.

We will also use the following notation:

• m_i denotes the i-th row of M;

• m^i denotes the i-th column of M.

Let

$$I_m = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{m\times m},$$

and

$$A_m = I_m^{<0,\,1>} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{m\times m}.$$

It is easy to see that $MA_m = [\mathbf{m}^n, \mathbf{m}^1, \ldots, \mathbf{m}^{n-1}] = M^{<0,\,1>}$ and

$$A_n^T M = \begin{bmatrix} \mathbf{m}_m \\ \mathbf{m}_1 \\ \vdots \\ \mathbf{m}_{m-1} \end{bmatrix} = M^{<1,\,0>}.$$

That is, MA_m represents a column cyclic shift (that is, $\tau_{0,\,1}(M)$) and $A_n^T M$ is a row cyclic shift (that is $\tau_{1,\,0}(M)$).

Therefore, combining the action of $\tau_{0,\,1}$ and $\tau_{1,\,0}$ we have the following interpretation of $\tau_{i,\,j}$ in terms of matrix product.

Proposition 2.2 Let $M \in \mathbb{R}^{n\times m}$ and let $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$. Then

$$\tau_{i,\,j}(M) = (A_n^T)^i M A_m^j, \tag{1}$$

where A^i denotes the i-th power of the matrix A.

Proof. The right translation of i positions in M is given by MA_m^i and the downward translation of i positions is given by $(A_n^T)^i M$. The assertion is proved as $\tau_{i,\,j}(M) = \tau_{0,\,j}(\tau_{i,\,0}(M))$. □

Suppose that $m = n$. We define the **rotation** of 90 degrees ρ of a matrix $M = \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \vdots \\ \mathbf{m}_n \end{bmatrix} \in \mathbb{R}^{n\times n}$ the matrix given as follows

$$\rho(M) = [\mathbf{m}_n^T,\ \cdots,\ \mathbf{m}_2^T,\ \mathbf{m}_1^T] \in \mathbb{R}^{n\times n},$$

whose fixed point is the center of the matrix, that is, if n is odd the only fixed position is $((n+1)/2, (n+1)/2)$ and if n is even there is no fixed position and the center of the rotation can be seen as the geometric center of the matrix. Moreover, also note that ρ^4 is the identity map.

Example 2.3 Let $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbb{R}^{2\times 2}$, then $\rho(M) = \begin{pmatrix} c & a \\ d & b \end{pmatrix}$. Let $M' = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \in \mathbb{R}^{3\times 3}$ then $\rho(M') = \begin{pmatrix} g & d & a \\ h & e & b \\ i & f & c \end{pmatrix}$. Note that in the first case, there is no fixed entry in the matrix M, whereas in the second case the only fixed entry is in the position $(2, 2)$.

Similarly to the case of translations, it is easy to see that ρ is a linear map in $\mathbb{R}^{n\times n}$.

One can see that

$$\rho(M) = M^T B_n, \tag{2}$$

where

$$B_n = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ & \vdots & & \\ 1 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

The rotations of degree 180 and 270 correspond to ρ^2 and ρ^3 and we have that

$$\rho^2(M) = (M^T B_n)^T B_n = B_n^T M B_n \text{ and } \rho^3(M) = (B_n^T M B_n)^T B_n = B_n^T M^T B_n^2.$$

By making use of translations, we can apply rotation having as fixed point any entry of the matrix.

Proposition 2.4 Suppose that n is odd and let $l = (n+1)/2$. The rotation $\rho_{i,j}$ of a matrix M of 90 degrees and centered in (i, j) is

$$\rho_{i,j}(M) = \tau_{a'',b''}(\rho(\tau_{a',b'}(M))) = (A_n^{b'} A_n^{a''})^T M^T (A_n^{a'} B_n A_n^{b''}),$$

where

$$a' = \begin{cases} l-i & \text{if } i \le l \\ \\ n+l-i & \text{if } i > l, \end{cases} \qquad b' = \begin{cases} l-j & \text{if } j \le l \\ \\ n+l-j & \text{if } j > l, \end{cases}$$

$a'' = n - a'$ and $b'' = n - b'$.

Proof. The rotation of 90 degrees centered in the entry (i, j) can be obtained as follows: first consider the translation $\tau_{a',b'}$ that sends the entry (i, j) to the entry (l, l), then we can apply the rotation ρ and then we apply the translation $\tau_{a'',b''}$ which maps the entry (l, l) into the entry (i, j). The last part follows from (1) and (2). □

Combining the results of this section, we can show how the factors of a factorization become after applying translations and rotations.

Theorem 2.5 Let M be a real $n \times m$ matrix and $M = FG$ with F an $n \times r$ matrix and G an $r \times m$ matrix. Then for any integers $i \in \{0, \ldots, n\}$ and $j \in \{0, \ldots, m\}$ the translation $\tau_{i,j}$ satisfies the following

$$\tau_{i,j}(M) = (A_n^T)^i M A_m^j = F'G',$$

where $F' = (A_n^T)^i F$ and $G' = GA_m^j$. If $m = n$, the rotation of 90 degrees ρ satisfies the following

$$\rho(M) = M^T B_n = F'' G'',$$

where $F'' = G^T$ and $G'' = F^T B_n$.

Remark 2.6 It is still possible to describe the rotation of a non-square matrix using ρ. In fact, suppose that $M \in \mathbb{R}^{n\times m}$ with $n < m$ (one can argue in a similar way with $n > m$). Consider $\overline{M} \in \mathbb{R}^{m\times m}$ obtained from M as follows: the first n rows are those of M and the last $m-n$ rows are zero. If $M = FG$, with $F \in \mathbb{R}^{n\times r}$ and $G \in \mathbb{R}^{r\times m}$, then $\overline{M} = \overline{F}G$, where $\overline{F}$ is equal to F in the first n rows and then in the last $m-n$ is zero. We can now apply ρ to $\overline{M}$, which by Equation (2) is

$$\rho(\overline{M}) = \overline{M}^T B_m = G^T \overline{F}^T B_m.$$

By removing the first $m-n$ columns of $\rho(\overline{M})$ (which are all zeros because of the definition of $\overline{M}$) we thus find that its factorization is given by $G^T F'^T$, where F' is obtained from $B_m \overline{F}$ by deleting the first $m-n$ columns. Therefore, the action of ρ can be extended to non-square matrices.

These transformations are very interesting as they are Frobenius invariant. In fact, recall that the Frobenius norm of a matrix $M \in \mathbb{R}^{n\times m}$ is $\|M\|_F^2 = \sum_{i=1}^n \sum_{j=1}^m M_{ij}^2$.

Proposition 2.7 Let $M \in \mathbb{R}^{n\times m}$, then for any $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$ we have

$$\|\tau_{i,j}(M)\|_F = \|M\|_F,$$

and also

$$\|\rho(M)\|_F = \|M\|_F.$$

Proof. The statement follows from the fact that the maps $\tau_{i,j}$ and ρ permute the entries of the matrix M, therefore the sum of the square of the entries of M is left invariant. □

As a final remark, we observe that the maps $\tau_{i,j}$ and ρ preserve also some properties of the factors, such as the non-negativity.

3. An NMF-based data augmentation procedure

In this section, we outline a data augmentation procedure based on the results of Section 2. In particular, our idea is to generate new feature vectors from shifted versions of the left factors of an NMF of the starting dataset, which we assume to be composed of two-dimensional data.

Given a matrix $M \in \mathbb{R}^{n\times m}$ with non negative entries, the NMF of M provides a low rank approximation of M by finding two non negative matrices $W \in \mathbb{R}^{n\times r}$, $H \in \mathbb{R}^{r\times m}$ such that $M \approx WH$, for some $r < \min\{m, n\}$. Formulated in optimization terms, the factorization is obtained by solving the problem

$$\min_{W\geq 0, H\geq 0} J(M||WH), \tag{3}$$

where J is some divergence function that measures the dissimilarity between M and WH.

The measure function J is commonly chosen as the Frobenius norm of error, and this is indeed the norm we will consider in this paper. More precisely, $J(M||WH) = \|M - WH\|_F^2$. For a reference, see [14].

As a consequence of Theorem 2.5 and Proposition 2.7, we obtain the following result.

Theorem 3.1 Let M be a real $n \times m$ matrix with non negative entries and WH be an NMF of M with W a non negative $n \times r$ matrix and H a non negative $r \times m$ matrix. For any integer $i \in \{0, \ldots, n\}$ and $j \in \{0, \ldots, m\}$ an NMF of the translation $\tau_{i,j}(M)$ is given by

$$W'H' = (A_n^T)^i W H^T A_m^j.$$

If $m = n$, an NMF of the rotation of 90 degrees $\rho(M)$ is provided by

$$W''H'' = H^T W^T B_n.$$

Proof. Let $E = M - WH$ and note that, by Proposition 2.1,

$$\tau_{i,j}(E) = \tau_{i,j}(M) - \tau_{i,j}(WH)$$

and by Proposition 2.7,

$$\|\tau_{i,j}(E)\|_F^2 = \|E\|_F.$$

Therefore if W and H minimize the error function $\|M - WH\|_F^2$ subject to $W \geq 0$ and $H \geq 0$ then W' and H' minimize the error function $\|\tau_{i,j}(M) - W'H'\|$ subject to $W' \geq 0$ and $H' \geq 0$. The same holds for ρ. □

Clearly, any composition of translations and rotations will preserve the NMF and thanks to the above theorem we can compute explicitly an NMF of the obtained matrix.

Remark 3.2 Note also that these operations preserve the boundedness property, in the sense of the above proposition; see e.g. [15] Theorem 1 for the boundedness property.

Remark 3.3 Given a matrix $M \in \mathbb{R}^{n \times m}$ with no symmetries, and a corresponding NMF, we can obtain a set of $4mn$ matrices with an associated NMF for each of them, by applying all the possible combinations of translations and rotations.

At this point, we can outline our data augmentation procedure. We first observe that given a matrix $M \in \mathbb{R}^{n \times m}$ with non negative entries, an NMF of M provides a low rank approximation $M \approx WH$. In fact, if by $\mathbf{m}^1, \ldots, \mathbf{m}^m$ we denote the columns of M, then they can be written as the linear combination of the columns $\mathbf{w}^1, \ldots, \mathbf{w}^r$ of the factor $W \in \mathbb{R}^{n \times r}$ as follows.

$$\mathbf{m}^j \approx \sum_{i=1}^{r} \mathbf{w}^i h_{ij}, \quad j = 1, \ldots, m,$$

where the h_{ij}'s denote the entries of H. This implies that the column space of M is contained in the column space of W. In this perspective, the NMF is extracting some features of the original matrix that are represented by the information stored in the left factor W. This property has been extensively studied [8, 16–18].

Then, by applying all possible transformations to the factor W we can generate new feature vectors, resulting in a data augmentation procedure. More formally, given a dataset $\mathcal{D} = \{S_i, i = 1, ..., N\}$ where S_i's are factors W derived from the NMF factorizations of some non negative matrices in $\mathbb{R}^{n\times m}$, we choose as transformation operations T_j's the maps $\tau_{i,j}$ and ρ^l with $i \in \{0, \dots, n\}$, $j \in \{0, \dots, m\}$, and $l \in \{0, 1, 2, 3\}$, to create additional training data $S_i^j = T_j(S_i)$. The final dataset results

$$\mathcal{D}_A = \{S_i, S_i^j,\ i = 1, \dots, N,\ j = 1, \dots, K\}.$$

From Remark 3.3, we can get a new dataset of size at most $4mn|\mathcal{D}|$.

4. Numerical examples

This section presents an application of the proposed procedure. First of all, we give some numerical evidences of the previous results on a toy example, then we will show how these properties can be exploited for the task of data augmentation in a machine learning framework.

In the previous sections, we proved that non negative factorizations of translations or rotations of a given matrix can be obtained by linearity starting from suitable transformations applied to the non negative factors of the original matrix. In this way, once an NMF of the given data is known, using Theorem 3.1 one can obtain non negative factorizations of geometric transformations of the original data without repeating the algorithmic procedure. A tangible insight of the theoretical claims is presented in the following.

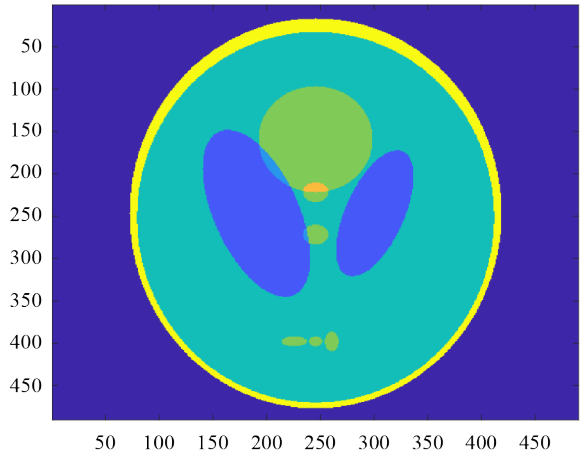


Figure 1. Original image of the Shepp-Logan phantom

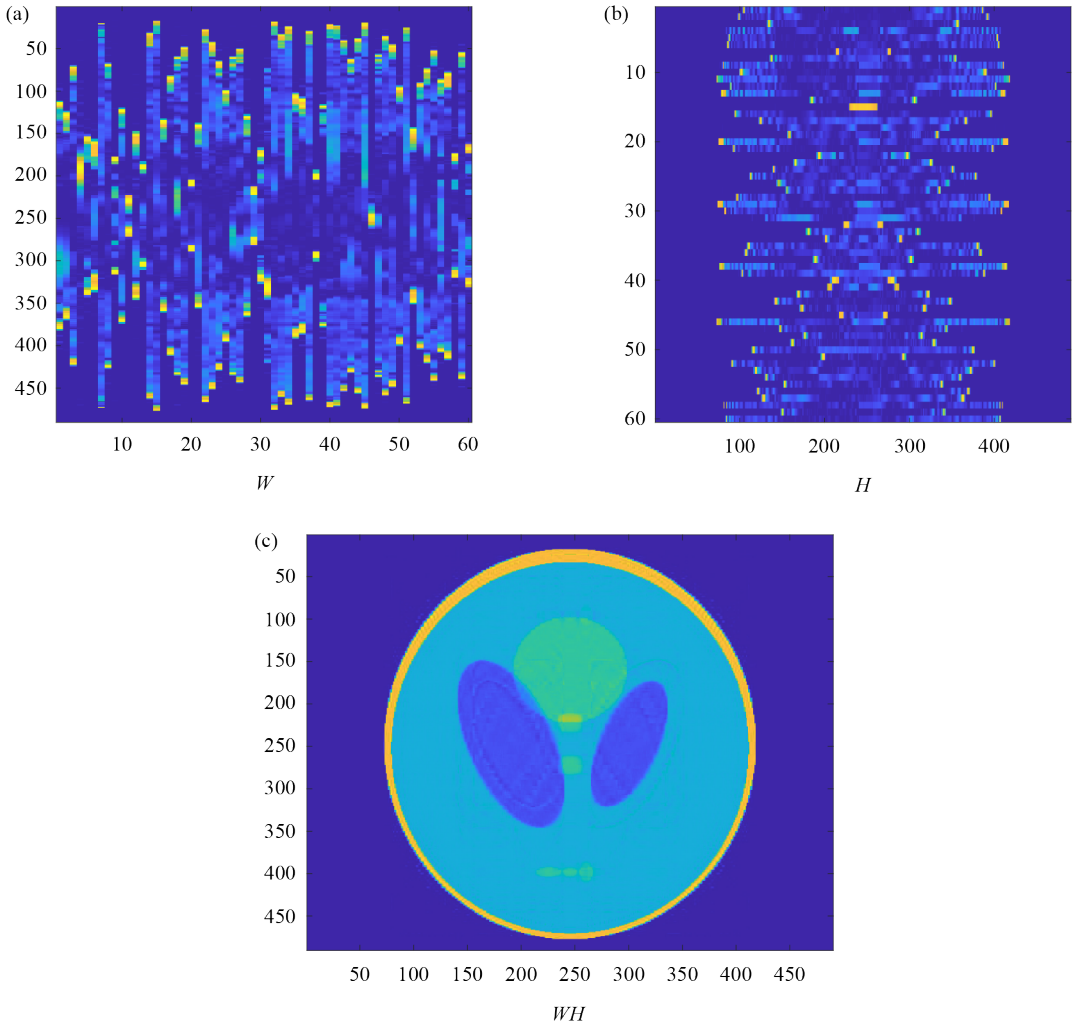
Algorithm 1 Alternating Non negative Least Squares (ANLS)
Input: M, W_0, H_0
$t \leftarrow 0$
while stopping condition is not satisfied **do**
  $H^{t+1} \leftarrow \arg\min_{H\geq 0}\|M - W^t H\|_F$
  $W^{t+1} \leftarrow \arg\min_{W\geq 0}\|M^T - H^{t+1^T} W^T\|_F$
  $t \leftarrow t + 1$
end while
Output: H^{t+1}, W^{t+1}.

Consider the 2D image M of size 490×490 shown in Figure 1, which is a standard test image known as the Shepp-Logan phantom. We computed an NMF of the image by means of the Alternating Non negative Least Squares (ANLS) method, which ensures the convergence to a stationary solution. A pseudocode of the method is given in Algorithm 1. In particular, to update the factors W and H we solve the least-squares subproblems by means of the Matlab built-in function lsqnonneg. Here, the choice of the rank of factorization ($r = 60$) is Singular Value Decomposition (SVD)-based and accounts for 80% of the largest singular values, following the selection rule proposed in [19]; W and H are initialized at random. The procedure ends when the relative change of the factors is less than a prefixed tolerance $\epsilon < 0$, i.e., when

$$\frac{\|W^{t+1} - W^t\|_F}{\|W^t\|_F} + \frac{\|H^{t+1} - H^t\|_F}{\|H^t\|_F} < \epsilon,$$

within a maximum number of 1,000 iterations. For this test, $\epsilon = 10^{-4}$ and the algorithm converges achieving the following relative reconstruction error in the Frobenius norm

$$\frac{\|M - WH\|_F}{\|M\|_F} \approx 0.08.$$



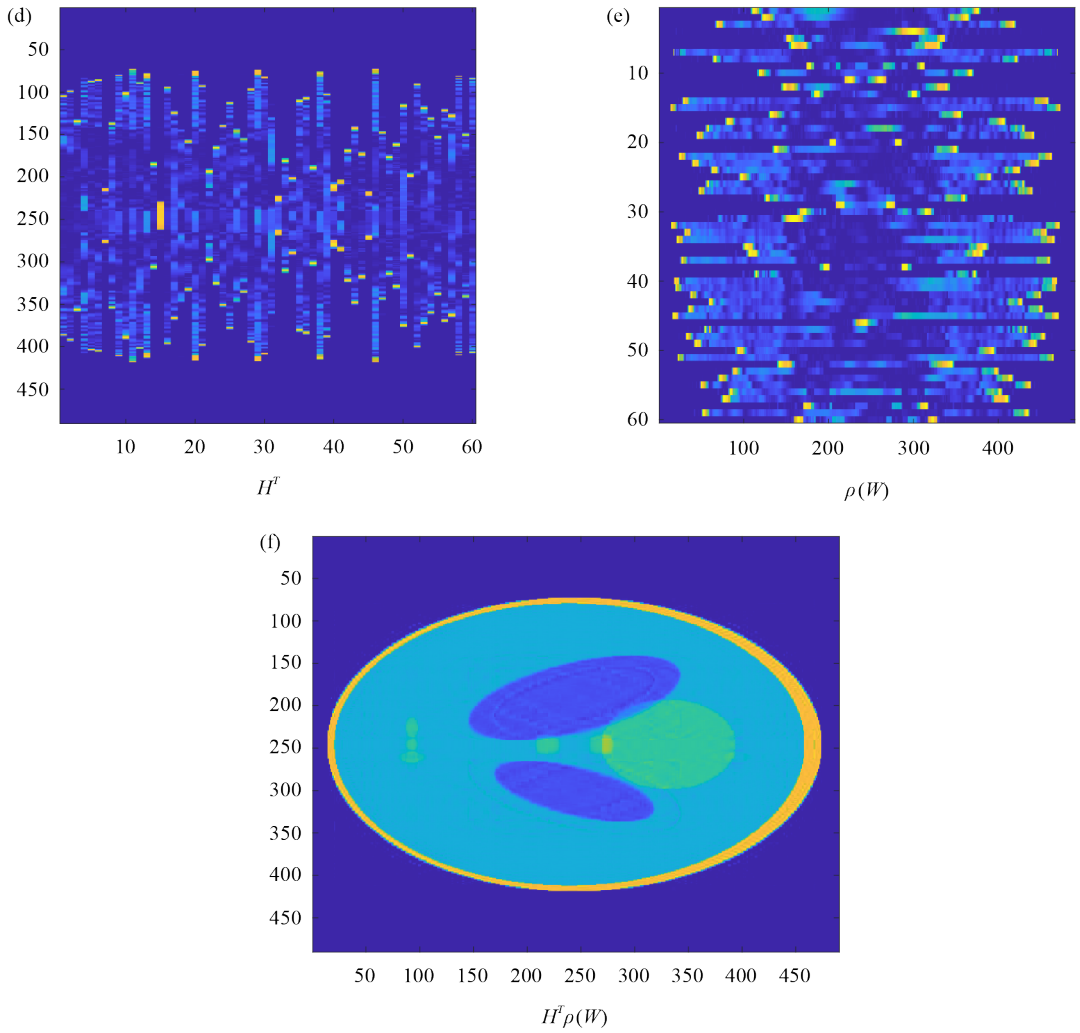


Figure 2. Example of geometric transformations on a 2D digital image

The two factors W and H obtained at the end of the procedure are shown in the Figure 2a, 2b together with the resulting low-rank approximation given by the product WH (Figure 2c). At this point, the product $H^T\rho(W)$, where $\rho(W)$ is a 90 degrees rotation of the factor W, provides an NMF approximation of $\rho(M)$, preserving the accuracy of the factorization, which will be the same as WH, in accordance with Theorem 3.1. The corresponding images are shown in the Figure 2d-f. To generate geometric transformations of the original image, translations are performed by applying circular shifts to the rows of W, implemented in Matlab using the circshift function. Rotations and reflections are achieved by multiplying W with a permutation or flipped identity matrix, e.g., fliprl(eye(...)), and the corresponding factor H is transformed accordingly. These operations are linear with respect to the factors, so they preserve non-negativity and allow efficient production of transformed datasets without recomputing the NMF. The computational complexity of these transformations is relatively low: for an image of size $n \times n$ and factor rank r, each circular shift or rotation of W requires $\mathcal{O}(n \cdot r)$ operations, while updating the corresponding H is a matrix multiplication of complexity $\mathcal{O}(r^2)$, which is negligible compared to the original NMF computation.

4.1 *Classification of glyph images*

In this subsection we apply our data augmentation technique in the context of binary classification of archaeological images by means of the supervised machine learning methodology known as Support Vector Machines (SVMs). The aim is to extract and analyse incisions and glyphs that are present on specific archeological sites, which may be difficult to reach or to move. Indeed, in archaeological research, the extraction of glyphs or petroglyphs from surfaces remains a common yet challenging task. As this procedure is generally performed manually, it is both labor-intensive and prone to inaccuracies. Another difficulty is the correct identification of inscriptions and glyphs. In order to use supervised machine learning tools for automatisation of the processes of classification and analysis, a suitable data set is needed for the training phase. In particular, our interest is directed towards the analysis of ancient graffiti found in Domus de Janas, a type of Neolithic tomb found in Sardinia, Italy [20, 21]. An example of such carvings is shown in Figure 3. The first step of our procedure consists in generating synthetic data that simulate the most important characteristics of the real surfaces of interest. As we can see in Figure 3, the main features of the shape are represented by spiral and jagged lines. Examples of simulated data are given in Figure 4. These data were obtained using software developed in [22], which provide a Matlab tool with a graphic interface to create three-dimensional realistic surfaces with petroglyphs of desired shape, position, and depths.

Figure 3. Engraving found in the Domus de Janas of Corongiu, in Pimentel (Sardinia, Italy) [21]

The panels in Figure 4 correspond to 2D representations of 3D surfaces, in the sense that each pixel value represents the distance of the corresponding point from the zero level.

In the current numerical test, the starting dataset is composed by the two matrices shown in Figure 4, both of size 599×574. For each of them, we computed an NMF approximation by means of the ANLS algorithm previously described. Here, the tolerance for the stopping criterion is set equal to 10^{-2}.

Let denote by M_1 and M_2 the matrices of square pixels representing, respectively, the images in Figure 4; moreover, let consider the corresponding NMF approximations $M_1 \approx W_1 H_1$ and $M_2 \approx W_2 H_2$, $W_1, W_2 \in \mathbb{R}^{n\times r}$, $H_1, H_2 \in \mathbb{R}^{r\times m}$, where the rank of factorization is selected by hand tuning as $r = 64$; the relative reconstruction errors (measured in the Frobenius norm) achieved with this choice are, respectively, 0.099 and 0.107, which means that the factorizations explain about 90% of the data, in both the cases. The starting dataset is composed of the factors $\{W_1, W_2\}$, with the corresponding labels $\{1, -1\}$. We considered the row circular shifts $W_1^{<i,0>}$ for a subset of row indices obtained by selecting $i = 1 : 10 : n$, thus providing a total number of 60 shifted version of the factor W_1, with label 1; the same translations are applied to W_2. The resulting dataset is composed of 120 couples

$$\mathcal{D} = \{(W_1^{<i,0>}, y_1^i), (W_2^{<i,0>}, y_2^i), \quad i = 1 : 10 : n\},$$

where $W_1^{<i,0>}$, $W_2^{<i,0>} \in \mathbb{R}^{nr}$ are the vectorized factors representing, respectively, the features of the associated images $M_1^{<i,j>}$, $M_2^{<i,j>}$, and y_1^i, $y_2^i \in \{1, -1\}$ are the corresponding labels. We randomly selected 80% of the elements of the dataset $\mathcal{D}$ for the training set, using the remaining 20% of data for testing.

Given the new dataset available, we apply the supervised machine learning methodology known as SVMs (see [23, 24]) with the aim of training a binary classifier that is able to automatically detect the presence of specific type of glyphs, that is, in our specific case, to distinguish between the presence of spirals and zigzag (like in Figure 4a) or spirals only (like in Figure 4b) whatever their are located within the surface. For the training phase, we considered the problem of minimizing the standard dual SVM formulation with Gaussian kernel, which we addressed by means of a gradient projection method with nonmonotone linesearch along the feasible direction and combined with a steplength selection strategy based on the adaptive alternation of Barzilai-Borwein-like rules introduced in [25].

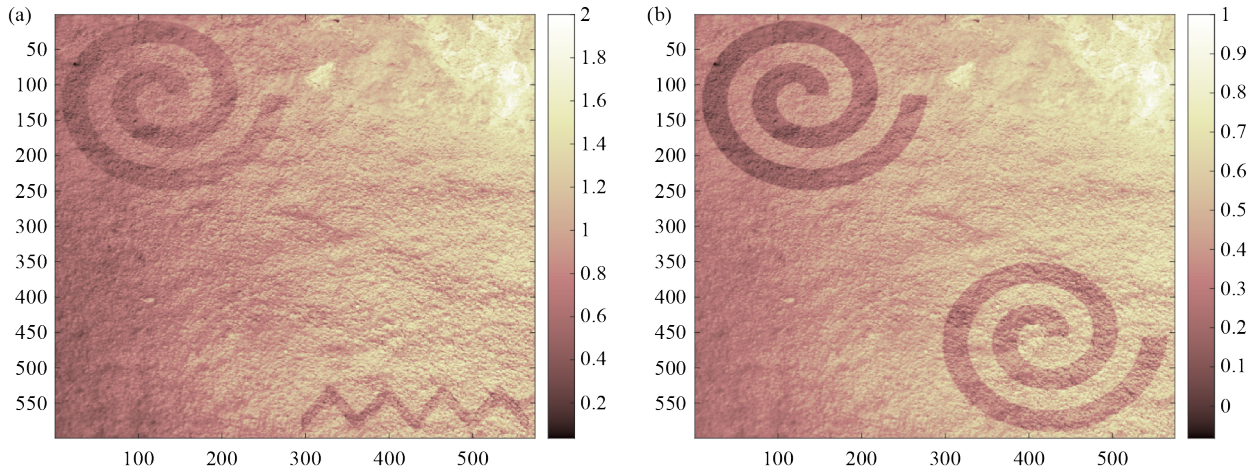


Figure 4. Synthetic images representing the starting dataset

The performance related to the quality of the prediction of the trained classifier are reported in Table 1 in terms of the performance scores given by Accuracy (Acc.), Sensitivity (Sens.), Precision (Prec.), and F1-score. These metrics are defined on the basis of the number of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) samples obtained in the testing phase, as follows:

- $\text{Acc.} = \dfrac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \cdot 100\%;$
- $\text{Prec.} = \dfrac{\text{TP}}{\text{TP}+\text{FP}} \cdot 100\%;$
- $\text{Sens.} = \dfrac{\text{TP}}{\text{TP}+\text{FN}} \cdot 100\%;$
- $\text{F1-score} = 2\,\dfrac{\text{precision} \cdot \text{sensitivity}}{\text{precision} + \text{sensitivity}} \cdot 100\%.$

Table 1. Performance scores of the classification obtained on testing set

Acc.	Prec.	Sens.	F1
100.0%	100.00%	100.00%	100.00%

In particular, the accuracy gives the number of correct predictions with respect to the total number of predictions, the precision provides the portion of positive samples that have been correctly predicted, while sensitivity computes the

number of actual positive samples with respect to the total number of positives (both true Positive and false negative) with the aim of identify the true positive that have been incorrectly predicted. Finally, F1-score is the harmonic means of sensitivity and precision.

We then performed further experiments, aimed at assessing the robustness of our approach with respect to the presence of noise, which is typical in real acquisitions, in order to account for random variations in the image caused by issues related to lighting, atmospheric conditions, or artifacts produced by acquisition devices. First of all, we computed the performance scores of the learned model on a testing set including factors arising from noisy images. To this aim, the original images (see Figure 4) were corrupted, respectively, by white gaussian noise with zero-mean and 0.01 variance, and by salt and pepper noise with noise density of 0.05, using the Matlab function imnoise. Then, the testing set used in the previous experiment was increased by 50% with data obtained as left shifted factors of the NMF of the corrupted images. The classification results on the new testing set, having trained the model on noise-free data, are shown in Table 2. As expected, we observe a decrease in the performance scores, which however still reveal acceptable generalization ability of the model on unseen data; moreover, since the two original images were corrupted by different types of noise, the sensitivity and precision scores might suggest that possible misclassifications are related to the presence of gaussian noise, which affects the class with positive labels.

Table 2. Performance scores of the classification obtained on a testing set with feature vectors arising from noisy images

Acc.	Prec.	Sens.	F1
82.50%	100.00%	74.07%	74.07%

To improve the generalization performance to account for noisy data, we generated a new dataset as follows. Starting from the original images M_1 and M_2, for each class we generated two additional synthetic images, changing shape, position, and depths of the petroglyphs; then, we perturbed each image with additive white Gaussian noise (zero mean, variance $\sigma = 0.01$ and 0.03), salt-and-pepper noise (noise densities 0.01, 0.03, 0.05, 0.08), and speckle noise (zero mean, variance $\sigma = 0.05$ and 0.08). For each of the 26 images thus obtained, we applied row circular shifts to the left factors of the corresponding NMF decompositions; this procedure results in a new augmented dataset of 1,560 samples, equally balanced between negative and positive samples. Notice that this final dataset incorporates a twofold augmentation, since noise injection represents another common technique for data augmentation. To evaluate model performance on this augmented dataset, we used the Matlab built-in functions fitcsvm and crossval. As cross-validation strategies, we considered k-fold with $k = 5$, holdout with 50% of the data used for training and 50% for testing, and leave-one-out, obtaining in all cases zero misclassification error, and 100% performance scores. As final assessment, we evaluated the classification ability on real-world images of the model trained on the augmented dataset. For this task, we considered the images illustrated in Figure 5, presenting: (Figure 5a) a section of a vase featuring engraved spiral motif, (Figure 5b) a petroglyph from the Fremont archaeological culture in Utah (U.S), (Figure 5c) a petroglyph from an archeological site in New Mexico (U.S), Figure 5d a petroglyph from the Chaco Culture National Historical Park in New Mexico. The images were first converted to grayscale and then resized to dimensions compatible with the model's input space. After that we evaluated the predicting ability of the model, obtaining the correct identification of the corresponding class.

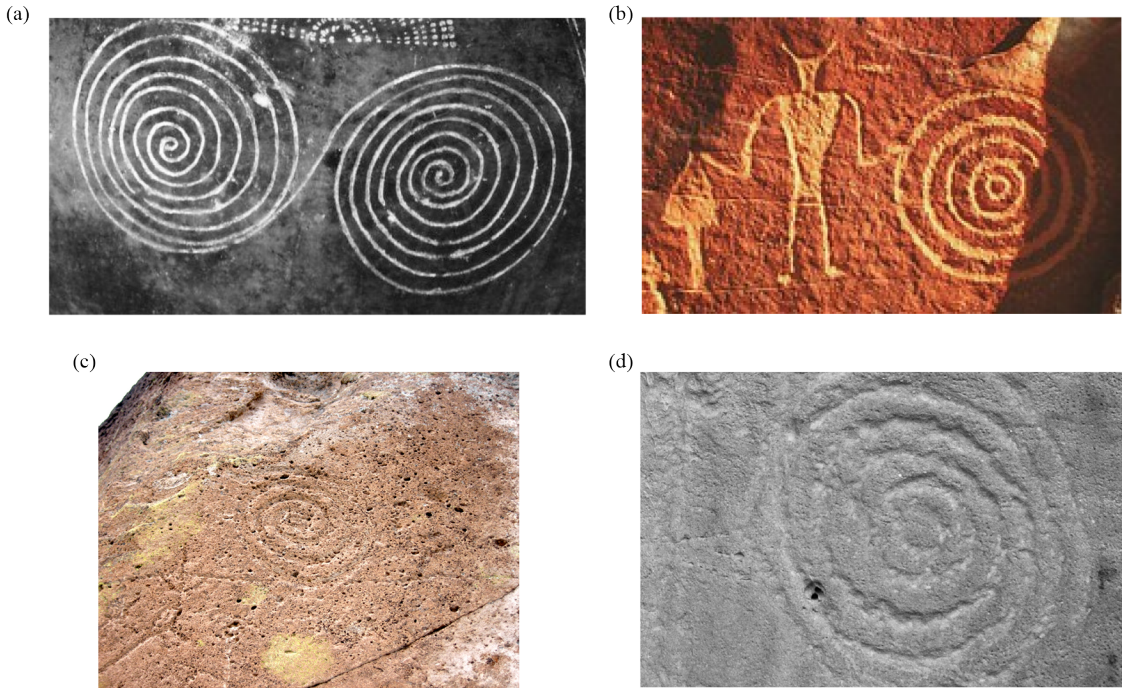


Figure 5. Real images of spiral petroglyphs from different sources

5. Conclusions

In this paper, we exploited the use of Non negative Matrix Factorization (NMF) for feature extraction for the task of data augmentation. By focusing on the columns of the left factors of the NMF decompositions, we were able to identify the main features inherent in the data. Based on that, we introduced a novel approach where linear transformations are applied to these factors, facilitating the simultaneous generation of a new example and its corresponding non negative factorization. Our methodology leverages the invariant properties of matrix decompositions under translations and rotations, ensuring robustness and consistency in the transformations applied. This approach not only enhances the interpretability of the extracted features but also preserve the nonnegativity of the data, which is crucial for various practical applications. The resulting data augmentation procedure is then tested on a binary classification problem in the framework of glyph/petroglyph identification. The proposed NMF-based data augmentation scheme is characterized by its feature-based approach, in which transformations are applied directly to the columns of the NMF decomposition rather than the raw images. This allows the method to generate new images while preserving the essential semantic features of the original data, ensuring interpretability and meaningful representations. Each augmentation simultaneously produces a new image and its corresponding factorization, which can be leveraged in neural network architectures to enhance feature diversity and improve generalization. By exploiting the invariance of NMF under translations and rotations, the scheme maintains structural integrity while enabling compact data representation, reducing storage and computational requirements compared to traditional augmentation techniques. However, the approach has some limitations: computing NMF on large datasets can be computationally intensive (although we compute it for a limited number of images), the effectiveness depends on the quality of the initial factorization, and it is inherently suited to non negative, additive data structures, which may restrict its applicability. Additionally, applying transformations in the factor space can sometimes introduce subtle artifacts if the relationship between factors and reconstructed images is not perfectly linear. Future work will explore the evaluation of this technique in accordance with state-of-the-art metrics for testing the effectiveness of data augmentation, and its application to more complex datasets. Furthermore, we will analyse the combination with other data dimensionality reduction methods to further improve performance and applicability in diverse domains.

Fundings

V. De Simone and F. Zullo acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 1409 published on 14.9.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union—NextGenerationEU—Project Title *A mathematical approach to inverse problems arising in cultural heritage preservation and dissemination*—CUP B53D23027900001-Grant Assignment Decree No. 1379 adopted on 1 September 2023 by the Italian Ministry of Ministry of University and Research (MUR). The research of F. Zullo was partially supported by the project COMBINE of of the program "Giovani Ricercatori" of University of Campania "L. Vanvitelli". S. Crisci aknowledges the support from the EU-FESR PON Ricerca e Innovazione 2014-2020, art. 24, comma3, lett. a) L. 240/2010 e s.m.i., D.M. 1062/2021 and the support of the project and the support by the project FEEDING of the program "Giovani Ricercatori" of University of Campania "L. Vanvitelli". This work has been partially supported by the Italian National research groups INdAM-GNCS and INdAM-GNSAGA.

Data availability

The datasets and codes generated during the current study are available from the authors on reasonable request. The tool used to generate the synthetic glyph dataset reported in Figure 4 and described in [22] can be found at the link https://sites.google.com/site/alessandobuccini/software. The images presented in Figure 5 are available as follows: panel (a)-online collection of the British Museum, 2025, Image ID: 262226001, licensed under CC BY-NC-SA 4.0; panel (b)-© J. Q. Jacobs, licensed under CC BY-SA 2.5, Wikimedia Commons; panel (c)-licensed under CC BY-SA 2.5, Wikimedia Commons, available at https://commons.wikimedia.org/wiki/File:Spiral_petroglyph_in_New_Mexico.jpg; panel (d)-public domain, National Park Service, available at https://www.nps.gov/media/photo/gallery-item.htm?gid=9DAC8A43-155D-451F-673D9AF1E702DBF9&id=9DAC8A9C-155D-451F-679EECBB4AC5B970.

Authors contribution

All the authors have contributed equally to this paper.

Ethical approval

This article does not contain studies with human.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Wang Z, Wang P, Liu K, Wang P, Fu Y, Lu CT, et al. A comprehensive survey on data augmentation. *arXiv:2405.09591*. 2024. Available from: https://doi.org/10.48550/arXiv.2405.09591.

[2] Xu M, Yoon S, Fuentes A, Park D. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*. 2023; 137: 109347. Available from: https://doi.org/10.1016/j.patcog.2023.109347.

[3] Lu X, Zheng B, Velivelli A, Zhai C. Enhancing text categorization with semantic-enriched representation and training data augmentation. *Journal of the American Medical Informatics Association*. 2006; 13(5): 526-535. Available from: https://doi.org/10.1197/jamia.M2051.

[4] Nanthini K, Sivabalaselvamani D, Chitra K, Gokul P, KavinKumar S, Kishore S. A survey on data augmentation techniques. In: *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*. Erode, India: IEEE; 2023. p.913-920. Available from: https://doi.org/10.1109/ICCMC56507.2023.10084010.

[5] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *Journal of Big Data*. 2019; 6(1): 1-48. Available from: https://doi.org10.1186/s40537-019-0197-0.

[6] Mumuni A, Mumuni F. Data augmentation: A comprehensive survey of modern approaches. *Array*. 2022; 16: 100258. Available from: https://doi.org/10.1016/j.array.2022.100258.

[7] Paatero P, Tapper U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*. 1994; 5(2): 111-126. Available from: https://doi.org/10.1002/env.3170050203.

[8] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999; 401: 788-791. Available from: https://doi.org/10.1038/44565.

[9] Fathi Hafshejani S, Moaberfard Z. Initialization for non-negative matrix factorization: A comprehensive review. *International Journal of Data Science and Analysis*. 2023; 16: 119-134. Available from: https://doi.org/10.1007/s41060-022-00370-9.

[10] Calvetti D, Somersalo E. *Mathematics of Data Science: A Computational Approach to Clustering and Classification*. SIAM; 2020.

[11] Gillis N. The why and how of nonnegative matrix factorization. *arXiv:1401.5226*. 2014. Available from: https://doi.org/10.48550/arXiv.1401.5226.

[12] Rice L, Wong E, Kolter JZ. Overfitting in adversarially robust deep learning. In: *ICML'20: Proceedings of the 37th International Conference on Machine Learning*. JMLR.org; 2020. p.8093-8104.

[13] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human level performance on imagenet classification. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE; 2019. p.1026-1034. Available from: https://doi.org/10.1109/ICCV.2015.123.

[14] Lu J. Matrix decomposition and applications. *arXiv:2201.00145*. 2022. Available from: https://doi.org/10.48550/arXiv.2201.00145.

[15] Wang YX, Zhang YJ. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*. 2013; 25(6): 1336-1353. Available from: https://doi.org/10.1109/TKDE.2012.51.

[16] Li T, Ma S. IFD: Iterative feature and data clustering. In: *Proceedings of the 2004 SIAM International Conference on Data Mining*. SIAM; 2004. p.472-476. Available from: https://doi.org/10.1137/1.9781611972740.49.

[17] Lazar C, Doncescu A. Non negative matrix factorization clustering capabilities; application on multivariate image segmentation. In: *2009 International Conference on Complex, Intelligent and Software Intensive Systems*. Fukuoka, Japan: IEEE; 2009. p.924-929. Available from: https://doi.org/10.1109/CISIS.2009.190.

[18] Ding C, He X, Simon HD. On the equivalence of nonnegative matrix factorization and spectral clustering. In: *Proceedings of the 2005 SIAM International Conference on Data Mining*. SIAM; 2005. p.606-610. Available from: https://doi.org/10.1137/1.9781611972757.70.

[19] Qiao H. New SVD based initialization strategy for non-negative matrix factorization. *Pattern Recognition Letters*. 2015; 63: 71-77. Available from: https://doi.org/10.1016/j.patrec.2015.05.019.

[20] Vanzi M, Mannu C, Dessì R, Rodriguez G, Tanda G. Photometric stereo for 3D mapping of carvings and relieves: Case studies on prehistorical art in Sardinia. In: *XVII MAÇÃO's International Rock Art Seminar*. Macao, Portugal; 2014. p.1-8.

[21] Crabu E, Pes F, Rodriguez G, Tanda G. Ascertaining the ideality of photometric stereo datasets under unknown lighting. *Algorithms*. 2023; 16(8): 375. Available from: https://doi.org/10.3390/a16080375.

[22] Buccini A, Crabu E. A tool for the construction of synthetic petroglyphs. In: *Computational Science and Its Applications—ICCSA 2025 Workshops*. Cham: Springer; 2026. p.417-429. Available from: https://doi.org/10.1007/978-3-031-97638-4_26.

[23] Cortes C, Vapnik V. Support vector network. *Machine Learning*. 1995; 20: 273-297. Available from: https://doi.org/10.1007/BF00994018.

[24] Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*. 1998; 2(2): 121-167. Available from: https://doi.org/10.1023/A:1009715923555.

[25] Crisci S, Porta F, Ruggiero V, Zanni L. Spectral properties of Barzilai-Borwein rules in solving singly linearly constrained optimization problems subject to lower and upper bounds. *SIAM Journal on Optimization*. 2020; 30(2): 1300-1326. Available from: https://doi.org/10.1137/19M1268641.